Mémoire pour soutenir l'Habilitation à Diriger des Recherches

Nicolas Petit

Janvier 2007

Table des matières

Résumé

Ce document contient une présentation de mes activités scientifiques dans le domaine du contrôle des systèmes et plus particulièrement des procédés. Les activités scientifiques sont détaillées : thèmes de recherche, publications, encadrement, enseignement, collaborations industrielles. En annexe sont reproduites certaines publications.

1 Formation et emploi

- depuis 2001 Maître-Assistant à l'École Nationale Supérieure des Mines de Paris.
  Centre Automatique et Systèmes.

- 2000-2001 Postdoctoral Scholar au California Institute of Technology
  Control and Dynamical Systems. Pasadena, California USA.
  Bourse postdoctorale INRIA.

- 1996-2000 Docteur de l'École des Mines de Paris
  Spécialité Mathématiques et Automatique. Centre Automatique et Systèmes.
  Directeur de thèse Pierre Rouchon. Mention très honorable avec félicitations du jury.
  Systèmes à retards. Platitude en génie des procédés et contrôle de certaines équations des ondes.

- 1995-1996 DEA à l'Université d'Orsay
  Automatique et Traitement du Signal. Mention très bien.

- 1992-1995 Ancien élève de l'École Polytechnique promotion X92

2 Activités scientifiques

2.1 Présentation

Les questions clefs de la commande en génie des procédés s'articulent autour des notions de boucle ouverte et boucle fermée. On peut ainsi chercher à résoudre les problèmes de transitions entre deux modes opératoires ou points stationnaires d'une part et les problèmes de stabilisation autour de points de fonctionnement ou de trajectoires. Classiquement, les problèmes de stabilisation ont une importance prédominante pour les applications. C'est lorsqu'on cherche des performances accrues qu'on commence à chercher à résoudre les problèmes de transition entre points de fonctionnement. Dans ce dernier cadre, on aboutit naturellement aux méthodes de la commande optimale. Enfin, une grande difficulté en pratique concerne le manque de mesures pertinentes dû à l'impossibilité d'utiliser des capteurs adéquats ou dû aux bruits de mesures importants présents.

J'ai cherché à approfondir certaines questions de la commande des procédés

- Stabilité et stabilisation
- Transition entre deux points de fonctionnement
- Commande optimale
- Estimation de variables non-mesurées
- Modélisation

J'ai étudié ces questions sur la base des procédés industriels (ATOFINA, IFP, APPRYL, TOTAL) suivants

- Mélangeuses de carburants
- Réacteur de polymérisation type continu parfaitement agité
- Réacteur de polymérisation type continu à écoulement piston
- Moteurs Diesel HCCI
- Puits pétroliers activés en gas-lift
- Unité d'alkylation à acide sulfurique

2.2 Stabilité et stabilisation

2.2.1 Étude des instabilités dans les puits pétroliers activés en gas-lift

Un puits de pétrole est dit éruptif dès lors que la pression dans son réservoir est suffisante pour soulever le poids de la colonne d'huile qu'il contient. Cependant dans beaucoup de cas la pression dans le réservoir n'est pas assez élevée, un remède est de recourir au gas-lift pour permettre la production. Cette technique consiste à injecter du gaz haute pression en fond de puits. Le gaz est introduit à la surface, comprimé dans un volume tampon (casing), il descend le long du puits avant de pénétrer dans la partie centrale (tubing) qui contient l'huile. Le gaz en pénétrant dans le tubing va créer un mélange plus léger dont le poids sera inférieur aux forces de pression. En pratique, l'introduction d'un volume de gaz tampon et l'injection de gaz dans l'huile induisent de nouvelles complexités et de possibles instabilités dans la dynamique du procédé. Peu de gaz étant disponible, l'allocation de gaz maximisant la production tout en minimisant le gaz peut contraindre certains puits à être opérés dans des zones où la production est instable. Or dans cette zone non seulement les puits sont soumis à des contraintes liées aux oscillations de fortes amplitudes pour lesquelles leurs équipements n'ont pas été conçus mais en plus leur production est en moyenne beaucoup plus faible que celle qu'ils auraient obtenus en régime stabilisé. Enfin tous les puits étant couplés via un réseau de gaz et un réseau d'huile produite il est important de maintenir les oscillations dans des limites raisonnables de peur de voir les instabilités se propager.

Principalement, deux catégories d'instabilités sont présentes. La première, le casing-heading, se caractérise par des oscillations de la pression au fond du puits de très grande amplitude, de l'ordre de 20 bar pour une valeur nominale de 70 et par une production par à coups. Elle trouve son origine dans l'interaction de la colonne d'huile avec le volume de gaz tampon. Si le casing-heading se caractérise par une injection intermittente de gaz entre le casing et le tubing il ne faut cependant pas en conclure trop rapidement que maintenir cette injection constante suffit à garantir la stabilité du puits. En effet il existe un autre phénomène oscillatoire, la density-wave instability. Cette instabilité est localisée dans le tubing et provient du déphasage entre l'injection d'huile et la pression de fond. Un premier modèle de la littérature (voir [EIF03]) permet de simuler le premier type d'instabilités. Nous avons développé un modèle permettant de mettre en évidence les mécanismes du second type d'instabilités. Nous avons analysé ces deux instabilités et montré dans [SPL$^+$05a] qu'elles se ramenaient dans le premier cas à une bifurcation type Hopf dont l'analyse peut être menée grâce au théorème de Poincaré-Bendixon dans un plan de phases. Les difficultés d'applications de ce théorème sont ici de deux ordres. En premier lieu, il nous a fallu déterminer un compact rentrant, c.-à-d. un ensemble positivement invariant par la dynamique du système. Ceci a pu être fait en invoquant des arguments physiques spécifiques. D'autre

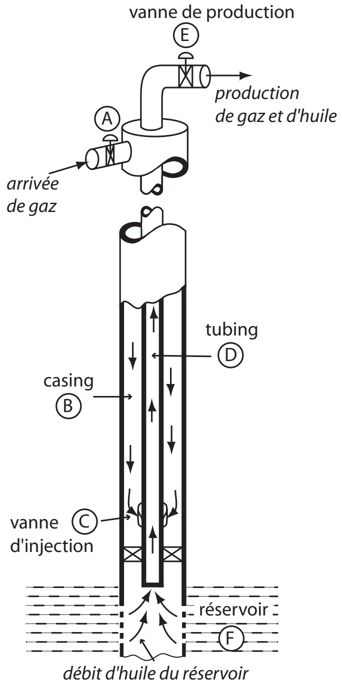


FIG. 1 – Schéma d'un puits activé en gas-lift

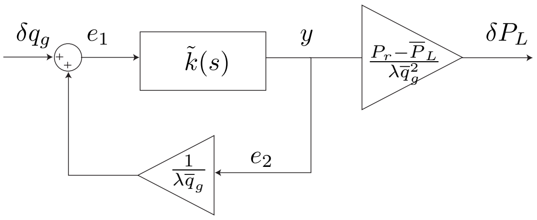
part, nous avons dû établir l'existence et l'unicité des trajectoires de ce système qui comporte de nombreux sauts de dynamique correspondant à l'ouverture et à la fermeture des vannes.

Dans le deuxième cas, nous avons montré que le phénomène trouvait son origine dans la contraction volumique issue du phénomène de propagation le long du tubing. Cette étude de décompose en deux étapes. Tout d'abord, nous montrons sous quelles hypothèses la dynamique de propagation peut se ramener par l'intermédiaire d'un invariant de Riemann à un système à retard couplé à une condition frontière non linéaire . Ensuite, nous étudions l'équation caractéristique associée et la localisation de ses racines suivant la valeur de certains paramètres physiques. Nous nous sommes aperçu que la quantité de gaz injecté peut être assimilé à un retard τ et que la stabilité du système dépend de la localisation des racines de son équation caractéristique dont la forme est donnée par

$$s = a + be^{-s\tau} + \frac{c}{s\tau}(1 - e^{-s\tau})$$

Nous avons pu mettre en évidence théoriquement l'existence d'une injection de gaz limite en dessous de laquelle le système présente une instabilité de type "density-wave". On trouvera l'analyse détaillée dans notre publication [SPM05].

La stabilisation est un autre point intéressant pour ce système. Comme nous l'avons dit il existe plusieurs types d'instabilités. Parmi lesquelles le casing-heading. Ce phénomène est maintenant bien connu, il existe beaucoup de travaux dans la littérature à son sujet et de nombreux remèdes existent. Ainsi il est possible d'avoir recours à une solution de contrôle en ligne, du type de celle que nous avons proposées [SPL$^+$05b, SPSP06] ou de procéder à des modifications de l'équipement du puits de façon à diminuer le couplage tubing/casing. Une modification consiste par exemple à remplacer la vanne d'injection par une vanne de diamètre plus petit ou par une vanne NOVA qui garantit un débit critique, c'est-à-dire un débit qui ne dépend que de la pression amont. En revanche pour la density-wave, la stratégie est autre. Il suffit d'injecter un ratio constant d'huile et de gaz en fond de puits pour assurer la stabilité.



δq_g
e_1
$+$
$+$
$\tilde{k}(s)$
y
$\frac{P_r - \overline{P}_L}{\lambda \overline{q}_g^2}$
δP_L
$\frac{1}{\lambda \overline{q}_g}$
e_2

FIG. 2 – Schéma bloc du sous-système tubing linearisé autour d'un point d'équilibre. Le bouclage positif correspond aux effets gravitationnels de la colonne de fluide non homogène. L'entrée, δq, correspond à l'injection de gaz et la sortie, δP_L, à la pression de fond.

Nous nous sommes ensuite intéressés à l'impact de certains autres paramètres physiques sur la stabilité du système ("casing-heading" et "density-wave"). Pour ce faire nous avons linéarisé autour d'un point de fonctionnement et calculé la fonction de transfert du système. Les deux instabilités apparaissent alors comme des boucles internes instables. La "density-wave" correspond au bouclage interne au tubing. L'évolution de la pression de fond dépend du poids de la colonne et donc de sa constitution. Or, cette constitution dépend de ce qui a été injecté dans le tubing précédemment (effet mémoire). La colonne d'huile joue le rôle d'une mémoire finie. Le casing-heading provient du couplage entre le casing et le tubing. Les figures 2 et 3 mettent en évidence ces bouclages.

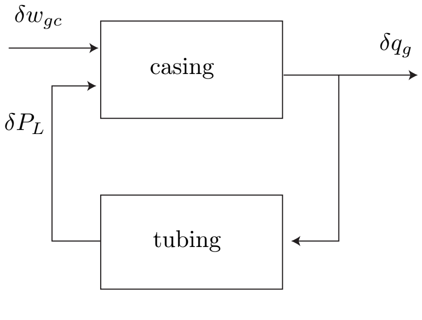


FIG. 3 – Schéma bloc du système interconnecté linéarisé autour d'un point d'équilibre. L'entrée, δ_{gc}, correspond à l'injection de gaz en tête de puits et la sortie, δq_g, à la quantité de gaz qui rentre dans le tubing.

Le théorème des petits gains (voir par exemple [Kha92]) nous donne une condition de stabilité. Il faut que le produit des gains du tubing et du casing soit inférieur à 1. Ainsi on montre que :

- Augmenter la pression dans le réservoir à tendance à stabiliser le puits.
- Un indice de productivité important est un facteur d'instabilité.
- Un débit de gaz élevé peut être nécessaire au maintient de la stabilité
- Une ouverture de vanne d'injection trop importante causera un casing-heading

On trouvera plus de détails dans notre publication [SPM06].

[EIF03] G. O. Eikrem, L. Imsland, and B. Foss. Stabilization of gas-lifted wells based on state estimation. In *Proc. of the ADCHEM 2003, International Symposium on Advanced Control of Chemical Processes*, 2003.

[Kha92] H. K. Khalil. *Nonlinear Systems.* MacMillan, 1992.

[SPL+05a] L. Sinègre, N. Petit, P. Lemétayer, P. Gervaud, and P. Ménégatti. Casing-heading phenomenon in gas-lifted well as a limit cycle of a 2d model with switches. In *Proc. of the* 16*th IFAC World Congress*, 2005.

[SPL+05b] L. Sinègre, N. Petit, P. Lemétayer, P. Gervaud, and P. Ménégatti. Contrôle des puits activés en gas-lift. In *10ème Congrès de la Société française de génie des procédés*, 2005.

[SPM05] L. Sinègre, N. Petit, and P. Ménégatti. Distributed delay model for density wave dynamics in gas lifted wells. In *Proc. of the* 44*th IEEE Conf. on Decision and Control*, 2005.

[SPM06] L. Sinègre, N. Petit, and P. Ménégatti. Predicting instabilities in gas-lifted wells simulation. In *Proc. of the 2006 American Control Conference*, 2006.

[SPSP06] L. Sinègre, N. Petit, and T. Saint-Pierre. Active control strategy for density-wave in gas-lifted wells. In *Proc. of the ADCHEM 2006, International Symposium on Advanced Control of Chemical Processes*, 2006.

2.2.2 Limite des régulateurs PI sur les systèmes à retards variables

Les systèmes à retards variables représentent un véritable défi pour le contrôle des procédés industriels. Mathématiquement, leur stabilisation est un problème difficile, et en pratique la mise en place d'un régulateur robuste capable de gérer le retard variable est compliquée. En dépit des avancées des techniques du contrôle des procédés, la simplicité et la robustesse du contrôleur PID en font un choix très fréquent, en particulier dans les raffineries. De nombreuses méthodes de réglage existent lorsque les modèles présentent un retard plus ou moins important. Les régulateurs de type "prédictif" prennent en compte les retards purs présents dans les modèles, mais ils ne sont pas spécifiquement adaptés aux retards variables. Le prédicteur de Smith [Smi58] est un de ces régulateurs qui permet d'accroître les performances obtenues lorsqu'on connaît précisément le retard. Il souffre, comme tous les contrôleurs utilisés sur les modèles à retard constant, d'une forte sensibilité face à l'identification et la variabilité du retard. En contrôle monovariable, les performances des prédicteurs sont généralement décevantes et poussent à l'utilisation de contrôleurs plus basiques de type PID dont on réduit les gains, les résultats sont alors peu séduisants, mais la robustesse est assurée dans la plupart des cas. Une des principales causes de la variabilité du retard est le transport de la charge dans un volume constant lorsque le débit de la charge fluctue.

Dans la publication [BCP04], nous avons évalué les performances des contrôleurs PI sur la classe de système

$$G(s) = \frac{Ke^{-\delta s}}{\tau s + 1}\,, \quad G_c(s) = K_c \left(1 + \frac{1}{sT_i}\right)$$

lorsque le contrôleur est réglé par les méthodes classiques (Ziegler-Nichols [ZN42, HÅH91], Cohen-Coon [CC53]) et par la méthode plus récente (Tavakoli-Fleming [TF03]). Cette évaluation a été conduite sur un modèle d'unité de déhydrodésulfuration (HDS) telle qu'on la trouve dans pratiquement toutes les raffineries (voir figure 4). Le retard est considéré comme variable et nous avons étudié les réponses en poursuite et en rejet, ce qui nous a permis de montrer un réel besoin dans des méthodes plus efficaces lorsque le retard est variable.

À la recherche d'une solution à ce problème nous avons proposé un contrôleur de type "prédicteur adaptatif" dans [BCP05]. Le prédicteur adaptatif s'avère plus stable qu'un prédicteur de Smith classique ; ses performances sont plus élevées que celles des contrôleurs classiques.

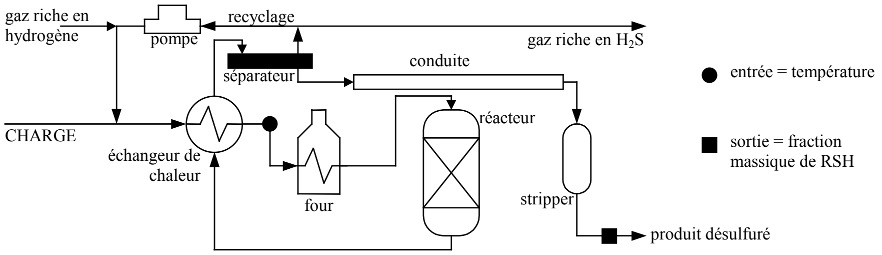


FIG. 4 – Schéma du procédé HDS. Des non-linéarités fournissent un gain et une constante de temps variable, le transport par la tuyauterie donne un retard variable.

[BCP04] J. Barraud, Y. Creff, and N. Petit. PI controllers performance for a process model with varying delay. In *Proc. of UKACC Int. Control Conference*, 2004.

[BCP05] J. Barraud, Y. Creff, and N. Petit. Performances d'un prédicteur de Smith adaptatif pour un modèle de procédé avec retard variable. In *10ème Congrès de la Société française de génie des procédés*, 2005.

[CC53] G. H. Cohen and G. A. Coon. Theoretical consideration of retarded control. *Trans. A.S.M.E.*, Vol. 75(No. 1) :pp. 827–834, 1953.

[HÅH91] C. C. Hang, K. J. Åström, and W. K. Ho. Refinements of the Ziegler-Nichols tuning formulas. *IEE Proceeding-D*, Vol. 138(No. 2) :pp. 111–118, 1991.

[Smi58] O. J. M. Smith. Closer control of loops with dead time. *Chemical Engineering Progress*, 53(5) :217–219, 1958.

[TF03] S. Tavakoli and P. Fleming. Optimal tuning of PI controllers for first order plus dead time/long dead time models using dimensional analysis. *Proc. of the* 7*th European Control Conf.*, (2003).

[ZN42] J. G. Ziegler and N. B. Nichols. Optimum settings for automatic controllers. *Trans. A.S.M.E.*, Vol. 64 :pp. 759–765, 1942. Available from www.driedger.ca.

2.3 Transition entre deux points de fonctionnement

2.3.1 Platitude des systèmes

Les transitoires sont un enjeu important en commande de procédés pour des raisons économiques en général. Ce point est détaillé dans les exemples qui suivent. Dans le cadre des systèmes d'équations différentielles ordinaires linéaires, on peut résoudre les problèmes de planification de trajectoires lorsque le système considéré possède la propriété de commandabilité. Cette propriété est équivalente à l'existence d'une sortie de Brunovsky [Bru70] permettant de mettre le système sous forme contrôleur. Dans le cadre non-linéaire qui nous intéresse, on peut également donner un sens à ces propriétés mais elles ne sont pas équivalentes. Pour construire le contrôleur, si c'est possible, on peut chercher à se ramener à une forme canonique contrôleur également mais par des **changements non-linéaires de coordonnées et des bouclages**. Nos investigations portent sur ces transformations non-linéaires. Dans ce cadre, une notion clef est la platitude [FLMR95, FLMR99] et ses extensions [Mou95, RM98].

Définition 1 ([FLMR95, FLMR99] Système plat)**.** On dit que le système défini par

$$\dot{x} = f(x,u), x \in \mathbb{R}^n, u \in \mathbb{R}^m$$

est **plat** s'il existe une application $h : \mathbb{R}^n \times (\mathbb{R}^m)^{r+1} \mapsto \mathbb{R}^m$, une application $\phi : (\mathbb{R}^m)^r \mapsto \mathbb{R}^n$ et une application $\psi : (\mathbb{R}^m)^{r+1} \mapsto \mathbb{R}^m$ telles qu'on puisse écrire :

$$\begin{aligned} y =& h(x,u,\dot{u},\ldots,u^{(r)}) \\ x =& \phi(y,\dot{y},\ldots,y^{(r-1)}) \\ u =& \psi(y,\dot{y},\ldots,y^{(r-1)},y^{r}). \end{aligned}$$

Tout le comportement dynamique du système est résumé par le comportement de sa sortie plate : toutes les trajectoires sont de la forme

$$\begin{aligned} x(t) =& \phi(y(t),\dot{y}(t),\ldots,y^{(r)}(t)) \\ u(t) =& \psi(y(t),\dot{y}(t),\ldots,y^{(r+1)}(t)) \end{aligned}$$

où r est un entier.

De manière générale il n'est pas facile de trouver des trajectoires d'un système donné. En effet une application quelconque $t \mapsto (x(t),u(t))$ n'est en général pas solution de $\dot{x} = f(x,u)$ c.-à-d. ne satisfait pas

$$\dot{x}(t) = f(x(t),u(t)).$$

En revanche pour un système dont on connaît une sortie plate y toutes les trajectoires sont de la forme

$$\begin{aligned} x(t) =& \phi(y(t),\dot{y}(t),\ldots,y^{(r)}(t)) \\ u(t) =& \psi(y(t),\dot{y}(t),\ldots,y^{(r+1)}(t)) \end{aligned}$$

où r est un entier différent suivant les cas. N'importe quelle fonction du temps $[0,T] \ni t \mapsto y(t)$ fournit une trajectoire du système $[0,T] \ni t \mapsto (x(t),u(t))$ (voir figure 5). On dit qu'il y a une correspondance bi-univoque entre les trajectoires du système et celles des sorties plates.

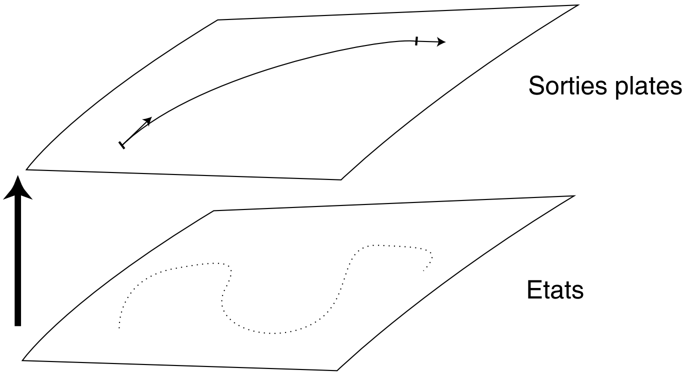


FIG. 5 – Équivalence des systèmes - Correspondance entre les trajectoires

[Bru70] P. Brunovsky. A classification of linear controllable systems. *Kibernetica*, 3 :173–187, 1970.

[FLMR95] M. Fliess, J. Lévine, P. Martin, and P. Rouchon. Flatness and defect of nonlinear systems : introductory theory and examples. *Int. J. Control*, 61(6) :1327–1361, 1995.

[FLMR99] M. Fliess, J. Lévine, P. Martin, and P. Rouchon. A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems. *IEEE Trans. Automat. Control*, 44 :922–937, 1999.

[Mou95] H. Mounier. *Propriétés structurelles des systèmes linéaires à retards : aspects théoriques et pratiques.* PhD thesis, Université Paris Sud, Orsay, 1995.

[RM98] M. Rathinam and R. M. Murray. Configuration flatness of Lagrangian systems underactuated by one control. *SIAM J. Control Optimization*, 36(1) :164–179, 1998.

2.3.2 Réacteur de polymérisation APPRYL PP2

Le réacteur APPRYL PP2 situé à Lavéra (Bouches du Rhône) est le cœur d'une unité de fabrication de polypropylène (PP). Ce réacteur produit environ $250kT/an$ de polypropylène. Il peut être considéré comme un réacteur parfaitement agité où le retard agit sur la commande en raison de la dynamique d'activation du catalyseur.

La prise en compte du retard et de la non-linéarité grâce à la platitude du système a permis de concevoir un régulateur très performant doté par construction de très bonnes performances dynamiques (temps de réponse court et pas de dépassement).

Ce régulateur est en service depuis juillet 1999. Ces travaux ont fait l'objet des publications [PRB$^+$02, PRB$^+$00]. La publication [PRB$^+$02] a recu un prix (voir Section 2.12).

La marche de l'unité est caractérisée par deux grandeurs : le débit de production et le melt-index (appelé aussi indice de fluidité ou grade). Cette dernière grandeur est une mesure des caractéristiques mécaniques du polymère produit ; elle est d'une grande importance pour les applications de mise en forme, de soufflage, *etc* qui interviennent dans la fabrication de pièces pour l'industrie automobile, d'emballages de produits cosmétiques, *etc*. La marche de l'unité est organisée en fonction d'un planning qui dépend lui-même du

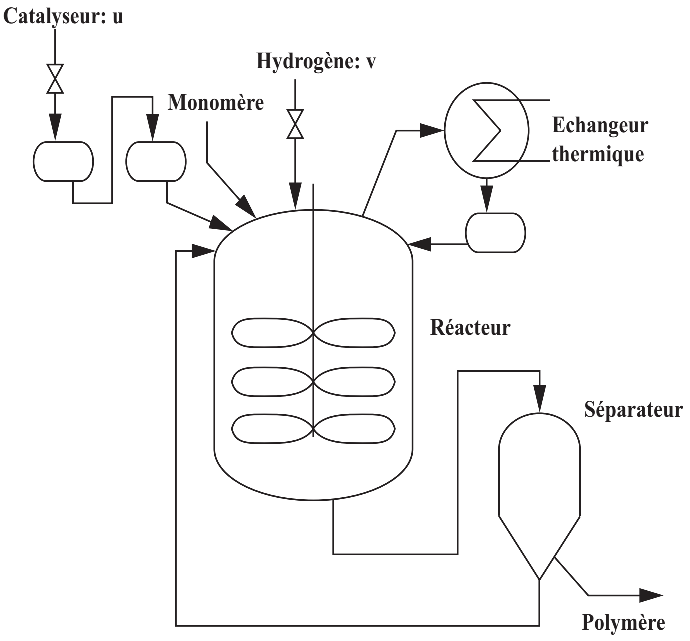


Fig. 6 – Schéma du procédé du réacteur APPRYL PP2

marché des polymères. Ceci implique de fréquents changements de point de fonctionnement. On souhaite réaliser les transitions entre les points de fonctionnement à l'aide d'un dispositif automatique de commande afin de gagner en précision et en rapidité d'éxécution par rapport au fonctionnement manuel.

La réaction de polymérisation se produit dans le réacteur représenté sur la figure 6. Le monomère se polymérise en présence du catalyseur (mud). Les terminaisons des chaînes se produisent grâce à l'hydrogène présent dans le réacteur. Le melt-index du produit est fonction de la concentration d'hydrogène dans le réacteur. L'exothermicité de la réaction est compensée par un dispositif d'évacuation thermique. À la sortie du réacteur les particules solides (polypropylène) sont séparées du liquide (propylène essentiellement) qu'on réintroduit dans le réacteur.

On dispose de deux commandes : le débit entrant de catalyseur et le débit entrant d'hydrogène. Notons qu'une variation de débit de catalyseur intervient dans le réacteur avec un certain retard, à cause de la présence de différents dispositifs d'activation du catalyseur. Ceci introduit un retard dans le modèle un constitué d'équations bilans.

En raison de la difficulté de la manœuvre, le changement de point de fonctionnement prend plusieurs heures. L'un des objectifs principaux du dispositif de contrôle est de permettre une réduction substantielle des temps de transition et de contrôler la qualité du produit durant la transition. On doit éviter les oscillations avant, pendant et après la transition. En outre on ne doit pas faire de dépassement (overshoot) lors des transitions.

Les dynamiques en jeu impliquent des phénomènes physiques et chimiques complexes. De simples régulateurs type PID peuvent être installés pour stabiliser le système autour d'un point stationnaire mais ne peuvent garantir de bonnes performances dynamiques entre deux points de fonctionnement.

En examinant un modèle du procédé à base d'équations bilan (voir [PRB$^+$02]), on s'paerçoit qu'il y a une relation biunivoque entre les trajectoires du système et les trajectoires des sorties plates qui sont le Melt-Index (MI) et le taux volumique de solide. On calcule les commandes boucle ouverte par des relations de platitude (en pratique on rajoute une boucle fermée utilisant ces trajectoires de référence). Par exemple, entre deux points stationnaires différents, on construit une trajectoire suffisamment régulière et douce pour que les entrées restent limitées.

Nous avons installé sur le réacteur industriel un régulateur, en service depuis 2000, qui assure les transitions entre deux points stationnaires en générant une trajectoire de référence boucle-ouverte complétée par des régulateurs assurant la stabilisation en boucle fermée. On trouvera les détails dans [PRB$^+$02].

[PRB$^+$00] N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, and P. Pinvidic. Control of an industrial polymerization reactor using flatness. In *Proc. of the International Symposium Mathematical Theory of Systems, Control, Network*, 2000.

[PRB$^+$02] N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, and P. Pinvidic. Control of an industrial polymerization reactor using flatness. *Journal of Process Control*, 12(5) :659–665, 2002. Best paper in the category "application" for the period 2002 to 2005.

2.3.3 Étude de la planification de trajectoires pour un problème de Stefan non linéaire

Dans les articles [DPRM03b, DPRM03a] nous avons montré comment calculer les trajectoires boucle ouverte pour un problème de Stefan non linéaire . Il s'agit d'un système régi par une équation aux dérivées partielles parabolique non linéaire à frontière libre. Nous cherchons à résoudre le problème inverse c.-à-d. que connaissant le comportement de la frontière libre *a priori* nous cherchons une solution, ici sous la forme d'une série convergente, permettant de calculer le contrôle et une description des trajectoires entre deux états stationnaires.

Le problème de Stefan classique représente une colonne en phase liquide en contact à 0 degrés avec une bande infinie de phase solide, tel que représenté sur la figure Figure 7. C'est un problème présenté en détail dans [Can84]. Une liste des problème se réduisant à celui-ci peut être trouvé dans [Rub71] : notamment de nombreux procédés de formation et de fonte des cristaux. Nous avons travaillé sur un problème de Stefan modifié en rajoutant un terme de diffusion et un terme non linéaire de réaction. Cela constitue un modèle simplifié de liquide réactant caloporteur entouré de phase solide tel qu'étudié dans [FS01].

Notons $(x,t) \mapsto u(x,t)$ la température dans la phase liquide et $t \mapsto y(t)$ la position de l'interface liquide/solide. Les fonctions $h(t)$ et $\psi(x)$ sont respectivement les températures à l'extrémité fixe ($x = 0$) et à l'instant initial ($t = 0$). Le problème de Stefan non linéaire consiste à déterminer $u(x,t)$ et $y(t)$, étant donné $h(t)$ et $\psi(x)$ satisfaisant

$$\left.\begin{array}{ll}
u_t = \ u_{xx} - \nu u_x - \rho u^2, & \forall (x,t) \in D_T \\
u(0,t) = h(t) \geq 0, & 0 < t \leq T \\
u(x,0) = \psi(x) \geq 0, & 0 \leq x \leq y(0) \\
u(y(t),t) = 0, \ u_x(y(t),t) = -\dot{y}(t), & 0 < t \leq T
\end{array}\right\} \qquad (1)$$

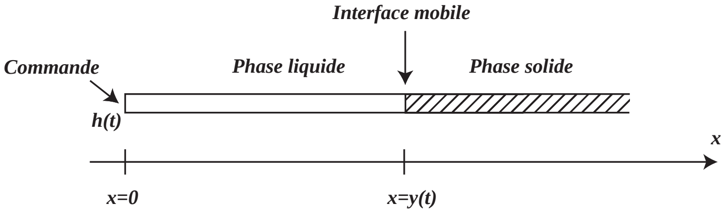


FIG. 7 – Problème de Stefan avec commande frontière. Phase liquide avec contrôle frontière gouvernée par une équation de réaction-diffusion non linéaire en contact avec une phase solide.

où

$$D_T \equiv \{(x,t) \ : \ 0 < x < y(t), \ 0 < t \leq T\}$$

avec les frontière notées

$$B_T \equiv \{(0,t) : 0 < t \leq T\} \ \cup \ \{(x,0) : 0 \leq x \leq y(0)\} \ \cup \ \{(y(t),t) : 0 < t \leq T\}$$

La condition frontière $u_x(y(t),t) = -\dot{y}(t)$ exprime que le flux thermique à l'interface est utilisé pour la fonte (ou la cristallisation) de la phase solide. Les paramètres de conductivité et de chaleur latente de liquéfaction sont ici normalisés mais, sans restrictions, on peut considérer des coefficients quelconques par des changements de variables sur x et t.

Le problème inverse consiste à calculer la commande frontière $h(t)$ permettant le transitoire entre deux états stationnaires. Comme le note Hill [Hil67], il s'agit d'un problème de Cauchy non caractéristique avec données de Cauchy. Nous avons résolu ce problème non linéaire par la méthode suivante. Nous montrons qu'on peut chercher des solutions de (1) sous la forme de la série suivante

$$u(x,t) = \sum_{n=0}^{\infty} \frac{a_n(t)}{n!}[x - y(t)]^n. \qquad (2)$$

où les coefficients $(a_n(t))$ satisfont les relations de récurrence nécessaires et suffisantes

$$a_n = \dot{a}_{n-2} - a_{n-1}\dot{y} + \nu a_{n-1} + \rho \sum_{k=0}^{n-2} \binom{n-2}{k} a_{n-2-k} \, a_k$$

pour $n \geq 2$, avec $a_0 = 0$ (d'après $u(y(t),t) = 0$) et $a_1 = -\dot{y}$ (d'après $-u_x(y(t),t) = \dot{y}(t)$).

Par des majorations, nous montrons que la série (2) converge absolument lorsqu'il existe des constantes strictement positives M, R, T telles que

$$\left| y^{(l+1)}(t) \right| \leq M \frac{l!^{\alpha}}{R^l}, \ \forall \ l = 0, 1, 2, ..., \forall t \in [0, T]$$

et calculons une borne inférieure à son rayon de convergence. Les principales difficultés résident dans le calcul par récurrence de bornes sur les dérivées successives des coefficients

$(a_n(t))$. Ceci implique des développements combinatoires des dérivées des termes croisés provenant de la non-linéarité en u^2, pour lesquels on peut utiliser des identités de Chu-Vandermonde (voir [PWZ96]). La borne inférieure sur le rayon de convergence est ensuite calculée par une analyse des racines d'un polynôme du troisième degré. Cette borne inférieure permet de justifier l'utilisation de cette solution sous forme de série pour résoudre le problème inverse de fonte (ou cristallisation) de la phase solide par la commande $h(t)$. Supposons que la phase liquide ait une longueur initiale L et qu'on souhaite atteindre en temps fini la longueur $L + \Delta L$. C'est un problème difficile car l'actionneur $h(t)$ est situé à l'extrémité fixe opposée à l'interface liquide-solide qui va se déplacer au cours du temps. La commande doit donc compenser la perte énergétique dûe à la fonte de solide et celle dûe à la diffusion et au terme de réaction. Pour résoudre ce problème, il suffit d'utiliser la fonction

$$y(\tau) = \begin{cases} L + \Delta L & \text{si } \tau \geq T, \\ L + \Delta L g(\tau/T) & \text{si } T > \tau > 0, \\ L & \text{si } \tau \leq 0, \end{cases}$$

où

$$g(\tau) = \frac{f(\tau)}{f(\tau) + f(1-\tau)}, \ \tau \in [0,1],$$

et

$$f(\tau) = \begin{cases} e^{-\frac{1}{\tau}} & \text{si } \tau > 0, \\ 0 & \text{si } \tau \leq 0. \end{cases}$$

Cette fonction définit une transition régulière entre les longueurs L et $L + \Delta L$. En choisissant le paramètre T en fonction des autres paramètres physiques, on peut garantir un rayon de convergence supérieur à $L + \Delta L$ prouvant ainsi que le développement en série, et donc la solution proposée au problème inverse sont valides.

Ces travaux font suite aux publications [LR00] pour une équation de réaction diffusion à frontière fixe. Outre la convergence de cette série pour une classe bien particulière de fonctions Gevrey (telles que définies dans [Gev18] et [Can84]) utilisables sous une hypothèse explicite dépendant des paramètres physiques du système, nous avons démontré un principe du maximum indiquant que le maximum de la température était toujours atteint au bord [DPRM03b] et une propriété asymptotique de positivité de la solution.

[Can84] J. R. Cannon. *The one-dimensional heat equation*, volume 23 of *Encyclopedia of Mathematics and its applications.* Addison-Wesley Publishing Company, 1984.

[DPRM03a] W. B. Dunbar, N. Petit, P. Rouchon, and P. Martin. Boundary control for a nonlinear Stefan problem. In *Proc. of the* 42*nd IEEE Conf. on Decision and Control*, 2003.

[DPRM03b] W. B. Dunbar, N. Petit, P. Rouchon, and P. Martin. Motion planning for a nonlinear Stefan problem. *ESAIM : Control, Optimisation and Calculus of Variations*, 9 :275–296, February 2003.

[FS01] M. Fila and P. Souplet. Existence of global solutions with slow decay and unbounded free boundary for a superlinear Stefan problem. *Interfaces and Free Boundaries*, 3 :337–344, 2001.

[Gev18] M. Gevrey. La nature analytique des solutions des équations aux dérivées partielles. *Ann. Sci. École Norm. Sup.*, 25 :129–190, 1918.

[Hil67] C. D. Hill. Parabolic equations in one space variable and the non-characteristic Cauchy problem. *Comm. Pure Appl. Math.*, 20 :619–633, 1967.

[LR00] A. F. Lynch and J. Rudolph. Flatness-based boundary control of a nonlinear parabolic equation modelling a tubular reactor. In A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, editors, *Lecture Notes in Control and Information Sciences 259 : Nonlinear Control in the Year 2000*, volume 2, pages 45–54. Springer, 2000.

[PWZ96] M. Petkovsek, H.S. Wilf, and D. Zeilberger. *A=B*. Wellesley, 1996.

[Rub71] L. I. Rubinstein. *The Stefan problem*, volume 27 of *Translations of mathematical monographs.* AMS, Providence, Rhode Island, 1971.

2.3.4 Planification de trajectoires pour des systèmes régis par des équations aux dérivées partielles hyperboliques

Une classe très riche de systèmes d'importance pratique sont les systèmes régis par des équations hyperboliques ou des équations des ondes. On les rencontre dans les phénomènes de transport (par exemple dans les réacteurs tubulaires ou dans les procédés d'activation par gas-lift ou de severe slugging [Dur05]). On peut étendre la notion de platitude à ce type de système, en montrant (quand c'est possible) qu'on peut paramétrer leurs trajectoires par leur sortie plate.

Nous avons traité différents exemples typiques : équation de Burgers linéaire [PCR98], équation des ondes pour un récipient rempli de liquide [PR02, DPR99, Nil03] (problème typique de l'industrie agro-alimentaire à la base de l'experience de laboratoire “Milk-race” à l'université Lund LTH inspiré par Tetra-Pak), équation des télégraphistes [FMPR99] et équation des ondes pour un système de pont roulant à cable pesant [PR01]. De manière générale, ces équations possèdent des vitesses de propagation finies [DL93]. Cette propriété importante nous permet d'étudier, dans les exemples que nous traitons, les relations liant à chaque fois l'entrée et la sortie plate du système grâce à des opérateurs avance et retard. On aboutit alors à une méthode complète de planification de trajectoire.

En nous plaçant dans le domaine du calcul opérationnel, ou des transformées de Laplace, nous pouvons ramener l'étude des équations aux dérivées partielles linéaires à une seule variable d'espace en l'étude d'équations différentielles ordinaires . Au lieu de nous intéresser à l'influence de la commande sur l'état, c.-à-d. la fonction de transfert (état)/(entrée), nous écrivons la “fonction de transfert” (état)/(sortie plate). Dans les cas traités on parvient à une écriture dans le domaine de Laplace de la forme

$$\hat{X}(x,s) = A_x(s)\hat{Y}(s)$$

où $\hat{X}$ est la transformée de Laplace de l'état, $\hat{Y}$ est la transformée de Laplace de la sortie plate et $A_x(\cdot)$ est une famille d'opérateurs indexée par la variable x. Naturellement on cherche à revenir dans le domaine temporel. La question est de savoir si $s \mapsto A_x(s)$ possède une transformée de Laplace inverse.

Dans les cas que nous traitons, on peut décomposer $A_x(s)$ sous la forme

$$A_x(s) = \sum_{i \in \{1,2\}} \alpha_i(x) \exp(\delta_i(x)s) + B_x(s)$$

en une somme finie d'opérateurs d'avance et de retard ponctuels et un autre opérateur. On montre ensuite que $s \mapsto B_x(s)$ satisfait aux hypothèses du théorème de Paley-Wiener ([Rud87]), c.-à-d. que $s \mapsto B_x(s)$ possède un original à support compact.

De retour dans le domaine temporel, on aboutit à une relation du type

$$X(x,t) = \sum_{i \in \{1,2\}} \alpha_i(x) Y(t + \delta_i(x)) + (b_x * Y)(t)$$

où $b_x(t) \supset B_x(s)$.

Cette dernière relation indique que l'état dépend de la sortie plate via des opérateurs à retards ponctuels et distribués à supports compacts. Le caractère compact du support de l'opérateur permet de raccorder des trajectoires du système, prouvant ainsi que ces systèmes sont commandables au sens du raccord des trajectoires de Willems [Wil91]. On peut être amené à faire également intervenir la dérivée temporelle de la sortie plate Y.

Une telle correspondance des trajectoires peut aussi être établie pour les équations hyperboliques du type

$$\begin{cases} v_t + \lambda(v) v_x = 0 & x \in [0,1] \\ v(0,t) = u(t) \end{cases}$$

puisqu'on a la relation entre $y(t) = v(1,t)$ et u et v

$$y(t) = u[t - 1/\lambda(y(t))], \quad y(t) = v[t - (1-x)/\lambda(y(t)), x]. \tag{3}$$

La formule (3), donne

$$v = y \circ (id - \frac{1-x}{\lambda(y)})^{-1}$$

où id est la fonction identité et $\circ$ est la loi de composition par rapport à la première variable.

Ces formules s'étendent notamment aux équations Lighthill-Whitham-Richard utilisées dans les problèmes de gestion de traffic sur axes monodimensionnels [ABSP05] .

Dans le cas des récipients remplis de liquide ("gamelles d'eau"), nous avons introduit la notion de "steady-state controllability" mettant en avant l'intérêt des transitions entre deux points stationnaires. Nous avons montré que les récipients rectangulaires et circulaires possédaient cette propriété. Il a été montré récemment [CCG05] que génériquement, ce n'est pas le cas pour des récipients à frontière régulière de type quelconque.

[ABSP05] J.-P. Aubin, A. M. Bayen, and P. Saint-Pierre. Computation and control of solutions to the burgers equation using viability theory. In *Proc. of the 2005 American Control Conference*, 2005.

[CCG05] Y. Chitour, J.-M. Coron, and M. Garavello. On conditions that prevent steady-state controllability of certain linear partial differential equations. In *Proc. of the 44th IEEE Conf. on Decision and Control*, 2005.

[DL93] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology*, volume 6. Springer-Verlag, 1993.

[DPR99] F. Dubois, N. Petit, and P. Rouchon. Motion planning and nonlinear simulations for a tank containing a fluid. In *European Control Conference, Karlsruhe*, 1999.

[Dur05] E. Duret. *Dynamique et contrôle des écoulements polyphasiques.* PhD thesis, École des Mines de Paris, 2005.

[FMPR99] M. Fliess, P. Martin, N. Petit, and P. Rouchon. Active signal restoration for the telegraph equation. In *Proc. of the 38th IEEE Conf. on Decision and Control*, 1999.

[Nil03] O. Nilsson. Physics based wave generation for the shallow water equations. Technical Report Masters thesis ISRN LUTFD2/TFRT--5710--SE, Department of Automatic Control, Lund Institute of Technology, Sweden, August 2003.

[PCR98] N. Petit, Y. Creff, and P. Rouchon. Motion planning for two classes of nonlinear systems with delays depending on the control. In *Proc. of the 37th IEEE Conf. on Decision and Control*, pages 1007– 1011, 1998.

[PR01] N. Petit and P. Rouchon. Flatness of heavy chain systems. *SIAM J. Control Optimization*, 40 :475–495, 2001.

[PR02] N. Petit and P. Rouchon. Dynamics and solutions to some control problems for water-tank systems. *IEEE Trans. Automat. Control*, 47(4) :594–609, 2002.

[Rud87] W. Rudin. *Real and Complex Analysis.* McGraw-Hill International Editions, third edition, 1987.

[Wil91] J. C. Willems. Paradigms and puzzles in the theory of dynamical systems. *IEEE Trans. Automat. Control*, 36 :259–294, 1991.

2.4 Commande optimale

2.4.1 Caractérisation des trajectoires optimales de systèmes de diélectrophorèse

Les systèmes de séparation de particules par diélectrophorèses présentent une dynamique assez spécifique. Par l'action d'un champ électrique non uniforme, un dipôle de crée à l'intérieur de chaque particule. La nature de ce dipôle dépend du type de particule considéré. Simultanément, le champs électrique agit sur le dipôle ainsi crée et déplace la particule, séparant les particules suivant la nature de leur dipôle. Les principes physiques de cette technique et son application au déplacement de particules ont été étudiés par Pohl [Poh78]. A l'époque, les champs électriques qu'on pouvait générer aux échelles spatiales intéressantes étaient assez limités. Il est aujourd'hui possible en utilisant des MEMS (micro-electro-mechanical systems), de générer des champs plus forts qu'auparavant à l'échelle de cellules, ADN, protéines et nanoparticules. Nous avons cherché, dans la publication [CP05], à formuler différentes problématiques intéressant la communauté de l'Automatique en mettant en lumière les propriétés de ces systèmes de diélectrophorèse. Principalement, nous nous sommes intéressés au problème de commande de déplacement de particules en temps minimal.

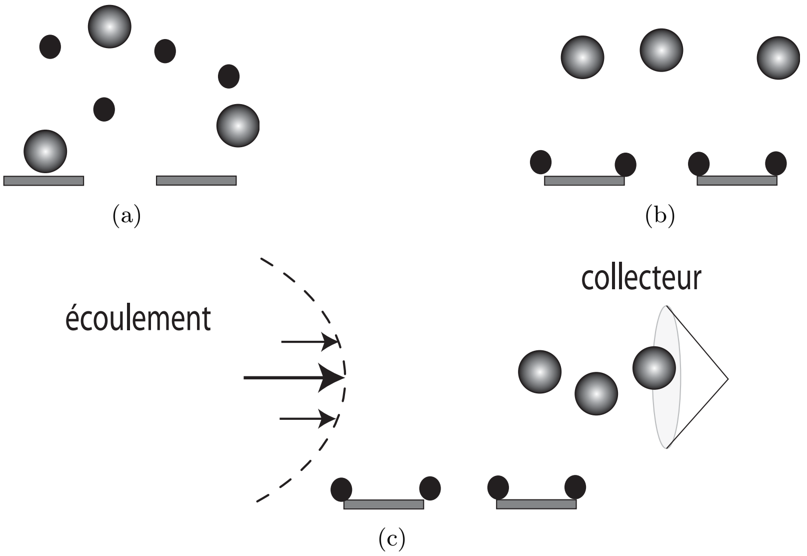


Fig. 8 – Séparation de particules par diélectrophorèse. (a) Mélange de particules de deux types différents à l'état initial. (b) Séparation verticale lorsque la force de diélectrophorèse est appliquée. (c) Collecte d'un des deux types de particules.

Nous avons considéré un modèle simple d'arrangement linéaire d'électrode, de façon à ramener l'étude du mouvement sur un axe monodimensionelle voir [CP05]. Étant donnée la dynamique à deux états (dipôle induit et position par rapport à l'électrode) et une commande (la différence de potentiel), nous avons cherché à caractériser les trajectoires optimales en temps pour un déplacement donné permettant une optimisation des procédés de séparation de particules utilisant cette technologie (par exemple, on pourrait réduire le temps d'évaluation du nombre de cellules cancéreuses dans un échantillon sanguin en optimisant leur séparation des cellules saines). Les équations en jeu sont

$$\begin{aligned} \dot{x} &= yu + \alpha u^2 \\ \dot{y} &= -cy + u \end{aligned} \tag{4}$$

avec $(x, y) \in \mathbb{R}^2$ et $u \in \mathbb{R}$ sous les contraintes

$$\begin{aligned} x(0) =&\ \text{donné}, \quad y(0) = 0, \\ x(t_f) =&\ \text{donné}, \quad y(t_f) = \text{ libre}, \\ |u| \leq&\ 1 \end{aligned}$$

Les paramètres α, c satisfont

$$\alpha < 0, \quad c > 0.$$

L'étude complète de la nature des extrémales aboutit à une caractérisation algorithmique très simple lorsque les contraintes portent sur la commande uniquement. Nous avons montré [CPR05, CP05], qu'à cause du terme quadratique dans la dynamique (4), les trajectoires optimales commencent toujours par une réponse inverse ("undershoot"). Les arguments utilisés sont assez élémentaires mais nombreux ce qui rend l'étude délicate : écriture d'un problème aux deux bouts par le principe du maximum de Pontryagin, preuve de la positivité du premier état adjoint, existence d'un point selle dans un plan de phase réduit, utilisation de trois symétries, minoration des temps de parcours et preuve d'existence et d'unicité. Ce travail figure dans [CPR06]. Le cas des contraintes d'état peut être partiellement traité analytiquement, ou traité directement numériquement comme dans [CPR05, CP05].

[CP05] D. E. Chang and N. Petit. Toward control of dielectrophoretic system. *International Journal of Robust and Nonlinear Control*, 15(16) :769–784, 2005.

[CPR05] D. E. Chang, N. Petit, and P. Rouchon. Time-optimal control of a particle in a dielectrophoretic system. In *Proc. of the* 16*th IFAC World Congress*, 2005.

[CPR06] D. E. Chang, N. Petit, and P. Rouchon. Time-optimal control of a particle in a dielectrophoretic system. *IEEE Trans. Automat. Control*, 51(7) :1100–1114, 2006.

[Poh78] H. A. Pohl. *Dielectrophoresis.* Cambridge University Press, 1978.

2.4.2 Technique de commande en temps minimum par platitude pour une unité d'alkylation

Nous présentons ici à travers une application industrielle une méthode générale de commande en temps minimum pour les systèmes linéaires à retards dont nous avons établi la propriété de platitude.

Le contrôleur que nous présentons est en service sur l'unité de régulation du circuit acide de l'unité d'alkylation de la raffinerie TOTAL de Feyzin (Rhône) depuis janvier 1997.

L'alkylation des butènes est une opération courante dans les raffineries pétrolières. Cette opération permet la synthèse d'alkylats, produits intéressants pour leur indice d'octane élevé et utilisé dans la composition de carburants. Le catalyseur acide alimente en série et de manière continue deux réacteurs. Partiellement détérioré au cours de l'alkylation, le catalyseur est soutiré du second réacteur et dirigé vers un ballon de stockage pour une régénération hors site. Il est nécessaire de maintenir une quantité minimale de catalyseur dans les réacteurs pour que la réaction se déroule correctement. On peut en fournir plus que le minimum requis pour éviter les risques de dysfonctionnement mais cela induit de coûteuses surconsommations. L'opérateur cherche donc à stabiliser l'unité légèrement au dessus de la quantité minimale requise. La détérioration du catalyseur est très lente, ce qui rend difficile un pilotage manuel. L'unité étant très lente, nous avons choisi de mettre en œuvre un algorithme de commande en temps minimal. Ce régulateur fonctionne depuis janvier 1997, avec un taux d'utilisation supérieur à 98%. Il a permis de diminuer d'environ 5% la consommation annuelle d'acide.

Le problème de contrôle s'écrit sous la forme d'un problème d'optimisation que nous résolvons de manière approchée à l'aide d'une discrétisation de la sortie plate du système, ce qui nous permet de nous ramener à un problème de dimension finie. Plus précisément, on se ramène à tester l'existence d'un point dans l'intérieur d'un polytope. Ceci est résolu par un algorithme de programmation linéaire. L'existence d'un tel point à l'intérieur du polytope dépend de manière monotone du temps de transition. Par une dichotomie, on cherche le temps le plus court fournissant une solution admissible.

Nous avons démontré analytiquement la convergence de cette méthode dans le cas précis que nous considérons. Ainsi peut-on établir que cette méthode de résolution du temps minimum sous contraintes par discrétisation converge effectivement vers la solution du problème continu, lorsqu'elle existe, lorsque le pas de discrétisation tend vers 0. Ces travaux ont été publié dans [PCLR01] où nous avons donnée des résultats d'exploitation sur une période de 6 mois.

[PCLR01] N Petit, Y. Creff, L. Lemaire, and P. Rouchon. Minimum time constrained control of acid strength on a sulfuric acid alkylation unit. *Chemical Engineering Science*, 56(8) :2767–2774, 2001.

2.4.3 Méthodes numériques utilisant l'inversion non-linéaire

Nous avons étudié l'utilisation de l'inversion non linéaire et de la platitude dans le calcul de trajectoires de référence optimales. La conclusion est que lorsqu'elle est exploitée, la réduction du nombre d'inconnues qu'elles procurent permet de réduire significativement les temps de calculs. Ces constatations sont expérimentales.

Considérons un problème général de commande optimale sous contraintes pour un système affine en la commande

$$\begin{cases} \min\limits_{(x,u)} J(x,u) \\ \dot{x} = f(x) + g(x)u, \\ lb \leq c(x,u) \leq ub. \end{cases}$$

où J, f, g, c sont des fonctions régulières de leurs arguments et lb et ub des vecteurs constants. Il est possible de résoudre ce problème par une technique de collocation type Hargrawes-Paris [HP87] en approximant ce problème en dimension finie en représentant les variables du système (états x de dimension n et entrées u de dimension m) par des splines $\hat{x}$, $\hat{u}$ définies sur une grille

$$t_0 = t_1 < t_2 < \ldots < t_N = t_f$$

Chaque spline est définie par N coefficients. Les contraintes $lb \leq c(x,u) \leq ub$ et les dynamiques sont approximativement satisfaites par $\hat{x}$ et $\hat{u}$ aux points de la grille et le problème devient un problème de programmation non linéaire

$$\begin{cases} \min\limits_{z \in \mathbb{R}^{(n+m)N}} F(z) = J(\hat{x}(z), \hat{u}(z)) \\ \dot{\hat{x}} - f(\hat{x}(z), \hat{u}(z)) = 0, \quad lb \leq c(\hat{x}(z), \hat{u}(z)) \leq ub, \end{cases}$$

où z est un vecteur de coefficients inconnus de dimension $M = (n+m)N$.

Il est pourtant possible lorsque le système est plat de paramétrer toutes les grandeurs du système (états et entrées) par les m sorties plates du système. En approximant uniquement les sorties plates par des splines définies sur la même grille que précédemment on obtient un autre problème de programmation non-linéaire de dimension beaucoup faible mN

$$\begin{cases} \min\limits_{y \in \mathbb{R}^{mN}} F(y) = J(\hat{x}(y), \hat{u}(y)) \\ lb \leq c(\hat{x}(y), \hat{u}(y)) \leq ub, \end{cases}$$

où en outre les équations de la dynamique ont disparu.

La complexité numérique de la résolution d'un problème de programmation non-linéaire est une fonction cubique du nombre d'inconnues en jeu [GMW81]. On peut donc prévoir une substantielle réduction du temps de calcul en exploitant la platitude du système.

En pratique on retrouve les gains ainsi espérés. On pourra se reporter aux expériences que nous avons réalisées avec NTG [MMM00] le logiciel de calcul de trajectoires optimales développé au California Institute of Technology [MHJ$^+$03]. De manière générale, c'est le degré relatif r (tel que défini dans [Isi89]) de la sortie considérée qui compte. On peut en effet éliminer r grandeurs du problème d'optimisation comme nous l'avons montré dans [PMM01]. Plus r est grand, plus on peut éliminer de variables et donc directement d'équations différentielles à satisfaire, et plus on réduit les temps de calculs nécessaires à la résolution.

Pour les systèmes régis par des équations aux dérivées partielles, il est également possible de tirer parti de l'élimination de variables à travers les équations de la dynamique. Pour ce faire, nous avons écrit une extension de NTG pour ces systèmes, en considérant des tensor-product B-Splines comme fonctions de bases. Ceci nous permet de traiter des cas sous contraintes à une dimension en espace et une dimension en temps comme par exemple les réacteurs tubulaires. Cette approche est détaillée dans [PMM02].

Nous avons également utilisé cette méthode pour des problèmes difficiles issus du domaine aérospatial [MP01, NTP03, MPM01] notamment le problème de réentrée atmosphérique et le vol en formation de microsatellites en présence d'effet J2. Ces derniers problèmes comportent des singularités dans le paramétrage des variables en fonction des sorties plates. Certaines de ces singularités sont de fausses singularités qui disparaissent après simplifications (typiquement dans les relations trigonométriques et réciproques), d'autres sont éliminables localement, au voisinage de solutions de référence. Dans les cas traités dans [MP01, MPM01], on a des systèmes plats. Dans le cas présenté dans [NTP03], nous avons montré que le problème n'est pas plat en utilisant la condition nécessaire de platitude de [Rou95] et sommes parvenus à une représentation minimale du problème d'optimisation en terme de nombre de variables.

Nous avons également travaillé sur l'utilisation de cette technique d'inversion pour les méthodes indirectes de l'optimisation de trajectoires (méthode avec adjoint et problème aux deux bouts) dans [CP03, CPre]. On montre qu'autant de variables adjointes que de variables primales peuvent être explicitement reconstruites à partir des dérivées de la sortie linéarisante. Ceci permet des raccourcis intéressants dans le calcul des extrémales de voisinages (telles que définies dans [BH69]) et ouvre des perspectives concernant la résolution numérique par les méthodes spécifiques pour les problèmes aux deux bouts d'ordre supérieur telles que présentée dans [AMR88].

[AMR88] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell. *Numerical solution of boundary value problems for ordinary differential equations.* Prentice Hall Series in Computational Mathematics. Prentice Hall, Inc., Englewood Cliffs, NJ, 1988.

[BH69] A. E. Bryson and Y. C. Ho. *Applied Optimal Control.* Ginn and Company, 1969.

[CP03] F. Chaplais and N. Petit. Inversion in indirect optimal control. In *Proc. of the 7th European Control Conf.*, 2003.

[CPre] F. Chaplais and N. Petit. Inversion in indirect optimal control of multivariable systems. *ESAIM : Control, Optimisation and Calculus of Variations*, 2007 (à paraître).

[GMW81] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Addison-Wesley, 1981.

[HP87] C. Hargraves and S. Paris. Direct trajectory optimization using nonlinear programming and collocation. *AIAA J. Guidance and Control*, 10 :338–342, 1987.

[Isi89] A. Isidori. *Nonlinear Control Systems.* Springer, New York, 2nd edition, 1989.

[MHJ$^+$03] R. M. Murray, J. Hauser, A. Jadbabaie, M. B. Milam, N. Petit, W. B. Dunbar, and R. Franz. Online control customization via optimization-based control. In T. Samad and G. Balas, editors, *Software-Enabled Control, Information technology for dynamical systems*, pages 149–174. Wiley-Interscience, 2003.

[MMM00] M. B. Milam, K. Mushambi, and R. M. Murray. A new computational approach to real-time trajectory generation for constrained mechanical systems. In *IEEE Conference on Decision and Control*, 2000.

[MP01] M. B. Milam and N. Petit. Constrained trajectory generation for a planar missile. Technical report, California Institute of Technology, Control and Dynamical Systems, 2001.

[MPM01] M. B. Milam, N. Petit, and R. M. Murray. Constrained trajectory generation for micro-satellite formation flying. In *AIAA Guidance, Navigation and Control Conference*, pages 328–333, 2001.

[NTP03] T. Neckel, C. Talbot, and N. Petit. Collocation and inversion for a reentry optimal control problem. In *Proc. of the 5th Intern. Conference on Launcher Technology*, 2003.

[PMM01] N. Petit, M. B. Milam, and R. M. Murray. Inversion based constrained trajectory optimization. In 5*th IFAC Symposium on Nonlinear Control Systems*, 2001.

[PMM02] N. Petit, M. B. Milam, and R. M. Murray. A new computational method for optimal control of a class of constrained systems governed by partial differential equations. In *Proc. of the* 15*th IFAC World Congress*, 2002.

[Rou95] P. Rouchon. Necessary condition and genericity of dynamic feedback linearization. *J. Math. Systems Estim. Control*, 5(3) :345–358, 1995.

2.4.4 Application au modèle d'un réacteur tubulaire

Comme nous l'avons montré dans [PMM02], il est intéressant d'utiliser l'élimination de variables pour les équations aux dérivées partielles où apparaît le contrôle. Nous avons utilisé l'extension du logiciel NTG pour de tels systèmes et étudié deux cas de la littérature consistant en deux équations aux dérivées partielles paraboliques monodimensionelle réprésentant un problème de transition de phase liquide-solide d'une part, et une équation de Ginzburg-Landau [Zwi97] d'autre part. Nous avons ainsi montré que nous obtenions des solutions très proches de celles obtenues par des techniques de collocation classique. En outre, il est possible, au prix d'une tolérance sur les conditions d'extrémalité, de réduire encore les temps de calculs pour un usage en commande prédictive (telle que définie dans [MRRS00]), voir [Mil03].

Nous avons également utilisé cette approche sur un modèle de réacteur tubulaire. Ce travail prolonge une étude industrielle réalisée sur le réacteur ATOFINA PS de Carling au cours de laquelle nous avons réglé les contrôleurs PI existants en utilisant un modèle de la cinétique et en explicitant les couplages ayant lieu à travers les actionneurs. Ces réglages ont permis d'augmenter significativement la production du réacteur (par plus de 10%).

Dans ce réacteur, le grade du polystyrène fabriqué dépend du profil de température présent le long du réacteur. La variable qui doit être contrôlée est précisément la grandeur qu'il faut contrôler (par opposition aux cas nombreux où on cherche prioritairement à contrôler le taux de conversion). Sur cette unité, les contraintes de qualités sont très serrées et imposent un contrôle précis de la température. Le production est organisée par une optimisation économique ce qui résulte en de fréquents changements de points de fonctionnement. Dans la publication [dVP05], nous avons cherché à comparer les techniques de contrôle décentralisé à base de contrôleurs PI, ainsi qu'une technique centralisée utilisant encore les contrôleurs PI mais dont les gains sont calculés par une méthode LQR mettant en avant les couplages du système et enfin par une méthode non linéaire utilisant l'optimisation de trajectoires pour équations aux dérivées partielles. Il s'agit en effet d'un réacteur tubulaire à échangeurs thermiques répartis. Chaque actionneur agit sur une zone assez étendue, et contrairement à ce qu'on peut penser, la dynamique du système n'est pas une simple cascade orientée par le sens de l'écoulement mais comporte bien un couplage par l'échangeur thermique.

Dans la publication [dVP05], nous avons élucidé plusieurs questions portant sur ce réacteur régi par une équation aux dérivées partielles sous-actionné : origine des instabilités, robustesse des différentes stratégies de commande monovariables, intérêt de la commande multivariable centralisée linéaire, intérêt de la commande par stabilisation autour d'une trajectoire optimale de l'équation aux dérivées partielles.

[dVP05] D. del Vecchio and N. Petit. Boundary control for an industrial under-actuated tubular chemical reactor. *Journal of Process Control*, 15(7) :771–784, 2005.

[Mil03] M. B. Milam. *Real-time optimal trajectory generation for constrained systems.* PhD thesis, California Institute of Technology, 2003.

[MRRS00] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control : stability and optimality. *Automatica*, 36 :789–814, 2000.

[PMM02] N. Petit, M. B. Milam, and R. M. Murray. A new computational method for optimal control of a class of constrained systems governed by partial differential equations. In *Proc. of the* 15*th IFAC World Congress*, 2002.

[Zwi97] D. Zwillinger. *Handbook of Differential Equations.* Academic Press, 3rd edition, 1997.

2.5 Estimation de variables non-mesurées

Les contraintes croissantes des normes environnementales ont placé le contrôle moteur au cœur de nombreuses préoccupations des constructeurs automobiles. L'objectif du contrôle moteur est la gestion en temps réel de l'alimentation en air et en carburant du moteur pour le pilotage actif de la combustion dans les cylindres. En pratique, on peut contrôler le circuit d'air et le système d'injection, en jouant sur les quantités admises ainsi que sur les lois horaires et leur synchronisation (voir [GLFP02]). Une volonté persistante de diminuer les polluants et d'augmenter la performance du moteur ont suscité la recherche de techniques de contrôle de plus en plus efficaces (voir [GA98, KN00] par exemple). Les stratégies de contrôle deviennent de plus en plus structurées et nécessitent des modèles de connaissances toujours plus précis prenant en compte des phénomènes haute fréquence. Désormais, un des principaux enjeux est le contrôle temps réel de la combustion. En supposant les actionneurs parfaits et les mesures suffisamment nombreuses et non bruitées, la tâche pourrait sembler relativement aisée, en tout cas, de nombreuses solutions sont envisageables. En pratique hélas, on est loin du compte. Si on laisse de côté les problèmes propres aux actionneurs (par exemple les problèmes de common rail comme ceux soulignés dans [CG03, OST$^+$03]), on peut se concentrer sur les problèmes d'observation. En vue du contrôle actif de la combustion, des observateurs haute fréquence sont à développer. Bien que nous travaillons sur des moteurs expérimentaux, nous n'utilisons que des informations provenant de capteurs utilisés sur les moteurs de série avec comme constant souci l'applicabilité en temps réel sur un système embarqué.

À l'échelle de temps que nous considérons (6^o vilebrequin soit typiquement 250 μs), le moteur est un système instationnaire, périodique dans la variable angulaire. Cette périodicité est mécanique, elle provient de la géométrie des nombreuses masses mobiles. C'est une hypothèse fondamentale pour notre travail, qui n'est pas remise en cause par les nombreuses perturbations auxquelles le moteur est soumis (défaut des injecteurs, variation de composition du carburant, de la charge et de la demande de couple...). De tels systèmes périodiques ont été largement étudiés par la communauté de l'Automatique (voir [BC01, BG86, AM81] par exemple). Il résulte de ces études que le point clef de la construction d'observateurs est la propriété d'observabilité uniforme. Dans les cas que nous considérons, c'est une propriété que nous pouvons établir. À partir de ce constat, on peut construire différents observateurs. Nos choix se sont portés sur un filtre de Kalman et un observateur non linéaire de type Luenberger (recopie de la dynamique avec injection de sortie et terme de contraction). Le premier choix est dicté par la grande confiance ressentie à l'égard de cette technique éprouvée, le second répond à un soucis de réduction de temps de calcul. Notre conclusion est que, bien que très populaire et efficace en pratique, le filtre de Kalman se révèle, dans notre contexte de contrôle moteur, bien plus consommateur de ressources que l'observateur non linéaire. On peut obtenir, sur banc moteur, grâce à l'observateur non linéaire des résultats comparables sans requérir une part trop importante des ressources du calculateur embarqué.

Nous avons étudié deux exemples de problèmes d'observation de la combustion : l'estimation du couple de combustion et l'estimation de la richesse cylindre à cylindre. Dans le premier cas, seul le capteur de vitesse angulaire du vilebrequin est utilisé, dans le second cas, un unique capteur de richesse placé derrière la turbine à l'échappement est utilisé (voir figure 9).

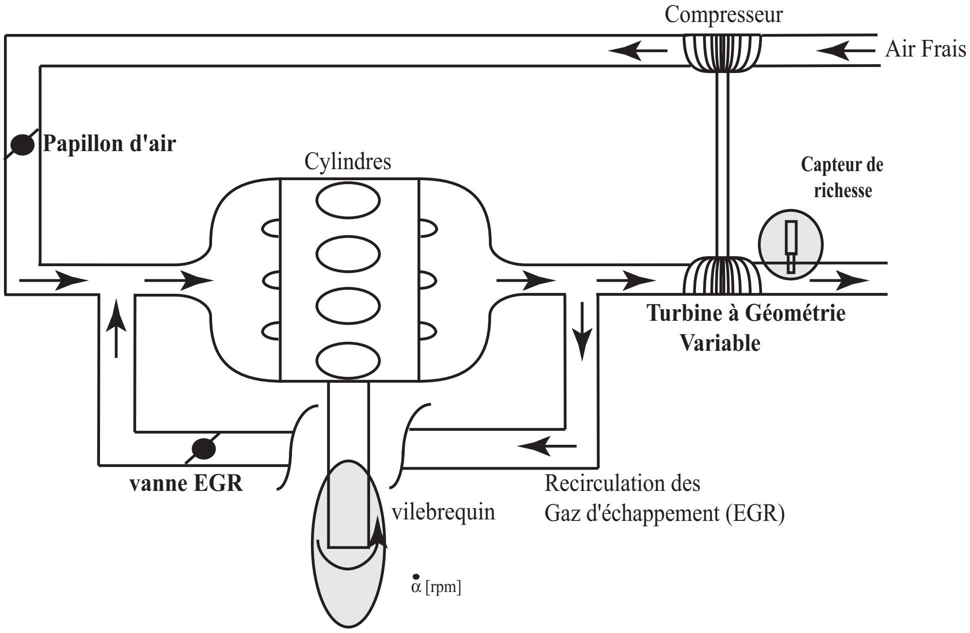


Fig. 9 – Schéma moteur avec boucle de recirculation et emplacements de capteurs utilisés pour le contrôle de la combustion.

Il assez facile, par symétrie, de ramener le nombre nécessaire de paramètres de réglages de l'observateur à seulement trois (on se reportera à [CCM$^+$04b] pour les détails). De manière intéressante, il est possible de conserver les réglages obtenus en simulation sur le banc moteur. Ceci est sans doute dû à la qualité du modèle utilisé pour la simulation (librairie de l'IFP) et également à la pertinence de notre modélisation réduite (équations bilans simples) utilisées pour le calcul des observateurs. Quantitativement et qualitativement, les résultats des deux estimateurs sont proches. En pratique l'observateur non linéaire est beaucoup plus facile à régler. On notera également que la loi horaire du couple dépend du point de fonctionnement considéré. Il est nécessaire de régler le filtre de Kalman pour chacun de ces points de fonctionnement par calibration. À l'inverse, aucune mise à jour des réglages n'est vraiment nécessaire pour l'observateur non linéaire. Mathématiquement, les preuves de convergence s'articulent comme-suit. On démontre que le système converge vers un cycle limite. Ceci est donné par l'inspection des propriétés de contraction du système et l'analyse de l'application de Poincaré. Ensuite, en travaillant dans le voisinage ce ce cycle limite, on montre l'observabilité uniforme du système instationnaire en tant que famille de système qui eux, de manière intéressante, ne sont pas observables. L'argument principale est l'indépendance linéaire des débits d'échappements des cylindres en tant que fonctions continues périodiques (les débits sont déphasés mécaniquement). On peut ainsi conclure dans ce cas précis à la convergence (locale) d'un filtre de Kalman instationnaire (périodique). Dans le cal de l'observateur non linéaire, on utilise une fonction de Lyapunov qui mène à l'étude d'un ensemble invariant par le théorème de LaSalle. C'est la même hypothèse d'indépendance linéaire des débits d'échappements des cylindres qui permet de conclure que cet ensemble est réduit à 0. De manière intéressante, on voit qu'il faut faire appel à des propriétés fondamentales et spécifiques aux moteurs pour prouver la convergence de deux observateurs assez naturels.

En pratique, de nombreux détails d'implémentations sont à prendre en compte : utilisation de cartographie pour les modèles de pression échappement, synchronisation des mesures pour prendre en compte les délais des phénomènes de transport, inversion du filtrage des capteurs (sonde λ par exemple) par des techniques d'estimation linéaire de leur décomposition de Fourier. On trouvera des développements ainsi que de nombreux résultats expérimentaux réalisés sur des moteurs HCCI dans [CC$^+$re, CCM$^+$06a, CCPR06b, CCM$^+$06b, CCPR06a, CCM$^+$05, CPRC06, CPR$^+$06a, CPR$^+$06b, CCP$^+$04, CCM$^+$04a, CCM$^+$04e, CCM$^+$04d, CCM$^+$04c].

[AM81] B.D.O. Anderson and J.B. Moore. Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM J. Contr. Optmiz.*, 19 :20–32, 1981.

[BC01] S. Bittanti and P. Colaneri. *Periodic Control Systems.* Pergamon, 2001.

[BG86] S. Bittanti and G. Gurdabassi. Optimal periodic control and periodic systems analysis : An overview. In *Proc. of the IEEE Conf. Decision and Control*, pages 1417–1423, 1986.

[CC$^+$re] J. Chauvin, G. Corde, , N. Petit, and P. Rouchon. Periodic input estimation for linear periodic systems : automotive engine applications. *Automatica*, 2006 (à paraître).

[CCM$^+$04a] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Observer design for torque balancing on a DI engine. In *Proc. of Society Automotive Engine World Congress*, 2004.

[CCM$^+$04b] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Real-time combustion torque estimation on a Diesel engine test bench using an adaptive Fourier basis decomposition. In *Proc. of the the* 43*rd IEEE Conf. Decision and Control*, 2004.

[CCM$^+$04c] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Real-time combustion torque estimation on a Diesel engine test bench using an adaptive Fourier basis decomposition. In *Proc. of the* 43*rd IEEE Conf. on Decision and Control*, 2004.

[CCM$^+$04d] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Real-time combustion torque estimation on a Diesel engine test bench using time-varying Kalman filtering. In *Proc. of the* 43*rd IEEE Conf. on Decision and Control*, 2004.

[CCM$^+$04e] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Time-varying linear observer for torque balancing on a DI engine. In *Proc. of the First IFAC Symposium on Advances in Automative Control*, 2004.

[CCM$^+$05] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon. Real-time nonlinear individual cylinder air fuel ratio observer on a Diesel engine. In *Proc. of the* 16*th IFAC World Congress*, 2005.

[CCM$^+$06a] J. Chauvin, G. Corde, P. Moulin, N. Petit, and P. Rouchon. High frequency individual cylinder estimation for control of Diesel engines. *Oil & Gas Science and Technology - Revue de l'Institut Français du Pétrole*, 61(1) :57–72, 2006.

[CCM+06b] J. Chauvin, G. Corde, P. Moulin, N. Petit, and P. Rouchon. Kalman filtering for real-time individual cylinder air fuel ratio observer on a Diesel engine test bench. In *Proc. of the 2006 American Control Conference*, 2006.

[CCP+04] J. Chauvin, G. Corde, N. Petit, P. Rouchon, P. Moulin, and M. Castagné. Nonlinear output injection observer of indicated torque for DI engine. In *Proc. of the FISITA 04 World Automotive Congress*, 2004.

[CCPR06a] J. Chauvin, G. Corde, N. Petit, and P. Rouchon. Experimental motion planning in airpath control for HCCI engine *(winner of the best presentation in session award)*. In *Proc. of the 2006 American Control Conference*, 2006.

[CCPR06b] J. Chauvin, G. Corde, N. Petit, and P. Rouchon. Filtre de Kalman ou observateur de Luenberger nonlinéaire ? comparaison expérimentale sur des exemples issus du contrôle moteur. In *Conférence Internationale Francophone d'Automatique*, 2006.

[CG03] O. Chiavola and P. Giulianelli. Modelling and simulation of common rail systems. In *Proc. of Society Automotive Engine World Congress*, 2003.

[CPR+06a] J. Chauvin, N. Petit, P. Rouchon, G. Corde, and C. Vigild. Air path estimation on Diesel HCCI engine. In *Proc. of Society Automotive Engine World Congress*, 2006.

[CPR+06b] J. Chauvin, N. Petit, P. Rouchon, P. Moulin, and G. Corde. Six degrees crankshaft individual air fuel ratio estimation of Diesel engines for cylinder balancing purpose. In *Proc. of Society Automotive Engine World Congress*, 2006.

[CPRC06] J. Chauvin, N. Petit, P. Rouchon, and G. Corde. Periodic input observer design : Application for imbalance diagnosis. In *Proc. of Society Automotive Engine World Congress*, 2006.

[GA98] L. Guzzella and A. Amstutz. Control of Diesel engines. *Proc. of the IEEE Control Systems Magazine*, 18 :53–71, 1998.

[GLFP02] G. Gissinger and N. Le Fort-Piat. *Contrôle-commande de la voiture.* Lavoisier, 2002.

[KN00] U. Kiencke and L. Nielsen. *Automotive Control Systems For Engine, Driveline, and Vehicle.* Springer, 2000.

[OST+03] T. Ogata, Y. Serizawa, H. Tsuchiya, K. Hayashi, and K. Mizuno. Further pressure pulsation reduction in fuel rails. In *Proc. of Society Automotive Engine World Congress*, 2003.

2.6 Modélisation

2.6.1 Mélanges en cuve

Nous avons travaillé sur la modélisation des systèmes de mélange industriels pour produits visqueux (type glucose, confiture, yaourt) dans des cuves munis d'hélices rubans tels que présentés dans [Ott89]. La contribution principale de ce travail est l'élaboration d'un modèle mixte type réacteur parfaitement agité en boucle fermée avec un réacteur type piston. La commande consiste à faire varier les frontières spatiales entre les deux domaines,

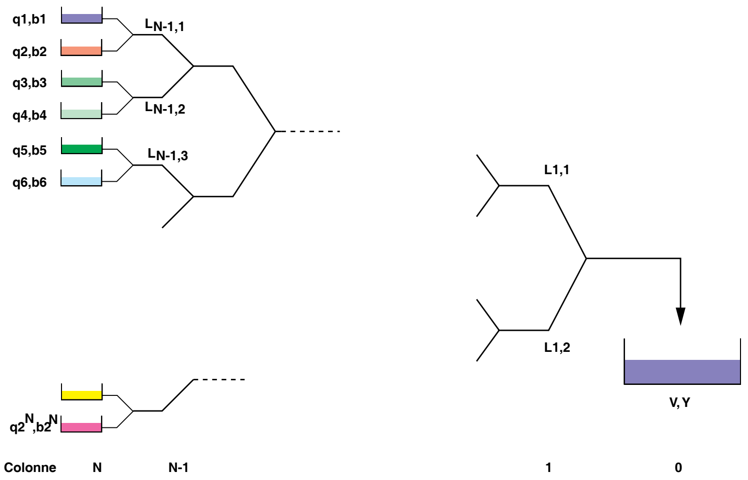


Fig. 10 – Réseau binaire de prémélanges.

procurant ainsi une “prime au mélange” telle qu’intuitivement remarquée par les expérimentateurs. Différentes variantes de ce modèle ont été proposées [DPRD05a, DPRD05b]. Nous avons montré qu’il s’agissait bien d’un modèle conservatif, et qu’il était possible de le valider sur des données expérimentales recueillies par une sonde de conductimétrie axiale sur une cuve de solution aqueuse de glucose. Nous continuons ces travaux en cherchant les stratégies de commande optimale de ces modèles.

[DPRD05a] J.-Y. Dieulot, N. Petit, P. Rouchon, and G. Delaplace. An arrangement of ideal zones with shifting boundaries as a way to model mixing processes in unsteady stirring conditions in agitated vessels. *Chemical Engineering Science*, 60(20) :5544–5554, 2005.

[DPRD05b] J.-Y. Dieulot, N. Petit, P. Rouchon, and G. Delaplace. A torus model containing a sliding well-mixed zone as a way to represent mixing process at unsteady stirring conditions in agitated vessels. *Chemical Engineering Communications*, 192 :805–826, 2005.

[Ott89] J. M. Ottino. *The kinematics of mixing : stretching, chaos, and transport.* Cambridge Texts in Applied Mathematics. Cambridge University Press, 1989.

2.6.2 Modèles à retard variables

Nous avons travaillé sur les réseaux de canalisations tels qu’ils sont utilisés dans les raffineries pour réaliser en continu ou en batch les mélanges de produits de bases produisant les carburants à destination commerciale. Une des spécificités de ces réseaux est l’utilisation de prémélanges pour éviter la redondance de canalisations.

Les différents bacs de produits de base sont, en fonction de leur éloignement de la mélangeuse, regroupés par une cascade de prémélanges. Sous l’hypothèse d’écoulement incompressible, le débit dans chacune des sous-sections peut se calculer comme combinaison des débits des produits de bases, c.-à-d. la recette. Celle-ci évolue dans le temps notamment lorsqu’on change de produit à fabriquer. Les retards dus aux transport dans les canalisations sont donc variables. Néanmoins, on peut les calculer explicitement par une formule du type

$$V = \int_{t-\delta(t)}^{t} q(s)ds$$

où V est le volume de la canalisation et q le débit au cours du temps. Cette formule est inversible dans chacune des branches du réseau. Un cas intéressant est celui de trois bacs de bases utilisant un prémélange. Dans ce cas que nous avons exposé dans [PCR98] ("mélange des couleurs"), on peut paramétrer toutes les trajectoires du système en utilisant comme sorties plates les quantités contenues dans le bac de sortie et ainsi planifier l'ouverture des vannes de commande en fonction de la production souhaitée dans le bac de sortie. Il s'agit encore de la notion de platitude. De manière générale, nous avons montré que cette propriété est vraie pour un réseau binaire quelconque tel que représenté sur la figure 10. La démonstration figure dans [Pet00].

[PCR98] N. Petit, Y. Creff, and P. Rouchon. Motion planning for two classes of nonlinear systems with delays depending on the control. In *Proc. of the* 37*th IEEE Conf. on Decision and Control*, pages 1007– 1011, 1998.

[Pet00] N. Petit. *Systèmes à retards, platitude en génie des procédés et contrôle de certaines équations des ondes.* PhD thesis, École des Mines de Paris, 2000.

2.7 Séminaires

- Séminaire d'Automatique de Paris ENS Cachan : *Platitude et équations des télégraphistes.* Avril 2000.
- Séminaire Control and Dynamical Systems California Institute of Technology : *Delay systems : open loop control techniques.* Novembre 2000.
- Journées Commandes Méthodes numériques en commande optimale Université d'Orleans : *Problèmes de commande optimale et systèmes plats (EDO et EDP).* Janvier 2002
- Séminaire CESAME Université Louvain-la-Neuve : *Commande optimale par platitude.* Mars 2002.
- GDR systèmes à retards : *Platitude des systèmes de mélanges en raffinage.* Octobre 2002
- Séminaire d'Automatique de Paris CNAM : *Planification de trajectoires pour un problème de Stefan non linéaire* . Novembre 2002
- Scientific Comitee Meeting TOTAL Petrochemicals : *Control of polymerization reactors, PDEs driven systems.* Mars 2005
- Séminaire Kolloquium Technische Kybernetik Université de Stuttgart : *Individual Cylinder Observers on Diesel Engines.* Juillet 2005

2.8 Applications industrielles

- **Logiciel TOTAL ANAMEL V4 et V5.**
  Algorithme de réalisation en temps réel des mélanges de carburants en présence d'incertitudes. Assure la réalisation de tous les carburants des raffineries de Donges, Feyzin, Gonfreville, Leuna (Allemagne), Grand-Puits. resp. 2001, 2002, 2003, 2004, 2005

- **Réacteur ATOFINA PS de Carling.**
  Algorithme de Contrôle temps réel (Réacteur tubulaire 120 KT/an). 2000
- **Réacteur APPRYL PP2 de Lavéra.**
  Algorithme de Contrôle temps réel (Réacteur parfaitement agité 250KT/an). 2000
- **Raffinerie TOTAL de Feyzin.**
  Algorithme de Contrôle temps réel de l'unité d'alkylation. 1996

2.9 Contrats d'études industrielles (responsable principal)

- **DGA/LRBA**
  Contrôle de formations d'hélicoptères miniatures en collaboration avec des engins terrestres autonomes. 2004-2007.
- **DGA/ONERA**
  Lauréat de la bourse pour le concours universitaire international de Mini-drones (projet Oiseau Artificiel). 2003-2005.
- **EDF**
  Modélisation pour la commande des circuits primaires et secondaires des réacteurs nucléaires à eau pressurisée. 2002-2005
  Commande optimale d'un parc de microturbines à gaz. 1999
- **IFP**
  Contrôle moteur Diesel. 2003-2006.
- **TOTAL**
  Contrôle de puits activés en gas-lift. 2003-2006
- **CNES**
  Optimisation de trajectoires de réentrée atmosphérique. 2003

2.10 Brevets (co-déposant)

- Titre : "METHODE D'ESTIMATION PAR UN FILTRE NON-LINEAIRE ADAPTATIF DE LA RICHESSE DANS UN CYLINDRE D'UN MOTEUR A COMBUSTION." numéro de dépôt de la demande de brevet : 05/05.442.
- Titre : "METHODE D'ESTIMATION PAR UN FILTRE DE KALMAN ETENDU DE LA RICHESSE DANS UN CYLINDRE D'UN MOTEUR A COMBUSTION." numéro de dépôt de la demande de brevet : 05/05.443.
- Titre : "METHODE D'ESTIMATION DU REGIME INSTANTANE PRODUIT PAR CHACUN DES CYLINDRES D'UN MOTEUR A COMBUSTION INTERNE." numéro de dépôt de la demande de brevet : 05/09.624.

2.11 Autres activités scientifiques

- depuis avril 2006 **Associate Editor pour la revue Automatica**
- depuis 2004 membre du conseil scientifique du projet recherche fédérateur Doom de l'ONERA.
- Reviewer pour Oxford University Press et les pour les revues internationales suivantes : Automatica, SIAM J. Control and Optimization, IEEE Tr. Automatic Control, Intern. Journal of Robust and Nonlinear Control, Journal of Process Control, Intern. Journal of Control, IEE Proceedings Control Theory and Applications

- Membre de l'International Program Committee du Fifth IFAC Workshop on Time-Delay Systems (Leuven 2004)
- depuis 2000 conseil scientifique groupe TOTAL : réacteurs de polymérisation et systèmes de mélangeuses.

2.12 Distinctions

- Journal of Process Control Paper Prize : Best article 2002-2005 (Application) pour l'article :

N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, & P. Pinvidic (2002). Control of an industrial polymerization reactor using flatness. *Journal of Process Control*, 12(5) :659–665.

- Lauréat du 4ème prix au Concours International de Drones Miniatures ONERA/DGA pour le projet oiseau artificiel Sep. 2005

3 Liste complète de publications

La majorité de ces documents est consultable sur

`http ://cas.ensmp.fr/˜petit/index.html`

3.1 Revues internationales avec comité de lecture

F. Chaplais & N. Petit (2007 (à paraître)). Inversion in indirect optimal control of multivariable systems. *ESAIM : Control, Optimisation and Calculus of Variations.*

D. E. Chang, N. Petit, & P. Rouchon (2006). Time-optimal control of a particle in a dielectrophoretic system. *IEEE Trans. Automat. Control*, 51(7) :1100–1114.

J. Chauvin, G. Corde, , N. Petit, & P. Rouchon (2006 (à paraître)). Periodic input estimation for linear periodic systems : automotive engine applications. *Automatica.*

J. Chauvin, G. Corde, P. Moulin, N. Petit, & P. Rouchon (2006). High frequency individual cylinder estimation for control of Diesel engines. *Oil & Gas Science and Technology - Revue de l'Institut Français du Pétrole*, 61(1) :57–72.

D. E. Chang & N. Petit (2005). Toward control of dielectrophoretic system. *International Journal of Robust and Nonlinear Control*, 15(16) :769–784.

D. del Vecchio & N. Petit (2005). Boundary control for an industrial under-actuated tubular chemical reactor. *Journal of Process Control*, 15(7) :771–784.

J.-Y. Dieulot, N. Petit, P. Rouchon, & G. Delaplace (2005a). A torus model containing a sliding well-mixed zone as a way to represent mixing process at unsteady stirring conditions in agitated vessels. *Chemical Engineering Communications*, 192 :805–826.

——— (2005b). An arrangement of ideal zones with shifting boundaries as a way to model mixing processes in unsteady stirring conditions in agitated vessels. *Chemical Engineering Science*, 60(20) :5544–5554.

W. B. Dunbar, N. Petit, P. Rouchon, & P. Martin (2003). Motion planning for a nonlinear Stefan problem. *ESAIM : Control, Optimisation and Calculus of Variations*, 9 :275–296.

N. Petit & P. Rouchon (2002). Dynamics and solutions to some control problems for water-tank systems. *IEEE Trans. Automat. Control*, 47(4) :594–609.

N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, & P. Pinvidic (2002). Control of an industrial polymerization reactor using flatness. *Journal of Process Control*, 12(5) :659–665. Best paper in the category "application" for the period 2002 to 2005.

N. Petit, Y. Creff, L. Lemaire, & P. Rouchon (2001). Minimum time constrained control of acid strength on a sulfuric acid alkylation unit. *Chemical Engineering Science*, 56(8) :2767–2774.

N. Petit & P. Rouchon (2001). Flatness of heavy chain systems. *SIAM J. Control Optimization*, 40 :475–495.

M. Petit & N. Petit (1993). The search of chaotic attractor in mental diseases. *Annales Médico-psychologiques*, 151(10) :701–705.

3.2 Conférences internationales avec comité de lecture

J. Chauvin, G. Corde, & N. Petit (2007 (à paraître)). Transient control of a Diesel engine airpath. In *Proc. of the 2007 American Control Conference.*

D. Vissière, D. E. Chang, & N. Petit (2007 (à paraître)). Experiments of trajectory generation and obstacle avoidance for a UGV. In *Proc. of the 2007 American Control Conference.*

J. Chauvin, G. Corde, P. Moulin, N. Petit, & P. Rouchon (2006a). Kalman filtering for real-time individual cylinder air fuel ratio observer on a Diesel engine test bench. In *Proc. of the 2006 American Control Conference.*

J. Chauvin, G. Corde, & N. Petit (2006b). Constrained motion planning for the airpath of a Diesel HCCI engine. In *Proc. of the* 45*th IEEE Conf. on Decision and Control.*

J. Chauvin, G. Corde, N. Petit, & P. Rouchon (2006c). Experimental motion planning in airpath control for HCCI engine *(winner of the best presentation in session award).* In *Proc. of the 2006 American Control Conference.*

J. Chauvin, N. Petit, P. Rouchon, & G. Corde (2006d). Periodic input observer design : Application for imbalance diagnosis. In *Proc. of Society Automotive Engine World Congress.*

J. Chauvin, N. Petit, P. Rouchon, G. Corde, & C. Vigild (2006e). Air path estimation on Diesel HCCI engine. In *Proc. of Society Automotive Engine World Congress.*

J. Chauvin, N. Petit, P. Rouchon, P. Moulin, & G. Corde (2006f). Six degrees crankshaft individual air fuel ratio estimation of Diesel engines for cylinder balancing purpose. In *Proc. of Society Automotive Engine World Congress.*

L. Sinègre, N. Petit, & P. Ménégatti (2006a). Predicting instabilities in gas-lifted wells simulation. In *Proc. of the 2006 American Control Conference.*

L. Sinègre, N. Petit, & T. Saint-Pierre (2006b). Active control strategy for density-wave in gas-lifted wells. In *Proc. of the ADCHEM 2006, International Symposium on Advanced Control of Chemical Processes.*

J. Barraud, Y. Creff, & N. Petit (2005). Performances d'un prédicteur de Smith adaptatif pour un modèle de procédé avec retard variable. In *10ème Congrès de la Société française de génie des procédés.*

D. E. Chang, N. Petit, & P. Rouchon (2005). Time-optimal control of a particle in a dielectrophoretic system. In *Proc. of the* 16*th IFAC World Congress.*

J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, & P. Rouchon (2005). Real-time nonlinear individual cylinder air fuel ratio observer on a Diesel engine. In *Proc. of the* 16*th IFAC World Congress.*

L. Sinègre, N. Petit, P. Lemétayer, P. Gervaud, & P. Ménégatti (2005). Contrôle des puits activés en gas-lift. In *10ème Congrès de la Société française de génie des procédés.*

L. Sinègre, N. Petit, P. Lemétayer, P. Gervaud, & P. Ménégatti (2005a). Casing-heading phenomenon in gas-lifted well as a limit cycle of a 2d model with switches. In *Proc. of the* 16*th IFAC World Congress.*

L. Sinègre, N. Petit, & P. Ménégatti (2005b). Distributed delay model for density wave dynamics in gas lifted wells. In *Proc. of the 44th IEEE Conf. on Decision and Control.*

J. Barraud, Y. Creff, & N. Petit (2004). PI controllers performance for a process model with varying delay. In *Proc. of UKACC Int. Control Conference.*

J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, & P. Rouchon (2004a). Time-varying linear observer for torque balancing on a DI engine. In *Proc. of the First IFAC Symposium on Advances in Automative Control.*

——— (2004b). Observer design for torque balancing on a DI engine. In *Proc. of Society Automotive Engine World Congress.*

——— (2004c). Real-time combustion torque estimation on a Diesel engine test bench using time-varying Kalman filtering. In *Proc. of the 43rd IEEE Conf. on Decision and Control.*

——— (2004d). Real-time combustion torque estimation on a Diesel engine test bench using an adaptive Fourier basis decomposition. In *Proc. of the 43rd IEEE Conf. on Decision and Control.*

J. Chauvin, G. Corde, N. Petit, P. Rouchon, P. Moulin, & M. Castagné (2004e). Nonlinear output injection observer of indicated torque for DI engine. In *Proc. of the FISITA 04 World Automotive Congress.*

F. Chaplais & N. Petit (2003). Inversion in indirect optimal control. In *Proc. of the 7th European Control Conf.*

W. B. Dunbar, N. Petit, P. Rouchon, & P. Martin (2003). Boundary control for a nonlinear Stefan problem. In *Proc. of the 42nd IEEE Conf. on Decision and Control.*

T. Neckel, C. Talbot, & N. Petit (2003). Collocation and inversion for a reentry optimal control problem. In *Proc. of the 5th Intern. Conference on Launcher Technology.*

N. Petit, M. B. Milam, & R. M. Murray (2002). A new computational method for optimal control of a class of constrained systems governed by partial differential equations. In *Proc. of the 15th IFAC World Congress.*

N. Petit & P. Rouchon (2002). Flatness of heavy chain systems. In *Proc. of the 41th IEEE Conf. on Decision and Control.*

M. B. Milam, N. Petit, & R. M. Murray (2001). Constrained trajectory generation for micro-satellite formation flying. In *AIAA Guidance, Navigation and Control Conference*, pages 328–333.

N. Petit, M. B. Milam, & R. M. Murray (2001). Inversion based constrained trajectory optimization. In *5th IFAC Symposium on Nonlinear Control Systems.*

N. Petit & P. Rouchon (2001). Systèmes plats et contrôle des procédés. In *GP2001, 8^{eme} congrès francophone de génie des procédés.*

P. Brault, H. Mounier, N. Petit, & P. Rouchon (2000). Flatness based tracking of a manoeuvrable vehicle : the π-car. In *Proc. of the International Symposium Mathematical Theory of Systems, Control, Network.*

N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, & P. Pinvidic (2000). Control of an industrial polymerization reactor using flatness. In *Proc. of the International Symposium Mathematical Theory of Systems, Control, Network.*

F. Dubois, N. Petit, & P. Rouchon (1999). Motion planning and nonlinear simulations for a tank containing a fluid. In *European Control Conference, Karlsruhe.*

M. Fliess, P. Martin, N. Petit, & P. Rouchon (1999). Active signal restoration for the telegraph equation. In *Proc. of the* 38*th IEEE Conf. on Decision and Control.*

H. Mounier, M. Mboup, N. Petit, P. Rouchon, & D. Seret (1998). High speed network congestion control with a simplified time-varying delay model. In *Proc. Sys. Struct. and Control Conf.*

N. Petit, Y. Creff, & P. Rouchon (1998). Motion planning for two classes of nonlinear systems with delays depending on the control. In *Proc. of the* 37*th IEEE Conf. on Decision and Control*, pages 1007– 1011.

——— (1997). δ-freeness of a class of linear delayed systems. In *European Control Conference, Brussels.*

3.3 Chapitres d'ouvrages

R. M. Murray, J. Hauser, A. Jadbabaie, M. B. Milam, N. Petit, W. B. Dunbar, & R. Franz (2003). Online control customization via optimization-based control. In T. Samad & G. Balas, editors, *Software-Enabled Control, Information technology for dynamical systems*, pages 149–174. Wiley-Interscience.

N. Petit & P. Rouchon (2000). Motion planning for heavy chain systems. In A. Isidori, F. Lamnabhi-Lagarrigue, & W. Respondek, editors, *Lecture Notes in Control and Information Sciences 259 : Nonlinear Control in the Year 2000*, volume 2, pages 229–236. Springer.

N. Petit, P. Rouchon, J.-M. Boueilh, F. Guérin, & P. Pinvidic (2000). Control of an industrial polymerization reactor using flatness. In A. Isidori, F. Lamnabhi-Lagarrigue, & W. Respondek, editors, *Lecture Notes in Control and Information Sciences 259 : Nonlinear Control in the Year 2000*, volume 2, pages 237–244. Springer.

3.4 Autres publications

J. Chauvin, G. Corde, N. Petit, & P. Rouchon (2006). Filtre de Kalman ou observateur de Luenberger nonlinéaire ? comparaison expérimentale sur des exemples issus du contrôle moteur. In *Conférence Internationale Francophone d'Automatique.*

B. Laroche, P. Martin, & N. Petit (2004). Commande par platitude, équations différentielles ordinaires et aux dérivées partielles. Notes de cours, 78 pages, École Nationale Supérieure de Techniques Avancées.

H. Mounier, N. Petit, & P. Rouchon (2004). Quelques problèmes en mécanique. *E-sta, revue électronique de la Société de l'Électricité, de l'Électronique et des Technologies de l'Information et de la Communication*, 1(1).

N. Petit (2004). Optimisation. Notes de cours, 62 pages, École Nationale Supérieure des Mines de Paris.

——— (2003). Commande prédictive. Notes de cours option procédé et environnement, 22 pages, École Centrale Paris.

N. Petit & P. Rouchon (2002). *Théorie du contrôle non linéaire*, pages 60–67. Tangente Sup Sciences POLE.

M. B. Milam & N. Petit (2001). Constrained trajectory generation for a planar missile. Technical report, California Institute of Technology, Control and Dynamical Systems.

N. Petit (2000). *Systèmes à retards, platitude en génie des procédés et contrôle de certaines équations des ondes.* Ph.D. thesis, École des Mines de Paris.

M. Fliess, P. Martin, N. Petit, & P. Rouchon (1999). Commande de l'équation des télégraphistes et restauration active d'un signal. In *Colloque en l'honneur de Bernard Picinbono.* Paris.

P. Martin & N. Petit (1997). Commande non linéaire, le point de vue des systèmes plats. Notes de cours, 36 pages, École Centrale Paris.

N. Petit (1996). *Systèmes δ-libres sous contraintes.* Master's thesis, Univ. d'Orsay. Rapport de stage de DEA.

4 Encadrement

Étudiants en thèse

Thèses en cours

- D. Vissière. Ingénieur de l'Armement. Thèse en cours depuis Oct. 04 en partenariat DGA/LRBA. Sujet : contrôle coopératif de drones par platitude.

- O. Lepreux. Thèse CIFRE depuis Oct. 06 en partenariat avec l'IFP. Sujet : systèmes de dépollution par post-traitement pour moteurs Diesel.

- M. Hillion. Thèse CIFRE depuis Nov. 06 en partenariat avec l'IFP. Sujet : commande de la boucle d'air et de la boucle fuel pour moteurs Diesel.

- T. Leroy Thèse CIFRE depuis Oct. 06 en partenariat avec l'IFP. Sujet : commande haute fréquence de moteurs essence.

Thèses soutenues

- J. Barraud. Thèse CIFRE Oct. 03-Sept. 06 en partenariat avec l'IFP. Sujet : commande prédictive en génie des procédés pétrochimiques.

- J. Chauvin. Thèse CIFRE Sept. 03-Sept. 06 co-encadrée avec P. Rouchon en partenariat avec l'IFP. Sujet : contrôle de moteur diesel HCCI.

- L. Sinègre. Thèse CIFRE en cours Oct. 03-Sept. 06 en partenariat avec TOTAL. Sujet : contrôle des écoulements en puits pétroliers activés par gas-lift.

Post-doc

- K. graichen. PhD Stuttgart, sous la direction du Pr. M. Zeitz. Optimisation de trajecoires et inversion. Jan. 07 -
- D. E. Chang. PhD CalTech, sous la direction du Pr. J. Marsden. Contrôle optimal de systèmes de diélectrophorèse. Oct. 03-Oct. 04.

Stagiaires de DEA Automatique et Traitement du Signal. Université d'Orsay

- 2002-2003. J. Chauvin. Co-encadrement de stage. Estimation de couple par mesure d'angle pour moteurs Diesel. Sujet en commun avec l'IFP.

- 2002-2003. L. Sinègre. Encadrement de stage. Commande des puits en gas-lift. Sujet en collaboration avec TOTAL Exploration-Production.

- 2002-2003. D. Vissière. Encadrement de stage du DEA Automatique et Traitement du signal. Commande frontière optimale de l'équation de la chaleur monodimensionnelle par représentation intégrale.

Étudiants hors-thèse

- **California Institute of Technology**
  2003-2004. D. del Vecchio. Doctorante du Pr. Richard Murray accueillie en stage. Encadrement sur le contrôle frontière optimal d'un réacteur piston de polymérisation non-linéaire.
  2001-2002. W. B. Dunbar. Doctorant du Pr. Richard Murray accueilli en stage. Encadrement sur le contrôle d'un problème de Stefan non linéaire.

- **Lund LTH**
  2002-2003. O. Nilsson. Étudiant de Master of Science accueilli en Master thesis. Encadrement sur une méthode du second ordre pour le contrôle par platitude des équations de Saint-Venant.

- **École Polytechnique**
  2002-2003. T. Neckel. Élève du programme étranger Technische Universität München. Encadrement sur l'optimisation de trajectoires de réentrée par platitude pour engin spatial (collaboration avec le CNES).
  2001-2002. D. Vissière. X99. Encadrement de stage d'option sur l'algorithme d'ANAMEL à la raffinerie de Feyzin.
  1999-2000. J. Freysz. Optionnaire Génie des Procédés. Co-encadrement stage en usine (APPRYL PP2 Lavéra). Commande et estimation de paramètres par modèle physique.

- **École Centrale Paris**
  2004-2005. Responsable du projet Oiseau artificiel 7 élèves de deuxième année.
  2003-2004. Responsable du projet Oiseau artificiel 12 élèves de deuxième année.
  2002-2003. Responsable du projet Oiseau artificiel 10 élèves de deuxième année.
  1998-1999. M. Jribi et I. Ben Hania. Promotion 2000. Co-encadrement projet long

de recherche. Simulation et contrôle linéaire d'une colonne à distiller binaire.

- **ISIA** (Institut Supérieur d'Informatique et Automatique)
  1998-1999. F. Guérin. Encadrement stage long d'Automatique sur l'usine APPRYL PP2 Lavéra.

5 Enseignement

- **California Institute of Technology.**
  Spring 2001.
  Responsable du cours "*Optimal Control*" (Graduate level). 30 h de cours.
  `http ://www.cds.caltech.edu/~npetit/CDS270-1/`

- **École des Mines de Paris.**
  depuis 2005 Co-Responsable du cours de Tronc Commun d'Automatique 30 h/an.
  depuis 2003 Co-Responsable Enseignement Spécialisé Optimisation 40 h/an.
  2002-2005 Chargé de cours en Probabilités 20 h/an.
  1998-2000. et depuis 2001- Chargé de cours en Automatique 15 h/an.

- **École Nationale Supérieure de Techniques Avancées.**
  depuis 2002 Responsable Module d'Automatique non linéaire 21 h/an.

- **École Centrale Paris.**
  2001-2003 Responsable du module thématique Commande Optimale 30 h/an.
  depuis 2001 Co-Responsable du cours de Commande en génie des procédés 16 h/an.
  1996-2000. 2001-2005. Chargé de TD en Automatique 20 h/an.
  1996-2000 Chargé de cours thématique en Commande Non-linéaire 16 h/an.

- **Université d'Orsay.**
  1996-1999. Moniteur en licence EEA, TD d'Analyse Fonctionnelle 50 h/an.

Sélection de publications

Time-Optimal Control of a Particle in a Dielectrophoretic System

Dong Eui Chang, Nicolas Petit, and Pierre Rouchon

***Abstract*—We study the time-optimal control of a particle in a dielectrophoretic system. This system consists of a time-varying nonuniform electric field which acts upon the particle by creating a dipole within it. The interaction between the induced dipole and the electric field generates the motion of the particle. The control is the voltage on the electrodes which induces the electric field. Since we are considering the motion of a particle on an invariant line in a chamber filled with fluid flowing at low Reynolds number, the dynamics have a two dimensional state; one for the particle position and the other for the induced dipole moment. In regard to time-optimal control, we address the issue of existence and uniqueness of optimal trajectories, and explicitly compute the optimal control and the corresponding minimum time. Finally, we cast our analysis in the framework of symplectic reduction theory in order to provide geometric insight into the problem.**

***Index Terms*—Biotechnology, dielectrophoresis, nanotechnology, time-optimal control.**

I. Introduction

WE STUDY the time-optimal control of the following system:

$$\dot{x} = yu + \alpha u^2 \tag{1}$$

$$\dot{y} = -cy + u \tag{2}$$

with the state $(x, y) \in \mathbb{R}^2$ and the single control u satisfying

$$x(0) = x_0 = \text{ given} \quad y(0) = 0 \tag{3}$$

$$x(t_f) = x_f = \text{ given} \quad y(t_f) = \text{ free} \tag{4}$$

$$|u| \leq 1 \tag{5}$$

where the parameters α and c satisfy

$$\alpha < 0 \quad c > 0. \tag{6}$$

These dynamics describe, after a nonlinear change of coordinates, the motion of a neutrally buoyant and neutrally charged particle on an invariant line in a chamber filled with fluid flowing at low Reynolds number and with a parallel electrode array at the bottom of the chamber. The existence of the invariant line is due to symmetry in the arrangement of electrodes and the boundary potential on electrodes. The motion is created by the interaction between a nonuniform electric field and the dipole moment induced in the particle. This motion is called dielectrophoresis (DEP) [11]. Dielectrophoresis has wide applications in nano/bio-technology, in particular, in manipulating, separating and identifying nano/bio-particles [7], [8].

A brief explanation of the dynamics is in order. The variable x describes the displacement of the particle. The variable y describes the exponentially decaying part of the induced dipole moment. Voltage u is given on every other electrode and $(-u)$ on the others. Parameters α and c depend on the permittivities and conductivities of the particle and the fluid medium. The positivity of c is imposed by physics, but the negativity of α is arbitrary. As one is not interested in the final value of the induced dipole moment, the final value of y is free in (4).

Our goal in this paper is to study the minimum time trajectories of this system. A complete solution of this problem in a general setup—for a set of particles in a three-dimensional space, for instance—would allow significant improvement in DEP-based devices for particle analysis such as detecting cancers cells and separating different cells. However, here we address a simple case, which still includes the key feature of dielectrophoresis. Various control problems on dielectrophoresis in nano/biotechnology are suggested in [5].

This paper is organized as follows. We first overview the main results of the paper. Second, we derive the dynamics from physics. Third, we study the nonexistence of Lebesgue measurable time-optimal control for $x_f < x_0$ even though the target point is reachable. Fourth, we show that $(1 + \alpha c) > 0$ is the necessary and sufficient condition, under assumption (6), for the existence of time-optimal controls when $x_f > x_0$. Fifth, we address the issue of uniqueness of time-optimal control when $x_f > x_0$. We find a condition on α and c which guarantees the uniqueness of optimal trajectories, and compute the minimum time and the optimal control for a given target point x_f. For the case that the uniqueness condition is not satisfied, we give a constructive algorithm with which we can easily find a time-optimal control. Sixth, we make a discussion on the case where the optimal trajectories derived above are still valid in the presence of a state constraint on x such as $x \geq 0$ or $x \leq 0$. Seventh, we give geometric insight into the problem by putting the previous analysis in the picture of symplectic reduction theory. Finally, we perform some simulations to demonstrate the result.

II. Overview of Main Results

Time optimal trajectories satisfy the following dynamics:

$$\dot{x} = yu + \alpha u^2 \tag{7}$$

$$\dot{y} = -cy + u \tag{8}$$

$$\dot{\lambda} = c\lambda - u \tag{9}$$

Manuscript received January 31, 2005; revised November 21, 2005. Recommended by Associate Editor U. Jonsson.

D. E. Chang is with the Department of Applied Mathematics, Faculty of Mathematics, the University of Waterloo, Waterloo, ON N2L 3G1, Canada.

N. Petit and P. Rouchon are with the Centre Automatique et Systèmes, the Ecole Nationale Supérieure des Mines de Paris, Paris 75272, France.

Digital Object Identifier 10.1109/TAC.2006.878748

0018-9286/$20.00 © 2006 IEEE

with conditions (3)–(6) and

$$\lambda(0) = \lambda_0 = \text{ to be found } \quad \lambda(t_f) = 0 \tag{10}$$

where

$$u(t) = \arg\max_{|v|\le 1} \tilde{H}(y(t), \lambda(t), v) \tag{11}$$

with

$$\tilde{H}(y, \lambda, v) = \alpha u^2 + (y+\lambda)u - cy\lambda. \tag{12}$$

The control u satisfying (11) is given by

$$u = \begin{cases} +1, & \text{if } y+\lambda > -2\alpha \\ \frac{y+\lambda}{-2\alpha}, & \text{if } |y+\lambda| \le -2\alpha \\ -1, & \text{if } y+\lambda < 2\alpha \end{cases}$$

because $\alpha < 0$. A symplectic reduction picture is hidden in (7)–(12).

When $x_f < x_0$, there are no Lebesgue measurable time-optimal control functions resulting in $x(t_f) = x_f$ even though x_f is reachable.

We now consider the case of $x_f > x_0$. Time-optimal control law exists if and only if the parameters, α and c satisfy $(1 + \alpha c) > 0$, which is assumed in the following of this section. Define the open interval Λ by

$$\Lambda = (-2\alpha\sqrt{1+\alpha c}, -2\alpha\sqrt{(1+\alpha c)/(-\alpha c)}) \tag{13}$$

if $(1 + 2\alpha c) \le 0$, and

$$\begin{aligned} \Lambda &= \Lambda_1 \cup \Lambda_2 \\ &= (-2\alpha\sqrt{1+\alpha c}, -2\alpha) \cup [-2\alpha, 1/c) \end{aligned} \tag{14}$$

if $(1+2\alpha c) > 0$. For the sake of convenience, let us define four sentences as follows:

- P1 := $[(1 + 2\alpha c) > 0] \wedge [\lambda_0 \in \Lambda_1]$;
- P2 := $[(1 + 2\alpha c) > 0] \wedge [\lambda_0 \in \Lambda_2]$;
- Q := $[(1 + 2\alpha c) \le 0] \wedge [\lambda_0 \in \Lambda]$;
- R := $[(3 + 4\alpha c) \ge 0] \wedge [\lambda_0 \in \Lambda]$;

where $\wedge$ is the logical connective AND. Let us define a strictly increasing onto function $X_1 : \Lambda \to (0, \infty)$ by

$$X_1(\lambda_0) = \begin{cases} \text{equation (63)}, & \text{if P1} \vee \text{Q} \\ \text{equation (67)}, & \text{if P2} \end{cases} \tag{15}$$

where $\vee$ is the logical connective OR. Let us define another function T_1 on Λ by

$$T_1(\lambda_0) = \begin{cases} \text{equation (73)}, & \text{if P1 } \vee \text{Q} \\ \text{equation (74)}, & \text{if P2.} \end{cases} \tag{16}$$

We call a trajectory of (7)–(9) satisfying (3)–(5) and (10)–(11), an *extremal*. Let us call an arc of an extremal a *basic arc* if the projection of the arc onto the $y - \lambda$ plane starts from $0 \times \Lambda$ (respectively, $0 \times (-\Lambda)$) and ends on $\Lambda \times 0$ (respectively, $(-\Lambda) \times 0$), going through the first (respectively, third) quadrant of the $y - \lambda$ plane. We call an extremal an *n-shot extremal* with $n \in \mathbb{N}$ if the maximum number of basic arcs in the extremal is n.

To decompose extremals into finite arcs when [P1 $\vee$ Q], let us introduce some notation. An arc associated with the linear control $u = (y+\lambda)/(-2\alpha)$ on a time interval of length $\Delta t_{AB}(\lambda_0)$ in (71), is denoted by $\gamma_{\mathrm{L}}^{\lambda_0}$. Let $\gamma_{+}^{\lambda_0}$ (respectively, $\gamma_{-}^{\lambda_0}$) denote an arc with $u = 1$ (respectively, $u = -1$) on a time interval of length $\Delta t_{BC}(\lambda_0)$ in (72). Define two arcs $\Gamma_{\pm}^{\lambda_0}$ by the concatenation

$$\Gamma_{\pm}^{\lambda_0} = \gamma_L^{\lambda_0} \star \gamma_{\pm}^{\lambda_0} \star \gamma_L^{\lambda_0} \tag{17}$$

where the concatenation $\star$ is defined such that the leftmost one comes first and the rightmost one comes last. An arc with the linear control $u = (y+\lambda)/(-2\alpha)$ is called an *idling arc* if its projection (y, λ) starts from the positive (respectively, negative) y-axis, goes through the fourth (respectively, second) quadrant in the $y - \lambda$ plane, and finally ends at the negative (respectively, positive) λ-axis. An idling arc is denoted by γ_{idling}; see Fig. 8. Its duration T_{idling} is given in (76). Hence, when [P1 $\vee$ Q], we can express n-shot extremals as $\Gamma_{\pm}^{\lambda_0,n}$ for $n \in \mathbb{N}$, which are defined as follows:

$$\Gamma_{\pm}^{\lambda_0,1} = \Gamma_{\pm}^{\lambda_0} \tag{18}$$

$$\Gamma_{\pm}^{\lambda_0,k} = \begin{cases} \Gamma_{\pm}^{\lambda_0,k-1} \star \gamma_{\text{idling}} \star \Gamma_{\mp}^{\lambda_0}, & \text{if } k \text{ is even} \\ \Gamma_{\pm}^{\lambda_0,k-1} \star \gamma_{\text{idling}} \star \Gamma_{\pm}^{\lambda_0}, & \text{if } k \text{ is odd} \end{cases} \tag{19}$$

for $k \ge 2$ with

$$\lambda(0) = \begin{cases} +\lambda_0, & \text{for } \Gamma_{+}^{\lambda_0,m} \\ -\lambda_0, & \text{for } \Gamma_{-}^{\lambda_0,m} \end{cases}$$

for $m \ge 1$.

We now discuss the existence and uniqueness of optimal trajectories. If $(3+4\alpha c) \ge 0$, there exist exactly two time-optimal trajectories for $x_f > x_0$, and they are basic arcs. Here is the procedure of constructing them.

[A.1.] Find $\lambda_0 = X_1^{-1}(x_f - x_0) \in \Lambda$.

[A.2.] Set $\lambda(0) = \pm\lambda_0$.

[A.3.] The minimum time cost is $T_1(\lambda_0)$ and the optimal trajectories γ_{opt} are

$$\gamma_{\text{opt}} = \begin{cases} \Gamma_{\pm}^{\lambda_0}, & \text{if R} \wedge \neg\text{P2} \\ \text{basic arc with } u = \pm 1, & \text{if P2} \end{cases}$$

where $\neg$ is the logical connective NOT.

If $(3 + 4\alpha c) < 0$, we do not have any general proof of the uniqueness of optimal control. However, we have a finite procedure of finding all optimal control laws for $x_f > x_0$ as follows.

[B.1.] Define two sequences, for $k \in \mathbb{N}$

$$\begin{aligned} \lambda_{0,k} &= X_1^{-1}\left(\frac{x_f - x_0}{k}\right) \\ T_k &= k\,T_1(\lambda_{0,k}) + (k-1)T_{\text{idling}}. \end{aligned}$$

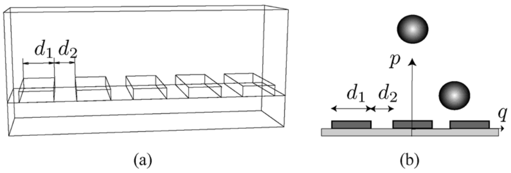


Fig. 1. Dielectrophoretic system is a chamber filled with a fluid medium where there is a parallel array of electrodes at the bottom.

[B.2.] Find $n = \arg\min_k\{T_k : k \in \mathbb{N}\}$ (such an n always exists and is less than $1 + (T_1/T_{\text{idling}})$).

[B.3.] Set $\lambda(0) = \pm\lambda_{0,n}$.

[B.4.] The minimum time cost is T_n and the corresponding optimal trajectories are $\Gamma_{\pm}^{\lambda_0,n}$.

If there are j integers in step [B.2.] to give the minimum time, then there are exactly $2j$ time-optimal trajectories. We remark that each basic arc in the k-shot extremal $\Gamma_{\pm}^{\lambda_{0,k},k}$ equally contributes $(x_f - x_0)/k$ to the increment of x and the idling arcs in between make no contribution.

In Section VI-H, we discuss the possibility that some optimal trajectories derived previously become optimal even with state constraints $x \geq 0$ or $x \leq 0$. In Section VII, we draw a symplectic reduction picture hidden in the problem and our solution.

III. Derivation of Dynamics

We briefly derive the dynamics in (1) and (2), and explain the conditions in (3)–(6); see [4] for more details. Consider a neutrally charged particle in a chamber with a fluid medium and a parallel electrode array at the bottom as in Fig. 1(a) where d_1 is the width of each electrode, and d_2 is the width of the gap between two electrodes. As the electrodes are very long compared with the size of particles, we may assume that there are infinite number of infinitely long electrodes. Due to this symmetry, we can consider the motion of the particle in the vertical plane as in Fig. 1(b).

Let $(q,p) \in \mathbb{R}^2$ be the coordinates in Fig. 1(b). We give the boundary voltage

$$V_{\text{bd}}(t) = V_0 \cdot u(t), \qquad |u(t)| \leq 1$$

on every other electrode and $(-V_{\text{bd}}(t))$ on the others. This creates potential $V(q,p,t)$ in $\{p \geq 0\}$. The electric field vector $\mathbf{E}(q,p,t) \in \mathbb{R}^2$ in $\{p \geq 0\}$ is given by

$$\mathbf{E}(q,p,t) = -\nabla V(q,p,t).$$

This electric field induces a dipole moment $\mathbf{m}$ in a single-layered spherical particle as follows:

$$\mathbf{m}(q,p,t) = g(t) * \mathbf{E}(q,p,t)$$

where $*$ denotes the usual convolution operator with respect to time t and the Laplace transform $G(s)$ of the (transfer) function $g(t)$ is given by

$$G(s) = a + \frac{b}{s+c}$$

where

$$\begin{aligned} a &= 4\pi r^3 \epsilon_m(\epsilon_p - \epsilon_m)/(\epsilon_p + 2\epsilon_m) \\ b &= a\left(\frac{\sigma_p - \sigma_m}{\epsilon_p - \epsilon_m} - \frac{\sigma_p + 2\sigma_m}{\epsilon_p + 2\epsilon_m}\right) \qquad (20) \\ c &= (\sigma_p + 2\sigma_m)/(\epsilon_p + 2\epsilon_m) \qquad (21) \end{aligned}$$

where r is the radius of the particle, ϵ_p (respectively, ϵ_m) is the permittivity of the particle (resp., medium) and σ_p (respectively, σ_m) is the conductivity of the particle (respectively, medium). The interaction between the electric field and the induced dipole moment creates a force $\mathbf{F}_{\text{dep}}$. It is called *dielectrophoretic force* and is given by

$$\mathbf{F}_{\text{dep}}(q,p,t) = (\mathbf{m}(q,p,t) \cdot \nabla)\mathbf{E}(q,p,t).$$

We restrict our interest to the motion of a particle on the p-axis because it can practically represent the vertical motion of all particles in the whole chamber. One can check that the dielectrophoretic force on the p-axis is parallel to this axis due to the symmetry in the boundary voltage. This vertical dielectrophoretic force on the p-axis is denoted by $F_{\text{dep}}(p,t)$. It is of the form

$$F_{\text{dep}}(p,t) = F(p)u(t)(g * u)(t)$$

where

$$F(p) = \frac{-\pi^3 V_0^2 e^{-\frac{\pi p}{2d}}\left(1 - e^{-\frac{\pi p}{d}}\right)}{2d^3 K^2\left(\cos\left(\frac{d_2\pi}{4d}\right)\right)\left[1 - 2e^{-\frac{\pi p}{2d}}\cos\left(\frac{d_1\pi}{2d}\right) + e^{-\frac{\pi p}{d}}\right]^2}$$

where K is the complete elliptic function of the first kind and $d = (d_1 + d_2)/2$; see [4] for the derivation of F. Notice that $F(p) \leq 0$ on $p \geq 0$, $\lim_{p\to\infty} F(p) = 0$, and $F(p) = 0$ only at $p = 0$.

Let us assume that *the particle is neutrally buoyant and the medium fluid flows at low Reynolds number*. Thus, the gravitational force and the buoyant force cancel and the inertial term $m\ddot{p}$ is trivial. The only forces on the particle are the drag and the DEP force. Hence, the motion of the particle on the p-axis can be described by

$$f\dot{p} + F(p)u(t)(g * u)(t) = 0. \qquad (22)$$

where $f > 0$ is the drag constant.

We assume that b *in* (20) *is nonzero*, which generically holds. Then, (1) and (2) come from (22) where x and y are defined by

$$\begin{aligned} x &= \int_\epsilon^p \frac{-f}{b\,F(z)}\mathrm{d}z \\ Y(s) &= \frac{1}{s+c}U(s) \qquad (23) \end{aligned}$$

for $p \geq \epsilon$ where ϵ is a positive number and $Y(s)$ and $U(s)$ are the Laplace transforms of $y(t)$ and $u(t)$, and α is defined by

$$\alpha = a/b.$$

If a particle is close to the electrode, then additional physical/chemical forces other than the DEP force start to appear in the dynamics [7], [8], [11], so the parameter ϵ in (23) defines the region where the dynamics (22) is valid. Physically, y is the exponentially induced part of the dipole moment, so we have the initial condition $y(0) = 0$. As we are not interested in the final state of the induced dipole moment, we have $y(t_f) = \text{free}$.

Depending on the sign of b, the original region $\{p \geq \epsilon\}$ is mapped to $\{x \geq 0\}$ or $\{x \leq 0\}$. In this paper, we ignore this state constraint on x, allowing for x to be on the whole real line. In Section VI-H, we discuss the possibility that the time optimal trajectories without the state constraints remain optimal with the state constraints. In a future publication, we will address the optimization problem with the state constraint on x.

We also make the following assumption on the signs of parameters α and c

$$\alpha < 0 \quad c > 0.$$

The assumption $c > 0$ is imposed by physics; see (21). However, the condition $\alpha < 0$ is chosen for convenience. The case where $\alpha \geq 0$ is left for future work.

IV. Nonexistence of Optimal Control for $x_f < x_0$

We will show that there are no time-optimal (Lebesque) measurable control functions for $x_f < x_0$ even though x_f is reachable.

Fix a $T > 0$. Let us define a sequence of functions $\{u_n^T : [0, T] \to \pm 1\}_{n \in \mathbb{N}}$ as follows:

$$u_n^T(t) = \operatorname{sign}(\sin(2\pi n t/T)). \tag{24}$$

It is straightforward to prove the following lemma.

Lemma 4.1: For any continuous function f on $[0, T]$

$$\lim_{n \mapsto \infty} \int_0^t f \cdot u_n^T = 0 \tag{25}$$

uniformly in $t \in [0, T]$.

The substitution of u in (2) to (1) yields

$$\dot{x} = y\dot{y} + cy^2 + \alpha u^2.$$

It follows that

$$\begin{aligned} &x(T) - x(0) \\ &\quad = \int_0^T \dot{x}\, dt \\ &\quad = \frac{1}{2}y(T)^2 + c\int_0^T y(\tau)^2 d\tau + \alpha \int_0^T u(\tau)^2\, d\tau \qquad (26) \\ &\quad \geq \alpha \int_0^T u(\tau)^2 d\tau \\ &\quad \geq \alpha T \qquad (27) \end{aligned}$$

because $\alpha < 0, x(0) = 0$, and $|u| \leq 1$. This implies that $x(0) + \alpha T$ is a lower bound of $x(T)$ for any admissible control u. Let $(x_n(t), y_n(t)), 0 \leq t \leq T$ be the solution to (1) and (2) with control u_n^T in (24). In particular

$$y_n(t) = e^{-ct} \int_0^t e^{c\tau} u_n^T(\tau)\, d\tau. \tag{28}$$

Given $\epsilon > 0$, by Lemma 4.1, there exists $N \in \mathbb{N}$ such that

$$|y_n(t)| \leq \sqrt{\epsilon} \tag{29}$$

for all $t \in [0, T]$ and all $n \geq N$. By (26), (29) and the definition of u_n^T, we have

$$|x_n(T) - x(0) - \alpha T| \leq \left(\frac{1}{2} + cT\right)\epsilon$$

for all $n > N$. Hence

$$\lim_{n \to \infty} x_n(T) = x(0) + \alpha T.$$

We have constructed a sequence $\{u_n^T\}$ of control laws such that the corresponding $\{x_n(T)\}$ converges to the lower bound $x(0) + \alpha T$ of the reachable point of x in time T. Notice that the sequence of functions $\{u_n^T\}$ does not converge to a measurable function. The following lemma addresses this issue.

Lemma 4.4: For a given $T > 0$, there exists no measurable control function $u : [0, T] \to [-1, 1]$ such that $x(T) = x(0) + \alpha T$.

Proof: The proof is by contradiction. Suppose that there is a measurable function $u : [0, T] \to [-1, 1]$ such that $x(T) = x(0) + \alpha T$. By (27), it follows that

$$y(t) = 0 \quad \text{a.e. on } [0, T] \tag{30}$$

$$u(t) = \pm 1 \text{ a.e. on } [0, T]. \tag{31}$$

Hence, for almost all $t \in [0, T]$

$$\begin{aligned} 0 = y(t) &= \int_0^t \dot{y}(s)\, ds \\ &= \int_0^t (-cy(s) + u(s))\, ds = \int_0^t u(s)\, ds. \end{aligned}$$

By [6, Th. 4.9], the function $t \mapsto \int_0^t u(s)\, ds$ is continuous. It follows that

$$\int_0^t u(s)\, ds = 0$$

for *all* $t \in [0, T]$. By [6, 4.11], $u(t) = 0$ for almost all $t \in [0, T]$ which contradicts (31). Thus, there exists no such measurable function u that produces $x(T) = x(0) + \alpha T$. ■

We have shown the following.

Claim 4.3: For $x_f < x_0$, the infimum of the time cost is $T = (x_0 - x_f)/(-\alpha)$, but there are no time-optimal (Lebesgue) measurable controls to reach x_f in time T.

Remark 4.4: It will become an interesting project to extend the technology in [10] and [3] in order to show the nonexistence of optimal trajectories in an alternative way.

Remark 4.5: For $x_f = x_0$, the control $u = 0$ with $t_f = 0$ is trivially the time-optimal control.

V. Pontryagin Maximum Principle

We derive, from the Pontryagin Maximum Principle (PMP), a necessary condition which time-optimal trajectories must satisfy. Let us define the PMP Hamiltonian H for the time-optimal control as follows (see [12] and [2]):

$$H(x, y, \lambda_x, \lambda_y, u) = \lambda_x \alpha u^2 + (\lambda_x y + \lambda_y)u - cy\lambda_y \tag{32}$$

where (λ_x, λ_y) is a covector. Let

$$M^\circ(x, y, \lambda_x, \lambda_y) = \max_{|u|\le 1} H(x, y, \lambda_x, \lambda_y, u). \tag{33}$$

The application of the PMP in [12] gives:

Theorem 5.1: Consider system (1), (2) with conditions (3)–(6). Let $u(t)$ be a time-optimal control and $(x(t), y(t))$ be the corresponding trajectory. Then, it is necessary that there exists a continuous covector $(\lambda_x(t), \lambda_y(t))$, which is not identically zero, such that

$$\begin{aligned} \dot{x} &= yu + \alpha u^2 && (34)\\ \dot{y} &= -cy + u && (35)\\ \dot{\lambda}_x &= 0 && (36)\\ \dot{\lambda}_y &= c\lambda_y - \lambda_x u. && (37) \end{aligned}$$

Additionally, the following must be satisfied:

1. $u(t) = \arg\max_{|v|\le 1} H(x(t), y(t), \lambda_x(t), \lambda_y(t), v)$ (38)
2. $M^\circ(x(t), y(t), \lambda_x(t), \lambda_y(t)) = \text{constant} \ge 0$ (39)
3. $\lambda_y(t_f) = 0$ (transversality condition). (40)

The boundary conditions of (x, y) and the signs of parameters are given in (3)–(6). The transversality condition in (40) comes from the free final boundary condition on y in (4).

Remark 5.2: Since $x(t) \in \mathbb{R}$, one can equivalently formulate the previous time-optimal control problem as follows:

$$\text{maximize } x(T)$$

for $T \ge 0$.

VI. Analysis of Extremals

We study the dynamics in (34)–(37). We call the trajectories satisfying the dynamics and all the conditions in Theorem 5.1, extremals.

A. The Necessary Positivity of Λ_x

By (36), $\lambda_x = \lambda_x(t)$ is constant in t, so there can be the following three cases:

$$\lambda_x = 0 \quad \lambda_x < 0 \quad \lambda_x > 0.$$

We will show that extremals exist only if $\lambda_x > 0$.

First, we assume $\lambda_x = 0$. Then (36) and (37) become a linear ordinary differential equation in (λ_x, λ_y) with $(\lambda_x(t_f), \lambda_y(t_f)) = (0, 0)$. By the uniqueness theorem of solutions of ordinary differential equations [1] or by direct computation, we have $(\lambda_x(t), \lambda_y(t)) = 0$ for all t. Hence, by the PMP, there are no optimal trajectories when $\lambda_x = 0$.

We now assume $\lambda_x < 0$. Let

$$\begin{aligned} \lambda &= \lambda_y/\lambda_x \\ \tilde{H}(y, \lambda, u) &= H(x, y, \lambda_x, \lambda_y, u)/\lambda_x \\ &= \alpha u^2 + (y + \lambda)u - cy\lambda. \end{aligned} \tag{41}$$

Then, (38) implies

$$u = \arg\min_{|v|\le 1} \tilde{H}(y, \lambda, v).$$

One can compute (recall that $\alpha < 0$)

$$u(t) = -\text{sign}(y(t) + \lambda(t)).$$

With this control, the dynamics of (y, λ) can be written as

$$\begin{bmatrix} \dot{y} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} y \\ \lambda \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \cdot \text{sign}(y + \lambda). \tag{42}$$

Suppose $\lambda(0) > 0$. Since $y(0) = 0, y(0) + \lambda(0) > 0$. One can compute the flow of (42)

$$y(t) = \frac{1}{c}(e^{-ct} - 1) \quad \lambda(t) = \lambda(0)e^{ct} + \frac{1}{c}(e^{ct} - 1).$$

It follows

$$y(t) + \lambda(t) = \lambda(0)e^{ct} + \frac{2}{c}(\cosh(ct) - 1) > 0.$$

Hence, $u(t) = -\text{sign}(y(t) + \lambda(t)) = -1$ for all $t \ge 0$, which implies that $\lambda(t) > 0$ for all $t \ge 0$. Hence, the transversality condition (40) cannot be satisfied. One can show in the similar manner that the transversality condition cannot be satisfied when $\lambda(0) < 0$ or $\lambda(0) = 0$. Therefore, there are no optimal trajectories if $\lambda_x < 0$. Hence, we have proved the following claim.

Claim 6.1: Along every extremal, we have $\lambda_x > 0$.

B. A Necessary Condition on Parameters for the Existence of Optimal Trajectories

We derive a necessary condition on the parameters α and c so that optimal trajectories exist. By Claim 6.1, we assume $\lambda_x > 0$ in the following.

Let

$$\begin{aligned} \lambda &= \lambda_y/\lambda_x && (43)\\ \tilde{H}(y,\lambda,u) &= H(x,y,\lambda_x,\lambda_y,u)/\lambda_x \\ &= \alpha u^2 + (y+\lambda)u - cy\lambda \\ M(y,\lambda) &= \max_{|v|\le 1} \tilde{H}(y,\lambda,v) && (44)\end{aligned}$$

and

$$\begin{aligned} R_n &= \{(y,\lambda) \mid y+\lambda < 2\alpha\}\\ R_p &= \{(y,\lambda) \mid y+\lambda > -2\alpha\}\\ R_l &= \{(y,\lambda) \mid |y+\lambda| \le -2\alpha\}.\end{aligned}$$

By (38) and (44)

$$u = \arg\max_{|v|\le 1} \tilde{H}(y,\lambda,v).$$

Let us compute u in each region of R_n, R_p, and R_l and study the dynamics in each region.

First, we consider the case where $(y,\lambda) \in R_n$. Then, $u = -1$. The (y,λ)-dynamics become

$$\dot{y} = -cy - 1 \quad \dot{\lambda} = c\lambda + 1 \tag{45}$$

where the equilibrium at $(y,\lambda) = (-1/c, -1/c)$ is a saddle.

Second, if $(y,\lambda) \in R_p$, then $u = 1$. The (y,λ)-dynamics become

$$\dot{y} = -cy + 1 \quad \dot{\lambda} = c\lambda - 1 \tag{46}$$

where the equilibrium at $(y,\lambda) = (1/c, 1/c)$ is a saddle.

Finally, if $(y,\lambda) \in R_l$, then $u = (y+\lambda)/(-2\alpha)$. The Hamiltonian M becomes

$$M(y,\lambda) = \frac{(y+\lambda)^2}{-4\alpha} - cy\lambda \tag{47}$$

and the (x,y,λ)-dynamics become

$$\dot{x} = \frac{y^2 - \lambda^2}{-4\alpha} \tag{48}$$

$$\begin{bmatrix}\dot{y}\\ \dot{\lambda}\end{bmatrix} = A \begin{bmatrix} y \\ \lambda\end{bmatrix} \tag{49}$$

where

$$A = \begin{bmatrix} -\left(c+\frac{1}{2\alpha}\right) & -\frac{1}{2\alpha} \\ \frac{1}{2\alpha} & \left(c+\frac{1}{2\alpha}\right)\end{bmatrix}. \tag{50}$$

The matrix A satisfies

$$\operatorname{tr} A = 0 \quad \det A = c(1+\alpha c)/(-\alpha).$$

The type of the equilibrium at $(y,\lambda) = (0,0)$ depends on the sign of $\det A$.

We now make qualitative phase portraits of the (y,λ)-dynamics in the following three different cases:

$$(1+\alpha c) < 0 \quad (1+\alpha c) = 0 \quad (1+\alpha c) > 0.$$

Suppose $(1+\alpha c) < 0$. Then, $\det A < 0$. The origin $(y,\lambda) = (0,0)$ is a saddle point of (49). The (real) eigenvalues of A are given by

$$\mu_s(A) = \frac{\sqrt{\alpha c(1+\alpha c)}}{\alpha} \quad \mu_u(A) = \frac{\sqrt{\alpha c(1+\alpha c)}}{-\alpha}$$

and the corresponding eigenvectors are given by

$$\begin{aligned} v_s(A) &= (1, (\sqrt{-\alpha c} - \sqrt{-(\alpha c+1)})^2)^T \\ v_u(A) &= (1, (\sqrt{-\alpha c} + \sqrt{-(\alpha c+1)})^2)^T.\end{aligned}$$

Notice that $\mu_s(A) < 0$ and $\mu_u(A) > 0$. The second (or λ-) components of $v_s(A)$ and $v_u(A)$ satisfy

$$0 < v_{s,\lambda}(A) < 1 < v_{u,\lambda}(A) \tag{51}$$

because $(1+\alpha c) < 0$. In addition, the two saddle points $\pm(1/c, 1/c)$ of the dynamics in (45) and (46) do not belong to $R_n \cup R_p$ but to R_l since $(1+\alpha c) < 0$. Gathering the information in each region of R_n, R_p, and R_l, we can draw a phase portrait of the (y,λ)-dynamics, qualitatively, in Fig. 2. By (51), the slope of the unstable (respectively, stable) manifold of the origin is greater (respectively, smaller) than 1. As $y(0) = 0$ and $\lambda(t_f) = 0$, time-optimal trajectories must start from the λ-axis and ends on the y-axis in the y-λ plane. However, there are no such orbits in Fig. 2 because the stable and unstable manifolds of the saddle points $\pm(1/c, 1/c)$ prevent it in $R_n \cup R_p$. This implies that there are no optimal trajectories when $\lambda_x > 0$ and $(1+\alpha c) < 0$.

If $(1+\alpha c) = 0$, then the matrix A in (50) becomes

$$A = \begin{bmatrix} -\frac{c}{2} & \frac{c}{2} \\ -\frac{c}{2} & \frac{c}{2}\end{bmatrix}.$$

The integration of (49) gives $\lambda(t) = y(t) + c$ where c is a constant. In particular, the line $\lambda = y$ in R_l is a set of equilibria. The phase portrait in the $y - \lambda$ plane is given in Fig. 3. One can see that no trajectories starting from the λ-axis reach the y-axis.

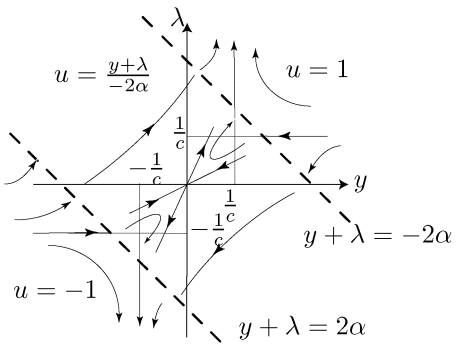


Fig. 2. Phase portrait in the y-λ plane when $\lambda_x > 0$ and $(1 + \alpha c) < 0$.

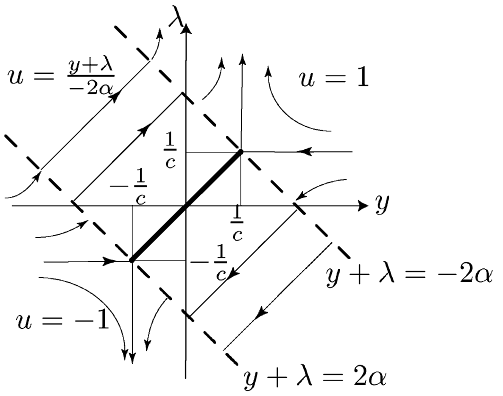


Fig. 3. Phase portrait in the $y - \lambda$ plane when $\lambda_x > 0$ and $(1 + \alpha c) = 0$. The segment between $(-1/c, -1/c)$ and $(1/c, 1/c)$ is the set of equilibria.

Hence, there are no time-optimal trajectories when $\lambda_x > 0$ and $(1 + \alpha c) = 0$.

We have, so far, proved the following.

Claim 6.2: Time-optimal trajectories exist only if $(1+\alpha c) > 0$.

When $(1+\alpha c) > 0$, the fixed point $(y, \lambda) = (0, 0)$ of the dynamics (49) is a center as $\det A > 0$ with A in (50). In this case, there are two possible qualitatively different phase portraits of the (y, λ)-dynamics depending on the position of the λ-intercept of the switching line $y + \lambda = -2\alpha$ relative to the point $(0, 1/c)$ on the λ-axis. They are given in Fig. 4 depending on the sign of $(1 + 2\alpha c)$, with control

$$u = \begin{cases} +1, & \text{if } y + \lambda > -2\alpha \\ \frac{y+\lambda}{-2\alpha}, & \text{if } |y + \lambda| \leq -2\alpha \\ -1, & \text{if } y + \lambda < 2\alpha. \end{cases} \tag{52}$$

We will show that the (y, λ)-projection of optimal trajectories must be contained in the shaded region in Fig. 4. In the rest of this paper, we assume that $(1 + \alpha c) > 0$.

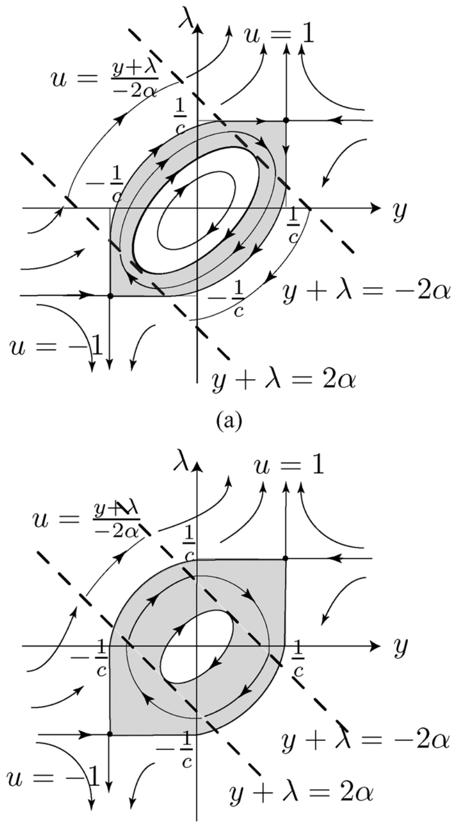
C. Discrete Symmetry

We will find $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry in the dynamics. Define the following maps:

$$S_1 : (x, y, \lambda) \mapsto (x, \lambda, y) \tag{53}$$

$$S_2 : (x, y, \lambda) \mapsto (x, -\lambda, -y). \tag{54}$$

$$S_3 := S_1 \circ S_2 : (x, y, \lambda) \mapsto (x, -y, -\lambda). \tag{55}$$



(a)

(b)

Fig. 4. Phase portrait in the $y - \lambda$ plane when $\lambda_x > 0$ and $(1 + \alpha c) > 0$. Depending on the sign of $(1+2\alpha c)$, the λ-intercept, -2α, of the switching line $y+\lambda = -2\alpha$ is greater or less than $1/c$. (a) $(1+2\alpha c) \leq 0$. (b) $(1+2\alpha c) > 0$.

Denoting by $Z_l(x, y, \lambda)$ the vector field in (7)–(9) with the linear control $u = (y+\lambda)/(-2\alpha)$ on $\mathbb{R} \times R_l$, i.e., that in (48) and (49), we obtain

$$S_i \circ Z_l = -Z_l \circ S_i, \qquad i = 1, 2. \tag{56}$$

The linear vector field Z_l is invariant under the reflections, S_1 and S_2, up to the time-reversal, and it is invariant under the reflection S_3 without time-reversal. The region $\mathbb{R} \times R_l$ is invariant under $Z_i, i = 1, 2$. Notice this symmetry in the phase portraits in region R_l of Fig. 4. This symmetry gives us useful information as follows. Consider a trajectory of Z_l whose (y, λ)-projected image is contained in R_l as in Fig. 5. The duration Δt_{AB} from A to B along the trajectory with $u = (y + \lambda)/(-2\alpha)$ in the $y - \lambda$ plane is the same as $\Delta t_{S_i(B)S_i(A)}, i = 1, 2$. Also, the corresponding (positive or negative) increments in x satisfy

$$\Delta x_{AB} = -\Delta x_{S_i(B)S_i(A)}, \qquad i = 1, 2.$$

This implies that there cannot be any optimal trajectories in the white region surrounded by the shaded region in Fig. 4 because an arbitrary trajectory starting from the λ-axis and ending at the y-axis can be decomposed into parts, each of which is invariant under S_1 or S_2, so $\Delta x = 0$ along the whole trajectory. However, the time-optimal control for $\Delta x = 0$ is $u = 0$ with

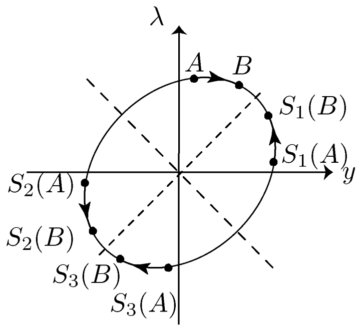


Fig. 5. $\mathbb{Z}_2 \times \mathbb{Z}_2$-symmetry in the linear region.

$t_f = 0$; see Remark 4.5. *Therefore, all time-optimal trajectories are contained in the shaded regions in* Fig. 4 *because only the trajectories in the shaded region can satisfy* $y(0) = 0$ *and* $\lambda(t_f) = 0$.

Let us consider the symmetry S_3 in (55). Let Z be the vector field in (7)–(9) on the whole domain, $\mathbb{R}^3 = \mathbb{R} \times (R_n \cup R_p \cup R_l)$. One can check

$$
\begin{aligned}
S_3(\mathbb{R} \times R_n) &= \mathbb{R} \times R_p \\
S_3(\mathbb{R} \times R_l) &= \mathbb{R} \times R_l \\
S_3 \circ Z &= Z \circ S_3 \\
\Delta t_{AB} &= \Delta t_{S_3(A) S_3(B)}.
\end{aligned}
$$

Hence, for example

$$\Delta x_{AB} = \Delta x_{S_3(A) S_3(B)}$$

in Fig. 5.

D. Definition, Monotonicity, and Positivity of X_1

Let us first define the basic region in the $y - \lambda$ plane. In each phase portrait in Fig. 4, we denote by $\mathcal{A}$ the interior of the shaded region. The *basic region* $\mathcal{B}$ is defined by

$$\mathcal{B} = \{(y, \lambda) \in \mathcal{A} \subset \mathbb{R}^2 \,|\, y \geq 0, \lambda \geq 0\}.$$

See Fig. 6 for an illustration of the basic region.

Given an extremal $\gamma(t) = (x(t), y(t), \lambda(t))$, $0 \leq t \leq T$, the arc $\gamma([t_1, t_2])$ with $0 \leq t_1 < t_2 \leq T$ is called a *basic arc* of γ if $y(t_1) = 0$, $\lambda(t_2) = 0$ and

$$\tilde{\pi}(\gamma([t_1, t_2])) \subset \mathcal{B} \cup \tilde{S}_3(\mathcal{B})$$

where $\tilde{\pi}(x, y, \lambda) = (y, \lambda)$ and $\tilde{S}_3(y, \lambda) = (-y, -\lambda)$.

We denote by Λ the open interval which is the intersection of the positive λ-axis with the basic region; see Fig. 6. It is given in (13) and (14), whose derivation will be made later. Let us construct a function X_1 of $\lambda_0 \in \Lambda$, which measures the (signed) increment of x along a basic arc starting with $\lambda(0) = \lambda_0$. The construction of X_1 and Λ is made in the following two separate cases:

$$(1 + 2\alpha c) \leq 0 \quad (1 + 2\alpha c) > 0.$$

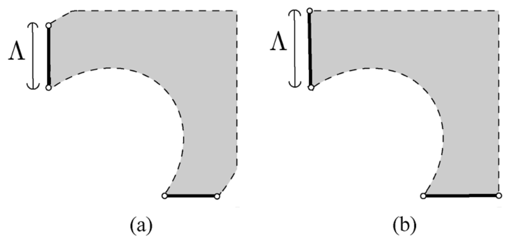


Fig. 6. Basic region and the domain Λ of the map X_1. (a) $(1 + 2\alpha c) \leq 0$. (b) $(1 + 2\alpha c) > 0$.

Case $(1 + 2\alpha c) \leq 0$*:* The domain Λ of X_1 is computed as follows:

$$
\begin{aligned}
\Lambda = \{\lambda \in \mathbb{R}_+ \,|\, & M(-\alpha, -\alpha) < M(0, \lambda) \\
& < M(-(1 + 2\alpha c)/c, 1/c)\} \\
= (& -2\alpha\sqrt{1 + \alpha c}, -2\alpha\sqrt{(1 + \alpha c)/(-\alpha c)}) \qquad (57)
\end{aligned}
$$

which is exactly (13), where the Hamiltonian M is given in (47). The switching line $\{y + \lambda = -2\alpha\}$ is tangent to the level set $\{M(y, \lambda) = M(-\alpha, -\alpha)\}$ at $(-\alpha, -\alpha)$. The level set $\{M(y, \lambda) = M(-(1 + 2\alpha c)/c, 1/c)\}$ goes through the intersection of the switching line $\{y + \lambda = -2\alpha\}$ and a stable manifold $\{\lambda = 1/c\}$ of $(1/c, 1/c)$; see Fig. 4(a). Notice in Fig. 4(a) that all the basic arcs are of form $\Gamma_{\pm}^{\lambda_0}$ defined in (17). Namely, basic arcs in the basic region are like arc $ABCD$ in Fig. 7(a). Let $A = (0, \lambda_0)$ in Fig. 7(a). Then, $D = (\lambda_0, 0)$ by the S_1 symmetry. The two points B and C in Fig. 7(a) are the intersection of the level set of M

$$D_1 = \{(y, \lambda) \mid M(y, \lambda) = M(0, \lambda_0)\} \qquad (58)$$

and the switching line

$$D_2 = \{(y, \lambda) \,|\, y + \lambda = -2\alpha\}. \qquad (59)$$

These points are symmetric to each other with respect to S_1 in (53) because D_1 and D_2 are invariant under S_1. The y-coordinates of B and C are given by

$$
\begin{aligned}
y_B(\lambda_0) &= -\alpha - \sqrt{\alpha^2 + (\lambda_0^2 - 4\alpha^2)/(-4\alpha c)} \qquad (60) \\
y_C(\lambda_0) &= -y_B(\lambda_0) - 2\alpha. \qquad (61)
\end{aligned}
$$

Due to the S_1 symmetry, the increments in x along AB and CD cancel each other, i.e.,

$$\Delta x_{AB} + \Delta x_{CD} = 0. \qquad (62)$$

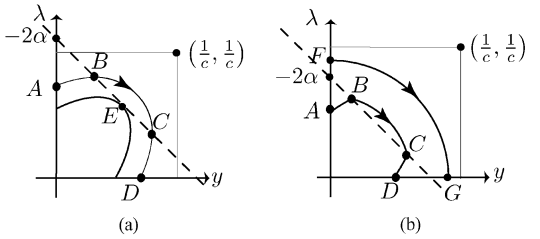


Fig. 7. Construction of the x-increment map X_1. (a) $(1+2\alpha c) \leq 0$. (b) $(1+2\alpha c) > 0$.

Hence

$$
\begin{aligned}
X_1(\lambda_0) = \Delta x_{BC} &= \int_{y_B(\lambda_0)}^{y_C(\lambda_0)} \frac{\dot{x}}{\dot{y}}\, dy \\
&= \int_{y_B}^{y_C} \frac{y+\alpha}{-cy+1}\, dy \\
&= \frac{1}{c}\Bigg[-2\alpha - 2y_C(\lambda_0) \\
&\qquad + \left(\alpha+\frac{1}{c}\right)\ln\left(\frac{1+2\alpha c + cy_C(\lambda_0)}{1-cy_C(\lambda_0)}\right)\Bigg]
\end{aligned}
\tag{63}
$$

with $y_C(\lambda_0)$ in (61). Regarding X_1 as a function of y_C, one can see

$$
\frac{dX_1}{dy_C} = \frac{2(y_C+\alpha)^2}{c\left(\frac{1}{c}+2\alpha+y_C\right)\left(\frac{1}{c}-y_C\right)} \geq 0 \tag{64}
$$

on $y_C(\Lambda) = (-\alpha, 1/c)$ because $(1+\alpha c) > 0$. Since y_C in (61) is a strictly increasing function of λ_0, it follows that X_1 is a strictly increasing function on Λ. Note that

$$
\begin{aligned}
\lim_{\lambda_0 \to -2\alpha\sqrt{1+\alpha c}} X_1(\lambda_0) &= \lim_{y_C \to -\alpha} X_1(y_C) = 0 \\
\lim_{\lambda_0 \to -2\alpha\sqrt{\frac{1+\alpha c}{-\alpha c}}} X_1(\lambda_0) &= \lim_{y_C \to 1/c} X_1(y_C) = +\infty.
\end{aligned}
\tag{65}
$$

It follows that $X_1(\Lambda) = (0,\infty)$.

Case $(1+2\alpha c) > 0$*:* The domain Λ of X_1 is given by

$$
\Lambda = \Lambda_1 \cup \Lambda_2 = (-2\alpha\sqrt{1+\alpha c}, -2\alpha) \cup [-2\alpha, 1/c) \tag{66}
$$

where the left end $-2\alpha\sqrt{1+\alpha c}$ of Λ is computed from $M(0,\lambda) = M(-\alpha,-\alpha)$ and the right end $1/c$ is just the intersection of the stable manifold of the fixed point $(1/c, 1/c)$, with the λ-axis in Fig. 4(b). As $-2\alpha\sqrt{1+\alpha c} < -2\alpha < 1/c$, the decomposition of Λ into Λ_1 and Λ_2 is valid. Notice in Fig. 4(b) that there are two different kinds of basic arcs in the basic region. If $\lambda_0 \in \Lambda_1$, then the trajectory is like $ABCD$ in Fig. 7(b), i.e., $\Gamma_+^{\lambda_0}$ in (17). In this case, the control is given by $u = (y+\lambda)/(-2\alpha)$ on AB and CD and $u = +1$ on BC. If $\lambda_0 \in \Lambda_2$, then the trajectory is like FG where the control is given by $u = +1$.

First, we consider the case where $\lambda_0 \in \Lambda_1$. Let y_B and y_C be the y-coordinates of the intersections points of D_1 and D_2 in (58) and (59). The formulas of y_B and y_C are given in (60) and (61). Due to the S_1 symmetry, relation (62) holds. Hence

$$
\begin{aligned}
X_1(\lambda_0) = \frac{1}{c}\Bigg[&-2\alpha - 2y_C(\lambda_0) \\
&+ \left(\alpha+\frac{1}{c}\right)\ln\left(\frac{1+2\alpha c + cy_C(\lambda_0)}{1-cy_C(\lambda_0)}\right)\Bigg]
\end{aligned}
$$

which coincides with (63). One can check that (64) is still valid on $y_C(\Lambda_1)$, and that y_C is a strictly increasing function of λ_0 on Λ_1. Hence, X_1 is a strictly increasing function of λ_0 on Λ_1.

We now consider the case that $\lambda_0 \in \Lambda_2$. We have

$$
\begin{aligned}
X_1(\lambda_0) &= \int_{y_F}^{y_G} \frac{\dot{x}}{\dot{y}}\, dy = \int_0^{\lambda_0} \frac{y+\alpha}{-cy+1}\, dy \\
&= \frac{1}{c}\left[-\lambda_0 + \left(\alpha+\frac{1}{c}\right)\ln\left(\frac{1}{1-c\lambda_0}\right)\right].
\end{aligned}
\tag{67}
$$

Notice the continuity of X_1 at $\lambda_0 = -2\alpha$ from (63) and (67). Since

$$
\frac{dX_1}{d\lambda_0} = \frac{\lambda_0+\alpha}{1-c\lambda_0} > 0
$$

for $\lambda_0 \in \Lambda_2$, the function $X_1(\lambda_0)$ is strictly increasing on Λ_2. Notice that

$$
\lim_{\lambda_0 \to 1/c} X_1(\lambda_0) = +\infty. \tag{68}
$$

From (65) and (68), it follows that $X_1(\Lambda) = (0,\infty)$. We conclude that the map $X_1(\lambda_0)$ is strictly increasing on Λ and its image is $(0,\infty)$.

We have proved the following.

Claim 6.3: Irrespective of the sign of $(1+2\alpha c)$, the map $X_1(\lambda_0)$ on Λ, which is the displacement in x along the basic arc with $\lambda(0) = \lambda_0$ in the basic region, is a strictly increasing function and its range is $(0,\infty)$.

E. Duration of Basic Arcs and Idling Arcs

We compute the duration of basic arcs. Given an extremal $(x(t), y(t), \lambda(t))$ with $(y(0), \lambda(0)) = (0, \lambda_0) \in 0 \times \Lambda$, let T_1 be the smallest $t > 0$ such that $(y(t), \lambda(t)) = (\lambda_0, 0)$ where Λ is defined in (13) and (14). We can regard T_1 as a function of $\lambda_0 \in \Lambda$. Recall the three logic sentences, P_1, P_2 and Q defined in Section II. We will compute T_1 separately for the following two cases:

$$
[\mathrm{P1} \vee \mathrm{Q}] \quad \text{and} \quad \mathrm{P2}.
$$

We begin with the case of $[\mathrm{P1} \ \vee \ \mathrm{Q}]$. The (y,λ)-projection of a basic arc is like the arc $ABCD$ in Fig. 7(a) and (b). It consists of the three arcs, AB, BC, and CD where $u = (y+\lambda)/(-2\alpha)$ on AB and CD and $u = +1$ on CD. Let us compute the flight time Δt_{AB} from A to B. By the S_1 symmetry, the flight time Δt_{CD} is the same as Δt_{AB}. The dynamics are given in (48)–(50) with the initial condition $(y(0), \lambda(0)) = (0, \lambda_0)$. The solution is given by

$$\begin{cases} y(t) = \dfrac{\lambda_0}{-2\alpha\omega}\sin(\omega t) \\ \lambda(t) = \lambda_0\cos(\omega t) + \frac{\lambda_0(1+2\alpha c)}{2\alpha\omega}\sin(\omega t). \end{cases} \tag{69}$$

where

$$\omega = \sqrt{\det A} = \sqrt{c(1+\alpha c)/(-\alpha)} > 0. \tag{70}$$

Thus

$$\Delta t_{AB}(\lambda_0) = \Delta t_{CD}(\lambda_0) = \frac{1}{\omega}\sin^{-1}\left(\frac{-2\alpha\omega y_B(\lambda_0)}{\lambda_0}\right) \tag{71}$$

where $y_B(\lambda_0)$ is given in (60). We now compute the flight time Δt_{BC} for BC. For this purpose, the corresponding y-dynamics can be written as

$$\dot{y} = -cy + 1 \quad y(0) = y_B \quad y(\Delta t_{BC}) = y_C$$

where y_B and y_C are given in (60) and (61). Direct integration yields

$$\Delta t_{BC}(\lambda_0) = \frac{1}{c}\ln\left(\frac{1 - cy_B(\lambda_0)}{1 - cy_C(\lambda_0)}\right). \tag{72}$$

By (71) and (72), the total flight time T_1 of $ABCD$ is given by

$$T_1(\lambda_0) = \frac{2}{\omega}\sin^{-1}\left(\frac{-2\alpha\omega y_B(\lambda_0)}{\lambda_0}\right) + \frac{1}{c}\ln\left(\frac{1 - cy_B(\lambda_0)}{1 - cy_C(\lambda_0)}\right). \tag{73}$$

We now consider the case of P2. The y-dynamics is

$$\dot{y} = -cy + 1 \quad y(0) = 0 \quad y(T_1) = \lambda_0.$$

Direct integration yields

$$T_1(\lambda_0) = \frac{1}{c}\ln\left(\frac{1}{1 - c\lambda_0}\right). \tag{74}$$

We have so far verified the formula of T_1 in (16).

We make a remark on the relation between X_1 and T_1. Recall that $X_1(\lambda_0)$ is a bijection from Λ to $(0, \infty)$. The corresponding flight time $T_1(\lambda_0)$ can be regarded as function of $X_1 \in (0, \infty)$ as follows:

$$T_1(X_1) = T_1 \circ \lambda_0(X_1). \tag{75}$$

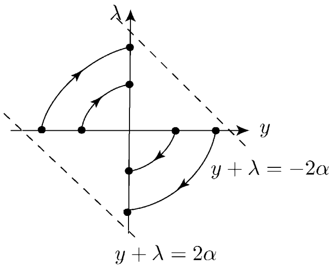


Fig. 8. Idling arcs do not contribute to the net displacement of a particle.

Recall the definition of *idling arcs* in Section II. An idling arc is an arc whose (y, λ)-projection starts from the positive (respectively, negative) y-axis, goes through the fourth quadrant (respectively, the second quadrant) in the region of R_l, and ends at the negative (respectively, positive) λ-axis; see Fig. 8. By the S_2 symmetry, they do not contribute to the displacement of x. Idling arcs occur in the case of $[\mathrm{P1} \quad \vee \quad \mathrm{Q}]$. The flight time, T_{idling}, which we call *idling time*, of idling arcs is given by

$$T_{\text{idling}} = \frac{1}{\omega}\sin^{-1}(-2\alpha\omega) \tag{76}$$

with ω in (70). Notice that the idling time is independent of the coordinates of the initial point on the y-axis.

F. Construction and Uniqueness of Optimal Trajectories

We investigate the issue of the construction and uniqueness of the time-optimal trajectory. Recall that the (y, λ) projection of optimal trajectories must start from the λ-axis and end at the y-axis as $y(0) = 0$ and $\lambda(t_f) = 0$. Because of the S_3 symmetry, if a trajectory with $\lambda(0) \in \Lambda$ is optimal, then its S_3 image is also optimal. Hence, without loss of generality, we will always give proofs only for optimal trajectories starting with $\lambda(0) \in \Lambda$.

We call an extremal an *n-shot extremal* where $n \in \mathbb{N}$ if the maximum number of basic arcs in the extremal is n. Equivalently, we say that an extremal is an n-shot extremal if its (y, λ) projection meets with the y-axis n times. For example, the extremal whose (y, λ) projection is in Fig. 9(a), is a one-shot extremal and that in Fig. 9(b) is a two-shot extremal. The one in Fig. 9(c) corresponds to a multishot extremal. Notice that the (y, λ) projection of an n-shot extremal with $n \geq 3$ is a closed curve by symmetry. By the discussion in Sections VI-C and VI-D, we know that for a given $x_f > x_0$ there always exists a unique n-shot extremal for each $n \in \mathbb{N}$ with $\lambda(0) = \lambda_0 \in \Lambda$ reaching x_f. Hence, we need to know how to find time-optimal ones among them. We divide our discussion into the following two cases.

$$(3 + 4\alpha c) \geq 0 \quad (3 + 4\alpha c) < 0.$$

Case $(3 + 4\alpha c) \geq 0$*:* We will show that the two one-shot extremals are *the* time-optimal ones for a given $x_f > x_0$. Recall

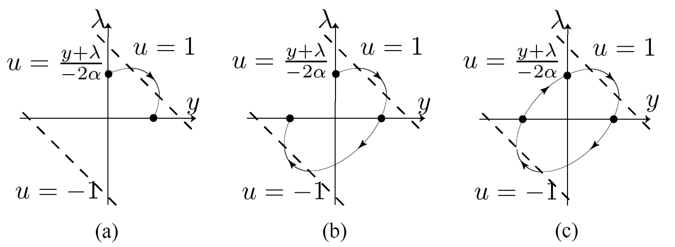


Fig. 9. There are several types of extremals. Their (y, λ)-projection must start from the λ-axis and end on the y-axis because of the boundary condition. (a) One-shot extremal, (b) two-shot extremal, and (c) multishot extremal.

the three logic sentences, P1, P2, and R defined in Section II. It is easy to verify

$$\text{P1} \vee \text{P2} \Rightarrow \text{R}.$$

We will divide our discussion into the following three cases:

$$\text{P2, P1}, [\text{R} \wedge \neg(\text{P1} \vee \text{P2})].$$

We first consider the case of P2. Suppose that the two-shot extremal, $FGHIJK$, in Fig. 10(a) is time-optimal with a final time $t_f > 0$ and a control u for $x_f > x_0$. Let us consider another control $v(t) = 1$ for $t \in [0, t_f]$. Let $FGHJ'K'$ be the trajectory due to v. By the S_2 symmetry, $\Delta x_{HI} = 0$. Hence, the increment in x due to the control u is given by

$$\Delta x_{FGHIJK}(u) = \Delta x_{FGH}(u = 1) + \Delta x_{S_3(I)FG}(u = 1)$$

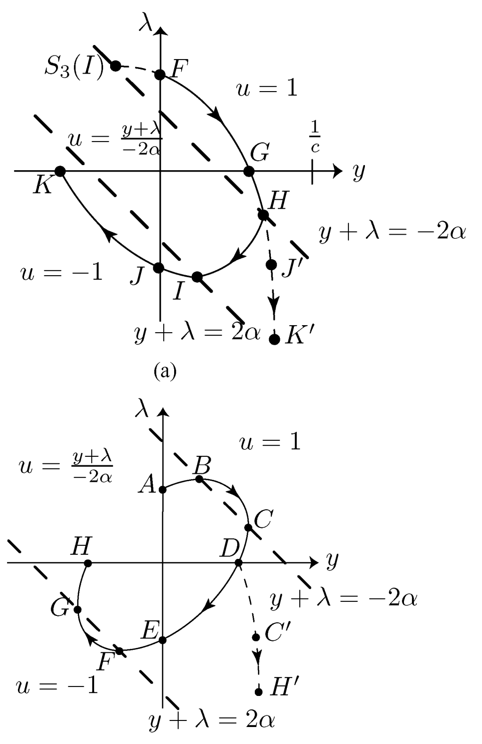
by the S_3 symmetry where $S_3(IJK) = S_3(I)FG$. The increment in x due to the control v is given by

$$\Delta x_{FGHJ'K'}(v) = \Delta x_{FGH}(v = 1) + \Delta x_{HJ'K'}(v = 1).$$

Along $S_3(I)FGHJ'K', \dot{y} = -cy + 1 > 0$. Let J' be the point such that the flight time on $S_3(I)FG$ is the same as that on HJ'. Namely, $\Delta t_{S_3(I)FG} = \Delta t_{HJ'}$. Then the flight time on HI is the same as that on $J'K'$. Notice that

$$\begin{aligned} &\Delta x_{S_3(I)FG}(u = 1) \\ &\quad = \int_0^{\Delta t_{S_3(I)FG}} (y_1(t) + \alpha)\, dt \\ &\quad < \int_0^{\Delta t_{HJ'}} (y_2(t) + \alpha)\, dt \\ &\quad = \Delta x_{HJ'}(v = 1) \end{aligned}$$

where $y_1(t) = y_{S_3(I)} e^{-ct} + (1)/(c)(1 - e^{-ct}), y_2(t) = y_H e^{-ct} + (1)/(c)(1 - e^{-ct})$, and $y_{S_3(I)} < y_H$. On $J'K', \dot{x} = y + \alpha > 0$ because $y_H \geq y_G = \lambda_0 \geq -2\alpha$ and $\dot{y} > 0$ on $J'K'$. Thus, $\Delta x_{J'K'}(v = 1) > 0$. Therefore, $\Delta x_{FGHJ'K'}(v) > \Delta x_{FGHIJK}(u)$. This implies, by the continuity of $t \mapsto x(t) \in \mathbb{R}$, that the trajectory due to the control v reaches x_f before $t = t_f$, which contradicts the time-optimality of the control u. Hence, no two-shot extremals can be time-optimal in the case of P2. Similarly, one can show that no multi-shot extremals can be time-optimal in the case of P2.



(a)

(b)

Fig. 10. One-shot extremal is the optimal one among extremals for a given $x_f > x_0$ when $(1 + 2\alpha c) > 0$. (a) $\lambda_0 \in \Lambda_2 = [-2\alpha, 1/c)$. (b) $\lambda_0 \in \Lambda_1 = (-2\alpha\sqrt{1 + \alpha c}, -2\alpha)$.

We now consider the case of P1. Suppose that a two-shot extremal with $\lambda_0 \in \Lambda_1$ is time-optimal with $t_f > 0$ and control u for $x_f > x_0$. Let $ABCDEFGH$ be the trajectory associated with u in Fig. 10(b). Let t_1 is the time when the trajectory reaches D. Let us construct another control v as follows:

$$v(t) = \begin{cases} u(t), & \text{if } t \in [0, t_1] \\ 1, & \text{if } t \in [t_1, t_f]. \end{cases}$$

Let $ABCDC'H'$ be the trajectory associated with the control v where $DC'H'$ corresponds to the time interval $[t_1, t_f]$, and we choose C' such that $\Delta t_{BC} = \Delta t_{DC'}$. By the S_1 and S_2 symmetry

$$\Delta x_{AB} + \Delta x_{CD} = \Delta x_{EF} + \Delta x_{GH} = 0.$$

The total increment in x due to the control u is given by

$$\Delta x(u) = 2\Delta x_{BC}. \tag{77}$$

The total increment in x due to the control v is

$$\Delta x(v) = \Delta x_{BC} + \Delta x_{DC'} + \Delta x_{C'H'}. \tag{78}$$

By the S_1 symmetry, $y_D = \lambda_0$, so

$$y_D = \lambda_0 > -2\alpha\sqrt{1+\alpha c} \geq -\alpha \geq y_B \tag{79}$$

where $\lambda_0 \in \Lambda$, $(3+4\alpha c) \geq 0$ and (60) were used for the three inequalities. As $\dot{y} = -cy + 1 > 0$ on BC and DC', we have

$$\begin{aligned}\Delta x_{BC} &= \int_0^{\Delta t_{BC}} (y_1(t)+\alpha)\,dt \\ &< \int_0^{\Delta t_{DC'}} (y_2(t)+\alpha)dt = \Delta x_{DC'}\end{aligned} \tag{80}$$

where $y_1(t) = y_B e^{-ct} + (1)/(c)(1-e^{-ct})$ and

$$y_2(t) = y_D e^{-ct} + \frac{1}{c}(1-e^{-ct}).$$

It is straightforward to see

$$\Delta x_{C'H'} > 0. \tag{81}$$

By (77), (78), (80), and (81),

$$\Delta x(v) > \Delta x(u).$$

This implies that the trajectory due to the control v reaches x_f before $t = t_f$, which contradicts the time-optimality of u. Therefore, there are no two-shot time-optimal trajectories in the case of P1. In the similar manner, one can show that no multishot extremals can be time-optimal in the case of P1.

We now consider the case of [R $\wedge\neg$(P1 $\vee$P2)]. In this case, all the basic arcs are of form $\Gamma_\pm^{\lambda_0}$ in (17). Notice that for the case of P1 we only used the fact that the basic arcs are of form $\Gamma_\pm^{\lambda_0}$ and $(3+4\alpha c) \geq 0$ where the latter was used in (79). Hence, the case of [R $\wedge\neg$(P1 $\vee$ P2)] can be handled in the same way as the case of P1, to conclude that no multishot extremals are time-optimal.

So far, we have showed that multishot extremals cannot be time-optimal when $(3+4\alpha c) \geq 0$. By the discussion in Section VI-D, we know that there exists a unique one-shot extremal with $\lambda_0 \in \Lambda$, or a basic arc, for a given $x_f > x_0$ such that $x(t_f) = x_f$. Hence, this basic arc and its image under the reflection S_3 are the only optimal trajectories.

Claim 6.4: If $(3+4\alpha c) \geq 0$, then there are exactly two time optimal trajectories for $x_f > x_0$. One is the one-shot extremal $\Gamma_+^{\lambda_0}$ (or, the basic arc) with $\lambda(0) = \lambda_0 = X_1^{-1}(x_f - x_0)$ and the other is its image by the reflection S_3 (see also Section II). The corresponding minimum time is $T_1(\lambda_0)$. The maps X_1 and T_1 are defined in (15) and (16).

Case $(3+4\alpha c) < 0$*:* In this case, unlike the case of $(3+4\alpha c) \geq 0$, we have no general proof that only one-shot extremals are time-optimal. Instead we provide a finite algorithm of finding all time-optimal trajectories for $x_f > x_0$.

Take $x_f > x_0$. For each $n \in \mathbb{N}$, there exists a unique n-shot extremal with $\lambda(0) = \lambda_{0,n} := X_1^{-1}((x_f - x_0)/n)$ reaching x_f. Since $(3+4\alpha c) < 0$ implies $(1+2\alpha c) \leq 0$, this n-shot extremal consists of n one-shot extremals and $(n-1)$ idling arcs between the n one-shot extremals; see (19), Figs. 7(a), and 9(c). Each of the n one-shot extremals contributes $(x_f - x_0)/n$ to the increment in x, and each idling arc makes zero contribution. By this decomposition, the total time cost T_n for this n-shot extremal is

$$T_n = n \cdot T_1 \circ X_1^{-1}((x_f - x_0)/n) + (n-1)\cdot T_{\text{idling}} \tag{82}$$

where $T_1 \circ X_1^{-1}((x_f - x_0)/n)$ is the flight time corresponding to the increment $(x_f - x_0)/n$ in x where X_1, T_1, and T_{idling} are given in (63), (73), and (76). There exists $N \in \mathbb{N}$ such that

$$T_1 = T_1 \circ X_1^{-1}(x_f - x_0) < N \cdot T_{\text{idling}}$$

which implies

$$T_1 < T_n \qquad \forall n \geq (N+1).$$

Thus, all possible time-optimal trajectories are among the first N extremals. Choose $k \in \{1,\ldots,N\}$ such that $T_k = \min\{T_i : 1 \leq i \leq N\}$. Then, the k-shot extremal is a time-optimal trajectory. Such k's give all the time-optimal trajectories corresponding to $x_f > x_0$. This proves the procedure [B.1]–[B.4] in Section II, which can be summarized as follows.

Claim 6.5: If $(3+4\alpha c) < 0$, then there is a finite and explicit procedure ([B.1]–[B.4] in Section II) of finding all time-optimal trajectories and the corresponding minimum time for $x_f > x_0$.

As a remark, we give a practical way of showing that the one-shot extremals are the unique time-optimal trajectories. First, with (63) and (73), one draws the graph $(X_1, T_1(X_1))$ in (75). Suppose that it is strictly concave. Then

$$T_1 = T_1(X_1 = \Delta x) < n \cdot T_1(X_1 = \Delta x/n) \qquad \forall n = 2,3,\ldots$$

which, with (82), implies that $T_1 < T_n$ for all $n \geq 2$. Hence, the two one-shot extremals are the only time-optimal trajectories if $T_1(X_1)$ is a strictly concave function of X_1.

G. Initial Undershoots

One can check that for $\lambda(0) \in \Lambda$, every extremal has $\dot{x}(t) \leq 0$ from $t = 0$ until $y(t)$, which is initially zero, becomes $-\alpha > 0$. In other words, $x(t)$ goes through an initial undershoot. We will compute this undershoot and its duration. Here, we do not give detailed computation because the methodology is very similar to those in Sections VI-D and VI-E. Recall the three logic sentences, P1, P2, and Q defined in Section II.

Let us first consider the case of [P1 $\vee$ Q]. The undershoot consists of two parts in Fig. 7(a) and (b): AB and $B \to (-\alpha, \bar{\lambda})$ for some $\bar{\lambda}$ where $(-\alpha, \bar{\lambda})$ lies between B and C on the trajectory. One can compute the duration $T_{\text{under}}(\lambda_0)$ of the undershoot as

$$T_{\text{under}}(\lambda_0) = \Delta t_{AB}(\lambda_0) + \frac{1}{c}\ln\left(\frac{1 - cy_B(\lambda_0)}{1+c\alpha}\right)$$

and the (negative-valued) amount of the undershoot

$$X_{\text{under}}(\lambda_0) = x(T_{\text{under}}(\lambda_0)) - x(0)$$

as

$$\begin{aligned}
&X_{\text{under}}(\lambda_0)\\
&\quad= \frac{\lambda_0^2}{8\omega^3\alpha^2}\left\{(\omega^3\alpha + \omega c + \omega\alpha c^2)t\right.\\
&\qquad + (1+2\alpha c)\omega\sin^2(\omega t)\\
&\qquad \left. + (\omega^2\alpha - c - \alpha c^2)\cos(\omega t)\sin(\omega t)\right\}|_{t=\Delta t_{AB}(\lambda_0)}\\
&\qquad + \frac{1}{c}\left\{y_B(\lambda_0) + \alpha + \left(\alpha + \frac{1}{c}\right)\right.\\
&\qquad \left. \times \ln\left(\frac{1 - cy_B(\lambda_0)}{1+c\alpha}\right)\right\}
\end{aligned}$$

where ω, Δt_{AB}, and $y_B(\lambda_0)$ are defined in (70), (71), and (60).

In the case of P2, the undershoot corresponds to the arc, $F \rightarrow (-\alpha, \bar{\lambda})$ for some $\bar{\lambda}$ in Fig. 7(b) where $(-\alpha, \bar{\lambda})$ lies between F and G. One can compute the duration $T_{\text{under}}(\lambda_0)$ of the undershoot as

$$T = \frac{1}{c}\ln\left(\frac{1}{1+\alpha c}\right)$$

and the amount $X_{\text{under}}(\lambda_0)$ of the undershoot as

$$X_{\text{under}}(\lambda_0) = \frac{1}{c}\left(\alpha + \left(\alpha + \frac{1}{c}\right)\ln\left(\frac{1}{1+\alpha c}\right)\right).$$

Notice that this undershoot is independent of λ_0.

H. Discussion on State Constraints on x

We have seen in Section III that physics imposes a constraint on x as either

$$x \geq 0 \quad \text{or} \quad x \leq 0. \tag{83}$$

We have not considered these constraints in computing optimal trajectories in this paper. We now make a remark that in some cases optimal trajectories derived without the state constraints are also optimal with the state constraints.

First, let us consider the result in Section IV again—this time with the state constraints in (83). In either case of (83), it is not hard to verify that the result in Section IV still holds without violating the constraint. Namely, there are no time optimal trajectories for $x_f < x_0$ even in the existence of either state constraint in (83).

Second, we consider the constraint $x \leq 0$ with the initial and final condition satisfying $x_f \geq x_0$. It is trivial to see that an optimal trajectory derived without the state constraint reaches x_f without violating the constraint, so it is still optimal in the existence of the state constraint.

Lastly, we consider the constraint $x \geq 0$ with the initial and final condition satisfying $x_f \geq x_0$. We divide the discussion into the two cases; $(3+4\alpha c) \geq 0$ and $(3+4\alpha c) < 0$. When $(3+4\alpha c) \geq 0$, we know from Section VI-F that there exists a unique $\lambda_0 \in \Lambda$ such that the two basic arcs $\Gamma_{\pm}^{\lambda_0}$ are the only optimal trajectories. Recall the discussion on initial undershoots in Section VI.G. Also, recall from Section VI-D that X_1 is a positive function. We have $\dot{x}(t) > 0$ after $x(t)$ passes through $x_0 + X_{\text{under}}(\lambda_0)$ along the basic arc. Hence, we have $\min_{t\in[0,T_1(\lambda_0)]} x(t) = x_0 + X_{\text{under}}(\lambda_0)$ where $T_1(\lambda_0)$ is the duration of the basic arc. It follows that in the case of $(3+4\alpha c) \geq 0$ if $x_0 \geq |X_{\text{under}}(\lambda_0)|$ then the optimal trajectories derived without the state constraint $x \geq 0$ are optimal with the state constraint. On the other hand, as seen in Section VI-F, when $(3+4\alpha c) < 0$ there can be more than two optimal trajectories for a given $x_f \geq x_0$. Some of the optimal trajectories can be multishot extremals. On an idling arc, one has $\dot{x} = (\lambda^2 - y^2)/(-4\alpha)$. Hence, during the first half of an idling arc $\dot{x}(t) \geq 0$ and during the second half $\dot{x}(t) \leq 0$. Recall that an idling arc does not contribute any net displacement in x. Let t_1 and $t_2 = t_1 + T_{\text{idling}}$ be the initial and final time of an idling arc of an optimal trajectory with $\lambda_0 \in \Lambda$ derived without the state constraint $x \geq 0$. Then, it follows that $\min_{t\in[t_1,t_2]} x(t) = x(t_1) = x(t_2) \geq x_0$. Considering the decomposition of multishot extremals in (19), one can see that in the case of $(3+4\alpha c) < 0$ if $x_0 \geq |X_{\text{under}}(\lambda_0)|$ then the optimal trajectories derived without the state constraint $x \geq 0$ remain optimal with the state constraint. Therefore, irrespective of the sign of $(3+4\alpha c)$, an optimal trajectory with $\lambda_0 \in \Lambda$ derived without the state constraint $x \geq 0$ is again optimal with the state constraint if $x_f > x_0 \geq |X_{\text{under}}(\lambda_0)|$.

VII. Symplectic Reduction Picture

We cast the analysis used in Sections V and VI in the framework of symplectic reduction theory in order to provide geometric insight into the problem. We refer readers to [1] and [9] for symplectic reduction theory.

Let $G = \mathbb{R}$ be the Lie group acting on $M = \mathbb{R}^2 = \{(x,y)\}$ by translation in x and on the cotangent bundle $T^*M = \mathbb{R}^4 = \{(x, y, \lambda_x, \lambda_y)\}$ by cotangent lift where T^*M is equipped with the canonical symplectic form, $\Omega = \mathrm{d}x \wedge \mathrm{d}\lambda_x + \mathrm{d}y \wedge \mathrm{d}\lambda_y$. The momentum map $J : T^*M \rightarrow \mathbb{R}$ corresponding to the G-action is $J(x, y, \lambda_x, \lambda_y) = \lambda_x$. We have $J^{-1}(\nu) \simeq \mathbb{R}^3$ for any $\nu \in \mathbb{R}$, and $J^{-1}(\nu)/G \simeq \mathbb{R}^2$. One can construct the symplectic projection $\pi : J^{-1}(\nu) \subset T^*M \rightarrow J^{-1}(\nu)/G \simeq \mathbb{R}^2$ by $\pi(x, y, \nu, \lambda_y) = (y, \lambda_y)$ where $J^{-1}(\nu)/G$ is given the canonical symplectic form $\omega = \mathrm{d}y \wedge \mathrm{d}\lambda_y$. Notice that the Hamiltonian H in (32) is G-invariant, that the control u in (38) is G-invariant, and that the Hamiltonian M° in (33) is also G-invariant. Hence, H and M° induce reduced Hamiltonians h and m° on $J^{-1}(\nu)/G$ as follows:

$$\begin{aligned}
h(y, \lambda_y, u) &= H(x, y, \nu, \lambda_y, u)\\
&= \lambda_x \alpha u^2 + (\lambda_x y + \lambda_y)u - cy\lambda_y \qquad (84)\\
m^\circ(y, \lambda_y) &= M^\circ(x, y, \nu, \lambda_y) = \max_{|u|\leq 1} h(y, \lambda_y, u). \qquad (85)
\end{aligned}$$

Notice that the control u maximizing $H(x, y, \nu, \lambda_y, u)$ is the same as the control maximizing $h(y, \lambda_y, u)$ because the control u maximizing $H(x, y, \nu, \lambda_y, u)$ is G-invariant and the group G is abelian. Equations (35) and (37) are the (reduced) dynamics of the reduced Hamiltonian m° (or h) on the based space $(J^{-1}(\nu)/G, \omega)$, and (34) can be regarded as a reconstruction equation to compute the displacement along the fiber $G = \mathbb{R}$. The map X_1 constructed in Section VI.D measures this displacement while $(y(t), \lambda(t))$ with $(y(0), \lambda(0)) \in 0 \times \Lambda$ stays in

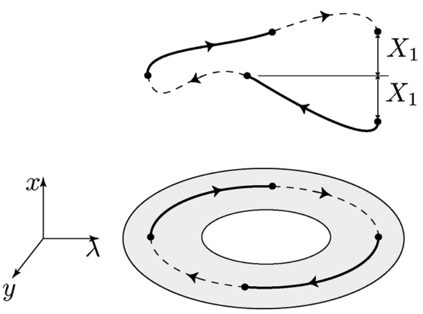


Fig. 11. Level set $J^{-1}(\nu)$ is diffeomorphic to $\mathbb{R}^3 = \{(x, y, \lambda)\}$ where the x-axis is the fiber, and $\pi : J^{-1}(\nu) \to \mathbb{R}^2$ is the projection to the y-λ plane. The map X_1 measures the displacement along the fiber. The dotted arcs are the idling arcs. The shaded region here corresponds to those in Fig. 4.

the first quadrant in the y-λ plane. In particular, it is interesting to notice that if $[(1+2\alpha c) \leq 0]$ or $[(1+2\alpha c) > 0$ and $\lambda(0) \in \Lambda_1]$, then X_1 measures the half of the total phase (or, holonomy with respect to the trivial connection on the principal bundle $\pi : J^{-1}(\nu) \to J^{-1}(\nu)/G$) corresponding to a closed trajectory $(y(t), \lambda(t))$ in the base space; see Fig. 11.

In Sections VI.A and VI.B, we did not use the reduced space $(\mathbb{R}^2, \omega)$ and the projection π constructed above when $J = \lambda_x = \nu \neq 0$, but we used the scaled variable $\lambda = \lambda_y/\lambda_x$ in (41) and (43) so as to show the discrete symmetry as reflection maps and to deal with all cases of $\lambda_x > 0$ simultaneously. To put this scaling in the reduction process, one can use $\nu\, \mathrm{d}y \wedge \mathrm{d}\lambda$ as a symplectic form on the reduced space $J^{-1}(\nu)/G = \{(y, \lambda)\}$, and define the projection $\pi_\nu : J^{-1}(\nu) \to J^{-1}(\nu)/G$ by

$$\pi_\nu : (x, y, \nu, \lambda_y) \mapsto (y, \lambda) = (y, \lambda_y/\nu).$$

In this case, the reduced Hamiltonians h and m° induced by H and M° are those in (84) and (85) with λ_y replaced by $\nu\lambda$. The vector field $(\dot{y}, \dot{\lambda})$ in Section VI-B is the Hamiltonian vector field of the reduced Hamiltonian m° (or h) on the reduced space $(\mathbb{R}^2, \nu\, \mathrm{d}y \wedge \mathrm{d}\lambda)$.

Remark 7.1: The $G = \mathbb{R}$ symmetry in the Hamiltonian dynamics was created by the nonlinear coordinate change in (23). This reduction process extends to the case where we consider many particles on the one-dimensional invariant line.

VIII. Simulations

We demonstrate the theoretical result by simulation. Suppose that the particle and the medium are given such that

$$\alpha = -3/4 \quad c = 1$$

and that the initial and final position of the particle is given by

$$x_0 = 1 \quad x_f = 2.$$

These numbers are chosen arbitrarily, but one can also do the same simulation with real data once they are given. Let $\Delta x = x_f - x_0$. One can check that the condition $(1+\alpha c) > 0$ for the existence of optimal trajectories holds. The idling time in (76) is given by $T_{\text{idling}} = 1.8138$. By (63) and (71)–(73), the flight time of the one-shot extremal T_1 is given by $T_1 = 7.8117$. By the algorithm in Claim 6.5, one learns that the minimum time is in $\{T_1, \ldots, T_5\}$. By (82) one computes these flight times in Table I.

TABLE I
Time Costs of the First Five Extremals

n	$\lambda_{0,n}$	$T_1 \circ X_1^{-1}(\Delta x/n)$	Time cost T_n
1	0.8649	7.8117	7.8117
2	0.8577	5.7950	13.4038
3	0.8491	5.1093	18.9556
4	0.8415	4.7597	24.4802
5	0.8350	4.5463	29.9867

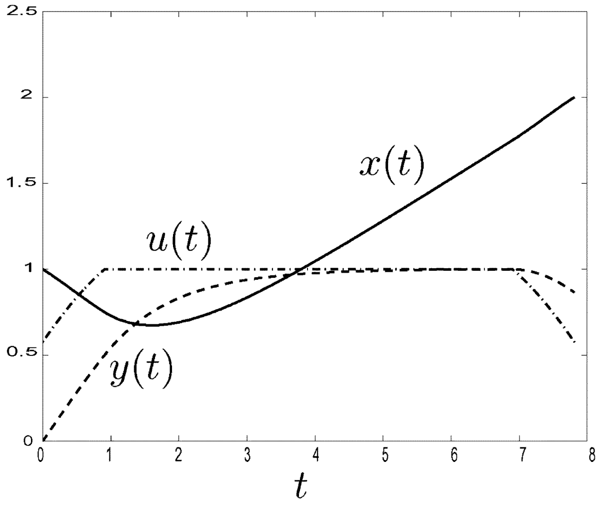


Fig. 12. State $(x(t), y(t))$ of the optimal trajectory and the corresponding optimal control $u(t)$ for $x_0 = 1$ and $x_f = 2$ with minimum time $t_f = 7.8117$.

Hence, the one-shot extremal is the optimal trajectory and the minimum time is $t_f = 7.8117$. This result agrees with the uniqueness result because the given α and c satisfy $(3 + 4\alpha c) \geq 0$. The optimal control $u(t)$ and the optimal trajectory $(x(t), y(t))$ are given in Fig. 12. Notice the initial undershoot predicted in Section VI-G.

IX. Conclusion and Future Work

The time-optimal control problem for the dielectrophoretic system studied in this paper has several interesting features. The existence of a term quadratic in control creates the non-existence of optimal trajectories when the final position of the particle is below the initial position. In contrast, when the final position is above the initial position we can show the existence and uniqueness of optimal trajectories in a range of the parameters α and c. In the other range of α and c we give a finite algorithm of finding all optimal trajectories instead. Both continuous and discrete symmetry in the problem simplifies the analysis.

In summary, the optimal trajectories are described as follows. There are three different types of optimal trajectories depending on the values of α and c, and the displacement of the particle. If $(1 + 2\alpha c) > 0$, basic arcs, or one-shot extremals, are the optimal trajectories. To move a particle a long distance, the optimal control is the saturated controls, i.e., $u = \pm 1$, but for a small displacement of the particle the optimal control consists of three parts; a linear control, a saturated control and a linear control again. If $(1 + 2\alpha c) \leq 0$ and $(3 + 4\alpha c) \geq 0$, then for any displacement of the particle the optimal control consists of three parts; a linear control, a saturated control and a linear control again. If $(3 + 4\alpha c) < 0$, optimal trajectories may have the structure of multi-shot extremals in (18) and (19).

As for future work, we will take into account a state constraint, and/or consider the time-optimal control of two different particles for the purpose of separating them which are initially close to each other. They have important applications in nano/bio-technology [5], [7], [8]. We believe that the use of control systems theory will refine and improve the manipulaton of particles in applications and that our work in this paper makes a first forward step in this direction.

ACKNOWLEDGMENT

The authors would like to thank B. Bonnard, J. Lévine, J. Marsden, L. Praly, B. Shashikanth, and the anonymous reviewers for invaluable comments.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. Reading, MA: Addison-Wesley, 1985, ch. 4, p. 62.

[2] B. Bonnard and M. Chyba, *Singular Trajectories and Their Role in Control Theory*. New York: Springer-Verlag, 2003.

[3] U. Boscain and B. Piccoli, *Optimal Syntheses for Control Systems on 2-D Manifolds*. New York: Springer-Verlag, 2004.

[4] D. E. Chang, S. Loire, and I. Mezic, "Closed-form solutions in the electrical field analysis for dielectrophoretic and travelling wave interdigitated electrode arrays," *J. Phys. D: Appl. Phys.*, vol. 36, no. 23, pp. 3073–3078, 2003.

[5] D. E. Chang and N. Petit, "Toward controlling dielectrophoresis," *Int. J. Robust Nonlinear Control*, vol. 15, no. 16, pp. 769–784, 2005.

[6] G. de Barra, *Measure Theory and Integration*. Upper Saddle River, NJ: Prentice-Hall, 1981, pp. 87–88.

[7] M. P. Hughes, *Nanoelecromechanics in Engineering and Biology*. Boca Raton, FL: CRC, 2002.

[8] T. B. Jones, *Electromechanics of Particles*. New York: Cambridge Univ. Press, 1995.

[9] J. E. Marsden, R. Montgomery, and T. S. Ratiu, "Reduction, symmetry and phases in mechanics," in *AMS Memoirs*. Providence, RI: AMS, 1990, vol. 436.

[10] B. Piccoli, "Time-optimal control problems for the swing and ski," *Int. J. Control*, vol. 62, no. 6, pp. 1409–1429, 1995.

[11] H. A. Pohl, *Dielectrophoresis*. Cambridge, U.K.: Cambridge Univ. Press, 1978.

[12] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*. New York: Wiley, 1962, pp. 20–50.

Dong Eui Chang received the B.S. degree in control and instrumentation engineering and the M.S. degree in electrical engineering, both from the Seoul National University, Korea, in 1994 and 1997, respectively, and the Ph.D. degree in control and dynamical systems from the California Institute of Technology, Pasadena, in 2002.

He was a Postdoctoral Fellow at the University of California, Santa Barbara, in 2003, at the Centre Automatique et Systèmes, Ecole Nationale Supérieure des Mines de Paris, France, in 2004, and at the University of Liège, Belgium. He joined the Department of Applied Mathematics at the University of Waterloo, Canada, in August, 2005, as an Assistant Professor. His research interests include geometric control theory and its application to robotics and nanotechnology.

Nicolas Petit was born in Paris, France, in 1972. He graduated from Ecole Polytechnique, Paris, France, in 1995, and received the Ph.D. degree in mathematics and control at Ecole Nationale Supérieure des Mines de Paris, France, in 2000.

In 2000–2001, he was a Postdoctoral Scholar in the Control and Dynamical Systems at the California Institute of Technology, Pasadena. Since 2001, he has held the position of Maitre-Assistant at Ecole des Mines de Paris in the Centre Automatique et Systèmes. His research interests include flatness theory for partial differential equations, numerical treatment of optimal trajectory generation problems for nonlinear systems, motion planning, observation of periodic systems, and analysis of distributed systems. On the application side, he is active in industrial process control, engine control, and embedded systems. He has developed the controllers of several industrial chemical reactors, including polystyrene and polypropylene reactors, and the ANAMELV4 and V5 softwares, currently used for closed-loop control of blending devices in numerous refineries. He is a coauthor of several patents in the field of engine control and process control.

Dr. Petit received the *Journal of Process Control* Paper Prize for Best Article 2002–2005 (Application). He has served as an Associate Editor for *Automatica*.

Pierre Rouchon was born in Saint-Etienne, France, in 1960. He graduated from Ecole Polytechnique, Paris, France, in 1983, and received the Ph.D. degree in chemical engineering from Ecole des Mines de Paris, France, in 1990. He received the "habilitation á diriger des recherches" in mathematics from the University Paris-Sud Orsay, France, in 2000.

From 1993 to 2005, he was an Associate Professor at École Polytechnique in Applied Mathematics. From 1998 to 2002, he was the Head of the Centre Automatique et Systèmes of École des Mines de Paris. He is currently a Professor at Ecole des Mines de Paris. His fields of interest include the theory and applications of dynamical systems, nonlinear control, and in particular differential flatness and its extension to infinite-dimensional systems. He has worked on many industrial applications, such as distillation columns, electrical drives, car equipments, and chemical reactors. One of his recent fields of interest is relative to the control and estimation of closed and open quantum systems.

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Process Control 15 (2005) 771–784

JOURNAL OF PROCESS CONTROL

www.elsevier.com/locate/jprocont

Boundary control for an industrial under-actuated tubular chemical reactor

D. Del Vecchio [a], N. Petit [b,*]

[a] Control and Dynamical Systems, California Institute of Technology, Mail Code 107-8l, 1200 E California Blvd. Pasadena, CA 91125, USA

[b] Centre Automatique et Systèmes, Ecole Nationale Supérieure des Mines de Paris, 60, boulevard Saint-Michel, 75272 Paris Cedex 06, France

Received 8 March 2004; received in revised form 28 February 2005; accepted 11 March 2005

Abstract

Several control strategies are presented and studied for an industrial under-actuated tubular chemical reactor. This work presents a case-study of the performance of a decentralized versus centralized control strategy. The tubular reactor under consideration is characterized by nonlinear kinetic laws, and it has some structural constraints on the location of the heat exchangers and of the sensors. For this system, a set of PI controllers is considered and a multivariable LQR controller is constructed to optimally choose the gains. The performance of these control strategies is studied. Finally, a direct numerical treatment of optimal control of the partial differential equations is presented. Industrial results are given for the linear controllers. Simulations emphasize the possible relevance of a direct numerical treatment of the nonlinear partial differential equations.

Keywords: Polystyrene; Tubular reactor; Control; Optimization; Industrial application

1. Introduction

The contribution of this paper is a study of different control strategies for a class of tubular reactors. The application underlying this study is a reactor in the ATOFINA PS (polystyrene) plant in Carling, France. We present a model of the tubular reactor used in this plant, which is characterized by nonlinear kinetic laws and by an under-actuated structure due to the choice of heat exchangers and sensor locations.

Around a steady production state (corresponding to an average of 120 kT/year), the grade of the polystyrene produced in this plant critically depends on the temperature profile along the reactor. In fact, the real control objective is temperature control. Other controlled variables such as molecular mass distribution are controlled using other inputs such as dilution. This is different from other reactors (see e.g. [1] for a survey of polymerization reactor control) where monomer conversion is usually considered as a critical value. In this plant, quality constraints (in connection to further injection and thermoforming applications) are tight and they directly translate to temperature constraints.

The grade of the produced polymer is scheduled with respect to economical considerations. This induces frequent changes in the setpoints that have to be precisely and quickly met to optimize profit and minimize off-spec products.

This paper proposes several control schemes to improve upon the results obtained with the existing PI controllers used in the plant. In particular, for a decentralized PI scheme, we show that the choice of the measurement and setpoints affects both transient and asymptotic performance. Then, a centralized PI scheme

[*] Corresponding author. Tel.: +33 1 4051 9330; fax: +33 1 4051 9165.
E-mail address: nicolas.petit@ensmp.fr (N. Petit).

0959-1524/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2005.03.004

is proposed, where the proportional gains are designed using an LQR design. This approach can be considered as a weighting of the input of a PI controller based on the model structure. Finally, a nonlinear centralized controller is proposed and its performance compared with the others. This work relies on the controller upgrade project that was carried out at the ATOFINA plant by a joint team of TOTAL engineers and researchers from École des Mines de Paris, which is reported in [2].

We propose three different control strategies ranging from fully decentralized to fully centralized. This work can be seen as an industrial scale case-study of the role of a decentralized versus centralized control strategy. This question was raised by several authors in various fields of the process industry [3–5], and we found it particularly relevant in this problem. From our point of view and from this particular study, we believe it needs to be answered with two facts on mind: performance requirements and availability of efficient numerical tools and accurate models.

This paper considers the problems as they appeared when trying to improve the existing PI controllers. In Section 2, we give the model of the plant. In Section 3, we underline the importance of the right choice of PI structure for performance improvement. In Section 4, we explain the design and tuning of an LQR controller that has been successful since when it was installed in 2000 (overall load was increased by more than 10%). Industrial results are given. Finally in Section 5, we propose an approach based on the direct treatment of the nonlinear partial differential equations that govern the system. This approach relies on the NTG optimal control software package designed for PDEs [6,7]. We compare the results of different control strategies and show that this last method, when a good knowledge of the process dynamics is available, is efficient and flexible.

2. Model of the reactor

The process under study is a polymerization tubular exothermic reactor with heat exchangers on the sides and with plug flow, see Fig. 1.

The styrene monomer enters the tubular reactor at a constant temperature at point I (see Fig. 1) along with the peroxide initiator. The monomer reacts inside the tubular reactor as it travels to point O. The tubular reactor is equipped with heat exchangers to evacuate the reaction exothermicity.

This tubular reactor is frequently subject to strong oscillations. These are usually interpreted as temperature perturbations propagating through the system (these perturbations can arise from various phenomena: in [8] it is shown that for such tubular reactors, jacket temperature perturbations may lead to oscillatory dynamic responses[1]).

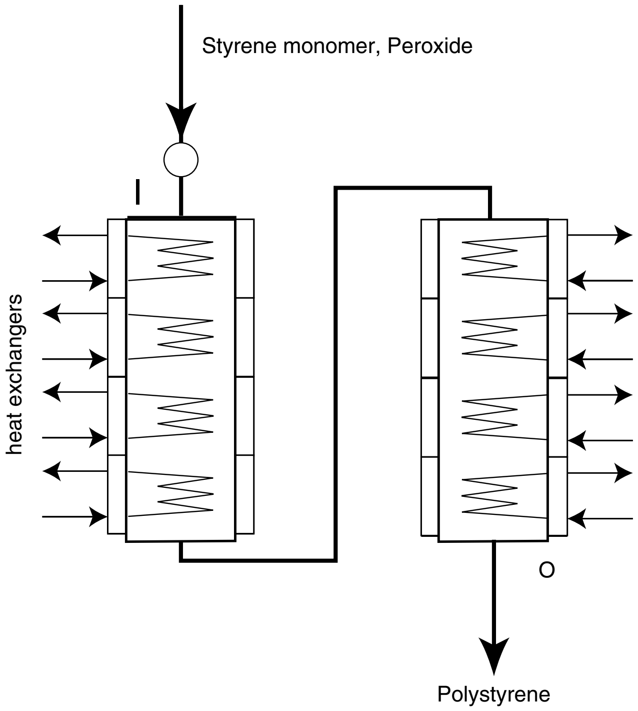


Fig. 1. Schematic of the tubular reactor.

2.1. Under-actuated structure

A close-up view of the reactor shows the heat exchangers and the temperature sensors (see Fig. 2). The reactor is divided in eight zones, each of which has two sensors and one heat exchanger. One is located at the middle of the zone, and one is located at its end. More complicated sensor configurations (with variable number of sensors and varying locations) could also be considered but these are out of the scope of this paper (optimal placement for such process is indeed an important topic as underlined in [9]). In this study, the total length of the tubular reactor is scaled to 1, and the velocity v of the flow inside the reactor is 0.01. This system can be considered as under-actuated since the eight heat exchangers can only produce piecewise constant control along the reactor's length. Classically, polymer viscosity is very high and laminar flow is assumed for modelling. The industrial tubular reactor underlying this study is very thin due to the heat transfer constraints. These factors lead us to model the reactor dynamics as a set of one-dimensional partial differential equations (PDE). Measurements of the temperature T are available at a

[1] Interested readers can also find in the previous reference developments of a model predictive control that allows the successful control of monomer conversion in a different context.

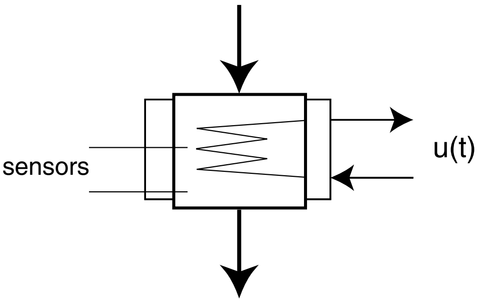


Fig. 2. Close-up view of the reactor.

finite number of locations along the reactor. Online measurement of monomer conversion x and peroxide concentration G is currently out of reach.

2.2. Model

Let $t \in \mathbb{R}^+$ represent the time and $z \in [0,1]$ the coordinate along the reactor, then the dynamics of the tubular reactor can be described by a set of three partial differential equations

$$\begin{cases} C_p(T,x)\left(\dfrac{\partial T(t,z)}{\partial t} + v\dfrac{\partial T(t,z)}{\partial z}\right) = \Delta H(T)r(T,x,G) + u, \\ \dfrac{\partial x(t,z)}{\partial t} + v\dfrac{\partial x(t,z)}{\partial z} = r(T,x,G), \\ \dfrac{\partial \log(G(t,z))}{\partial t} + v\dfrac{\partial \log(G(t,z))}{\partial z} = f(T), \end{cases} \tag{1}$$

with boundary and initial conditions

$$T(t,0) = T_{t,0}, \quad x(t,0) = x_{t,0}, \quad G(t,0) = G_{t,0}, \tag{2}$$

$$T(0,z) = T_{ic}(z), \quad x(0,z) = x_{ic}(z), \quad G(0,z) = G_{ic}(z) \tag{3}$$

where notation is defined in Table 1. The terms on the right hand side of (1) are as follows. We have

$$\Delta H(T) = a_0 + \frac{a_1}{T+273} + a_2(T+273),$$

$$r(T,x,G) = k_1(T,x,G)k_2(T,x),$$

where a_0, a_1 and a_2 are constant coefficients and k_1 is of the form

$$k_1(T,x,G) = \sqrt{F_1(T,x) + GF_2(T,x)}, \tag{4}$$

with F_1 and F_2 two smooth functions, while

$$f(T) = e_0 \exp\left(\frac{e_1}{T+273}\right),$$

where e_0 and e_1 are constant coefficients. A polynomial fit is used for $C_p(T,x)$ (affine function in x with third order polynomial in T as coefficients). This kinetic model arises from the classic Hui and Hamielec approach [10]. Typical values of the $r(T,x,G)\Delta H$ term are shown in Fig. 3 (where scales are omitted for confidentiality reasons). One clearly sees that the reaction mostly takes place in the middle of the tubular reactor.

The first equation in (1) is a heat balance, the second describes the conversion of monomer, and the third one represents the organic peroxide initiator dynamics that is thermally activated. Details about the kinetic scheme of this polymerization reaction and the role of the peroxide initiator can be found in [11]. The velocity v is considered constant. We thus neglect the effect of density variations.

Table 1
Nomenclature

Symbols	Quantity	Unit
$T(t,z)$	Temperature in the reactor	Celsius degrees
$x(t,z)$	Monomer conversion	–
$G(t,z)$	Peroxide concentration	$\mathrm{mol\,m^{-3}}$
v	Velocity of the particles along the reactor	$\mathrm{m\,s^{-1}}$
$C_p(T,x)$	Heat capacitance of the fluid	$\mathrm{J\,K^{-1}}$
$u(t,z)$	Heat exchange (control)	$\mathrm{J\,s^{-1}}$
$r(T,x,G)$	Rate of reaction	$\mathrm{s^{-1}}$
$\Delta H(T)$	Reaction enthalpy	K

2.3. Control objectives

Two control problems are considered. In the first place, the problem of regulating the temperature profile along the reactor to a given profile T_{sp} is considered, which guarantees good product quality at the end of the reactor. Then, we consider the problem of allowing fast transitions between desired temperature profiles corresponding to good product quality of different materials.

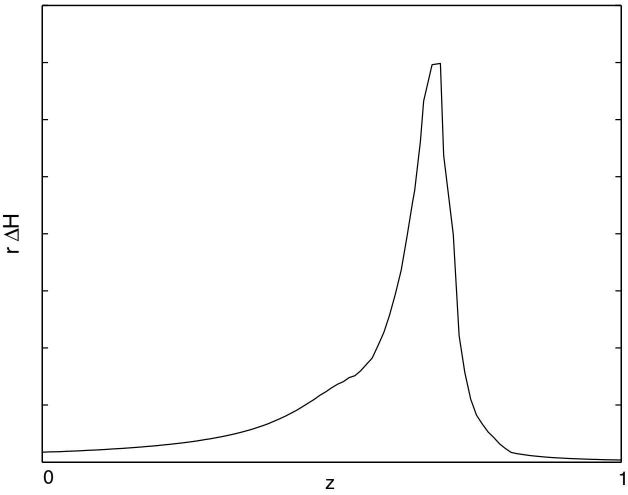


Fig. 3. Exothermicity along the reactor at steady state. Exact scales are omitted for confidentiality reasons.

2.4. Simulation setup

In this section, the testbed we base all our study on is considered. The three PDEs in Eq. (1) are discretized along the z variable to obtain three systems of ordinary differential equations (following the tanks-in-series model [12]). This approach is alike other methods found in the literature. Our model does not imply parabolic PDEs (i.e. with spatial diffusion operators) and so is not easily described by small finite-dimensional systems. In [13], a discretization of the coupled temperature and conversion PDEs similar to ours is used, but a different control structure is considered (control of temperature and concentration is achieved from the inlet side of the reactor). Radial variations are not taken into account either. Advanced methods could also be considered but are out of the scope of this study: e.g. in [14] a very fine representation of the model reactor (considering radial variation of all variables) is solved by the method of lines using either finite volume discretization or global spline orthogonal collocation.

Spatial discretization. The partial derivatives with respect to z are discretized using finite backward differences according to

$$\left.\frac{\partial W(t,z)}{\partial z}\right|_{(t,z_i)} \cong \frac{W(t,z_i)-W(t,z_{i-1})}{\delta_z},$$

for any variable of time and space W, where $(z_1,\ldots,z_n)$ are the cells, equally distributed in space, in which we discretize the reactor, and $\delta_z=(z_i-z_{i-1})$. As a result, letting $T_i(t)=T(t,z_i)$, $x_i(t)=x(t,z_i)$, $G_i(t)=G(t,z_i)$, $r_i=\Delta H(T_i)r(T_i,x_i,G_i)$, $k_i=r(T_i,x_i,G_i)$, $f_i=f(T_i)$, $C_{p,i}=C_p(T_i,x_i)$, we obtain the three systems of ordinary differential equations

$$\begin{cases}\dot{T}=AT+bc_T+C_p^{-1}(r+Bu),\\ \dot{x}=Ax+bc_x+k,\\ \dfrac{\mathrm{d}\log G}{\mathrm{d}t}=A\log G+bc_G+f,\end{cases} \tag{5}$$

where $T=(T_1,\ldots,T_n)^{\mathrm{T}}$, $x=(x_1,\ldots,x_n)^{\mathrm{T}}$, $G=(G_1,\ldots,G_n)^{\mathrm{T}}$, $r=(r_1,\ldots,r_n)^{\mathrm{T}}$, $k=(k_1,\ldots,k_n)^{\mathrm{T}}$, $f=(f_1,\ldots,f_n)^{\mathrm{T}}$, $bc_T=(T_{t,0},0,\ldots,0)^{\mathrm{T}}$, $bc_x=(x_{t,0},0,\ldots,0)^{\mathrm{T}}$, $bc_G=(\log(G_{t,0}),\ 0,\ldots,0)^{\mathrm{T}}$, and C_p is a diagonal matrix with diagonal entries $C_{p,i}$. $A=\frac{v}{\delta_z}(a_{ij}=(-1)\delta_i^j+\delta_{i-1}^j)$ is the $n\times n$ backward differences matrix, B is the $n\times 8$ input matrix, and $u=(u_1,\ldots,u_8)^{\mathrm{T}}$. Each actuator i acts on its zone of competence, which will be referred to as zone i. In particular, if the ith zone has n_i cells, with $n_1+\cdots+n_8=n$, then in the ith column of B the first $n_1+\cdots+n_{i-1}$ elements are zeros, the elements from $n_1+\cdots+n_{i-1}+1$ to $n_1+\cdots+n_i$ are ones, and the remaining elements are zeros.

Other possible choices for the spatial discretization method include forward difference equations, centered difference equations and second order methods such as the Lax–Wendroff numerical scheme (see [15]). Forward difference approximation results into an unstable A matrix when $v>0$, which is our case. The centered difference approximation produces a matrix that has imaginary eigenvalues and therefore is not asymptotically stable. The Lax–Wendroff second order method produces a A matrix with complex eigenvalues causing unrealistic oscillations.

Simulation setup. The value of n is chosen to be 100. The reason of this choice is a compromise between the time needed for simulating the system and the accuracy to which the spatial partial derivative is approximated. The numerical damping induced on the transport phenomena drops by only 2% (using a standard Runge–Kutta solver) when n is increased from 100 to 200, while the required computational effort rises by 100%.

3. Basic PI control designs

Based on the model in (5), two PI schemes used in the industrial setting were considered. These are diagonal control structures. Classically, derivatives term (D) were omitted to prevent temperature sensor noise from being amplified. In these two schemes, only one measurement in each zone is used. In the first scheme, the measurement is taken at the center of the zone, while in the second scheme the measurement is taken at the end of the zone. From a theoretical point of view, these choices induce strong constraints on the controller structure and lead to performance deterioration when compared to the system with a full controller matrix (as pointed out in [16]). However, this deterioration has to be weighted against design simplicity and failure tolerance. In fact, each block controller can be designed for the isolated subsystem, and fewer controller parameters need to be chosen than for the full system. Further, stability and performance are preserved to some degree when individual sensors or actuators fail. This failure tolerance is an attractive feature in the industrial framework. Part of such a controller can be turned off without dramatically affecting the system.

3.1. PI controller with measurement at the center of the zone

In this context, only the eight sensors at the center of the zone are used. Let n_i be the number of cells in zone i, and let $T_{m_1},\ldots,T_{m_8}$ denote the measured temperatures, then we have that $m_i=\sum_{j=1}^{i-1}n_j+n_i/2$. Let $T_{\mathrm{sp}}=(T_{\mathrm{sp},1},\ldots,T_{\mathrm{sp},n})$ and $u_{\mathrm{ref}}=(u_{\mathrm{ref},1},\ldots,u_{\mathrm{ref},8})^{\mathrm{T}}$ denote the reference temperature profile and the corresponding constant reference input respectively. The closed loop control laws are

$$u_i = -K_{\mathrm{P},i}(T_{m_i} - T_{\mathrm{sp},i}) - K_{\mathrm{I},i}\int (T_{m_i} - T_{\mathrm{sp},i}) + u_{\mathrm{ref},i},\quad i \in \{1,\ldots,8\}. \tag{6}$$

The gains $K_{\mathrm{P},i}$ and $K_{\mathrm{I},i}$ are tuned in descending cascade (we tune first the PI of the first zone, and then the PI of the following zones by leaving the already tuned PI on) by means of the Ziegler–Nichols closed loop PID tuning rule (see for example [17]). When a step disturbance of amplitude 10% is applied at the entrance of the reactor, the response of the closed loop system is given in Fig. 4. The left plot of the same figure shows the asymptotic temperature profile along the reactor. The performance at locations different from the ones at which the measurement occurs is not satisfactory: only the measured temperatures are well tracked. The right plot shows the control effort $u_i - u_{\mathrm{ref},i}$.

3.2. PI controller with measurement at the end of the zone

In this setup, only the eight end of zone temperature sensors $T_{m_1},\ldots,T_{m_8}$ are used. With n_i the number of cells in zone i, we have $m_i = \sum_{j=1}^{i} n_j$. The closed loop control law u is given again by expression (6), where the reference value $T_{\mathrm{sp},i}$ and $u_{\mathrm{sp},i}$ are appropriately computed. Proportional and integral gains are still tuned in descending cascade using Ziegler–Nichols method. Closed loop response are shown in Fig. 5.

By contrast with results in Fig. 4, this control design produces an asymptotic temperature profile that is satisfactory not only where the measurement is performed. An analysis of such a performance difference is explained in the following section.

3.3. Comparisons of the two measurement schemes

Spatially distributed offsets in the asymptotic temperature profile still persist when the integral part of the

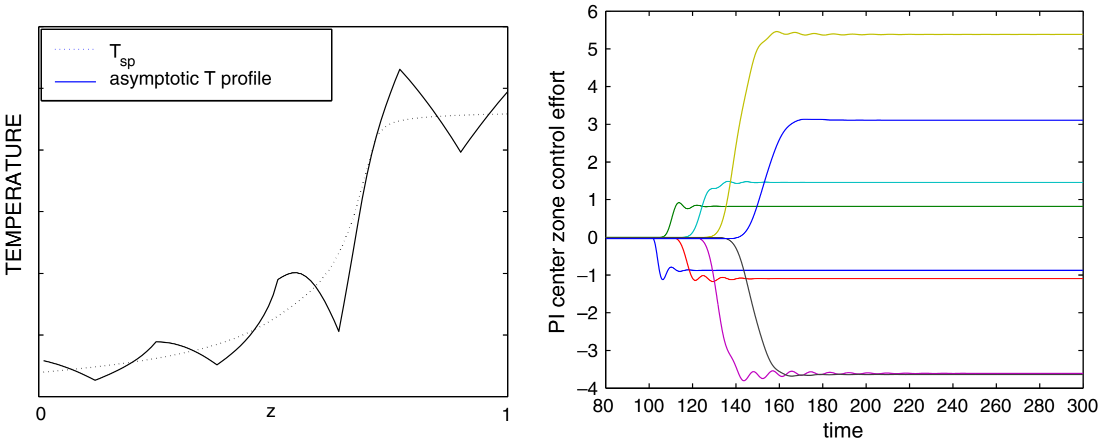


Fig. 4. PI controller with measurement at the center of the zone. Performance of the closed loop system when a step of amplitude 10% of input temperature is applied at the entrance of the reactor. Left: asymptotic temperature profile. Right: control effort.

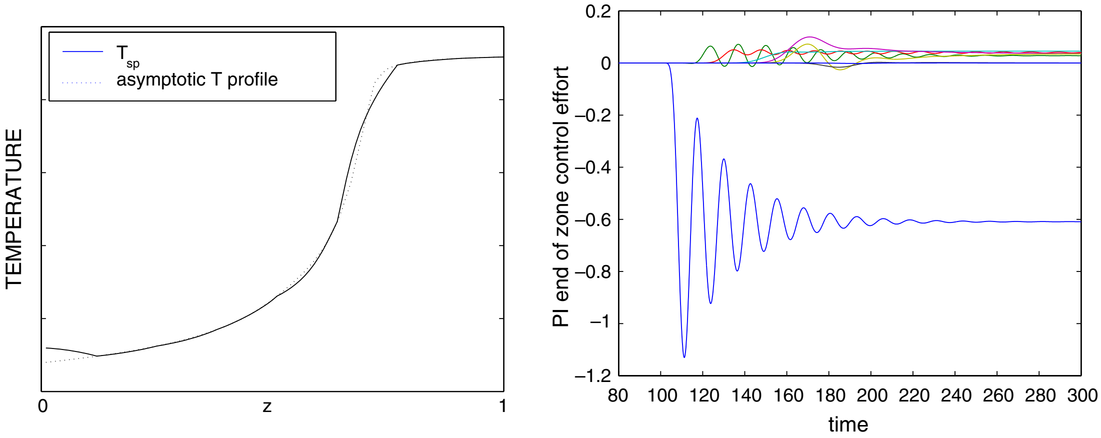


Fig. 5. PI controller with measurement at the end of the zone. Performance of the closed loop system when a step of amplitude 10% of input temperature is applied at the entrance of the reactor. Left: asymptotic temperature profile. Right: control effort.

controller is turned off. Thus for the sake of simplicity, we neglect the integral terms.

Perturbation analysis of disturbance rejection. Consider two adjacent zones $i-1$ and i, we wish to estimate the attenuation of a step disturbance of amplitude d entering at the beginning of zone $i-1$. One can ask where the best location $\alpha \in [0,1]$ for the temperature measurement is for having the highest attenuation onto the next zone. Having $\alpha = 1$ or $\alpha = 0$ means that the measurement is performed at the beginning or at the end of the zone respectively. Once controlled by one of the proposed controllers, the temperature PDE is a transport equation in first order approximation. Assuming that stability is achieved by the previous control loops, perturbation in the control δu affects the temperature by the following integral formula for all $x \in [0,1]$

$$\delta T(i-x,t) = \delta T(i-1,t-(1-x)) + \int_{t-(1-x)}^{t} \tilde{B}(t-s)\delta u_i(s)\,\mathrm{d}s,$$

with $\tilde{B} > 0$. Now assume that δu_i is a closed loop signal $\delta u_i(t) = -K_i T(i-\alpha,t)$. Steady state values satisfy

$$T(i-x,\infty) = \frac{1 + K_i\int_{1-x}^{1-\alpha}\tilde{B}(s)\,\mathrm{d}s}{1 + K_i\int_{0}^{1-\alpha}\tilde{B}(s)\,\mathrm{d}s} T(i-1,\infty)$$

and in particular

$$T(i,\infty) = \frac{1 - K_i\int_{1-\alpha}^{1}\tilde{B}(s)\,\mathrm{d}s}{1 + K_i\int_{0}^{1-\alpha}\tilde{B}(s)\,\mathrm{d}s} d.$$

As K_i increases (strong gains seem often a good option, especially with the Ziegler–Nichols tuning rules), the disturbance is attenuated from the entrance of the zone to the exit by a factor asymptotically equal to

$$-\int_{1-\alpha}^{1}\tilde{B}(s)\,\mathrm{d}s \bigg/ \int_{0}^{1-\alpha}\tilde{B}(s)\,\mathrm{d}s.$$

Since $\alpha > 0$, this term is negative. The optimum is to choose $\alpha = 0$, meaning the best measurement location in terms of disturbance rejection is at the end of the zone. With this choice, the disturbance does not propagate to the next zone. If a different choice is made (e.g. centre of the zone measurement), then the disturbance propagates with an opposite sign to the next zone. This explains the steady state reached with the center of the zone measurement in Fig. 4. On the contrary, one can clearly see in Fig. 5 that the disturbance affects mainly the first zone and then is strongly attenuated by the end of the zone measurement controller. Similarly, the control effort is focused on the first zone in Fig. 5.

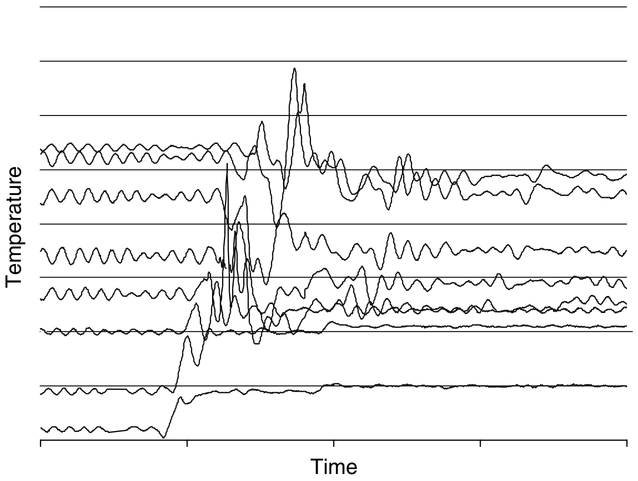


Fig. 6. Industrial results with an end of the zone PI controller (unsatisfactory).

3.4. Industrial results with a PI controller

In Fig. 6, the results of a grade transient are shown. Actual scales are omitted for confidentiality reasons. Yet, it is possible to represent this transient as a non uniform shift in temperature for the different zones ranging from +10% to −5%. The measurement are all performed at the center of the zone except one that is measured at the end. While the regulation is satisfactory before the transient, some oscillations persist. This is mostly due to the high value of the gains chosen for the PI to get a strong hold on the system during transients. The transient itself gives rise to oscillations. In the first zone, the system is steered smoothly by the controller but the next zone is harder to control as the reaction is the strongest. Finally, the last zones are suffering from the disturbances travelling from the first zones. These industrial results are consistent with the study presented in Section 3.1. This behavior was considered a serious bottleneck for the plant productivity and was at the origin of our study on the control upgrade.

4. LQR controller design

In this section, a centralized PI controller is proposed in which the proportional gains are optimally chosen using LQR design. Weights of the input channels are computed through this LQR design. In a future work, we could investigate the use of higher order controller (that may include some explicit or implicit observer for instance). In the sequel, we will refer to this centralized PI scheme as LQR to remind the way the proportional gain was designed.

4.1. Control setup

The LQR is designed to regulate the temperature about the desired profile T_{sp}. To this end, the first equation of (5) is linearized about $(T_{\mathrm{sp}}, x_{\mathrm{sp}}, G_{\mathrm{sp}}, u_{\mathrm{sp}})$. This yields

$$\dot{T} = \left(A + C_p^{-1} \frac{\mathrm{d}r}{\mathrm{d}T}\bigg|_{(T_{\mathrm{sp}}, x_{\mathrm{sp}}, G_{\mathrm{sp}})} \right) T + Bu. \tag{7}$$

The values of C_p range from 0.4 to 1, and its variation can be roughly seen as a multiplicative disturbance acting on u. We therefore neglect C_p^{-1} multiplying Bu in the control design. As we will explain in a later section, this does not cause a problem due to the robustness properties of the LQR. Define $\overline{A} = (A + C_p^{-1} \frac{\mathrm{d}r}{\mathrm{d}T}|_{(T_{\mathrm{sp}}, x_{\mathrm{sp}}, G_{\mathrm{sp}})})$. The LQR problem is solved in each one of the eight zones separately, neglecting the propagation of perturbations from zone i to zone $i+1$. Perturbations are very small in zone $i+1$ if the LQR in zone i is properly designed. This choice is due to numerical issues that arise when considering the problem of assigning the eigenvalues of the relatively large dimensional entire system. Thus, for the control design purpose we have $\overline{A} \cong \mathrm{block}-\mathrm{diag}(\overline{A}_1, \ldots, \overline{A}_8)$, with $\overline{A}_1 \in \mathbb{R}^{n_i \times n_i}$ previously defined. Then, we have eight identical LQR problems for the pairs $(\overline{A}_i, B_i)$, with $B_i \in \mathbb{R}^{n_i}$ a vector of ones. For each i the functional

$$J(u) = \int_0^\infty \left(u_i(t)^2 + T_{\sum_{j=1}^{i} n_j}(t)^2 \right) \mathrm{d}t, \tag{8}$$

is minimized, where $T_{\sum_{j=1}^{i} n_j}(t)$ is the temperature at the end of zone i, and $u_i(t)$ is the control input of the same zone. Let $K_{\mathrm{P},i} \in \mathbb{R}^{n_i}$ denote the optimal vector of proportional gains in zone i, we use

$$u_i = -K_{\mathrm{P},i}^{\mathrm{T}}(I_p T - T_{\mathrm{sp}}) - K_{\mathrm{I},i} \int_0^t (T_{n_i}(s) - T_{\mathrm{sp},i})\,\mathrm{d}s + u_{i,\mathrm{ref}}, \tag{9}$$

where the integral term (tuned a posteriori) guarantees zero asymptotic error with respect to step disturbances at least at the end of the zone. Since we assume to have at most three possible measurements in each zone (i.e. at the beginning which corresponds to the end of the previous zone, at the center, and at the end), we linearly interpolate the measurements we have in each zone in order to do the feedback from the interpolated temperature. The $n_i \times n_i$ matrix I_p models the interpolations, that is $T_{\mathrm{interp}} = I_p T$. Ideally, the larger the number of measurements the closer I_p to the identity matrix.

Fig. 7 reports the behavior of the closed loop system when a step disturbance of 10% is applied at the entrance of the reactor. The left plot shows the asymptotic temperature profile along the reactor. The right plot shows the control effort. Comparisons with Fig. 5 stress that the transients are smoother and the control signal does not oscillate.

4.2. Industrial results with a LQR controller

Fig. 8 shows industrial results obtained with the LQR design that has been in service since 2000 (see [2]). Again, a grade transient is considered, which is different from the one presented in the PI Section but is just as difficult to achieve. Before the transient, the system is well controlled. Residual oscillations are very small compared to the PI results in Fig. 6. This is due to the better suited choice of the gains. The transient itself is satisfactory. It is fast with mostly monotonic trajectories and no propagation of undesirable perturbations between the zones. After the transient, the system is well controlled. This behavior has been considered successful since this new controller was designed and installed. Indirectly, it also allowed to increase the productivity by an upgrade of the total amount of monomer that can be processed (changing the velocity of the flow and reference temperature profiles) without changing any actuators or sensors.

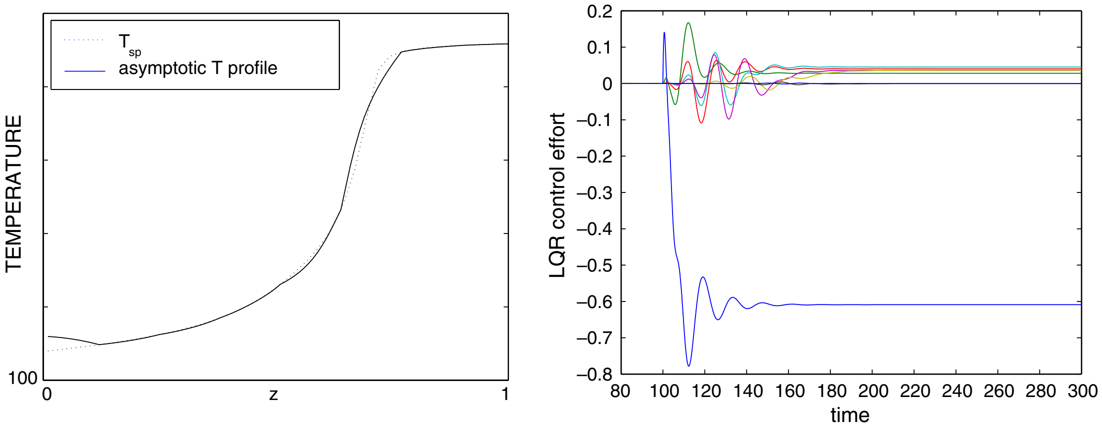


Fig. 7. LQR. Performance of the closed loop system when a step of 10% is applied at the entrance of the reactor.

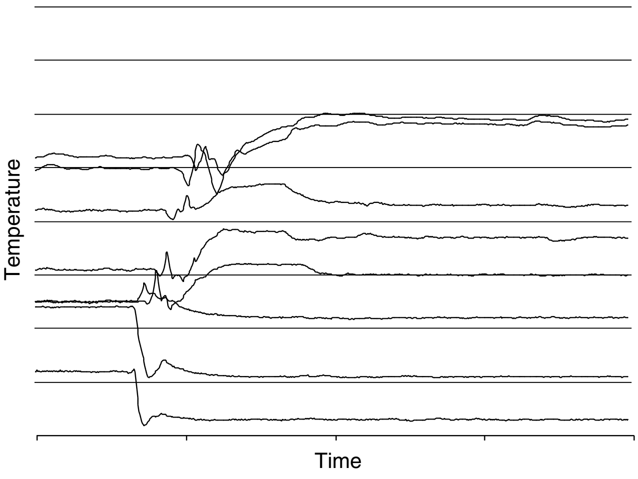


Fig. 8. Industrial results with a LQR controller.

4.3. Comparison between the LQR and the PI with measurement at the end of the zone

The LQR controller is a natural evolution of the decentralized PI control scheme in the case in which a full measurement is possible. It simply provides a methodology for optimally choosing the gains of the proportional controller, and thus it leads to a centralized PI controller structure. A better performance of the LQR controller is to be expected because for designing the proportional gains an optimization is run, and because the number of measurements considered is larger than the one in the decentralized PI scheme. The asymptotic performance of the decentralized PI is comparable to the one of the centralized scheme obtained with LQR design. The transient performance of the LQR design is instead better as expected. To investigate this feature further on, the system was simulated starting from a temperature profile 20% higher than the desired one. The results with the decentralized PI and with the LQR are shown in Fig. 9. The decentralized PI gives rise to instability as its gains were designed around the desired temperature profile. The LQR instead steers the system to the desired profile. This robustness property with respect to unmodelled dynamics is due to the centralized nature of this controller and to the number of measurements. In Fig. 10 (right), the Nyquist plots of $\det[I + H(\mathrm{j}\omega)]$ and of $\det[I + \widetilde{H}(\mathrm{j}\omega)]$, where $H(s) = K(sI - A)^{-1}B$ and $\widetilde{H}(s) = K(sI - \widetilde{A})^{-1}B$, with $\widetilde{A} = A + BK(I - I_p)$ are depicted. The effect of a poor interpolation is to reduce the robustness margins of the system with respect to input perturbations. In particular, if eight measurements are considered (e.g. only end of the zone sensors are available) the LQR controller results in an unstable closed loop system.

In conclusion, the LQR design turns out to be easy because the exact expression of the kinetics does not need to be known (the x and G dynamics can be neglected), and thus the design considers only the linearized version of the T dynamics. It is more robust than the PI with measurement at the end of the zone, this being due to its centralized structure and to the quality of the measurement interpolation.

5. Nonlinear trajectory generation control approach

In the latest years, optimal control problems with systems governed by partial differential equations subject to control and state constraints have been extensively studied. We refer for instance to [18] for necessary optimality conditions for special cases of elliptic problems and to [19,20] for numerical studies. One may think to use these tools as a closed loop controller as in estab-

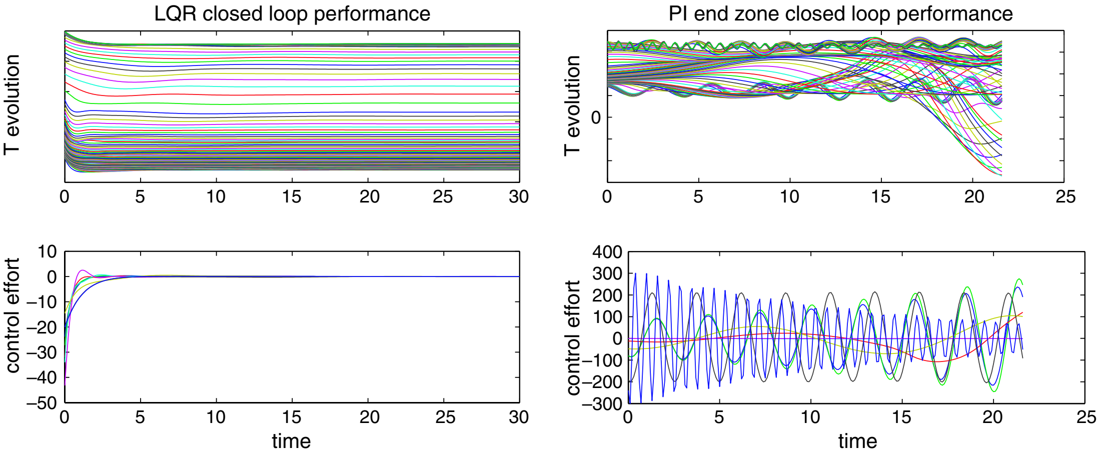


Fig. 9. Performance of the LQR controller with 17 uniformly distributed measurements (left plot) and of the decentralized PI with measurements at the end of the zone (right plot), when the system is started at a temperature profile 20% higher that the desired one T_{sp}.

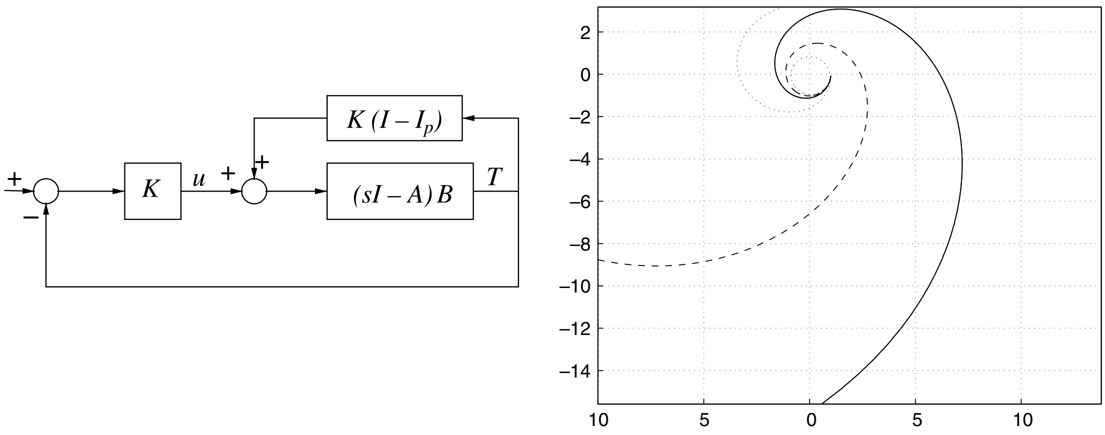


Fig. 10. Linearized temperature control loop (left). Nyquist plot of det$[I + H(\mathrm{j}\omega)]$ (solid line), and of det$[I + \widetilde{H}(\mathrm{j}\omega)]$ for two different interpolations. The dashed plot corresponds to a number of 24 uniformly distributed measurements along the reactor, while the dotted plot corresponds to a number of 17 uniformly distributed measurements.

lished receding horizon control strategies for systems governed by ordinary differential equations (see [21,22] for instance).

A direct method to solve these problems is to use finite dimensional approximations for both the control and the state and to enforce constraints at some prescribed grid points (see [23] for an overview of this direct collocation approach). This results in a nonlinear program, see [24,25]. In [6], a different methodology is proposed. For optimal control of nonlinear ordinary differential equations of the form $\dot{x} = f(x) + g(x)u$, where $\mathbb{R} \ni t \mapsto x \in \mathbb{R}^n$ and $\mathbb{R} \ni t \mapsto u \in \mathbb{R}^m$, it is shown [26] that it is possible and computationally efficient to reduce the dimension of the nonlinear programming problem by using inversion to reduce the number of dynamic constraints in the problem. In this approach, variables are eliminated through explicit substitutions. The "inversion" concept can be extended to the field of partial differential equations. For numerical implementation, the variables can be parameterized by tensor-product B-splines (among other basis functions). Their partial derivatives can be easily (analytically) computed, combined, and substituted to as many components of the states and the control as possible in both the cost functions and the constraints.

After the variables have been parameterized in terms of B-spline surfaces, the coefficients $C^l_{i,j}$ of the B-spline basis functions will be found using sequential quadratic programming. This problem is stated as

$$\min_{y \in \mathbb{R}^{N_c}} F(y) \text{ subject to } l_b \leqslant c(y) \leqslant u_b,$$
$$\text{where } y = (C^1_{1,1}, C^1_{1,2}, \ldots, C^p_{p_t,p_x}) \quad \text{and} \quad N_c = p_t * p_x * p. \tag{10}$$

$F(y)$ is the discrete approximation of the chosen objective function. We then use NPSOL [27] as the sequential quadratic programming to solve this new problem.

5.1. Optimal control problem formulation

We consider the variables x and T only, while we neglect the variation of the G value that we assume fixed to its reference G_{sp}. Then we formulate the problem of shifting the temperature profile from a starting profile $T(0,z) = T_{ic}(z)$ to the desired final profile $T(t_{\mathrm{f}}, z) = T_{\mathrm{sp}}(z)$ for a given transient time t_{f} as a constrained minimization problem. In this setup we relax the underactuated model by assuming full actuation (we could add constraints on u but this could be expensive in terms of computation time). In particular we want to find the $(t,z) \mapsto (T(t,z),\ x(t,x),\ u(t,z))$ that minimizes the cost

$$J(T,x,u) = \int_{\tau=0}^{t_{\mathrm{f}}} \int_{s=0}^{1} c_1 u(\tau,s)^2 + c_2 (T(\tau,s) - T_{\mathrm{sp}}(s))^2 \,\mathrm{d}\tau\,\mathrm{d}s, \tag{11}$$

subject to the boundary constraints

$$T(0,z) = T_{ic}(z), \quad T(t_{\mathrm{f}},z) = T_{\mathrm{sp}}(z), \quad T(t,0) = T_{\mathrm{inlet}},$$
$$T(0,z) = x_{ic}(z), \quad x(t,0) = x_{\mathrm{inlet}} \tag{12}$$

and to the domain constraints

$$\left(\frac{\partial T(t,z)}{\partial t} + v\frac{\partial T(t,z)}{\partial z}\right) = \frac{r(T,x,G)}{C_p(T,x)} + \frac{u}{C_p(T,x)}, \tag{13}$$

$$\frac{\partial x(t,z)}{\partial t} + v\frac{\partial x(t,z)}{\partial z} = k(T,x,G), \tag{14}$$

$$l_b \leqslant u(t,z) \leqslant u_b. \tag{15}$$

We reduce the number of variables involved in the constrained minimization problem to two, by rewriting u as a function of x and T by means of equation (13). Therefore, the cost in (11) becomes

$$J(T,x) = \int_{\tau=0}^{t_{\mathrm{f}}} \int_{s=0}^{1} c_1 \Bigg(\left(\frac{\partial T(\tau,s)}{\partial \tau} + v\frac{\partial T(\tau,s)}{\partial s}\right) C_p(T,x)$$
$$-r(T,x,G)\Bigg)^2 + c_2 T(\tau,s)^2 \,\mathrm{d}\tau\,\mathrm{d}s,$$

subject to the boundary constraints given in Eq. (12), and to the domain constraints

$$\frac{\partial x(t,z)}{\partial t} + v\frac{\partial x(t,z)}{\partial z} = k(T,x,G), \tag{16}$$

$$0 \leqslant x(t,z) \leqslant 1 - c_0, \tag{17}$$

$$l_b \leqslant \left(\frac{\partial T(t,z)}{\partial t} + v\frac{\partial T(t,z)}{\partial z}\right)C_p(T,x) - r(T,x,G) \leqslant u_b, \tag{18}$$

where we added the domain constraint on the x values to avoid any sign change for the argument of the square root in the reaction expression given in (4). This term would become negative if this constraint is violated and thus the numerical solver would fail.

Once the optimal solution $T^{\text{opt}}(t,z)$, $x^{\text{opt}}(t,z)$ has been found, the optimal input is computed as

$$u^{\text{opt}}(t,z) = \left(\frac{\partial T^{\text{opt}}(t,z)}{\partial t} + v\frac{\partial T^{\text{opt}}(t,z)}{\partial z}\right)C_p(T^{\text{opt}},x^{\text{opt}}) - r(T^{\text{opt}},x^{\text{opt}},G).$$

5.2. Optimal solutions

The NTG software package [26,6] is used to solve the constrained optimization problem explained in the previous section. Typical problems that can be solved with this package include the example given in Fig. 11 where the parameters are set as $c_1 = c_2 = 1$, $l_b = -20$, $u_b = 20$, constraints are enforced on a 16×16 uniform grid, Eq. (14) is satisfied with a tolerance of 10^{-5}. The total number of spline coefficients is 144. In this problem the initial temperature offset is distributed along the reactor with a peak at 25. The optimal control strategy saturates the constraints.

5.3. Using NTG with a closed loop controller

The optimal control input computed by NTG is then used in our simulator. In addition to the open loop optimal control, we use also a closed loop controller to be able to track the optimal temperature evolution T^{opt}. Let $\mathbf{u}^{\text{opt}}(t) = (u^{\text{opt}}(t,z_1),\ldots,u^{\text{opt}}(t,z_n))^{\text{T}}$, and $\mathbf{T}^{\text{opt}}(t) = (T^{\text{opt}}(t,z_1),\ldots,T^{\text{opt}}(t,z_n))^{\text{T}}$, then the control input $u \in \mathbb{R}^8$ used in model (5) is

$$u(t) = (B^{\text{T}}B)^{(-1)}B^{\text{T}}(\mathbf{u}^{\text{opt}}(t)) - K(I_pT(t) - \mathbf{T}^{\text{opt}}(t)),$$

with K a scalar constant. This last expression takes into account the under-actuated structure of our model by doing a least square approximation of the optimal control value. This strategy addresses the problem of regulation of distributed process with spatially-distributed control actuators and measurement sensors. In NTG the system is treated as a fully distributed system, the feedback term takes care of the imperfections of this

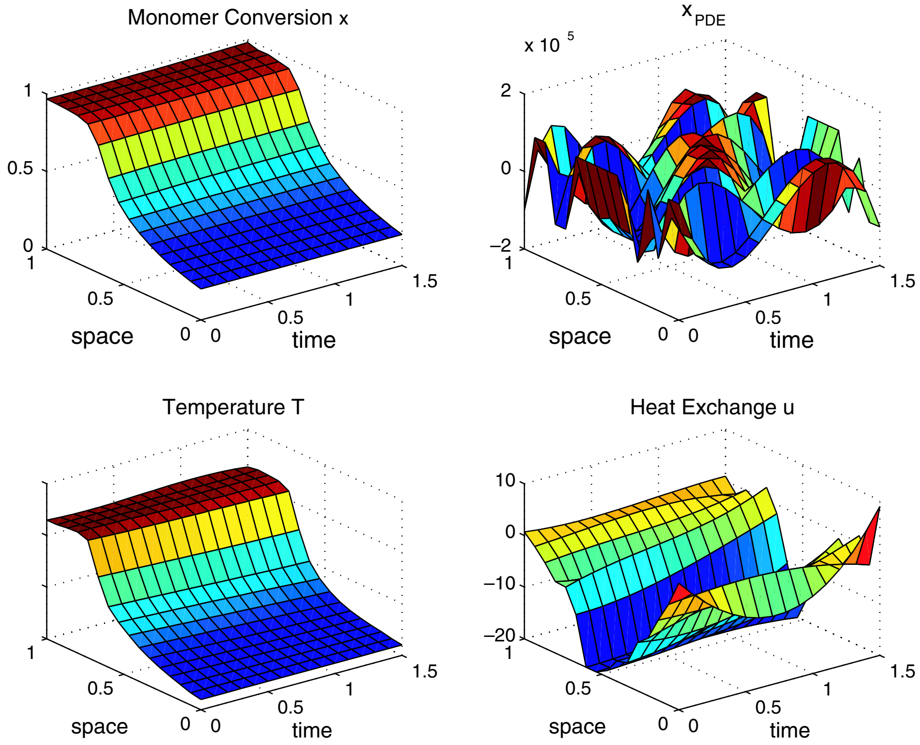


Fig. 11. Optimal solution found by the NTG software. We show the two dimensional plots of the optimal quantities on the grid where the constraints were enforced and the cost was minimized.

model. Further, this feedback term is needed for attenuating numerical and initialization errors in addition to errors due to the G unmodelled dynamics. The linearization about points on the optimal trajectory of the overall system, as this appears in equations (5), gives rise to complex eigenvalues with negative real parts. Even if the system is stable around the optimal trajectory, the presence of a non zero imaginary part gives rise to oscillations when a small error is present, due to the above explained factors. The amplitude of such oscillations become small after a time, which depends on the dynamics of the system, that is larger than the target final time t_f. The feedback term eliminates these oscillations within the final time t_f. As we can see from Fig. 12 the open loop control and the close loop one are very similar, meaning that only a small amount of error needs to be corrected.

Numerical setup. In the cost given in (11), we chose $c_1 = c_2 = 1$. The bounds on the input in (15) have been chosen to be $l_b = -100$, $u_b = 100$. The tolerance on the satisfaction of the x PDE given in (14) was chosen to be 10^{-2}. The two-dimensional t, z plots of the quantities $T^{\rm opt}(t,z)$, $x^{\rm opt}(t,z)$, $u^{\rm opt}(t,z)$ are reported in Fig. 11.

The computational time needed for computing the solutions is 4 minutes when the number of spline

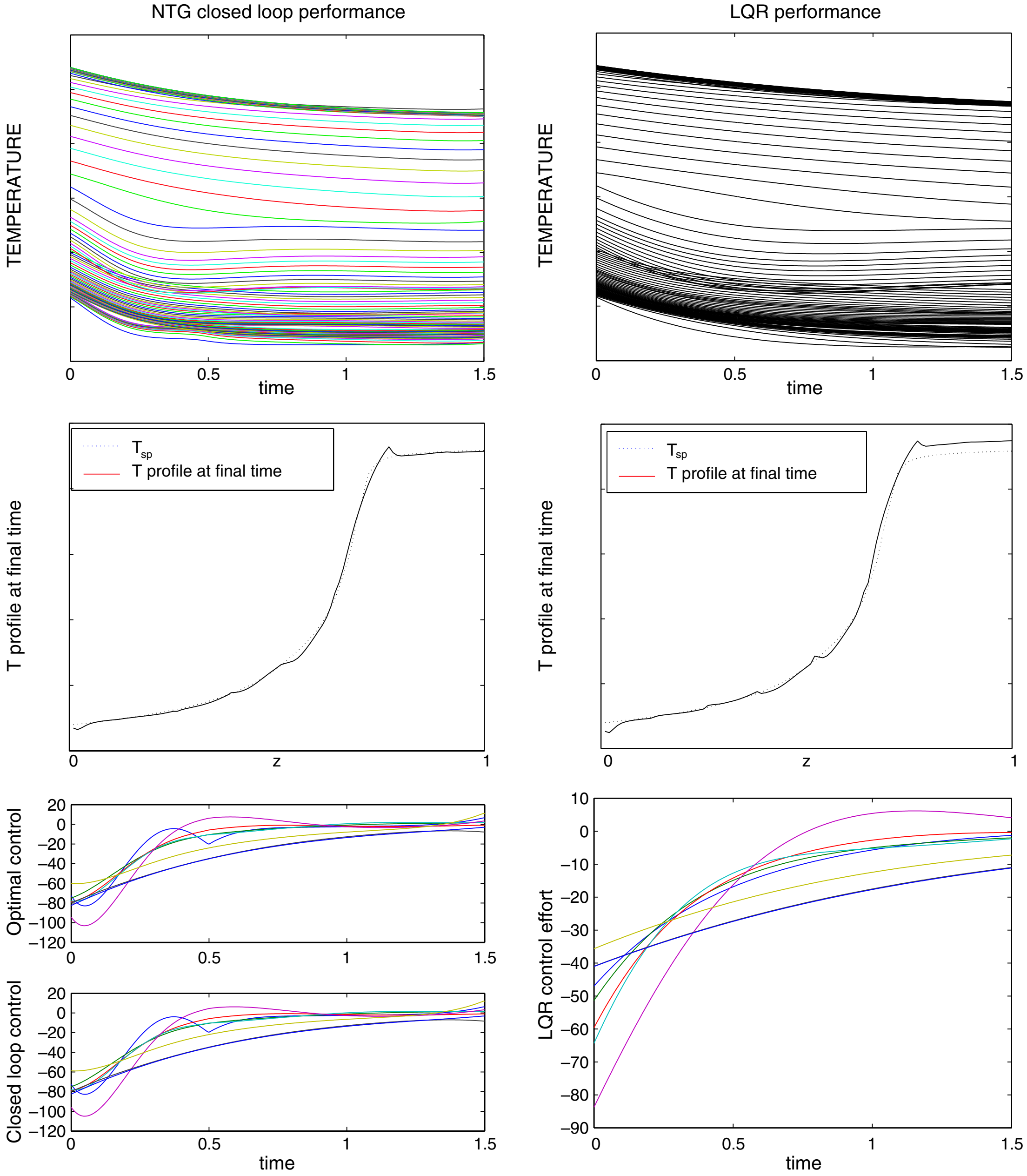


Fig. 12. Performance of the NTG controller (left plots), and of the LQR controller (right plots).

coefficients is 616. The x and T variables are approximated with 6th order B-splines using 6 and 4 knots for the z and t directions respectively. First partial derivatives are continuous across the knots (multiplicity of 2). The constraints are enforced on a 20×10 (z, t) mesh grid. This computation time may seem high when compared to results from the literature (see [28] where SNOPT is used instead of NPSOL) but these refer to systems of parabolic equations. Diffusion terms regularize the solutions of these dynamics and make collocation easier to run thanks to smoothness of the unknown variables. In the problem addressed here, the unknown variables are not very smooth, constraints are thus difficult to enforce and the SQP requires many iterations.

The simulations were run with a fixed step equal to 1.5/300. We show the simulation results in Fig. 12 next to the performance of the LQR designed in the previous section. The values of the cost given in (11) are 1.098×10^3 for the optimal solutions computed by the NTG, it is 1.0978×10^3 for the closed loop quantities obtained from the Simulink simulation, and it is 0.82×10^3 for the LQR. Cost functions are not comparable in this transient mode, but results are of the same order of magnitude. The smaller cost of the LQR is due to a smaller value of the controller, as it appears from Fig. 12, but the behavior of the temperature with the NTG controller is better since it has a faster transient, and it achieves the desired final profile within the target time t_f. This does not happen for the LQR controller whose transient is slower.

5.4. Perspectives of receding horizon control

It is very tempting to use such a numerical tool in the context of receding horizon control (RHC) as detailed in [29,30]. Our control problem incorporates a zero terminal constraint which is consistent with the RHC strategy proposed in [31,32]. One may also use a terminal cost instead, as proposed in [33]. This process has many of the interesting features that make RHC attractive for industrial applications (see [23]): frequent grade changes, possibly large disturbances. Because computational efficiency is vital to the success of such an online optimization, it seems important to test whether our numerical approach can be improved further. One way to achieve shorter computations time is to use well chosen initial guesses. As an example, we investigated the time required to solve the problem given in Fig. 11 with perturbed initial conditions. Results are as follow: the reference solution is computed in 560 s, perturbation of the initial condition offset of 25%, 50% and, 80% are solved using the first run as initial guess in 40, 60 and 80 s respectively (on average). It thus possible to significantly reduce the computational load provided a large enough set of initial guesses is computed offline. In this context, the use of an efficient tool is critical. One could also save more information than just good initial guess. We refer to [34] for a methodology that uses precalculated reference trajectories along with Hessians and gradients information in a real-time embedding strategy. It can also be interesting to consider alternative control configuration to propose a fault tolerant control strategy (as in [35]). This can be done by computing relevant initial guesses for such configurations. More work needs to be done though. A set of costs and terminal constraints has to be well chosen to provide stability in closed loop (see [32] for a discussion on this topic in the case of ordinary differential equations). Robustness is also an issue since computation time cannot be upper bounded (the number of SQP iterations is not limited [23]). It might be interesting to use feasible SQP to be able to exit in a prescribed time with a (possibly suboptimal) feasible solution (see [29] for a discussion on this subject).

6. Conclusion

In this paper, we have shown the main differences between four control schemes for the example of an underactuated exothermic tubular reactor. In particular, we considered a decentralized PI design with two different measurement schemes, and a centralized PI design in which the gains have been computed by means of an LQR design for the problem of regulating the temperature about a desired profile along the reactor. We then considered a nonlinear control scheme, the NTG controller, for the purpose of shifting fast the temperature between desired profiles along the reactor. Our study confirms the industrial results obtained with PI and LQR controllers and gives insight into the experimentally obtained performance. The measurement at the center of the zone scheme performs badly at locations different from the measured ones, while the measurement at the end of the zone scheme allows a good regulation performance everywhere along the reactor. Further, the LQR design allows faster transients than the decentralized PI controller, and its robustness characteristics allow to well reject uncertainties on the initial conditions. This suggests that such a control scheme can be successfully applied to the problem of shifting fast the operating point between temperature profiles along the reactor. This is confirmed by our simulation and industrial results. Finally, we showed how a nonlinear control scheme, the NTG, can be used to impose constraints on the input values and to shift the operating point between different temperature profiles along the reactor within an established time.

In the industrial framework, the key issue we encountered is the compromise between the need for performance and robustness and the model knowledge, availability of measurements, and limited actuation available for control. This is according to the results of [1] where the authors also highlight that actual implementation of advanced control theory in the polymerization area requires the improvement of measurement and state estimation techniques. To the light of this study, our recommendations are as follows. If only local temperature measurements are available, and one lacks knowledge of the kinetic law, we recommend using the end of the zone measurement scheme. If interpolation of the sensor values is accurate and the knowledge of the kinetics law appears reliable, then we recommend using the LQR for the sake of performance improvement. Finally, if a distributed control system (DCS) is available on site, and if the kinetics law are accurately known, then we suggest that a numerical tool, such as the one presented here, be used. The advantages of using such an approach are: constraints handling, and optimization with respect to the true nonlinear dynamics.

Acknowledgements

We gratefully thank Mr. Creff for his scientific support, Mr. Scarselli, Mr. Germain, Mr. Noel, Mr. de Linares, and Mr. Boueilh from the ATOFINA company for this work and TOTAL for his financial support.

References

[1] M. Embirucu, E.L. Lima, J.C. Pinto, A survey of advanced control of polymerization reactors, Polymer Engineering and Science 36 (4) (1996) 433–447.

[2] N. Petit, Systèmes à retards, platitude en génie des procédés et contrôle de certaines équations des ondes, Ph.D. thesis, Ecole des Mines de Paris, 2000.

[3] T. Larsson, S. Skogestad, Plantwide control—a review and new design procedure, Modeling Identification and Control 21 (4) (2000) 209–240.

[4] H. Cui, E.W. Jacobsen, Performance limitations in decentralized control, Journal of Process Control 12 (4) (2002) 485–494.

[5] W. Wang, D.E. Rivera, K.G. Kempf, Comparison of centralized versus decentralized model predictive control strategies to semiconductor manufacturing supply network, in: Proceedings of the 2003 American Control Conference, 2003.

[6] N. Petit, M.B. Milam, R.M. Murray, A new computational method for optimal control of a class of constrained systems governed by partial differential equations, in: Proceedings of the 15th IFAC World Congress, 2002.

[7] R.M. Murray, J. Hauser, A. Jadbabaie, M.B. Milam, N. Petit, W.B. Dunbar, R. Franz, Online control customization via optimization-based control, in: G. Balas, T. Samad (Eds.), Software-enabled Control, Information Technology for Dynamical Systems, Wiley-Interscience, 2003, pp. 149–174.

[8] M.P. Vega, E.L. Lima, J.C. Pinto, Modeling and control of tubular solution polymerization reactors, Computers and Chemical Engineering 21 (13) (1997) 1049–1054.

[9] C. Antoniades, P.D. Christofides, Integrating nonlinear output feedback control and optimal actuator/sensor placement for transport-reaction processes, Chemical Engineering Science 56 (2001) 4517–4535.

[10] A.W.T. Hui, A.E. Hamielec, Thermal polymerization of styrene at high conversion and temperatures. an experimental study, Journal of Applied Polymer Science 16 (1972) 749–769.

[11] G. Odian, Principles of Polymerization, third ed., John Wiley & Sons, 1991.

[12] O. Levenspiel, Chemical Reaction Engineering, third ed., John Wiley & Sons, Inc., 1999.

[13] D.M. Bošković, M. Krstić, Backstepping control of chemical tubular reactors, Computers and Chemical Engineering 26 (7–8) (2002) 1077–1085.

[14] E.F. Costa Jr., P.L.C. Lage, E.C. Biscaia Jr., On the numerical solution and optimization of styrene polymerization in tubular reactors, Computers and Chemical Engineering 28 (1–2) (2004) 27–35.

[15] R. Dautray, J.-L. LionsMathematical Analysis and Numerical Methods for Science and Technology, vol. 6, Springer-Verlag, 1993.

[16] M. Morari, E. Zafiriou, Robust Process Control, Prentice-Hall, Englewood Cliffs, 1989.

[17] K. Astrom, T. Hagglund, PID Controllers: Theory, Design, and Tuning, second ed., ISA—The Instrumentation, Systems, and Automation Society, 1995.

[18] J.-L. Lions, Optimal Control of Systems Governed by Partial Differential Equations, Springer-Verlag, 1971.

[19] H. Maurer, H. Mittelmann, Optimization techniques for solving elliptic control problems with control and state constraints. Part 2: Distributed control, Computational Optimization and Applications 18 (2001) 141–160.

[20] L.T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders, Large-scale PDE-constrained optimization, Lecture Notes in Computational Science and Engineering, vol. 30, Springer Verlag, 2003.

[21] S.J. Qin, T. Badgwell, A survey of industrial model predictive control technology, Control Engineering Practice 11 (2003) 733–764.

[22] J.B. Rawlings, Tutorial overview of model predictive control, IEEE Control Systems Magazine (2000) 38–52.

[23] T. Binder, L. Blank, H.G. Bock, R. Bulirsch, W. Dahmen, M. Diehl, T. Kronseder, W. Marquardt, J.P. Schlöder, O. von Stryk, Introduction to model based optimization of chemical processes on moving horizons, in: M. Grötschel, S.O. Krumke, J. Rambau (Eds.), Online Optimization of Large Scale Systems: State of the Art, Springer, 2001, pp. 295–340.

[24] O. von Stryk, R. Bulirsch, Direct and indirect methods for trajectory optimization, Annals of Operations Research 37 (1992) 357–373.

[25] L.T. Biegler, Efficient solution of dynamic optimization and NMPC problems, in: F. Allgöwer, A. Zheng (Eds.), Nonlinear Model Predictive Control, Birkhäuser, 2000, pp. 219–245.

[26] M.B. Milam, K. Mushambi, R.M. Murray, A new computational approach to real-time trajectory generation for constrained mechanical systems, in: IEEE Conference on Decision and Control, 2000.

[27] P.E. Gill, W. Murray, M.A. Saunders, M.A. Wright, User's guide for NPSOL 5.0: a Fortran package for nonlinear programming, Systems Optimization Laboratory, Stanford University, Stanford, CA 94305, 1998.

[28] L. Petzold, J.B. Rosen, P.E. Gill, L.O. Jay, K. Park, Numerical optimal control of parabolic PDEs using DASOPT, in: M. Grötschel, S.O. Krumke, J. Rambau (Eds.), Large Scale Optimization with Applications, Part II, vol. 93, Springer, 1997, pp. 271–300.

[29] F. Allgöwer, T.A. Badgwell, J.S. Qin, J.B. Rawlings, S.J. Wright, Nonlinear predictive control and moving horizon estimation—an introductory overview, in: P.M. Frank (Ed.), Advances in Control, Highlights of ECC'99, Springer, 1999, pp. 391–449.

[30] F. Allgöwer, A. Zheng, Nonlinear predictive control, in: Progress un Systems Theory, vol. 26, Birkhäuser, 2000.

[31] D.Q. Mayne, H. Michalska, Receding horizon control of nonlinear systems, IEEE Transactions on Automatic Control 35 (7) (1990) 814–824.

[32] D.Q. Mayne, J.B. Rawlings, C.V. Rao, P.O.M. Scokaert, Constrained model predictive control: stability and optimality, Automatica 36 (2000) 789–814.

[33] A. Jadbabaie, J. Hauser, On the stability of unconstrained receding horizon control with a general terminal cost, in: Proceedings of the 40th IEEE Conference on Decision and Control, 2001.

[34] M. Diehl, H.G. Bock, J.P. Schlöder, R. Findeisen, Z. Nagy, F. Allgöwer, Real-time optimization and nonlinear model predictive control of processes governed by differential–algebraic equations, Journal of Process Control 12 (2002) 577–585.

[35] N.H. El-Farra, P.D. Christofides, Coordinating feedback and switching for control of spatially distributed processes, Computers and Chemical Engineering 28 (2004) 111–128.

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2005; **15**:769–784
Published online 22 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/rnc.1023

Toward controlling dielectrophoresis

Dong Eui Chang*,† and Nicolas Petit

*Centre Automatique et Systèmes, Ecole Nationale Superieure des Mines de Paris 60,
bd Saint-Michel, 75272 Paris, Cedex 06, France*

SUMMARY

Dielectrophoresis is the motion of a particle due to the interaction between a non-uniform electric field and its induced dipole moment in the particle. With the advent of the fabrication technology at micro/nanoscale, dielectrophoresis is actively applied in manipulating, separating, and characterizing micro/nano-sized particles such as DNA, cells, proteins, nanotubes and nanoparticles. In this paper we introduce control engineers to dielectrophoresis by suggesting several possible research topics and performing a case study: a time-optimal control of a dielectrophoretic system with a state constraint. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: dielectrophoresis; nanotechnology; biotechnology; systems theory; time-optimal control

1. INTRODUCTION

Dielectrophoresis (DEP) refers to the motion of a particle due to the force exerted on the induced dipole moment of the particle by a non-uniform electric field. The study of dielectrophoresis and its application to the manipulation of small and biological particles was first thoroughly investigated out by Pohl [1]. At that time there was a limit to the magnitude of electric fields that could be generated with small voltages. With the advent of MEMS and nanotechnology, one can now generate a large electric field with weak voltages so that dielectrophoresis may be actively applied to manipulating, separating, and characterizing micro/nano-sized particles such as cells, DNA, proteins, nanotubes, and nanoparticles [2–7]. An advantage of dielectrophoresis over electrophoresis is that it can also work on neutrally charged particles [8].

The objective of this paper is to turn the attention of control engineers to this area of dielectrophoresis so that they can not only find many interesting control problems but also contribute to DEP-based engineering applications. The connection between control theory and dielectrophoresis is not new. The interpretation of a simple model of the induced dipole moment due to an electric field *as a control system* was briefly mentioned by Daniel [9] in 1967. To our

*Correspondence to: D. E. Chang, Centre Automatique et Systèmes, Ecole Nationale Superieure des Mines de Paris 60, bd Saint-Michel, 75272 Paris, Cedex 06, France.
†E-mail: dchang@cas.ensmp.fr

Accepted 22 July 2005

knowledge, the first attempt to apply control theory to a DEP problem was made by Kaler *et al.* [10, 11] for the purpose of locally stabilizing levitation of biological particle with dielectrophoresis. Their main approach was as follows. They linearized the original nonlinear dynamics around an equilibrium of interest, applied a sinusoidal voltage on electrodes at an appropriate frequency, (naively) averaged the resultant equations over the period of the sinusoidal voltage so that the equations become time-invariant, and then finally modulate the amplitude of the boundary voltage, which was initially assumed constant, with a linear feedback controller. Their clever but *ad hoc* procedure proved effective experimentally [10, 11]. Since their work has been unknown to much of the control community, the approach has neither been formalized nor improved by modern control theory. Only recently, the issue of applying control technology to DEP applications was raised in Reference [12]. Hence, it is time for the control theory, which has advanced for the last 40 years, to make contributions to this area. For an overview of various control issues in other (non-DEP-related) nano-scale systems, we refer to the report available on the web page [13].

This paper is organized as follows. First, we explain the physics of dielectrophoresis and review a traditional method of manipulating particles with dielectrophoresis in Section 2. Second, we provide several possible research topics for control engineers: the system identification, the boundary control of DEP systems governed by PDEs, effect of a term quadratic in a control variable, and higher-order hidden dynamics. Third, we perform a case study: a time-optimal control of a DEP system with a state constraint that arises from the existence of electrodes. According to Chang *et al.* [14], all the time-optimal trajectories in the system *without* the state constraint begin with undershoots. Hence, one needs to do the time-optimal study *with* the state constraint to prevent particles near the electrodes from trying to go through electrodes. Finally, we conclude in Section 5.

2. BACKGROUND MATERIALS

2.1. *Physics of dielectrophoresis*

We briefly explain basic physics of dielectrophoresis; see References [1, 8, 15] for more details. When a particle is immersed in a medium and an electric field, $\mathbf{E}(\mathbf{x}, t)$, is imposed, then an effective dipole moment, $\mathbf{m}(\mathbf{x}, t)$, is induced in the particle, where $\mathbf{x} \in \mathbb{R}^3$ is the position vector and t is the time; see Figure 1(a). The relation between $\mathbf{E}$ and $\mathbf{m}$ is linear and given by

$$\mathbf{m}(\mathbf{x}, t) = g(t) * \mathbf{E}(\mathbf{x}, t) \tag{1}$$

where $*$ denotes time convolution. The Laplace transform $G(s)$ of $g(t)$ is called the Clausius-Mossotti function (up to a constant) [1, 8, 15] and it depends on the physical structure and electric properties of the particle and the electric properties of the medium in which the particle is immersed. When the particle is a sphere, $G(s)$ is rational, generically of relative degree 0, where the degree of the denominator (or numerator) is the number of layers in the particle; see Figure 1(b) and Appendix C of Reference [8]. For example, when the particle is spherical and homogeneous, $G(s)$ is given by

$$G(s) = a + \frac{b}{s + c} \tag{2}$$

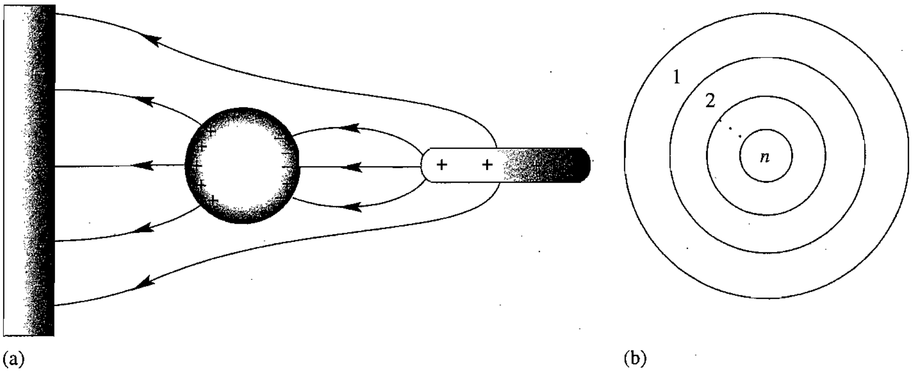


Figure 1. (a) An electric field redistributes the charges in the particle so that a dipole moment is induced in the particle; and (b) a multi-layer shell model of a spherical dielectric particle. The Clausius–Mossotti (transfer) function of the n-layered spherical particle in a medium is a rational function of relative degree 0 where the degree of the denominator is n.

with

$$a = 4\pi r^3 \varepsilon_m \frac{\varepsilon_p - \varepsilon_m}{\varepsilon_p + 2\varepsilon_m}$$

$$b = a\left(\frac{\sigma_p - \sigma_m}{\varepsilon_p - \varepsilon_m} - \frac{\sigma_p + 2\sigma_m}{\varepsilon_p + 2\varepsilon_m}\right) \tag{3}$$

$$c = \frac{\sigma_p + 2\sigma_m}{\varepsilon_p + 2\varepsilon_m} \tag{4}$$

where r is the radius of the particle, ε_p (resp., ε_m) is the permittivity of the particle (resp., medium) and σ_p (resp., σ_m) is the conductivity of the particle (resp., medium). The frequency dependence of the Clausius–Mossotti function $G(s)$ is at the heart of DEP applications since most methods of separating particles with DEP make use of the fact that different types of particles have different frequency dependences [4, 5, 8].

The dielectrophoretic force, $\mathbf{F}_{\text{dep}}$, due to the interaction between the induced dipole moment $\mathbf{m}$ and the electric field $\mathbf{E}$ is given by

$$\mathbf{F}_{\text{dep}}(\mathbf{x}, t) = (\mathbf{m}(\mathbf{x}, t) \cdot \nabla)\mathbf{E}(\mathbf{x}, t) \tag{5}$$

The dielectrophoretic torque, τ_{dep}, is given by

$$\tau_{\text{dep}}(\mathbf{x}, t) = \mathbf{m}(\mathbf{x}, t) \times \mathbf{E}(\mathbf{x}, t) \tag{6}$$

In applications of dielectrophoresis there are electrodes that govern the boundary voltage, which induces the electric field $\mathbf{E}(\mathbf{x}, t)$, so the boundary voltage plays the role of control.

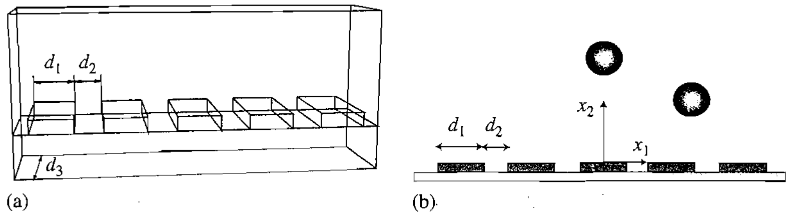


Figure 2. (a) The arrangement of a linear electrode array; and (b) the front view of the arrangement of a linear electrode array where the origin of the x_1x_2 axis lies in the middle of an electrode.

2.2. *Parallel array of linear electrodes*

We consider the configuration with a parallel array of linear electrodes in Figure 2(a). This configuration is often used to separate one type of bioparticles from the rest in the mixture of particles, or to sort bioparticles [3, 5, 16]. As particles are usually relatively small compared with the electrodes, one may assume that each electrode is infinitely long (or, $d_3 \gg d_1, d_2$) and that there are infinite number of them. Then, the problem reduces to a planar case as in Figure 2(b). On electrodes we give the following boundary value of the potential function (or, the voltage):

$$V_{\rm bd}(x_1, t) = u_l(t), \quad x_1 \in [l(d_1 + d_2) - d_1/2, l(d_1 + d_2) + d_1/2]$$

with $u_l(t) \in \mathbb{R}, l \in \mathbb{Z}$. We choose $u_l(t)$'s such that the function $V_{\rm bd}$ is even and periodic in x of period $N(d_1 + d_2)$ with a fixed $N \in \mathbb{N}$, i.e.

$$V_{\rm bd}(x_1, t) = V_{\rm bd}(-x_1, t) = V_{\rm bd}(x_1 + N(d_1 + d_2), t)$$

It is practical to assume that the boundary value of the potential function between electrodes changes linearly as follows:

$$V_{\rm bd}(x_1, t) = \frac{u_{l+1}(t) - u_l(t)}{d_2}\left(x_1 - l(d_1 + d_2) - \frac{d_1}{2}\right) + u_l(t)$$

$$x_1 \in [l(d_1 + d_2) + d_1/2, (l + 1)(d_1 + d_2) - d_1/2]$$

with $l \in \mathbb{Z}$. This assumption is acceptable when the gap between electrodes is small (see 10.3.2 of Reference [5] and references therein).

The potential function $V(x_1, x_2, t)$ in the region $x_2 > 0$ is derived by solving the Dirichlet problem, $\nabla^2 V = 0$ with the boundary condition given above. As the boundary value of the voltage is a linear combination of u_l's, the potential function $V(x_1, x_2, t)$ is also a linear combination of u_l's. It can be written as follows:

$$V(x_1, x_2, t) = \sum_{l=1}^{N} u_l(t) V_l(x_1, x_2)$$

Hence, the electric field $\mathbf{E}(x_1, x_2, t)$ is also a linear combination of $u_l(t)$:

$$\mathbf{E}(x_1, x_2, t) = \sum_{l=1}^{N} u_l(t) \mathbf{E}_l(x_1, x_2)$$

with $\mathbf{E}_l = -\nabla V_l$. By (1)–(6), the induced dipole moment, the dielectrophoretic force and the dielectrophoretic torque on a particle with the Clausius–Mossotti function $G(s)$, are, respectively, given by

$$\mathbf{m}(x_1, x_2, t) = \sum_{l=1}^{N} (g * u_l)(t)\mathbf{E}_l(x_1, x_2) \tag{7}$$

$$\mathbf{F}_{\text{dep}}(x_1, x_2, t) = \sum_{l=1}^{N} \sum_{m=1}^{N} (g * u_l)(t)u_m(t)(\mathbf{E}_l(x_1, x_2) \cdot \nabla)\mathbf{E}_m(x_1, x_2) \tag{8}$$

$$\tau_{\text{dep}}(x_1, x_2, t) = \sum_{l=1}^{N} \sum_{m=1}^{N} (g * u_l)(t)u_m(t)\mathbf{E}_l(x_1, x_2) \times \mathbf{E}_m(x_1, x_2) \tag{9}$$

where $g(t)$ is the inverse Laplace transform of the Clausius–Mossotti function $G(s)$.

2.3. *Traditional methods of manipulating particles*

In the current application area of dielectrophoresis, sinusoidal signals are often used for the boundary potential to manipulate/separate particles [16, 17]. Sinusoidal signals have a couple of advantages in that they are not only easy to generate but also make use of the *linear* relation between the induced dipole moment and the electric field in (1).

We consider the case of controlling particles with a travelling wave array from References [17, 18]. Notice the four-phase travelling wave electrode array in Figure 2 with the boundary potentials, $\phi_1(x_1, 0)$ and $\phi_2(x_1, 0)$ in Figure 3, where we assume that the potentials change linearly between neighbouring electrodes on the boundary [5, 17]. The potential on the boundary is time-modulated as follows:

$$\phi(x_1, 0, t) = \phi_1(x_1, 0)\cos(\omega t) + \phi_2(x_1, 0)\sin(\omega t)$$

One computes the corresponding dielectrophoretic force by (5) or (8), and then takes the *naive* average of it over the period $2\pi/\omega$ (when one wants to justify the use of the averaging method,

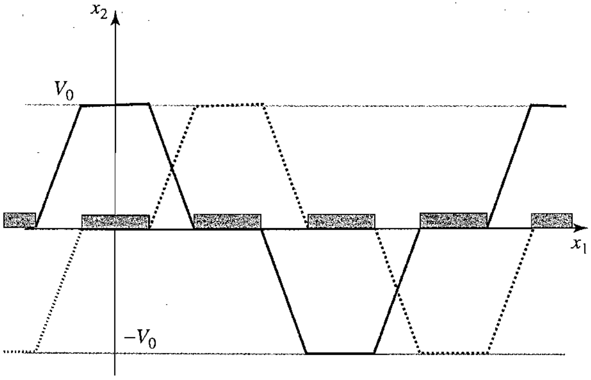


Figure 3. The boundary condition for the potential at $x_2 = 0$ for the travelling wave electric field. The potential $\phi_2(x_1, 0)$ (···) is a one-phase shift of the potential, $\phi_1(x_1, 0)$ (—).

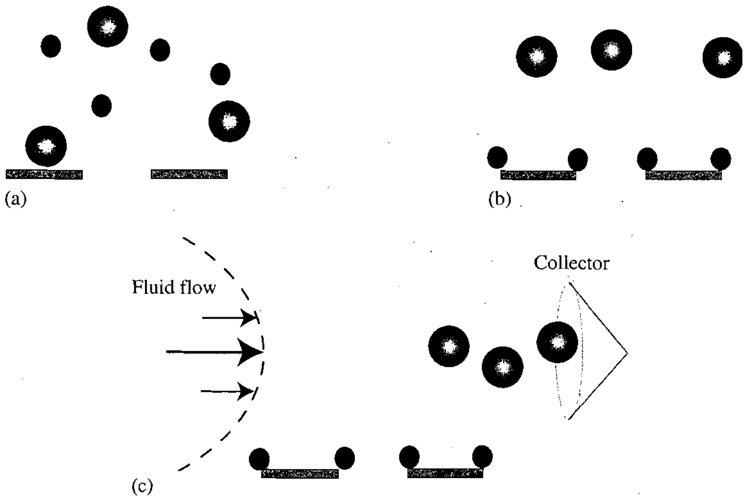


Figure 4. Separation of particles with dielectrophoresis and field flow fractionation: (a) a mixture of two different types of particles before the DEP force is given; (b) the vertical separation is achieved with the DEP force; and (c) fluid flow collects one type of particles while the other type stays attracted to the electrodes.

the dynamics needs to be transformed to a standard form [19], but this procedure is missing in this traditional approach). The averaged dielectrophoretic force $\langle \mathbf{F}_{\mathrm{dep}} \rangle$ is of the form

$$\langle \mathbf{F}_{\mathrm{dep}} \rangle(x_1, x_2) = \mathrm{Re}[G(\mathrm{j}\omega)]F_c(x_1, x_2) + \mathrm{Im}[G(\mathrm{j}\omega)]F_s(x_1, x_2) \tag{10}$$

which can be checked in References [17, 18] for more details. In general, the term $\mathrm{Re}[G(\mathrm{j}\omega)]F_c(x_1, x_2)$ in (10) creates a vertical force and the term $\mathrm{Im}[G(\mathrm{j}\omega)]F_s(x_1, x_2)$ creates a horizontal force [4, 19]. Consider a mixture of two different types of particles immersed in a fluid medium in Figure 2. Each type will have different Clausius–Mossotti functions $G(s)$. By choosing an appropriate frequency ω, one can separate these two kinds of particles. One can also employ additional fluid flow to move particles horizontally instead of using the term $\mathrm{Im}[G(\mathrm{j}\omega)]F_s(x_1, x_2)$. This method with fluid flow is called the *field flow fractionation* [4, 5]; see Figure 4. This traditional method works well experimentally. Its formalization and improvement are left for control engineers.

3. RESEARCH DIRECTIONS FOR CONTROL ENGINEERS

We now propose several possible research topics for control engineers in the field of dielectrophoretic systems. This section is inspired by Jones [8], a standard reference in dielectrophoresis.

First, we consider a simple system which has all the key features of dielectrophoretic systems. The configuration is given in Figure 2 with the boundary voltage as in Figure 5. Notice that we here consider the exact (not approximate) boundary condition, $\partial\phi/\partial n = 0$ between electrodes.

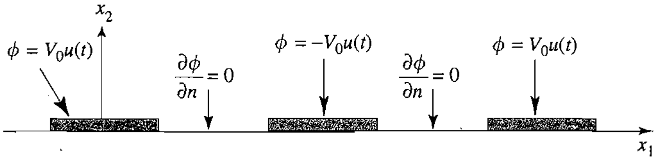

x_2
$\phi = V_0u(t)$
$\frac{\partial\phi}{\partial n} = 0$
$\phi = -V_0u(t)$
$\frac{\partial\phi}{\partial n} = 0$
$\phi = V_0u(t)$
x_1

Figure 5. The exact boundary condition for the potential at $y = 0$ for the standing wave electric field. The normal derivative of the potential is zero on the boundary between neighbouring electrodes.

We impose the boundary potential $V_0u(t)$ on every other electrode and $(-V_0u(t))$ on the others where $u(t)$ is the control. There is a neutrally charged spherical particle in a fluid medium in the chamber. We assume that the particle is homogenous such that the Clausius–Mossotti functions $G(s)$ is given by (2) with $a\neq 0$. The dielectrophoretic force $\mathbf{F}_{\text{dep}}(x, y, t)$ is of the form:

$$\mathbf{F}_{\text{dep}}(x_1, x_2, t) = u(t)(g * u)(t)(\mathbf{E}(x_1, x_2)\cdot\nabla)\mathbf{E}(x_1, x_2) \tag{11}$$

where $u(t)\mathbf{E}(x_1, x_2)$ is the electric field in $x_2 > 0$; see Reference [18] to verify this. Function $g(t)$ in (11) is the inverse Laplace transform of $G(s)$ in (2) and given by

$$g(t) = a\delta(t) + be^{-ct}$$

where $\delta(t)$ is the Dirac delta function. Thus, the DEP force in (11) can be written as

$$\mathbf{F}_{\text{dep}}(x_1, x_2, t) = (au(t)^2 + by(t)u(t))(\mathbf{E}(x_1, x_2)\cdot\nabla)\mathbf{E}(x_1, x_2)$$

where the new variable y satisfies

$$\dot{y} + cy = u \tag{12}$$

To simplify the dynamics, we will make two assumptions. First, we assume that the particle and its surrounding fluid are such that the Reynolds number is low. Under this assumption the drag force on a sphere is linear in velocity by the Stokes law; see Section 3.8.1 of Reference [5]. Second, we assume that the term $m\ddot{x}$ is relatively small compared with other forces, which is reasonable as the particle is small and light. For large or heavy particles this assumption fails and the dynamics cannot be simplified. The resulting study requires further investigations which fall out of the scope of this paper. We refer to References [5,20] to help understand these two assumptions. Then, the only remaining forces acting on the particle are the drag, the gravitational plus buoyant force and the dielectrophoretic force. The equation of motion is of the form

$$\underbrace{m\begin{bmatrix}\ddot{x}_1\\ \ddot{x}_2\end{bmatrix}}_{\text{inert.}=0}+\underbrace{k\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}}_{\text{drag}}+\underbrace{w\begin{bmatrix}0\\ x_2\end{bmatrix}}_{\text{grav.+buoy.}}+\underbrace{(au^2+byu)\begin{bmatrix}H_1(x_1,x_2)\\ H_2(x_1,x_2)\end{bmatrix}}_{\text{DEP}}=0 \tag{13}$$

with $(\mathbf{E}\cdot\nabla)\mathbf{E} = (H_1, H_2)$. Let

$$\mathbf{x} = [x_1, x_2, y]^{\mathrm{T}}$$

Then, the dynamics can be written in the form familiar to control engineers as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = X_0(\mathbf{x}) + uX_1(\mathbf{x}) + u^2X_2(\mathbf{x}) \tag{14}$$

We now discuss features of the dielectrophoretic systems and suggest possible future research directions to control engineers.

1. *Quadratic in control.* Notice that the system in (14) is not an affine control system. There exists a term which is quadratic in control u. In general, the term quadratic in control comes from the fact that the Clausius–Mossotti function $G(s)$ is a rational function of *relative degree* 0. The existence of this quadratic term makes dielectrophoretic systems challenging from a control point of view because it does not allow both plus and minus signs of the term.
2. *Bounded control.* The control u is always bounded in its magnitude because it is a voltage—divided by V_0, precisely speaking—on electrodes.
3. *Boundary control.* One can also view DEP systems from the viewpoint of the partial differential equations (PDEs). The PDE involved here is the Laplace equation in computing the potential function from the boundary value where the boundary value is regarded as control. When there are large number of particles, one can also employ a density function to describe their overall movement. For example, the Fokker–Planck equation is used with a periodic potential to separate particles in Reference [21]; see also Section 8.4 of Reference [5] and references therein. Hence, PDE control theory will be useful for this direction of research.
4. *System identification.* Different types of biological cells or small particles have different physical/electrical characteristics such as the number of layers in the shell model in Figure 1(b), the permittivity, and the conductivity of the particle. Namely, one needs to model the Clausius–Mossotti function $G(s)$; see Appendix C of Reference [8]. Existing measurement techniques of $G(s)$ in application have not fully taken advantage of system identification theory. In Appendix E of Reference [8], one can see that there have been some approximate methods in identifying $G(s)$. For example, they assume that all the poles of $G(s)$ are simple and sufficiently distant from one another so that they can reduce the system identification problem to the case of a single pole. In addition, the Argand diagram used in Appendix E of Reference [8] is, in principle, the same as the Nyquist plot in control theory. One can see that the system identification technology in control theory will contribute a lot to the study of nanoparticles and bioparticles [7, 10, 11]. In particular, the work in References [10, 11] is noteworthy because the concept of feedback control through the boundary value was employed to identify the Clausius–Mossotti function of a given bioparticle. In the same references, one can learn that the dielectrophoretic levitation of particles is closely linked with system identification problems in the sense that one needs to levitate and trap a particle to measure its electrical/physical properties; see References [8, Section 3.4, 22] and references therein.
5. *Interaction among particles or between particles and electrodes.* When particles are close to each other, sometimes the interaction between them is no longer negligible. This interaction creates clustering or chaining of particles; see Chapters 6, 7 of Reference [8]. Likewise, when a particle is close to an electrode, a new chemical force starts to appear due to the interaction between the charges on the electrode and the induced dipole moment in

the particle. In these cases, one needs to modify the dynamics taking into account these interaction effects; see Chapter 3 of Reference [5], in particular, Section 3.4.

6. *Multipoles*. Non-uniform electric fields induce not only dipole moments but also multipole moments in a particle; see Chapter 4 of Reference [15] or Appendix B of Reference [8]. One can add this into the dynamics for a more precise model or employ a robust control technique, regarding this higher-order effect as uncertainty.

This list is far from exhaustive. We believe that control technology, which has advanced for the last forty years, will make many contributions to the applications of dielectrophoresis.

4. A CASE STUDY: A TIME-OPTIMAL CONTROL PROBLEM

We now consider a time-optimal control problem of a dielectrophoretic system because time-optimal control is one of the useful and challenging optimal control problems. Ideally, time-optimal control will reduce the process time in manipulating particles in labs-on-a-chip systems. As an initial step, we will deal with a simple case of (14). For this case the time-optimal control problem was studied in Reference [14] *without* the state constraint which comes from the fact that particles cannot go through electrodes. In Reference [14], it was discovered that due to the existence of the quadratic term u^2 in (14), optimal trajectories without the state constraint always start with an undershoot. Because of this phenomenon, it is necessary to consider the state constraint because the particle starting close to electrodes and following the time-optimal trajectory, which is derived without the state constraint considered, will violate the state constraint. We hope that this case study will provide a good example of exchanging problems and solutions between control theory and engineering application.

4.1. Derivation of equations of motion

We derive the dynamics for which we will investigate the time-optimal control. First, recall the equation of motion in (12) and (13). From Reference [18] one can check that $H_1(0, x_2) = 0$ in (13). Hence, the x_2-axis is an invariant set of the dynamics. As the vertical motion of particles in the whole chamber can be practically represented by that of particles on the x_2-axis, we will restrict ourselves to this invariant line. Let us assume that the particle is neutrally buoyant, so that the coefficient $w = 0$ in (13). Then, the dynamics of (x_2, y) on the x_2-axis can be written as

$$k\dot{x}_2 + (byu + au^2)H_2(0, x_2) = 0 \tag{15}$$

$$\dot{y} + cy = u \tag{16}$$

where one can verify that $H_2(0, x_2)$ satisfies $H_2(0, x_2) < 0$ for $x_2 > 0, H_2(0, 0) = 0$ and $\lim_{x_2 \to \infty} H_2(0, x_2) = 0$; see Reference [18] to verify this. Let

$$x = \int_{\varepsilon}^{x_2} \frac{-k}{bH_2(0, x_2)} \mathrm{d}x_2 \tag{17}$$

for $x_2 \geqslant \varepsilon$ where ε is a positive number. If a particle is close to the electrode, then additional forces other than the DEP force start to appear in the dynamics (for example the Stern layer effect; see Section 3.4 of Reference [5]), so the parameter ε in (17) defines the region $\{x_2 \geqslant \varepsilon\}$

where the dynamics (15) is valid. As a function of x_2, $x(x_2)$ is strictly monotone on $\{x_2 \geqslant \varepsilon\}$ since $x'(x_2) = -k/(bH_2(0, x_2))$ is sign definite on $\{x_2 \geqslant \varepsilon\}$. Hence, we can use x as a new co-ordinate in place of x_2. This new co-ordinate not only simplifies the dynamics but also makes the dynamics independent of the physical size of electrodes (such as d_1 and d_2 in Figure 2) and the maximum value of the boundary voltage, V_0. In the state (x, y), the equations in (15) and (16) are written as

$$\dot{x} = yu + \alpha u^2 \tag{18}$$

$$\dot{y} = -cy + u \tag{19}$$

where

$$\alpha = a/b \tag{20}$$

We consider the following conditions:

$$x(0) = \text{specified}, \quad y(0) = 0 \tag{21}$$

$$x(t_f) = \text{specified}, \quad y(t_f) = \text{free} \tag{22}$$

$$|u| \leqslant 1 \tag{23}$$

and

$$\alpha < 0, \quad c > 0 \tag{24}$$

Initially the induced dipole is zero, so we have $y(0) = 0$. Since we are only interested in the position of the particle and not interested in the final state of the induced dipole, we have $y(t_f) = \text{free}$. Because the available voltage has a magnitude limit, we require $|u| \leqslant 1$. The condition $c > 0$ comes from (4), but the condition $\alpha < 0$ is arbitrary. The case of $\alpha > 0$ can be handled similarly. When $\alpha = 0$, then the system becomes an affine system, which is relatively easy to deal with. Because $b \neq 0$ generically in (3), the coefficient α in (20) is generically well-defined. Notice in (17) that depending on the sign of b the original region $\{x_2 \geqslant 0\}$ is mapped to

$$\{x \geqslant 0\} \quad \text{or} \quad \{x \leqslant 0\} \tag{25}$$

This gives a state constraint to the dynamics in (18) and (19). Equations (18) and (19) with (21)–(24) and a state constraint (25) are our final dynamics.

4.2. Statement of the time-optimal problem

We address the following time-optimal control problem:

> Consider the system (18) and (19) with conditions (21)–(24). Find a time-optimal trajectory with the state constraint $x \geqslant 0$ (or $x \leqslant 0$).

The same time-optimal control problem *without the state constraint* was fully and analytically studied in Reference [14], summarized as follows. First, when $x(t_f) < x(0)$, there are no time-optimal trajectories even though all $x(t_f)$ ($< x(0)$) are reachable. Second, for $x(t_f) > x(0)$, time-optimal trajectories exist if and only if the parameters satisfy $(1 + \alpha c) > 0$. Moreover, when $(1 + 2\alpha c) > 0$, the existence and uniqueness of time-optimal trajectories were proved, and the formula of optimal control was constructed. However, in the case of $(1 + 2\alpha c) \leqslant 0$, only existence was shown. Instead of uniqueness, a finite algorithm for finding optimal trajectories was provided. Irrespective of the sign of $(1 + 2\alpha c)$, a feature of all time-optimal trajectories is that there is an initial undershoot in $x(t)$. One can guess this from (18), (19) and $y(0) = 0$ in (21). Because of this initial undershoot, when the initial position $x(0)$ of x is close to $x = 0$, the time-optimal trajectory without the state constraint violates the state constraint $x \geqslant 0$. This phenomenon leads us to study the time-optimal control of the same system *with the state constraint*.

4.3. Numerical algorithm to construct optimal trajectories

We make a numerical study of the time-optimal control problem given in Section 4.2. For convenience, we will only consider the case of the state constraint $x \geqslant 0$.

Let us introduce a time-scaling

$$s = t/T$$

for some $T > 0$. Let $(z_1(s), z_2(s)) = (x(sT), y(sT))$. We use $'$ to denote the derivative with respect to s. Let us first reformulate the time-optimal problem such that the control variable u disappears and the time interval is normalized to $[0, 1]$. The new idea of removing the control variable was effectively employed in the software package called nonlinear trajectory generation (NTG) to solve optimal control problems; see References [23, 24]. Along these lines, the time-optimal control problem is given by

$$\left\{ \begin{array}{l} \min\limits_{(z_1, z_2) \in \mathbb{R}^2} T \\ \text{subject to} \\ \quad F(z_1, z_2, z_1'/T, z_2'/T) = 0 \\ \quad |z_2'/T + cz_2| \leqslant 1 \\ \quad z_1 \geqslant 0 \\ \quad z_1(0), z_1(1) = \text{specified} \\ \quad z_2(0) = 0 \\ \quad T \geqslant \varepsilon. \end{array} \right. \tag{26}$$

for a small $\varepsilon > 0$ where

$$F(x, y, \dot{x}, \dot{y}) = \dot{x} - y(\dot{y} + cy) - \alpha(\dot{y} + cy)^2 \tag{27}$$

In (26), ε can be chosen to be any sufficiently small positive number that ensures T is positive. Notice that $(x(t), y(t), \dot{x}(t), \dot{y}(t))$ is replaced by $(z_1(s), z_2(s), z_1'(s)/T, z_2'(s)/T)$, which normalizes the time interval to $[0, 1]$. The constraint $F = 0$ in (26) comes from the substitution of $u = \dot{y} + cy$ in (19) to (18).

We now approximate this (continuous-time) optimal control problem by a (discrete) nonlinear dynamic programming. First, we represent (z_1, z_2) with B-splines as follows:

$$z_1(s; q) = \sum_{i=1}^{p_1} B_{i,k_1}(s) C_i^1, \quad z_2(s; q) = \sum_{i=1}^{p_2} B_{i,k_2}(s) C_i^2$$

with

$$\begin{aligned} p_j &= l_j(k_j - m_j) + m_j \\ q &= (C_1^1, \ldots, C_{p_1}^1, C_1^2, \ldots, C_{p_2}^2) \in \mathbb{R}^{p_1+p_2} \end{aligned} \tag{28}$$

where $\{B_{i,k_j}(t),\ i = 1,2\}$ is the B-spline basis function defined in Reference [25] for z_i with order k_j, C_j^i are the coefficients of the B-spline, l_j is the number of knot intervals, and m_j is number of smoothness conditions at the knots. The curve (z_1, z_2) is thus represented by the coefficient vector q. B-splines have the advantage that it is easy to enforce continuity across knot points and to compute their derivatives.

We then discretize the time interval [0, 1] into $(N-1)$ subintervals

$$0 = s_1 < s_2 < \cdots < s_N = 1 \tag{29}$$

In general N collocation points $\{s_1, \ldots, s_N\}$ are chosen uniformly over the time interval [0, 1] for convenience although optimal knots placements may also be considered. Both dynamics and constraints will be enforced at the collocation points. The problem in (26) can be approximated by the following nonlinear programming form: subject to

$$\begin{cases} \min_{q \in \mathbb{R}^{p_1+p_2}} \quad T \\ \text{subject to} \\ \quad \left.\begin{array}{l} F(z_1(s;q), z_2(s;q), z_1'(s;q)/T, z_2'(s;q)/T) = 0 \\ |z_2'/T + cz_2| \leqslant 1 \\ z_1 \geqslant 0 \end{array}\right\} \quad \text{for every } s \in \{s_1, \ldots, s_N\} \\ \quad z_1(0;q), z_1(1;q) = \text{specified} \\ \quad z_2(0;q) = 0 \\ \quad T \geqslant \varepsilon \end{cases} \tag{30}$$

The coefficients of the B-spline basis functions can be optimized with nonlinear programming. We note that the resultant control law is sub-optimal because we allow only polynomials for (z_1, z_2) and $u(s;q) = z_2'(s;q)/T + cz_2(s;q)$. However, as any continuous function on a closed interval can be uniformly approximated by polynomials according to the Stone–Weierstrass theorem [26], we can find sub-optimal trajectories which are sufficiently close to optimal ones.

We make a remark on the non-flatness of the system in (18) and (19); see References [27, 28] for the theory of flatness. A system is flat if one can find a set of outputs (equal in number to the number of inputs) such that all states and inputs can be determined from these outputs without integration (thus, differentiation is allowed). Hence, if systems (18)–(19) were flat, we could reformulate (26) with only one function (a flat output) and represent it with B-splines in (30), which would reduce the numerical load in nonlinear programming optimization [24, 29].

However, systems (18)–(19) is not flat. It can be checked by the ruled-manifold criterion which is given in the following:

Theorem 4.1 (Martin *et al.* [28], Rouchon [30])
Assume the system $\dot{z} = f(z, u)$ is flat. The projection on the p-space of the submanifold $p = f(z, u)$, where z is considered as a parameter, is a ruled manifold for all z.

Eliminating u from the dynamics $\dot{z} = f(z, u)$, $z \in \mathbb{R}^n$ yields a set of equations $F(z, \dot{z}) = 0$ that defines a ruled manifold. In other words for all $(z, p) \in \mathbb{R}^{2n}$ such that $F(z, p) = 0$, there exists a direction $d \in \mathbb{R}^n$, $d \neq 0$ such that

$$\forall \lambda \in \mathbb{R}, \quad F(z, p + \lambda d) = 0$$

One can check that there is no such direction for systems (18)–(19), and thus our system is not flat. This non-flatness of systems (18)–(19) explains why we used both states x and y (or, z_1 and z_2) in (26) (or (30)).

4.4. Simulations

We now perform a simulation to illustrate the difference of the time-optimal problem with the state constraint ($x \geqslant 0$) and without it.

Consider the specification:

$$\alpha = -3/4, \quad c = 1; \quad x(0) = 0.1, \quad x(t_f) = 1.1$$

We choose this arbitrarily for the purpose of comparison between the time-optimal control with and without the state constraint. If one wants to use a set of real data, then one needs to recall that x in (18) is the transformed variable in (17). Also, one might need to modify (30) (or (26)) with a time-rescaling, a change of control bound, etc for the purpose of numerics.

According to Reference [14], the minimum time cost, $T_{\text{w.o.s.c.}}$ without the state constraint is

$$T_{\text{w.o.s.c.}} = 7.8117 \tag{31}$$

A plot of $(x(t), y(t), u(t))$ in this case is given in Figure 6(a). Notice that the trajectory $x(t)$ has such an undershoot that it violates the state constraint. This initial undershoot is due to the existence of the term αu^2 ($\leqslant 0$) in (18) and the initial condition $y(0) = 0$.

We then performed a numerical computation of a time-optimal control with the state constraint. We choose

$$k_i = 6, \quad m_i = 4, \quad l_i = 9, \quad N = 80$$

with $i = 1, 2$ for the B-splines parameters in (28) and (29). The computed time cost is

$$T_{\text{w.s.c.}} = 8.6482$$

The corresponding plot of $(x(t), y(t), u(t))$ is given in Figure 6(b). Notice in this case that the trajectory respects the state constraint, $x \geqslant 0$. We remark that only the control $u(t)$ was computed with (30). Then, we ran the simulation of the dynamics with this $u(t)$, so $(x(t), y(t))$ in Figure 6(b) is not the curve directly from (30) respecting the dynamics only on the N collocation points, but the real trajectory satisfying the dynamics for all t. The comparison between the two plots in Figure 6 shows the necessity of the state constraint in finding time-optimal trajectories. Let us now consider a traditional approach. We assume that the signs of a and b in (2) are given

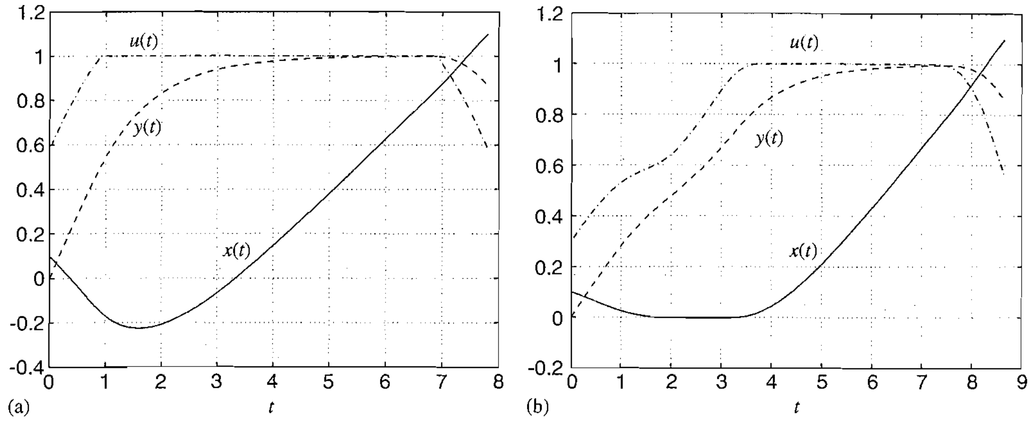


Figure 6. (a) A time-optimal trajectory $(x(t), y(t))$ corresponding to the optimal control $u(t)$ without the state constraint, $x \geqslant 0$; and (b) a time-optimal trajectory $(x(t), y(t))$ corresponding to the optimal control $u(t)$ with the state constraint, $x \geqslant 0$.

by $a<0$ and $b>0$ so that $\alpha = a/b = -0.75$ as given above. Then,

$$\mathrm{Re}[G(\mathrm{j}\omega)] = a + \frac{bc}{\omega^2 + c^2} = b\left(-0.75 + \frac{1}{\omega^2 + 1}\right)$$

As $\Delta x > 0$, one would choose ω which maximizes $\mathrm{Re}[G(\mathrm{j}\omega)]$ because the real part of $G(\mathrm{j}\omega)$ in (10) is a gain to the vertical DEP force as explained in Section 2.3. Thus, one would choose $\omega = 0$. That is, $u = 1$ or -1. In either case, simple integration yields

$$x(t) = 0.25t + \mathrm{e}^{-t} - 0.9 \tag{32}$$

One can check that this trajectory violates the state constraint $x \geqslant 0$ as the lowest point along the trajectory is $x = -0.3034$ at $t = \ln 4$ and that the time cost $T_{\text{trad.}}$ to reach $x_f = 1.1$ is

$$T_{\text{trad.}} = 7.9985 \tag{33}$$

Notice that in the traditional method it is not clear how to incorporate the state constraint into the control design procedure, but the state constraint is well treated by the time optimal control technique. We now compare the trajectory derived from the traditional method and the time-optimal trajectory without the state constraint considered. From (31) and (33) we see that the time-optimal control improves the time cost by

$$\frac{T_{\text{trad.}} - T_{\text{w.o.s.c.}}}{T_{\text{trad.}}} \times 100 = 2.335\%$$

It is interesting to notice that along the time-optimal trajectory without the state constraint considered one uses less energy ($\int |u|^2$) and the magnitude of the undershoot is smaller than along the trajectory with $u = 1$ or -1.

5. CONCLUSIONS

Since the initially significant study by Pohl [1], dielectrophoresis has been used for manipulating, separating and characterizing micro-/nano-/bio-particles. The objective of this paper is to invite control engineers to this application. After suggesting a list of future research directions for control engineers, we made a case study of the time-optimal control of a particle with dielectrophoresis. We derived the dynamics, and stated the time-optimal control problem *with* a state constraint, provided an NTG-approach nonlinear programming optimization algorithm to compute optimal trajectories and performed a simulation. The time-optimal control problem of the same system *without* the state constraint was already studied in Reference [14]. With the simulation, we compared the two cases: with or without the state constraint. The case study in Section 4 provides a good example of the synergy of engineering application and control theory. The former inspires the latter by providing new problems and the latter helps the former by providing solutions. We hope that this article stimulates control engineers so that they can enjoy the interdisciplinary research in nano/biotechnology through dielectrophoresis.

ACKNOWLEDGEMENTS

The first author would like to thank professor Igor Mezic for having introduced him to the topic of dielectrophoresis and having given him insight during his post-doctoral period at UCSB, and professor Noah Cowan at the Johns Hopkins University and professor Banavara Shashikanth at NMSU, Las Cruces for invaluable comments.

REFERENCES

1. Pohl HA. *Dielectrophoresis*. Cambridge University Press: Cambridge, 1978.
2. Green NG, Morgan H. Dielectrophoretic separation of nano-particles. *Journal of Physics D: Applied Physics* 1997; **30**(11):L41–L44.
3. Huang Y, Ewalt KL, Tirado M, Haigis TR, Forster A, Ackley D, Heller MJ, O'Connell JP. Electric manipulation of bioparticles and macromolecules on microfabricated electrodes. *Analytical Chemistry* 2001; **73**:1549–1559.
4. Hughes MP. Strategies for dielectrophoretic separation in laboratory-on-a-chip systems. *Electrophoresis* 2002; **23**(16):2569–2582.
5. Hughes MP. *Nanoelectromechanics in Engineering and Biology*. CRC Press: Boca Raton, 2002.
6. Video Library by the Dielectrophoresis Group in the Department of Molecular Pathology at the University of Texas, M.D. Anderson Cancer Center in Houston, Texas. http://www.dielectrophoresis.org/PagesMain/Video Library.htm [10 May 2005].
7. Zheng L, Brody JP, Burke PJ. Electronic manipulation of DNA, proteins, and nanoparticles for potential circuit assembly. *Biosensors and Bioelectronics* 2004; **20**(3):606–619.
8. Jones TB. *Electromechanics of Particles*. Cambridge University Press: Cambridge, 1995.
9. Daniel VV. *Dielectric Relaxation*. Academic Press: New York, 1967.
10. Kaler KVIS, Jones TB. Dielectrophoretic spectra of single cells determined by feedback-controlled levitation. *Biophysical Journal* 1990; **57**(1):173–182.
11. Kaler KVIS, Xie J-P, Jones TB, Paul R. Dual-frequency dielectrophoretic levitation of Canola protoplasts. *Biophysical Journal* 1992; **63**(1):58–69.
12. Chang DE, Loire S, Mezic I. Separation of bioparticles using the travelling wave dielectrophoresis with multiple frequencies. *Proceedings of the IEEE Conference on Decision and Control*, Hawaii, 2003.
13. Shapiro B. (ed.). Report from the NSF workshop on Control and Systems Integration of Micro- and Nano-scale Systems, 2004. http://www.isr.umd.edu/CMN-NSFwkshp/ [10 May 2005].
14. Chang DE, Petit N, Rouchon P. Time-optimal control of a particle in a dielectrophoretic system. *IFAC Congress* 2005, Prague, Czech Republic.
15. Jackson JD. *Classical Electrodynamics* (3rd edn). Wiley: New York, 1999.

16. Cui L, Holmes D, Morgan H. The dielectrophoretic levitation and separation of latex beads in microchips. *Electrophoresis* 2001; **22**:3893–3901.
17. Morgan H, Izquierdo AG, Bakewell D, Green NG, Ramos A. The dielectrophoretic and travelling wave forces generated by interdigitated electrode arrays: analytical solution using Fourier series. *Journal of Physics D: Applied Physics* 2001; **34**:1553–1561.
18. Chang DE, Loire S, Mezic I. Closed-form solutions in the electrical field analysis for dielectrophoretic and travelling wave interdigitated electrode arrays. *Journal of Physics D: Applied.Physics* 2003; **36**(23):3073–3078.
19. Guckenheimer J, Holmes P. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer: Berlin, 2000.
20. Purcell EM. Life at low Reynolds number. *American Journal of Physics* 1977; **45**:3–11.
21. Ajdari A, Prost J. Mouvement induit par un potentiel périodique de basse symétrie: diélectrophorèse pulsée. *Comptes Rendus de l'Académie Sciences, Paris* 1992; **315**(Série II):1635–1639.
22. Qian L, Scott M, Kaler KVIS, Pual R. Integrated planar concentric ring dielectrophoretic (DEP) levitator. *Journal of Electrostatics* 2002; **55**:65–79.
23. Milam MB. Real-time optimal trajectory generation for constrained dynamical systems. *Ph.D. Thesis*, California Institute of Technology.
24. Milam MB, Mushambi K, Murray RM. A new computational approach to real-time trajectory generation for constrained mechanical systems. *Proceedings of the IEEE Conference on Decision and Control*, Sydney, Australia, 2000.
25. de Boor C. *A Practical Guide to Splines*. Springer: Berlin, 1978.
26. Marsden JE, Hoffman MJ. *Elementary Classical Analysis* (2nd edn). W. H. Freeman: New York, 1993.
27. Fliess M, Lévine J, Martin P, Rouchon P. Flatness and defect of non-linear systems: introductory theory and examples. *International Journal of Control* 1995; **61**(6):1327–1360.
28. Martin P, Murray RM, Rouchon P. Flat systems. *Proceedings of the 4th European Control Conference*, Plenary Lectures and Mini-Courses, Brussels, 1997; 211–264.
29. Petit N, Milam MB, Murray RM. Inversion based trajectory optimization. *Proceedings of the IFAC Symposium on Nonlinear Control Systems Design (NOLCOS)*, 2001.
30. Rouchon P. Necessary condition and genericity of dynamics feedback linearization. *Journal of Mathematical Systems and Estimation* 1995; **5**(3):345–358.

ELSEVIER

Journal of Process Control 12 (2002) 659–665

JOURNAL OF PROCESS CONTROL

www.elsevier.com/locate/jprocont

Control of an industrial polymerization reactor using flatness

N. Petit[a,*], P. Rouchon[a], J.-M. Boueilh[b], F. Guérin[b], P. Pinvidic[c]

[a]*Centre Automatique et Systèmes, École Nationale Supèrieure des Mines de Paris, 60, boulevard Saint-Michel, 75272 Paris Cedex 06, France*
[b]*Centre Technique ATOFINA, Chemin de la Lône BP 32, 69492 Pierre Bénite Cedex. France*
[c]*APPRYL, Usine PP2, BP 21, 13117 Lavéra, France*

Received 8 November 2000; received in revised form 26 May 2001; accepted 21 June 2001

Abstract

The aim of this paper is to report the design and use of a controller for the world's largest polypropylene reactor. This is the first industrial process-controller to use the so-called flatness property of the system, which is presented here in a concise and application oriented manner. Industrial results are given and the control strategy is presented in the context of today's fast and competitive market of polymers. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Polypropylene; Continuous stirred tank reactor; Control; Flatness; Industrial application

1. Introduction

The aim of this paper is to report the design and use of a unique controller in an industrial framework. This controller is worth mentioning because it is the first application of *flatness* in industrial process control and also because the system under consideration is the largest propylene polymerization plant in the world [15].

Originally studied for mechanical systems [3,4], *flatness* exposes important issues in nonlinear control theory such as interpretations of controllability and feedback linearization [4,5,7,11]. Flatness implies a one-to-one correspondence between the trajectories of the system and those of a reduced set of variables called *flat outputs*. Many of the problems that are known to be very difficult to solve for nonlinear systems such as trajectory generation and tracking are thus transposed into a lower dimensional space, where they become straightforward. This is the methodology followed in this paper.

Two quantities are of particular interest when producing polypropylene in this plant located in Lavéra (south of France): the amount of production and the melt-index of the polymer. The melt-index indicates some of the mechanical properties of the polymer and is critical for injection and thermoforming transformations [2] (see also the http://www.appryl.fr). These two quantities depend in a nonlinear way on the amount of catalyst and hydrogen that are present in the reactor. The amount of production and the melt-index are planned with respect to economical considerations (i.e. the market of polymers). This induces frequent changes in the setpoints that must be met fast and with precision to optimize profit. This critical issue arises in different polymerization processes, see for instance [9] and [16].

Thus, the main challenge is to control the system for a wide range of setpoints with high accuracy and dynamical performance. These requirements suggest that controllers based on linear approximations of the system are unlikely to be very successful.

This system is very complex. Precise simulations models of this continuous stirred tank reactor involve thousands of variables. Yet, for control purpose, we concentrate on a reduced set of 4 differential equations and 2 nonlinear mappings, originating from balance equations and statistical studies. From a mathematical point of view, this can be seen as a two input two output model with a delay on one input. This model is both compact and rich enough to represent with accuracy the behaviour of the reactor.

The controller designed here takes into account these nonlinearities and the delay. It is capable of doing fast and precise transients, fulfilling the requirements above.

The paper is organized as follows. In Section 1 we give a model of the process. In Section 2 we recall the

[*] Corresponding author. Tel.: +33-1-40-51-93-29; fax: +33-1-40-51-91-65.
E-mail address: petit@cas.ensmp.fr (N. Petit).

PII: S0959-1524(01)00049-X

Nomenclature

Q_a	(in kg) is the amount of catalyst in the reactor
X	(dimensionless) is the rate of solid ($0 \leqslant X \leqslant 1$) (ratio between the mass of solid and the mass of solid + the mass of liquid particles)
Prod	(in kg s^{-1}) is the instantaneous amount of produced polymer
C_{H_2}	(in mol m^{-3}) is the hydrogen concentration
MI	(dimensionless) is the melt-index of the polymer in the reactor
u	(in kg s^{-1}) is the amount of catalyst coming in the reactor per unit of time
v	(in mol m^{-3} s^{-1}) is the amount of hydrogen coming in the reactor per unit of time per unit of volume
τ	(in s) is the residence time. It is a constant.

definition of the flatness property and show that this model is flat. We use this property to design control strategies. In Section 3, we present industrial results of our controller and discuss comparisons with other possible approaches. In conclusion we underline the tradeoff between the difficulty of building up a relevant nonlinear model and the simplicity of the resulting controller.

2. Modeling

The polymerization process is depicted in Fig. 1. The hydrogen enters the reactor directly while the catalyst enters the reactor after a delay due to activation preprocessing. Roughly speaking, the catalyst acts upon the amount of production, while the hydrogen acts upon the melt-index of the polymer. using the following nomenclature it is possible to write a nonlinear model.

2.1. Model

The process depicted in Fig. 1 can be represented by

$$\frac{\mathrm{d}}{\mathrm{d}t}(Q_a) = u(t-\delta) - \frac{Q_a}{\tau} \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(X) = Q_a(\alpha X + \beta) - \gamma X + \xi\frac{X}{1-X} \tag{2}$$

$$y_1 = \mathrm{Prod} = \varphi\frac{X}{1-X} \tag{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(C_{H2}) = v - g\big(C_{H_2}, Q_a\big) \tag{4}$$

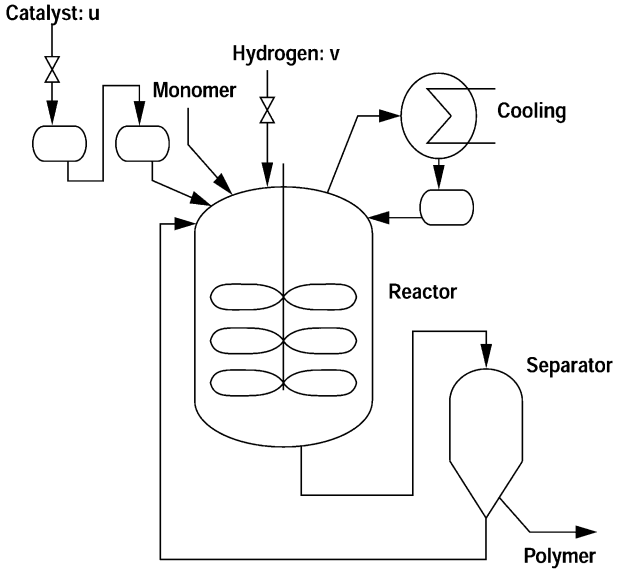


Fig. 1. The polymerization process: 2 inputs (u, v), 2 outputs (melt-index and amount of production).

$$\frac{\mathrm{d}}{\mathrm{d}t}(\log MI) = \frac{a\log C_{H_2} + b - \log MI}{\tau} \tag{5}$$

$$y_2 = MI \tag{6}$$

where a, b are constant dimensionless coefficients, α, β (both in s^{-1} kg^{-1}) and γ, ξ, (both in s^{-1}) and φ (in kg s^{-1}) are combinations of densities and other known operating parameters (omitted here for sake of clarity), δ (in s) is a constant delay. As mentioned before the effect of u is primarily on the amount of production. Still, one can clearly see the interaction of u on MI that appears through $g(C_{H_2},Q_a)$. Finally, the residence time τ is assumed to be a constant thanks to a low level regulatory loop acting upon the level of the reactor.

Eq. (1) is a dilution equation with a constant delay δ on the input. Eq. (2) is a mass balance equation. Eq. (4) is a balance equation and includes a nonlinear inference. Eq. (5) is a mixing equation where the source term arises from theoretical chemical studies of polymer growth, (see [8] for a similar study).

This model captures the essential elements of the dynamics. It is quite precise: we represent in Fig. 2 a comparison between real-time measurements of the production Prod and simulation results obtained with this model. These results were obtained for a one day period and are representative.

3. Flatness of the model

The flatness property of a (nonlinear) dynamical system $\dot{x}=f(x, u)$ with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $(n, m) \in \mathbb{N}$ is described as follows [3,4]

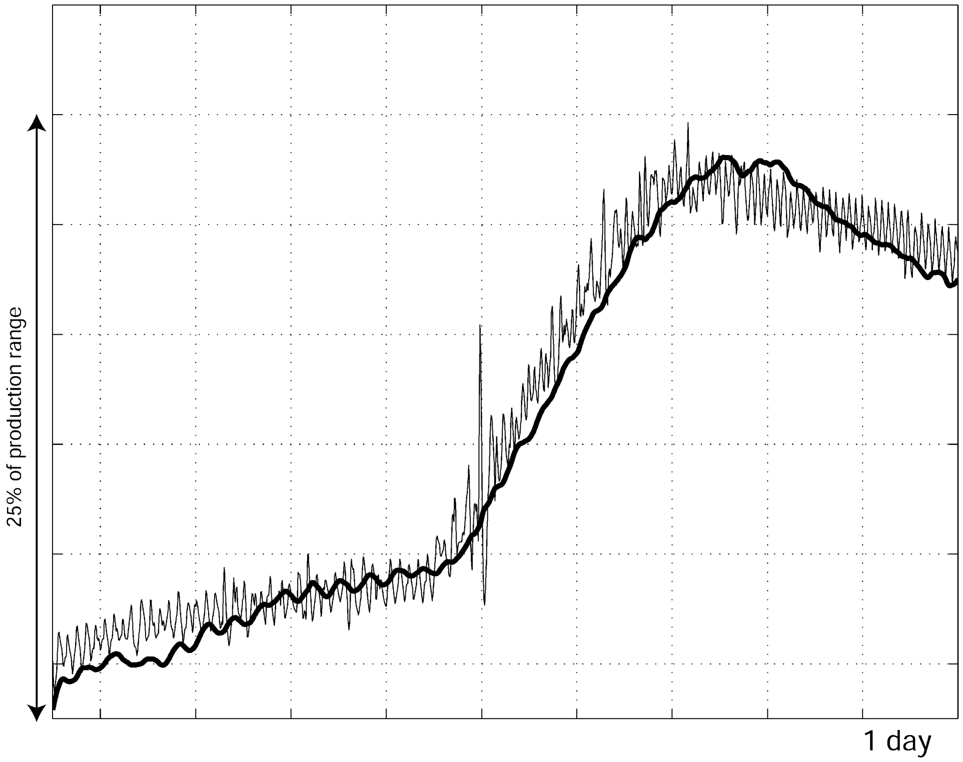


Fig. 2. Accuracy of the model. Comparison between real-time measurements of the production (Prod) and simulation results obtained with the model (1, 2, 3, 4, 5, 6). Time period = 1 day.

Definition 1. ([3,4]). The system $\dot{x} = f(x, u)$, $y = h(x)$ with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $(n, m) \in \mathbb{N}$, is flat if and only if there exists a variable z called the flat output such that

$$x = A\big(z, \dot{z}, \dots z^{(n-1)}\big)$$
$$y = B\big(z, \dot{z}, \dots z^{(n-1)}\big)$$
$$u = C\big(z, \dot{z}, \dots z^{(n)}\big)$$

where A, B and C are three mappings (depending on f and h), and $z^{(i)}$ denotes the ith derivative of the output z.[1]

In the previous definition the equations mean that there exists a quantity z that summarizes the behaviour of the whole system via the mappings A and B. The trajectories of the system, i.e. (x, u), are easily computed by the trajectories of z and its derivatives without integrating any differential equation.

To see how this property appears in our particular problem, one may write the previous equations in this form

$$\dot{x}_1 = u(t - \delta) - \frac{x_1}{\tau} \qquad (7)$$

$$\dot{x}_2 = x_1 f(x_2) + h(x_2) \qquad (8)$$

$$\dot{x}_3 = v - g(x_3, x_1) \qquad (9)$$

$$\dot{x}_4 = \frac{a\ \log(x_3) + b - x_4}{\tau} \qquad (10)$$

$$y_1 = k(x_2) \qquad (11)$$

$$y_2 = \exp(x_4) \qquad (12)$$

where f is a strictly positive function (on its interval of definition [0, 1]). It is easy to see that this system is flat:[2] all the variables are parameterized by the flat outputs $x_2 = X$, $x_4 = MI$ and their derivatives.

More precisely

$$x_3 = \exp\left(\frac{\tau\dot{x}_4 + x_4 - b}{a}\right) \qquad (13)$$

$$x_1 = \frac{\dot{x}_2 - h(x_2)}{f(x_2)} \qquad (14)$$

$$y_1 = k(x_2) \qquad (15)$$

$$y_2 = \exp(x_4) \qquad (16)$$

[1] This definition is very general. In particular the order of the flat output need not be n in the multi input multi output case.

[2] δ-Flatness is the precise definition. This notion introduced in [11], see also [12], addresses the particular case of delay systems.

and

$$u(t-\delta) = \frac{\ddot{x}_2 - \dot{x}_2 h'(x_2)}{f(x_2)} - (\dot{x}_2 - h(x_2))\frac{\dot{x}_2 f'(x_2)}{f^2(x_2)} + \frac{\dot{x}_2 - h(x_2)}{\tau f(x_2)} \tag{17}$$

$$v = \exp\left(\frac{\dot{x}_4\tau - b + x_4}{a}\right)\frac{\ddot{x}_4\tau + \dot{x}_4}{a} + g(x_4, \dot{x}_4, x_2, \dot{x}_2). \tag{18}$$

3.1. Open-loop control strategy

A general property of flat systems [3,4] is that it suffices to control the flat outputs to control the whole system. In our case once x_2 and x_4 are controlled, so are x_3, x_1, y_1, y_2 because of Eqs. (13)–(16). The open loop controls are given by Eqs. (17) and (18).

Example. We detail here an example of an open-loop control calculation. Assume that the operator wishes to increase the setpoint for the amount of production from $\text{Prod}_{\text{initial}}$ to $\text{Prod}_{\text{objective}}$ while keeping the melt-index MI_{initial} constant.

The trajectory of the amount of production is transformed to the flat outputs: x_2 must go from X_{initial} to $X_{\text{objective}}$ where

$$\text{Prod}_{\text{initial}} = \varphi\left(\frac{X_{\text{initial}}}{1 - X_{\text{initial}}}\right),$$
$$\text{Prod}_{\text{objective}} = \varphi\left(\frac{X_{\text{objective}}}{1 - X_{\text{objective}}}\right),$$

while x_4 will remain constant. A transition in finite time T between X_{initial} and $X_{\text{objective}}$ is prescribed by any function joining these two setpoints, e.g. a polynomial, denoted by $[0, T] \ni t \mapsto x_2^{\text{ref}}(t)$. Then the open-loop control is computed via (17) and (18) as

$$\begin{aligned} u^{ol}(t) = {} & \frac{\ddot{x}_2^{\text{ref}}(t+\delta) - \dot{x}_2^{\text{ref}}(t+\delta)h'\left(x_2^{\text{ref}}(t+\delta)\right)}{f\left(x_2^{\text{ref}}(t+\delta)\right)} \\ & - \left(\dot{x}_2^{\text{ref}}(t+\delta) - h\left(x_2^{\text{ref}}(t+\delta)\right)\right)\frac{\dot{x}_2^{\text{ref}}(t+\delta)f'\left(x_2^{\text{ref}}(t+\delta)\right)}{f^2\left(x_2^{\text{ref}}(t+\delta)\right)} \\ & + \frac{\dot{x}_2^{\text{ref}}(t+\delta) - h\left(x_2^{\text{ref}}(t+\delta)\right)}{\tau f\left(x_2^{\text{ref}}(t+\delta)\right)} \\ v^{ol}(t) = {} & g(\log(MI_{\text{initial}}), 0, x_2(t), \dot{x}_2(t)). \end{aligned} \tag{19}$$

In Fig. 3 one can see an example of such a calculation. Given a polynomial transition function for the flat output x_2 we compute the control u via (17). One can clearly see the effect of the advance in formula (19): the control starts increasing before one can expect the flat output to increase (with exactly a δ advance). The (input) overshoot occurs while the output is still far from the setpoints.

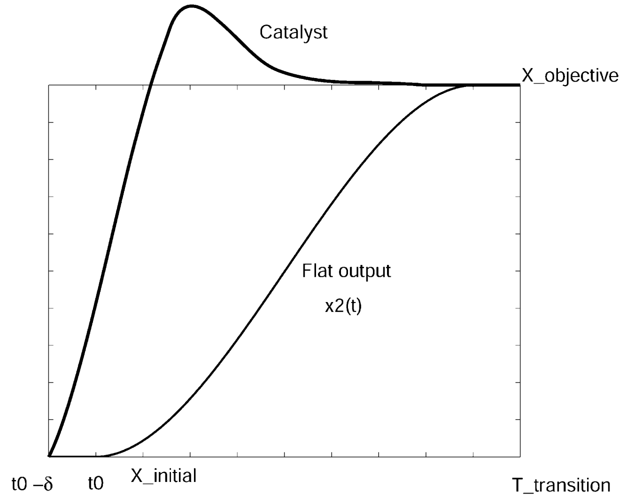


Fig. 3. Open-loop control strategy. The operator's request is expressed in terms of a transition for the flat output X and the open-loop control is computed via (17).

3.2. Closing the loop

In fact the open-loop strategy must be complemented by a feedback control law.

As mentioned before, once the flat outputs are stabilized, the whole system is stabilized because all the variables of the system are expressed in terms of the flat outputs via Eqs. (13)–(16).

The dynamics of the flat outputs are given by (17) and (18) We can stabilize them. To satisfy the following stable closed loop equations for the flat output

$$\left(\ddot{x}_2 - \ddot{x}_2^{\text{ref}}\right) = -k_1\left(\dot{x}_2 - \dot{x}_2^{\text{ref}}\right) - k_2\left(x_2 - x_2^{\text{ref}}\right)$$
$$\left(\ddot{x}_4 - \ddot{x}_4^{\text{ref}}\right) = -k_3\left(\dot{x}_4 - \dot{x}_4^{\text{ref}}\right) - k_4\left(x_4 - x_4^{\text{ref}}\right)$$

where k_1, k_2, k_3, k_4 are constants, it suffices to substitute these desired $\ddot{x}_2$ and $\ddot{x}_4$ in Eqs. (17) and (18). This gives the closed loop controller

$$u(t-\delta) = u^{ol}(t-\delta) + h_1\left(x_2 - x_2^{\text{ref}}, \dot{x}_2 - \dot{x}_2^{\text{ref}}, x_2, \dot{x}_2\right) \tag{20}$$

$$v = v^{ol} + h_2\left(x_4 - x_4^{\text{ref}}, \dot{x}_4 - \dot{x}_4^{\text{ref}}, x_4, \dot{x}_4\right). \tag{21}$$

Some required variables in Eqs. (20) and (21) are not available: x_2 is not measured and x_4 is measured at discrete times (with a delay due to the necessary laboratory analysis). To overcome this we use estimators based on classical least-squares methods, predictors, and Luenberger-style observer. These observers give naturally stable dynamics that do not interfere with the stability of the closed-loop controller. The derivatives were approximated by passive low-pass filters.

4. Industrial results

Our controller has been in full service since July 1999 and allows optimization of profit. As one can see in Fig. 4, the controller allows very fast and precise transients. On the same figure one can clearly see the effect of the delay compensation by a "time advance" in the controller design [see Eq. (19)]. Before the system meets the setpoints, the controller stops changing the value of the input (catalyst), preventing the overshoot in the production rate. These results are representative of the overall behavior of our controller.

We give industrial results for melt-index transitions on Fig. 5. The controller is capable of simultaneous transitions for the amount of production Prod and for the melt-index *MI*.

4.1. Comparisons with other techniques

Of course it is not possible to compare these results with every possible controller. Yet, we tried to tune some basic linear controllers (PI and LQR) for various simulations. Unfortunately we did not experience good results on the real plant when dealing with large changes in the setpoints and then decided to shift to another solution (the flatness approach presented here): it was difficult to combine good dynamical performances and robustness to perturbations. It is possible to sketch that with a linear controller the system may take about twice as long to converge as with the flatness controller, and that the overshoots would be very difficult to prevent without any serious deterioration of the dynamical performances. We represent in Fig. 6 a comparison between such a LQR controller acting on a simulator (using the model presented before under similar conditions, i.e. using real data for the coefficients γ, ξ and φ) and the real-time results of our controller on the plant.

Another question of relevance is: how does it compare to the well established model predictive controller (MPC) approach (see again [9] for instance)? First it should be noted that here the control objective does not really express in terms of a well-defined cost function: the main goal is to get transitions as repeatable as possible. In other terms, provided that the starting and ending setpoints are the same, a transition should always take the same time and be as accurate. Yet, as usual with flat systems, see again [3,4], the flatness Eqs, namely (13)–(18) express all the trajectories of the system. Should the control objective be expressed as a cost, it would have been possible, and computationally profitable as it has been pointed out in other applications [1,6,10,14], to solve the optimal control problem through these flatness equations. The unknown would be the shapes of the transition functions for the flat outputs $[0, T] \ni t \mapsto x_2^{\mathrm{ref}}(t)$ and $[0, T] \ni t \mapsto x_4^{\mathrm{ref}}(t)$. This would have been a flatness-based implementation of an MPC controller. But here, it seemed more suitable to take advantage of years of practice of the operators and to mimic some of their reactions: we translated these in terms of the flat output via the flatness equations and ended up with an arbitrary choice of transition functions for the flat outputs. In the end we have a controller with a predictable behavior. It should be noted also that the computational effort required of the flatness based controller is extremely light compared to an MPC optimization-based technique: here the control is computed through two analytic expressions. Besides it

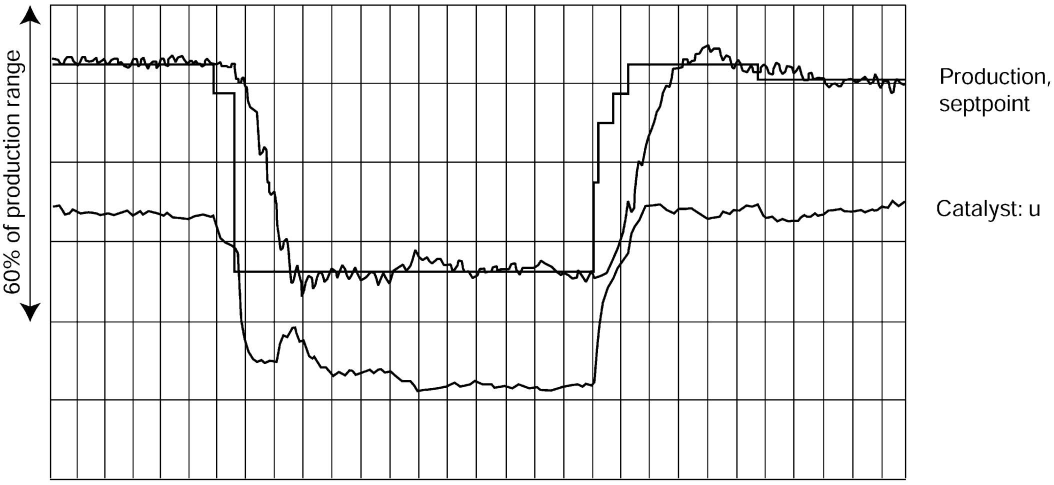


Fig. 4. Industrial results over 2 days (tests) with production (Prod) transients. The transients are fast and precise. Precise scales are omitted for confidentiality reasons.

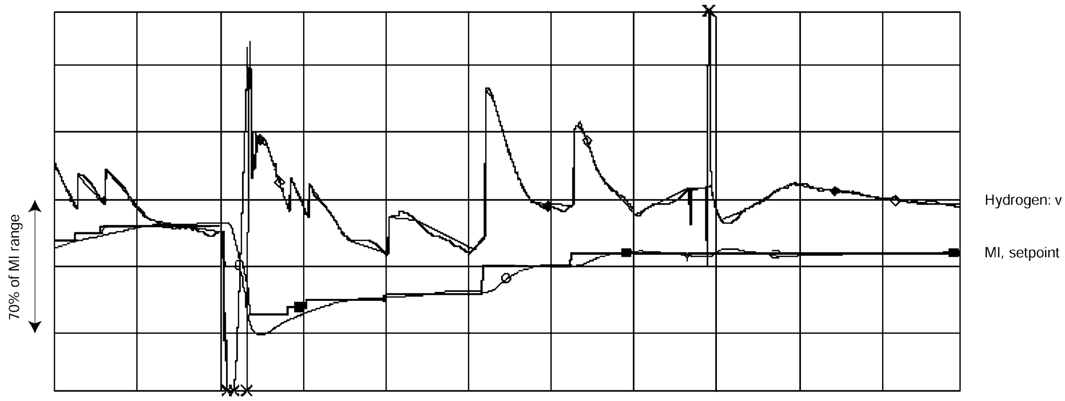


Fig. 5. Industrial results 10 h (tests) for melt-index (*MI*) transients. The transients are fast and precise. Precise scales are omitted for confidentiality reasons. (The off-limit values are sensor failures.) The hydrogen actuator is much faster than the catalyst actuator, explaining the fast response of the system to a setpoint change.

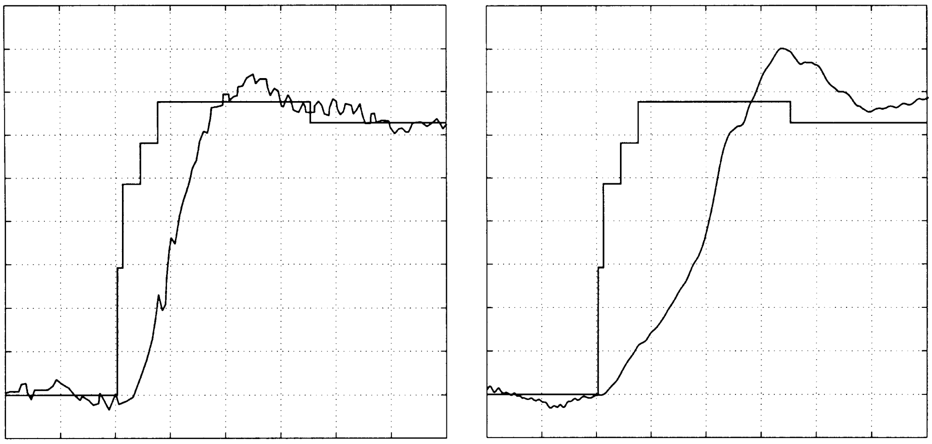


Fig. 6. Comparisons between our controller (test case real time results from history in Fig. 4 on the left) and a LQR controller for a simulation under similar conditions (right.) Prod transition. Time period 1/2 day.

is possible to satisfy some min–max constraints on the inputs by saturating them without compromising the stability of the closed-loop system. On the other hand, this flatness based controller can not, as-is, handle stringent constraints of general forms. The MPC techniques would definitely be better suited for such cases.

5. Conclusion

The flatness of the system allows us to take into account the nonlinearities and the delay of the system. Though there is a tradeoff. On one hand we had to build a nonlinear model of the process which is time consuming and requires a good knowledge of the unit, on the other hand this allowed us to design an efficient controller in a relatively simple way. For such an accuracy demanding application we believe that this methodology is relevant and recommend it. A first order approximation of the unit would be less appropriate.

The key to our approach is the use of the flat outputs of the system. We found them easily thanks to the compactness of our model. It is true that there does not exist any “algorithm” to find the flat outputs. In the field of process control, at least, this is often not a big deal (see [13] for flat plug-flow reactor, flat mixing systems). As in mechanical engineering (see the flat pendulum [3]) the flat outputs always seem to have a strong physical meaning. In the present case this is also true: they are the rate of solid and the melt-index. Currently we are investigating different processes, trying to build relevant models and find their flat outputs.

More details about this particular application and other industrial control realizations in process control using flatness can be found in [13].

References

[1] S.K. Agrawal, N. Faiz, A new higher-order method for optimization of a class of nonlinear dynamic systems without Lagrange

multipliers, Journal of Optimization Theory and Applications 97 (1) (1998) 11–28.

[2] B. Elvers, S. Hawkins, G. Schulz (Eds.), Ullmann's Encyclopedia of Industrial Chemistry. VCH, 1993.

[3] M. Fliess, J. Lévine, P. Martin, P. Rouchon, Flatness and defect of nonlinear systems: introductory theory and examples, Int. J. Control 61 (6) (1995) 1327–1361.

[4] M. Fliess, J. Lévine, P. Martin, P. Rouchon, A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems, IEEE Trans. Automat. Control 44 (1999) 922–937.

[5] M. Fliess, H. Mounier, P. Rouchon, J. Rudolph. Systèmes linéaires sur les opérateurs de Mikusiński et commande d'une poutre flexible, in: ESAIM Proc. " Élasticité, viscolélasticité et contrôle optimal", 8ème entretiens du centre Jacques Cartier, Lyon, 1996, pp. 157–168.

[6] R. Mahadevan, S. K. Agrawal, F. J. Doyle, Differential flatness based nonlinear predictive control of fed-batch bioreactors, in: IFAC Symposium on Advanced Control of Chemical Processes, 2000.

[7] P. Martin, R. M. Murray, P. Rouchon, Flat systems, in: Proc. of the 4th European Control Conf., 1997, pp. 211–264, Plenary lectures and Mini-courses.

[8] K.B. McAuley, J.F. MacGregor, On-line inference of polymer properties in an industrial polyethylene reactor, AIChE J. 37 (6) (1991) 825–835.

[9] K.B. McAuley, J.F. MacGregor, Optimal grade transitions in a gas phase polyethylene reactor, AIChE J. 38 (10) (1992) 1564–1576.

[10] M.B. Milam, K. Mushambi, R.M. Murray, A new computational approach to real-time trajectory generation for constrained mechanical systems, in: IEEE Conference on Decision and Control, 2000, pp. 845–852.

[11] H. Mounier, Propriétés Structurelles des Systèmes Linéaires à Retards: Aspects Théoriques et pratiques, PhD thesis, Université Paris Sud, Orsay, 1995.

[12] H. Mounier, J. Rudolph, Flatness based control of nonlinear delay systems: a chemical reactor example, Int. J. Control 71 (1998) 871–890.

[13] N. Petit, Systèmes à Retards. Platitude en Génie des procédés et contrôle de Certaines Èquations des Ondes. PhD thesis, Ècole des Mines de Paris, 2000.

[14] N. Petit, Y. Creff, P. Rouchon, Minimum time constrained control of acid strength on a sulfuric acid alkylation unit, Chemical Engineering Science 56/8 (2001) 2767–2774.

[15] M. Roberson, APPRYL investments worth 1 billion french francs, Hydrocarbon Processing 77 (6) (June 1998).

[16] K. Wang, T. Loehl, M. Stobbe, S. Engell. A genetic algorithm for online scheduling of a mutiproduct polymer batch plant, in: 7th International Symposium on Process System Engineering, Computer and Chemical Engineering 24, 2000, pp. 393–400.

ELSEVIER

Chemical Engineering Science 60 (2005) 5544–5554

Chemical Engineering Science

www.elsevier.com/locate/ces

An arrangement of ideal zones with shifting boundaries as a way to model mixing processes in unsteady stirring conditions in agitated vessels

J.-Y. Dieulot[a,*], N. Petit[b], P. Rouchon[b], G. Delaplace[c]

[a]*L.A.G.I.S., Laboratoire d'Automatique Génie Informatique et Signal, UMR CNRS 8146, I.A.A.L Ecole Polytechnique Universitaire de Lille, 59655 Villeneuve d'Ascq-France*

[b]*Centre Automatique et Systèmes, École des Mines de Paris, 60 Boulevard Saint Michel, 75272 Paris Cedex 06, France*

[c]*I.N.R.A. Institut National de la Recherche Agronomique, Laboratory for food Process Engineering and Technology, 369, rue Jules Guesde, B.P. 39, 59651 Villeneuve d'Ascq-France*

Received 19 July 2004; received in revised form 2 March 2005; accepted 2 March 2005
Available online 28 June 2005

Abstract

This paper investigates the way modelling mixing phenomena occur in unsteady stirring conditions in agitated vessels. In particular, a new model of torus reactor including a well-mixed zone and a transport zone is proposed. The originality of the arrangement of ideal reactors developed here lies in the time-dependent location of the boundaries between the two zones. This concept is applied to model the positive influence of unsteady stirring conditions on homogenization process: the model avoids a mass balance discontinuity when the transition from steady to unsteady stirring conditions is performed.

To ascertain the reliability of the model proposed, experimental runs with highly viscous fluids have been carried out in an agitated tank. The impeller used was a non-standard helical ribbon impeller, fitted with an anchor at the bottom. The degree of homogeneity in the tank was observed using a conductivity method after a tracer injection.

It is shown that for a given agitated fluid and mixing system, model parameters are easy to estimate and that modelling results are in close agreement with experimental ones. Moreover, it would appear that this model allows the easy derivation of a control law, which is a great advantage when optimizing the dynamics of a mixing process.

Keywords: Mixing; Modelling; Nonlinear dynamics; Parameter identification; Unsteady stirring; Torus model

1. Introduction

1.1. Enhancement of mixing with unsteady flows

High viscosity mixing operations in agitated vessels are commonly encountered in chemical and food industries. Since batch mixing operations are both time and energy consuming, their optimization remains an important challenge. Depending on whether the design of the mixing system is set or not, there are various possible ways to improve mixing:

- The first one is to select, for a given mixing system, the geometrical parameters (wall-clearance, shape of the bottom, bottom clearance, number of baffles) which optimize the overall homogenization efficiency. Of course, this has already been largely covered in the literature. For example, one may mention papers concerning the determination of power consumption and mixing times, under steady rotational speeds, for mixing systems equipped with close clearance impellers such as screw or helical ribbon agitators (Tatterson, 1994; Delaplace et al., 2000a), which are known to be the best suited to achieve mixing of highly viscous media.

[*] Corresponding author. Tel.: +33 3 28 76 74 95; fax: +33 3 28 76 74 01.
E-mail addresses: jean-yves.dieulot@polytech-lille.fr (J.-Y. Dieulot), delapla@lille.inra.fr (G. Delaplace).

0009-2509/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ces.2005.03.067

- The second way is to consider that, for a given mixing system, the ability of a flow to homogenize viscous products can be significantly enhanced with the help of unsteady time-varying stirring approaches. Efficient mixing in laminar regime has been shown to be related to the amount of stretching and folding generated within the tank by the agitator (Ottino, 1989; De La Villeon et al., 1998; Alvarez-Hernández et al., 2002). When stirring conditions are steady, initially designated fluid material will follow closed streamlines in the vessel and consequently the mixing efficiency will be rather poor since such regular flows will induce a linear evolution of intermaterial area with time (Niederkorn and Ottino, 1994). However, when a suitable perturbation is superimposed on the steady velocity field, flows reorientations will appear, fluid elements will be no longer trapped by closed steady streamlines and will become free to wander throughout chaotic flow domains. As the stretching rate is higher in these flow regions, the inter-material area will grow faster (Niederkorn and Ottino, 1994; Alvarez-Hernández et al., 2002) and higher than average values of the efficiency will be obtained.

However, there is a lack of systematic studies that provide us with quantitative information about the conditions under which these chaotic flows are produced within a stirred tank and their actual benefits on mixing efficiency. Consequently, the design of a sequence of flows which involves a reorientation of material elements (for instance, when periodic or co-reverse rotation of the impeller is performed) has yet to be clearly identified.

Moreover, most unsteady stirring approaches used to improve laminar mixing in batch reactors (Nomura et al., 1997; Lamberto et al., 1996; Yao et al., 1998) deal with small-diameter agitators which are usually devoted to work in turbulent regime and not suited for the batch mixing of viscous fluids; their purpose being to prevent the formation of isolated mixed regions (Metzner and Taylor, 1960) with co-reverse or periodic rotational speed sequences. Such a work has not been carried out for systems equipped with efficient closed-clearance impellers.

1.2. Flow modelling in batch reactors

From this survey, it would appear clearly that there is a strong need for rational studies which quantify the efficiency of a stretching process for a given mixing system under unsteady operating conditions. Numerical studies using computational fluid dynamics (CFD) methods allow the determination of the whole velocity field for laminar mixing within the tank at steady and unsteady rotational speeds and thus point out the well-mixed and stagnant zones (e.g. Zalc et al., 2002; Arratia et al., 2004; Harvey and Rogers, 1996; Campolo et al., 2003). However, these finite element methods require a long computation time (e.g. for a vessel with close clearance impeller, see de la Villeon et al., 1998 for details). With these models, one cannot extrapolate the behaviour of the mixing device for a new rotational speed sequence, and no quantitative indication is given as to what rotational speed pattern should be used to optimize mixing (Alvarez-Hernández, et al., 2002). Consequently, one cannot design an optimal control that minimizes energy or time expense to achieve a given degree of homogeneity.

It is, therefore, essential to design a proper and simplified flow model for such mixing processes, incorporating the significant features of partially chaotic phenomena and usable to assess the combined effects of unsteady and steady stirring approaches on mixing efficiency, thereby allowing fast prediction and eventually the derivation of a control law.

Networks of ideal reactors have been used since the 1960s to model mixing with steady stirring approaches. Khang and Levenspiel's (1976) model consists of a plug flow reactor in series with a single continuous stirred tank reactor, with total recycling, in which the fluid flows with a constant flow rate $\dot{Q}$. Assuming that both the volume of these two ideal reactors (V_p for the plug flow reactor and V_d for the well mixed zone) are constant and that the flow rate $\dot{Q}$ which appears in the model is proportional to the rotational speed of the impeller N, it was possible by one experimental run (one tracer injection) to determine the space–time parameters of each ideal mixers (time delay θ for a plug flow reactor $\theta = V_p/\dot{Q}$ and mean residence time $T = V_d/\dot{Q}$ for a CSTR).

This simple model has now been extended (Dieulot et al., 2002) to unsteady mixing, along with an additional CSTR in the recycle loop which represents the benefit (due to additional stretching) of mixing at an unsteady rotational speed which was observed experimentally. As has been previously discussed (Dieulot et al., 2002), this model allows us to use the same network of ideal mixers to simulate the mixing performances of the agitated vessel for both the steady and unsteady approaches. The model allows fast prediction and involves only three geometrical parameters that can be easily determined from only two experimental runs (one at constant impeller speed and the two others using unsteady rotational speed experiments). However, the extension to unsteady flow is not straightforward: the expression of the time delay in the plug flow zone is complicated and, moreover, the introduction of the additional volume does not allow the mass balance to be respected.

This has been the motivation for the model presented in the next section. In order to respect the mass balance, the decision was taken to add no further ideal reactors (as the additional CSTR in the previous study) to account for changes in mixing conditions when transition from steady to unsteady stirring approaches is carried out. On the contrary, an attempt was carried out to model the increase in mixing efficiency due to unsteady stirring conditions both by adapting the relative volume ratios of the ideal zones which compose the final model and by keeping the volume of each ideal zone unchanged. This was achieved by using a juxtaposition of a plug flow zone and a well-mixed zone contained in a torus volume with time varying boundaries. In the following section, we will give more details about the basis of the model and mathematical expressions of the space–time parameters for the ideal reactors contained in the torus volume. Experimental mixing runs (for steady and unsteady stirring approaches) were used to ascertain the validity and to compare performances of this model with those found in the literature. Note that the ultimate framework of this scientific programme is to determine an optimal controller, i.e., the rotational speed profile that minimizes the mixing energy for a given mixing time.

2. Principle of the torus reactor model

Consider a torus of fixed volume V divided into two ideal reactors (a constant stirred tank reactor of volume V_d and a plug flow zone of volume $V_p = V - V_d$) in which flows a Newtonian fluid with a uniform time-varying flow rate $\dot{Q}$ in a clockwise direction (Fig. 1). $y(t)$ refers to the fluid concentration (kg/m^3) in component y (tracer) which varies with time and space. It is assumed that the total material quantity of the component y in the reactor remains constant.

The originality of the torus reactor arises from the time-dependent position of the boundaries (S_1 and S_2) which separate the two ideal flow zones. Indeed, it is assumed that S_1 and S_2 move alternately in a counter-clockwise direction to the flow rate fluctuations. Consequently, when the flow rate is non-steady, the volumes (V_d and V_p) of the two ideal reactors are time variant. In particular, it is assumed that S_1 (respectively, S_2) move only when positive (respectively, negative) variations in the flow rate occur in the torus volume and is otherwise motionless. Note also, that when a variation of flow rate occurs, not only the volume of the zones vary but their location within the torus evolves counter-clockwise.

Assuming that at each time t the flow rate $\dot{Q}(t)$ is proportional to the impeller rotational speed $N(t)$ (via α (m^3), a constant: $\dot{Q}(t) = \alpha N(t)$), the torus model proposed is likely to describe the response curve after a tracer injection, whatever the stirring approach adopted. Indeed, for steady approaches the network of ideal mixers used to simulate the mixing process becomes similar to that of those used by Khang and Levenspiel (1976) whose reliability have been previously shown. Moreover, in the case of unsteady stirring, the model is also supposed to account for the experimental observation that an improvement in mixing occurs when a positive variation in the rotational speed is enforced. For example, in the case of a positive variation in impeller rotational speed, the volume of the stirred tank reactor increases while that of the plug flow decreases. As the whole volume of the torus loop is supposed to be unchanged, an enhancement in mixing is expected.

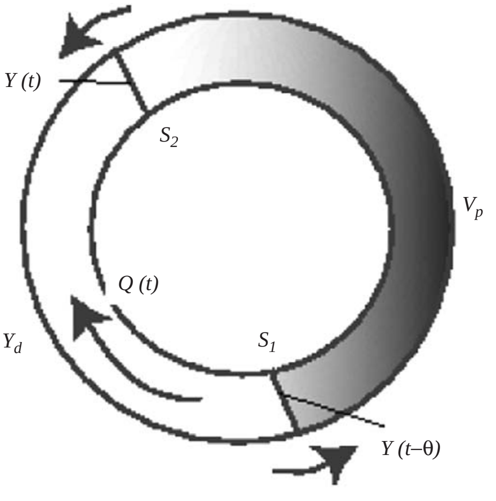


Fig. 1. Sketch of torus model proposed in this study.

Note that the model structure should not be confused with real toroidal reactors (e.g. Benkhelifa et al., 2000).

Let us define $\dot{V}_d^+$ (resp., $\dot{V}_d^-$) as the variation of volume V_d due to the motion of S_1 (resp., S_2) in the torus, and let θ be the residence time of the particle leaving the plug flow zone at time t. Using notations previously introduced, the whole system can be characterized by the following differential equations (see Appendix A):

$$\begin{aligned}
&V = V_d(t) + V_p(t),\\
&\int_{t-\theta}^{t} \dot{Q}(\sigma)\,\mathrm{d}\sigma = V - V_d(\dot{Q}(t-\theta)) - \int_{t-\theta}^{t} \dot{V}_d^+(\sigma)\,\mathrm{d}\sigma,\\
&V_d(\dot{Q}(t))\frac{\mathrm{d}[y(t)]}{\mathrm{d}t} = (\dot{Q}(t) + \dot{V}_d^+)[y(t-\theta) - y(t)],\\
&\dot{Q}(t) = \alpha N(t). \qquad (1)
\end{aligned}$$

2.1. Theorem

The mass balance in the species $y(t)$ within the torus reactor defined by Eq. (1) is respected (see proof in Appendix B).

In our study, it is assumed that in the case of steady mixing (constant rotational speed), the volume of the well-mixed zone does not depend on the amplitude of the rotational speed and has a constant value V_{d_1}. Note that integrating Eq. (1), we obtain

$$V_d(t) = \int_0^t \dot{V}_d^+\,\mathrm{d}t - \int_0^t \dot{V}_d^-\,\mathrm{d}t + V_{d_1}, \qquad (2)$$

where V_{d_1} is the initial volume of the well-stirred zone.

Assuming that the total volume of torus reactor V corresponds to the volume of the agitated fluid, the proposed system involves five unknown variables or parameters α, V_{d_1}, $\dot{V}_d^+(t)$, $\dot{V}_d^-(t)$ and $y(t)$.

Providing two prerequisites, a simulation algorithm can be used to predict the output $y(t)$:

- the two constant parameters (α and V_{d_1}), which are not influenced by the time-dependent rotational speed, are known.
- the effects of stirring conditions $N(t)$ on boundary motions (S_1 and S_2) are established. Indeed, such knowledge

will allow $\dot{V}_d^+$ and $\dot{V}_d^-$ to be obtained at each time. Consequently, V_d and V_p can also be computed.

The assumptions used in this paper concerning evolutions of $\dot{V}_d^+$ and $\dot{V}_d^-$ with stirring conditions are the following, which requires a third constant parameter k for the model:

$$\dot{V}_d^+ = k\,\frac{\mathrm{d}N}{\mathrm{d}t} \ \text{ if } \frac{\mathrm{d}N}{\mathrm{d}t} > 0, \quad \dot{V}_d^+ = 0 \ \text{ if } \frac{\mathrm{d}N}{\mathrm{d}t} < 0,$$
$$\dot{V}_d^- = -k\,\frac{\mathrm{d}N}{\mathrm{d}t} \ \text{ if } \frac{\mathrm{d}N}{\mathrm{d}t} < 0, \quad \dot{V}_d^- = 0 \ \text{ if } \frac{\mathrm{d}N}{\mathrm{d}t} > 0. \tag{3}$$

The simulation of a system involving an input-dependent transport delay is not always trivial, since the delay is defined by an implicit equation. In particular, Zenger and Ylinen (1994) have shown that for most flow rate fluctuations, the expression of $\theta(t)$ cannot be obtained analytically but must be computed by numerical methods. In this work, for the sake of simplicity, it has been chosen not to deal with this issue in detail. More information about the computational methods used in this work can be found in the original publication (Zenger and Ylinen, 1994) or in a previous paper (Dieulot et al., 2002).

Finally, note that the simulation algorithm has been developed considering the torus model as a discrete automaton. First, the torus has been divided into a large number of cells. At each simulation step, the values of the concentration should move from one cell of the plug flow zone to the next one, using the definition of a plug flow reactor (pure transport). The concentration in the well-mixed zone can then be computed using a total mass balance and the fact that the concentrations in each cell of the zone are equal. The boundaries are then updated. The time step is variable and corresponds to the residence time in a cell, which depends on the flow rate values (rotational speed).

3. Material and methods

3.1. Apparatus used to monitor mixing experiments

The mixing equipment used appears in Fig. 2. During all the experiments, the level of the liquid at rest was maintained at a constant level of 0.402 m in height for a total volume of $30 \times 10^{-3}\,\mathrm{m}^3$. Experiments were carried out with the helix pumping upward (counter-clockwise direction of rotation). Additional information about the flow pattern produced by the mixing system is given elsewhere (Delaplace et al., 2000a,b).

The agitated fluid is an aqueous solution of glucose with a viscosity of 1.8 Pa s at 26 °C. A controlled speed rotational viscometer (CONTRAVES, Rheomat 30) was used to determine the Newtonian viscosity of the viscous medium. The shear rate ranged from 0.1–500 s^{-1} and the dependence of viscosity and density on temperature was taken into account.

A conductivity probe (SOLEA-TACCUSSEL, type CD 78) was used to obtain the circulation curves in the vessel after a tracer injection. The signal was amplified by a converter (Type AT40, SFERE), and recorded with the help of an I/O board (PCL-812 PG, ADVANTECH) plugged into a PC. The sampling rate was 200 Hz.

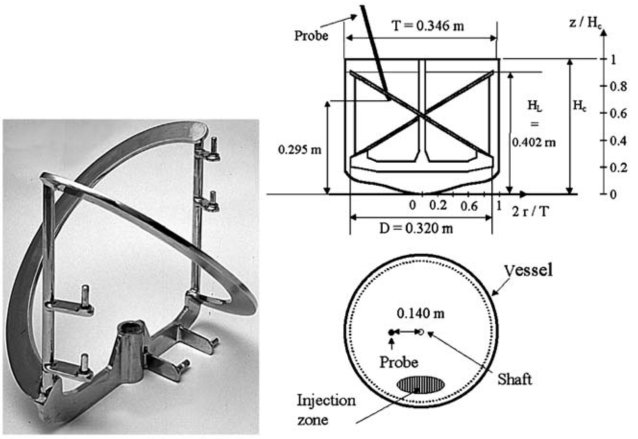


Fig. 2. Picture and geometrical parameter of the mixing equipment investigated (other geometrical parameters of PARAVISC® mixing system: blade width, $w = 0.032$ m; impeller pitch, $p = 0.560$ m; impeller height, $L = 0.340$ m; tank height, $H_c = 0.443$ m).

The tracer pulse injected had the same physical properties as the fluid in the tank (composition and temperature), with an additional quantity of NaCl at a concentration of 100 g/l. The incorporation was performed with the help of a pneumatic system with pistons (type DACO, PCM DOSYS) equipped with a duct (DACC 48/40, DOSYS) which holds the product at the end of the pipe. This device was able to inject 72 ml (0.24% of the tank volume) of viscous tracer into the tank with an accuracy of 2%. The injection duration is by a fraction of a second. It was checked, measuring a sample of the injected fluid before and after each injection, that the influence of the addition of salt on density and viscosity was negligible for a limited (40) number of successive trials. The volume of the tank was brought back to $30 \times 10^{-3}\,\mathrm{m}^3$ after each experiment. The conductivity probe and the injection locations were kept unchanged throughout the experiments (Fig. 2).

The I/O board allows the operating conditions to be accurately controlled, i.e., the injection time, the departure and the magnitude of speed variations that were enforced on the agitation system. The rotational speed and the conductivity signal were recorded throughout the mixing process. Recording was activated 3 s before the tracer injection. Each experiment (for one set of experimental conditions) was repeated four times to ensure repeatability.

The values of the rotational speed varied from 0.16 to 1.5 rev/s. Mixing and circulation times were determined from the response signal recorded after tracer injection. The mixing time is defined as the duration needed for the signal to reach 95% of its final value (Fig. 3). The circulation time is defined as the signal period, when mixing at constant impeller rotational speed (Fig. 3). When the conductivity

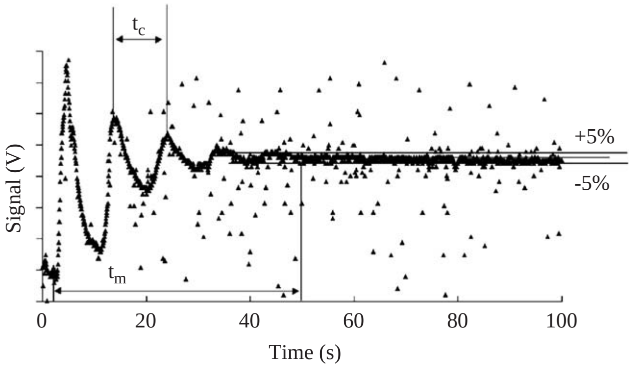


Fig. 3. A typical probe response curve.

method was used, it is clear that the values of circulation and mixing times depend significantly on the location of the injection point and measurement probe. Nevertheless, for the experimental conditions tested, the local values of axial circulation times obtained are in close agreement with the global values obtained by following the movement of freely suspended particles. Moreover, global values of axial circulation times deduced from CFD velocity field (Delaplace et al., 2000a) were also in close agreement with those obtained by the conductivity method.

The conductivity signals (axial circulation curves) were particularly noisy, owing to recording problems and high-frequency environmental noise (Fig. 3). Filtering consisted, firstly, in the elimination of scatters. Measuring points with a derivative higher than a threshold value (empirically five times the signal derivative standard deviation) were replaced by an average value of their neighbours. The sampling period selected was 1 s. This choice was important for parametric identification and has been already justified and discussed elsewhere (Dieulot et al., 2002).

3.2. Operating stirring conditions tested

Table 1 shows the different types of operating stirring conditions tested after tracer injection. Trials (1) and (2) refer to well-known steady stirring approaches, whereas trials (3) to (8) concern unsteady stirring approaches. In the context of this paper, trials (3) to (4) will be called “speed ramps”, trials (5) to (7) “speed pulses” and trial (8) “speed step”.

Note that for each type of perturbation, different operating conditions were adopted (e.g. various lapses of time between tracer injection and the start of the impeller rotational speed fluctuations (PS)). The various operating conditions tested are also reported in Table 1.

3.3. Parameter identification: V_{d_1}, α, k

The torus model proposed requires the estimation of three constant parameters V_{d_1}, α, k (defined by Eqs. (1)–(3)), which depend on the characteristics of the mixing device and on the viscous media (which are maintained at a constant level in this study).

Parameters V_{d_1} and α have been estimated from one tracer experiment when mixing at constant impeller speed (0.667 rev/s—trial number 1 in Table 1). Using the values of the parameters V_{d_1} and α previously estimated, an additional injection was performed with unsteady stirring conditions (a speed pulse—trial number five in Table 1) to obtain the value of parameter k.

The set of model parameters were estimated using an optimization algorithm (simplex method). The optimization algorithm is based on the minimization of the mean absolute error criterion defined in Eq. (4).

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=0}^{M-1} |\varepsilon(i.T_e)|. \tag{4}$$

This criterion represents the sum of the absolute differences, $|\varepsilon(i.T_e)|$, between the experimental points and the estimated points, T_e is the sampling period (1 s) and M is the number of samples required to describe the homogenization process. The importance of this criterion was discussed by Dieulot et al. (2002), where it was shown that it leads to a good compromise between minimizing the shifting between real and modelling curves (due to time delay estimation mismatch) and other errors due to unmodelled non-linearities.

3.4. Reliability of the model

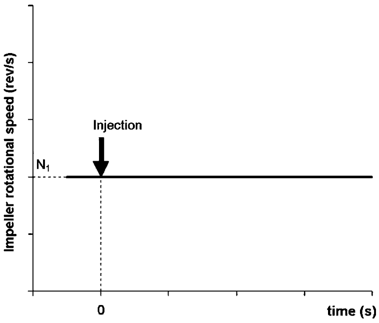
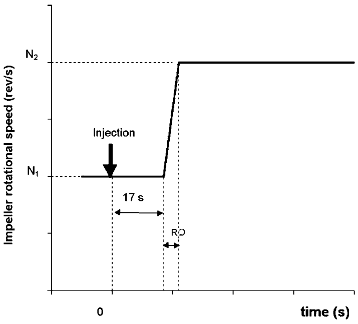
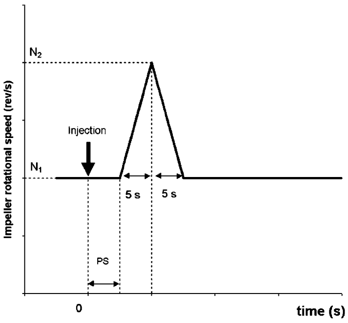
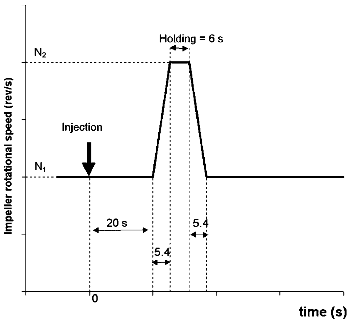
Using different operating conditions (trials 2–4 and 6–8 in Table 2) to those adopted for parameter estimation (trials 1 and 5), the validity of the model was tested. The reliability procedure consists of comparing experimental and predicted mixing times (obtained with the help of estimated parameters). The mean absolute error between experimental and model data was also computed and its value was compared to those obtained for the trials used for fitting.

4. Results

4.1. Efficiency of mixing using unsteady stirring conditions

The positive influence of unsteady stirring condition on mixing efficiency is recalled in Table 2. It can be observed that the mixing work required for unsteady stirring is less significant than those calculated for those mixing procedures which would give identical mixing times at constant RPM. These values of energy consumed and their determinations have already been discussed (Dieulot et al., 2002) and are not the key consideration of this work. Note simply that, as presented in previous works, depending upon the type of unsteady stirring conditions adopted, the energy savings vary from 30% to 60% and justify the interest of introducing time-dependent perturbations for a homogenization process.

Table 1
Operating conditions (impeller rotational speed fluctuations) adopted during the mixing process after tracer injection

Trial number	Name and type of impeller rotational speed fluctuation		N_1 (rev/s)	N_2 (rev/s)	Time parameters (s)
1	Steady Stirring		0.667	—	—
2	Steady stirring		0.833	—	—
3	Ramp Speed		0.667	1.333	RD 5
4	Ramp Speed		0.667	1.333	15
5	Pulse Speed		0.667	1.333	PS 17
6	Pulse Speed		0.667	1.333	10 4
7	Pulse Speed		0.667	1.333	
8	Step speed		0.667	1.333	—

4.2. Validity of the model

The predictive model developed in this study has been tested on our mixing equipment. As mentioned before, one trial at constant impeller speed (trial 1) and one run at unsteady impeller speed (a pulse—trial 5—see Table 3) were necessary to determine the various ideal zone parameters. The values of the parameters estimated are $V_{d_1} =$

Table 2
Experimental mixing performances of the helical mixing system studied using various stirring conditions (starting impeller rotational speed = 0.667 rev/s except trial 2)

	Trial 1 Steady stirring	Trial 2 Steady stirring	Trial 3 Ramp (RD = 5 s)	Trial 4 Ramp (RD = 15 s)	Trial 5 Pulse (PS = 17 s)	Trial 6 Pulse (PS = 10 s)	Trial 7 Pulse (PS = 4 s)	Trial 8 Step
Experimental mixing time (s)	80.7	72	33	38.5	60.7	58.5	65.1	53.3
Experimental mixing work (J)	627.9	937.8	793.6	763.6	510.9	525.6	385	579.7
Values of mixing work (J) for the mixing process which would give same mixing time at constant impeller rotational speed[12]	627.9	937.8	1535.5	1214.3	740.4	814.6	1035.7	980.9
Energy savings	—	—	48.3	37.1	31.0	35.5	62.8	40.9

Table 3
Values of MAE and predicted values of mixing times obtained by the model for the helical mixing system studied using various stirring conditions

	Operating conditions used for parameter identification		Operating conditions used for model validation					
	Trial 1 Steady stirring $N = 0.667$ rev/s	Trial 5 Pulse (PS = 17 s)	Trial 2 Steady stirring $N = 0.833$ rev/s	Trial 3 Ramp (RD = 5 s)	Trial 4 Ramp (RD = 15 s)	Trial 6 Pulse (PS = 10 s)	Trial 7 Pulse (PS = 4 s)	Trial 8 Step
Experimental mixing time (s)	80.7	60.7	72	33	38.5	58.5	65.1	53.3
Predicted values of mixing time (s)	81.5	66.7	65.6	31.5	37.4	66.7	66.5	47.5
Values of criterion MAE (V)	0.24	0.26	0.27	0.24	0.28	0.20	0.18	0.24

$6.0\,10^{-3}\,\mathrm{m}^3$; $\alpha = 1.61\,10^{-3}\,\mathrm{m}^3$; $k = 4.5\,10^{-3}\,\mathrm{m}^3\,\mathrm{s}$ and were then used with other stirring conditions (see Table 3) to validate the proposed model.

Examples of curve fitting obtained by this approach are given in Figs. 4–9. These figures show that the estimated response curve after tracer injection is close to the experimental one, despite the high noise observed for the experimental curves and the non-linearities (such as the non-periodicity of signals which sometimes occurs at constant impeller speeds). Moreover, in order to test the accuracy of the model, the values of measured mixing times and values calculated by the model are reported in Table 3. We can note that there are close agreements between the experimental and predicted values of mixing times, whatever the stirring conditions adopted (mean error 6.8%). As previously explained, another criterion has also been computed to estimate the validity of the model: the sum of the absolute differences between the calculated and experimental outlet curves. Values of the mean absolute error (MAE) between experimental and model data are also reported in Table 3.

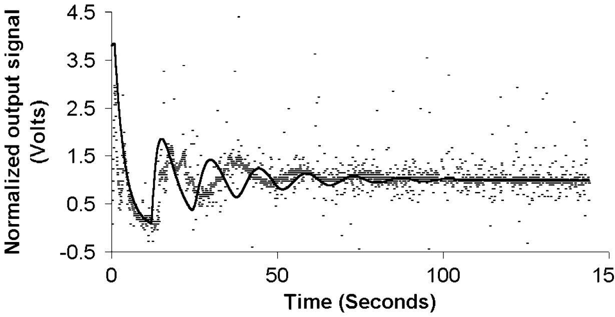


Fig. 4. Predicted (-) and experimental (.) circulation curves for steady speed at 40 rpm.

The values of MAE deduced from trials used for model validation (0.18, 0.28) are not significantly different from

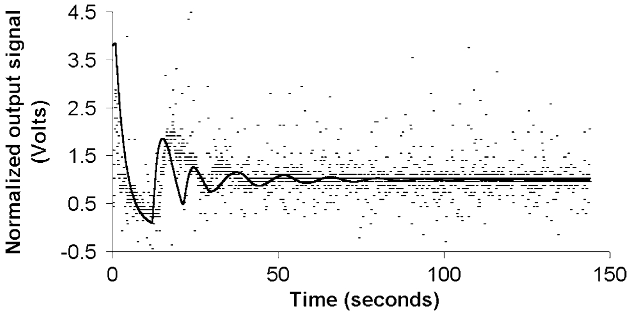


Fig. 5. Predicted (-) and experimental (.) circulation curves for speed pulse from 40 to 80 rpm, starting at 17 s, duration 5 s (trial 5 in Table 2).

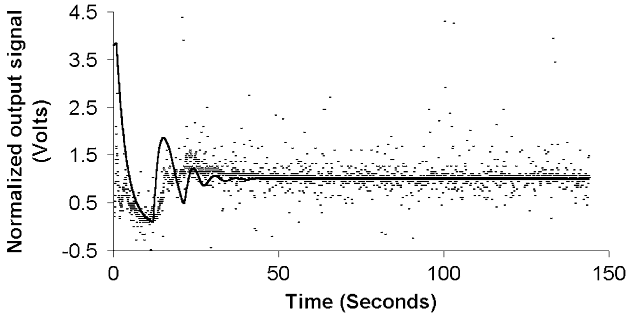


Fig. 6. Predicted (-) and experimental (.) circulation curves for speed ramp from 40 to 80 rpm, starting at 17 s, ramp duration RD = 5 s (trial 4 in Table 2).

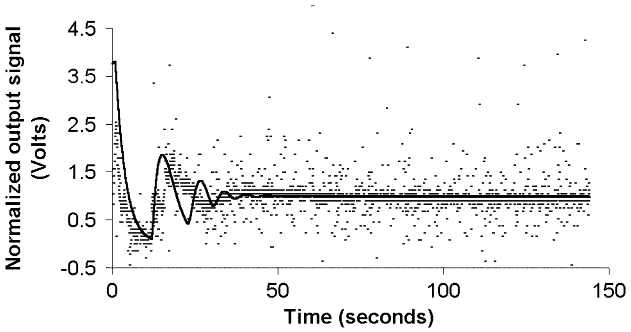


Fig. 7. Predicted (-) and experimental (.) circulation curves for speed ramp from 40 to 80 rpm, starting at 17 s, ramp duration RD = 15 s. (trial 3 in Table 2).

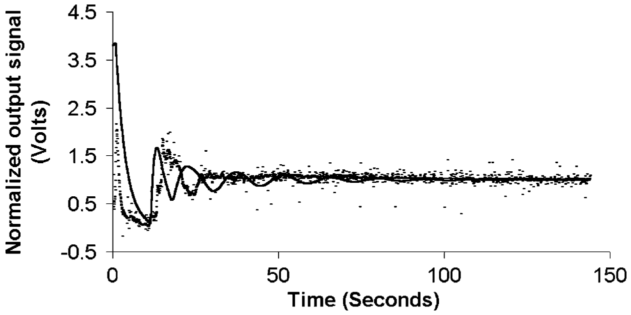


Fig. 8. Predicted (-) and experimental (.) circulation curves for speed pulse from 40 to 80 rpm, starting at 10 s, duration 5 s. PS = 10 s (trial 6 in Table 2).

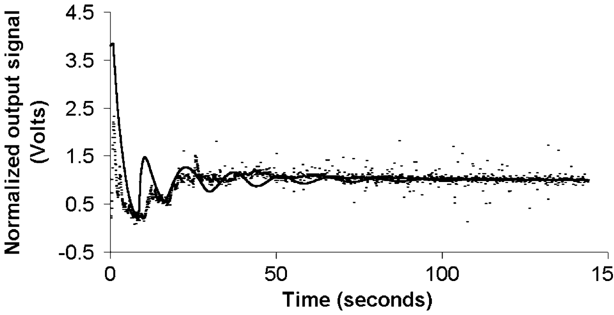


Fig. 9. Predicted (-) and experimental (.) circulation curves for speed pulse from 40 to 80 rpm, starting at 4 s, duration 5 s. PS = 4 s (trial 7 in Table 2).

those deduced from trial used for parameter estimation (0.25), and show that the model is in accordance with the experimental curves.

All these experimental results concerning modelling show us that it is possible to describe the mixing process which occurs under steady or unsteady stirring using the structure of the proposed model. It is thus possible to perform fast mixing time computations (less than 1 s on a PC) for any rotational speed profile. One major interest of the torus reactor compared to previous studies (Dieulot et al., 2002) is that the mass balance in the species within the batch reactor is respected. Note that the toroidal reactor model is not a limited concept only applied for a specific mixing system design. On the contrary, the proposed model can be generalized to describe other mixing processes at steady or unsteady rotational speeds in stirred tanks.

Another motivation to use the torus reactor relies on deriving a control law from mathematical equations (for the rotational speed of the impeller) which thereby optimizes mixing dynamics. Indeed, introducing the following change in time-scale:

$$\mathrm{d}s = (\dot{Q}(t) + \dot{V}_d^{+})\,\mathrm{d}t, \tag{5}$$

the mass-balance equations become

$$\begin{aligned} s(t) - s(t-\theta) = V_p \quad \text{and} \quad & V_d(U(s))\frac{\mathrm{d}Y}{\mathrm{d}s}(s) \\ & = Y(s-\varsigma) - Y(s), \end{aligned} \tag{6}$$

where $\dot{Q} = U(s(t))$; $y(t) = Y(s(t))$ and ς is defined by $\varsigma = V_p(U(s-\varsigma))$.

From Eqs. (5) and (6), it can be seen that, if V_d is an increasing function of u, then there is a difference when $u > 0$ ($\dot{V}_d^{+} \neq 0$) and $u < 0$ ($\dot{V}_d^{+} = 0$), and that for $u > 0$ the new "time" $s(t)$ passes faster. Mixing is thus more efficient when the flow accelerates, which is consistent with experimental observations. This can be illustrated by considering a flow with a saw tooth profile, for which the volumes will return to their initial value after the saw tooth is completed. In a first instance, the boundary S_1 moves and V_d expands. When S_2 moves in turn and V_p expands as u decreases, the plug flow zone will move counter-clockwise and will overlap an area which was previously in the well-mixed zone. The effect of $u < 0$ is thus more limited than in the case where $u > 0$.

Finally, defining ς by $W = V_d(U)$ as the control parameter and using volume balance in the torus, it can be written

$$W \frac{\mathrm{d}Y}{\mathrm{d}s}(s) = Y(s - \varsigma) - Y(s) \quad \text{and} \\ \varsigma + V_d(U(s - \varsigma)) = V. \tag{7}$$

By construction $W \in [0, V]$ and a positive solution for ς always exists when W is a continuous function of s. When the equation has several roots, ς should be chosen as the smallest. Consequently, using the torus model, an optimal solution for the control should be quite simple to obtain using algebraic methods. Preliminary results have been obtained (Dieulot and Richard, 2001), which will be extended in future work.

5. Conclusion

A torus model has been developed to describe a mixing process at unsteady rotational speeds. The combination of ideal reactors proposed includes a well-mixed and a plug flow zone contained in a torus volume. The boundaries between the two zones vary with the flow rate (proportional to impeller rotational speed) and are supposed to represent the enhancement of mixing efficiency, experimentally observed when using unsteady stirring conditions. Only the knowledge of three constant parameters V_{d_1}, α, k is required for the model proposed. Moreover, only two trials are necessary to estimate the three fixed parameters (one at constant impeller speed V_{d_1}, α, and one at unsteady rotational speed k). Finally, the model proposed gives a close agreement between predicted and experimental circulation curves and allows us to estimate the mixing times, for any kind of time-dependent rotational impeller speed tested.

Of course, the model proposed fails to demonstrate that the use of dynamic flow perturbations (time-dependent revolution per minute) contributes to generate a more global chaotic flow which reduces segregated regions and enhances mixing as Tanguy et al. (1998) and Lamberto et al. (2001) have done with CFD applications. The model proposed does not allow to obtain the time-dependent map of the segregated regions. However, our model is quite complementary to CFD applications and very useful since according to us, so far, there was no way to predict by an arrangement of ideal reactors the enhancement of mixing when using time-dependent stirring conditions. In this sense the model proposed succeeds in quantifying quickly the gain in mixing time and energy provided by applying time-dependent RPM.

Moreover, this study has been conducted with a helical ribbon impeller but the approach proposed is not limited to this kind of agitators and can be extended to other mixing systems. It would be even possible to propose a new classification of mixing systems based on their homogenization performances during unsteady stirring and would at last allow to propose new mixers that have an appropriate behaviour when mixing under such operating conditions.

Finally, the mathematical equations of the system are indeed easily tractable which allows to define an optimal control strategy for the torus model. This will be tackled in a future work. The optimal control would be a compromise between the additional energy required to damp down quickly the degree of homogeneity and additional energy required to create dynamic flow perturbations (unsteady rotational speed).

Notation

D	impeller diameter, m
H_c	tank height, m
H_L	liquid height, m
k, α	model parameters (see units in text)
L	impeller height, m
N	impeller rotational speed, rev/s
p	helical ribbon pitch, m
$\dot{Q}$	fluid flow rate, m^3/s
S_1, S_2	moving boundaries for the torus volume, m^2
t	time, s
t_m	mixing time, s
T	tank diameter, m
T_e	sampling period used for estimation, s
V	vessel or torus reactor volume, m^3
V_d	volume of the well-mixed zone for the torus volume, m^3
V_p	volume of the plug flow zone for the torus volume, m^3
w	blade width, m
W_m	mixing work, J
$y(t)$	tracer concentration, kg/m^3

Greek letters

α	proportionality constant, m^3
θ	time-varying delay, s
μ	viscosity of Newtonian fluid, Pa s
ρ	fluid density, kg/m

Appendix A. Derivation of the toroidal reactor model equations

A.1. Space–time for the constant stirred tank zone in the torus loop

Defining $\dot{V}_d^+$ (resp., $\dot{V}_d^-$) as the variation of volume V_d due to the motion of S_1 (resp., S_2) in the torus, variation in volume V_d with time can be written as

$$\frac{\mathrm{d}[V_d(\dot{Q}(t))]}{\mathrm{d}t} = \dot{V}_d^{+} - \dot{V}_d^{-}. \tag{8}$$

Using notations previously developed, the material balance in the well-mixed zone is

$$\begin{aligned}\frac{\mathrm{d}[V_d(\dot{Q}(t))y(t)]}{\mathrm{d}t} &= (\dot{Q}(t) + \dot{V}_d^{+}(t))y(t-\theta) \\ &\quad - (\dot{Q}(t) + \dot{V}_d^{-}(t))y(t),\end{aligned} \tag{9}$$

where θ is the residence time of the particle leaving the plug flow zone at time t.

Another expression of material balance in the well-mixed zone is

$$\frac{\mathrm{d}[V_d(\dot{Q}(t))y(t)]}{\mathrm{d}t} = V_d(\dot{Q}(t))\frac{\mathrm{d}y(t)}{\mathrm{d}t} + y(t)\frac{\mathrm{d}[V_d(\dot{Q}(t))]}{\mathrm{d}t}, \tag{10}$$

combining Eqs. (8) and (9) with Eq. (10):

$$V_d(\dot{Q}(t))\frac{\mathrm{d}[y(t)]}{\mathrm{d}t} = (\dot{Q}(t) + \dot{V}_d^{+})[y(t-\theta) - y(t)]. \tag{11}$$

A.2. Space–time modelling for the plug flow zone in the torus loop

The particle which enters the plug flow zone at the instant $t = t - \theta$ and transported at non-steady flow rate $\dot{Q}$ in a clockwise direction must go through the plug flow volume ahead of it, before leaving at time t. The transport delay θ is defined by the implicit Eq. (12):

$$\int_{t-\theta}^{t} \dot{Q}(\sigma)\,\mathrm{d}\sigma = V_p(\dot{Q}(t-\theta)) - \int_{t-\theta}^{t} \dot{V}_d^{+}(\sigma)\,\mathrm{d}\sigma. \tag{12}$$

Due to the clockwise flow direction, the plug flow volume ahead of the particle can only decrease during the route. This decrease corresponds to the second right term of Eq. (12).

Appendix B. Material balance in the torus reactor (proof of the theorem)

Since $\dot{V}_d + \dot{V}_p = 0$,

$$\begin{aligned}V_p(t) - V_p(t-\theta) &= \int_{t-\theta}^{t} \dot{V}_p(\sigma)\,\mathrm{d}\sigma \\ &= -\int_{t-\theta}^{t} \dot{V}_d(\sigma)\,\mathrm{d}\sigma \\ &= \int_{t-\theta}^{t} (\dot{V}_d^{-} - \dot{V}_d^{+})\,\mathrm{d}\sigma,\end{aligned}$$

and Eq. (12) can be rewritten as

$$\int_{t-\theta}^{t} \dot{Q}(\sigma)\,\mathrm{d}\sigma = V_p(t) - \int_{t-\theta}^{t} \dot{V}_d^{-}(\sigma)\,\mathrm{d}\sigma.$$

At time t, the quantity $A(t)$ of the species y inside the reactor is the sum of that in the plug flow and the well-mixed zones,

$$A(t) = V_d(t)y(t) - \int_{plug\ flow} y(t - \theta(z,t))S\,\mathrm{d}z,$$

where S is the constant surface of the torus section and $\theta(z,t)$ is the time delay of a particle whose position in the plug flow zone is z. The abscissa z ranges from 0 to $V_p(t)/S$. The particles which are at position z at time t have entered the plug flow zone at time $t - \theta(z,t)$. These particles had to travel the distance $z - \frac{1}{S}\int_{t-\theta(z,t)}^{t} \dot{V}_d^{-}(\sigma)\,\mathrm{d}\sigma$ which yields the following relation which in turn generalizes Eq. (12):

$$\int_{t-\theta(z,t)}^{t} \dot{Q}(\sigma)\,\mathrm{d}\sigma = Sz - \int_{t-\theta(z,t)}^{t} \dot{V}_d^{-}(\sigma)\,\mathrm{d}\sigma.$$

Now let us show that the derivative of $A(t)$ is zero.

Deriving the equation above with respect to t and z, we obtain the useful relations

$$\begin{aligned}&\dot{Q}(t) + \dot{V}_d^{-}(t) = (\dot{Q}(t - \theta(z,t)) \\ &\quad + \dot{V}_d^{-}(t - \theta(z,t)))\left(1 - \frac{(\partial\theta(t - \theta(z,t), z))}{\partial t}\right), \\ &\frac{\partial\theta(z,t)}{\partial t}(\dot{Q}(t - \theta(z,t)) + \dot{V}_d^{-}(t - \theta(z,t))) = S.\end{aligned}$$

First we calculate

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int_{plug\ flow} y(t - \theta(z,t))Sdz \\ &\quad = \dot{V}_p(t)y(t-\theta) + \int_0^{V_p/S} \dot{y}(t - \theta(z,t)) \\ &\qquad \times \left(1 - \frac{(\partial\theta(t - \theta(z,t), z))}{\partial t}\right)S\,\mathrm{d}z,\end{aligned}$$

which becomes, using previous equations,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int_{plug\ flow} y(t - \theta(z,t))S\,\mathrm{d}z \\ &\quad = \dot{V}_p(t)y(t-\theta) + \int_0^{V_p/S} \dot{y}(t - \theta(z,t)) \\ &\qquad \times \frac{\dot{Q}(t) + \dot{V}_d^{-}(t)}{(\dot{Q}(t - \theta(z,t)) + \dot{V}_d^{-}(t - \theta(z,t)))}S\,\mathrm{d}z\end{aligned}$$

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int_{plug\ flow} y(t - \theta(z,t))S\,\mathrm{d}z \\ &\quad = \dot{V}_p(t)y(t-\theta) + (\dot{Q}(t) + \dot{V}_d^{-}(t)) \\ &\qquad \times \int_0^{V_p/S} \dot{y}(t - \theta(z,t))\frac{\partial\theta(z,t)}{\partial z}S\,\mathrm{d}z,\end{aligned}$$

and, integrating the last equation

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{plug\ flow} y(t - \theta(z,t))S\,\mathrm{d}z \\ &\quad = \dot{V}_p(t)y(t-\theta) + (\dot{Q}(t) \\ &\qquad + \dot{V}_d^{-}(t))(y(t) - y(t - \theta(t))).\end{aligned}$$

Replacing in the derivative of $A(t)$ yields

$$\dot{A}(t) = (\dot{Q}+\dot{V}_d^-)y(t-\theta)-(\dot{Q}+\dot{V}_d^-)y(t)+\dot{V}_p(t)y(t-\theta)$$
$$+ (\dot{Q}(t) + \dot{V}_d^-(t))(y(t) - y(t - \theta(t))),$$
$$\dot{A}(t) = (\dot{V}_d^+ - \dot{V}_d^- + \dot{V}_p)y(t - \theta) = 0,$$

which completes the proof.

References

Alvarez-Hernández, M.M., Shinbrot, T., Zalc, J., Muzzio, F.J., 2002. Practical chaotic mixing. Chemical Engineering Science 57, 3749–3753.

Arratia, P.E., Lacombe, J.P., Shinbrot, T., Muzzio, F.J., 2004. Segregated regions in continuous laminar stirred tank reactors. Chemical Engineering Science 59, 1481–1490.

Benkhelifa, H., Legrand, J., Legentilhomme, P., Montillet, A., 2000. Study of the hydrodynamic behaviour of the batch and continuous torus reactor in laminar and turbulent flow regimes by means of tracer methods. Chemical Engineering Science 55, 1871–1882.

Campolo, M., Sbrizzai, F., Soldati, A., 2003. Time-dependent flow structures and Lagrangian mixing in Rushton-impeller baffled-tank reactor. Chemical Engineering Science 58, 1615–1629.

Delaplace, G., Leuliet, J.-C., Relandeau, V., 2000a. Circulation and mixing times for helical ribbon impellers. Review and experiments. Experiments in Fluids 28 (2), 170–182.

Delaplace, G., Torrez, C., André, C., Belaubre, N., Loisel, P., 2000b. Numerical simulation of flow of Newtonian fluids in an agitated vessel equipped with a non standard helical ribbon impeller, Proceedings of 10th European Conference on Mixing, (Delft 2–5 july), (Elsevier Amsterdam): 289–296.

De La Villeon, J., Bertrand, F., Tanguy, P.A., Labrie, R., Bousquet, J., Lebouvier, D., 1998. Numerical investigation of mixing efficiency of helical ribbons. A.I.Ch.E. Journal 44 (4), 972–977.

Dieulot, J.-Y., Richard, J.-P., Tracking control of a nonlinear system with input-dependent delay. IEEE Int. Conf. Decision and Control. CDC 01, Orlando, 2001.

Dieulot, J.-Y., Delaplace, G., Guérin, R., Brienne, J-P., Leuliet, J.-C., 2002. Laminar mixing performances of a stirred tank equipped with helical ribbon agitator subjected to steady and unsteady rotational speed. Transactions of Institution of Chemical Engineers 80 (Part A), 335–344.

Harvey, A.D., Rogers, S.E., 1996. Steady and unsteady computation of impeller stirred tank reactors. A.I.Ch.E. Journal 42, 2701–2712.

Khang, S.J., Levenspiel, O., 1976. New scale-up and design method for stirrer agitated batch mixing vessels. Chemical Engineering Science 31, 569–577.

Lamberto, D.J., Muzzio, F.J., Swanson, P.D., Tonkovich, A.L., 1996. Using time-dependent RPM to enhance mixing in stirred vessels. Chemical Engineering Science 51 (5), 733–741.

Lamberto, D.J., Alvarez, M.M., Muzzio, F.J., 2001. Computational analysis of regular and chaotic mixing in a stirred tank reactor. Chemical Engineering Science 56 (16), 4887–4899.

Metzner, A.B., Taylor, J.S., 1960. Flow patterns in agitated vessels. A.I.Ch.E. Journal 6 (1), 109–114.

Niederkorn, T.C., Ottino, J.M., 1994. Chaotic mixing of shear-thinning fluids. A.I.Ch.E. Journal 40 (11), 1782–1793.

Nomura, T., Uchida, T., Takahashi, K., 1997. Enhancement of mixing by unsteady agitation of an impeller in an agitated vessel. Journal of Chemical Engineering of Japan 30 (5), 875–879.

Ottino, J.M., 1989. The Kinematics of Mixing. Stretching, Chaos and Transport. Cambridge University Press, Cambridge.

Tanguy, P.A., Thibault, F., Brito-De La Fuente, Espinosa Solares, T., Tecante, A., 1998. Mixing performance induced by coaxial flat blade-helical ribbon impellers rotating at different speeds. Chemical Engineering Science 52 (11), 1733–1741.

Tatterson, G.B., 1994. Scale Up and Design of Industrial Mixing Processes. McGraw-Hill, New York.

Yao, W.G., Sato, H., Takahashi, K., Koyama, K., 1998. Mixing performance experiments in impeller stirred tanks subjected to unsteady rotational speeds. Chemical Engineering Science 53 (17), 3031–3043.

Zalc, J.M., Szalai, E.S., Alvarez, M.M., Muzzio, F.J., 2002. Using CFD To Understand Chaotic Mixing in Laminar Stirred Tanks. A.I.Ch.E. Journal 48, 2124–2134.

Zenger, K., Ylinen, R., 1994. Simulation of variable delays in material transport models. Mathematics and Computers in Simulation 37, 57–72.

ESAIM: Control, Optimisation and Calculus of Variations February 2003, Vol. 9, 275–296
DOI: 10.1051/cocv:2003013

MOTION PLANNING FOR A NONLINEAR STEFAN PROBLEM

William B. Dunbar[1], Nicolas Petit[2], Pierre Rouchon[2] and Philippe Martin[2]

Abstract. In this paper we consider a free boundary problem for a nonlinear parabolic partial differential equation. In particular, we are concerned with the inverse problem, which means we know the behavior of the free boundary *a priori* and would like a solution, *e.g.* a convergent series, in order to determine what the trajectories of the system should be for steady-state to steady-state boundary control. In this paper we combine two issues: the free boundary (Stefan) problem with a quadratic nonlinearity. We prove convergence of a series solution and give a detailed parametric study on the series radius of convergence. Moreover, we prove that the parametrization can indeed can be used for motion planning purposes; computation of the open loop motion planning is straightforward. Simulation results are given and we prove some important properties about the solution. Namely, a weak maximum principle is derived for the dynamics, stating that the maximum is on the boundary. Also, we prove asymptotic positiveness of the solution, a physical requirement over the entire domain, as the transient time from one steady-state to another gets large.

Mathematics Subject Classification. 93C20, 80A22, 80A23.

Received October 31, 2002. Revised January 24, 2003.

1. Introduction

In this paper we consider a free boundary problem for a nonlinear parabolic partial differential equation. In particular, we are concerned with the inverse problem, which means we know the behavior of the free boundary *a priori* and would like a solution, *e.g.* a convergent series, in order to determine what the trajectories of the system should be for steady-state to steady-state boundary control.

The classical Stefan problem models a column of liquid in contact at 0 degrees with an infinite strip of ice, as depicted in Figure 1. The problem is thoroughly explored in [1] and a catalogue of various problems reducing to problems of the Stefan type is given in [14]. We investigate a modified Stefan problem that includes a diffusion term and a nonlinear reaction term. This can be seen as a simple model of a chemically reactive and heat diffusive liquid surrounded by ice, as considered under a more general form in [2].

Define $(x, t) \mapsto u(x, t)$ as the temperature in the liquid and $t \mapsto y(t)$ as the position of the liquid/solid interface, given a position x and time t. The functions $h(t)$ and $\psi(x)$ are the temperatures at the fixed end $(x = 0)$ and at initial time $(t = 0)$, respectively. The nonlinear Stefan problem is to determine a $u(x, t)$ and $y(t)$,

Keywords and phrases: Inverse Stefan problem, flatness, motion planning.

[1] Control and Dynamical Systems, California Institute of Technology, Mail Code 107-81, 1200 E California Blvd., Pasadena, CA 91125, USA.

[2] Centre Automatique et Systèmes, École Nationale Supérieure des Mines de Paris, 60 boulevard Saint-Michel, 75272 Paris Cedex 06, France; e-mail: petit@cas.ensmp.fr

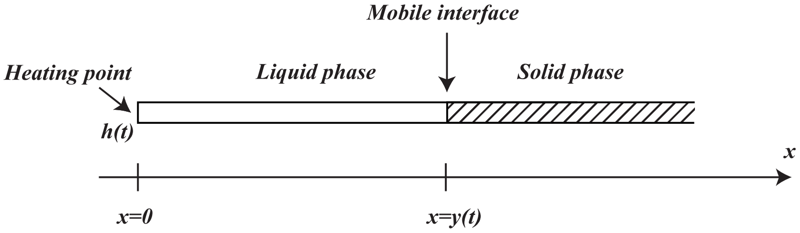


Figure 1. The system under consideration: liquid phase governed by a nonlinear heat equation with boundary control, in contact with an infinite solid phase.

given $h(t)$ and $\psi(x)$, that satisfy

$$\left.\begin{array}{ll} u_t = \ u_{xx} - \nu u_x - \rho u^2, & \forall (x,t) \in D_T, \\ u(0,t) = h(t) \geq 0, & 0 < t \leq T \\ u(x,0) = \psi(x) \geq 0, & 0 \leq x \leq y(0) \\ u(y(t),t) = 0, \ u_x(y(t),t) = -\dot{y}(t), & 0 < t \leq T \end{array}\right\} \tag{1}$$

where

$$D_T \equiv \{(x,t) \ : \ 0 < x < y(t), \ 0 < t \leq T\},$$

and the boundaries defined in the last three conditions are denoted in order as

$$B_T \equiv \{(0,t) : 0 < t \leq T\} \ \cup \ \{(x,0) : 0 \leq x \leq y(0)\} \ \cup \ \{(y(t),t) : 0 < t \leq T\} \equiv B_T^1 \cup B_T^2 \cup B_T^3.$$

The notation $\dot{y}(t)$ is the time derivative of $y(t)$ and ν, $\rho \geq 0$, $T > 0$. This model arises from a classical energy balance. The equation $u_x(y(t),t) = -\dot{y}(t)$ expresses the fact that all of the heat energy arriving at the liquid-solid interface is utilized in the melting process. In the model, the thermal conductivity coefficient and a parameter equal to the product of the solid density and the latent heat of fusion are normalized to one. Of course it is possible to use different values for these coefficients using changes of scales for x, t and u, as described in [1] (p. 282).

As in [1], we refer to D_T and B_T as the parabolic interior and parabolic boundary, respectively. Figure 2 gives a graphical 2-D representation of the interior and boundary. The *inverse* problem and its solution are stated here as a definition.

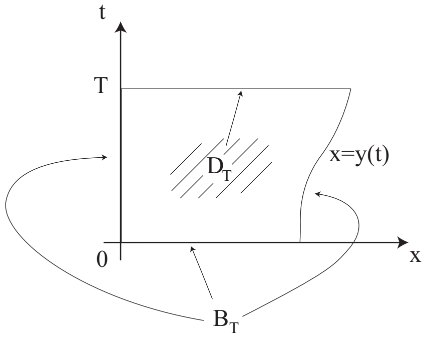


Figure 2. Picture of parabolic interior and boundary for free boundary problem.

Definition 1.1. A solution of (1) for a known function $y \in C^\infty[0,T]$, with all derivatives known, is a function $u = u(x,t)$ defined in $D_T \cup B_T$ such that u_{xx}, $u_t \in C(D_T)$, u is bounded, satisfies the conditions of (1) and $u \in C(D_T \cup B_T)$.

The inverse problem is precisely a non-characteristic Cauchy problem with Cauchy data: $u(y(t),t) = 0$, $u_x(y(t),t) = -\dot{y}(t)$ [7]. Given $y(t)$ and a solution u to the inverse problem, we have the initial profile $\psi(x)$ and the boundary control $h(t)$, both of which must be non-negative according to (1). Physically, the temperature everywhere in the liquid column should be non-negative; we will return to this point in the latter sections of the paper.

The flatness approach [3, 4] for partial differential equations are a means of solving inverse problems. The explicit parametrization of the trajectories of the systems is the key to straightforward motion planning strategies that can incorporate optimization [11, 12].

Recent work on these parametrization include [9,10] where approximate controllability of any initial condition to steady state is explored. The series expansion techniques used for heat equations have been used since the work of Gevrey, while inverse problems have been addressed as early as the work by Hill [7]. Specifically, Hill gives a complete solution to the inverse Stefan problem with a linear heat equation, *i.e.* $u_t = u_{xx}$. Recently a nonlinear heat equation over a constant spatial domain, with a quadratic reaction term, was examined by Rudolph and Lynch [10].

In this paper we combine the two issues: the free boundary (Stefan) problem with a quadratic nonlinearity. Using the work in [10] as a starting point, we prove convergence of a series solution. Then a detailed parametric study on the series radius of convergence is carried out. Moreover, we prove that the parametrization can indeed can be used for motion planning purposes; computation of the open loop motion planning is straightforward. Simulation results are given and we prove some important properties about the solution. Namely, a weak maximum principle is derived for the equation in (1), stating that the maximum is on the boundary. Also, we prove asymptotic positiveness of the solution, a physical requirement over the entire domain, as the transient time from one steady-state to another gets large.

The Stefan problem we consider is a first step towards a more complex problem for multidimensional reaction-diffusion systems investigated in [5], where three chemical species balance equations contain second order reaction terms. Extension of our approach to a fourth order radiation term, *i.e.* the Stefan–Boltzmann condition, is a subject for future work. A model with this condition arises in crystal growth furnaces, where a solid phase is surrounded by an infinite liquid phase. Note that in our work, either phase may be modelled with equivalent computations (after the appropriate sign changes). We feel the work presented here can highlight some of the difficulties and challenges of problems that arise in motion planning for crystal growth models.

2. Series solutions and convergence

2.1. Outline

In Section 2.2 we establish a lower bound on the radius of convergence of a series solution, denoted η^*, that depends upon the physical constants of the model ρ and ν. The radius η^* also depends upon M and R, the Gevrey constants of the function $\dot{y}(t)$. The definition of a Gevrey function is given implicitly in Theorem 2.1 (and also in [8]), from which it is clear that the associated constants (M and R in this case) characterize the aggressiveness of the trajectories of the system. In Section 2.3 we make use of several lemmas to construct parametric expressions for lower bounds on the radius. Specifically, Lemma 2.4 allows us to bound η^* from below by a unique root to a quartic polynomial that depends only upon ρ, ν, R, where the root is compared to M directly. In Lemma 2.5 we use a convexity argument to bound the root from Lemma 2.4 by an analytic expression that also depends only upon ρ, ν, R. The two lemmas are then combined in Theorem 2.6 to give the main result regarding convergence of the series solution; namely, the radius of convergence is bounded by an analytic expression.

2.2. Series solution

Consider the series solution

$$u(x,t) = \sum_{n=0}^{\infty} \frac{a_n(t)}{n!}[x - y(t)]^n. \tag{2}$$

For the solution (2) to satisfy

$$u_t = \ u_{xx} - \nu u_x - \rho u^2, \forall (x,t) \in D_T,$$

the $a_n(t)$ coefficients must satisfy the recurrence equation

$$a_n = \dot{a}_{n-2} - a_{n-1}\dot{y} + \nu a_{n-1} + \rho \sum_{k=0}^{n-2} \binom{n-2}{k} a_{n-2-k}\ a_k, \quad n \geq 2, \tag{3}$$

with $a_0 = 0$ (arising from the $u(y(t),t) = 0$ condition) and $a_1 = -\dot{y}$ (arising from the Stefan condition $-u_x(y(t),t) = \dot{y}(t)$). From (3) it is clear that given $y(t)$, all the series coefficients $a_n(t)$ and therefore the temperature $u(x,t)$ and boundary conditions $h(t)$ and $\psi(x)$ are uniquely determined.

By majorizing the series in (2), we will prove that this solution converges absolutely. We now state the first of two main theorems in the paper. The proof makes use of two lemmas stated in the Appendix.

Theorem 2.1. *Given that $\dot{y} \in C^\infty[0,T]$ is Gevrey of order $(\alpha - 1)$ for $1 \leq \alpha \leq 2$, i.e.*

$$\exists M, R > 0 \quad \textit{such that} \quad \left| y^{(l+1)}(t) \right| \leq M \frac{l!^\alpha}{R^l}, \ \forall\ l = 0,1,2,..., \forall t \in [0,T]$$

the radius of convergence of the series has as a lower bound the unique positive root $\eta = \eta^*$ of the polynomial

$$\left(\frac{\rho M}{2}\right)\eta^3 + \left(\frac{1}{R}\right)\eta^2 + \left(\frac{\nu + M}{2}\right)\eta - 1 = 0. \tag{4}$$

Proof. By induction on n, we prove that for all $n = 0,1,2,...$, the coefficients satisfy the bound

$$\left| a_n^{(l)}(t) \right| \leq \frac{MA^{n-1}}{R^{l+n-1}} \frac{(l+n)!^\alpha}{n!^{\alpha-1}}, \ \forall\ l = 0,\ 1,\ 2,\ ..., \tag{5}$$

for some $A > 0$. The coefficient $a_0 = 0$ satisfies (5) trivially and we examine $n = 1$ as the base case, since the recurrence is defined for $n \geq 2$. Namely, for $a_1 = -\dot{y}$,

$$\left| a_1^{(l)}(t) \right| = \left| y^{(l+1)}(t) \right| \leq M \frac{l!^\alpha}{R^l} \leq \frac{M}{R^l}\ l!^\alpha (l+1)^\alpha,$$

and the last inequality is strict when $l > 0$. By inductive hypothesis, we assume now that (5) holds for all $i = 0,1,...,n-1$ and prove that it must also hold for $i = n$. Taking an absolute value and l time derivatives of (3), after the triangle inequality we have

$$\left| a_n^{(l)} \right| \leq \left| a_{n-2}^{(l+1)} \right| + \nu \left| a_{n-1}^{(l)} \right| + \sum_{m=0}^{l} \binom{l}{m} \left| a_{n-1}^{(l-m)} \right| \left| y^{(m+1)} \right| + \rho \sum_{k=0}^{n-2} \sum_{r=0}^{l} \binom{n-2}{k} \binom{l}{r} \left| a_{n-2-k}^{(r)} \right| \left| a_k^{(l-r)} \right|. \tag{6}$$

The first two terms in (6) can be majorized using (5) as

$$\begin{aligned} \left| a_{n-2}^{(l+1)} \right| &\leq \frac{MA^{n-3}}{R^{l+n-2}} \frac{(l+n-1)!^\alpha}{(n-2)!^{\alpha-1}} &= \frac{MA^{n-1}}{R^{l+n-1}} \frac{(l+n)!^\alpha}{n!^{\alpha-1}} \left[\frac{R}{A^2} \frac{(n(n-1))^{\alpha-1}}{(l+n)^\alpha} \right], \\ \nu \left| a_{n-1}^{(l)} \right| &\leq \nu \frac{MA^{n-2}}{R^{l+n-2}} \frac{(l+n-1)!^\alpha}{(n-1)!^{\alpha-1}} &= \frac{MA^{n-1}}{R^{l+n-1}} \frac{(l+n)!^\alpha}{n!^{\alpha-1}} \left[\nu \frac{R}{A} \frac{n^{\alpha-1}}{(l+n)^\alpha} \right]. \end{aligned}$$

The third term in (6) is majorized as

$$\begin{aligned}\sum_{m=0}^{l}\binom{l}{m}\left|a_{n-1}^{(l-m)}\right|\left|y^{(m+1)}\right| &\le \sum_{m=0}^{l}\binom{l}{m}\frac{M^2A^{n-2}}{R^{l+n-m-2}R^m}\frac{(l+n-m-1)!^\alpha m!^\alpha}{(n-1)!^{\alpha-1}}\\ &= \frac{MA^{n-1}}{R^{l+n-1}}\frac{(l+n)!^\alpha}{n!^{\alpha-1}}\left[\frac{MR}{A}\frac{n!^{\alpha-1}}{(l+n)!^\alpha}\frac{1}{(n-1)!^{\alpha-1}}\sum_{m=0}^{l}\binom{l}{m}(l+n-m-1)!^\alpha m!^\alpha\right].\end{aligned}$$

Using Lemma A.1 and Lemma A.2, we can bound the term

$$\sum_{m=0}^{l}\binom{l}{m}(l+n-m-1)!^\alpha m!^\alpha \le \left[\sum_{m=0}^{l}\binom{l}{m}(l+n-m-1)!m!\right]^\alpha = \left[\frac{(n-1)!(n+l)!}{n!}\right]^\alpha,$$

resulting in

$$\sum_{m=0}^{l}\binom{l}{m}\left|a_{n-1}^{(l-m)}\right|\left|y^{(m+1)}\right| \le \frac{MA^{n-1}}{R^{l+n-1}}\frac{(l+n)!^\alpha}{n!^{\alpha-1}}\left[\frac{MR}{A\,n}\right].$$

The fourth (nonlinear) term in (6) is majorized as

$$\begin{aligned}\rho\sum_{k=0}^{n-2}\sum_{r=0}^{l}\binom{n-2}{k}\binom{l}{r}\left|a_{n-2-k}^{(r)}\right|\left|a_{k}^{(l-r)}\right| &\le \rho\sum_{k=0}^{n-2}\sum_{r=0}^{l}\binom{n-2}{k}\binom{l}{r}\frac{M^2A^{n-4}}{R^{l+n-4}}\frac{(n+r-k-2)!^\alpha(l+k-r)!^\alpha}{(n-k-2)!^{\alpha-1}k!^{\alpha-1}}\\ \le \frac{MA^{n-1}}{R^{l+n-1}}\frac{(l+n)!^\alpha}{n!^{\alpha-1}}&\left[\frac{\rho MR^3}{A^3}\frac{n!^{\alpha-1}}{(l+n)!^\alpha}\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{1}{(n-k-2)!^{\alpha-1}k!^{\alpha-1}}\left\{\sum_{r=0}^{l}\binom{l}{r}(n+r-k-2)!(l+k-r)!\right\}^\alpha\right],\end{aligned}$$

where the last inequality makes use of Lemma A.2. Using Lemma A.1 we have

$$\begin{aligned}&\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{1}{(n-k-2)!^{\alpha-1}k!^{\alpha-1}}\left\{\sum_{r=0}^{l}\binom{l}{r}(n+r-k-2)!(l+k-r)!\right\}^\alpha\\ &\quad=\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{1}{(n-k-2)!^{\alpha-1}k!^{\alpha-1}}\left\{\frac{k!(n-k-2)!(n+l-1)!}{(n-1)!}\right\}^\alpha\\ &\quad=\sum_{k=0}^{n-2}\binom{n-2}{k}(n-k-2)!\;k!\left\{\frac{(n+l-1)!}{(n-1)!}\right\}^\alpha=\sum_{k=0}^{n-2}(n-2)!\left\{\frac{(n+l-1)!}{(n-1)!}\right\}^\alpha\\ &\quad=(n-1)(n-2)!\left\{\frac{(n+l-1)!}{(n-1)!}\right\}^\alpha=\frac{(n+l-1)!^\alpha}{(n-1)!^{\alpha-1}},\end{aligned}$$

resulting in

$$\begin{aligned}\rho\sum_{k=0}^{n-2}\sum_{r=0}^{l}\binom{n-2}{k}\binom{l}{r}\left|a_{n-2-k}^{(r)}\right|\left|a_{k}^{(l-r)}\right| &\le \frac{MA^{n-1}}{R^{l+n-1}}\frac{(l+n)!^\alpha}{n!^{\alpha-1}}\left[\frac{\rho MR^3}{A^3}\frac{n!^{\alpha-1}}{(l+n)!^\alpha}\frac{(n+l-1)!^\alpha}{(n-1)!^{\alpha-1}}\right]\\ &= \frac{MA^{n-1}}{R^{l+n-1}}\frac{(l+n)!^\alpha}{n!^{\alpha-1}}\left[\frac{\rho MR^3}{A^3\,n}\left(\frac{n}{n+l}\right)^\alpha\right].\end{aligned}$$

Collecting the terms for (6) and noticing that for $n \geq 1$, $l \geq 0$, and $\alpha \geq 0$, $\left[\frac{n}{n+l}\right]^{\alpha} \leq 1$ holds, we have

$$\left|a_n^{(l)}\right| \leq \frac{MA^{n-1}}{R^{l+n-1}} \frac{(l+n)!^{\alpha}}{n!^{\alpha-1}} \left[\frac{R}{A^2}\frac{(n-1)^{\alpha-1}}{n} + \frac{(\nu+M)R}{A\,n} + \frac{\rho M R^3}{A^3\,n}\right] \cdot$$

The terms in the square brackets are bounded as

$$\max_{n\geq 2,\ \alpha\in[1,2]} \frac{(n-1)^{\alpha-1}}{n} = \left.\frac{(n-1)^1}{n}\right|_{(n\geq 2)} \leq 1, \quad \max_{n\geq 2} \frac{1}{n} = \frac{1}{2} \cdot$$

With these bounds, we have

$$\left|a_n^{(l)}\right| \leq \frac{MA^{n-1}}{R^{l+n-1}} \frac{(l+n)!^{\alpha}}{n!^{\alpha-1}} \left[\frac{1}{R}\left(\frac{R}{A}\right)^2 + \frac{(\nu+M)}{2}\left(\frac{R}{A}\right) + \frac{\rho M}{2}\left(\frac{R}{A}\right)^3\right] \cdot$$

Given $(M,\ R,\ \rho,\ \nu)$, the parameter A is chosen such that

$$\left[\frac{1}{R}\left(\frac{R}{A}\right)^2 + \frac{(\nu+M)}{2}\left(\frac{R}{A}\right) + \frac{\rho M}{2}\left(\frac{R}{A}\right)^3\right] = 1, \tag{7}$$

implying that (5) is proven by induction. Using (5) and the *Cauchy–Hadamard formula*, the radius of convergence is given by

$$\frac{1}{\limsup_{n\to\infty} |a_n/n!|^{1/n}} \geq \left[\lim_{n\to\infty} \left|\frac{MA^{n-1}}{R^{n-1}}\right|^{1/n}\right]^{-1} = \lim_{n\to\infty} \frac{R}{A}\left[\frac{A}{MR}\right]^{1/n} = \frac{R}{A} \cdot$$

Denoting this lower bound on the radius of convergence as $\eta \equiv R/A$ and substituting into (7) yields (4). Existence and uniqueness of the positive root $\eta = \eta^*$ are now proven. First, given $(M,\ R,\ \nu,\ \rho) > 0$ define the positive, analytic and strictly increasing function $\eta \mapsto f(\eta)$ as

$$f(\eta) = \left(\frac{\rho M}{2}\right)\eta^3 + \left(\frac{1}{R}\right)\eta^2 + \left(\frac{\nu+M}{2}\right)\eta. \tag{8}$$

The positive root η^* of the equation $f(\eta^*) - 1 = 0$ exists and is unique since $(f(\cdot) - 1)(\eta)$ is analytic and strictly increases from -1 to $+\infty$ as η grows from 0 to $+\infty$. □

Remark 2.2. We give here analytic expressions of the first five coefficients of the series (2) so one can see how the successive derivatives of y appear

$$\begin{aligned}
a_1 &= -\dot{y}\\
a_2 &= -\dot{y}(\nu + \dot{y})\\
a_3 &= -\ddot{y} + \dot{y}^3 - \nu^2\dot{y}\\
a_4 &= -\ddot{y}(2\nu + \dot{y}) - \dot{y}^4 + \nu\dot{y}^3 + (\nu^2 + 2\rho)\dot{y}^2 - \nu^3\dot{y}\\
a_5 &= -y^{(3)} - 3\nu^2\ddot{y} + \dot{y}^5 - 2\nu\dot{y}^4 + 4\rho\dot{y}^3 + \left(4\ddot{y} + 2\nu(\nu^2 + 4\rho)\right)\dot{y}^2 + (\nu\ddot{y}^2 - \nu^4)\dot{y}.
\end{aligned}$$

2.3. Parameterizations of radius of convergence

This section is concerned with constructing parametric lower bounds on η^*. We first derive by an easy calculation a lower bound that is suitable for most values of the physical parameters ρ and ν. This bound is then complemented with another lower bound that is more tedious to derive but is less conservative for specific values of the physical parameters, namely when ρ is large and ν is small.

The first bound is achieved using the following lemma:

Lemma 2.3. *For all a, b, c strictly positive real parameters, the unique positive root η^0 of*

$$a\eta^3 + b\eta^2 + c\eta - 1 = 0$$

is lower bounded by

$$\frac{-c + \sqrt{c^2 + 4(a/c + b)}}{2(a/c + b)}.$$

Proof. The function $\eta \mapsto a\eta^3 + b\eta^2 + c\eta - 1$ is analytic and strictly increases from -1 to $+\infty$ as η grows from 0 to $+\infty$. Define $h_1(\eta) = a\eta^3 + b\eta^2$, $h_2(\eta) = 1 - c\eta$. The graphs of h_1 and h_2 intersect at η^0. Since $h_1 > 0$ on $]0, +\infty[$ it is clear that $\eta^0 < 1/c$.

On $]0, 1/c[$ it is easy to check that $h_1(\eta) < (a/c + b)\eta^2$. On this interval h_1 is a strictly increasing function while h_2 is strictly decreasing. Let $\hat{\eta}$ be the unique positive root of $(a/c+b)\eta^2 = 1 - c\eta$ we get that $h_1(\hat{\eta}) < h_2(\hat{\eta})$ which yields $\hat{\eta} < \eta^0$. Finally

$$\eta^0 > \frac{-c + \sqrt{c^2 + 4(a/c + b)}}{2(a/c + b)}. \qquad \square$$

When $a = \rho M/2$, $b = 1/R$ and $c = (\nu + M)/2$, it is clear that η^0 corresponds to η^*.

The second lower bound is derived from the following lemmas:

Lemma 2.4. *Given R, ν, $\rho > 0$, define the function $(\eta, M) \mapsto f(\eta, M)$ as*

$$f(\eta, M) = \left(\frac{\rho M}{2}\right)\eta^3 + \left(\frac{1}{R}\right)\eta^2 + \left(\frac{\nu + M}{2}\right)\eta. \tag{9}$$

Let M^* be the unique positive root of $f(M^*, M^*) = 1$. Then, for any given M with $0 < M \leq M^*$, the root η^* of $f(\eta^*, M) = 1$ satisfies $\eta^* \geq M^*$.

Proof. The positive root M^* of the equation $f(M^*, M^*) - 1 = 0$ exists and is unique since $(f(\cdot) - 1)(M, M)$ is analytic and strictly increases from -1 to $+\infty$ as M grows from 0 to $+\infty$. Since f is a strictly increasing function of η and M, $f(\eta^*, M) = 1 = f(M^*, M^*)$ with $M \leq M^*$ implies that $M^* \leq \eta^*$. $\square$

From the lemma, $M = M^*$ is the unique positive root of the polynomial

$$\left(\frac{\rho}{2}\right)M^4 + \frac{1}{2}\left(\frac{R+2}{R}\right)M^2 + \left(\frac{\nu}{2}\right)M - 1 = 0, \tag{10}$$

giving a lower bound on the radius of convergence provided $M \leq M^*$. The next lemma gives a strict lower bound on M^* using a convexity argument. Defining the functions

$$g_1(M) = aM^4 + bM^2, \quad g_2(M) = 1 - cM, \quad \text{where } a = \frac{\rho}{2},\ b = \frac{1}{2}\left(\frac{R+2}{R}\right),\ c = \frac{\nu}{2},$$

and it is clear that $g_1(M^*) = g_2(M^*)$. The functions, their intersection point M^* and the line cdM are shown in Figure 3, where $d = g_1(1/c) > 0$.

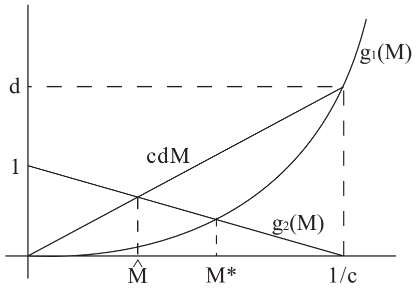


Figure 3. Illustration of convexity argument for strict lower bound on M^*.

The point $\hat{M}$ satisfies $cd\hat{M} = g_2(\hat{M})$, so after some algebra we have that

$$d = \frac{a+c^2b}{c^4}, \quad \hat{M} = [c(d+1)]^{-1} = \left[c + bc^{-1} + ac^{-3}\right]^{-1}.$$

Substituting in for the parameters and treating $\nu \mapsto \hat{M}(\nu)$ as a function given $(R,\ \rho)$, we have

$$\hat{M}(\nu) = \left[\frac{\nu}{2} + \frac{4\rho}{\nu^3} + \frac{R+2}{\nu R}\right]^{-1}. \tag{11}$$

Lemma 2.5. *Given* $R,\ \rho > 0$ *and defining* $\hat{\nu} = \arg\max_{\nu>0} \hat{M}(\nu)$, *we have the following strict bounds on* M^* *defined in Lemma 2.4:*

$$\begin{aligned} M^* &> \hat{M}(\nu), \quad \textit{for all } \nu \in [\hat{\nu}, +\infty[\\ M^* &> \hat{M}(\hat{\nu}), \quad \textit{for all } \nu \in [0, \hat{\nu}[\ . \end{aligned}$$

Proof. We first compute $\hat{\nu}$ by taking the derivative of $\hat{M}$ with respect to ν, yielding

$$\frac{\mathrm{d}\hat{M}(\nu)}{\mathrm{d}\nu} = -\hat{M}(\nu)^2 \left[\frac{1}{2} - \frac{12\rho}{\nu^4} - \frac{R+2}{\nu^2 R}\right].$$

Over the domain $\nu \in]0, +\infty[$, $\hat{M}(\nu) > 0$ so setting the derivative to zero to find the extremum requires the bracketed expression to be zero, which is equivalently

$$\nu^4 - \frac{2(R+2)}{R}\nu^2 - 24\rho = 0.$$

The unique positive solution to this squared-quadratic is

$$\hat{\nu} = \left[\left(\frac{R+2}{R}\right) + \left[\left(\frac{R+2}{R}\right)^2 + 24\rho\right]^{1/2}\right]^{1/2}. \tag{12}$$

The second derivative at $\hat{\nu}$ is negative, verifying that the extremum is indeed a maximum. Observe that for all $R > 0$ and $\rho \geq 0$, $\hat{\nu} \geq \sqrt{2}$. Also, $\hat{\nu} = \sqrt{2}$ if and only if $(R, \rho) = (+\infty, 0)$. For a given $(R, \rho) > 0$, $g_1(M)$ is a strictly convex function, *i.e.*

$$cdM > g_1(M), \quad \forall M \in]0, 1/c[.$$

It is clear that M^* and $\hat{M}$ are in the domain $]0, 1/c[$. By the convexity of g_1 then,

$$cdM^* > g_1(M^*) = 1 - cM^* \;\Rightarrow\; M^* > [c(d+1)]^{-1} = \hat{M}(\nu),$$

which holds for all $\nu \in]0, +\infty[$. The trouble with $\hat{M}(\nu)$ as a bound for M^* over $\nu \in [0, \hat{\nu}[$ is that as $\nu \to 0$, $\hat{M}(\nu) \to 0$ while M^* is increasing up to the solution of the equation $g_2(M) = 1$. Since we know $M^* > \hat{M}(\nu)$ for $\nu = \hat{\nu}$ and M^* is increasing as ν decreases from $\hat{\nu}$ to 0, we can (conservatively) choose the lower bound on M^* to remain at the maximum $\hat{M}(\hat{\nu})$ for all $\nu \in [0, \hat{\nu}[$. □

We now give the main analytic result of the paper regarding convergence of the proposed solution, making use of the two lower bounds derived above.

Theorem 2.6. *Given $\nu, \rho > 0$ and assuming that $\dot{y} \in C^\infty[0, T]$ is Gevrey of order $(\alpha - 1)$ for $1 \le \alpha \le 2$, i.e.*

$$\exists M,\ R > 0 \quad \text{such that} \quad \left| y^{(l+1)}(t) \right| \le M \frac{l!^\alpha}{R^l},\ \forall\, l = 0,\ 1,\ 2,\ ...,$$

the radius of convergence of the series (2) *is greater than*

$$\hat{\eta} = \frac{-R(\nu+M)^2 + \sqrt{R^2(\nu+M)^4 + 16(\rho M R^2(\nu+M) + R(\nu+M)^2)}}{4(\rho R M + \nu + M)} \cdot \tag{13}$$

Furthermore, if $M^* \ge M$, M^* given by (10)*, then the radius of convergence of the series* (2) *is greater than the following quantities*

$$\left.\begin{array}{ll} \hat{M}(\nu), & \text{for all } \nu \in [\hat{\nu}, +\infty[\\ \hat{M}(\hat{\nu}), & \text{for all } \nu \in [0, \hat{\nu}[\end{array}\right\} \tag{14}$$

where $\hat{M}$, $\hat{\nu}$ are given by (11)*,* (12)*, respectively.*

Proof. We get the first lower bound (13) from Lemma 2.3 where we make $a = \rho M/2$, $b = 1/R$, $c = (\nu + M)/2$.

From Theorem 2.1, η^* is a lower bound on the radius of convergence. Since $M^* \ge M$ by assumption we can apply Lemma 2.4, and with Lemma 2.5 we have

$$\eta^* \ge M^* > \hat{M},$$

giving the stated result in (14). □

Remark 2.7. Let us here characterize the different lower bounds on the radius of convergence, given the parameter values ρ, ν, M and R. Calculating η^* numerically will of course result in the least conservative bound. Of the two analytic lower bounds, equation (13) is less conservative than (14) for most values of the physical parameters ρ and ν. Only when ρ is large and $\nu < \hat{\nu}$ does (14) become less conservative. Specifically, in that case, $\hat{\eta} \propto \rho^{-1/2}$ and $\hat{M} \propto \rho^{-1/4}$ and thus $\hat{\eta}$ approaches zero faster than $\hat{M}$ as ρ approaches infinity. Numerical comparisons between these bounds are given in the simulation studies in Section 3.2.

3. Properties of solution application

The output domain is $y : [0, T] \to \mathbb{R}$ and we define $y(t) = \phi(t/T)$. Then, given $\phi : [0, 1] \to \mathbb{R}$ with Gevrey bounds

$$\exists M_\phi, R_\phi > 0 \quad \text{such that} \quad \left| \phi^{(l+1)}(t) \right| \le M_\phi \frac{l!^\alpha}{R_\phi^l},\ \forall\, l = 0,\ 1,\ 2,\ ..., \tag{15}$$

and observing that

$$y^{(l)}(t) = \frac{1}{T^l}\phi^{(l)}(t/T),$$

the Gevrey bounds on $y(t)$ become,

$$\left|y^{(l+1)}(t)\right| \leq \frac{M_\phi}{T}\frac{l!^\alpha}{(R_\phi T)^l} = M\frac{l!^\alpha}{R^l}, \quad M \equiv M_\phi/T, \ R \equiv R_\phi T.$$

Thus, given the function $\phi : [0,1] \to \mathbb{R}$, the conditions of Theorem 2.6 can be checked by substituting $M = M_\phi/T$ and $R = R_\phi T$.

In the following, we give numerical simulations and discuss theoretical properties about the results. The time bound parameter T is shown to be important for guaranteeing desirable properties of the solution. Intuition about the physical problem suggests that smaller T will require more aggressive trajectories for steady-state to steady-state boundary control. Using the recurrence relation, we prove that as T grows we can approximate the series solution to second order by an analytic expression. The expression guarantees that the temperature profile in the liquid column remains above the thawing point of ice (0 degrees which is the temperature of the interface).

3.1. Physical aspects of solution

The classical Stefan problem assumes the given initial profile satisfies $\psi(x) \geq 0$ in B_T^2 and the boundary control satisfies $h(t) \geq 0$ in B_T^1. In this case, from a Weak Maximum/Minimum Principle, the temperature in D_T never exceeds or drops below the temperature of the column on B_T (see Sect. 1.6 in [1]). Thus, it is a given that the temperature everywhere is non-negative since $u \geq 0$ in B_T.

For the inverse nonlinear problem, we are faced with guaranteeing that our solution satisfies this physical requirement that the temperature in the liquid column never drops below the freezing point. The specific Gevrey function chosen for $\dot{y}(t)$ will be required to satisfy certain properties. In Section 17.2 of [1], $y(t)$ is shown to be an increasing function and strictly increasing if ψ or h are nonzero in every neighborhood of $t = 0$. Of course, we cannot expect a converse-like result without a similar assumption imposed on $y(t)$. Specifically we assume that given any $\epsilon \in]0, 1/2[$,

$$\dot{y}(t) > 0, \ \forall t \in [T\epsilon, T(1-\epsilon)], \tag{16}$$

and $\dot{y}(t) = 0 \iff t = 0$ or $t = T$. Moreover, for steady-state behavior we assume that

$$y(0) = L, \ y(T) = L + \Delta L, \quad \text{where } L, \Delta L \in \mathbb{R}^+, \quad \text{and } y^{(m)}(0) = y^{(m)}(T) = 0, \ \forall m = 1, \ 2, \ 3, \ ...$$

From (3), $a_n(0) = a_n(T) = 0$, for all $n = 0, \ 1, \ 2, \ ...$ As a result, $u(x, 0) = u(x, T) = 0$. So $u = 0$ on the open line $\partial D_T \equiv \{(x, T) : 0 < x < y(T)\}$ and on B_T^2 and B_T^3.

Proving that the temperature satisfies $u(x, t) \geq 0$ for all (x, t) in $D_T - \partial D_T$ and for all t in B_T^1 would thus assure that $u(x, t) \geq 0$ in $D_T \cup B_T$. Given a Weak Minimum Principle, we could assure the same if we proved $u(x, t) \geq 0$ in B_T^1. We do not establish a Weak Minimum Principle here. Instead we derive a Weak Maximum Principle, that serves as a sanity check for numerical experiments (*i.e.* the interior temperature should never exceed the maximum boundary temperature) and focus on a detailed analytic study. It is shown that approximate steady-state to steady-state boundary control can guarantee non-negativeness of the temperature in the entire domain.

3.2. Numerical simulations

For practical purposes of course the series solution (2) is truncated for implementation. Specifically, for $y(t)$ defined by the function given in Appendix B, we here take the first 10 terms to approximate the solution $u(x, t)$.

Steady-state to steady-state simulations are considered and so, as shown above, $\psi \equiv 0$. In Figure 4 we show the temperature profile for $T = 100$, $\rho = 1.5$, $\nu = 0.5$, $L = 1$, and $L + \Delta L = 2$. For this case, we compute $\eta^* = 2.1849$, which guarantees convergence of the desired domain. To give a case where the more conservative, analytic lower bounds (13) and (14) in Theorem 2.6 guarantee convergence, see Figure 5. For this temperature profile, the parameters are $T = 100$, $\rho = 1.2$, $\nu = 0.5$, $L = 0.25$, and $L + \Delta L = 0.5$. The analytic expressions yield $\hat{\eta} = 2.5165$ and $\hat{M} = 0.50302$. For this case, we also computed $\eta^* = 2.586$ showing the conservatism of (14) quantitatively.

The lower bounds on the radius of convergence can also be compared analytically as T becomes large. In that case, η^* and $\hat{\eta}$ approach $2/\nu$, while $\hat{M}$ approaches $(\nu/2 + 4\rho/\nu^3 + 1/\nu)^{-1}$, again displaying the relative conservatism of (14). In both simulations, the time bound T was chosen large enough such that the temperature in the column remained non-negative. An asymptotic analysis of how large T needs to be to ensure non-negativity is detailed in Section 3.4.

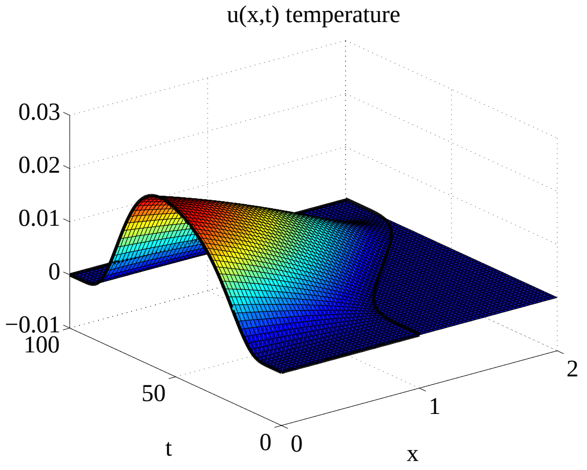


FIGURE 4. Temperature profile for transition from column length 1 to 2.

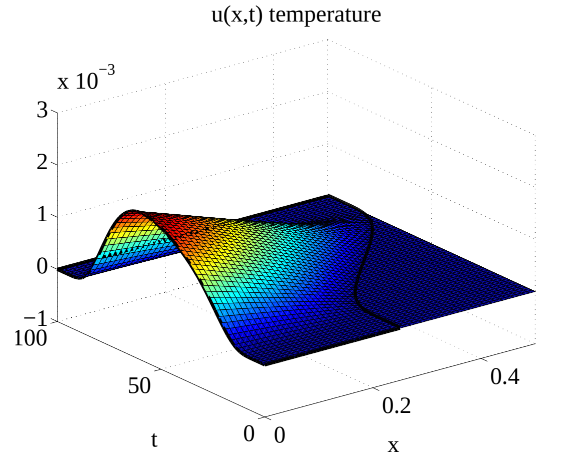


FIGURE 5. Temperature profile for transition from column length 0.25 to 0.5.

3.3. The Weak Maximum Principle (WMP)

A powerful tool for the study of linear initial-boundary-value problems is the Weak Maximum Principle [1]. We now state this tool as it applies to (1).

Theorem 3.1. *For a solution u of $u_t = u_{xx} - \nu u_x - \rho u^2$ in a bounded D_T, with $\nu, \rho \geq 0$, which is continuous in $D_T \cup B_T$,*

$$\max_{D_T \cup B_T} u = \max_{B_T} u.$$

Proof. The proof follows along the lines of the proof of the Weak Maximum Principle (Th. 1.6.1) for $u_t = u_{xx}$ given in [1]. From Definition 1.1, the value of u is bounded in $D_T \cup B_T$ and we denote this bound $|u| \leq U$. Define

$$v(x,t) = u(x,t) + \varepsilon \exp[W(x - C)]$$

over $D_T \cup B_T$, where $\varepsilon,\ W > 0$ and $|y(t)| \leq C \equiv L + \Delta L$. Then v assumes its maximum on B_T. Otherwise, there exists a point $(x_0, t_0) \in D_T$ such that

$$\max_{D_T \cup B_T} v = v(x_0, t_0).$$

Hence at $(x_0, t_0) \in D_T$,

$$\mathcal{L}(v) \equiv v_{xx} - v_t - \nu v_x - \rho v^2 \leq 0,$$

since $0 < x < y(t),\ 0 < t \leq T$ in D_T and

$$\begin{aligned}
&v_{xx}(x_0,t_0) \leq 0 \text{ (at a maximum)},\\
&v_x(x_0,t_0) = 0 \text{ (at an extremum)},\\
&v_t(x_0,t_0) \geq 0 \text{ (at a maximum over a right-closed domain)},
\end{aligned}$$

where $v_t(x_0, t_0)$ may be positive in D_T only when $t_0 = T$. However, throughout D_T,

$$\begin{aligned}
\mathcal{L}(v) &= u_{xx} - u_t - \nu u_x - \rho \left\{u + \varepsilon \exp[W(x-C)]\right\}^2\\
&\quad + \varepsilon W^2 \exp[W(x-C)] - \nu \varepsilon W \exp[W(x-C)]\\
&= \varepsilon W(W-\nu) \exp[W(x-C)] - \rho \left\{2u\varepsilon \exp[W(x-C)] + \varepsilon^2 \exp[2W(x-C)]\right\}\\
&= \varepsilon \exp[W(x-C)] \left\{W(W-\nu) - \rho\left(2u + \varepsilon \exp[W(x-C)]\right)\right\}\\
&\quad > \varepsilon \exp[-WC] \left\{W(W-\nu) - \rho(2U + \varepsilon)\right\},
\end{aligned}$$

where the inequality follows from $0 < x < C$ in D_T, which implies $\exp(W(x-C)) \in]\exp(-WC), 1[$, and from choosing W large enough, namely such that

$$W(W - \nu) > \rho(2U + \varepsilon).$$

As a result, $\mathcal{L}(v) > 0$. Thus by contradiction v assumes its maximum value on B_T.

Since $u \leq v$ on $D_T \cup B_T$ and $v \leq \max_{B_T} v$,

$$u \leq \max_{B_T} v \leq \max_{B_T} u + \varepsilon \max_{B_T} \exp[W(x - C)].$$

As ε can be chosen arbitrarily small,

$$u \le \max_{B_T} u. \qquad \square$$

The Weak Maximum Principle is typically accompanied by its dual, the Weak Minimum Principle, as the latter principle can be directly derived from the former by replacing u with $(-u)$ in the heat equation. With a quadratic nonlinearity of course, $(-u)$ does not satisfy the same equation so we do not have a dual principle as directly. The simulation studies have indicated that there is a Weak Minimum Principle, but we do not pursue the theoretical version here. In the next section, to assure positivity of the temperature in the liquid column, we examine asymptotic behavior of the solution u as our time bound T is made large.

3.4. **Asymptotics of the solution**

In this section, it is shown that positivity of the temperature everywhere in the column can be guaranteed with as much resolution, up to the boundaries of the domain, as desired given a sufficiently large T.

Lemma 3.2. *Consider the problem* (1) *and series solution* (2)*, where the output* $y : [0,T] \to \mathbb{R}$ *is defined as* $y(t) = \phi(t/T)$*, given* $\phi : [0,1] \to \mathbb{R}$ *satisfying* (15)*. When* T *is large enough, the series is uniformly approximated to second order in* $(1/T)$ *by*

$$-\frac{\dot{y}(t)}{\nu}\left(\exp\left[\nu(x-y(t))\right]-1\right).$$

Proof. First, we prove by induction that the recurrence equation (3) can be expressed as

$$a_n = -\nu^{n-1}\dot{y} + \hat{F}_n(y^{(g(n))}, y^{(g(n)-1)}, ..., \dot{y}), \quad n \ge 1, \tag{17}$$

where $\hat{F}_n$ is a multivariate polynomial, in $\dot{y}$ and its time derivatives up to order $g(n)$, that has no constant terms and no terms affine in $\dot{y}$. It is also clear from the (3) that $\hat{F}$ is time-invariant, *i.e.* t does not appear explicitly. By induction it can be shown that $g(n) \in \mathbb{N}$ is given by $g(n) = 1 + \text{floor}((n-1)/2)$, which means that $g(n \pm 2) = g(n) \pm 1$. The base case $a_1 = -\dot{y}$, and the next term $a_2 = -\nu\dot{y} - \dot{y}^2$, satisfy (17). By inductive hypothesis we assume (17) holds for $i = 2, ..., n-1$ and show that it also holds for $i = n$. From (3) we have

$$\begin{aligned} a_n =& \frac{\mathrm{d}}{\mathrm{d}t}\left[-\nu^{n-3}\dot{y} + \hat{F}_{n-2}(y^{(g(n-2))}, ..., \dot{y})\right] + (\nu - \dot{y})\left[-\nu^{n-2}\dot{y} + \hat{F}_{n-1}(y^{(g(n-1))}, ..., \dot{y})\right] \\ &+ \rho\sum_{k=0}^{n-2}\binom{n-2}{k}\left[-\nu^{n-3-k}\dot{y} + \hat{F}_{n-2-k}(y^{(g(n-2-k))}, ..., \dot{y})\right]\left[-\nu^{k-1}\dot{y} + \hat{F}_k(y^{(g(k))}, ..., \dot{y})\right]. \end{aligned}$$

Given the properties on $\hat{F}_m$, for any $m = 1, ..., n-1$, we can rewrite it as

$$\hat{F}_m\left(y^{(g(m))}, ..., \dot{y}\right) = \dot{y}\,\hat{F}_m^1\left(y^{(g(m))}, ..., \dot{y}\right) + \hat{F}_m^2\left(y^{(g(m))}, ..., \ddot{y}\right),$$

where $\hat{F}_m^1$ and $\hat{F}_m^2$ have no constant terms. Observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{F}_m\left(y^{(g(m))}, ..., \dot{y}\right) = \ddot{y}\hat{F}_m^1\left(y^{(g(m))}, ..., \dot{y}\right) + \dot{y}\bar{F}_{m+2}^1\left(y^{(g(m+2))}, ..., \dot{y}\right) + \bar{F}_{m+2}^2\left(y^{(g(m+2))}, ..., \ddot{y}\right),$$

where $\bar{F}_{m+2}^i \equiv \mathrm{d}/\mathrm{d}t(\hat{F}_m^i)$, $i = 1, 2$. As $\hat{F}_m^i$ has no explicit dependence on t, $\bar{F}_{m+2}^i$ can have no constant terms. Thus, $\mathrm{d}/\mathrm{d}t(\hat{F}_m(y^{(g(m))}, ..., \dot{y}))$ has no constant terms or terms that are affine in $\dot{y}$, and we rewrite it as

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{F}_m\left(y^{(g(m))}, ..., \dot{y}\right) = \bar{F}_{m+2}\left(y^{(g(m+2))}, ..., \dot{y}\right).$$

Returning to the expression for a_n and applying the previous result for $m = n-2$ we have

$$\begin{aligned} a_n &= \left[-\nu^{n-3}\ddot{y} + \bar{F}_n\left(y^{(g(n))}, ..., \dot{y}\right)\right] - \nu^{n-1}\dot{y} + \nu^{n-2}\dot{y}^2 + (\nu - \dot{y})\,\hat{F}_{n-1}\left(y^{(g(n-1))}, ..., \dot{y}\right) \\ &\quad + \rho\sum_{k=0}^{n-2}\binom{n-2}{k}\left[-\nu^{n-3-k}\dot{y} + \hat{F}_{n-2-k}\left(y^{(g(n-2-k))}, ..., \dot{y}\right)\right]\left[-\nu^{k-1}\dot{y} + \hat{F}_k\left(y^{(g(k))}, ..., \dot{y}\right)\right] \\ &= \, -\nu^{n-1}\dot{y} + \hat{F}_n\left(y^{(g(n))}, ..., \dot{y}\right), \end{aligned}$$

where we have defined $\hat{F}_n$ as the collected terms and it is clear that $\hat{F}$ also has no terms affine in $\dot{y}$ or constant terms. Note that we do not claim that collecting terms in the last step does not result in the cancellation of terms. We only prove that it is not possible to generate new terms that are affine in $\dot{y}$ or constant. This concludes that induction and proves (17).

The next step is rewrite (17) with $\hat{F}_n$ as a function of the given output ϕ and its derivatives, as

$$a_n(t) = -\nu^{n-1}\dot{y}(t) + \hat{F}_n\left(\frac{\phi^{(g(n))}(t/T)}{T^{g(n)}}, ..., \frac{\dot{\phi}(t/T)}{T}\right), \quad n \geq 1,\ t \in [0, T].$$

Given the properties on the multivariate polynomial $\hat{F}$, the lowest order terms in $1/T$ that are possible are $\dot{y}^2$ and $\ddot{y}$, which means that as $T \to +\infty$, $\hat{F}_n = O(1/T^2)$ uniformly and

$$a_n(t) = -\nu^{n-1}\dot{y}(t) + O(1/T^2), \quad n \geq 1,\ t \in [0, T].$$

Using this as the recurrence equation for the series solution, we can write (2) as

$$\begin{aligned} u(x,t) &= -\frac{\dot{y}(t)}{\nu}\sum_{n=1}^{\infty}\frac{1}{n!}[\nu(x - y(t))]^n + O(1/T^2)\sum_{n=1}^{\infty}\frac{1}{n!}[x - y(t)]^n \\ &= -\frac{\dot{y}(t)}{\nu}\left(\exp\left[\nu(x - y(t))\right] - 1\right) + O(1/T^2)\left(\exp\left[x - y(t)\right] - 1\right) \\ &= -\frac{\dot{y}(t)}{\nu}\left(\exp\left[\nu(x - y(t))\right] - 1\right) + O(1/T^2), \end{aligned}$$

which concludes the proof. □

We can visualize the implications of Lemma 3.2 by looking at the a_n coefficients. Figure 6 shows the first ten coefficients for the parameter case study corresponding to Figure 4 with $T = 10$, in which case the temperature in the liquid does go negative. Figure 7 shows the coefficients for T increased to 100, corresponding exactly to the case in Figure 4. From the proof of the lemma, equation (17) implies that as T increases the coefficients approach $a_n \simeq -\nu^{n-1}\dot{y}(t) = -\nu^{n-1}\dot{\phi}(t/T)/T$. So as T increases, all of the coefficients approach the shape of $-\dot{y}$ (a negative definite, symmetric function) and decrease in amplitude, both trends observable from the figures.

The following result guarantees that up to arbitrarily small precision, we can achieve steady-state to state-state boundary control while maintaining a positive temperature in the entire liquid column, provided the upper bound on time T is large enough.

Lemma 3.3. *Consider the problem* (1) *and series solution* (2)*, where we take* $\nu > 0$*. The output* $y : [0, T] \to \mathbb{R}$ *is defined as* $y(t) = \phi(t/T)$*, given* $\phi : [0, 1] \to \mathbb{R}$ *satisfying* (15)*. Assume* ϕ *is a strictly increasing function, so* $y(t)$ *satisfies* (16)*. Given any* $1/2 > \varepsilon_t > 0$ *and* $L > \varepsilon_x > 0$*, there exists* $T_{\min} > 0$ *such that for all* $T > T_{\min}$ *the temperature satisfies*

$$u(x,t) > 0, \quad \forall \ \ (x,t) \in D_T^{\varepsilon} \cup B_T^{\varepsilon},$$

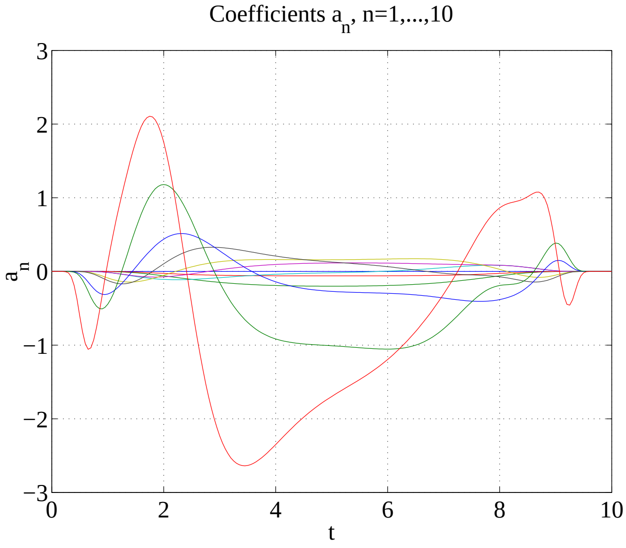


FIGURE 6. First ten a_n coefficients for $T = 10$.

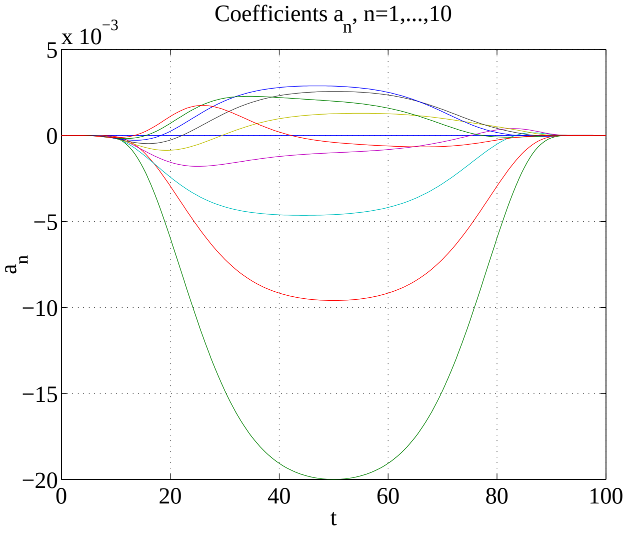


FIGURE 7. First ten a_n coefficients for $T = 100$.

where

$$\begin{aligned} D_T^\varepsilon \equiv & \{(x,t) \ : \ 0 < x < y(t) - \varepsilon_x, \ T\varepsilon_t < t \le T(1-\varepsilon_t)\} \\ B_T^\varepsilon \equiv & \{(0,t) : T\varepsilon_t < t \le T(1-\varepsilon_t)\} \\ & \cup \ \{(x, T\varepsilon_t) : 0 \le x \le y(T\varepsilon_t) - \varepsilon_x\} \\ & \cup \ \{(y(t) - \varepsilon_x, t) : T\varepsilon_t < t \le T(1-\varepsilon_t)\} \cdot \end{aligned}$$

A graphical 2-D representation of the ε interior and boundary $D_T^\varepsilon, \ B_T^\varepsilon$ is given in Figure 8.

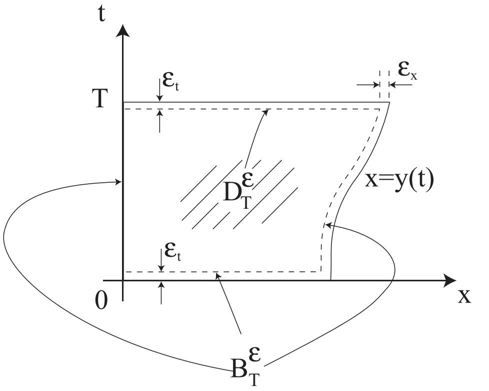


Figure 8. Picture of parabolic ε interior and boundary for Stefan problem.

Proof. By assumption,

$$\dot{\phi}(s) > 0, \ \forall s \in [\varepsilon_t, (1-\varepsilon_t)].$$

Define $\gamma \equiv \min_s \dot{\phi}(s) > 0$, with $s \in [\varepsilon_t, (1-\varepsilon_t)]$. Also, observe that

$$x - \phi(t/T) \leq -\varepsilon_x < 0, \quad \forall \ (x,t) \in D_T^\varepsilon \cup B_T^\varepsilon.$$

Now, we apply Lemma 3.2 to get

$$\begin{aligned} u(x,t) &= -\frac{\dot{\phi}(t/T)}{T\nu}\left(\exp\left[\nu(x-\phi(t/T))\right]-1\right) + O(1/T^2) \\ &\geq -\frac{\gamma}{T\nu}\left(\exp\left[-\nu\varepsilon_x\right]-1\right) + O(1/T^2), \quad \forall \ (x,t) \in D_T^\varepsilon \cup B_T^\varepsilon. \end{aligned}$$

Since $\gamma, \nu, \varepsilon_x > 0$, the term $-\gamma(\exp[-\nu\varepsilon_x]-1)/\nu$ is strictly positive. Therefore, there exists a $T_{\min}$ that depends on γ (*i.e.* on ε_t) and ε_x such that for $T > T_{\min}$, the positive term dominates and $u(x,t) > 0$ for all $(x,t) \in D_T^\varepsilon \cup B_T^\varepsilon$. □

For the simulation parameters used for the case of Figure 4, we numerically investigated values for ε_t as T increases, with the results given in Table 1. The values for ε_t reported are lower bounds, meaning we (conservatively) estimate the largest distance from the boundary at which negative temperatures occur. The distance is a lower bound for the following reason: if for a given ε_t we have a T that satisfies Lemma 3.3, for any $\varepsilon > \varepsilon_t$, the same T will result in a positive u in $D_T^\varepsilon \cup B_T^\varepsilon$. The trend of the table is that as ε_t decreases, T must increase, as implied by Lemma 3.3. From the table, the simulation corresponding to Figure 4 ($T = 100$) is guaranteed to have positiveness up to $\varepsilon_t = 0.058$.

Table 1. Estimation of ε_t parameter for Lemma 3.3.

T	5	10	50	100	500
ε_t	0.267	0.190	0.0851	0.0580	0.0217

4. Conclusions and future work

The problem treated in this paper includes two technical difficulties: the moving boundary and a quadratic reaction term. When combined, these issues make convergence substantially more difficult to study. We derived conservative results that can indeed be used in practice, as shown in the simulation section. We underlined that the solution we propose has to be used very carefully or the formal solution might not satisfy the physical requirement of the model (non-negativity of the temperature in the liquid phase).

An issue for future work is the study of approximate controllability of the model studied in this paper. Replacing the zero initial condition $u(x, 0) = 0$ by an arbitrary initial condition $u(x, 0) = \psi(x)$, is it still possible to steer the system approximately to zero in finite time, *i.e.* $u(x, T) \approx 0$?

Usually, such a result arises from a projection of the initial condition onto the formal series expression (refer to [9]). Straightforward conditions are thus available for formal series of the form $u(x,t) = \sum_{i=0}^{\infty} y^{(n)}(t)\frac{x^{2n}}{(2n)!}$ using the fact that the set of polynomials x^{2n}, $n \in \mathbb{N}$ is dense in the set of L^2 functions. In a simple manner, the conditions result in specified values for all the derivatives of y at the initial time 0.

In our case the series expansion is much different. Due to the moving boundary and the nonlinear effect, no simple identification between the (a_i) coefficients and the derivatives of y can be achieved. Moreover, both even and odd polynomials appear for most derivatives of y, ruling out classical density results, *e.g.* Stone–Weierstrass theorem [15]. All this makes the situation more convoluted and difficult to handle. Gathering terms it seems possible to derive solvable conditions in terms of the successive derivatives of y and some projections of the initial condition $\psi(x)$. This point is currently under investigation but we sketch an explicit procedure here.

Treating a non zero initial condition

Let us sketch here how one can derive the conditions on the successive derivatives of y. From the series solution (2) and remembering that $a_0 = 0$ from the boundary condition, one gets

$$\begin{aligned}\psi(x) = u(x,0) &= \sum_{n=1}^{\infty} \frac{a_n(0)}{n!}(x - y(0))^n \\ &= \sum_{n=1}^{\infty} a_{2n-1}(0)\frac{(x-y(0))^{2n-1}}{(2n-1)!}\left(1 + \frac{a_{2n}(0)}{2n\ a_{2n-1}(0)}(x - y(0))\right)\end{aligned}$$

where the last line assumes that every ratio $\frac{a_{2n}}{a_{2n-1}}$, for $n = 1, 2, ...,$ has a limit around zero.

Denoting $P_n(x) = \frac{(x-y(0))^{2n-1}}{(2n-1)!}\left(1 + \frac{a_{2n}(0)}{2n\ a_{2n-1}(0)}(x - y(0))\right)$, the initial condition reads

$$\psi(x) = \sum_{n=1}^{\infty} a_{2n-1}(0)P_n(x). \tag{18}$$

This idea is then to project ψ onto a basis of such polynomials. To do so it is necessary to derive an orthonormal set of polynomials from the set P_n, $n \in \mathbb{N}^*$. Step by step we follow the Gram–Schmidt procedure. Let us use $\langle f, g\rangle = \int_0^{y(0)} f(s)g(s)\mathrm{d}s$ as a dot product and $\|f\|^2 = \langle f, f\rangle$ as a norm.

First let us define $\tilde{P}_1 = P_1/\|P_1\|$ and $\pi_1 = \langle\psi, \tilde{P}_1\rangle$ the projection of ψ onto $\tilde{P}_1$. It is easy to check that $\pi_1 \in \mathbb{R}$ depends only on $\dot{y}(0)$. We shall note $\langle\psi, \tilde{P}_1\rangle = \pi_1(\dot{y}(0))$.

Then, following the orthonormalization procedure, we define $\tilde{P}_2 = g_2/\|g_2\|$ where $g_2 = P_2 - \langle P_2, \tilde{P}_1\rangle\tilde{P}_1$. We want to project ψ onto $\tilde{P}_2$. Before doing so, let us consider the following result: for $p > 1$ the coefficients a_{2p-1} and a_{2p} are of the polynomial form

$$a_{2p-1} = -y^{(p)} + h_{2p-1}\left(\dot{y}, ..., y^{(p-1)}\right) \tag{19}$$

$$a_{2p} = -y^{(p)}\left(p\ \nu + (3 - p)\dot{y}\right) + h_{2p}\left(\dot{y}, ..., y^{(p-1)}\right) \tag{20}$$

where h_{2p-1} and h_{2p} are polynomials of their variables. This result is complemented by the special cases $a_1 = -\dot{y}$ and $a_2 = -\dot{y}(\nu + \dot{y})$. The proof follows from an easy induction. For sake of numerical experiments, exact expressions of the first five a_i are given in the Appendix.

This last property allows us to state that the coefficients of the P_n polynomial depend upon $\dot{y}(0), ...,$ $y^{(n)}(0)$ only. Furthermore, by the orthonormal construction of the $\tilde{P}_n$, the coefficients of the $\tilde{P}_n$ depend upon $\dot{y}(0), ..., y^{(n)}(0)$ only. Thus the projection $\pi_n = \langle \psi, \tilde{P}_n \rangle$ depend only upon $\dot{y}(0), ..., y^{(n)}(0)$, and so

$$\pi_n = \langle \psi, \tilde{P}_n \rangle = \pi_n \left(\dot{y}(0), ..., y^{(n)}(0) \right).$$

So far we have a projection of ψ onto an orthonormal basis. We have now to recombine the obtained coefficients to derive conditions upon the (a_i) coefficients. From

$$\psi(x) = \sum_{i=1}^{\infty} \pi_i \left(\dot{y}(0), ..., y^{(i)}(0) \right) \tilde{P}_i$$

and

$$\tilde{P}_i = \sum_{j=i}^{\infty} \langle \tilde{P}_i, P_j \rangle P_j$$

we get after recombination

$$\psi(x) = \sum_{j=1}^{\infty} \sum_{i=1}^{j} \pi_i \left(\dot{y}(0), ..., y^{(i)}(0) \right) \langle \tilde{P}_i, P_j \rangle P_j.$$

So it is possible to identify the coefficients in (18) for $j = 1, 2, ...$ as

$$a_{2j-1} = \sum_{i=1}^{j} \pi_i \left(\dot{y}(0), ..., y^{(i)}(0) \right) \langle \tilde{P}_i, P_j \rangle \cdot$$

Finally we substitute the expressions (19) and (20) in these last relations to get a set of equations to be solved in terms of the successive derivatives of y:

$$\left. \begin{aligned} -\dot{y}(0) &= \pi_1 \left(\dot{y}(0) \right) \langle \tilde{P}_1, P_1 \rangle \\ -y^{(2)}(0) + h_3 \left(\dot{y}(0) \right) &= \pi_1 \left(\dot{y}(0) \right) \langle \tilde{P}_1, P_2 \rangle + \pi_2 \left(\dot{y}(0), \ddot{y}(0) \right) \langle \tilde{P}_2, P_2 \rangle \\ &\vdots \end{aligned} \right\} . \tag{21}$$

These equations are solvable for "small" initial conditions $x \mapsto \psi(x)$ in the L^2 sense. Indeed, the projection of such initial conditions onto the $\tilde{P}_i$ orthonormal polynomials are small, so the π_i , $i \in \mathbb{N}^*$ are small. This last property makes the set of equations solvable, since its jacobian gets closer to (minus) identity.

This procedure must be looked at in greater detail prior to any implementation. Density of the $\tilde{P}_i$ polynomials obtained through the Gram Schmidt procedure is an open issue (unfortunately the use of Stone–Weierstrass theorem is not straightforward and a dedicated approach seems required). Nonetheless numerical evaluations of the successive derivatives of y corresponding to a prescribed initial condition seem tractable. While the number of such coefficients to be computed to approximate ψ within some given tolerance is not known, we feel that such a result would be important for stabilization purposes.

Appendix A. Technical lemmas

Lemma A.1.

$$\frac{i!j!(i+j+l+1)!}{(i+j+1)!} = \sum_{r=0}^{l} \binom{l}{r} (j+r)!(i+l-r)!, \quad i,\ j,\ l \geq 0.$$

Proof. This result directly follows from the Chu–Vandermonde identity [13] that gives

$$(i+j+2)_l = \sum_{r=0}^{l} \binom{l}{r} (j+1)_r (i+1)_{l-r}$$

where $(a)_n = a(a+1)...(a+n-1)$ is the Pochhammer Symbol. One can use

$$\begin{aligned} (i+j+2)_l &= \frac{(i+j+l+1)!}{(i+j+1)!} \\ (j+1)_r &= \frac{(j+r)!}{j!} \\ (i+1)_{l-r} &= \frac{(i+l-r)!}{i!} \end{aligned}$$

and get after substitution

$$\frac{i!j!(i+j+l+1)!}{(i+j+1)!} = \sum_{r=0}^{l} \binom{l}{r} (j+r)!(i+l-r)! \qquad \square$$

Lemma A.2. *For $\alpha, c_k \geq 1$ and $b_k \geq 0$, $k = 0, 1, ..., l$,*

$$\sum_{k=0}^{l} c_k (b_k)^\alpha \leq \left(\sum_{k=0}^{l} c_k b_k \right)^\alpha, \quad l \geq 0.$$

The proof is an easy extension of that for the case $c_k = 1$, $\forall k = 0, 1, ..., l$, given in [6].

Appendix B. Gevrey functions bounds

In this section we give the derivation of the constants M_ϕ and R_ϕ. The Gevrey function $\dot{\phi}$ used in the simulations is given by the function ϕ, defined as

$$\phi(\tau) = \begin{cases} L + \Delta L & \text{if } \tau \geq 1, \\ L + \Delta L g(\tau) & \text{if } 1 > \tau > 0, \\ L & \text{if } \tau \leq 0, \end{cases}$$

where,

$$g(\tau) = \frac{f(\tau)}{f(\tau) + f(1-\tau)}, \ \tau \in [0,1] \quad \text{and} \quad f(\tau) = \begin{cases} \mathrm{e}^{-\frac{1}{\tau}} & \text{if } \tau > 0, \\ 0 & \text{if } \tau \leq 0. \end{cases}$$

The function ϕ defines a smooth transition from L to $L + \Delta L$ in liquid column length. The function chosen above is based upon an unpublished work of François Malrait done at École des Mines, which guarantees that

$$g(\tau) \leq \frac{f(\tau)}{m} \text{ where } 0 < m \leq f(\tau) + f(1-\tau).$$

Given f defined above, it is easy to show that $m = 2\mathrm{e}^{-2}$. The function $\dot{g}(\tau)$ is symmetric ($\dot{g}(\tau) = \dot{g}(1-\tau)$) so in estimating Gevrey bounds we can restrict the domain of τ to $[0, 1/2]$. Also, $\mathbb{C} \ni z \mapsto g(z)$ is holomorphic in the infinite strip $\{z = x + iy \in \mathbb{C} \ : \ 0 < x < 1\}$. To identify estimates for the bounds on the Gevrey constants we utilize Cauchy's integral formula, namely for $g(\tau)$ we have

$$g^{(k)}(\tau) = \frac{k!}{2\pi r^k} \int_{-\pi}^{+\pi} \mathrm{e}^{-ik\theta} g\left(\tau + r\mathrm{e}^{i\theta}\right) \mathrm{d}\theta,$$

where $r \in]0, 1/2[$ and $k = 1, 2, 3, ...$

Taking the absolute value we have

$$\left| g^{(k)}(\tau) \right| \leq \frac{k!}{2\pi r^k} \int_{-\pi}^{+\pi} \left| g\left(\tau + r\mathrm{e}^{i\theta}\right) \right| \, \mathrm{d}\theta \ \leq \ \frac{k!\mathrm{e}^2}{4\pi r^k} \int_{-\pi}^{+\pi} \left| f\left(\tau + r\mathrm{e}^{i\theta}\right) \right| \, \mathrm{d}\theta.$$

Bounding the integrand

$$\left| \mathrm{e}^{-1/\{\tau + r\mathrm{e}^{i\theta}\}} \right| = \exp\left[\frac{-(\tau + r\cos(\theta))}{\tau^2 + 2r\tau\cos(\theta) + r^2} \right].$$

Choosing $r = \mu\tau$, $0 < \mu < 1$, we can simplify as

$$\exp\left[\frac{-(\tau + r\cos(\theta))}{\tau^2 + 2r\tau\cos(\theta) + r^2} \right] = \exp\left[\frac{-(1 + \mu\cos(\theta))}{\tau\left(1 + 2\mu\cos(\theta) + \mu^2\right)} \right].$$

The last expression in the brackets is an even function of $\theta \in [-\pi, +\pi]$ so we need only consider the behavior for $\theta \in [0, +\pi]$. Over this range, the expression is a decreasing function of θ; to verify this, the derivative with respect to θ yields $-(\mu - \mu^3)\sin(\theta)/\tau$, which is always negative since $1 > \mu > 0$. Thus we can maximize the bracketed expression by evaluating it at $\theta = 0$ as

$$\exp\left[\frac{-(1 + \mu\cos(\theta))}{\tau\left(1 + 2\mu\cos(\theta) + \mu^2\right)} \right] \leq \exp\left[\frac{-(1+\mu)}{\tau(1+\mu)^2} \right] = \exp\left[\frac{-1}{\tau(1+\mu)} \right].$$

Returning to the integral equation with the definition for r gives

$$\frac{k!\mathrm{e}^2}{4\pi\mu^k\tau^k} \int_{-\pi}^{+\pi} \exp\left[\frac{-1}{\tau(1+\mu)} \right] \mathrm{d}\theta = \frac{k!\mathrm{e}^2}{2\mu^k\tau^k} \exp\left[\frac{-1}{\tau(1+\mu)} \right].$$

It is easy to show that

$$\max_{\tau \in [0,1/2]} \left\{ \tau^{-k} \exp\left[\frac{-1}{\tau(1+\mu)} \right] \right\} \leq \mathrm{e}^{-k} k^k (1+\mu)^k.$$

The bound now becomes

$$\frac{k!\mathrm{e}^2}{2\mu^k\tau^k} \exp\left[\frac{-1}{\tau(1+\mu)} \right] \leq \frac{k!\mathrm{e}^2}{2} \left(\frac{k}{e} \right)^k \left(\frac{1+\mu}{\mu} \right)^k, \quad \text{for any } \mu \in]0, 1[.$$

The best bound we can achieve is derived from the limiting behavior as $\mu \to 1$, resulting in

$$\left|g^{(k)}(\tau)\right| \leq \frac{k!\mathrm{e}^2}{2}\left(\frac{2k}{e}\right)^k \simeq k!^2\ \frac{\mathrm{e}^2}{2\sqrt{2\pi}}\ \frac{2^k}{\sqrt{k}}, \quad \forall\ k = 1, 2, 3, ..., \ \tau \in [0, 1],$$

where the last step utilizes Stirling's approximation. For reference, we also have the bounds

$$\left|f^{(k)}(\tau)\right| \leq \frac{k!^2}{\sqrt{2\pi}}\ \frac{2^k}{\sqrt{k}}, \quad \forall\ k = 1, 2, 3, ...$$

These parameterizations verify that g and f are Gevrey order 1. As we want the Gevrey constants for $\dot{\phi}$, we can write the bounds as

$$\left|\phi^{(k+1)}(\tau)\right| = \Delta L\left|g^{(k+1)}(\tau)\right| \leq \Delta L\frac{(k+1)!^2\ \mathrm{e}^2\ 2^{k+1}}{2\sqrt{2\pi(k+1)}} = \frac{\Delta L\mathrm{e}^2}{\sqrt{2\pi}}\ k!^2\ 2^k(k+1)^{3/2},$$

where $k = 0, 1, 2, ...$ With the bounds $2^k(k+1)^{3/2} \leq 2^k(k+1)^2 \leq 4^k$, we have bounds for $\dot{\phi}$ Gevrey order 1, namely

$$\left|\phi^{(k+1)}(\tau)\right| \leq M_\phi\frac{k!^2}{R_\phi^k}, \quad M_\phi \equiv \frac{\Delta L\mathrm{e}^2}{\sqrt{2\pi}},\ R_\phi \equiv \frac{1}{4}\cdot$$

The value of the column length increase ΔL is the only non-constant in these Gevrey bounds. To characterize these bounds, consider normalizing the model coefficients $\rho = \nu = 1$ and examine the radius of convergence η^* as a function of column length as shown in Figure 9, where we assume $0 < L \ll 1$. For guaranteed convergence of the solution over the domain, $\eta^* > \Delta L$ and the plot shows that this is the case when $\Delta L \in [0.0, 0.45]$. For an initial column length $L > 0.6$, a time bound T larger than 1 is also clearly required for even small column length increases.

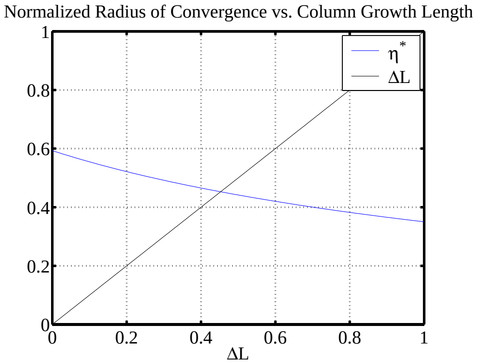


FIGURE 9. For normalized model coefficients ($\rho = \nu = 1$), the locus of the radius of convergence η^* as a function of the column length increase ΔL.

References

[1] J.R. Cannon, *The one-dimensional heat equation.* Addison-Wesley Publishing Company, *Encyclopedia Math. Appl.* **23** (1984).

[2] M. Fila and P. Souplet, Existence of global solutions with slow decay and unbounded free boundary for a superlinear Stefan problem. *Interfaces Free Boundaries* **3** (2001) 337-344.

[3] M. Fliess, J. Lévine, Ph. Martin and P. Rouchon, Flatness and defect of nonlinear systems: Introductory theory and examples. *Int. J. Control* **61** (1995) 1327-1361.

[4] M. Fliess, J. Lévine, Ph. Martin and P. Rouchon, A Lie–Bäcklund approach to equivalence and flatness of nonlinear systems. *IEEE Trans. Automat. Control* **44** (1999) 922-937.

[5] A. Friedman and B. Hu, A Stefan problem for multidimensional reaction-diffusion systems. *SIAM J. Math. Anal.* **27** (1996) 1212-1234.

[6] M. Gevrey, La nature analytique des solutions des équations aux dérivées partielles. *Ann. Sci. École Norm. Sup.* **25** (1918) 129-190.

[7] C.D. Hill, Parabolic equations in one space variable and the non-characteristic Cauchy problem. *Comm. Pure Appl. Math.* **20** (1967) 619-633.

[8] Chen Hua and L. Rodino, General theory of partial differential equations and microlocal analysis, in *Proc. of the workshop on General theory of PDEs and Microlocal Analysis, International Centre for Theoretical Physics, Trieste*, edited by Qi Min-You and L. Rodino. Longman (1995) 6-81.

[9] B. Laroche, Ph. Martin and P. Rouchon, Motion planing for the heat equation. *Int. J. Robust Nonlinear Control* **10** (2000) 629-643.

[10] A.F. Lynch and J. Rudolph, *Flatness-based boundary control of a nonlinear parabolic equation modelling a tubular reactor*, edited by A. Isidori, F. Lamnabhi–Lagarrigue and W. Respondek. Springer, *Lecture Notes in Control Inform. Sci.* **259:** Nonlinear Control in the Year 2000, Vol. 2. Springer (2000) 45-54.

[11] M.B. Milam, K. Mushambi and R.M. Murray, A new computational approach to real-time trajectory generation for constrained mechanical systems, in *IEEE Conference on Decision and Control* (2000).

[12] N. Petit, M.B. Milam and R.M. Murray, A new computational method for optimal control of a class of constrained systems governed by partial differential equations, in *Proc. of the* 15*th IFAC World Congress* (2002).

[13] M. Petkovsek, H.S. Wilf and D. Zeilberger, $A = B$. Wellesley (1996).

[14] L.I. Rubinstein, *The Stefan problem.* AMS, Providence, Rhode Island, *Transl. Math. Monogr.* **27** (1971).

[15] W. Rudin, *Real and Complex Analysis.* McGraw-Hill International Editions, Third Edition (1987).

Dynamics and solutions to some control problems for water-tank systems

Nicolas Petit , Pierre Rouchon

Abstract

We consider a tank containing a fluid. The tank is subjected to directly controlled translations and rotations. The fluid motion is described by linearized wave equations under shallow water approximations. For irrotational flows, a new variational formulation of Saint-Venant equations is proposed. This provides a simple method to establish the equations when the tank is moving. Several control configurations are studied: one and two horizontal dimensions; tank geometries (straight and non-straight bottom, rectangular and circular shapes), tank motions (horizontal translations with and without rotations). For each configuration we prove that the linear approximation is steady-state controllable and provide a simple and flatness-based algorithm for computing the steering open-loop control. These algorithms rely on operational calculus. They lead to second order equations in space variables whose fundamental solutions define delay operators corresponding to convolutions with compact support kernels. For each configurations several controllability open-problems are proposed and motivated.

Keywords:. Wave equations, boundary control, flatness, controllability, motion planning, delay operators.

Introduction

The following study is derived from an industrial problem for which tanks filled with liquid are to be moved to different steady-state workbenches as fast as possible. For such start and stop motions, the fluid mass has a significant contribution in the dynamics of the whole system. Several recent publications deal with this question, see, for example, [1], [2], [3], [4], [5]. This paper is a first attempt to base the control design on wave equations describing the fluid surface dynamics .

We concentrate on finding open-loop tank trajectories such that if the liquid is initially at rest then it returns to rest when the tank stops. This is a typical motion planning problem: finding open-loop control steering in finite time from one steady-state to another one. For finite dimensional systems, flatness based methods [6], [7] are very efficient to solve this problem. In [8], [9], [10], [11], [12], [13], infinite dimensional extensions are proposed for several systems described by partial differential equations with boundary control. We employ such a "flatness based" methodology, working on physical models of the system and we establish several controllability results: positive results consider exact steady-state controllability in finite time T, i.e., proving that there exists a control $[0, T] \ni t \mapsto u(t)$ steering the system from any steady-state to any other one, on the other hand negative results describe the lack of approximate controllability.

The first contribution of the paper consists of models. The major modelling difficulty lies in the fact that the fluid surface is unknown. A "rigorous" modelling involving Euler or Navier-Stoke equations with free surface boundaries is out of reach. Thus we restrict our study to classical modelling based on shallow water approximation [14]. Even for such restrictive modelling, the motion equations are not so simple to derive when the tank is moving. Thus in a first step, we propose a variational formulation of the Saint-Venant equations for irrotational flows and fixed tank: their solutions are extremal of the action under the constraint formed by the mass conservation equation. Then, when the tank is moving, we derive a similar formulation by adding the contribution of the tank motion in the kinetic and potential energy and proceed as before to get the dynamics.

The second contribution is relative to motion planning. The Saint-Venant equations are nonlinear hyperbolic equations. Only few results are available concerning their nonlinear stability, stabilization and controllability (see, e.g., [15], [16] and a recent result in [17]). Preliminary results [18] sketched in appendix lead us to thinking that, when used properly, linearized Saint-Venant equations can be an insighful approximation for motion planning purposes. We restrict our study to such linearized wave equations and show how to obtain open-loop control algorithms that are computationally straightforward. They are derived from formulas presented in lemmas 3, 4, 5 and 6 and are based on symbolic computations and involve operational calculus, Bessel functions, and Paley-Wiener theorem.

We would like to emphasize that although wave equations with Neuman boundary control have been intensively studied and many precise and general results are available on their controllability and stabilization (see, e.g., [19], [20], [21], [22], [23]), the classical results do not apply in a simple way to the problem presented in this paper. First reason: we have a linear wave equation controlled via Neuman boundary control but, the control is not distributed on the boundary; even for the simple tank described by system (26), the same control u, the acceleration of the tank, appears at both edges $x = -a$ and $x = a$. Second reason: the controlled wave equation is coupled with a double integrator $\ddot{D} = u$; one has to control not only the surface waves inside the tank but also the position and velocity of the tank.

N. Petit is with the Centre Automatique et Systèmes, École des Mines de Paris, France. Email: petit@cas.ensmp.fr.

P. Rouchon is the director of the Centre Automatique et Systèmes, École des Mines de Paris, France. Email: rouchon@cas.ensmp.fr

The paper is organized as follows. The dynamics of the systems under consideration are the subject of section I. Sections II and III are devoted to control problems for the one-dimensional and two-dimensional cases respectively.

More precisely, in section I we detail the variational formulation: the Saint-Venant equations for a moving tank are established. In section II, we consider one-dimensional cases: the translation case with a straight bottom is treated in details; the non-straight bottom is also addressed; combination of tank translation and rotation are investigated. Section III is devoted to two-dimensional cases: the translation of rectangular and circular tanks are solved; the combination of tank translation and rotation for an arbitrary geometry are also solved. In conclusion, we show that shallow water approximation is essential to ensure an explicit formulation of the dynamics: a more general modelling with a non-horizontal fluid velocity leads to an implicit formulation, an infinite dimensional analogue of an index one differential-algebraic system. In the appendix we prove a technical lemma devoted to symbolic analysis of the one dimensional wave equations when the speed depends on space: it can be seen as a generalization of d'Alembert formulas. We also recall in appendix some preliminary results and nonlinear simulations based on Godunov scheme for the Saint-Venant equation.

In this paper we pin point several open problems. As far as we know, the techniques presented in this paper give only partial answer, if any, in these tricky situations. We hope researchers in this area will welcome these challenges.

Some preliminary results relative to the horizontal translation of a tank in a vertical plane and with a straight bottom can be found in [18]. A preliminary version of this paper can be found in [24].

I. Variational formulations

The Saint-Venant equations describe the motion of a perfect fluid under gravity g with a free boundary (the shallow water assumption). We provide a variational formulation of these equations that, up to our knowledge, is new although not surprising (see, e.g., [25], [26], [14]). This variational formulation is interesting: it gives directly the dynamics equations when the tank is moving (translation and rotation).

A. The one-dimensional cases

A.1 One-dimensional straight bottom fixed tank

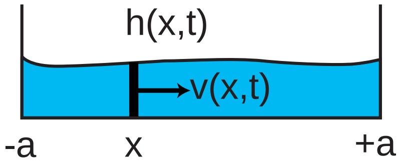


Fig. 1. The one dimensional tank.

We assume that the tank is at rest and study the motion of the fluid.

A.1.a Notations. As displayed on figure 1, the system is described by the following quantities

- a horizontal coordinate $x \in [-a, a]$ where $2a$ is the length of the tank;
- the height profile $[-a, a] \ni x \mapsto h(x, t)$ with $h(x, t) > 0$;
- the velocity profile $[-a, a] \ni x \mapsto v(x, t)$ with respect to the tank
- g is the gravity, ρ is the specific mass of the fluid.

A.1.b Physics. The mass conservation equation is

$$\frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial x} = 0. \tag{1}$$

The momentum conservation equation is

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = -g\frac{\partial h}{\partial x}. \tag{2}$$

The boundary condition is

$$v(-a, t) = v(a, t) = 0. \tag{3}$$

The kinetic energy T is

$$T(h, v) = \frac{\rho}{2}\int_{-a}^{a} h(x, t)v^2(x, t)\ dx. \tag{4}$$

The potential energy is

$$U(h, v) = \frac{\rho g}{2}\int_{-a}^{a} h^2(x, t)\ dx. \tag{5}$$

Under these hypothesis the following lemma holds

Lemma 1: Take a positive time $\tau > 0$. Equation (2) , i.e. the momentum conservation equation, results from the Euler-Lagrange first-order stationarity conditions deduced from

$$\delta\left(\int_0^\tau (T(h,v) - U(h,v))\ dt\right) = 0 \tag{6}$$

under the constraints formed by the mass equation (1), the boundary conditions (3) and fixed initial and final values for h and v: $h(x,0) = h_0(x)$, $v(x,0) = v_0(t)$, $h(\tau,x) = h_\tau(x)$, $v(\tau,x) = v_\tau(x)$.

Proof: Denote by $\lambda(x,t)$ the multiplier associated to the constraint (1) and by $\mathcal{L}(h,v,\lambda)$ the Lagrangian[1]

$$\mathcal{L} = \int_0^\tau (T(h,v) - U(h,v))\ dt + \int_0^\tau \int_{-a}^{a} \lambda(x,t)\left(\frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial x}\right)\ dx\ dt.$$

The condition $\delta\mathcal{L} = 0$ for any small variation δh of h such that $\delta h(x,0) = \delta h(x,\tau) = 0$ yields

$$\int_0^\tau \int_{-a}^{a} \left[\rho(v^2/2 - gh)\delta h + \lambda\left(\frac{\partial(\delta h)}{\partial t} + \frac{\partial(v\delta h)}{\partial x}\right)\right] dx\ dt = 0,$$

and then thanks to an integration by parts

$$\int_0^\tau \int_{-a}^{a} \left[\rho(v^2/2 - gh) - \frac{\partial\lambda}{\partial t} - v\frac{\partial\lambda}{\partial x}\right]\ \delta h\ dx\ dt = 0.$$

Thus

$$\frac{\partial\lambda}{\partial t} + v\frac{\partial\lambda}{\partial x} = \rho(v^2/2 - gh).$$

Similarly, variation δv of v such that $\delta v(x,0) = \delta v(x,\tau) = 0$ and $\delta v(-a,t) = \delta v(a,t) = 0$, gives

$$\frac{\partial\lambda}{\partial x} = \rho v.$$

Gathering these last two stationarity equations we get

$$\frac{\partial\lambda}{\partial t} + \frac{\rho}{2}v^2 = -g\rho h.$$

A differentiation with respect to x yields

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = -g\frac{\partial h}{\partial x} \tag{7}$$

which is indeed identical to (2). $\blacksquare$

A.2 One-dimensional non-straight bottom moving tank

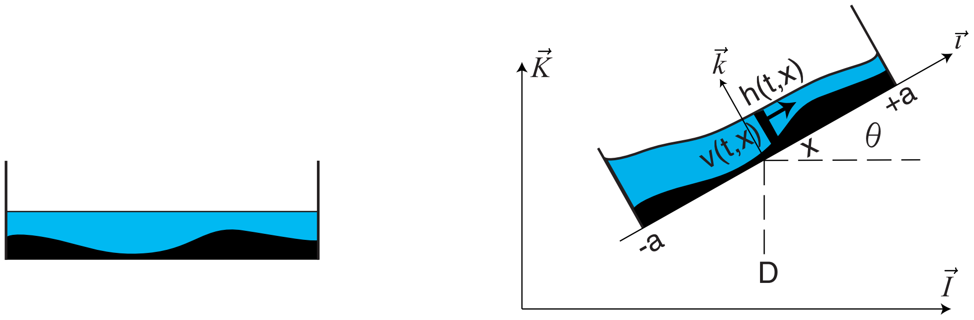


Fig. 2. The non-straight bottom tank at rest (right) and in movement (translation and rotation) (left).

We assume that the tank is moving.

[1] The Lagrangian we use is the classical Lagrangian as used in optimization: the constraints are adjoined with their Lagrange multipliers to the function (or functional) that is minimized.

A.2.a Notations. $(\vec I, \vec K)$ is the fixed frame with $\vec I$ horizontal and $(\vec\imath, \vec k)$ is the tank frame: $\vec\imath = \cos\theta \vec I + \sin\theta \vec K$ and $\vec k = -\sin\theta \vec I + \cos\theta \vec K$.

As displayed on figure 2, the tank motion is described by an horizontal position $\mathbf{R}^+ \ni t \mapsto D(t) \in \mathbf{R}$ and a rotation angle $\theta(t)$ around the horizontal axis orthogonal to the translation axis.

We still assume that the fluid can be described by $[-a, a] \times \mathbf{R}^+ \ni (x, t) \mapsto h(x, t)$ and $[-a, a] \times \mathbf{R}^+ \ni (x, t) \mapsto v(x, t)$, the velocity with respect to the tank. Notice that the space coordinate x is relative to the tank. Moreover we assume that the tank bottom is not straight but described by a smooth profile $[-a, a] \ni x \mapsto b(x) \in \mathbf{R}$. As before g is the gravity, ρ is the specific mass of the fluid.

A.2.b Physics and derivation of the model. The momentum conservation equation is derived from the variational formulation of lemma 1 with the following kinetic and potential energies (the boundary and constraint conditions remain unchanged)

$$T(h, v) = \frac{\rho}{2} \int_{-a}^{a} h \left(\dot D \vec I + v \vec\imath + x \dot\theta \vec k \right)^2 \, dx \tag{8}$$

$$U(h, v) = \rho g \int_{-a}^{a} \int_{b}^{b+h} \left(x \vec\imath + z \vec k \right) \cdot \vec K \; dz \; dx. \tag{9}$$

Denote by $\lambda(x, t)$ the multiplier associated to the mass conservation constraint and by $\mathcal{L}$ the resulting Lagrangian

$$\mathcal{L}(h, v, \lambda) = \int_0^\tau (T(h, v) - U(h, v)) \; dt + \int_0^\tau \int_{-a}^{a} \lambda(x, t) \left(\frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial x} \right) \; dx \; dt.$$

Stationary condition of $\mathcal{L}$ with respect to any small variation δv of v such that $\delta v = 0$ for $t = 0, \tau$ and $x \in \{-a, a\}$, yields

$$\frac{\partial \lambda}{\partial x} = \rho(\dot D \cos\theta + v). \tag{10}$$

Stationary condition of $\mathcal{L}$ with respect to any small variation δh of h such that $\delta h = 0$ for $t = 0, \tau$ yields

$$\frac{\partial \lambda}{\partial t} + v \frac{\partial \lambda}{\partial x} = \frac{\rho}{2} \left(\dot D \vec I + v \vec\imath + x \dot\theta \vec k \right)^2 - \rho g \left(x \vec\imath + (b + h) \vec k \right) \cdot \vec K. \tag{11}$$

We differentiate (11) with respect to x and substitute $\frac{\partial \lambda}{\partial x}$ by the righthand side of (10). We obtain the momentum conservation equation for v. The full dynamics are then described by the following set of equations

$$\left\{ \begin{aligned} & \frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial x} = 0 \\ & \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = -\ddot D \cos\theta - g \sin\theta + x \dot\theta^2 - g \cos\theta \frac{\partial (b + h)}{\partial x} \\ & \quad v(-a, t) = v(a, t) = 0. \end{aligned} \right. \tag{12}$$

Notice that these equations are indeed invariant under Galilean transformations, i.e. uniform translations, $D \mapsto D + p_1 t + p_0$ with p_1 and p_0 arbitrary constants.

Assume now that $\dot\theta$ is small ($\dot\theta^2 a \ll g$), that $h(x, t) = \bar h(x) + H(x, t)$, with $\bar h(x) = \varpi - b(x) > 0$ (where ϖ is a constant) is the steady-state height profile and with $|H| \ll \bar h$, and $|v| \ll \sqrt{g \bar h}$. Notice that we neither assume θ small nor $|\ddot D| \ll g$. Up to second order-terms the "linearized" dynamics read

$$\left\{ \begin{aligned} \frac{\partial H}{\partial t} &= -\frac{\partial (\bar h v)}{\partial x} \\ \frac{\partial v}{\partial t} &= -\ddot D \cos\theta - g \sin\theta - g \cos\theta \frac{\partial H}{\partial x} \\ v(-a, t) &= v(a, t) = 0. \end{aligned} \right.$$

We end up with the following model
Model 1: Elimination of v yields to a wave equation for H

$$\left\{ \begin{aligned} \frac{\partial^2 H}{\partial t^2} &= \frac{\partial}{\partial x} \left[\bar h \left(\ddot D \cos\theta + g \sin\theta + g \cos\theta \frac{\partial H}{\partial x} \right) \right] \\ g \cos\theta \frac{\partial H}{\partial x}(a, t) &= g \cos\theta \frac{\partial H}{\partial x}(-a, t) = -\ddot D \cos\theta - g \sin\theta \end{aligned} \right. \tag{13}$$

where $[-a,a] \ni x \mapsto \bar{h}(x)$ is the steady-state height profile and $h(x,t) = \bar{h}(x) + H(x,t)$ is up-to second order terms the liquid height. The control variables are $\ddot{D}(t)$ the horizontal acceleration of the tank, and $\dot{\theta}(t)$ its angular velocity. At any given time t, $[-a,a] \ni x \mapsto (H(x,t), \frac{\partial H}{\partial t}(x,t))$, $D(t)$, $\dot{D}(t)$ and $\theta(t)$ constitute the state of the system.

These equations are a good approximation as soon as

$$\dot{\theta}^2 a \ll g, \quad |H| \ll \bar{h}, \quad |v| \ll \sqrt{g\bar{h}}.$$

B. The two-dimensional cases

B.1 Two-dimensional straight bottom fixed tank

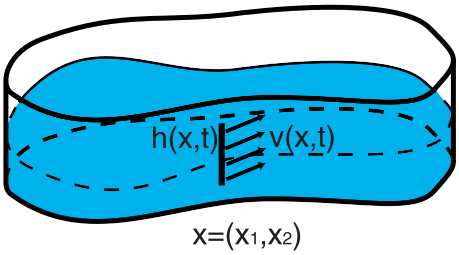


Fig. 3. The two-dimensional tank.

B.1.a Notations. As displayed on figure 3, the system is described by the following quantities:

- two horizontal coordinates $x = (x_1, x_2) \in \Omega$ where Ω is an open bounded connected domain of $\mathbf{R}^2$ with smooth boundary $\partial\Omega$;
- the height profile $\Omega \ni x \mapsto h(x,t)$ with $h(x,t) > 0$;
- the velocity profile $\Omega \ni x \mapsto \vec{v}(x,t) \in \mathbf{R}^2$.

We assume that the tank is at rest. As usual we denote by ∇ the operator

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \end{pmatrix}.$$

The mass conservation equation is

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\vec{v}) = 0 \tag{14}$$

where $\vec{v}$ is the velocity field of coordinates (v_1, v_2).

B.1.b Physics. The momentum conservation equation is

$$\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} = -g\nabla h. \tag{15}$$

The boundary condition is

$$\vec{v} \cdot \vec{n} = 0 \text{ on } \partial\Omega \tag{16}$$

where $\vec{n}$ is the normal to $\partial\Omega$.

We restrict our study to potential flow $\vec{v}$, i.e., solutions of (14,15,16) such that $\nabla \times \vec{v} = 0$. This makes sense since if the initial velocity profile is irrotational, it remains irrotational

$$\frac{\partial(\nabla \times \vec{v})}{\partial t} + \nabla \times [(\nabla \times \vec{v}) \times \vec{v}) = 0.$$

For irrotational $\vec{v}$, (15) reads also

$$\frac{\partial \vec{v}}{\partial t} + \frac{1}{2}\nabla(\vec{v}^2) = -g\nabla h.$$

The kinetic energy T is

$$T(h, \vec{v}) = \frac{\rho}{2} \int_\Omega h(x,t)\vec{v}^2(x,t)\ dx_1 dx_2. \tag{17}$$

The potential energy is

$$U(h, \vec{v}) = \frac{\rho g}{2} \int_\Omega h^2(x,t)\ dx_1 dx_2. \tag{18}$$

As for the one dimensional case, we have the following lemma

Lemma 2: Take a positive time $\tau > 0$ and consider irrotational solutions of (14,15,16) ($\nabla \times \vec{v} = 0$). Then equation (15), i.e. the momentum conservation equation, results from the Euler-Lagrange first-order stationarity conditions deduced from

$$\delta \left(\int_0^\tau (T(h, \vec{v}) - U(h, \vec{v}))\, dt = 0 \right) \tag{19}$$

under the constraints formed by the mass conservation equation (14), the boundary conditions (16) and fixed initial and final values for h and v, $h(x, 0) = h_0(x)$, $\vec{v}(x, 0) = \vec{v}_0(t)$, $h(\tau, x) = h_\tau(x)$, $\vec{v}(\tau, x) = \vec{v}_\tau(x)$.
The proof is very similar to the one dimensional one and is left to the reader.

B.2 Two-dimensional non-straight bottom moving tank

In the following we sketch the main steps to derive from lemma 2 the dynamical equations when the tank is moving with an irrotational velocity profile $\vec{v}$.

B.2.a Notations. The fixed frame is denoted by $(\vec{I}_1, \vec{I}_2, \vec{K})$ where $\vec{K}$ is the upwards vertical unit vector. The tank frame is $(\vec{\imath}_1, \vec{\imath}_2, \vec{k})$ where the liquid height is along the $\vec{k}$ axis. The rotation of the tank is described by the instantaneous rotation vector $\vec{\omega}$ defined by

$$\dot{\vec{\imath}}_1 = \vec{\omega} \times \vec{\imath}_1, \quad \dot{\vec{\imath}}_2 = \vec{\omega} \times \vec{\imath}_2, \quad \dot{\vec{k}} = \vec{\omega} \times \vec{k}.$$

Since we are looking for irrotational flows, we will see that, necessarily, $\vec{\omega} \cdot \vec{k} = 0$: the tank cannot spin around axis $\vec{k}$.

The fluid remains described by the height profile $\Omega \ni x \mapsto h(x, t) > 0$, where Ω is as before an open bounded connected domain of $\mathbf{R}^2$ with a piecewise smooth boundary $\partial\Omega$, and the velocity profile $\vec{v}(x, t) = v_1 \vec{\imath}_1 + v_2 \vec{\imath}_2$ where $x = (x_1, x_2)$ are Cartesian coordinates along a plane attached to the tank and parallel to $(\vec{\imath}_1, \vec{\imath}_2)$. We assume that the tank bottom is given by the profile $\Omega \ni x \mapsto b(x)$. We denote by (D_1, D_2, Z) the coordinates in the fixed frame of the point D attached to the tank. Vertical acceleration can be included by changing gravity g into $g + \ddot{Z}$ and just considering horizontal motions for D. Without loss of generality, we assume in the sequel that $Z \equiv 0$ and

$$\dot{D} = \dot{D}_1 \vec{I}_1 + \dot{D}_2 \vec{I}_2, \quad \ddot{D} = \ddot{D}_1 \vec{I}_1 + \ddot{D}_2 \vec{I}_2.$$

Once more, g denotes the gravity and ρ is the specific mass of the fluid.

B.2.b Physics and derivation of the model. With the above notations the kinetic and potential energies are

$$T(h, \vec{v}) = \frac{\rho}{2} \int_\Omega h \left(\dot{D} + \vec{v} + \vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) \right)^2 dx_1 dx_2 \tag{20}$$

$$U(h, \vec{v}) = \rho g Z + \rho g \int_\Omega \int_b^{b+h} \left(x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2 + z \vec{k} \right) \cdot \vec{K}\, dz dx_1 dx_2. \tag{21}$$

Denote by $\lambda(x, t)$ the multiplier associated to the mass conservation constraint and by $\mathcal{L}$ the resulting Lagrangian

$$\mathcal{L}(h, \vec{v}, \lambda) = \int_0^\tau (T(h, \vec{v}) - U(h, \vec{v}))\, dt + \int_0^\tau \int_\Omega \lambda(x, t) \left(\frac{\partial h}{\partial t} + \nabla \cdot (h \vec{v}) \right) dx_1 dx_2\, dt.$$

Stationary condition of $\mathcal{L}$ with respect to any small variation $\delta\vec{v} = \delta\vec{v}_1 \vec{\imath}_1 + \delta\vec{v}_2 \vec{\imath}_2$ of $\vec{v}$ such that $\delta\vec{v} = 0$ for $t = 0, \tau$ and $\delta\vec{v} \cdot \vec{n} = 0$ for $x \in \partial\Omega$, yields

$$\frac{\partial \lambda}{\partial x_\sigma} = \rho \left(\dot{D} + \vec{v} + \vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) \right) \cdot \vec{\imath}_\sigma, \qquad \sigma = 1, 2. \tag{22}$$

Stationary condition of $\mathcal{L}$ with respect to any small variation δh of h such that $\delta h = 0$ for $t = 0, \tau$ yields

$$\begin{aligned} \frac{\partial \lambda}{\partial t} + \vec{v} \cdot \nabla \lambda \;&=\; \frac{\rho}{2} \left(\dot{D} + \vec{v} + \vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) \right)^2 \\ &\quad - \rho g \left(x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2 + (b + h) \vec{k} \right) \vec{K}. \end{aligned} \tag{23}$$

According to (22)

$$\vec{v} \cdot \nabla \lambda = \rho \vec{v} \cdot \left(\dot{D} + \vec{v} + \vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) \right).$$

We then apply $\frac{\partial}{\partial x_1}$ on (23) and substitute $\frac{\partial \lambda}{\partial x_1}$ by the righthand side of (22):

$$\begin{aligned} &\frac{\partial v_1}{\partial t} + \vec{v} \cdot \frac{\partial \vec{v}}{\partial x_1} = x_2 \frac{d(\vec{\omega} \cdot \vec{k})}{dt} + \ldots \\ &\ldots + \frac{\partial}{\partial x_1} \left(\frac{1}{2} (\vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2))^2 - (\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) - g(b + h) \vec{k} \cdot \vec{K} \right). \end{aligned}$$

Similarly we have

$$\begin{aligned}
&\frac{\partial v_2}{\partial t} + \vec{v} \cdot \frac{\partial \vec{v}}{\partial x_2} = -x_1 \frac{d(\vec{\omega} \cdot \vec{k})}{dt} + \dots \\
&\dots + \frac{\partial}{\partial x_2} \left(\frac{1}{2} (\vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2))^2 - (\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) - g(b+h) \vec{k} \cdot \vec{K} \right).
\end{aligned}$$

This provides the vectorial momentum conservation equation for

$$\begin{aligned}
&\frac{\partial v}{\partial t} + \frac{1}{2} \nabla \vec{v}^2 = \frac{d(\vec{\omega} \cdot \vec{k})}{dt} (x_2 \vec{\imath}_1 - x_1 \vec{\imath}_2) + \dots \\
&\dots + \nabla \left(\frac{1}{2} (\vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2))^2 - (\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) - g(b+h) \vec{k} \cdot \vec{K} \right).
\end{aligned}$$

$\vec{v}$ must be kept irrotational to apply lemma 2. Thus we restrict rotations by $\vec{\omega} \cdot \vec{k} \equiv 0$.

The full dynamics is then described by the following set of equations

$$\left\{ \begin{aligned}
&\frac{\partial h}{\partial t} + \nabla \cdot (h \vec{v}) = 0 \\
&\frac{\partial \vec{v}}{\partial t} + \frac{1}{2} \nabla \vec{v}^2 = \frac{1}{2} \nabla \left((\vec{\omega} \times (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2))^2 \right) + \dots \\
&\dots + \nabla \left(-(\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) - g(b+h) \vec{k} \cdot \vec{K} \right) \\
&\vec{v} \cdot \vec{n} = 0 \text{ on } \partial\Omega.
\end{aligned} \right. \tag{24}$$

Assume now that $\vec{\omega}$ is small ($\vec{\omega}^2 a \ll g$, where a is the typical size of Ω), that $h(x,t) = \bar{h}(x) + H(x,t)$, with $\bar{h}(x) = cte - b(x) > 0$ is the steady-state height profile and with $|H| \ll \bar{h}$, and that $|\vec{v}| \ll \sqrt{g\bar{h}}$. Up to second order-terms the "linearized" dynamics read

$$\left\{ \begin{aligned}
\frac{\partial H}{\partial t} &= -\nabla \cdot (\bar{h} \vec{v}) \\
\frac{\partial \vec{v}}{\partial t} &= \nabla \left(-(\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) - g\vec{k} \cdot \vec{K} H \right).
\end{aligned} \right.$$

We end up with the following model

Model 2: Elimination of $\vec{v}$ yields to a wave equation for H

$$\left\{ \begin{aligned}
&\frac{\partial^2 H}{\partial t^2} = \nabla \cdot \left(\bar{h} \nabla \left[(\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) + gH \vec{k} \cdot \vec{K} \right] \right) \\
&\nabla \left[(\ddot{D} + g\vec{K}) \cdot (x_1 \vec{\imath}_1 + x_2 \vec{\imath}_2) + gH \vec{k} \cdot \vec{K} \right] \cdot \vec{n} = 0 \quad \text{ on } \partial\Omega
\end{aligned} \right. \tag{25}$$

where $\bar{h}(x)$ is the steady-state height profile and $h(x,t) = \bar{h}(x) + H(x,t)$ is up-to second order terms the liquid height.

The control variables are $\ddot{D}$ the tank acceleration and $\vec{\omega}$ its instantaneous rotation vector (remember that $\vec{\omega} \cdot \vec{k} = 0$). At any given time t, $[-a,a] \ni x \mapsto (H(x,t), \frac{\partial H}{\partial t}(x,t))$, $D(t)$, $\dot{D}$ and the three Euler angles defining the orientation of the tank constitute the state of the system.

These equations are a good approximation as soon as

$$\vec{\omega}^2 a \ll g, \quad |H| \ll \bar{h}, \quad \|\vec{v}\| \ll \sqrt{g\bar{h}}.$$

II. Several control problems and their solutions for the one-dimensional linearized model

A. Translation and straight bottom

Model 3: Assume that $[-a,a] \ni x \mapsto b(x) = 0$ and $\theta = 0$. Then $[-a,a] \ni x \mapsto \bar{h}(x)$ is constant and model 1 reads

$$\left\{ \begin{aligned}
&\frac{\partial^2 H}{\partial t^2} = \bar{h} g \frac{\partial^2 H}{\partial x^2} \\
&\frac{\partial H}{\partial x}(a,t) = \frac{\partial H}{\partial x}(-a,t) = -\frac{u}{g} \\
&\ddot{D} = u
\end{aligned} \right. \tag{26}$$

with $(H, \frac{\partial H}{\partial t}, D, \dot{D})$ as state and u as control.

The controllability of the above system can be studied directly by considering the dual system and its observability (see, e.g., [19], [21], [23]). The dual system reads

$$\begin{cases} \dfrac{\partial^2 P}{\partial t^2} = \bar{h} g \dfrac{\partial^2 P}{\partial x^2} \\ \dfrac{\partial P}{\partial x}(a,t) = \dfrac{\partial P}{\partial x}(-a,t) = 0 \\ \ddot{\xi} = 0 \end{cases}$$

with output $y = P(a,t) - P(-a,t) + \xi$ and is clearly not observable (any even solution $x \mapsto P(x,t)$ with $\xi \equiv 0$ gives $y = 0$). The approximate controllability is not even valid. Nevertheless, the system is steady-state controllable. This results from the following elementary lemma.

Lemma 3 (Parametrization of the trajectories) Denote $c = \sqrt{g\bar{h}}$ the velocity of the waves. The general solution of (26) is given by

$$\begin{cases} H(x,t) = \dfrac{c}{2g}\left(\dot{y}(t-\dfrac{x}{c}) - \dot{y}(t+\dfrac{x}{c})\right) + \dfrac{1}{2}\left(F(t+\dfrac{x}{c}) + F(t-\dfrac{x}{c})\right) + k_0 t \\ D(t) = \dfrac{1}{2}\left(y(t+\dfrac{a}{c}) + y(t-\dfrac{a}{c})\right) \\ u(t) = \dfrac{1}{2}\left(\ddot{y}(t+\dfrac{a}{c}) + \ddot{y}(t-\dfrac{a}{c})\right) \end{cases} \tag{27}$$

where k_0 is an arbitrary constant, F an arbitrary $2a/c$-periodic time function and y an arbitrary time function. Moreover

$$\begin{cases} k_0 = \dfrac{c}{2a}(H(0,2a/c) - H(0,0)) \\ F(t) = H(0,t) - \dfrac{c}{2a}(H(0,2a/c) - H(0,0))t \\ y(t) = D(t) + \dfrac{1}{2\bar{h}}\left(\displaystyle\int_0^a H(x,t)\,dx - \int_{-a}^0 H(x,t)\,dx\right). \end{cases} \tag{28}$$

Proof: When H and D are given by (27), standard computations show that they satisfy (26). Let us prove in details the converse: any solution of (26) admits the form (27) with k_0, F and y defined by (28).

The general solution of

$$\frac{\partial^2 H}{\partial t^2} = \bar{h} g \frac{\partial^2 H}{\partial x^2}$$

is given by the d'Alembert's formula

$$H(x,t) = \varphi(t+\frac{x}{c}) + \psi(t-\frac{x}{c})$$

where φ and ψ are smooth functions.

The idea of the proof is to turn the boundary conditions of the model into functional equations with φ and ψ as variables, and then to solve these equations.

The boundary conditions can be expressed as

$$\begin{cases} \dot{\varphi}(t+\dfrac{a}{c}) - \dot{\psi}(t-\dfrac{a}{c}) = -\dfrac{c}{g}\ddot{D}(t) \\ \dot{\varphi}(t-\dfrac{a}{c}) - \dot{\psi}(t+\dfrac{a}{c}) = -\dfrac{c}{g}\ddot{D}(t). \end{cases} \tag{29}$$

Elimination of D yields

$$\dot{\varphi}(t+\frac{a}{c}) + \dot{\psi}(t+\frac{a}{c}) = \dot{\varphi}(t-\frac{a}{c}) + \dot{\psi}(t-\frac{a}{c}).$$

Thus $f \equiv \dot{\varphi} + \dot{\psi}$ is a periodic function with period $2\frac{a}{c}$. Since $\frac{\partial H}{\partial t}(0,t) = f(t)$. F defined in (28) is a $2a/c$ periodic function and

$$\dot{F}(t) = f(t) - \frac{c}{2a}\int_0^{2a/c} f.$$

Consider y defined in (28). Since H is solution of (26), we have

$$\ddot{y}(t) = -g\frac{\partial H}{\partial x}(0,t).$$

Simple computations show also that

$$c\frac{\partial H}{\partial x}(0,t) = \dot{\varphi}(t) - \dot{\psi}(t).$$

So we have

$$\begin{aligned}\dot{\varphi}(t) - \dot{\psi}(t) &= -\frac{c}{g}\ddot{y}\\ \dot{\varphi}(t) + \dot{\psi}(t) &= \dot{F}(t) + \frac{c}{2a}\int_0^{2a/c} f.\end{aligned}$$

Thus

$$\begin{aligned}\dot{\varphi}(t) &= \frac{1}{2}\dot{F}(t) - \frac{c}{2g}\ddot{y}(t) + \frac{c}{4a}\int_0^{2a/c} f\\ \dot{\psi}(t) &= \frac{1}{2}\dot{F}(t) + \frac{c}{2g}\ddot{y}(t) + \frac{c}{4a}\int_0^{2a/c} f\end{aligned}$$

that is

$$\begin{aligned}\varphi(t) &= m + \frac{1}{2}F(t) - \frac{c}{2g}\dot{y}(t) + \frac{ct}{4a}\int_0^{2a/c} f\\ \psi(t) &= n + \frac{1}{2}F(t) + \frac{c}{2g}\dot{y}(t) + \frac{ct}{4a}\int_0^{2a/c} f\end{aligned}$$

where n and m are two constants. Yet $F(0) = H(0,0)$ and $H(0,0) = \varphi(0) + \psi(0)$. Thus $m + n = 0$ and

$$H(x,t) = \frac{1}{2}\left(F(t+\frac{x}{c}) + F(t-\frac{x}{c})\right) + \frac{c}{2g}\left(\dot{y}(t-\frac{x}{c}) - \dot{y}(t+\frac{x}{c})\right) + \frac{ct}{2a}\int_0^{2a/c} f.$$

With this relation, we compute $\int_0^a H(x,t)dx$ and $\int_{-a}^0 H(x,t)dx$ and derive D via $D(t) = y(t) - (\int_0^a H(x,t)dx - \int_{-a}^0 H(x,t)dx)/(2\bar{h})$. This gives

$$D(t) = \frac{1}{2}\left(y(t+\frac{a}{c}) + y(t-\frac{a}{c}))\right).$$

■

Remark 1 (Inspection of the controllability) The explicit parameterization (27) implies that (26) is not controllable neither exact nor approximate. To see this, take an initial state $(H_0(x), \dot{H}_0(x))$, $x \in [-a,a]$ that is zero. This means that ϕ and ψ are zeros on $[-a/c, a/c]$ since

$$2\varphi(t) = H_0(ct) + \int_0^{ct} \dot{H}_0(x)dx, \quad 2\psi(t) = H_0(ct) - \int_0^{ct} \dot{H}_0(x)dx.$$

Thus $F(t) = 0$ on $[-a/c, a/c]$. Take any final state $(H_T(x), \dot{H}_T(x))$ at time T. It will provide another $F(t)$ that will not vanish over $[T - a/c, T + a/c]$, in general. Since F is $2a/c$-periodic, there does not exist a trajectory joining such two states: F is an invariant quantity that cannot be modified by control. From an algebraic point of view, it corresponds to the torsion sub-module of the module attached to (29) (see [27] for more details). This means (26) is not controllable.

If we assume the initial state is zero then both F and k_0 vanish. We have an explicit description of all the trajectories passing though $(H_0, \dot{H}_0) = 0$. It suffices to take (27) with $k_0 = F = 0$. This provides a very simple way to steer the system from any steady-position in $D = p$ to any other steady-position in $D = q$. The system is steady-state controllable. More precisely there is a one-to-one correspondence between the trajectories starting from the steady position $D = p$ at time $t = 0$ and arriving at time $T > 2a/c$ at the steady position $D = q$, and the smooth functions $t \mapsto y(t)$ such that

$$y(t) = \begin{cases} p & \text{if } t \leq a/c\\ \text{arbitrary} & \text{if } a/c < t < T - a/c\\ q & \text{if } t \geq T - a/c\end{cases} \tag{30}$$

via the following formulas

$$\begin{cases} D(t) = \frac{1}{2}\left(y(t+\frac{a}{c}) + y(t-\frac{a}{c})\right)\\ H(x,t) = \frac{c}{2g}\left(\dot{y}(t+\frac{x}{c}) - \dot{y}(t-\frac{x}{c})\right).\end{cases} \tag{31}$$

Remark 2: The reader might believe that the problem of finding $t \mapsto D(t)$ with $D(t \leq 0) = p$ and $D(t \geq T) = q$ such that the solution of the Cauchy problem

$$\begin{cases} \dfrac{\partial^2 H}{\partial t^2} = \bar{h} g \dfrac{\partial^2 H}{\partial x^2} \\ \dfrac{\partial H}{\partial x}(a,t) = \dfrac{\partial H}{\partial x}(-a,t) = -\dfrac{\ddot{D}}{g} \end{cases}$$

starting from zeros at $t = 0$ and arriving at zero at $t = T$ could be obtained via basic symmetry arguments and invariance with respect to $t \mapsto -t$ and $x \mapsto -x$: this is false. The fact that $\ddot{D}(t) = -\ddot{D}(T-t)$ does not ensure that H and $\dot{H}$ return to zero at time T. The proposed method does.

Remark 3 (Physical meaning of the flat output) The quantity y appearing in (27) is the position of a particular point of the system. It is the center of gravity of the two punctual masses M^+ (the mass at the front of the tank) and M^- (the mass at the rear of the tank) placed at the edges of the tank ($x = a$ and $x = -a$):

$$M^+ = \int_0^a (\bar{h} + H(x,t))dx, \quad M^- = \int_{-a}^0 (\bar{h} + H(x,t))dx, \quad y(t) = D(t) + \frac{M^+ - M^-}{2\bar{h}}$$

(remember that $M^+ + M^- = 2a\bar{h}$, by mass conservation).

We have thus proved that the first-order linear approximation of

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial (hv)}{\partial x} = 0 \\ \dfrac{\partial v}{\partial t} + v\dfrac{\partial v}{\partial x} = -\ddot{D}(t) - \dfrac{\partial (h)}{\partial x} \\ v(-a,t) = v(a,t) = 0 \end{cases}$$

with $\ddot{D} = u$ as control is steady-state controllable but not controllable. Coron [17] has proved very recently using first return and fixed-point methods that the above nonlinear model itself is also steady-state controllable.

Remark 4 (Relevance of the linearization approach) Nonlinear simulations, see [18], show that the motions computed via formulas (26), i.e. parameterization of the trajectories of the linearized model, approximate the trajectories of the nonlinear system.

B. Translation and non-straight bottom

Model 4: When $\theta = 0$, model 1 reads

$$\begin{cases} \dfrac{\partial^2 H}{\partial t^2} = \dfrac{\partial}{\partial x}\left[\bar{h}(x)\ddot{D}(t) + g\bar{h}(x)\dfrac{\partial H}{\partial x}\right] \\ \dfrac{\partial H}{\partial x}(a,t) = \dfrac{\partial H}{\partial x}(-a,t) = -\dfrac{u(t)}{g} \\ \ddot{D}(t) = u(t) \end{cases} \tag{32}$$

where $\bar{h}(x)$ the the steady-state height profile and where $(H, \frac{\partial H}{\partial t}, D, \dot{D})$ is the state.

We will not study the controllability of (32) in details as for (26). We will just prove that for any $\bar{h}(x)$, this system is steady-state controllable: one can steer the system from the steady-position $D = p$ to another steady-position $D = q$ in finite time.

Lemma 4 (Steady-state controllability) Take p and q two reals, and $T > 2\Delta$ where

$$\Delta = \int_{-a}^{a} \frac{dx}{\sqrt{g\bar{h}(x)}}$$

is the propagation time between the two edges. There exists a smooth control $t \mapsto D(t)$ such that $D(t) = p$ for $t \leq 0$, $D(t) = q$ for $t \geq T$ and the solution of (32) starting from $(H, \dot{H}) = 0$ at time $t = 0$ returns to 0 at time T.

Proof: It is based on symbolic computations and the technical lemma 7 given in appendix. The proof is constructive in the sense that the control D is obtained via convolutions with L^2 kernels of compact support and deduced from the function $\mathcal{B}(x,\xi)$ of lemma 7. Just for this proof, we will assume that the liquid is between $x = 0$ and $x = 2a$. In the Laplace domain we have the following second order differential system[2] :

$$\begin{cases} (g\bar{h}H')' = s^2 H - s^2 \bar{h}' D \\ gH'(0,t) + s^2 D = gH'(2a,t) + s^2 D = 0 \end{cases} \tag{33}$$

[2] We do not consider extra terms such as $D(0)$ $\dot{D}(0)$ since s is just a formal variable that represents the derivation.

where $'$ is the derivation with respect to x. The general solution of $(g\bar{h}H')' = s^2H - s^2\bar{h}'D$ reads

$$H = s^2(X + D\beta)A - s^2(Y + D\alpha)B \tag{34}$$

where X and Y are the integration constants, A and B the solutions of $(g\bar{h}A')' = s^2A$ and $(g\bar{h}B')' = s^2B$ with $A(0) = 1$, $A'(0) = 0$, $B(0) = 0$, $g\bar{h}(0)B'(0) = 1$, and

$$\alpha(x, s) = \int_0^x \bar{h}'(x)A(x)dx, \quad \beta(x, s) = \int_0^x \bar{h}'(x)B(x)dx.$$

The fact that H given by (34) is solution results from the classical Wronskian identity

$$\begin{vmatrix} A & B \\ g\bar{h}A' & g\bar{h}B' \end{vmatrix} \equiv 1.$$

Since

$$H' = s^2(X + D\beta)A' - s^2(Y + D\alpha)B'$$

the boundary conditions read

$$\begin{cases} \bar{h}(0)D = Y \\ \quad D/g = -(X + D\beta_+)A'_+ + (Y + D\alpha_+)B'_+ \end{cases} \tag{35}$$

where $A'_+(s) = A'(2a, s)$, ... Notice that $A(0, s) = 1$, $\alpha(0, s) = \beta(0, s) = 0$ and $B'(0, s) = 1/(g\bar{h}(0))$. Elimination of D yields

$$PX = QY$$

where

$$\begin{aligned} P(s) &= g\bar{h}(0)A'_+ \\ Q(s) &= -(1 + g(\beta_+A'_+ - \alpha_+B'_+)) + g\bar{h}(0)B'_+. \end{aligned}$$

Let us examine in details the structure of the operators $P(s)$ and $Q(s)$. According to lemma 7, see appendix, with $c^2(x) = g\bar{h}(x)$, A and B read

$$\begin{aligned} A(x, s) &= \sqrt{\frac{c(0)}{c(x)}}\cosh(s\sigma(x)) + \int_{-\sigma(x)}^{\sigma(x)} \mathcal{A}(x, \xi)\exp(\xi s)d\xi \\ B(x, s) &= \int_{-\sigma(x)}^{\sigma(x)} \mathcal{B}(x, \xi)\exp(\xi s)d\xi \end{aligned}$$

where $\mathcal{A}$ and $\mathcal{B}$ are L^2 functions of ξ. Since

$$g\bar{h}A' = s^2\int_0^x A(\xi, s)\ d\xi, \quad g\bar{h}B' = 1 + s^2\int_0^x B(\xi, s)\ d\xi,$$

we have

$$\begin{aligned} A'(x, s) &= s^2\int_{-\sigma(x)}^{\sigma(x)} \bar{A}(x, \xi)\exp(\xi s)d\xi \\ B'(x, s) &= \frac{1}{g\bar{h}(x)} + s^2\int_{-\sigma(x)}^{\sigma(x)} \bar{B}(x, \xi)\exp(\xi s)d\xi \end{aligned}$$

for some L^2 functions of ξ, $\bar{A}$ and $\bar{B}$. Since

$$\begin{aligned} \alpha(x, s) &= \bar{h}(x)A(x, s) - \bar{h}(0) - \int_0^x \bar{h}A' \\ \beta(x, s) &= \int_0^x \bar{h}'B \end{aligned}$$

we have similarly

$$\alpha(x,s) = \bar{h}(x)A(x,s) - \bar{h}(0) + s^2 \int_{-\sigma(x)}^{\sigma(x)} \bar{\alpha}(x,\xi)\exp(s\xi)d\xi$$

$$\beta(x,s) = \int_{-\sigma(x)}^{\sigma(x)} \bar{\beta}(x,\xi)\exp(s\xi)d\xi$$

where $\bar{\alpha}$, $\bar{\beta}$ are L^2 functions of ξ with $\bar{\alpha}(0,\xi) = \bar{\beta}(0,\xi) \equiv 0$. Thus

$$P = s^2 \int_{-\sigma(2a)}^{\sigma(2a)} \bar{P}(\xi)\exp(\xi s)d\xi$$

$$Q = g\bar{h}(2a)A_+B'_+ - 1 + s^2 \int_{-\sigma(a)}^{\sigma(a)} \bar{Q}(\xi)\exp(\xi s)d\xi$$

where $\bar{P}$ and $\bar{Q}$ are L^2 functions of ξ. Thanks to the identity $g\bar{h}(AB' - A'B) = 1$, $g\bar{h}(2a)A_+B'_+ - 1$ is equal to $gB_+A'_+$ and can be represented as

$$s^2 \int_{-\sigma(2a)}^{\sigma(2a)} \bar{f}(\xi)\exp(\xi s)d\xi$$

via some L^2 function f. Thus Q reads

$$Q = s^2 \left(\int_{-\sigma(2a)}^{\sigma(2a)} (\bar{Q}(\xi) + \bar{f}(\xi))\exp(\xi s)d\xi \right).$$

and we have the following factorization $P(s) = s^2R(s)$ and $Q(s) = s^2S(s)$ with

$$\begin{cases} R = \displaystyle\int_{-\sigma(2a)}^{\sigma(2a)} \bar{P}(\xi)\exp(\xi s)d\xi \\ S = \displaystyle\int_{-\sigma(2a)}^{\sigma(2a)} (\bar{Q}(\xi) + \bar{f}(\xi))\exp(\xi s)d\xi. \end{cases} \tag{36}$$

The operators R and S correspond to convolution with L^2 kernels whose supports are included in $[-\sigma(2a), \sigma(2a)]$. For any quantity $Z(s)$

$$X = SZ, \quad Y = RZ, \quad D = \frac{R}{\bar{h}(0)}Z$$

formally satisfies the boundary conditions (35) and $H(x,s)$ defined by (34) is a solution of (33). Yet, we have seen that for each x, the operators A, B, α and β are also convolutions with compact kernels. In the time domain, all this machinery defines, for any arbitrary smooth time function $t \mapsto Z(t)$, a solution of $(t,x) \mapsto H(x,t)$ and $t \mapsto D(t)$ of (32). Moreover $D(t)$ depends on the values of Z over the interval $[t-\Delta, t+\Delta]$ where $\Delta = \sigma(2a)$ is the propagation time between the two edges. When Z is constant for $t < 0$, D is constant for $t < -\Delta$ and $H(x,t)$ is 0 for t small enough, i.e., $t < -3\Delta$ since

$$H = s^2[SA - RB + (\beta)A - \alpha B)R/\bar{h}(0)] \;\; Z$$

and for each x, each operator, S, R, A, B, α, β is a convolution with a kernel of support included in $[-\Delta, +\Delta]$. In fact $H(x,t)$ is 0 for $t < -\Delta$. This results from Holgrem uniqueness theorem: every quantity is smooth and $H = 0$ is also solution of (32) over $[-d, -\Delta]$ with $D = cte$ and $H_{t=-d} = 0$ and $\dot{H}_{t=-d} = 0$ for any $d > 3\Delta$. For Z constant we have

$$D = \frac{4a}{\bar{h}(2a)}Z$$

since $D = gA'(2a,s)/s^2Z$ and for $s = 0$, $A'(2a,s)/s^2$ is a well defined function of x, $\Lambda(x)$, solution of the differential equation

$$(g\bar{h}\Lambda')' = 2A(x,0) = 2,$$

the second s-derivative of (33) in $s = 0$ with 0 initial values ($\Lambda'(x) = 2x/(g\bar{h}(x))$). This relation explains the factorization by s^2 between P, Q and R, S: without it, we will not be able to steer the system from different steady-states via smooth functions Z that are constant outside $[-\Delta, \Delta]$; with P instead of R, D will always return to 0 when Z becomes constant; with R, the motion planning problem can be solved as in section II using a sigmoid function similar to (30). ■

Remark 5: For a general bottom profile, one can conjecture[3] that the minimum transition time is 2Δ, i.e., the double of the travelling time from one edge to the over one. This is to compare with the straight bottom case where the minimum transition case is just Δ. Notice that, when the bottom profile is symmetric one can prove that the minimum transition is also Δ . It suffices to define $A(x,s)$ and $B(x,s)$ such that A is symmetric and B is anti-symmetric and to adapt the above proof: with such A and B computations simplify and provide Δ as minimum transition time.

C. Translation and rotation

Assume that we have two controls D and θ and that we want to steer the system from rest to rest, i.e. from D_0 at time $t=0$ to D_T at time $t=T>0$. Take any smooth function $[0,T] \ni t \mapsto D(t)$ such that $D(0)=D_0$, $D(T)=D_T$ and $D^{(i)}(0)=D^{(i)}(T)=0$, $i=1,2$. Set $\theta = -\arctan(\ddot{D}/g)$. The solution $t \mapsto H(x,t)$ of (13) starting from 0 satisfies

$$\begin{cases} \dfrac{\partial^2 H}{\partial t^2} = \dfrac{\partial}{\partial x}\left[\bar{h}(x) g \cos\theta \dfrac{\partial H}{\partial x}\right] \\ g\cos\theta \dfrac{\partial H}{\partial x}(a,t) = g\cos\theta\dfrac{\partial H}{\partial x}(-a,t) = 0. \end{cases}$$

We can deduce from that $H(x,t)=0, \forall x \in [-a,a], \forall t \in [0,T]$. The control $\theta(t) = -\arctan(\ddot{D}(t)/g)$ steers the system from rest to rest. In practice such open-loop control will be valid if $\dot{\theta}^2 a \ll g$, i.e., for all $t \in [0,T]$,

$$|D^{(3)}(t)| \ll \frac{g^2 + (D^{(2)}(t))^2}{\sqrt{ga}}.$$

D. Open problems

D.1 Controllability of the non-straight bottom system

With the single control D ($\theta \equiv 0$), we have seen that, when the bottom is straight, the system is not controllable. Is it still true for a non-straight bottom ? This suggests the following problem: characterize in term of $\bar{h}(x)$, the controllability of

$$\begin{cases} \dfrac{\partial^2 H}{\partial t^2} = \dfrac{\partial}{\partial x}\left[\bar{h}(x)u(t) + g\bar{h}(x)\dfrac{\partial H}{\partial x}\right] \\ \dfrac{\partial H}{\partial x}(a,t) = \dfrac{\partial H}{\partial x}(-a,t) = -\dfrac{u(t)}{g} \\ \ddot{D}(t) = u(t). \end{cases}$$

Since $\frac{d^2}{dt^2}\left(\int_{-a}^{a} H(x,t)dx\right) = 0$ we assume that $\int_{-a}^{a} H \equiv 0$: this is just the global conservation of the fluid in the tank. An interesting fact is that one can prove, from the observability of the adjoint system [19], that, when $[-a,a] \ni x \mapsto h(x)$ is even ($h(x)=h(-x)$), the system is not controllable: the adjoin system

$$\begin{aligned} \frac{\partial^2 P}{\partial t^2} &= \frac{\partial}{\partial x}\left(g\bar{h}(x)\frac{\partial P}{\partial x}\right), \\ \frac{\partial P}{\partial x}(-a,t) &= \frac{\partial P}{\partial x}(a,t) = 0 \\ \ddot{x}i &= 0 \end{aligned}$$

with

$$y(t) = \xi \bar{h}(a)P(a,t) - \bar{h}(-a)P(-a,t) - \int_{-a}^{a} P(x,t)\bar{h}'(x)dx$$

as output is not observable ($y \equiv 0$ for solutions $P(x,t)$ that are even x-function and $\xi = 0$). This particular case is important in practice, but more precisely speaking, what are, if any, the necessary and sufficient conditions on $[-a,a] \ni x \mapsto h(x)$ for the system to be controllable?

D.2 Use of an extra control

We know that the straight bottom tank with the single control D is not controllable. Is-it still true with the additional control θ? This suggests the study of the controllability of the following system where the nonlinearity is due to the

[3] This conjecture has been suggested by Jean-Michel Coron.

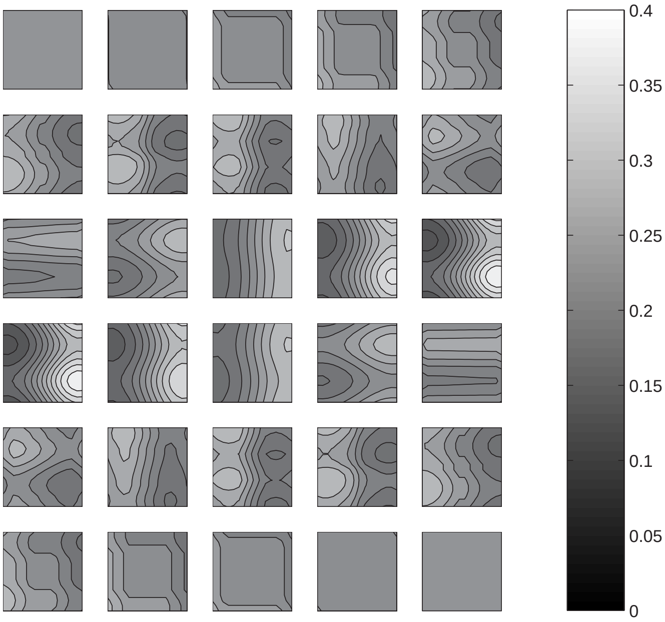


Fig. 4. Height contours sequence of the free surface of a square tank filled with fluid. Finite-time excursion from a steady point ($a_1 = a_2 = .5$, $\bar{h} = .2$, $g = 10$, $time = 2.2$)

control:

$$\begin{cases} \dfrac{\partial^2 H}{\partial t^2} = \dfrac{\partial}{\partial x}\left[\bar{h}\left(\ddot{D}\cos\theta + g\sin\theta + g\cos\theta\dfrac{\partial H}{\partial x}\right)\right] \\ g\cos\theta\dfrac{\partial H}{\partial x}(a,t) = g\cos\theta\dfrac{\partial H}{\partial x}(-a,t) = -\ddot{D}\cos\theta - g\sin\theta \end{cases}$$

with $\ddot{D} = u(t)$ and $\dot{\theta} = \omega$ as control variables (we still assume that $\int_{-a}^{a} H \equiv 0$).

III. Control of the two-dimensional linearized model: first issues

A. Translation of the rectangular tank

Model 5: When $Z \equiv 0$, $\vec{\omega} \equiv 0$ and $(\vec{\imath}_1, \vec{\imath}_2, \vec{k}) \equiv (\vec{I}_1, \vec{I}_2, \vec{K})$, model 2 becomes for a straight bottom ($\bar{h}$ constant):

$$\begin{cases} \ddot{H} = g\bar{h}\Delta H \\ g\nabla H \cdot \vec{n} = -u.\vec{n} \quad \text{on } \partial\Omega \\ \ddot{D} = u \end{cases} \tag{37}$$

Assume that Ω is the rectangle $[-a_1, a_1] \times [-a_2, a_2]$. The following lemma provides a constructive answer to the motion planing problem.

Lemma 5 (Flatness of the rectangular tank) Take two arbitrary C^3 time functions y_1 and y_2. Then D and H defined by ($c^2 = g\bar{h}$)

$$\begin{cases} D_1(t) = \dfrac{1}{2}\left(y_1(t+\dfrac{a_1}{c}) + y_1(t-\dfrac{a_1}{c})\right) \\ D_2(t) = \dfrac{1}{2}\left(y_2(t+\dfrac{a_2}{c}) + y_2(t-\dfrac{a_2}{c})\right) \\ H(x_1,x_2,t) = \dfrac{c}{2g}\left(\dot{y}_1(t+\dfrac{x_1}{c}) - \dot{y}_1(t-\dfrac{x_1}{c}) + \dot{y}_2(t+\dfrac{x_2}{c}) - \dot{y}_2(t-\dfrac{x_2}{c})\right). \end{cases} \tag{38}$$

satisfy (37) automatically.

The proof is straightforward. When y_1 and y_2 are constant, $D_1 = y_1$, $D_2 = y_2$ and $H = 0$. Steering from steady position (p_1, p_2) to steady position (q_1, q_2) can then be solved as in section II with a sigmoid function for y_1 and y_2 similar to (30).

In fact, equations (38) can be seen as the superposition of solutions of two one-dimensional wave equations whose boundary conditions are decoupled, see [12]. We represent on figure 4 successive contours of the free surface of a rectangular tank filled with fluid steered from two different steady points, using bump functions in equations (38).

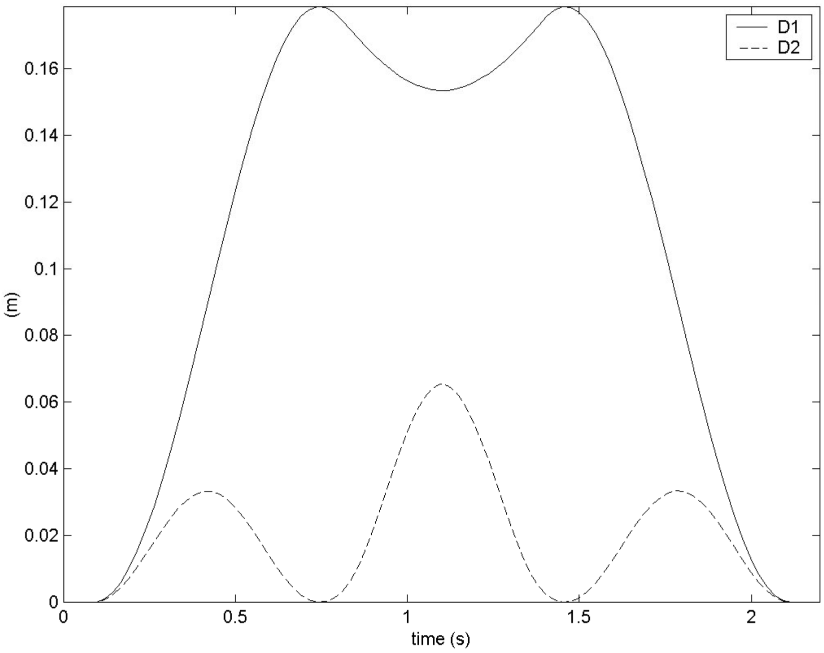


Fig. 5. The tank position D_1 and D_2 associated to the contours sequence of figure 4

B. Translation of the circular tank: the tumbler

Model 6: When $Z \equiv 0$, $\vec{\omega} \equiv 0$ and $(\vec{\imath}_1, \vec{\imath}_2, \vec{k}) \equiv (\vec{I}_1, \vec{I}_2, \vec{K})$, model 2 becomes for a straight bottom ($\bar{h}$ constant):

$$\begin{cases} \ddot{H} = g\bar{h}\Delta H \\ g\nabla H \cdot \vec{n} = -u.\vec{n} \quad \text{on } \partial\Omega \\ \ddot{D} = u \end{cases} \tag{39}$$

Assume that Ω is the disk of radius l and D its center. We denote by (r, θ) the polar coordinates with respect to the center of Ω. The following lemma provides a simple positive and constructive answer to the motion planing problem.

Lemma 6: Take two arbitrary C^3 time functions y_1 and y_2. Then D and H defined by

$$\begin{cases} D_1(t) = \dfrac{1}{\pi} \displaystyle\int_0^{2\pi} y_1 \left(t - \dfrac{l \cos\varphi}{\sqrt{g\bar{h}}} \right) \cos^2\varphi \, d\varphi \\ D_2(t) = \dfrac{1}{\pi} \displaystyle\int_0^{2\pi} y_2 \left(t - \dfrac{l \cos\varphi}{\sqrt{g\bar{h}}} \right) \cos^2\varphi \, d\varphi \\ H(r,\theta,t) = \dfrac{\cos\theta}{\pi} \sqrt{\dfrac{\bar{h}}{g}} \displaystyle\int_0^{2\pi} \dot{y}_1 \left(t - \dfrac{r}{\sqrt{g\bar{h}}} \cos\varphi \right) \cos\varphi \, d\varphi \\ \qquad + \dfrac{\sin\theta}{\pi} \sqrt{\dfrac{\bar{h}}{g}} \displaystyle\int_0^{2\pi} \dot{y}_2 \left(t - \dfrac{r}{\sqrt{g\bar{h}}} \cos\varphi \right) \cos\varphi \, d\varphi \end{cases} \tag{40}$$

satisfy automatically (39).

When y_1 and y_2 are constant, $D_1 = y_1$, $D_2 = y_2$ and $H = 0$. Steering from steady position (p_1, p_2) to steady position (q_1, q_2) can then be solved as for the rectangular tank.

Figure 6 shows the shape of the free surface during a transition between two steady points. This snapshot was computed using lemma 6. The corresponding `Matlab` code can be obtained upon request to the authors.

Proof: The direct proof which consists in verifying (39) is left to the reader. The proof given below is much more instructive. It explains the method used to obtain (40). Moreover it can be generalized to any variable depth profile $\bar{h}$ depending only on r. This proof uses some classical computations detailed in [28].

Let us perform a Laplace transform with respect to the time variable (with zero initial conditions)

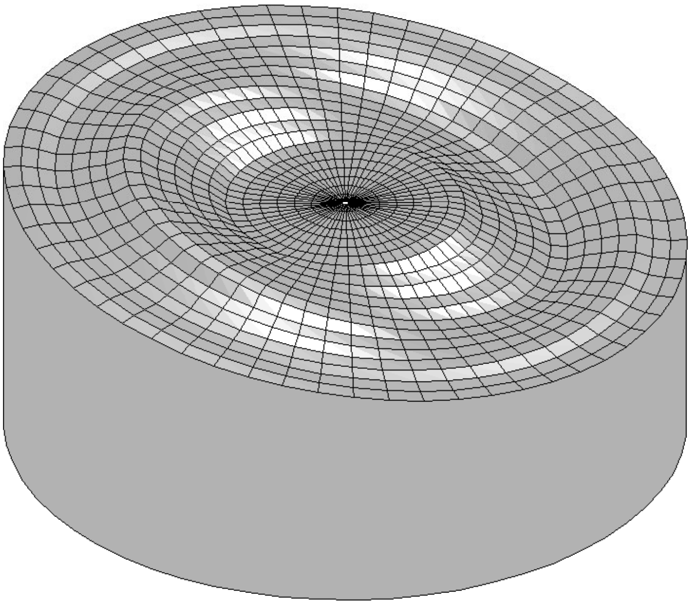


Fig. 6. The tumbler in movement. Snapshot of an animation computed using the explicit parameterization (40).

$$\Delta \hat{H}(x,y,s) - \frac{s^2}{g\bar{h}} \hat{H}(x,y,s) = 0 \tag{41}$$

where s can be considered as a parameter. Let us find a solution of equation (41) in cylindrical coordinates in the form $\bar{H}(r,\theta) = R(r)\Theta(\theta)$. By differentiation we get

$$\frac{1}{R}\left(\frac{d^2R}{dr^2} + \frac{1}{r}\frac{dR}{dr}\right) + \frac{1}{r^2\Theta}\frac{d^2\Theta}{d\theta^2} - \frac{s^2}{g\bar{h}} = 0. \tag{42}$$

The variable θ appears only in the second term of this equation. So the term $\frac{1}{\Theta}\frac{d^2\Theta}{d\theta^2}$ is independent of r and θ. It just depends on s and it can be denoted by $-\nu^2(s)$, $\nu \in \mathbf{C}$. Thus

$$\frac{1}{\Theta}\frac{d^2\Theta}{d\theta} = -\nu^2(s)$$

and

$$\Theta = A(s)\cos(\nu(s)\theta) + B(s)\sin(\nu(s)\theta)$$

for some integration constant $A(s)$ and $B(s)$. Then (42) writes

$$\frac{d^2R}{dr^2} + \frac{1}{r}\frac{dR}{dr} + R\left(-\frac{s^2}{g\bar{h}} - \frac{\nu(s)^2}{r^2}\right) = 0.$$

This is a Bessel equation. Its general solution is a combination of $J_{\nu(s)}\left(\frac{isr}{\sqrt{g\bar{h}}}\right)$ and $Y_{\nu(s)}\left(\frac{is\rho}{\sqrt{g\bar{h}}}\right)$. Since $Y_{\nu(s)}\left(\frac{isr}{\sqrt{g\bar{h}}}\right)$ is not bounded for $r = 0$, we only consider solutions involving $J_{\nu(s)}$. A set of bounded solution of (41) is given by

$$\hat{H}(r,\theta,s) = J_{\nu(s)}\left(\frac{isr}{\sqrt{g\bar{h}}}\right)(A(s)\cos(\nu(s)\theta) + B(s)\sin(\nu(s)\theta)) \tag{43}$$

The boundary conditions in the Laplace domain is

$$g\frac{\partial \hat{H}}{\partial r} = -\hat{u}_1(s)\cos\theta - \hat{u}_2(s)\sin\theta \text{ for } r = l.$$

Via (43) the boundary conditions also read

$$\frac{\partial \hat{H}}{\partial r} = \frac{is}{\sqrt{g\bar{h}}} {J'}_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right)(A(s)\cos\theta + B(s)\sin\theta) \ \text{ for } r = l.$$

By identification we have $\nu(s) = 1$ and

$$\begin{cases} \hat{u}_1(s) = & -\, isA(s)\sqrt{\frac{g}{\bar{h}}}\ {J'}_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right) \\ \hat{u}_2(s) = & -\, isB(s)\sqrt{\frac{g}{\bar{h}}}\ {J'}_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right) \\ \hat{H}(r,\theta,s) = & (A(s)\cos\theta + B(s)\sin\theta)\ J_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right). \end{cases}$$

Transforming these equations back into the time domain using the Poisson integral representations

$$J_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right) = \frac{1}{2i\pi}\int_0^{2\pi} e^{-\frac{sr\cos\varphi}{\sqrt{g\bar{h}}}}\cos\varphi\ d\varphi$$
$${J'}_1\left(\frac{isr}{\sqrt{g\bar{h}}}\right) = \frac{1}{2\pi}\int_0^{2\pi} e^{-\frac{sr\cos\varphi}{\sqrt{g\bar{h}}}}\cos^2\varphi\ d\varphi$$

with

$$A(s) = 2is\sqrt{\frac{\bar{h}}{g}}\ \hat{y}_1, \quad B(s) = 2is\sqrt{\frac{\bar{h}}{g}}\ \hat{y}_2,$$

yields (40). ■

C. Translation and rotation

We consider a general tank with an arbitrary domain Ω and assume that the dynamics are described by model 2. We will prove that the method used for the one-dimensional tank can be extended to the two-dimensional one.

Assume that $t \mapsto D = (D_1, D_2, Z)$ is a given smooth time function. We can adjust the tank rotations such that the term

$$(\ddot{D} + g\vec{K}) \cdot (x_1\vec{\imath}_1 + x_2\vec{\imath}_2)$$

appearing in (25) vanishes identically. With the three Euler angles (φ, θ, ψ) (see, e.g.,[29, pages 10,16]) this gives the following two equations

$$\begin{aligned} -(g+\ddot{Z})\cos\phi\sin\theta =& \ddot{D}_1(\cos\phi\cos\theta\cos\psi - \sin\phi\sin\psi) \\ & - \ddot{D}_2(\cos\phi\cos\theta\sin\psi + \sin\phi\cos\psi) \\ -(g+\ddot{Z})\sin\phi\sin\theta =& \ddot{D}_1(\sin\phi\cos\theta\cos\psi + \cos\phi\sin\psi) \\ & + \ddot{D}_2(-\sin\phi\cos\theta\sin\psi + \cos\phi\cos\psi) \end{aligned}$$

that must be completed by the non-holonomic constraint $\vec{\omega} \cdot \vec{k} = 0$ (the tank cannot spin around $\vec{k}$)

$$\dot{\psi} + \dot{\phi}\cos\theta = 0.$$

Simple computations give $\psi \in [0, 2\pi[$ and $\theta \in] - \pi/2, \pi/2[$ directly

$$\cos\psi = \frac{\ddot{D}_1}{\sqrt{\ddot{D}_1^2 + \ddot{D}_2^2}}, \quad \sin\psi = -\frac{\ddot{D}_2}{\sqrt{\ddot{D}_1^2 + \ddot{D}_2^2}}, \quad \tan\theta = -\frac{\sqrt{\ddot{D}_1^2 + \ddot{D}_2^2}}{g + \ddot{Z}}.$$

The remaining angle $\phi \in [0, 2\pi[$ is then obtained by integrating

$$\dot{\phi} = -\dot{\psi}/\cos\theta.$$

This method is just a compensation of accelerations by tank rotations. With such rotations the vector $\vec{k}$ that is orthogonal to the liquid surface at rest always remains co-linear to the total acceleration $\ddot{D} + g\vec{K}$. As for the one-dimensional tank, we can move the tank from one steady-state position to another one. Expected for simple motions $t \mapsto D(t)$ such as straight line ones, the orientation of the tank is not preserved between two steady-state positions: θ always returns to 0 after the motion, whereas the net rotation around the vertical axis $\vec{K}$, i.e., the total variation of $\phi + \psi$, does not. This results from the non-holonomic constraint $\dot{\psi} + \dot{\phi}\cos\theta = 0$.

Notice that, if the problem is to steer the tank from $D_0 = (p_1, p_2)$ at time 0 to $D_T = (q_1, q_2)$ at time $T > 0$ and to preserve its initial and final orientations, such method works when we take the straight trajectory $D(t) = (1 - \sigma(t))D_0 + \sigma(t)D_T$ with $[0, T] \ni t \mapsto \sigma(t) \in [0, 1]$ a smooth function such that $\sigma(0) = 0$, $\sigma(T) = 1$ and $\dot{\sigma} = \ddot{\sigma} = 0$ at $t = 0$ and $t = T$.

D. Open problem: beyond rectangular and circular shapes

For special geometries of the fluid domain Ω (namely rectangle and disk) and bottom profile ($\bar{h}$ constant) we have seen that

$$\begin{aligned} \frac{\partial^2 H}{\partial t^2} &= \nabla \cdot \left(\bar{h} \left(\ddot{D} + g\nabla H \right) \right) \quad \text{on } \Omega \\ g\nabla H \cdot \vec{n} &= -u \cdot \vec{n} \quad \text{on } \partial\Omega \\ \ddot{D} &= u \end{aligned}$$

is steady-state controllable with the two controls $\ddot{D}_1 = u_1$ and $\ddot{D}_2 = u_2$. We have also seen that in the one-dimensional case it is steady-state controllable for arbitrary bottom profile $\bar{h}(x)$. Is-it still true in the two dimensional case with an arbitrary domain Ω? As far as we know the ellipsoidal case is problematic. Using the technique we detailed for the circular case, we are left with Mathieu equations instead of Bessel equations. The fundamental solutions of these Mathieu equations do not have handy integral representations that would give a constructive proof of controllability when turned back into the time-domain. Up to now this seems a major obstruction to our method.

IV. Conclusion

The results presented in this paper are all based on linear control models deduced from shallow water approximations. This is a major restriction but we would like to emphasize the difficulties one would encounter dealing with a non-horizontal fluid velocity. For an arbitrary liquid height, a correct description of the dynamics around steady-states could be obtained as follows. For the translation of the tank of figure 1 with irrotational 2D flows, we linearize the Euler equations and the free boundary conditions. Following [14, page 436] the system is described by a scalar potential $\phi(x, z, t)$ depending on the horizontal coordinate x, the vertical one z and the time t, that satisfies

$$\begin{aligned} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial z^2} &= 0 \quad \text{for } (x, z) \in [-a, a] \times [0, \bar{h}] \\ \frac{\partial \phi}{\partial z}(x, 0, t) &= 0 \quad \text{for } x \in [-a, a] \\ g\frac{\partial \phi}{\partial z}(x, \bar{h}, t) &= -\frac{\partial^2 \phi}{\partial t^2}(x, \bar{h}, t) \quad \text{for } x \in [-a, a] \\ \frac{\partial \phi}{\partial x}(-a, z, t) &= \dot{D}(t) \quad \text{for } z \in [0, \bar{h}] \\ \frac{\partial \phi}{\partial x}(a, z, t) &= \dot{D}(t) \quad \text{for } z \in [0, \bar{h}] \end{aligned}$$

where $D(t)$ is the control, the horizontal tank position. The fluid velocity with respect to the tank admits two components, v the horizontal one and w the vertical one given by

$$v(x, z, t) = \frac{\partial \phi}{\partial x} - \dot{D}(t), \quad w(x, z, t) = \frac{\partial \phi}{\partial z}.$$

The liquid height is also derived from ϕ via

$$h(x, t) = \bar{h} - \frac{1}{g}\frac{\partial \phi}{\partial t}(x, \bar{h}, t).$$

This implicit formulation of the dynamics is very similar to differential-algebraic systems of index 1 [30], [31], [32]

$$\frac{dX}{dt} = f(X,Y,U), \quad 0 = g(X,Y,U)$$

often encountered for finite dimensional systems ($\frac{\partial g}{\partial Y}$ invertible). Set

$$X \equiv (\phi(x,\bar{h},t))_{-a\leq x\leq a}, \quad Y \equiv \phi, \quad U \equiv D.$$

Then the algebraic part $g(X,Y,U) = 0$ reads

$$\begin{cases} \dfrac{\partial^2\phi}{\partial x^2} + \dfrac{\partial^2\phi}{\partial z^2} = 0 \quad \text{for } (x,z) \in [-a,a]\times[0,\bar{h}] \\ \dfrac{\partial\phi}{\partial z}(x,0,t) = 0 \quad \text{for } x\in[-a,a] \\ \phi(x,\bar{h},t) = X(x,t) \quad \text{for } x\in[-a,a] \\ \dfrac{\partial\phi}{\partial x}(-a,z,t) = \dot{U} \quad \text{for } z\in[0,\bar{h}] \\ \dfrac{\partial\phi}{\partial x}(a,z,t) = \dot{U} \quad \text{for } z\in[0,\bar{h}]. \end{cases} \tag{44}$$

The differential part $dX/dt = f$ corresponds to

$$\frac{\partial^2\phi}{\partial t^2}(x,\bar{h},t) = -g\frac{\partial\phi}{\partial z}(x,\bar{h},t) \quad \text{for } x\in[-a,a]$$

The system is of “index 1” since the “algebraic part” is invertible with respect to the “algebraic variables” Y: ϕ is a linear function of X and U by solving (44). Such implicit formulations of “index one” remain valid when the fluid is irrotational and described by the following nonlinear Euler equations (see, e.g., [14, pp:431–436]):

$$\begin{aligned} \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial z^2} &= 0 \quad \text{for } -a\leq x\leq a, \quad 0\leq z\leq h(x,t) \\ \frac{\partial\phi}{\partial z}(x,0,t) &= 0 \quad \text{for } x\in[-a,a] \\ \left[\frac{\partial\phi}{\partial t} + \frac{1}{2}\left(\left(\frac{\partial\phi}{\partial x}\right)^2 + \left(\frac{\partial\phi}{\partial z}\right)^2\right)\right]_{(x,h(x,t),t)} + gh(x,t) &= 0 \quad \text{for } x\in[-a,a] \\ \frac{\partial\phi}{\partial z}(x,h(x,t),t) - \frac{\partial h}{\partial x}(x,t)\frac{\partial\phi}{\partial x}(x,h(x,t),t) - \frac{\partial h}{\partial t}(x,t) &= 0 \quad \text{for } x\in[-a,a] \\ \frac{\partial\phi}{\partial x}(-a,z,t) &= \dot{D}(t) \quad \text{for } z\in[0,h(-a,t)] \\ \frac{\partial\phi}{\partial x}(a,z,t) &= \dot{D}(t) \quad \text{for } z\in[0,h(a,t)] \end{aligned}$$

with $z = h(x,t)$ the free surface equation (the profiles $\zeta(x,t) = \phi(x,h(x,t),t)$ and $h(x,t)$ corresponding then to the “differential variables” X).

Very few results (see [33] for a first result on a closely related problem) are available concerning the controllability and stabilization of such implicit systems of infinite dimension. Are such systems steady-state controllable?

Acknowledgements

This work has been inspired by fruitful discussions with K.J. Åström during a visit organized at the University of Lund by the “Conférence des Grandes Écoles”. We thank Mattias Grundelius from University of Lund for useful references on the slosh problem.

We gratefully thank Jean-Michel Coron from the Université Paris-Sud and Michel Fliess from the École Normale Supérieure de Cachan for many important discussions and advice on this work.

References

[1] J. T. Feddema, C. R. Dohrmann, Gordon G. Parker, R. D. Robinett, V. J. Romero, and D. J. Schmitt, “Control for slosh-free motion of an open container,” *IEEE Control Systems*, vol. 17, no. 1, pp. 29–36, Feb. 1997.

[2] K. Yano, T. Yoshida, M. Hamaguchi, and K. Terashima, “Liquid container transfer considering the suppression of sloshing for the change of liquid level,” in *Proceedings of the 13th IFAC World Congress*, San Francisco, California, July 1996.

[3] R. Venugopal and D. S. Bernstein, "State space modeling and active control of slosh," in *Proceedings of the 1996 IEEE International Conference on Control Applications*, Dearborn, Michigan, Sept. 1996, pp. 1072–1077.

[4] M. Grundelius and B. Bernhardsson, "Motion control of open containers with slosh constraints," in *Proceedings of the 14th IFAC World Congress*, Beijing, P.R. China, July 1999.

[5] M. Grundelius and B. Bernhardsson, "Control of liquid slosh in an industrial packaging machine," in *Proceedings of the 1999 IEEE International Conference on Control Applications and IEEE International Symposium on Computer Aided Control System Design*, Kohala Coast, Hawaii, Aug. 1999.

[6] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon, "Flatness and defect of nonlinear systems: introductory theory and examples," *Int. J. Control*, vol. 61, no. 6, pp. 1327–1361, 1995.

[7] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon, "A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems," *IEEE Trans. Automat. Control*, vol. 44, pp. 922–937, 1999.

[8] H. Mounier, *Propriétés structurelles des systèmes linéaires à retards: aspects théoriques et pratiques*, Ph.D. thesis, Université Paris Sud, Orsay, 1995.

[9] M. Fliess, H. Mounier, P. Rouchon, and J. Rudolph, "Systèmes linéaires sur les opérateurs de Mikusiński et commande d'une poutre flexible," in *ESAIM Proc. "Élasticité, viscolélasticité et contrôle optimal", 8ème entretiens du centre Jacques Cartier, Lyon*, 1996, pp. 157–168.

[10] Ph. Martin, R. M. Murray, and P. Rouchon, "Flat systems," in *Proc. of the* 4*th European Control Conf.*, Brussels, 1997, pp. 211–264, Plenary lectures and Mini-courses.

[11] B. Laroche, Ph. Martin, and P. Rouchon, "Motion planing for the heat equation," *Int. Journal of Robust and Nonlinear Control*, vol. 10, pp. 629–643, 2000.

[12] N. Petit, *Systèmes à retards, platitude en génie des procédés et contrôle de certaines équations des ondes*, Ph.D. thesis, Ecole des Mines de Paris, 2000.

[13] P. Rouchon, "Motion planning, equivalence, infinite dimensional systems," *Int. J. Applied Mathematics and Computer Science*, vol. 11, no. 1, pp. 165–188, 2001.

[14] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, Inc., 1974.

[15] J. Glimm and P. D. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, vol. 101 of *Memoirs of the American Mathematical Society*, AMS, Providence, Rhode Island, 1970.

[16] J.-M. Coron, G. Bastin and B. D'Andréa-Novel, "A Lyapunov approach to control irrigation canals modeled by saint-venant equations," in *Proc. European Control Conference, Karlsruhe*, 1999.

[17] J.-M. Coron, "Return method: some applications to flow control," Sept. 2000, Université d'Orsay, Paris-Sud.

[18] F. Dubois, N. Petit, and P. Rouchon, "Motion planing and nonlinear simulations for a tank containing a fluid," in *European Control Conference, Karlsruhe*, 1999.

[19] D. Russel, "Controllability and stabilization theory for linear partial differential equations: recent progress and open questions," *SIAM reviews*, vol. 20, no. 4, pp. 639–739, 1978.

[20] J.-L. Lions, "Exact controllability, stabilization and perturbations for distributed systems," *SIAM Rev.*, vol. 30, pp. 1–68, 1988.

[21] V. Komornik, *Exact Controllability and Stabilization; the Multiplier Method*, vol. 36 of *Res. Appl. Math.*, Wiley-Masson, 1994.

[22] I. Lasiecka and R. Triggiani, "Exact controllability of the wave equation with Neuman boundary control," *Appl. Math. Optim.*, vol. 19, pp. 243–290, 1989.

[23] R. F. Curtain and H. J. Zwart, *An Introduction to infinite-Dimensional Linear Systems Theory*, Text in Applied Mathemtics, 21. Springer-Verlag, 1995.

[24] N. Petit and P. Rouchon, "Dynamics and solutions to some control problems for water-tank systems," CDS Technical Memo CIT-CDS 00-004, California Institute of Technology, Pasadena, CA 91125, Nov. 2000.

[25] V. I. Arnol'd, "Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits," *Ann. Inst. Fourier*, vol. 16, pp. 319–361, 1966.

[26] Y. Yourgrau and S. Mandelstam, *Variational Principles in Dynamics and Quantum Theory*, Dover, New-York, third edition, 1979.

[27] M. Fliess and H. Mounier, "Controllability and observability of linear delay systems: an algebraic approach," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 3, pp. 301–314, 1998.

[28] A. Angot, *Compléments de mathématiques*, Editions de la revue d'optique, Paris, third edition, 1957.

[29] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particules and Rigid Bodies (4th edition)*, Cambridge University Press, Cambridge, 1937.

[30] W. C. Rheinboldt, "Differential-algebraic systems as differential equations on manifolds," *Mathematics of Computation*, vol. 43, pp. 473–482, 1984.

[31] R. F. Sincovec, A. M. Erisman, E. L. Yip, and M.A. Epton, "Analysis of descriptor systems using numerical algorithms," *IEEE Trans. Automat. Control*, vol. 26, pp. 139–147, 1981.

[32] M. Fliess, J. Lévine, and P. Rouchon, "Index of an implicit time-varying differential equation: a noncommutative linear algebraic approach," *Linear Algebra and its Applications*, vol. 186, pp. 59–71, 1993.

[33] S. Mottelet, "Controllability and stabilization of a canal with wave generators," *SIAM J. Control Optim.*, vol. 38, no. 3, pp. 711–735, 2000.

[34] K. Yosida, *Lectures on Differential and Integral Equations*, Dover, New York, 1960.

Appendix

Technical lemma

Lemma 7: Take $\mathbf{R} \ni x \mapsto c(x)$ a strictly positive smooth function and consider for each $s \in \mathbf{C}$, $x \mapsto A(x,s)$, the solution of

$$\frac{\partial}{\partial x}\left(c^2(x)\frac{\partial A}{\partial x}\right) = s^2 A$$

$$A(0,s) = a$$

$$\frac{\partial A}{\partial x}(0,s) = b$$

with $(a,b) \in \mathbf{R}^2$. Set $\sigma(x) = \displaystyle\int_0^x \frac{d\xi}{c(\xi)}$. Then for each x, there exists an L^2 function $[-\sigma(x), \sigma(x)] \ni \xi \mapsto \mathcal{B}(x,\xi) \in \mathbf{R}$ such that

$$A(x,s) = a\sqrt{\frac{c(0)}{c(x)}}\cosh(s\sigma(x)) + \int_{-\sigma(x)}^{\sigma(x)} \mathcal{B}(x,\xi)\exp(\xi s)d\xi.$$

Proof: The proof of this result is organized as follows

1. A Liouville transform, $x \mapsto z$ and $A \mapsto u$, is performed.
2. Using a majoring series we prove that, for each z, $s \mapsto u(z,s)$ is an entire functions of exponential kind.
3. We show that for any given $z \in [0,1]$, $\imath\mathbf{R} \ni s \mapsto u(z,s)$ is, up to some addition of exponentials, in L^2
4. We conclude thanks to the Paley-Wiener theorem.

Remark 6: $\mathcal{B}$ depends on a and b as detailed below. $\mathcal{B} = 0$ if $(a,b) = (0,0)$.

Liouville transform

The Liouville transform

$$(x, A) \mapsto (z, u)$$

(see for instance [34, page 110]) turns the equations

$$\frac{d}{dx}\left(p(x)\frac{dA}{dx}\right) + (\lambda r(x) - q(x))\,A = 0,$$

where $p(x) > 0$ into the following form

$$\frac{d^2u}{dz^2} + (\rho^2 - h(z))u = 0$$

where ρ depends only on λ and can be considered as a parameter.

Here

$$p(x) = c^2(x), \quad \lambda = -s^2, \quad r(x) = 1, \quad q(x) = 0, \quad x \in [0, L].$$

With the change of variables

$$z = \int_0^x \frac{1}{c}, \quad u(z,s) = (c(x))^{1/2}A(x,s)$$

we obtain

$$H(z) = \frac{F''(z)}{F(z)} \quad \text{with } F(z) = \sqrt{c(x)}.$$

We have turned

$$\begin{cases} \dfrac{\partial}{\partial x}\left(c^2(x)\dfrac{\partial A}{\partial x}\right) = s^2 A \\ A(0,s) = a \\ \dfrac{\partial A}{\partial x}(0,s) = b \end{cases} \tag{45}$$

into

$$\begin{cases} \dfrac{d^2u}{dz^2} - (h(z) + s^2)u = 0 \\ u(0,s) = \alpha \\ \dfrac{du}{dz}(0,s) = \beta \end{cases} \tag{46}$$

with

$$\alpha = u(0,s) = a(c(0))^{1/2}, \quad \beta = \frac{du}{dz}(0,s) = \frac{c'(0)(c(0))^{1/2}}{2}a + c(0)^{3/2}b$$

Proving that $\mathbf{C} \ni s \mapsto u(z,s)$ is an entire function of exponential type

Let $W(z,s)$ the 2×2 matrix solution of

$$\frac{dW}{dz} = \begin{pmatrix} 0 & 1 \\ h(z)+s^2 & 0 \end{pmatrix} W \text{ with } W(0,s) = I$$

Then $u(z,s) = \begin{pmatrix} 1 & 0 \end{pmatrix} W(z,s) \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. Let us show that W is entire with respect to s. By the classical fixed point technique $W(z,s) = \sum_{i\geq 0} W_i(z,s)$ with the following recurrence

$$W_0(z,s) = I, W_{i+1}(z,s) = \int_0^z \begin{pmatrix} 0 & 1 \\ h(z)+s^2 & 0 \end{pmatrix} W_i(\xi,s)d\xi$$

Each $W_i(z,s)$ is a polynomial in s^2. Its degree is $2i$ and its coefficients depend only on z. Reordering all the terms we get

$$\sum_{0\leq i\leq k} W_i(z,s) = \sum_{0\leq j\leq k} W^{j,k}(z)s^{2j}.$$

From step k to $k+1$ we have

$$W^{j,k+1}(z) = W^{j,k}(z) + \mathcal{W}^{j,k+1}(z)$$

where $\mathcal{W}^{j,k+1}$ is the coefficient of s^{2j} in W_{k+1}.

Take $K>0$ and $z\in[0,K]$. Set $m = \sup_{[0,K]} \mid h \mid$ and define the following majoring series by the recurrence

$$M_0(z,s) = I, M_{i+1}(z,s) = \int_0^z \begin{pmatrix} 0 & 1 \\ m+s^2 & 0 \end{pmatrix} M_i(\xi,s)\; d\xi$$

As previously we define

$$\sum_{0\leq i\leq k} M_i(z,s) = \sum_{0\leq j\leq k} M^{j,k}(z)s^{2j}, M^{j,k+1}(z) = M^{j,k}(z) + \mathcal{M}^{j,k+1}(z)$$

By classical matrix computations we get

$$M(z,s) = \begin{pmatrix} \cosh(z\sqrt{m+s^2}) & \sinh(z\sqrt{m+s^2})/\sqrt{m+s^2} \\ \sinh(z\sqrt{m+s^2})\sqrt{m+s^2} & \cosh(z\sqrt{m+s^2}) \end{pmatrix}.$$

For each j, the matrices $M^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{M}^{j,l}$ converge as k tends to ∞. Denote by M^j the limit. By construction, $M = \sum_{j\geq 0} M^j(z)\; \rho^{2j}$ and this series has an infinite radius of convergence in ρ, since, for each z, the functions $s\mapsto \cosh(z\sqrt{m+s^2})$, $s\mapsto \sinh(z\sqrt{m+s^2})/\sqrt{m+s^2}$ and $s\mapsto \sinh(z\sqrt{m+s^2})\sqrt{m+s^2}$ are entire functions of s^2.

But, for each i, j and k, the matrices $M^{j,k}$ and $\mathcal{M}^{j,k+1}$ whose entries are always non-negative, dominate the absolute values of the entries of $W^{j,k}$ and $\mathcal{W}^{j,k+1}$, respectively. Thus for each j, the matrices $W^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{W}^{j,l}$ converge as k tends to ∞. Denote by W^j the limit. By construction, $W = \sum_{j\geq 0} W^j(z)\rho^{2j}$ and this series has an infinite radius of convergence in ρ, since M has one. In other words, W is an entire function of ρ. Moreover the entries of M are upper bounds of the entries of W. Thus W is of exponential type in ρ: for each $z\in[0,K]$, there exists $E>0$ such that

$$\forall s\in\mathbf{C},\quad |W(z,s)| \leq E\exp(z|s|).$$

We have proven that, for each $z\in[0,\pi]$, $u(z,s)$ is an entire function of s with exponential type :

$$\forall s\in\mathbf{C},\quad u(z,s) \leq b(z)\exp(z|s|) \tag{47}$$

for some $b(z)>0$ well chosen.

Proving that “a part” of $i\mathbf{R} \ni s \mapsto u(z,s)$ belongs to L^2

From the Volterra equation of the second kind satisfied by u (see for instance [34, p. 111]),

$$u(z,s) = \alpha \cosh(sz) + \beta \frac{\sinh(sz)}{s} + \frac{1}{s} \int_0^z \sinh(s(z-\xi)) h(\xi) u(z,s) d\xi \tag{48}$$

Denote

$$w(z,s) = u(z,s) - \alpha \cosh(sz) \tag{49}$$

From (48) we deduce

$$w(z,s) = \phi(z,s) + \frac{1}{s} \int_0^z \sinh(s(z-\xi)) h(\xi) w(z,s) d\xi \tag{50}$$

with

$$\phi(z,s) = \beta \frac{\sinh(sz)}{s} + \frac{1}{s} \int_0^z \sinh(s(z-\xi)) h(\xi) \alpha \cosh(s\xi) d\xi.$$

Clearly, there exists D such that for all $z \in [0,K]$ and $s \in \imath\mathbf{R}$,

$$\mid \phi(z,s) \mid \leq \frac{D}{1+ \mid s \mid}$$

(h is bounded). Let us show that for any given z, $i\mathbf{R} \ni s \mapsto w(z,s)$ is in L^2. To prove this we use the following classical majoring arguments (see [34, p. 112] for instance). Denote

$$\mu(z,s) = \sup_{0 \leq \xi \leq z} \mid w(\xi,s) \mid$$

We deduce from (50) that

$$\mu(z,s) \leq \frac{D}{1+ \mid s \mid} + \frac{1}{\mid s \mid} m \mu(z,s) K$$

for $m = \sup_{[0,K]} \mid h \mid$. So

$$\mu(z,s) \leq \frac{D}{(1+ \mid s \mid)} \frac{1}{1 - \frac{mK}{|s|}}.$$

And for $\mid s \mid \geq 2mK$

$$\mu(z,s) \leq \frac{2D}{1+ \mid s \mid}$$

which proves that $i\mathbf{R} \ni s \mapsto w(z,s)$ is in L^2.

Use of the Paley-Wiener theorem

$i\mathbf{R} \ni s \mapsto w(z,s)$ is in L^2 and is an entire function of s of exponential type such that $\mid w(z,s) \mid \leq d(z) \exp(z \mid s \mid)$. Thanks to the Paley-Wiener theorem we can conclude that for each z there exists $[-z,z] \ni \xi \mapsto \mathcal{K}(z,\xi)$ in $L^2[-z,z]$ such that

$$w(z,s) = \int_{-z}^{z} \mathcal{K}(z,\xi) \exp(\xi s) \ d\xi.$$

Then

$$u(z,s) = \alpha \cosh(sz) + \int_{-z}^{z} \mathcal{K}(z,\xi) \exp(\xi s) d\xi.$$

and

$$\begin{aligned} A(x,s) &= \frac{1}{\sqrt{c(x)}} u\left(\sigma(x), s\right) \\ &= \frac{\alpha}{\sqrt{c(x)}} \cosh(s\sigma(x)) + \frac{1}{\sqrt{c(x)}} \int_{-\sigma(x)}^{\sigma(x)} \mathcal{K}(\sigma(x),\xi) \exp(\xi s) d\xi \\ &= a\sqrt{\frac{c(0)}{c(x)}} \cosh(s\sigma(x)) + \int_{-\sigma(x)}^{\sigma(x)} \mathcal{B}(x,\xi) \exp(\xi s) d\xi \end{aligned} \tag{51}$$

with $z = \sigma(x) = \displaystyle\int_0^x \frac{1}{c}$ and $\mathcal{B}(x,\xi) = \mathcal{K}(\sigma(x), s)/\sqrt{c(x)}$. ■

I. Simulations

In this section, we report some results of [18]. They correspond to numerical simulations (Godunov scheme) of the $1D$ nonlinear Saint-Venant equations 12 with $\theta = 0$ with the open-loop control $u = \ddot{D}$ of formula (31) and based on the linear tangent equations. This simulation indicates that, when the tank motion is not to fast the neglected nonlinearity are not very important. Several other simulation show that our open-loop control design is effective when $\sup |\ddot{D}|/ga \ll \bar{h}$.

In the following, Δ, which is the required time for a wave to meet a boundary starting from the opposite one, is equal to 1. The vertical scale of the figures has been enlarged by a factor 3 for the reader to see the details.

A. Transfer time T=4.0

The prediction of a slow move is rather close to the numerical results of a Godunov scheme simulation. Results are shown on figure 7.

B. Transfer time T=2.5

Yet as the move speeds up the prediction results get more different from the numerical simulation. Results are shown on figure 8.

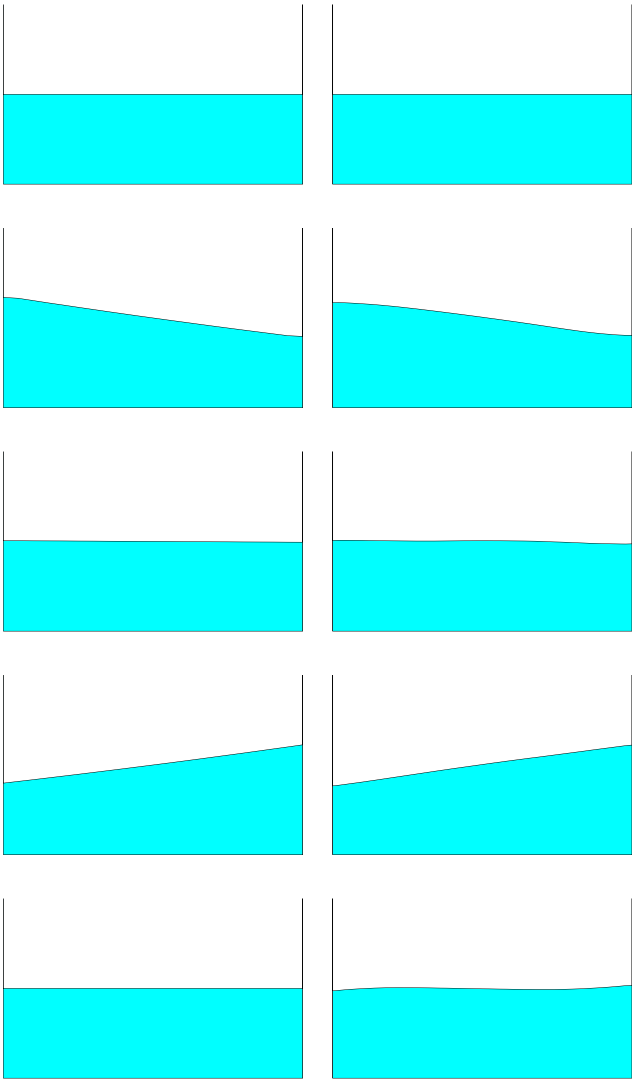


Fig. 7. T=4.0; snapshots at t=0, t=T/4, t=T/2, t=3T/4 and t=T. Left: linear prediction. Right: nonlinear simulation.

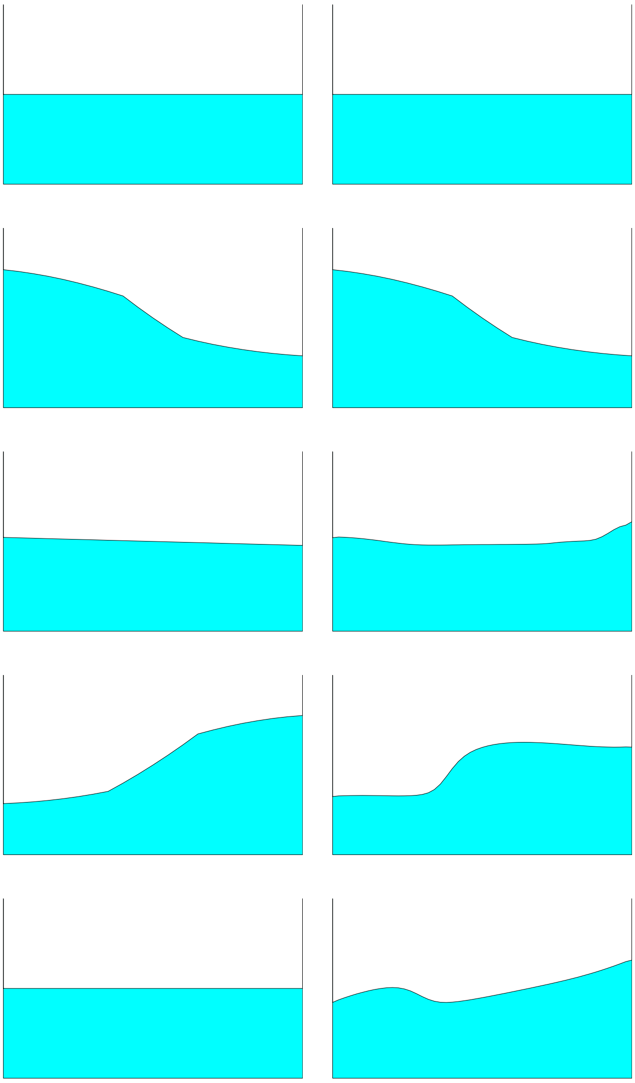


Fig. 8. T=2.5; snapshots at t=0, t=T/4, t=T/2, t=3T/4 and t=T. Left: linear prediction. Right: nonlinear simulation.

SIAM J. Control Optim.
Vol. 40, No. 2, pp. 475–495

FLATNESS OF HEAVY CHAIN SYSTEMS*

NICOLAS PETIT† AND PIERRE ROUCHON†

Abstract. In this paper the *flatness* [M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Internat. J. Control*, 61 (1995), pp. 1327–1361, M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *IEEE Trans. Automat. Control*, 44 (1999), pp. 922–937] of heavy chain systems, i.e., trolleys carrying a fixed length heavy chain that may carry a load, is addressed in the partial derivatives equations framework. We parameterize the system trajectories by the trajectories of its free end and solve the motion planning problem, namely, steering from one state to another state. When considered as a finite set of small pendulums, these systems were shown to be flat [R. M. Murray, in *Proceedings of the IFAC World Congress*, San Francisco, CA, 1996, pp. 395–400]. Our study is an extension to the infinite dimensional case.

Under small angle approximations, these heavy chain systems are described by a one-dimensional (1D) partial differential wave equation. Dealing with this infinite dimensional description, we show how to get the explicit parameterization of the chain trajectory using (distributed and punctual) advances and delays of its free end.

This parameterization results from symbolic computations. Replacing the time derivative by the Laplace variable s yields a second order differential equation in the spatial variable where s is a parameter. Its fundamental solution is, for each point considered along the chain, an entire function of s of exponential type. Moreover, for each, we show that, thanks to the Liouville transformation, this solution satisfies, modulo explicitly computable exponentials of s, the assumptions of the Paley–Wiener theorem. This solution is, in fact, the transfer function from the flat output (the position of the free end of the system) to the whole state of the system. Using an inverse Laplace transform, we end up with an explicit motion planning formula involving both distributed and punctual advances and delays operators.

Key words. wave equation, delay systems, flatness, motion planning

AMS subject classification. 99C20

PII. S0363012900368636

Introduction. The notion of *flatness* [3, 4] has proven to be relevant in many problems where motion planning problems have been solved [10, 5]. The existence of a *flat output* is the key to explicit formulas that can be implemented as open-loop controllers. Many systems of engineering interest are flat. So far the dynamics under consideration have been nonlinear ordinary differential equations, constant of varying delay equations, or even partial differential equations. In these cases the open-loop controller expression involved algebraic computations, punctual advances and delays [11, 6, 12], distributed advance and delay operators [12, 5, 14, 16], composition of functions [15], etc. In this paper we use both distributed and punctual advances and delays operators.

The heavy chain systems under consideration in this paper are defined by a trolley carrying a fixed length heavy chain to which a load may be attached. The dynamics are studied in a fixed vertical plane. When approximated as a finite set of small pendulums, such heavy chain systems were shown to be flat (see [13]). Their trajectories can be explicitly parameterized by the trajectories of their free ends. These parameterizations involve numerous derivatives (twice as many as the number of pendulums). When this number goes to infinity, the derivative order goes to infinity as

*Received by the editors February 24, 2000; accepted for publication (in revised form) February 26, 2001; published electronically July 19, 2001.

http://www.siam.org/journals/sicon/40-2/36863.html

†Centre Automatique et Systèmes, École des Mines de Paris, 60, bd. Saint-Michel, 75272, Paris Cedex 6, France (petit@cas.ensmp.fr, rouchon@cas.ensmp.fr).

well, yielding series expansions. This makes these relations difficult to handle and to use in practice.

In order to overcome these difficulties, we consider infinite dimensional descriptions of heavy chain systems. Around the stable vertical steady-state and under the small angle assumption, the dynamics are described by second order ordinary differential equations (dynamics of the load at position $y(t)$) coupled with one-dimensional (1D) wave equations (dynamics of the chain $X(x,t)$), where wave speed depends on x, the spatial variable along the chain length.

This combined ordinary and partial differential equation description turns out to be a significant shortcut to an explicit motion planning formula. Instead of an infinite number of derivatives, the explicit parameterization of the trajectories involves a small number of both distributed and punctual advances and delays. The controllability of such hybrid systems could be analyzed via Hilbert's uniqueness method [8, 9], as done in [7]. The work presented here is also a constructive proof of the controllability of these systems in the sense that it provides the open-loop control for steering the system from any given state to any other state. In a real application it should be used as a feedforward term complemented by a closed-loop controller using the energy method as proposed in [2].

In the case of a single homogeneous heavy chain as depicted in Figure 1.1 (see section 1 for details), our explicit parameterization shows that the general solution of

$$\frac{\partial}{\partial x}\left(gx\frac{\partial X}{\partial x}\right) - \frac{\partial^2 X}{\partial t^2} = 0$$

is given by the integral

$$X(x,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\sqrt{x/g}\sin\theta)\ d\theta, \tag{0.1}$$

where $t \mapsto y(t)$ is any smooth-enough time function: $X(0,t) = y(t)$ corresponds then to the free end position; the control $u(t) = X(L,t)$ is the trolley position.

For the general cases, we show here that relationships similar to (0.1) exist. They are expressed by (2.2) and (3.2). The structure is similar, but the moving averages involve weights (i.e., kernels) depending on the mass distribution. More precisely, given any mass distribution along the chain and any punctual mass at $x = 0$, we prove that there is a one-to-one correspondence between the trajectory of the load $t \mapsto y(t) = X(0,t)$ and the trajectory of the whole system (namely, the cable and the trolley): $t \mapsto X(x,t)$ and $t \mapsto u(t) = X(L,t)$. This correspondence yields the explicit parameterization of the trajectories: $X(x,\cdot) = \mathcal{A}_x y$, where $\{\mathcal{A}_x\}$ is a set of operators including time derivations, advances, and delays. In other words, $(x,t) \mapsto (\mathcal{A}_x y)(t)$ verifies the system equations for any smooth function $t \mapsto y(t)$. For each x, the operator $\mathcal{A}_x$ admits compact support. Thus it is possible to steer the system from any initial point to any other point in finite time.

This parameterization results from symbolic computations. Replacing the time derivative by the Laplace variable s yields a second order differential equation in x with s as a parameter. For each x, its fundamental solution A_x is an entire function of s of exponential type. Furthermore, for each x we show, thanks to the Liouville transformation, that $s \mapsto A_x(s)$ satisfies the assumptions of the Paley–Wiener theorem, modulo explicitly computable exponentials of s.

The paper is organized as follows.

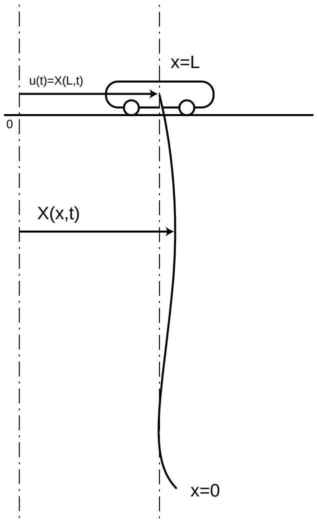


FIG. 1.1. *The homogeneous chain without any load.*

1. In section 1 we consider the case of a homogeneous chain without any load. Although it is the easiest case by far, it is explanatory, and it helps in understanding the meaning and control interest of our results.
2. In section 2 we address the case of an inhomogeneous chain without any load. The problem of the singularity at $x = 0$ of the second order differential equation receives special treatment. We prove the flatness of this system by Theorem 1.
3. In section 3 we solve the general problem of an inhomogeneous chain carrying a punctual load. By contrast with the previous case, the corresponding second order differential is not singular. Flatness of this system is proven by Theorem 2.

1. The homogeneous chain without any load. The computations are simple and explicit and summarize the goal of this paper.

Consider a heavy chain in stable position as depicted in Figure 1.1. Under the small angle approximation it is ruled by the dynamics[1]

$$\begin{cases} \dfrac{\partial}{\partial x}\left(gx\dfrac{\partial X}{\partial x}\right) - \dfrac{\partial^2 X}{\partial t^2} = 0, \\ X(L,t) = u(t), \end{cases} \tag{1.1}$$

where $x \in [0, L]$, $t \in \mathbb{R}$, $X(x,t) - X(L,t)$ is the deviation profile, g is the gravitational acceleration, and the control u is the trolley position.

Thanks to the classical mapping $y = 2\sqrt{\frac{x}{g}}$, we get

$$y\frac{\partial^2 X}{\partial y^2}(y,t) + \frac{\partial X}{\partial y}(y,t) - y\frac{\partial^2 X}{\partial t^2}(y,t) = 0.$$

[1] This model was used in the historical work of D. Bernoulli on a heavy chain system where the zero-order Bessel functions appear for the first time; see [18, pp. 3–4].

Use Laplace transform of X with respect to the variable t (denoted by $\hat{X}$ and with zero initial conditions, i.e., $X(.,0) = 0$ and $\frac{\partial X}{\partial t}(.,0) = 0$) to get

$$y \frac{\partial^2 \hat{X}}{\partial y^2}(y,s) + \frac{\partial \hat{X}}{\partial y}(y,s) - ys^2 \hat{X}(y,s) = 0.$$

Less classically, the mapping $z = \imath sy$ gives

$$z \frac{\partial^2 \hat{X}}{\partial z^2}(z,s) + \frac{\partial \hat{X}}{\partial z}(z,s) + z\hat{X}(z,s) = 0. \tag{1.2}$$

This is a Bessel equation. Its solution writes in terms of J_0 and Y_0 the zero-order Bessel functions. Using the inverse mapping $z = 2\imath s\sqrt{\frac{x}{g}}$, we get

$$\hat{X}(x,s) = A\ J_0(2\imath s\sqrt{x/g}) + B\ Y_0(2\imath s\sqrt{x/g}).$$

Since we are looking for a bounded solution at $x = 0$, we have $B = 0$. Then

$$\hat{X}(x,s) = J_0(2\imath s\sqrt{x/g})\hat{X}(0,s), \tag{1.3}$$

where we can recognize the Clifford function $\mathcal{C}_\prime$ (see [1, p. 358]). Using Poisson's integral representation of J_0 [1, formula 9.1.18],

$$J_0(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp(\imath z \sin\theta)\ d\theta,$$

we have

$$J_0(2\imath s\sqrt{x/g}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp(2s\sqrt{x/g}\sin\theta)\ d\theta.$$

In terms of Laplace transforms, this last expression is a combination of delay operators. Turning (1.3) back into the time-domain, we get

$$X(x,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\sqrt{x/g}\sin\theta)\ d\theta \tag{1.4}$$

with $y(t) = X(0,t)$.

Relation (1.4) means that there is a one-to-one correspondence between the (smooth) solutions of (1.1) and the (smooth) functions $t \mapsto y(t)$. For each solution of (1.1), set $y(t) = X(0,t)$. For each function $t \mapsto y(t)$, set X by (1.4) and u as

$$u(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\sqrt{L/g}\sin\theta)\ d\theta \tag{1.5}$$

to obtain a solution of (1.1).

Finding $t \mapsto u(t)$, steering the system from the steady-state $X \equiv 0$ at $t = 0$ to the other one $X \equiv D$ at $t = T$ becomes obvious. Our analysis shows that T must be larger than 2Δ, where $\Delta = 2\sqrt{L/g}$ is the travelling time of a wave between $x = L$ and $x = 0$. It consists only in finding $t \mapsto y(t)$ that is equal to 0 for $t \leq \Delta$ and to D for $t > T - \Delta$ and in computing u via (1.5).

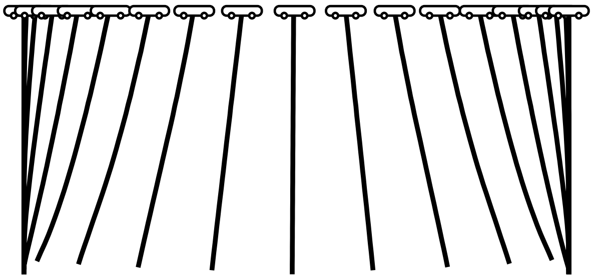


FIG. 1.2. *Steering from* 0 *to* $3L/2$ *in finite time* $T = 4\Delta$. *Regularly time-spaced positions of the heavy chain system are represented. The Matlab simulation code can be obtained from the second author via email.*

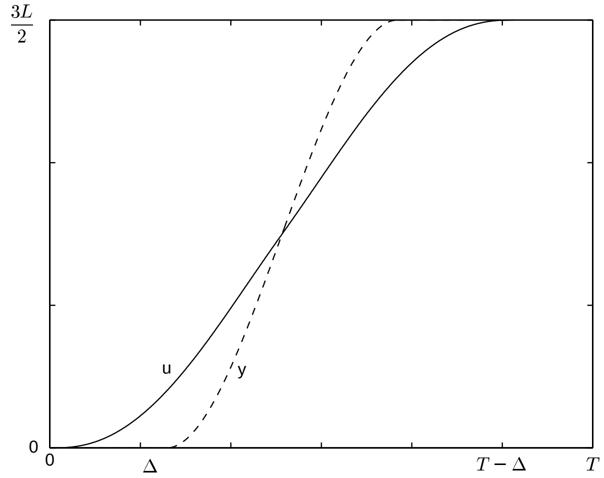


FIG. 1.3. *The steering control, trolley position* u, *and the "flat output," the free end* y.

Figure 1.2 illustrates computations based on (1.4) with

$$y(t) = \begin{cases} 0 \text{ if } t < \Delta, \\ \frac{3L}{2}\left(\frac{t-\Delta}{T-2\Delta}\right)^2 \left(3 - 2\left(\frac{t-\Delta}{T-2\Delta}\right)\right) \text{ if } \Delta \leq t \leq T - \Delta, \\ \frac{3L}{2} \text{ if } t > T - \Delta, \end{cases}$$

where the chosen transfer time T equals 4Δ. For $t \leq 0$ the chain is vertical at position 0. For $t \geq T$ the chain is vertical at position $D = 3L/2$.

Plots of Figure 1.3 show the control $[0, T] \ni t \mapsto u(t)$ required for such motion. Notice that the support of $\dot{u}$ is $[0, T]$, while the support of $\dot{y}$ is $[\Delta, T - \Delta]$. To be consistent with the small angle approximation, the horizontal acceleration of the end point $\ddot{y}$ must be much smaller than g. In our computations the maximum of $|\ddot{y}|$ is chosen rather large, $9g/16$. This is just for tutorial reasons. In practice, a reasonable transition time is $T = 5\Delta$ yielding $|\ddot{y}| \leq g/4$.

2. The inhomogeneous (i.e., variable section) chain without any load. Formula (1.4) can be extended to a heavy chain with variable section and carrying no load (see Figure 2.1). Such an extension deserves special consideration because of the singularity of the partial differential system at $x = 0$.

Such a system is governed by the equations

$$\left\{\begin{aligned} &\frac{\partial}{\partial x}\left(\tau(x)\frac{\partial X}{\partial x}\right) - \frac{\tau'(x)}{g}\frac{\partial^2 X}{\partial t^2} = 0,\\ &\qquad\qquad X(L,t) = u(t), \end{aligned}\right. \tag{2.1}$$

where $x \in [0, L]$, $t \in \mathbb{R}$, and u is the control. The tension of the chain is $\tau(x)$ with $\tau(0) = 0$ and $\tau(x) = gx + \mathcal{O}(x^2)$, while $\tau'(x)/g > 0$ is the mass distribution along the chain. Furthermore, we assume that there exists $a > 0$ such that $\tau(x) \geq ax \geq 0$.

THEOREM 1. *Consider* (2.1) *with* $[O, L] \ni x \mapsto \tau(x)$ *a smooth increasing function with* $\tau(0) = 0$ *and* $\tau' > 0$. *There is a one-to-one correspondence between the solutions* $[0, L] \times \mathbb{R} \ni (x, t) \mapsto (X(x, t), u(t))$ *that are* C^3 *in* t *and the* C^3 *functions* $\mathbb{R} \ni t \mapsto y(t)$ *via the formulas*

(2.2)

$$\begin{aligned} X(x,t) =& \frac{L^{1/4}\sqrt{g}}{2\pi^{3/2}(\tau(x)\tau'(x))^{1/4}}\sqrt{G(2\sqrt{\tau(x)/g})}\int_{-\pi}^{\pi} y\left(t + KG(2\sqrt{\tau(x)/g})\sin\theta\right) d\theta\\ &+ \frac{1}{(\tau(x)\tau'(x)/g)^{1/4}}\int_{-2\sqrt{\frac{\tau(x)}{ag}}}^{2\sqrt{\frac{\tau(x)}{ag}}} \mathcal{K}(G(2\sqrt{\tau(x)/g}),\xi)\ \dot{y}\left(t + \xi\right)\ d\xi,\\ u(t) =& X(L,t) \end{aligned}$$

with

$$y(t) = X(0, t),$$

where the constant K *and the functions* G *and* $\mathcal{K}$ *are defined by the function* τ *via formulas* (2.15) *and* (2.29).

The proof of this result is organized as follows.

1. A simple time-scaling simplifies the system. We shift from X to Y.
2. Symbolic computations where time derivatives are replaced by the Laplace variable s are performed.
3. The solution $Y(x, s)$ is factorized as $Y(x, s) = Y(0, s)A(x, s)$. A partial differential system is derived for $A(x, s)$.
4. A Liouville transformation is performed.
5. In these new coordinates the preceding transformed equation is compared to an equation that we have already solved in section 1, namely, the equation of a single homogeneous chain. We denote by $D(x, s)$ the difference between these two solutions.
6. $D(x, s)$ is proven to be an entire function of s and of exponential type.
7. A careful study of the Volterra equation satisfied by $D(x, s)$ shows that, for each x, the restriction to $D(x, s)/s$ to the imaginary axis is in L^2.
8. Thanks to the Paley–Wiener theorem, we prove that, for each x, $D(x, s)/s$ can be represented as a compact sum (discrete and continuous) of exponentials in s.
9. Gathering all the terms of $A(x, s)$, we get an expression involving the Bessel function J_0 (the solution for a homogeneous chain) and exponentials in s multiplied by s. This gives (2.2).

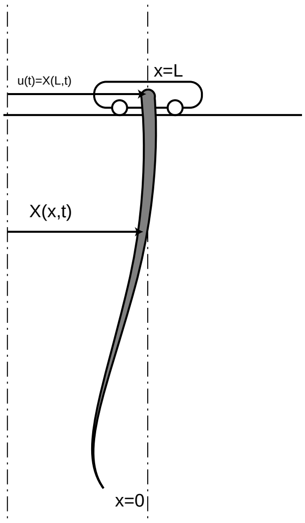


FIG. 2.1. *The inhomogeneous chain without any load.*

Proof. Simple change of coordinates Let[2] $Y(x,t) = X\left(\tau(x)/g, t\right)$.

[2] One may easily show the following result: if Y satisfies

$$\frac{\partial}{\partial x}\left(x\tau' \circ \tau^{-1}(gx)\frac{\partial Y}{\partial x}\right) - \frac{\partial^2 Y}{\partial t^2} = 0, \tag{2.3}$$

then $X(x,t) = Y(\tau(x)/g, t)$ satisfies

$$\frac{\partial}{\partial x}\left(\tau(x)\frac{\partial X}{\partial x}\right) - \frac{\tau'(x)}{g}\frac{\partial^2 X}{\partial t^2} = 0. \tag{2.4}$$

To show this, denote $\circ$ the composition operator with respect to the first variable. Thus $X = Y\circ(\tau/g)$. Then

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) = \frac{\partial}{\partial x}\left(\tau\tau'/g\frac{\partial Y}{\partial x}\circ(\tau/g)\right). \tag{2.5}$$

On the other hand, a factorization of (2.3) gives

$$\begin{aligned}\frac{\partial^2 Y}{\partial t^2} &= \frac{\partial}{\partial x}\left(\left(\tau/g\tau'\frac{\partial Y}{\partial x}\circ(\tau/g)\right)\circ\tau^{-1}(gx)\right)\\ &= \frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right)\frac{\partial}{\partial x}\left(\tau\tau'/g\frac{\partial Y}{\partial x}\circ(\tau/g)\right)\circ\tau^{-1}(gx).\end{aligned}$$

So by using (2.5)

$$\frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right)\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right)\circ\tau^{-1}(gx) = \frac{\partial^2 Y}{\partial t^2}.$$

Yet

$$\frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right) = \frac{g}{\tau'\circ\tau^{-1}(gx)},$$

so

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right)\circ\tau^{-1}(gx) = \frac{1}{g}\tau'\circ\tau^{-1}(gx)\frac{\partial^2 Y}{\partial t^2},$$

or, equivalently,

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) = \frac{\tau'}{g}\frac{\partial^2 Y}{\partial t^2}\circ(\tau/g) = \frac{\tau'}{g}\frac{\partial^2 X}{\partial t^2},$$

which gives the conclusion.

Now (2.1) gives

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial Y}{\partial x}\right) - \frac{\partial^2 Y}{\partial t^2} = 0, \tag{2.6}$$

where $\tau_1(x) = x\tau'(\tau^{-1}(gx))$.

Symbolic computations. Replacing the time derivation by s gives

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial Y}{\partial x}\right) - s^2 Y = 0. \tag{2.7}$$

Factorization. It is very easy to check that $Y(x,s) = Y(0,s)A(x,s)$ is the solution of (2.7), provided that $A(x,s)$ is solution of the following partial differential system:

$$\begin{cases} \dfrac{\partial}{\partial x}\left(\tau_1(x)\dfrac{\partial A}{\partial x}\right) - s^2 A = 0, \\ \quad A(0,s) = 1. \end{cases} \tag{2.8}$$

Existence of a solution. System (2.8) admits a smooth solution that is an entire function of exponential type in s. This solution reads

$$A(x,s) = \sum_{i\geq 0} \frac{s^{2i}}{i!} f_i(x), \tag{2.9}$$

where

$$\begin{cases} f_0 = 1, \\ f_i(x) = \displaystyle\int_0^x \frac{1}{\tau_1(l)} \int_0^l i f_{i-1}(s) ds \; dl. \end{cases} \tag{2.10}$$

It is very easy to check that, formally, $\sum_{i\geq 0} \frac{s^{2i}}{i!} f_i(x)$ is solution of (2.8): since

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial}{\partial x} f_i(x)\right) = i f_{i-1}(x),$$

we can write

$$\begin{cases} \dfrac{\partial}{\partial x}\left(\tau_1(x)\dfrac{\partial}{\partial x}\displaystyle\sum_{i\geq 0}\frac{s^{2i}}{i!} f_i(x)\right) = s^2 \displaystyle\sum_{i\geq 0}\frac{s^{2i}}{i!} f_i(x), \\ \displaystyle\sum_{i\geq 0}\frac{s^{2i}}{i!} f_i(0) = f_0(0) = 1. \end{cases} \tag{2.11}$$

Now let us address the convergence by proving that for all i

$$|f_i(x)| \leq \frac{1}{i!}\left(\frac{x}{a}\right)^i. \tag{2.12}$$

Suppose that (2.12) is true for a given i. (It is obviously the case for $i = 0$.) Let us inductively prove that it is also true for $i+1$. From (2.10) we get

$$|f_{i+1}(x)| \leq \int_0^x \frac{l^{i+1}}{\tau_1(l) a^i i!} dl.$$

Yet $\tau' \geq a$, so $\tau_1(x) \geq ax \geq 0$, and then

$$\begin{aligned}|f_{i+1}(x)| &\leq \int_0^x \frac{l^i}{a^{i+1}i!}dl\\ &\leq \frac{1}{(i+1)!}\left(\frac{x}{a}\right)^{i+1},\end{aligned}$$

which is (2.12) at rank $i+1$.

So, gathering (2.9) and (2.12) and using $\frac{1}{(i!)^2} \leq \frac{2^{2i}}{(2i)!}$, we get

$$A(x,s) \leq \sum_{i\geq 0} \frac{s^{2i}x^i}{(i!)^2 a^i} \leq \sum_{i\geq 0} \frac{s^{2i}2^{2i}x^i}{(2i)!a^i} \leq \exp\left(2s\sqrt{\frac{x}{a}}\right). \tag{2.13}$$

This proves that, for each x, $s \mapsto A(x,s)$ is an entire function of s of exponential type.

Liouville transformation. The Liouville transformation

$$(x, A) \mapsto (z, u)$$

(see, e.g., [19, p. 110]) turns equations of the form

$$\frac{d}{dx}\left(p(x)\frac{dA}{dx}\right) + (\lambda r(x) - q(x))\,A = 0$$

with $p(x) > 0$ into

$$\frac{d^2u}{dz^2} + (\rho^2 - h(z))u = 0,$$

where ρ is depending only on λ and can be considered as a parameter.

Here

$$p(x) = \tau_1(x), \quad \lambda = -s^2, \quad r(x) = 1, \quad q(x) = 0, \quad x \in [0, L],$$

and the transformation is defined for each $x > 0$. Nevertheless, it can be extended to $x = 0$ because around 0, $\tau_1(x) \approx gx$ with $g > 0$. It turns (2.8) into

$$\frac{d^2u}{dz^2} - K^2s^2u = \bar{h}(z)u \tag{2.14}$$

with

$$z = \frac{1}{K}\int_0^x \sqrt{\frac{1}{\tau_1}} \equiv G(2\sqrt{x}), \quad K = \frac{1}{\pi}\int_0^L \sqrt{\frac{1}{\tau_1}}, \tag{2.15}$$

$$u(z,s) = (\tau_1(x))^{1/4} A(x,s), \tag{2.16}$$

$$\bar{h}(z) = \frac{F''(z)}{F(z)} \quad \text{with } F(z) \equiv (\tau_1(x))^{1/4}. \tag{2.17}$$

Notice that since $\tau_1(x) \geq ax$ with $a > 0$, $\int_0^x 1/\tau_1$ is a smooth function of $\sqrt{x}$, and thus G is well defined and invertible. Similar arguments imply that $\bar{h}$ is, in fact, a function of z^2. Thus $\bar{h}(z) = h(z^2)$, and we have the following Laurent series around 0:

$$\bar{h}(z) = h(z^2) = \frac{-1}{4z^2} + \mathcal{O}(1).$$

Comparison to a simpler solution. We know from [1, formula 9.1.49, p. 362] that

$$u_0(z,s) = (Lg)^{1/4}\sqrt{\frac{z}{\pi}}\,J_0(iKsz) \tag{2.18}$$

satisfies

$$\frac{d^2u_0}{dz^2} - K^2s^2u_0 = \left(\frac{-1}{4z^2}\right)u_0. \tag{2.19}$$

According to the Laurent series of $\bar{h}$, we compare the solutions of (2.14), namely, $u(z,s)$, and (2.19), namely, $u_0(z,s)$. Let $D(z,s) = u(z,s) - u_0(z,s)$. We deduce from (2.14) and (2.19) that

$$\frac{d^2D}{dz^2} - K^2s^2D = \left(h(z^2) + \frac{1}{4z^2}\right)u_0 + h(z^2)D. \tag{2.20}$$

Since $z = G(2\sqrt{x})$ with G smooth and invertible, we have from (2.9) and (2.16)

$$u(z,s) = (Lg)^{1/4}\sqrt{\frac{z}{\pi}} + \mathcal{O}(z^{5/2}).$$

Then it is easy to check that for each s, D is a C^1 function of z around 0 with $D(0,s) = 0$ and $D'(0,s) = 0$. Equation (2.20) can be turned into the following integral equation (see [19, p. 111]):

$$\begin{aligned} D(z,s) =& \frac{1}{Ks}\int_0^z \sinh(Ks(z-t))\left(h(z^2) + \frac{1}{4t^2}\right)u_0(t,s)dt \\ &+ \frac{1}{Ks}\int_0^z \sinh(Ks(z-t))h(t^2)D(t,s)dt. \end{aligned} \tag{2.21}$$

Proving that $\mathbb{C} \ni s \mapsto D(z,s)$ is an entire function of exponential type. We already know that $A(x,s)$ and thus $u(z,s)$ (by (2.16)) are entire functions of exponential type in s. On the other hand, for each z, $s \mapsto u_0(z,s)$ is also an entire function of s of exponential type as J_0 is. This gives the conclusion.

Proving that $i\mathbb{R} \ni s \mapsto D(z,s)/s$ belongs to L^2. For each z, we need only an estimation of $D(z,iw)$ as w tends to ∞. For the sake of simplicity, we consider here $w \mapsto D(z,iw)$ for $w > 0$ large enough. The case $w < 0$ is similar. Classically (see, for instance, [19, p. 112]), let $M(z,w) = \sup_{0\le\zeta\le z}|D(\zeta,iw)|$. Using (2.21), we will get an estimation of $M(z,w)$. This gives

$$KwM(z,w) \le I_1(z,w) + I_2(z,w) \tag{2.22}$$

with

$$\begin{aligned} I_1(z,w) &= \int_0^z \left|h(t^2) + \frac{1}{4t^2}\right| |u_0(t,iw)|\,dt, \\ I_2(z,w) &= \int_0^z \left|h(t^2)\right| |D(t,iw)|\,dt. \end{aligned}$$

We know that

$$0 \le z \le \pi, \quad |u_0(t,iw)| \le (Lg)^{1/4}$$

since J_0 is bounded by 1 on the real axis. We know also that $h(t^2)+1/4t^2$ is bounded on $[0,\pi]$. Thus the integral I_1 is bounded by a constant $K_1 > 0$, independent of $z \in [0,\pi]$ and w,

$$I_1(z,w) \leq K_1. \tag{2.23}$$

Next, to majorate I_2 we split it into

$$I_2(z,w) = \underbrace{\int_0^{\gamma/w} \left|h(t^2)\right| |D(t,iw)|\, dt}_{I_2'(z,w)} + \underbrace{\int_{\gamma/w}^{z} \left|h(t^2)\right| |D(t,iw)|\, dt}_{I_2''(z,w)},$$

where $\gamma > 0$ is a parameter we will choose afterwards. A simple but quite tedious computation gives (using $J_0(z) = 1 - \frac{1}{4}z^2 + \circ(z^2)$)

$$D(z,s) = \sqrt{z}\; cs^2z^2(1+\mu(s^2z^2)),$$

where c is a constant and μ is a smooth function such that $\mu(0)=0$. Using this last expression in I_2', we get

$$I_2'(z,w) \leq \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right), \tag{2.24}$$

where $b > 0$ is such that $\left|h(t^2)\right| \leq b/(4t^2)$ for all $t \in]0,\pi]$. On the other hand, it is easy to check that

$$I_2''(z,w) \leq \frac{bw}{4\gamma}M(z,w). \tag{2.25}$$

Gathering (2.24) and (2.25), we get

$$I_2(z,w) \leq \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right) + \frac{bw}{4\gamma}M(z,w). \tag{2.26}$$

Thanks to the majorations (2.23) and(2.26), we get

$$KwM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right) + \frac{bw}{4\gamma}M(z,w).$$

This majoration is valid for $z \in]0,\pi]$, $w > 0$, and $\gamma > 0$ such that $\gamma/w \leq z$. Now we take

$$\gamma = \frac{b}{2K}.$$

Thus for each $z > 0$ and each $w > \gamma/z$, we have

$$(K - b/4\gamma)wM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right).$$

Since $K - b/4\gamma = K/2$, we have

$$\frac{1}{2}\; KwM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right). \tag{2.27}$$

Thus there exists $C_0 > 0$ such that for each $z \in]0, \pi]$ and for every $w > \gamma/z$,

$$|D(z, iw)| \leq \frac{C_0}{\sqrt{|w|}}. \tag{2.28}$$

Since $D(z, 0) \equiv 0$, we deduce for each $z > 0$ that $s \mapsto D(z, s)/s$ remains an entire function of s (of exponential type), and the above majoration says that $i\mathbb{R} \ni s \mapsto D(z, s)/s$ belongs to L^2.

Using the Paley–Wiener theorem. The Paley–Wiener theorem [17, p. 375] ensures that, for any $z \in [0, \pi]$, there exists $[-\frac{G^{-1}(z)}{\sqrt{a}}, \frac{G^{-1}(z)}{\sqrt{a}}] \ni t \mapsto \mathcal{K}(z, t)$ in L^2 such that

$$D(z, s)/s = \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} \mathcal{K}(z, \xi) \exp{(s\xi)} d\xi. \tag{2.29}$$

The integral bounds results from the following facts.

1. Via (2.16), $2\sqrt{x} = G^{-1}(z)$, and (2.13), we have

$$\forall s \in \mathbb{C}, \quad |(u(z, s)| \leq N(z) \exp\left(|s| \frac{G^{-1}(z)}{\sqrt{a}}\right)$$

   for some $N(z) > 0$.
2. A well-known property on J_0 implies that

$$\forall s \in \mathbb{C}, \quad |(u_0(z, s)| \leq N_0(z) \exp(|s|\, zK)$$

   for some $N_0(z) > 0$.
3. Since $\tau_1 x \geq ax$, (2.15) implies that $zK < \frac{G^{-1}(z)}{\sqrt{a}}$.
4. Thus

$$\forall s \in \mathbb{C}, \quad |D(z, s)| = |u(z, s) - u_0(z, s)| \leq (N(z) + N_0(z)) \exp\left(|s| \frac{G^{-1}(z)}{\sqrt{a}}\right).$$

Conclusion.

$$(u(z, s) - u_0(z, s))/s = \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} \mathcal{K}(z, \xi) \exp{(s\xi)} d\xi.$$

This gives

$$u(z, s) = \frac{(Lg)^{1/4}}{\sqrt{\pi}} \sqrt{z} J_0(iKsz) + \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} s\mathcal{K}(z, \xi) \exp(s\xi) d\xi.$$

Pulling back this relation in the (x, A) coordinates, we deduce using (2.16) that

$$\begin{aligned} A(x, s) =& \frac{(Lg)^{1/4}}{\sqrt{\pi}} \frac{1}{(\tau_1(x))^{1/4}} \sqrt{G(2\sqrt{x})} J_0(iKsG(2\sqrt{x})) \\ &+ \frac{1}{(\tau_1(x))^{1/4}} \int_{-2\sqrt{\frac{x}{a}}}^{2\sqrt{\frac{x}{a}}} s\mathcal{K}(G(2\sqrt{x}), \xi) \exp{(s\xi)} d\xi. \end{aligned}$$

Then we quickly get $Y(x,s) = Y(0,s)A(x,s)$. This gives in the time domain

$$\begin{aligned} Y(x,t) =& \frac{(Lg)^{1/4}}{\sqrt{\pi}} \frac{1}{(\tau_1(x))^{1/4}} \sqrt{G(2\sqrt{x})}\ \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(0, t + KG(2\sqrt{x}) \sin\theta) d\theta \\ &+ \frac{1}{(\tau_1(x))^{1/4}} \int_{-2\sqrt{\frac{x}{a}}}^{2\sqrt{\frac{x}{a}}} \mathcal{K}(G(2\sqrt{x}), \xi) \left[\frac{\partial}{\partial t} Y(0, t+\xi) \right] d\xi. \end{aligned}$$

Then substituting

$$\begin{aligned} X(x,t) &= Y\left(\tau(x)/g, t\right), \\ Y(0,t) &= X(0,t), \\ \frac{\partial Y}{\partial t}(0,t) &= \frac{\partial X}{\partial t}(0,t), \end{aligned}$$

we get

(2.30)
$$\begin{aligned} X(x,t) =& \frac{L^{1/4}\sqrt{g}}{2\pi^{3/2}(\tau(x)\tau'(x))^{1/4}} \sqrt{G(2\sqrt{\tau(x)/g})}\ \int_{-\pi}^{\pi} y(t + KG(2\sqrt{\tau(x)/g}) \sin\theta) d\theta \\ &+ \frac{1}{(\tau(x)\tau'(x)/g)^{1/4}} \int_{-2\sqrt{\frac{\tau(x)}{ag}}}^{2\sqrt{\frac{\tau(x)}{ag}}} \mathcal{K}(G(2\sqrt{\tau(x)/g}), \xi)\ \dot{y}\,(t+\xi)\ \ d\xi \end{aligned}$$

with $y(t) = X(0,t)$. □

Remark. In the case of a homogeneous chain, we can substitute

$$\begin{aligned} &\tau(x) = gx, \quad \tau'(x) = g, \quad \tau_1(x) = gx = \tau(x), \\ &K = \frac{2}{\pi}\sqrt{\frac{L}{g}}, \quad z = G(2\sqrt{x}) = \pi\sqrt{\frac{x}{L}}, \mathcal{K} = 0, \end{aligned}$$

and (2.30) reads

$$X(x,t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} y\left(t + 2\sqrt{\frac{x}{g}} \sin\theta\right) d\theta,$$

which is indeed identical to (1.4).

3. The inhomogeneous chain with punctual load. The system of Figure 3.1 consists of a heavy chain with a variable section carrying a punctual load m. Small deviations $X(x,t) - u(t)$ from the vertical position are described by the partial differential system

$$\text{(3.1)} \qquad \left\{ \begin{aligned} &\frac{\partial}{\partial x}\left(\tau(x)\frac{\partial X}{\partial x}\right) - \frac{\tau'(x)}{g}\frac{\partial^2 X}{\partial t^2} = 0, \\ &\frac{\partial^2 X}{\partial t^2}(0,t) = g\frac{\partial X}{\partial x}(0,t), \\ &X(L,t) = u(t), \end{aligned} \right.$$

where u is the control. The tension in the chain writes $\tau(x)$: $\tau(0) = mg$, and $\tau'(x)/g > 0$ is the mass distribution along the chain.

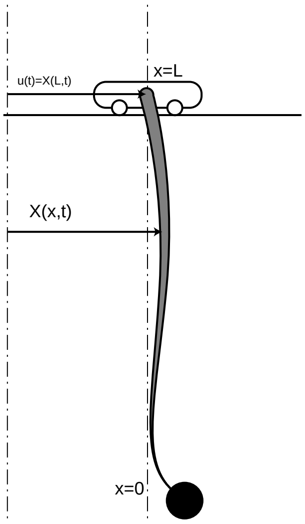


FIG. 3.1. *The inhomogeneous (variable section) chain with punctual load.*

THEOREM 2. *Consider* (3.1) *with* $[0, L] \ni x \mapsto \tau(x)$ *a smooth increasing function with* $\tau(0) = m$. *There is a one-to-one correspondence between the solutions* $[0, L] \times \mathbb{R} \ni (x, t) \mapsto (X(x, t), u(t))$ *that are* C^3 *in* t *and the* C^3 *functions* $\mathbb{R} \ni t \mapsto y(t)$ *via the following formulas:*

$$\left\{\begin{aligned} X(x,t) &= \phi(x)\ [y(t+\theta(x)) + y(t-\theta(x))] + \psi(x)\ [\dot{y}(t+\theta(x)) - \dot{y}(t-\theta(x))] \\ &\qquad\qquad + \int_0^x \mathcal{B}(x,\xi)[y(t+\theta(\xi)) + y(t-\theta(\xi))]\ d\xi, \\ u(t) &= X(L,t) \end{aligned}\right. \tag{3.2}$$

with

$$\begin{aligned} y(t) &= X(0,t), \\ \theta(x) &= \int_0^x \sqrt{\frac{\tau'}{g\tau}}, \\ \psi(x) &= \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}} \frac{1}{2}\sqrt{\frac{\tau(0)}{g\tau'(0)}}, \\ \phi(x) &= \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}} \cdots \\ &\quad \times \left[1 + \frac{1}{8}\sqrt{\frac{\tau(0)}{\tau'(0)}}\left(\left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)(x) - \left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)(0) \right.\right. \\ &\qquad\qquad\qquad \left.\left. + \cdots + \frac{1}{4}\int_0^x \left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)^2 \sqrt{\frac{\tau'}{\tau}}\right)\right], \end{aligned}$$

$B(x,\xi)$ a smooth function of x, and ξ defined by the function τ via formula (3.15).

Correspondence (3.2) defines a family of linear operators $\mathcal{A}_x$ with compact support such that, for any C^3 time function, $X(x,t) = \mathcal{A}_x y|_t$ is automatically the solution of (3.1) with $u(t) = X(L,t)$ and $X(0,t) = y(t)$.

The proof relies on the following points.

1. Symbolic computations where the time derivation is replaced by the Laplace variable s are performed. This yields a second order differential equation with nonconstant coefficients in the space variable x.
2. The solution $X(x,s)$ is factorized as $X(x,s) = X(0,s)A(x,s)$. A partial differential system is derived for $A(x,s)$.
3. The study of $s \mapsto A(x,s)$ is simplified by a Liouville transformation $(x,A) \mapsto (z,u)$.
4. The solution $A(x,s)$ of this differential equation is proven to be an entire function of s and of exponential type. (Volterra expansion and majoring series arguments are used.)
5. A careful study of the Volterra equation of the second kind satisfied by A shows that modulo some functions (exponentials of s, depending on x and explicitly calculated), for each x, the restriction of $A(x,s)$ to the imaginary axis is in L^2.
6. Thanks to the Paley–Wiener theorem and the last two properties of A, we prove that, for each x, A can be represented as a compact sum (discrete and continuous) of exponentials in s. This gives (3.2).

Proof. Symbolic computation. Replacing the time derivation by s gives

$$\left\{\begin{array}{l} \dfrac{\partial}{\partial x}\left(\tau(x)\dfrac{\partial X}{\partial x}\right) - \dfrac{\tau'(x)}{g}s^2 X = 0, \\ \qquad s^2 X(0,s) = gX'(0,s). \end{array}\right. \tag{3.3}$$

We do not consider the other boundary condition since u is the control and can be obtained explicitly from X via $u(t) = X(L,t)$.

Factorization. It is very easy to check that $X(x,s) = X(0,s)A(x,s)$ is the solution of (3.3), provided that $A(x,s)$ is the solution of the following partial differential system:

$$\left\{\begin{array}{l} \dfrac{\partial}{\partial x}\left(\tau(x)\dfrac{\partial A}{\partial x}\right) - \dfrac{\tau'(x)}{g}s^2 A = 0, \\ \qquad A(0,s) = 1, \\ \qquad gA'(0,s) = s^2. \end{array}\right. \tag{3.4}$$

Liouville transformation. This time we perform a Liouville transformation (already used in section 2)

$$(x,A) \mapsto (z,u)$$

with

$$p(x) = \tau(x), \quad \lambda = -\frac{s^2}{g}, \quad r(x) = \tau'(x), \quad q = 0, \quad x \in [0,L].$$

The new variables (z, u) are defined by the following formulas:

$$z = \frac{1}{K}\int_0^x \sqrt{\frac{\tau'}{\tau}}, \quad 0 \le z \le \pi, \quad K = \frac{1}{\pi}\int_0^L \sqrt{\frac{\tau'}{\tau}}, \tag{3.5}$$

$$u(z,s) = (\tau(x)\tau'(x))^{1/4} A(x,s). \tag{3.6}$$

System (3.4) is turned into

$$\frac{d^2u}{dz^2} + (\rho^2 - h(z))u = 0 \quad \text{with} \quad \frac{du}{dz}(0) = (a + b\rho^2), \quad u(0) = 1, \tag{3.7}$$

where

$$\begin{aligned} \rho &= \imath\, \frac{K}{\sqrt{g}} s, \quad \imath = \sqrt{-1}, \\ h(z) &= \frac{f''(z)}{f(z)} \quad \text{with } f(z) = (\tau(x)\tau'(x))^{1/4}, \\ a &= \frac{f'(0)}{f(0)}, \quad b = \frac{1}{K}\sqrt{\frac{\tau(0)}{\tau'(0)}}. \end{aligned}$$

Proving that $\mathbb{C} \ni \rho \mapsto u(z,\rho)$ is an entire function of exponential type. We claim that, for each z, $\rho \mapsto u(z,\rho)$ is an entire function of exponential type.

Denote by $W(z,\rho)$ the 2×2 matrix solution of

$$\frac{dW}{dz} = \begin{pmatrix} 0 & 1 \\ h(z) - \rho^2 & 0 \end{pmatrix} W$$

with $W(0,\rho) = I$. Since

$$u(z,\rho) = \begin{pmatrix} 1 & 0 \end{pmatrix} W(z,\rho) \begin{pmatrix} 1 \\ a + b\rho^2 \end{pmatrix},$$

it suffices to prove that W is entire in ρ and of exponential type. Using the classical fixed point technique, W can be expressed as an absolutely convergent series of iterated integrals (Volterra expansion)

$$W(z,\rho) = \sum_{i \ge 0} W_i(z,\rho)$$

with

$$W_0(z,\rho) = I, \quad W_{i+1}(z) = \int_0^z \begin{pmatrix} 0 & 1 \\ h(\sigma) - \rho^2 & 0 \end{pmatrix} W_i(\sigma,\rho)\, d\sigma. \tag{3.8}$$

For each $i > 0$, $W_i(z,\rho)$ is a polynomial in ρ^2 of degree i with coefficients depending on z. Thus we have

$$\sum_{0 \le i \le k} W_i(z,\rho) = \sum_{0 \le j \le k} W^{j,k}(z)\, \rho^{2j}.$$

From step k to $k+1$, we add to $W^{j,k}(z)$ the coefficient of ρ^{2j} in W_{k+1}, say, $\mathcal{W}^{j,k+1}$, to obtain $W^{j,k+1}(z)$:

$$W^{j,k+1}(z) = W^{j,k}(z) + \mathcal{W}^{j,k+1}(z).$$

Let $\alpha = \sup_{[0,\pi]} |h|$. Then the absolute value of each entry of $W_i(z,\rho)$ is bounded by the corresponding entries in the following *majoring series* $M_i(z,\rho)$ defined by the induction (to be compared to (3.8)):

$$M_0(z,\rho) = I, \quad M_{i+1}(z) = \int_0^z \begin{pmatrix} 0 & 1 \\ \alpha + \rho^2 & 0 \end{pmatrix} M_i(\sigma,\rho)\, d\sigma. \tag{3.9}$$

As for W, we can define $M = \sum_{i\geq 0} M_i$ and, for each $k > 0$, the matrices $M^{j,k}$ and $\mathcal{M}^{j,k+1}$ satisfying

$$\sum_{0\leq i\leq k} M_i(z,\rho) = \sum_{0\leq j\leq k} M^{j,k}(z)\ \rho^{2j}, \quad M^{j,k+1}(z) = M^{j,k}(z) + \mathcal{M}^{j,k+1}(z).$$

Standard matrix computations show that

$$\begin{aligned} M(z,\rho) = I &+ \sum_{i>0} \frac{z^{2i}}{(2i)!} \begin{pmatrix} (\rho^2+\alpha)^i & 0 \\ 0 & (\rho^2+\alpha)^i \end{pmatrix} \\ &+ \sum_{i>0} \frac{z^{2i+1}}{(2i+1)!} \begin{pmatrix} 0 & (\rho^2+\alpha)^i \\ (\rho^2+\alpha)^{i+1} & 0 \end{pmatrix}. \end{aligned}$$

That is,

$$M(z,\rho) = \begin{pmatrix} \cosh(z\sqrt{\rho^2+\alpha}) & \sinh(z\sqrt{\rho^2+\alpha})/\sqrt{\rho^2+\alpha} \\ \sinh(z\sqrt{\rho^2+\alpha})\sqrt{\rho^2+\alpha} & \cosh(z\sqrt{\rho^2+\alpha}) \end{pmatrix}. \tag{3.10}$$

For each j, the matrices $M^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{M}^{j,l}$ converge as k tends to ∞. Denote by M^j the limit. By construction, $M = \sum_{j\geq 0} M^j(z)\ \rho^{2j}$, and this series has an infinite radius of convergence in ρ, since, for each z, the functions $\rho \mapsto \cosh(z\sqrt{\rho^2+\alpha})$, $\rho \mapsto \sinh(z\sqrt{\rho^2+\alpha})/\sqrt{\rho^2+\alpha}$, and $\rho \mapsto \sinh(z\sqrt{\rho^2+\alpha})\sqrt{\rho^2+\alpha}$ are entire functions of ρ^2.

But, for each i, j, and k, the matrices $M^{j,k}$ and $\mathcal{M}^{j,k+1}$, whose entries are always nonnegative, dominate the absolute values of the entries of $W^{j,k}$ and $\mathcal{W}^{j,k+1}$, respectively. Thus for each j, the matrices $W^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{W}^{j,l}$ converge as k tends to ∞. Denote by W^j the limit. By construction, $W = \sum_{j\geq 0} W^j(z)\rho^{2j}$, and this series has an infinite radius of convergence in ρ, since M has one. In other words, W is an entire function of ρ. Moreover, the entries of M are upper bounds of the entries of W. Thus W is of exponential type in ρ: for each $z \in [0,\pi]$, there exists $E > 0$ such that

$$\forall \rho \in \mathbb{C}, \quad |W(z,\rho)| \leq E \exp(z|\rho|).$$

We have proven that, for each $z \in [0,\pi]$, $u(z,\rho)$ is an entire function of ρ of exponential type with

$$\forall \rho \in \mathbb{C}, \quad |u(z,\rho)| \leq b(z) \exp(z|\rho|)$$

for some $b(z) > 0$ well-chosen.

Proving that "a part" of $\mathbb{R} \ni \rho \mapsto u(z,\rho)$ *belongs to* L^2. In general, $\mathbb{R} \ni \rho \mapsto u(z,\rho)$ does not belong to L^2. Thus the Paley–Wiener theorem does not apply directly. Removing some appropriate terms, the remaining is in L^2.

Let

$$v(z,\rho) = u(z,\rho) + b\rho\sin(\rho z) - \left(1 + \frac{b\int_0^z h}{2}\right)\cos(\rho z). \tag{3.11}$$

In the following we prove that this entire function of exponential type is such that $\mathbb{R} \ni \rho \mapsto v(z,\rho)$ belongs to L^2.

From the Volterra equation of the second kind satisfied by u (see [19, p. 111]),

$$u(z,\rho) = \left(\cos(\rho z) + (a - b\rho^2)\frac{\sin(\rho z)}{\rho}\right) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))\ h(\zeta)\ u(\zeta,\rho)\ d\zeta,$$

we quickly derive a similar equation satisfied by v,

$$v(z,\rho) = \phi(z,\rho) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))\ h(\zeta)\ v(\zeta,\rho)\ d\zeta,$$

where $\phi = \phi_1 - b\phi_2$ with

$$\begin{aligned}\phi_1(z,\rho) =& a\frac{\sin(\rho z)}{\rho} + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))h(\zeta)\cos(\rho\zeta)\left(1 + (b/2)\int_0^\zeta h\right)\ d\zeta,\\ \phi_2(z,\rho) =& \cos(\rho z)\int_0^z h/2 + \int_0^z \sin(\rho(z-\zeta))h(\zeta)\sin(\zeta)\ d\zeta.\end{aligned}$$

Clearly, there exists $D_1 > 0$ such that for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$,

$$|\phi_1(z,\rho)| \le \frac{D_1}{1+|\rho|}$$

(h is bounded). With $2\sin(\rho(z-\zeta))\sin(\zeta) = \cos(\rho(z-2\zeta)) - \cos(\rho z)$, we have

$$\phi_2(z,\rho) = \int_0^z \cos(\rho(z-2\zeta))h(\zeta)\ d\zeta.$$

The integration by part (by assumption τ is C^4 thus h is C^1)

$$\int_0^z \cos(\rho(z-2\zeta))h(\zeta)\ d\zeta = \frac{h(0)+h(z)}{2\rho}\sin(\rho z) + \frac{1}{2\rho}\int_0^z \sin(\rho(z-2\zeta))h'(\zeta)\ d\zeta$$

shows that for large $|\rho|$, ϕ_2 tends to zero at least as $1/|\rho|$. Thus there exists $D_2 > 0$ such that for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$,

$$|\phi_2(z,\rho)| \le \frac{D_2}{1+|\rho|}.$$

This proves that v satisfies

$$v(z,\rho) = \phi(z,\rho) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))h(\zeta)v(\zeta,\rho)\ d\zeta \tag{3.12}$$

with $|\phi(z,\rho)| \le D/(1+|\rho|)$ for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$. ($D > 0$ is a well-chosen constant independent of z and ρ.)

This last inequality gives the desired conclusion by the following classical computation (see [19, p. 112], for instance).

Let $\beta(z,\rho) = \sup_{0\leq\zeta\leq z} |v(\zeta,\rho)|$. By (3.12) we have for each z_1 and z_2 in $[0,\pi]$, $z_1 \leq z_2$

$$|v(z_1,\rho)| \leq \frac{D}{1+|\rho|} + \frac{\alpha z_1 \beta(z_2,\rho)}{|\rho|} \leq \frac{D}{1+|\rho|} + \frac{\alpha\pi}{|\rho|}\,\beta(z_2,\rho).$$

(Remember that $\alpha = \sup_{[0,\pi]} |h|$.) In particular, when $z_1 = z_2 = z$, we have

$$\beta(z,\rho)\left(1 - \frac{\alpha\pi}{|\rho|}\right) \leq \frac{D}{1+|\rho|}. \tag{3.13}$$

Finally, for $|\rho| \geq 2\alpha\pi$, $\beta(z,\rho) \leq 2D/(1+|\rho|)$. This proves that $\mathbb{R} \ni \rho \mapsto v(z,\rho)$ belongs to L^2.

Using the Paley–Wiener theorem. At last, the Paley–Wiener theorem ensures that the Fourier transform of $\rho \mapsto v(z,\rho)$ has a compact support included in $[-z,z]$ since for all $\rho \in \mathbb{C}$, $|v(z,\rho)| \leq N \exp(z|\rho|)$ for some constant $N > 0$. This means that, for each $z \in [0,\pi]$, there exists $[-z,z] \ni \zeta \mapsto \mathcal{K}(z,\zeta)$ in $L^2([-z,z])$ such that

$$v(z,\rho) = \int_{-z}^{+z} \mathcal{K}(z,\zeta) \exp(\imath\zeta\rho)\, d\zeta.$$

Since v is an even function of ρ, $\mathcal{K}$ is also an even function of ζ. Thus we have, finally,

$$v(z,\rho) = \int_{0}^{+z} \mathcal{K}(z,\zeta)(\exp(\imath\zeta\rho) + \exp(-\imath\zeta\rho))\, d\zeta. \tag{3.14}$$

Conclusion. Pulling back this last relation in the (x,A) coordinates, noticing that $\rho = \imath Ks/\sqrt{g}$, that $\exp(-\theta s)$ is the Laplace transform of the θ-delay operator, and that $u(0,\rho)$ is, up to a constant, the Laplace transform of $X(0,t)$, we deduce after some standard but tedious computations formulae (3.2). The new function $\mathcal{B}(x,\xi)$ is related to $\mathcal{K}(z,\zeta)$ via

$$K\sqrt{\frac{\tau(\xi)}{\tau'(\xi)}}\,\mathcal{B}(x,\xi) = \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}} \mathcal{K}\left(\frac{\sqrt{g}}{K}\theta(x), \frac{\sqrt{g}}{K}\theta(\xi)\right). \tag{3.15}$$

At last,

$$\begin{aligned} A(x,s) = \varphi(x)\left(\exp\theta(x)s + \exp\theta(x)s\right) + \psi(x)s\left(\exp\theta(x)s - \exp\theta(x)s\right) \\ + \int_0^x \mathcal{K}(x,\zeta)(\exp(\theta(\zeta)s) + \exp(-\theta(\zeta)s))\, d\zeta, \end{aligned}$$

so $X(x,s) = X(0,s)A(x,s)$ when turned back into the time-domain does give formulae (3.2). □

4. Conclusion. We have shown that, around the stable vertical position, heavy chain systems with or without load, with constant or variable section, are "flat": the trajectories of these systems are parameterizable by the trajectories of their free ends. Relations (1.4), (2.2), and (3.2) show that such parameterizations involve operators of compact supports.

It is surprising that such parameterizations can also be applied around the inverse and unstable vertical position. For the homogenous heavy chain, we have only to replace g by $-g$ to obtain a family of smooth solutions to the elliptic equation (singular at $x = 0$)

$$\frac{\partial}{\partial x}\left(gx\frac{\partial X}{\partial x}\right) + \frac{\partial^2 X}{\partial t^2} = 0$$

by the integral

$$X(x,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\imath\sqrt{x/g}\sin\theta)\ d\theta,$$

where y is now a holomorphic function in $\mathbb{R} \times [-2\sqrt{L/g}, +2\sqrt{L/g}]$ that is real on the real axis. This parameterization can still be used to solve the motion planning problem in spite of the fact that the Cauchy problem associated to this elliptic equation is not well-posed in the sense of Hadamard.

Acknowledgments. The authors are indebted to Michel Fliess and Philippe Martin for fruitful discussions relative to the Paley–Wiener theorem, series expansions, majoring series arguments, and Liouville transformations.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] F. Boustany, *Commande Nonlinéaire Adaptative de Systèmes Mécaniques de Type Pont Roulant, Stabilisation Frontière d'EDP*, Ph.D. thesis, École des Mines de Paris, Paris, France, 1992.

[3] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.

[4] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 922–937.

[5] M. Fliess, P. Martin, N. Petit, and P. Rouchon, *Active signal restoration for the telegraph equation*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Computer Society, Los Alamitos, CA, 1999, pp. 1007–1011.

[6] M. Fliess and H. Mounier, *Controllability and observability of linear delay systems: An algebraic approach*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 301–314.

[7] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.

[8] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Masson, Paris, 1988.

[9] J.-L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[10] P. Martin and P. Rouchon, *Flatness and sampling control of induction motors*, in Proceedings of the IFAC World Congress, San Francisco, CA, 1996, pp. 389–394.

[11] H. Mounier, *Propriétés Structurelles des Systèmes Linéaires à Retards: Aspects Théoriques et Pratiques*, Ph.D. thesis, Université Paris Sud, Orsay, France, 1995.

[12] H. Mounier, J. Rudolph, M. Fliess, and P. Rouchon, *Tracking control of a vibrating string with an interior mass viewed as delay system*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 315–321.

[13] R. M. Murray, *Trajectory generation for a towed cable flight control system*, in Proceedings of the IFAC World Congress, San Francisco, CA, 1996, pp. 395–400.

[14] N. Petit, *Systèmes à Retards. Platitude en Génie des Procédés et Contrôle de Certaines Équations des Ondes*, Ph.D. thesis, École des Mines de Paris, Paris, France, 2000.

[15] N. Petit, Y. Creff, and P. Rouchon, *Motion planning for two classes of nonlinear systems with delays depending on the control*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Computer Society, Los Alamitos, CA, 1998, pp. 1107–1111.

[16] N. Petit and P. Rouchon, *Dynamics and Solutions to Some Control Problems for Water-Tank Systems*, CDS Technical Memo CIT-CDS 00-004, California Institute of Technology, Pasadena, CA, 2000.

[17] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, St. Louis, Paris, 1974.

[18] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1958.

[19] K. Yosida, *Lectures on Differential and Integral Equations*, Interscience, New York, 1960.

PERGAMON

Chemical Engineering Science 56 (2001) 2767–2774

Chemical Engineering Science

www.elsevier.nl/locate/ces

Minimum time constrained control of acid strength on a sulfuric acid alkylation unit

N. Petit[a], Y. Creff[b,*], L. Lemaire[c], P. Rouchon[d]

[a]Centre Automatique et Systèmes, École Nationale Supérieure des Mines de Paris, 60, boulevard Saint-Michel, 75272 Paris Cedex 06, France
[b]Elf Antar France, Centre de Recherches Elf à Solaize, BP 22, 69360 Solaize Cedex, France
[c]Elf Antar France, Raffinerie de Feyzin, BP 6, 69551 Feyzin Cedex, France
[d]Centre Automatique et Systèmes, École Nationale Supérieure des Mines de Paris, 60, boulevard Saint-Michel, 75272 Paris Cedex 06, France

Received 14 May 2000; received in revised form 28 July 2000; accepted 20 September 2000

Abstract

We detail here the controller of the acid strength that we implemented in the Elf-Antar France refinery in Feyzin (France). The control technique used is new. It relies on the *flatness* property of the system to solve a constrained minimum time objective as successive linear-programming problems. This controller is in full service since January 1997. We detail the control technique, including the estimation and numerical problems and then give industrial results over 6 months. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Alkylation unit; Process control; Optimization

1. Introduction

The alkylation of butenes is a common operation in oil refineries. It allows the synthesis of an interesting product, suitable to enter the composition of gasolines: the alkylate has a good octane number. Many kinds of units exist, but we concentrate here in the unit operated at the Elf Antar France's Feyzin refinery, which uses sulfuric acid as a catalyst.

The acid catalyst feeds two reactors in series. This feed is continuous. Partially destroyed during alkylation, the catalyst is withdrawn from the second reactor to feed a storage tank for off-site regeneration. A minimum amount of catalyst must be provided for the reactors to operate correctly. Providing more catalyst than the required minimum decreases the risks. But this implies expensive over-consumptions. The operator then tries to stabilize the unit just above the minimum. But the deterioration of the catalyst is very slow, and this makes such a manual driving difficult.

In 1996, the refinery decided to install a controller in order to limit acid consumption. The unit being very slow, we have decided to implement a minimum time control algorithm, applying results of a current collaboration between Elf and the "Centre Automatique et Systèmes" of the École des Mines de Paris. This controller is being used since January 1997, with a service factor higher than 98%. Under similar unit environments, it brings about 5% savings, which corresponds to a return time on investment of approximately 6 months.

2. Process description

The alkylation is made of two principal flow paths displayed in Fig. 1: a flow path for hydrocarbons and another for acid. The unit organizes the reaction of butenes and iso-butanes to form iso-octanes. Flows, either mainly containing butenes (olefins) or isobutane (recycle), are mixed before feeding two reactors in parallel, where the reaction takes place, catalyzed by sulfuric acid. The product of the reaction is flashed. The gas phase is condensed to generate a cold recycle. Mixed with the olefins and recycle, it helps to compensate for the exothermicity of the reaction. The liquid phase is washed before feeding a deisobutanizer. Propane is inert and accumulates in the unit. It is withdrawn after the flash. At its top, the deisobutanizer concentrates the isobutane, either coming from the so-called saturated feed or remaining in the liquid phase of the flash. The bottom

[*] Corresponding author.

0009-2509/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0009-2509(00)00472-3

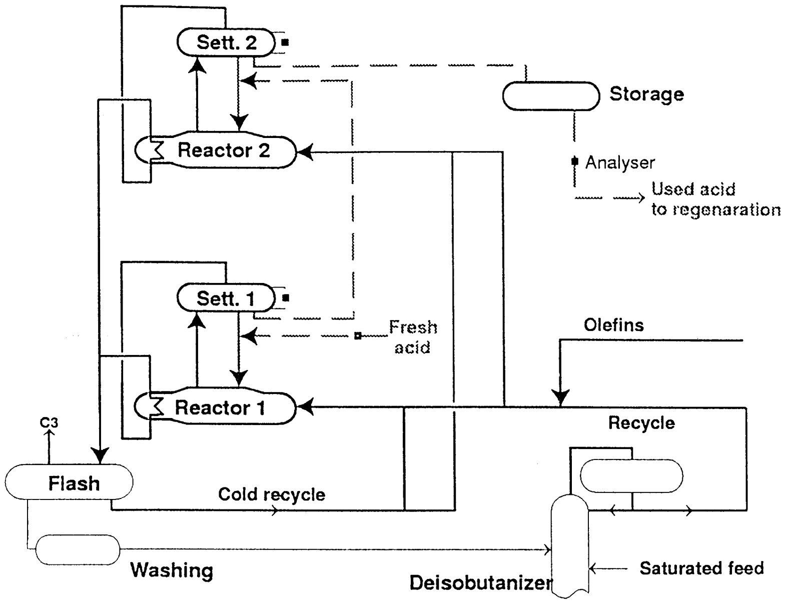


Fig. 1. Alkylation process.

product of this column essentially contains the alkylate and normal-butane (another inert), separated downstream.

The acid path flow is organized in series for the two reactors. A fresh acid flow (that is to say with a large acid concentration) feeds the first reactor and circulates between it and an associated settler. Secondary effects as reactions with impurities coming in small amounts with a flow of hydrocarbons induce a deconcentration of the acid (referred to as acid consumption). A flow, at a rate equivalent to the fresh acid flowrate, is withdrawn from the first settler and feeds the second reactor where a similar circulation is implemented with an associated settler. The second reactor also induces acid consumption. The used acid is withdrawn from the second settler and enters full storage tank. The acid from this tank is regenerated off-site. The concentration is analyzed at its output: this provides information about the nature of the forthcoming regeneration.

The fresh acid flowrate must be tuned to compensate for the variations of the acid consumption. Under a minimum concentration threshold, undesired reactions become important and induce serious malfunctions that must be avoided. Due to the way the unit is built, the concentration is the lowest at the output of the second settler. If its value is kept correct, good operating conditions are guaranteed for the two reactors. But allowing large security margins implies a large fresh acid flowrate that increases operating costs. It is better to work near the required minimum.

The slow variations of the acid concentration characterize this unit. A modification of the fresh acid flowrate is fully transmitted after about 1 week. Such a modification furthermore implies different residence times in the storage tank: they roughly vary from 8 to 24 h.

3. The control problem

The acid flow path is first modeled. A linear model sufficient for control purposes is derived from this physical model. It is better to control the output concentration of the second settler rather than the measured output concentration of the storage tank. But this implies the construction of an estimator, because of the location of the analyzer. We also use the physical model for this purpose.

3.1. Modeling

The acid flow path is viewed as two blocks in series followed by a full storage tank considered as an ideal plug flow reactor. Each block consists in a reactor and a settler. It is considered as a perfectly mixed reactor, which is a sounded assumption. The acid flow path is then made of the concatenation of the two block models and the model for the storage tank.

For each block, we consider the partial mass variation of sulfuric acid:

$$\rho V \frac{dx}{dt} = -(u + A(p))x + ux_{in},$$

where ρ is the acid density (mass/volume), assumed constant, V is the volume of acid phase in the block, assumed constant, x is the acid mass fraction in the block. As the block is perfectly mixed, it is the output concentration, x_{in} is the acid mass fraction of the input flow, u is the input and output flowrate, and A is a consumption term. It depends on a set of 12 disturbances p and takes into account all the effects implying a measurable acid deconcentration.

The storage tank is considered as a plug flow reactor. Dynamically, this tank introduces in the system a delay that is equal to the ratio between the mass of the acid contained in the tank and the flowrate. Finally, the model is

$$\rho V_1 \frac{\mathrm{d}x_1}{\mathrm{d}t} = -(u + A_1(p))x_1 + ux_f, \tag{1}$$

$$\rho V_2 \frac{\mathrm{d}x_2}{\mathrm{d}t} = -(u + A_2(p))x_2 + ux_1, \tag{2}$$

$$y = x_2\left(t - \frac{K}{u}\right), \tag{3}$$

where, x_1 and x_2, respectively, denote the sulfuric acid mass fractions at the output of the first and second blocks, u is the control, that is to say the acid flowrate feeding the first reactor, x_f is the sulfuric acid mass fraction of the fresh acid flow, A_1 and A_2, respectively, represent the acid consumptions in the first and second blocks, ρ is the acid density, V_1 and V_2, respectively, represent the acid phase volumes in the first and second blocks, y is the measured output. The dependency of delay on the control is denoted by K/u.

3.2. Control design

3.2.1. Control model

To act as efficiently as possible, we control an estimation of x_2. As y corresponds to the delayed value of x_2, if x_2 is correctly controlled, so is y. We shall see in the sequel how the physical model is used to build an estimation of x_2. Because of the very slow dynamics, when the situation is analyzed on a time range of a few hours, it is possible to ignore the drift of the system and summarize information for the control in the linear approximation

$$\frac{\mathrm{d}}{\mathrm{d}t}(x - x^{\text{mean}}) = a(u - u^{\text{mean}}) + \pi - \pi^{\text{mean}},$$

we use as a model and where

- x is the acid mass fraction at the output of the second block and x^{mean} its average value (we simplify notations: x corresponds to x_2 in the previous sections),
- u is the control and u^{mean} its average value,
- π denotes the contribution of measured disturbances and π^{mean} its average value,
- a denotes the gain by time unit.

Or, denoting, $P = -au^{\text{mean}} + \pi - \pi^{\text{mean}}$ and considering that x^{mean} is constant,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = au + P.$$

The gain a and the value of P are computed from a tangent approximation of the physical model (initial slope of a step response).

3.2.2. Minimum time constrained control

3.2.2.1. Flatness property and control algorithm. The idea relies upon the possibility to explicitly parameterize via x all the trajectories of the system. According to Fliess, Levine, Martin, and Rouchon (1995) and Martin, Murray, and Rouchon (1997), the system is *flat* and x is its *flat* output. Assuming x is known, u is derived immediately. Constraints on x, on the control and its variations are all linearly expressed with respect to x. Discretizing the model, we are led to the question of existence of solutions for a linear-programming problem. In case of multiple solutions, we choose the one allowing x to reach its setpoint in a minimum time.

Let us denote by x^i the values of x at the $n-1$ future sampling times and express the constraints that must be fulfilled over this horizon, exponent 1 denoting the current value. The n constraints on x are $x^{\min} \leqslant x^i \leqslant x^{\max}$. The $n-1$ constraints on u are $a\Delta u^{\min} + \Delta P^i \leqslant x^{i+1} - x^i \leqslant a\Delta u^{\max} + \Delta P^i$, where Δ is the sampling period and P^i the contribution of disturbances at time i (a constant equal to P^1 if no information is available about future disturbances). The sign of a impacts these inequalities. Here it is strictly positive. Constraints on the variations of u lead to similar expressions. Reaching the setpoint as an equilibrium point is achieved owing to the constraints $x^n = x^{n-1} = x^{\text{setpoint}}$. Finally, the current value x^{mes} of x (or its estimation in our case) is taken into account by $x^1 = x^{\text{mes}}$. All the constraints are summarized by $AX \leqslant B$, where X is the vector of the x^i. Every X obeying this inequality allows the construction of an admissible control profile:

$$u^i = \frac{x^{i+1} - x^i - \Delta P^i}{a\Delta} \qquad \forall i = 1,\ n-1.$$

As many solutions might exist for X, we must find a way to get a unique solution. We use a dichotomy on n to find the vector X with the lowest dimension that satisfies all the constraints. This is a minimum time control. Other approaches are possible. Only u^1 is applied and all the operations are computed at each sampling times, to partially compensate for nonmeasurable disturbances and modeling errors. At each sampling time, we also compute a prediction for the next time:

$$x^{\text{pred}} = x^{\text{mes}} + \Delta a(u - u^{\text{mean}}) + \Delta(\pi - \pi^{\text{mean}}).$$

The filtered difference between the prediction and the "measure" is added to P. This is a standard compensation method.

3.2.2.2. Generalization. The idea of this method was originally described in Petit (1996). Note that it might be extended to (controllable) multivariable linear systems. This brings an alternative formulation of the classical linear predictive control algorithms (Richalet (1993)). This method lies in a natural framework for efficient nonlinear extension of these algorithms, namely the *flatness* framework (Fliess et al., 1995; Martin et al., 1997). When a system is *flat*, it is possible to directly work on its parameterized trajectories and doing so to avoid solving ordinary differential equations, that penalizes nonlinear predictive control and other approaches in dynamic optimization.

3.2.2.3. Numerical solving. To solve the successive linear programming problems we used a standard commercial simplex-based algorithm. Though these algorithm are known to have a nonpolynomial complexity (Klee & Minty, 1972), it has been noted by many authors and specialists that in practice they behave very well and are very robust numerically speaking (Nering & Tucker, 1993). This robustness combined to the breadth of the commercial packages and the relative low dimension of the problem led us not to consider interior-point algorithms.

3.3. Estimation

The control law described above assumes that x_2 is known. But only y is measured: we have to construct an estimation of x_2 based on delayed measurements, furthermore with variable delays. We have tested many approaches before finding the following satisfactory answer:

$$\rho V_1 \frac{\mathrm{d}z_1}{\mathrm{d}t} = -(u + A_1(p))z_1 + ux_f, \tag{4}$$

$$\rho V_2 \frac{\mathrm{d}z_2}{\mathrm{d}t} = -(u + A_2(p))z_2 + uz_1, \tag{5}$$

$$y' = z_2 + \varphi, \tag{6}$$

$$\dot{\varphi} = \text{sat}\left(z_2\left(t - \frac{K}{u}\right), y\right) - \frac{\varphi}{\tau_f}, \tag{7}$$

where the estimated state is z, φ denotes a filter of the difference between the delayed observation and the measured value of y. The first two equations of this system are a copy of those of the original model. For confidential reasons it is not possible to describe the sat function which is roughly speaking a linear saturated function.

Strictly speaking, it is possible to prove that if the system is not perturbed the z_1 and z_2 converge to x_1 and x_2. Though the system is time-varying, its triangular form allows to prove that z_1 tends exponentially to x_1 since $(u + A_1(p))$ is lower-bounded by a positive constant. Then one can prove that z_2 converges exponentially to x_2 since $(u + A_2(p))$ is lower-bounded by a positive constant and $(z_1 - x_1)$ is an exponentially decreasing function (see, for instance, Khalil, 1992).

It should have been possible to use the classical high-gain observer approach (see, Gauthier, Hammouri, & Othman, 1992). But here the perturbations prevented us from implementing it successfully.

3.4. Robustness

Strictly speaking, we cannot prove the overall mathematical robustness of our approach.

On the one hand, the above observer gives good results, despite the perturbations, and is robust. Practically, this method insures that y' converges to x_2.

On the other, the control algorithm we use is numerically robust.

To show the experimental robustness of our approach are shown in Fig. 2 industrial results over 6 months with our controller.

3.5. Implementation

Implementing this control law took about 6 months. It constitutes a fast transfer between academic work and application in industry. The Feyzin refinery and the Elf research center have first developed and validated the model (we thank MM. Dajczman -Feyzin- and Djenab -CRES- for their fruitful participation). Together, we have then adapted to this problem the first results of N. Petit's Ph.D. thesis, detailed in Petit (1993). Finding a good estimator has revealed more time-consuming, because of the need for robustness in face of inaccuracies on variable delays.

The algorithm runs on a HP1000 computer. Its execution period is 15 min.

4. Conclusions

The controller was rapidly accepted by the operators. Since its implementation, it has been used almost full-time (service factor higher than 98%). The operator have first observed the way it was working with a setpoint above the final objective. They were convinced by its ability to safely react in order to stabilize the unit (modifying the fresh acid flowrate before variations on the measured concentration was surprising at the beginning). After a short period, they accepted to decrease the setpoint, then decreasing the required fresh acid flow

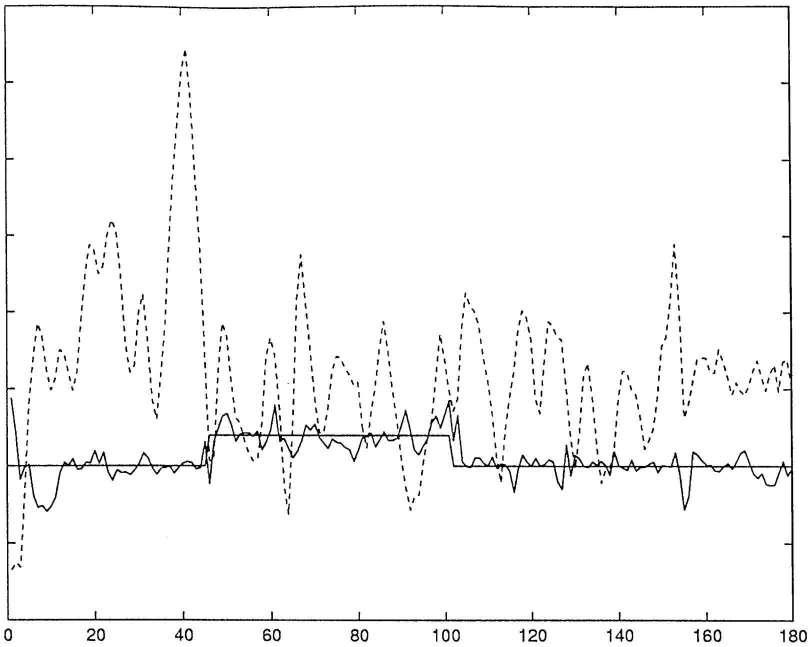


Fig. 2. Daily averages over 6 months: setpoint, result with our controller, and without it in dashed line (from history).

rate. Results of our controller over a period of 6 months are shown in Fig. 2. One can compare these results with results over a similar period of 6 months without our controller.

The benefits of this implementation are as follows. Stabilizing the unit allows the operators to concentrate on more difficult tasks. Furthermore, limiting acid consumption brings about 5% savings on the costs associated with the use of sulfuric acid.

Appendix

In the following, we show how to turn a solution to the discrete time optimization problem into a solution to a continuous time problem. This regularization is achieved owing to a convolution with a C^∞ kernel and a time scaling (details about the classical technique of regularization can be found in Schwartz (1973, pp. 165–167)). This demonstrates Proposition 1.

Next, we show (Propositions 4 and 5) that both continuous and discrete time problem have a unique minimum time solution.

In the end, we conclude (Theorem A.1) that when the time step decreases to zero, the solution of the discrete time problem tends towards the solution of the continuous time problem.

Notation A.1. Given a set of real numbers $Y^{(0)}_{\min}$, $Y^{(0)}_{\max}$, $Y^{(1)}_{\max}$, $Y^{(2)}_{\max}$, where we assume $Y^{(1)}_{\max} \geqslant 0$, $Y^{(2)}_{\max} \geqslant 0$, let $C(T)$ be the subset of functions $Y \in C^2([O,T])$ satisfying the following conditions:

$$C(T): \begin{cases} \forall t \in]0,T[:, \\ Y^{(0)}_{\min} \leqslant Y(t) \leqslant Y^{(0)}_{\max}, \\ |\dot{Y}(t)| \leqslant Y^{(1)}_{\max}, \\ |\ddot{Y}(t)| \leqslant Y^{(2)}_{\max}, \\ Y(0)=0, \quad \dot{Y}(0)=0, \quad \ddot{Y}(0)=0, \\ Y(T)=1, \quad \dot{Y}(T)=0, \quad \ddot{Y}(T)=0. \end{cases}$$

Besides, let $D(N,\delta t)$ be the set of samples $Y_N = [Y_N(0), Y_N(1), \ldots, Y_N(N)]$ satisfying the following conditions:

$$D(N,\delta t): \begin{cases} \forall i \in \mathbb{N} \text{ (where if necessary } Y_N(i<0) = Y_N(0)), \\ Y_N(i>N) = Y_N(N): \\ Y^{(0)}_{\min} \leqslant Y_N(i) \leqslant Y^{(0)}_{\max}, \\ \left|\dfrac{Y_N(i+1) - Y_N(i)}{\delta t}\right| \leqslant Y^{(1)}_{\max}, \\ \left|\dfrac{Y_N(i+2) - 2Y_N(i+1) + Y_N(i)}{(\delta t)^2}\right| \leqslant Y^{(2)}_{\max}, \\ Y_N(0)=0, \; Y_N(N)=1. \end{cases}$$

In the following, $[x]$ *denotes the largest integer less or equal to x.*

Proposition A.1. $\forall (N,\delta t)$, $D(N,\delta t) \neq \emptyset$ *implies* $C((N+3/2)\delta t(1+\delta t)) \neq \emptyset$.

Proof. Let Y^{aff} be the function continuing the affine interpolation of Y_N to the left and to the right of $[\delta t, (N+1)\delta t]$:

$$Y^{\mathrm{aff}} = \begin{cases} Y_N(0) & \text{if } t < \delta t, \\ Y_N(i-1) + \dfrac{Y_N(i) - Y_N(i-1)}{\delta t}(t - i\delta t) & \text{with } i = [t/\delta t] \text{ if } \delta t \leqslant t < (N+1)\delta t, \\ Y_N(N) & \text{if } t \geqslant (N+1)\delta t. \end{cases}$$

Let χ_ε be an approximation to the unit, i.e., a positive function, the support of which is $[-\varepsilon/2, \varepsilon/2]$ and such that $\int_{-\varepsilon/2}^{\varepsilon/2} \chi_\varepsilon(s)\,\mathrm{d}s = 1$.

We regularize Y^{aff} into Y^r by the following convolution:

$$Y^r = Y^{\mathrm{aff}} * \chi_\varepsilon.$$

The support of Y^r is $[\delta t - \varepsilon/2, (N+1)\delta t + \varepsilon/2]$. □

Let us assume that $0 < \varepsilon/2 \leqslant \delta t$.

Lemma A.1.

$$Y^r(0) = 0, \quad Y^r\left((N+1)\delta t + \frac{\varepsilon}{2}\right) = 1, \quad \textit{and}$$
$$\forall t,\ |Y^r(t)| \leqslant Y^{(0)}_{\max}.$$

Proof. $|Y_{\mathrm{aff}}| \leqslant Y^{\max}$ and $Y^r = Y^{\mathrm{aff}} * \chi_\varepsilon$, then $|Y^r| \leqslant Y^{(0)}_{\max}$. At last, $Y^r(0) = 0$ since $Y^{\mathrm{aff}}(t < \delta t) = 0$. Likely, $Y^r((N+1)\delta t + \varepsilon/2) = 1$ since $Y^{\mathrm{aff}}(t > (N+1)\delta t) = 1$. □

Lemma A.2.

$$\dot{Y}^r(0) = 0, \quad \dot{Y}^r\left((N+1)\delta t + \frac{\varepsilon}{2}\right) = 0, \quad \textit{and}$$
$$\forall t, |\dot{Y}^r(t)| \leqslant Y^{(1)}_{\max}.$$

Proof.

$$\begin{aligned}\dot{Y}^r(t) &= \dot{Y}^{\mathrm{aff}} * \chi_\varepsilon(t) \\ &= \frac{Y_N(i-1) - Y_N(i-2)}{\delta t}\int_{-\varepsilon/2}^{\eta} \chi_\varepsilon(s)\,\mathrm{d}s \\ &\quad + \frac{Y_N(i) - Y_N(i-1)}{\delta t}\int_{\eta}^{\varepsilon/2} \chi_\varepsilon(s)\,\mathrm{d}s,\end{aligned}$$

where $\eta \in [-\varepsilon/2, \varepsilon/2]$, $i = [t/\delta t]$, and $Y_N(i < 0) = Y_N(0)$ $Y_N(i > N) = Y_N(N)$ if necessary.

$\dot{Y}^r(t)$ can be seen as the barycentre of $(Y_N(i-1) - Y_N(i-2))/\delta t$ and $(Y_N(i) - Y_N(i-1))/\delta t$. Thus, $|\dot{Y}^r| \leqslant Y^{(1)}_{\max}$. The last formula directly implies that $\dot{Y}^r(0) = 0$ and $\dot{Y}^r((N+1)\delta t + \varepsilon/2) = 0$. □

Lemma A.3. *One may choose* χ_ε *such as*

$$\dot{Y}^r(0) = 0, \quad \dot{Y}^r\left((N+1)\delta t + \frac{\varepsilon}{2}\right) = 0,$$
$$\textit{and} \quad \forall t |\ddot{Y}^r| \leqslant (1 + \delta t) Y^{(2)}_{\max}.$$

Proof.

$$\begin{aligned}\ddot{Y}^r(t) &= \dot{Y}^{\mathrm{aff}} * \dot{\chi}_\varepsilon(t) \\ &= \frac{Y_N(i-1) - Y_N(i-2)}{\delta t}\int_{-\varepsilon/2}^{\eta} \dot{\chi}_\varepsilon(s)\,\mathrm{d}s \\ &\quad + \frac{Y_N(i) - Y_N(i-1)}{\delta t}\int_{\eta}^{\varepsilon/2} \dot{\chi}_\varepsilon(s)\,\mathrm{d}s \\ &= -\chi_\varepsilon(\eta)\frac{Y_N(i) - 2Y_N(i-1) + Y_N(i-2)}{\delta t},\end{aligned}$$

where $\eta \in [-\varepsilon/2, \varepsilon/2], i = [t/\delta t]$, $Y_N(i < 0) = Y_N(0)$ and $Y_N(i > N) = Y_N(N)$ if necessary. This yields

$$|\ddot{Y}^r| \leqslant |\chi_\varepsilon(r)| Y^{(2)}_{\max} \delta t$$

and

$$\ddot{Y}^r(0) = 0,$$

$$\ddot{Y}^r\left((N+1)\delta t + \frac{\varepsilon}{2}\right) = 0.$$

Let us choose χ_ε such as $\chi_\varepsilon \leqslant (1+\varepsilon)/\varepsilon$, which is compatible with $\int_{\varepsilon/2}^{\varepsilon/2} \chi_\varepsilon(s)\,\mathrm{d}s = 1$. Then choose $\varepsilon = \delta t$. This gives

$$|\ddot{Y}^r| \leqslant (1 + \delta t) Y^{(2)}_{\max}. \qquad \square$$

In the end, let us use a time scaling to define Y^d: $Y^d(t) = Y^r(t/(1+\delta t))$.

Lemma A.4. *According to the previous notations,* $Y^d \in C((N + \frac{3}{2})\delta t(1 + \delta t))$.

Proof. As shown by Lemmas A.1–A.3:

$$Y_{\min} \leqslant Y^d \leqslant Y_{\max},$$

$$|\dot{Y}^d| \leqslant Y^1_{\max},$$

$$|\ddot{Y}^d| \leqslant Y^{(2)}_{\max}$$

and

$$Y^d(0) = 0,$$

$$\dot{Y}^d(0) = 0, \quad \ddot{Y}^d(0) = 0,$$

$$Y^d((N+3/2)\delta t(1+\delta t)) = 1, \quad \dot{Y}^d((N+3/2)\delta t(1+\delta t)) = 0,$$

$$\ddot{Y}^d((N+3/2)\delta t(1+\delta t)) = 0.$$

Finally, the support of Y^r is $[\delta t/2, (N+\frac{3}{2})\delta t]$. This means that the support of Y_d is included into $[0,(N+3/2)\delta t(1+\delta t)]$. □

Lemma A.4 gives the conclusion of Proposition A.1.

Proposition A.2. $\forall T$, $C(T) \neq \emptyset$ *implies* $\forall \Delta T \leqslant 0$, $C(T+\Delta) \neq \emptyset$.

Proof. Assume $C(T) \neq \emptyset$, then there exists $Y \in C(T)$. For all $\Delta > 0$, let us continue Y into $\hat{Y}$:

$$\hat{Y} = \begin{cases} Y(t) & \text{if } t \leqslant T, \\ Y(T) & \text{if } T < t \leqslant T + \Delta. \end{cases}$$

Obviously, $\hat{Y} \in C(T+\Delta)$ which is not empty. □

Proposition A.3. $C(T) \neq \emptyset \Rightarrow \forall N \in \mathbb{N}^*$, $D(N, T/N) \neq \emptyset$.

Proof. Assume $C(T) \neq \emptyset$, then there exists $Y \in C(T)$. Let $\delta t = T/N$. Consider $Y_N = [Y(0), Y(\delta t), Y(2\delta t), \ldots, Y(T)]$. In the following $Y_N(j)$ denotes the $(j+1)$th coordinate of Y_N, the value of which is $Y(j\delta t)$.

(i) Obviously

$$Y_{\min} \leqslant Y_N(j) = Y(j\delta t) \leqslant Y_{\max}. \tag{A.1}$$

(ii) Let us consider the differences $Y_N(j+1) - Y_N(j) = 0 + \delta t \dot{Y}(j\delta t + \theta_j \,\delta t)$ where $\theta_j \in]0,1[$ from MacLaurin's formula,

Yet

$$|\dot{Y}(j\delta t + \theta_j \,\delta t)| \leqslant Y^{(1)}_{\max}$$

$$\text{so } \forall j, \quad \left|\frac{Y_N(j+1) - Y_N(j)}{\delta t}\right| \leqslant Y^{(1)}_{\max}, \tag{A.2}$$

(iii) at last, let us consider the differences

$$\begin{aligned} &Y_N(j+2) - 2Y_N(j+1) + Y_N(j) \\ &\quad = Y((j+2)\delta t) - 2Y((j+1)\delta t) + Y(j\delta t) \\ &\quad = \delta t \dot{Y}((j+1)\delta t) + \tfrac{1}{2}(\delta t)^2 \ddot{Y}((j+1)\delta t + \delta t\ \theta^+) \\ &\qquad \text{where } \theta^+ \in]0,1[\\ &\quad - \delta t \dot{Y}((j+1)\delta t) + \tfrac{1}{2}(\delta t)^2 \ddot{Y}((j+1)\delta t + \delta t\ \theta^-) \\ &\qquad \text{where } \theta^- \in]-1,0[\\ &\quad = \tfrac{1}{2}(\delta t)^2(\ddot{Y}((j+1)\delta t + \delta t\ \theta^+) + \ddot{Y}((j+1)\delta t + \delta t\ \theta^-)) \end{aligned}$$

which ends up in

$$\frac{Y_N(j+2) - 2Y_N(j+1) + Y_N(j)}{(\delta t)^2}$$

$$= \tfrac{1}{2}(\ddot{Y}((j+1)\delta t + \delta t\ \theta^+ + \ddot{Y}((j+1)\delta t + \delta t\ \theta^-)).$$

Yet, as we already know

$$\forall t, \quad |\ddot{Y}(t)| \leqslant Y^{(2)}_{\max}$$

so

$$\forall j, \quad \left|\frac{Y_N(j+2) - 2Y_N(j+1) + Y_N(j)}{(\delta t)^2}\right| \leqslant Y^{(2)}_{\max}. \tag{A.3}$$

In the end, Eqs. (A.1)–(A.3) ensure that $Y_N \in D(N, T/N)$ which is not empty. □

Proposition A.4. *There exists a unique minimum time, which we denote T_{mini}, such as $C(T) \neq \emptyset$.*

Proof. This is a direct conclusion from Proposition A.2. □

Proposition A.5. *For any given δt, there exists a unique minimum integer, which we denote $N_{\text{mini}}(\delta t)$ such as $D(N, \delta t) \neq \emptyset$.*

Proof. The proof is similar to the one of Proposition A.2: let $Y \in D(N, \delta t)$, then it is clear that $[Y(0)Y(1)\ldots Y(N)Y(N)]$ is an element of $D(N+1, \delta t)$. □

Theorem A.1. *The required time for the solution to the discrete time problem, $N_{\text{mini}}(\delta t)\delta t$, tends towards T_{mini} as δt tends towards zero. In other words, the discrete time problem tends to the continuous time problem as δt tends towards zero.*

Proof. From Proposition A.4, we know that there exists a unique minimum time T_{mini} such that $C(T_{\text{mini}}) \neq \emptyset$. Given δt, one may write

$$\left[\frac{T_{\text{mini}}}{\delta t}\right]\delta t < T_{\text{mini}} \leqslant \left(\left[\frac{T_{\text{mini}}}{\delta t}\right] + 1\right)\delta t.$$

This last equation means that

$$C\left(\left[\frac{T_{\text{mini}}}{\delta t}\right]\delta t\right) = \emptyset \tag{A.4}$$

$$C\left(\left(\left[\frac{T_{\text{mini}}}{\delta t}\right] + 1\right)\delta t\right) \neq \emptyset. \tag{A.5}$$

Then, we deduce from Proposition A.3 that $D([T_{\text{mini}}/\delta t] + 1, \delta t) \neq \emptyset$. Let $\delta t' = (-1 + \sqrt{1 + 4\delta t})/2$, i.e., $\delta t'(1 + \delta t') = \delta t$. We must have $D([T_{\text{mini}}/\delta t] - 2, \delta t') = \emptyset$ otherwise, Proposition A.1 insures that $C(([T_{\text{mini}}/\delta t] - \frac{1}{2})\ \delta t'(1 + \delta t')) \neq \emptyset$ which would mean that $C([T_{\text{mini}}/\delta t]\,\delta t) \neq \emptyset$ which is not true as we know from Eq. (A.4). So

$$D\left(\left[\frac{T_{\text{mini}}}{\delta t}\right] - 2, \frac{-1 + \sqrt{1 + 4\delta t}}{2}\right) = \emptyset,$$

$$D\left(\left[\frac{T_{\text{mini}}}{\delta t}\right] + 1, \delta t\right) \neq \emptyset.$$

Besides, as we know from Proposition A.5, for all δt there exists a unique minimum integer $N_{\text{mini}}(\delta t)$ such that $D(N_{\text{mini}}(\delta t), \delta t) \neq \emptyset$.

The last two relations imply that:

$$N_{\text{mini}}\left(\frac{-1 + \sqrt{1 + 4\delta t^2}}{2}\right) \leqslant \left[\frac{T_{\text{mini}}}{\delta t}\right] - 2,$$

$$N_{\text{mini}}(\delta t) \leqslant \left[\frac{T_{\text{mini}}}{\delta t}\right] + 1.$$

We deduce that

$$\lim_{\delta t \to 0} N_{\text{mini}}(\delta t)\delta t \leqslant T_{\text{mini}},$$

$$\lim_{\delta t \to 0} N_{\text{mini}}\left(\frac{-1 + \sqrt{1 + 4\delta t}}{2}\right)\delta t \leqslant T_{\text{mini}}$$

which gives

$$\lim_{\delta t \to 0} N_{\text{mini}}(\delta t)\,\delta t = T_{\text{mini}}.$$

Solving the discrete time problem gives a solution to the continuous time problem owing to a regularization and a time scaling. The support of the obtained solution is $[0, (N_{\text{mini}}(\delta t) + \frac{3}{2}\delta t(1 + \delta t)]$ and it tends to $[0, T_{\text{mini}}]$. □

References

Fliess, M., Lévine, J., Martin, Ph., & Rouchon, P. (1995). Flatness and defect of nonlinear systems: introductory theory and examples. *International Journal of Control*, *61*(6), 1327–1361.

Gauthier, J. -P., Hammouri, H., & Othman, S. A. (1992). Simple observer for nonlinear systems — applications to bioreactors. *IEEEtac*, *37*, 875–880.

Khalil, H. K. (1992). *Nonlinear systems*. New York: MacMillan.

Klee, V., & Minty, G. J. (1972). How good is the simplex algorithm? In O. Shisha (Ed.), *Inequalities*, *III* (pp. 159–175), New York: Academic Press.

Martin, Ph., Murray, R. M., & Rouchon, P. (1997). Flat systems. In *Proceedings of the fourth European control conference* (pp. 211–264). Plenary Lectures and Mini-courses.

Nering, E. D., & Tucker, A. W. (1993). *Linear programs and related problems*. New York: Academic Press.

Petit, N. (1996). *Systèmes δ-libres sous contraintes.* Rapport de stage de DEA, Université d'Orsay.

Richalet, J. (1993). *Pratique de la commande prédictive*. Hermès.

Schwartz, L. (1973). *Théorie des distributions*. Hermann.

Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Distributed delay model for density wave dynamics in gas lifted wells

Laure Sinègre*, Nicolas Petit
Centre Automatique et Systèmes
École des Mines de Paris
60, Bd Saint-Michel, 75272 PARIS Cedex
Email: sinegre,petit@cas.ensmp.fr
*corresponding author

Philippe Ménégatti
Centre Scientifique et Technique Jean Féger
TOTAL
Avenue Larribeau, 64000 PAU
Email: philippe.menegatti@total.com

Abstract— **Oil well instabilities cause production losses. One of these instabilities, referred to as the "density-wave" is an oscillatory phenomenon occurring on gas-lift artificially lifted well. We propose a distributed delay model of this dynamics. In order to interpret the observed oscillations we study the corresponding characteristic equation. Stabilization of this system is performed through a simple control law. Its performance is studied through realistic simulations.**

I. Introduction

Producing oil from deep reservoirs and lifting it through wells to surface facilities often requires activation to maintain oil output at a commercial level. In the gas-lift activation technique [3], gas is injected at the bottom of the well through the injection valve (point C in Figure 1) to lighten up the fluid column and to lower the gravity pressure losses. High pressure gas is injected at well head through the gas valve (point A in Figure 1), then goes down into the annular space between the drilling pipe (casing, point B) and the production pipe (tubing, point D) where it enters. Oil produced from the reservoir (point F) and injected gas mix in the tubing. They flow through the production valve E located at the surface.

As wells and reservoirs get older, liquid rates begin to decrease letting wells be more sensitive to flow instabilities commonly called headings. These induce important oil production losses (see [8]) along with possible facilities damages. Preventing instabilities through closed loop control has been an active field of research (see [10], [8] and [1]). These instabilities are defined as a flow regime characterized by regular and perhaps irregular cyclic changes in pressure at any point in the tubing string D (see [2]). Among these, one finds the "casing-heading" and the density-wave instability. "Casing heading" consists of a succession of pressure build-up phases in the casing without production and high flow rate phases due to intermittent gas injection rate from the casing to the tubing (see [10] for a complete description). The dynamics of the "casing heading" is well represented by a three balance ordinary differential equations model (proposed in [9], [6] and used in [12]). In the density-wave instability, which existence was first demonstrated in [8], oscillations are confined in the tubing D while the gas injection rate through valve C remains constant. Out-of-phase effects between the well influx and the total pressure drop along the tubing are usually reported at the birth of this phenomenon. In [8], dynamical choking is used to stabilise the density wave instability. In this paper, we propose a distributed delay model to represent and analyse the observed oscillations. Two applications of the model are presented: first a rigorous stability analysis demonstrating the impact of the gas flow rate and then an alternative control solution to [8] using the gas inlet A as an input and the downhole pressure measurements.

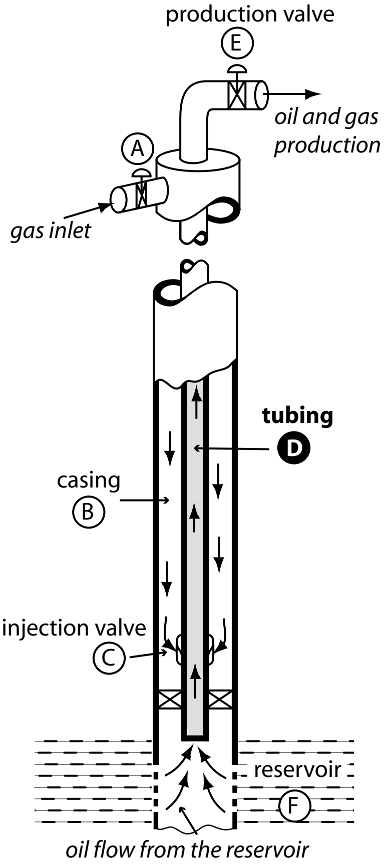


Fig. 1. Gas-lift activated well. Density-wave takes place in the tubing D.

The paper is organized as follows. In Section II, we detail the observed oscillating phenomenon in gas-lift operations. In Section III, we derive a reference distributed delay model for the density propagation in the tubing. Main assumptions

0-7803-9568-9/05/$20.00 ©2005 IEEE

and the use of Riemann invariant are explicited along with boundary conditions. In Section IV, stability analysis of the corresponding characteristic equations is performed. Comparisons with OLGA®2000 are conducted and stress the role of the amount of injected gas. In Section V, we propose a control strategy relying on the model. Realistic simulations show that we can stabilize the flow.

II. Gas-lift operations

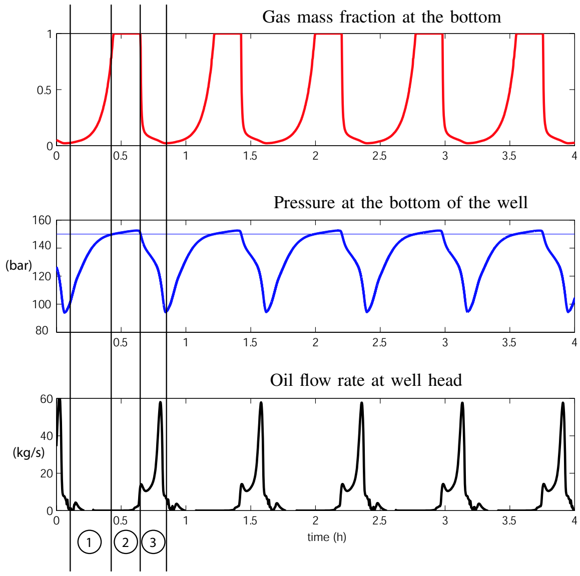


Fig. 2. Density wave simulated with OLGA®2000.

Figure 2 shows an example of density wave instability simulated with the transient multiphase flow simulator OLGA®2000. Typically, the depth of the well is 2500 m and the reservoir pressure is 150 bar. Oil production has an oscillating behavior consisting of 3 phases. In phase 1, there is no oil production at the surface but P_L, the pressure at the bottom of the well, is less than the reservoir pressure. Oil enters the pipe, letting P_L get closer to 150 bar. This is the self regulating mechanism of the well: the more is produced from the reservoir, the greater P_L becomes and eventually the less is produced. P_L is going to reach a constant which, in this case, is greater than 150 bar. The system switches to phase 2. This phase is characterized by zero oil production at the surface and from the reservoir (saturation of the oil flow rate at the bottom of the well). The gas mass fraction, which is close to 0 in phase 1, gets to a strictly positive constant in phase 2. Finally, the oil produced from the reservoir in phase 1 reaches the surface creating pressure drop in the well. This is phase 3. P_L decreases below 150 bar, oil flow rate at the bottom of the well increases and brings the fall of the gas mass fraction.

In summary, the density wave can be interpreted as the propagation of the mass fraction at the bottom of the well which is a result of a switching boundary condition.

Symb.	Constants	Values	Units
R	Gas perfect constant	287	S.I.
T	Temperature of the well	293	K
PI	Productivity Index	$4e-6$	kg/s/Pa
P_r	Reservoir pressure	$150e5$	Pa
P_0	Separator pressure	$10e5$	Pa
g	Gravity constant	9.81	m/s^2
ρ_l	Density of oil	800	kg/m^3
V^∞	Slip velocity constant	-	m/s
V_g	Gas velocity	0.8	m/s
β	Threshold parameter	0.03	
u_{min}	Saturation value of u	0.1	bar
u_{ref}	Reference value of u	10	bar
q_g^{min}	Saturation value of q_g	0.3	kg/s
L	Length of the pipe	2000	m

Symb.	Variables	Expressions	Units
$V_l(t,z)$	Oil velocity	$V_g + V^\infty/R_l$	m/s
$Rg(t,z)$	Gas volume fraction		
$R_l(t,z)$	Oil volume fraction	$R_g + R_l = 1$	
$x(t,z)$	Gas mass fraction		
$P(t,z)$	Pressure of the well		Pa
$x^L(t)$	Gas mass fraction at $z = L$		
$P_L(t)$	Pressure at $z = L$		Pa
$\rho_g(t,z)$	Gas density		kg/m^3
$\rho_m(t,z)$	Mixture density		kg/m^3
$q_l(t,z)$	Oil mass flux	$R_l\rho_l V_l$	kg/s/m^2
$q_g(t,z)$	Gas mass flux	$R_g\rho_g V_g$	kg/s/m^2
$u(t)$	Control $\simeq$ gas injection	$q_g/PI(1/\beta-1)$	bar

TABLE I

Nomenclature.

III. Proposed Model

We propose to study the density wave instability as a two phases flow problem in a vertical pipe filled with a mixture of oil and gas. The pressure at both ends are considered constant. Flows (gas and oil) enters the pipe at the bottom. The oil flow is given by the difference of pressure between the bottom of the pipe and the reservoir. The gas injection rate is considered constant (its value can be arbitrary updated for control purposes). Notations are given in Table I. Thanks to the choice of the slip velocity law (following [5]), we demonstrate the existence of a Riemann invariant. This lets the evolution of the distributed variables be summarized by the evolution of a single variable: the pressure at the bottom of the pipe.

A. Physics reduction

Pressure law: Using Bernoulli's law we get

$$P(t,z) = P_0 + \int_0^z \rho_m(t,\zeta) g d\zeta \qquad (1)$$

Model (1) implies no friction term, it is consistent with the observed low flow rates for density wave instability (see [8]). Density of the mixture is given by

$$1/\rho_m = x/\rho_g + (1-x)/\rho_l$$

To work with a linear expression of ρ_m we assume that

$$\rho_m \sim x\rho_g + (1-x)\rho_l \qquad (2)$$

Equivalently, we assume that the gas mass density is close to the gas volume density. Further, in the derivation of the gas density, gas is considered perfect and the temperature T is constant. Besides we assume that the pressure gradient $\frac{P_r-P_0}{L}$ along the tubing is also constant and computed from boundary condition. Simulations have shown that this simplification improves the tractability while saving the oscillatory behavior. Using the expressions in (2) and after substitution in (1), we get

$$P(t,z) = \\ P_0 + \rho_l g z + \int_0^z x(t,\zeta) g \left(\frac{(L-\zeta)P_0 + \zeta P_r}{LRT} - \rho_l \right) d\zeta \tag{3}$$

Slip velocity and Riemann invariant: We define the slip velocity (see [5]) as follows

$$V_g - V_l = \frac{V_\infty}{R_l}$$

Mass conservation laws write

$$\frac{\partial \rho_g R_g}{\partial t} + \frac{\partial q_g}{\partial z} = 0 \tag{4}$$

$$\frac{\partial \rho_l R_l}{\partial t} + \frac{\partial q_l}{\partial z} = 0 \tag{5}$$

As

$$x = \frac{R_g \rho_g}{R_g \rho_g + R_l \rho_l} \tag{6}$$

one can combine (4), (5) and (6), to obtain

$$\frac{\partial x}{\partial t} + V_g \frac{\partial x}{\partial z} = 0$$

meaning that x is a Riemann invariant (see [4]). For sake of simplicity we assume V_g to be constant. On real wells it is not as simple and we shall discuss the implications of this hypothesis in Section V-C. This implies

$$x(t,z) = x\left(t - \frac{L-z}{V_g}, L\right) = x^L\left(t - \frac{L-z}{V_g}\right)$$

Therefore, knowing bottom well gas mass fraction $t \mapsto x^L(t)$, we get the profile $(t,z) \mapsto x(t,z)$ in the tubing. Replacing this expression in equation (3) and denoting $P_L(t) = P(t,L)$, we find

$$P_L(t) = P_L^\star + \int_{t-\delta}^t k(t-\tau) x^L(\tau) d\tau \tag{7}$$

with

$$\delta = L/V_g \tag{8}$$

$$P_L^\star = P_0 + \rho_l g L \tag{9}$$

and

$$[0,\delta] \ni t \mapsto k(t) \triangleq V_g g \left(\frac{tP_0 + (\delta - t)P_r}{\delta RT} - \rho_l \right) < 0 \tag{10}$$

Notice that k is a strictly decreasing affine function.

Boundary condition: Classically, (see [3]), the oil rate q_l is given at the reservoir boundary by the Productivity Index (PI) through

$$q_l(t,L) = PI \max(P_r - P_L(t), 0) \tag{11}$$

By definition,

$$x^L(t) = \frac{1}{1 + PI/q_g \max(P_r - P_L(t), 0)} \tag{12}$$

We want to simplify this last expression in the case of large PI. On one hand, as $P_r - P_L$ begins to be positive, x^L goes to zero. Let β denote a threshold parameter. In particular $x^L < \beta$ is equivalent to $P_L < P_r - \frac{q_g}{IP}(1/\beta - 1)$. We denote

$$u \triangleq q_g \frac{1}{PI}(1/\beta - 1) \tag{13}$$

On the other hand, when $P_L > P_r$, $x^L = 1$. Therefore, we consider x^L as constant, equal to 1 when $P_L > P_r$ and equal to 0 when $P_L < P_r - u$. Finally, the considered expression of x^L reduces to

$$x^L = h(X), \ X \triangleq 1 - \frac{P_r - P_L}{u} \tag{14}$$

with

$$h(\cdot) = \max(\min(1,\cdot), 0)$$

Equation (14) is the definition we use instead of Equation (12) from now on.

B. Density-wave as a distributed delay model

We now gather equations (7) and (14), and consider an initial condition $[-\delta, 0] \ni t \mapsto \phi(t) \in \mathbb{R}$. The following model represents the density wave phenomenon by the evolution of the pressure at the bottom of the well P_L

$$\begin{cases} P_L(t) = P_L^\star + \displaystyle\int_{t-\delta}^t k(t-\tau) h\left(1 - \frac{P_r - P_L(\tau)}{u(\tau)}\right) d\tau \\ P_L(t) = \phi(t), \ t \in [-\delta, 0] \end{cases} \tag{15}$$

where δ is the transport delay defined in (8), $P_L^\star$, given in (9), is the pressure at the bottom of the pipe when it is full of oil and P_r is the pressure of the reservoir. k is an affine function, given in (10). It depends on the considered fluids. u is proportional to q_g (see equation (13)). It can be arbitrarily updated and thus can be considered as a control.

C. Simulation results

Figure 3 shows the simulations results of (15). When $P_r = 150$ bar and $u = 10$ bar, we get an oscillating trajectory which presents similarities with Figure 2. Indeed, the periodic behavior consist of 3 phases. Alternatively, out of phase switches of $h(X(t))$ and $h(X(t-\delta))$ result in 4 slope changes of P_L. These reproduce the 3 phases observed in Figure 2: oil production from the reservoir (1), followed by pressure buildup (2), and eventual pressure drop (3).

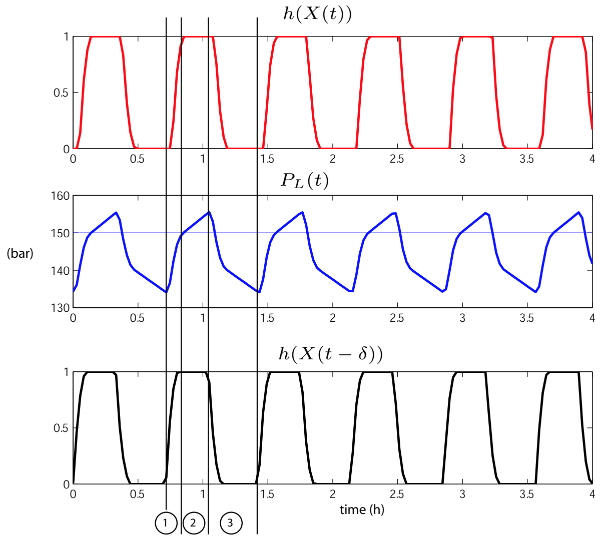


Fig. 3. Density wave simulated with equation (15) in Matlab. The reservoir pressure P_r is 150 bar and u is set at 10 bar.

D. Reference model for stability analysis

Model (15) will be used in Section V-A to design the control law for u. To study stability it is equivalent (but more convenient) to consider X as defined in equation (14). It follows from (15)

$$X(t) = 1 - \frac{P_r - P_L^\star}{u} + \frac{1}{u}\int_{t-\delta}^{t} k(t-\tau)h(X(\tau))d\tau \quad (16)$$

By derivation (assuming u constant), we get

$$u\dot{X}(t) =$$

$$k(0)h(X(t)) - k(\delta)h(X(t-\delta)) + k'(0)\int_{t-\delta}^{t} h(X(\tau))d\tau \quad (17)$$

We consider the system (17) with initial condition ϕ defined and continuous on $[-\delta, 0]$, satisfying:

$$\phi(0) = 1 - \frac{P_r - P_L^\star}{u} + \frac{1}{u}\int_{0-\delta}^{0} k(t-\tau)h(\phi(\tau))d\tau$$

For this class of initial conditions, equations (16) and (17) have the same solutions.

IV. Stability

We first study the stability of the trivial solution of the following saturation-free model derived from equation (17). We denote

$$\tau \triangleq \delta/u \quad (18)$$

and $\mathcal{C}$ the (Banach) space of continuous function mapping the interval $[-\tau, 0]$ into $\mathbb{R}$. We define $x_t \in \mathcal{C}$ as

$$[-\tau, 0] \ni \theta \mapsto x_t(\theta) = x(t+\theta)$$

By derivation and time scaling, equations (17) rewrites

$$\begin{cases} \dot{x}(t) = f(x_t) \text{ for } t \geq 0 \\ x(t) = \phi(t) \text{ for } t \in [-\tau, 0] \end{cases} \quad (19)$$

with $\phi \in \mathcal{C}$ and $f : \mathcal{C} \to \mathbb{R}$ defined as,

$$f(x_t) = ax(t) + bx(t-\tau) + \frac{c}{\tau}\int_{t-\tau}^{t} x(\zeta)d\zeta \quad (20)$$

with $a + b + c = 0$, $b > 0$, $c < 0$ and $b + c > 0$ (by equation (10)). Referring to the formulation used in [7], one can rewrite equation (19) as

$$f(x_t) = \int_{-\tau}^{0} d(\eta(\theta))x_t(\theta) \quad (21)$$

With

$$\begin{cases} \eta(\theta) = (c/\tau)\theta, \quad \theta \in]-\tau, 0[\\ \eta(0) = a \\ \eta(-\tau) = -(c+b) \end{cases}$$

As η is continuous on $]-\tau, 0[$ and has bounded variation on $[-\tau, 0]$, given any $\phi \in \mathcal{C}$, there exists a unique function x_t, continuous, that satisfies system (19). We now study stability of (19) through the solutions of its characteristic equation. As will appear, stability depends on τ.

A. Characteristics equation solutions

The characteristic equation associated with (19) writes

$$s = a + be^{-s\tau} + \frac{c}{s\tau}(1 - e^{-s\tau}) \quad (22)$$

This equation is well defined by continuity at 0 and for all $\tau \geq 0$, 0 is an isolated solution. Referring to the necessary condition expressed in [13], as, for all $\tau \geq 0$

$$\det(\eta(-\tau) - \eta(0)) = -(a + b + c) = 0 \leq 0,$$

the trivial solution is not asymptotically stable.

In the following, we characterize the location of the non zero roots with respect to τ. In Proposition 1, we exhibit a family $(\tau_k)_{k\in\mathbb{N}}$ at which two roots hit the imaginary axis. Then we show that, for small τ, roots are lying on the left half plane (Proposition 2). Further, proving that the roots cross the imaginary axis from left to right, we conclude towards the existence of $\tau^\star > 0$ (Proposition 3) such that

- for $\tau \in [0, \tau^\star[$, all roots except 0 have strictly negative real part
- for $\tau > \tau^\star$, there is at least one root with strictly positive real part.

Proposition 1. *Consider the following system*

$$\dot{x}(t) = ax(t) + bx(t-\tau) + \frac{c}{\tau}\int_{t-\tau}^{t} x(\zeta)d\zeta \quad (23)$$

with $a + b + c = 0$, $b > 0$, $c < 0$, $b + c > 0$ and $\tau > 0$. Let $\lambda = c/b$. There exists $(\tau_k, \omega_k)_{k\in\mathbb{N}} \in \mathbb{R}^+ \times \mathbb{R}^+$ such that, for $\tau = \tau_k$, besides 0 which is always a solution, the pure imaginary roots of the characteristic equation of (23) are $\pm j\omega_k$. This family (τ_k, ω_k) is defined by

$$\begin{cases} \cos(\omega_k\tau_k) = 1 + \dfrac{\lambda\sigma_k}{\sigma_k - (2+\lambda)} \\ \omega_k \sin(\omega_k\tau_k) = \dfrac{c\sigma_k(2+\lambda)}{\sigma_k - (2+\lambda)} \\ \omega_k^2 = b^2(2+\lambda)^2\left(\dfrac{-\lambda}{2+\lambda}\dfrac{\sigma_k}{\sigma_k - 2}\right) \end{cases} \quad (24)$$

with $\sigma_k = (2b+c)\tau_k + 2 > 3 + \lambda$.

Proof: We are now looking for pure imaginary roots of equation (22). If there exists $\tau \geq 0$ such that $j\omega$ is solution then $-j\omega$ is also a solution. Therefore, we restrict our study to $(\tau, \omega) \in \mathbb{R}^+ \times \mathbb{R}^+ \backslash \{0\}$. Equation (22) yields

$$\begin{cases} b\cos(\omega\tau) + \dfrac{c}{\tau\omega}\sin(\omega\tau) = b + c \\ \dfrac{c}{\tau\omega}\cos(\omega\tau) - b\sin(\omega\tau) = \omega + \dfrac{c}{\tau\omega} \end{cases} \tag{25}$$

This implies

$$\omega^2 = -\frac{c}{\tau}\left(2b\tau + c\tau + 2\right) \tag{26}$$

By construction, $\lambda \in]-1, 0[$. Note $\sigma = (2b+c)\tau + 2 \geq 0$. Equation (25) leads to

$$\cos\left(\sqrt{-\frac{\lambda}{2+\lambda}\sigma(\sigma-2)}\right) = 1 + \frac{\lambda\sigma}{(\sigma-2)-\lambda} \tag{27}$$

$$\sin\left(\sqrt{-\frac{\lambda}{2+\lambda}\sigma(\sigma-2)}\right) = \frac{1}{\omega}\frac{c\sigma(2+\lambda)}{\sigma-2-\tau} < 0 \tag{28}$$

We derive from inequality (28) that $\sigma \in \bigcup_{k\in\mathbb{N}} \left[1 + \sqrt{1 + \frac{2+\lambda}{-\lambda}(2k+1)^2\pi^2}, 1 + \sqrt{1 + \frac{2+\lambda}{-\lambda}4k^2\pi^2}\right]$. Right hand side of equation (27) approaches $0 < 1 + \lambda < 1$ as σ goes to infinity. The left hand side is oscillating thanks to the $\cos$ function and equation (27) has an infinite number of solutions. Among these, we keep those compatible with equation (28) and gather them in $(\sigma_i)_{i\in\mathbb{N}}$, an increasing sequence. By construction,

$$\lim_{i\to\infty} \sigma_i = +\infty$$

and

$$\sigma_i \sim_{i+\infty} \sqrt{\frac{2+\lambda}{-\lambda}}2i\pi$$

Further, for all $k \in \mathbb{N}$

$$1 + \sqrt{1 + \frac{2+\lambda}{-\lambda}(2k+1)^2\pi^2} > 1 + \sqrt{1+\pi^2} > 3 + \lambda$$

The set $(\sigma_i)_{i\in\mathbb{N}}$ is thus bounded by below

$$\forall k \in \mathbb{N}, \quad \sigma_k > 3 + \lambda \tag{29}$$

This set defines a family of solutions of equation (25), $(\tau_k, \omega_k)_{k\in\mathbb{N}}$ using equation (26)

$$\tau_k = \frac{\sigma_k - 2}{2b + c}$$
$$\omega_k^2 = b^2(2+\lambda)^2\left(\frac{-\lambda}{2+\lambda}\frac{\sigma_k}{\sigma_k - 2}\right)$$

■

Lemma 1. *Define s a non zero root of the characteristic equation* (22)*. For all $\alpha > -1$, $\beta > 0$ and $\tau_\alpha > 0$ there exists $\tau \leq \tau_\alpha$ such that:*

$$|s(\tau)| > \beta\tau^\alpha$$

Proof: Assume that one can find $(\alpha, \tau_\alpha, \beta)$ ($\alpha > -1$, $\beta > 0$ and $\tau_\alpha > 0$) such that for all $\tau \leq \tau_\alpha$

$$|s| \leq \beta\tau^\alpha$$

Thus $|s\tau| \to 0$ as $\tau \to 0$. A second order development of (22) yields

$$\frac{1}{\tau} = -b - \frac{c}{2} + \left(\frac{b}{2} + \frac{c}{6}\right)s\tau + o(s\tau)$$

The right hand side of this development goes to $-b - c/2$ as $\tau \to 0$ and the left hand side to $+\infty$. This cannot be, therefore, the assumption is false. ■

Lemma 2. *Define s a non zero root of the characteristic equation* (22)*. For all τ_r, there exists $\tau \leq \tau_r$ such that*

$$\textbf{\textit{Re}}(s(\tau)) < 0$$

Proof: Assume that there exists τ_r such that for all $\tau \leq \tau_r$

$$\mathbf{Re}(s(\tau)) \geq 0$$

It follows that $|e^{-s\tau}| \leq 1$ and that $|\frac{1-e^{-s\tau}}{s\tau}| \leq 1$. Using Equation (22) we get

$$|s(\tau)| \leq |a| + |b| + |c|$$

which is in contradiction with Lemma 1. ■

Proposition 2. *There exists $\underline{\tau} > 0$ such that for all $\tau \leq \underline{\tau}$ the roots of the characteristic equation* (22) *that are not zero are strictly lying on the left half plane.*

Proof: Consider a non zero root. From Proposition 1, we know that, for $\tau < \tau_1$, it does not intersect the imaginary axis. Further, we know, from Lemma 2, that there exists $\tau < \tau_1$ such that the root lies on the left half plane. As the real part of the root is continuous with respect to τ (by the implicit function theorem), the root cannot go to the right part without crossing the imaginary axis. This implies that, for all $\tau < \tau_1$ the root is in the left half plane. Finally, for all $\tau \in [0, \tau_1[$, all roots except 0 have strictly negative real part. This concludes the proof. ■

Proposition 3. *There exists τ^* such that for all $\tau \in [0, \tau^*[$ the characteristic equation* (22) *has one root at* 0 *and all other roots strictly in the left part of the complex plane. For all $\tau > \tau^*$ there exists at least one root lying strictly on the right half plane.*

Proof: Let τ be positive. As proven in Proposition 2, for small τ all roots except 0 lie in the left half plane. To know whether these roots become unstable or come back to the left hand side, we compute $\mathbf{Re}\frac{\partial s}{\partial \tau}_{|\pm j\omega_k}$. We use equations (24) and after some computations we get

$$\frac{\partial s}{\partial \tau}_{|s=j\omega} = \frac{-b\omega^2 e^{-j\omega\tau} + \frac{c}{\tau^2}(1 - e^{-j\omega\tau}) - \frac{cj\omega}{\tau}e^{-j\omega\tau}}{-2j\omega - b - c + be^{-j\omega\tau} - bj\omega\tau e^{-j\omega\tau} + ce^{-j\omega\tau}}$$

and

$$\textbf{Re}\frac{\partial s}{\partial \tau}_{|\pm j\omega_k} = - \frac{\lambda b^2(2+\lambda)^3(\sigma_k - 3 - \lambda)}{(\sigma_k^2 + (2\lambda^2 - 2 + 5\lambda)\sigma_k - 2\lambda(3+\lambda)^2)(u_k - 2)}$$

From (29) and noticing that $\sigma_k^2 + (2\lambda^2 - 2 + 5\lambda)\sigma_k - 2\lambda(3+\lambda)^2 > 0$ for $\sigma_k > 2$, we have

$$\textbf{Re}\frac{\partial s}{\partial \tau}_{|\pm j\omega_k} > 0$$

Therefore, after crossing the imaginary axis the roots always go to the right half plane. Thus simply, $\tau^* = \tau_1 = \frac{\sigma_1 - 2}{2b+c}$. ■

B. Conclusion

Parameter τ has a direct impact on the roots location of the characteristic equation. Increasing the time delay τ or letting the roots be unstable are equivalent. Recalling $\tau = \frac{\delta}{u}$, this last remark means that there exists a minimal gas injection rate that guarantees stability of the roots. Study of the characteristic equation is a key to the interpretation of the observed oscillating behavior. Depending on u trajectories of model (17) behave as follows. Unstable solutions of the model (17), which, initially match with unstable solutions of a linear system of type (20), finally reach saturation yielding behaviors depicted in Figure 3. Stable solutions remain bounded and if the initial condition is well chosen (e.g. constant) they do not reach the saturation.

V. Control

In this section, we design control laws to steer system (15) to a predefined steady state.

A. Control laws definition

We look for control laws u such that P_L converges to a chosen constant $P_{\text{ref}} \in \left]P_L^\star + \int_0^\delta k(\tau)d\tau, P_L^\star\right[$. The corresponding steady state value of X defined in (14) is

$$X_{\text{ref}} = -\frac{P_L^\star - P_{\text{ref}}}{\int_0^\delta k(\tau)d\tau} \in]0,1[\tag{30}$$

We note u_{ref} the value of u at steady state. It satisfies

$$X_{\text{ref}}(u_{\text{ref}}) = \frac{P_L^\star - P_r + u_{\text{ref}}}{u_{\text{ref}} - \int_0^\delta k(\tau)d\tau} \tag{31}$$

Our (closed-loop) control law is, simply,

$$u(t) = \frac{P_r - P_L(t)}{1 - X_{\text{ref}}} \tag{32}$$

This control strategy feeds system (15), which has finite memory δ, with a constant term. By direct computation, this straightforward approach provides convergence. We can state the following proposition.

Proposition 4. *With control law* (32), *P_L which dynamics is defined by system* (15) *converges to P_{ref} in finite time δ for any initial condition $[-\delta, 0] \ni t \mapsto \phi(t) \in \mathbb{R}$.*

Yet, the expression u defined in (32) does not take into account actuation saturations. Here, the most limiting factor in practice is a lower bound $u_{\min} > 0$ on the control. It is often reached with this naive approach. We now propose the saturated control law

$$\begin{cases} u = \dfrac{P_r - P_L(t)}{1 - X_{\text{ref}}}, \text{ for } P_L < P_r - u_{\min}(1 - X_{\text{ref}}) \\ \qquad u = u_{\min}, \text{ for } P_L \geq P_r - u_{\min}(1 - X_{\text{ref}}) \end{cases} \tag{33}$$

Proposition 5. *Assume that*

$$X_{ref} \geq \frac{P_L^\star - P_r + u_{min}}{u_{min} - \int_0^\delta k(\tau)d\tau} \tag{34}$$

With the (saturated) control law (33), *P_L which dynamics is defined by system* (15), *converges to P_{ref} in finite time 2δ for any initial condition $[-\delta, 0] \ni t \mapsto \phi(t) \in \mathbb{R}$.*

Proof: We now show that for $t \geq \delta$, the control law is unsaturated. Indeed, for $t > 0$

$$h\left(1 - \frac{P_r - P_L(t)}{u(t)}\right) \geq X_{\text{ref}}$$

Therefore, for all $t \in [\delta, +\infty[$,

$$P_L(t) \leq P_L^\star + X_{\text{ref}} \int_0^\delta k(\tau)d\tau$$

Assuming (34), a simple computation yields

$$\forall t \geq \delta,\ P_L(t) \leq P_r - u_{\min}(1 - X_{\text{ref}})$$

By equation (33), we get that, for all $t \geq \delta$, u is simply defined by

$$u = \frac{P_r - P_L(t)}{1 - X_{\text{ref}}}$$

The control is thus unsaturated and, by Proposition 4, we conclude that system (15) converges towards P_{ref} in 2δ. ■

In practice, one must choose P_{ref} in accordance to the minimum value $u_{\min}$ such that equation (31) holds. This choice implies that assumption (34) holds.

Indeed, if $P_r > P_L^\star + \int_0^\delta k(\tau)d\tau$, which simply means that the pressure at the bottom of pipe when it is full of gas is smaller than the reservoir pressure, then $u_{\text{ref}} \mapsto X_{\text{ref}}(u_{\text{ref}})$, given in (31), is increasing. Therefore, if $u > u_{\min}$

$$X > \frac{P_L^\star - P_r + u_{\min}}{u_{\min} - \int_0^\delta k(\tau)d\tau}$$

The meaning of assumption (34) is that one should not define P_{ref} outside the range of X that can be reached with $u > u_{\min}$.

B. Simulation

Figure 4 shows an example of stabilization with the saturated control law (33). Choosing $u = 10$ bar and using equations (30) and (31) we compute the corresponding steady states $P_{\text{ref}} = 145$ bar and $X_{\text{ref}} = 0.464$. We define $u_{\min} = 0.1$, which satisfies assumption (34). Until t_c the system is left open loop. At $t = t_c$, the controller is turned on. From t_c to $t_c + \delta$, the gas mass fraction $h(X(t))$ remains between X_{ref} and 1. Therefore, for $t > t_c + \delta$, $P_l(t)$ remains below $P_{\max} = P_r - u_{\min}(1 - X_{\text{ref}}) = 150$ bar and $h(X(t)) = X_{\text{ref}}$. Pressure P_L converges to P_{ref} in finite time.

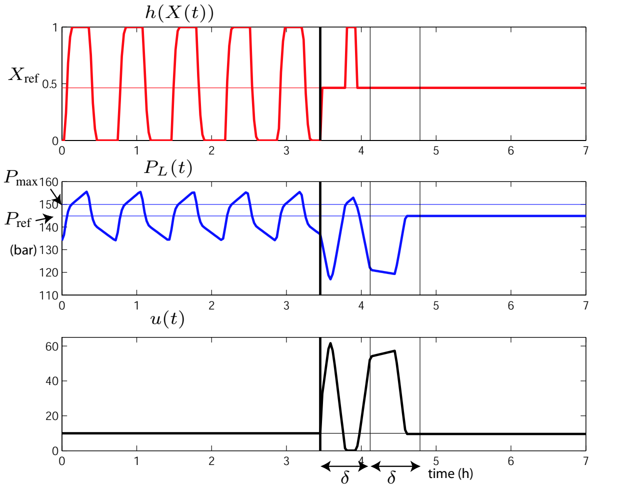


Fig. 4. Stabilization of equation (15) using the saturated control law (33). Control is switched on after approximatively 3.4 hours of open loop. P_L reaches P_{ref} and X reaches X_{ref} in finite time 2δ.

An alternative view is given in Figure 5. Left three snapshots describe the open-loop behavior. Gas mass fraction profile, $x(t, z)$, is represented in white (complementary black part stands for oil mass fraction). Boundary condition q_g is constant and q_l is defined by equation (11). Finally, the right scheme represents the transient obtained with closed loop control. The feeds keep the gas mass ratio constant at X_{ref}. During the transient, q_g is permanently adapted to counteract the effect of the state $x(t, z)$, $z \in [0, L]$, onto q_l. This yields a constant $X(t, L)$ which progressively steers the system to steady state through the transport equation.

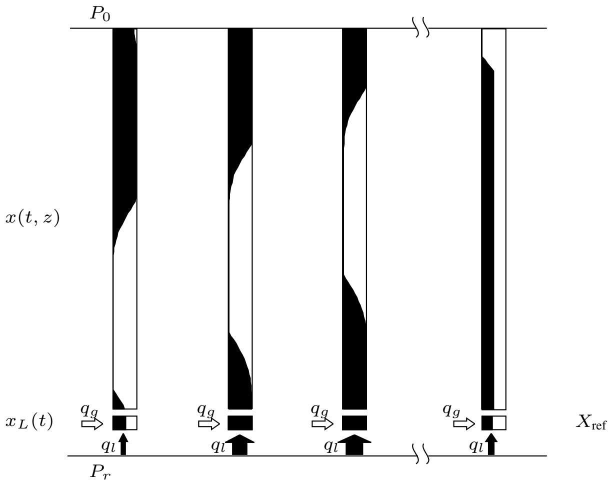


Fig. 5. Comparison of open loop (3 schemes on the left) and closed loop behavior.

C. Stabilization of the well simulated in OLGA®2000

The closed loop control law can be tested in OLGA®2000 Transient Multiphase Flow Simulator. A realistic dynamic oil-gas model is used along with semi-implicit numerical solver (see [11] for details).

In Section III-A, we assume the gas velocity to be constant, i.e. we neglect the impact of the gas mass fraction on the gas velocity. Therefore, when the gas mass flow rate is high enough, this assumption only results in a time depending time dilatation. But when the gas inlet is too low the well production eventually stops, which is not represented by the simple model. Therefore we want the gas injection rate to remain above a minimum, $q_g^{\min}$, guaranteeing the flow in the pipe. This defines a lower bound for our control law. Following the same lines as in the previous section, we define the control q_g, corresponding to u (see equation (13)), to keep x^L at a predefined constant, X_{ref}. Using x^L given in equation (12), our control law writes

$$q_g(t) = \max\left(\frac{X_{ref}}{1 - X_{ref}} IP(P_r - P_L(t)), q_g^{\min}\right)$$

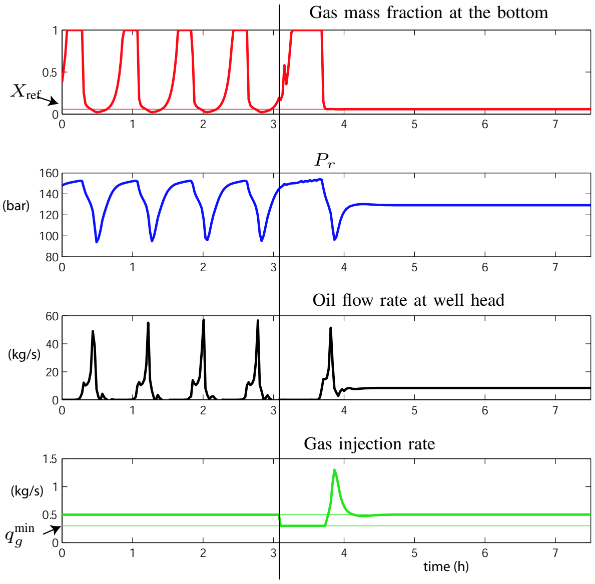


Fig. 6. Stabilization of density wave instability simulated in OLGA®2000. $X_{\text{ref}} = 0.0568$ and $q_g^{\min} = 0.3$.

Figure 6 shows an example of stabilization of density wave instability. We define $q_g^{\min} = 0.3$ kg/s and $X_{\text{ref}} = 0.0568$. The controller is switched on at the black line and steers the well to the steady state corresponding to the initial gas injection rate. As the period of the oscillations corresponds approximately to the travel time δ, we see in Figure 6 that the well is stabilized in 2δ. As shown in Proposition 5, 2δ corresponds to the time needed by the well to forget its initial condition.

VI. Conclusion

In this paper we propose an interpretation of the observed oscillations in the tubing of gas lifted wells. A distributed parameter model has been derived for the propagation of pressure (system (15)). It describes the dynamics as a transport phenomenon with state dependent boundary condition.

This equation is shown to be equivalent to a saturated linear delay model (equation (17)) involving the gas fraction. Analysis of the underlying characteristic equation is performed (for unsaturated solutions) and show that the critical parameter is the amount of injected gas. This is consistent with state-of-the-art and suggests a simple control strategy. Performance of the derived control strategy is demonstrated through OLGA® 2000 simulations, proving that it is possible to obtain a steady flow with the same amount of injected gas, after a finite time transient during which the oscillation is cancelled. The main restriction of this strategy is that downhole measurements are often not available. Therefore we are investigating a way to maintain the gas mass fraction constant at the entrance of the tubing using only topside measurements.

Acknowledgements: The authors are indebted to Emmanuel Duret and Pierre Rouchon for insightful discussions.

REFERENCES

[1] O. M. Aamo, G. O. Eikrem, H. Siahaan, and B. Foss, "Observer design for multiphase flow in vertical pipes with gas-lift - theory and experiments," *Journal of Process Control*, vol. 15, pp. 247–257, 2005.

[2] E. P. Blick, P. N. Enga, and P. C. Lin, "Theoretical stability analysis of flowing oil wells and gas-lift wells," *SPEPE*, pp. 508–514, 1988.

[3] K. E. Brown, *Gas lift theory and practice*. Petroleum publishing CO., Tulsa, Oklahoma, 1973.

[4] A. J. Chorin and J. E. Marsden, *A mathematical introduction to fluid mechanics*. Springer-Verlag, 1990.

[5] E. Duret, "Dynamique et contrôle des écoulements polyphasiques," Ph.D. dissertation, École des Mines de Paris, 2005.

[6] G. O. Eikrem, L. Imsland, and B. Foss, "Stabilization of gas-lifted wells based on state estimation," in *Proc. of the* 2*nd IFAC Symposium on System, Structure and Control*, 2004.

[7] J. K. Hale and S. M. V. Lunel, *Introduction to functional differential equations*. Springer-Verlag, 1993.

[8] B. Hu and M. Golan, "Gas-lift instability resulted production loss and its remedy by feedback control: dynamical simulation results," in *SPE International Improved Oil Recovery Conference in Asia Pacific*, no. SPE 84917, Kuala Lumpur, Malaysia, 2003.

[9] L. S. Imsland, "Topics in nonlinear control - ouput feedback stabilization and control of positive systems," Ph.D. dissertation, Norwegian University of Science and Technology, Department of Engineering and Cybernetics, 2002.

[10] B. Jansen, M. Dalsmo, K. Havre, L. Nøkleberg, V. Kritiansen, and P. Lemétayer, "Automatic control of unstable gas-lifted wells," *SPE Annual technical Confererence and Exhibition.*, no. SPE 56832, October 1999.

[11] Scandpower, *OLGA®2000 User's Manual*. Scandpower, 2004.

[12] L. Sinègre, N. Petit, P. Lemétayer, P. Gervaud, and P. Ménégatti, "Casing heading phenomenon in gas lifted well as a limit cycle of a 2d model with switches," in *Proc. of the* 16*th IFAC World Congress*, 2005.

[13] G. Stépán, *Retarded dynamical systems: stability and characteristic functions*, ser. Pitman Research Notes in Math. Series. UK: Longman Scientific, 1989.

CASING-HEADING PHENOMENON IN GAS-LIFTED WELL AS A LIMIT CYCLE OF A 2D MODEL WITH SWITCHES

Laure Sinègre [*] **Nicolas Petit** [*] **Pierre Lemétayer** [**]
Philippe Gervaud [**] **Philippe Ménégatti** [**]

[*] *CAS, École des Mines de Paris, France*
[**] *CSTJF, TOTAL Exploration-Production, Pau, France*

Abstract: Oil well instabilities cause production losses. One of these instabilities, referred to as the "casing-heading" is an oscillatory phenomenon occurring on gas-lift artificially lifted well. This behavior is well represented by a 2D model with a vector field that is not continuously differentiable across several switching curves. These correspond to switches in flow rate functions describing the valves. In order to interpret the observed oscillations as a limit cycle we use the Poincaré-Bendixon theorem with a detailed study of uniqueness of trajectories and the derivation of a positive invariant set. Apart from the general case considered here, an illustrative example is given. The vector field is explicited and a similar limit cycle appears.

Keywords: Process Control, Dynamic Systems, Limit Cycles, Switching System, Gas-Lifted Well.

1. INTRODUCTION

Producing oil from deep reservoirs and lifting it through wells to the surface facilities often requires activation to maintain the oil output at a commercial level. In the gas-lift activation technique (Brown, 1973), gas is injected at the bottom of the well through the injection valve (point C in Figure 1) to lighten up the fluid column and to lower the gravity pressure losses. High pressure gas is injected at the well head through the gas valve (point A in Figure 1), then goes down into the annular space between the drilling pipe (casing, point B) and the production pipe (tubing, point C) where it enters. The oil produced from the reservoir (point F) and the injected gas mix in the tubing. They flow through the production valve E located at the surface.

Since 1986, a system for automatic handling of such wells, FCW (Full Control of Wells) has been developed by TOTAL. Wells have been operated by FCW since 1988. This tool schedules the opening of valves A and E following a sequential logic algorithm which steers the system to a prescribed setpoint. These can be stable or unstable. Details can be found in (Lemetayer and Miret, 1991).

High yield setpoints (low gas and high oil output) lie in an unstable region (Jansen *et al.*, 1999). A periodic phenomenon called "casing-heading" can appear. It consists of a succession of pressure build-up phases in the casing without production and high flow rate phases. These oscillations reduce the overall oil production and may damage the reservoir well interface and the facilities. Currently FCW does not fully address such dynamical instabilities.

This "casing-heading" instability is accurately represented by multiphase partial differential equations models (such as those implemented in Indiss$^{\mathrm{TM}}$-IProd or Olga$^{\circledR}$2000). Yet, simpler mod-

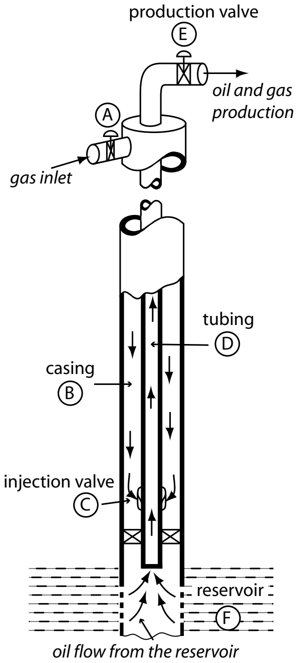


Fig. 1. Scheme of a gas-lift activated well.

els can be used. In (Imsland, 2002; Eikrem *et al.*, 2003) a three balance ordinary differential equations model is used as the well dynamics. Numerical simulations prove the relevance of this approach. Further studies reveal that, as it is assumed that the gas mass fraction is constant with respect to the depth, the 3D model can be reduced to a 2D one (the masses of oil and gas in the tubing are highly correlated). This assumption eliminates possible instabilities due to propagation and thus let us focus on the casing-heading phenomenon. This representation is handy to interpret the casing-heading oscillations as a limit cycle. The contribution of this paper is to explain the observed planar limit cycle (e.g see Figure 2 for a sample Indiss$^{\mathrm{TM}}$-IProd multiphase well simulation – exact scales are omitted for confidentiality reasons) through the Poincaré-Bendixon theorem. This system is related to other work on hybrid systems, such as the two-tank example addressed in (Hiskens, 2001), or the generalization of the Poincaré-Bendixon theorem to planar hybrid systems by (Simić *et al.*, 2002). Yet, several specific issues have to addressed here. The model includes two switching curves. These model the flow rate through the two valves (A and E). According to classic Saint-Venant laws (refer to (*Standard Handbook of Petroleum and Natural Gas Engineering*, 1996)) the flow rate is non differentially smooth around zero. The model is thus non differentially smooth across the switching curves. Therefore proving existence and uniqueness of the trajectories requires special care and does not directly derive from a Lipschtiz-continuity assumption.

The article is organized as follows. The system under consideration is presented in Section 2. In Section 3 a positive invariant set is constructed. In Section 4 existence and uniqueness of the trajectories are addressed through detailed studies around switching curves and their intersections. A future goal is to stabilize the system to the inner setpoint or to shrink the limit cycle. For that purpose a normalized sample problem is given for further reference. Its dynamics are explicited in Section 5. It exhibits a similar limit cycle. We hope it can serve as a test bench for various control techniques.

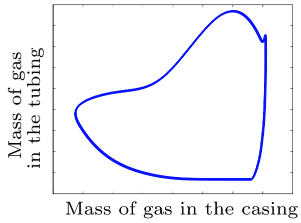


Fig. 2. Projection of a limit cycle obtained with the Indiss$^{\mathrm{TM}}$-IProd multiphase simulator.

2. DYNAMICS DEFINITION

2.1 Notations

We represent the behavior of the well around an unstable setpoint by the following dynamics over $[\underline{x},\overline{x}] \times [\underline{y},\overline{y}] \subset \mathbb{R}^+ \times \mathbb{R}^+$

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \varepsilon w_{gc}(x) - w_{iv}(x,y) \\ w_{iv}(x,y) - \mu w_{pg}(y) \end{pmatrix} \tag{1}$$

We note $\mathcal{X} \triangleq [\underline{x},\overline{x}]$, $\mathcal{Y} \triangleq [\underline{y},\overline{y}]$, $X \triangleq (x,y)^t$ and $\dot{X} = F(X) = (F_1(x),F_2(x))^T$. This 2D dynamics is a restriction of the 3D one defined in (Eikrem *et al.*, 2003). w_{gc}, w_{iv} and w_{pg} are the gas flow rate through the gas valve A, through the injection valve C and through the production valve E. x and y represent the mass of gas in the casing and in the tubing. The positive parameters ε and μ stand for the openings of valves A and E. $\phi(\cdot,X_0)$ denote the solution of Equation (1) with X_0 as initial condition.

2.2 Hypothesis

We assume that both w_{iv} and w_{pg} vanish over their definition intervals. Let $\partial\mathcal{F}^o_{iv}$ and $\partial\mathcal{F}^o_{pg}$ be the boundaries of the sets $w^{-1}_{iv}(0)$ and $w^{-1}_{pg}(0)$. We assume the following hypothesis hold.

(H1) $w_{gc} : \mathbb{R} \to \mathbb{R}$ is C^1, strictly decreasing and does not vanish.

(H2) $w_{iv} = g_{iv} \circ \tau_{iv}$

- $\tau_{iv} : \mathbb{R}^2 \to \mathbb{R}$ is C^2, and strictly increasing w.r.t x and y.

- $g_{iv} : \mathbb{R} \to \mathbb{R}^+$, is C^0, strictly increasing over $\mathbb{R}^+$, C^1 over $\mathbb{R}/\{0\}$, and non Lipschitz at 0. $g_{iv}(0) = 0$. g'_{iv} is decreasing over $\mathbb{R}^+\backslash\{0\}$. $g'_{iv} \sim t^\lambda$ with $-1/2 < \lambda < 0$.

(H3) $w_{pg} = g_{pg} \circ \tau_{pg}$

- $\tau_{pg} : \mathbb{R}^2 \to \mathbb{R}$ is C^1, strictly increasing w.r.t. y, and does not depend on x.
- $g_{pg} : \mathbb{R} \to \mathbb{R}^+$, is C^0, strictly increasing over $\mathbb{R}^+$ and C^1 over $\mathbb{R}/\{0\}$, non Lipschitz at 0. $g_{pg}(0) = 0$.

(H4) τ_{iv} and τ_{pg} vanish over $\mathcal{X} \times \mathcal{Y}$. We define $\partial\mathcal{F}^o_{iv} \triangleq \tau^{-1}_{iv}(0)$ and $\partial\mathcal{F}^o_{pg} \triangleq \tau^{-1}_{pg}(0)$.

In order to construct a polygon $\mathcal{P}$ such as defined later on in Section 3.1 we need some further assumptions.

(H5) $\forall x \in \mathcal{X}$, $\dot{y}(x, \overline{y}) < 0$
(H6) $\dot{x}(\overline{x}, y_{pg}) < 0$
(H7) $\forall x \in \mathcal{X}$, $\tau_{iv}(x, \underline{y}) \leq 0$
(H8) $\forall y \in \mathcal{Y}$, $\tau_{iv}(\underline{x}, \underline{y}) \leq 0$

where, thanks to the continuity of w_{pg}, $y_{pg} \triangleq \max\{y / w_{pg}(y) = 0\}$.

One last assumption (H9) is that a constant K uniquely defined later on in Section 4.3 by the functions above is not zero.

2.3 Existence conditions of a limit cycle

Let $\Omega(\phi)$ be the limit set of ϕ. According to the Poincaré-Bendixon theorem as expressed in (Miller and Michel, 1982), the fact that $\Omega(\phi)$ contains no critical point combined to the uniqueness of the solution of Equation (1) is sufficient to guarantee the existence of a limit cycle. On the other hand, exhibiting a positive invariant set containing no stable equilibrium implies that $\Omega(\phi)$ contains no critical point. Therefore we can simply check that

- there exists a positive invariant set (this will be shown in Section 3),
- given an particular initial condition the solution is uniquely defined (this will be addressed in Section 4).

3. POSITIVE INVARIANCE

3.1 Some useful lemmas

Let $\mathcal{P}$ be a polygon ($(P_i)_{i\in[1,N]}$ its vertexes) such that

$$\forall i \in [1, N], \exists \lambda \text{ such that } \overrightarrow{P_iP_{i+1}} = \lambda F(P_i) \quad (2)$$

Classically, $\mathcal{P}$ is a positive invariant set if and only if

$$\forall X_0 \in \partial\mathcal{P}, \exists t > 0 \text{ s.t. } \forall \epsilon \in [0, t] : \phi(\epsilon, X_0) \in \mathcal{P} \quad (3)$$

Lemma 1. Assume that F is C^n on a neighborhood of X_0, with $X_0 \in [P_i, P_{i+1}]$. Define $u = \frac{P_1 \times P_2}{\|P_1 \times P_2\|}$. If there exists $k \in [1, n]$ s.t.

$$\begin{cases} F(P_i) \times \dfrac{d^j\phi}{dt^j}(0, X_0) \cdot u = 0, j = 1..k-1 \\ F(P_i) \times \dfrac{d^k\phi}{dt^k}(0, X_0) \cdot u > 0 \end{cases}$$

then condition (3) holds.

Proof 1. A sufficient condition for condition (3) to be satisfied is that

$$\overrightarrow{P_iP_{i+1}} \times \overrightarrow{P_i\phi(\epsilon, X_0)} \cdot u > 0$$

This is equivalent to

$$A(\epsilon, X_0) = F(P_i) \times \overrightarrow{X_0\phi(\epsilon, X_0)} \cdot u > 0 \quad (4)$$

Since F is C^n on a neighborhood of X_0, an expansion of $A(\cdot, X_0)$ is

$$A(\epsilon, X_0) = \epsilon^{k-1}(F(P_i) \times \frac{d^k\phi}{dt^k}(0, X_0) \cdot u + o(1))$$

Therefore $A(\cdot, X_0)$ is strictly positive and condition (3) is satisfied. ∎

Similarly one can prove that

Lemma 2. Let $X_0 \in [P_i, P_{i+1}]$ and $(j, l) \in \{(1, 2); (2, 1)\}$. Assume that $F_j(P_i) = 0$. If F_l is continuous around X_0 and F_j is C^1, a sufficient condition leading to (3) is

$$\left.\begin{matrix} (-1)^j \dot{x}_l(P_i)\dot{x}_j(X_0) > 0 \text{ or} \\ \begin{cases} \dot{x}_l(P_i)\dot{x}_j(X_0) = 0 \\ (-1)^j \dot{x}_l(P_i)\ddot{x}_j(X_0) > 0 \end{cases} \end{matrix}\right\} \quad (5)$$

Corollary 1. If $F_j(P_i) = 0$ and if F_j and F_l are only C^0, a more restrictive condition is

$$(-1)^j \dot{x}_l(P_i)\dot{x}_j(X_0) > 0$$

3.2 Positive invariant set candidate

Two curves play a key role in the construction of the candidate rectangle $\mathcal{P} = (P_1P_2P_3P_4)$. These are the set $\{(x, y) / \dot{x} = 0\}$ and the set $\{(x, y) / \dot{y} = 0\}$. We show that this rectangle, which is illustrated in Figure 3, satisfies Equation (2).

P_1, P_2 and P_3 construction Let ψ be defined by

$$\psi(x) \triangleq \varepsilon w_{gc}(x) - w_{iv}(x, y_{pg})$$

From $(H6)$ and $(H8)$, $\psi(\underline{x}) > 0$ and $\psi(\overline{x}) < 0$. Since ψ is continuous, increasing, we can uniquely define

$$x_1 = \max\{x / \psi(x) = 0\}$$

We note $P_1 \triangleq (x_1, y_{pg})$. At that point $\dot{x}$ and w_{pg} vanish. Further, similar arguments relying on (H5), and (H2)-(H8) respectively, uniquely define $P_2 \triangleq (x_1, y_2)$ with $y_2 \triangleq \min\{y / \dot{y}(x_1, y) = 0\}$ and $P_3 \triangleq (x_3, y_2)$ with $x_3 \triangleq \max\{x / \dot{x}(x, y_2) = 0\}$.

P_4 construction Let $P_4 \triangleq (x_3, y_{pg})$. $[P_3, P_4]$ is tangent to the field at P_3. Further, $[P_4, P_1]$ is tangent to the field at P_4. This arises from the the following argument. As w_{iv} is cancelling at $(\underline{x}, y_{pg})$ and strictly positive at P_1, we can choose ε parameter in Equation (1) such that $[P_4, P_1] \cap \partial\mathcal{F}_{iv}^o \neq \varnothing$. Therefore $w_{iv}(P_4) = 0$. As a consequence $\dot{x}(P_4) > 0$ and $\dot{y}(P_4) = 0$.

3.3 Intersections with switching lines

Let $X_{iv}^2 \triangleq (x_{iv}, y_{pg})$ with $x_{iv} = \max\{x/(x, y_{pg}) \in [P_4, P_1] \cap \partial\mathcal{F}_{iv}^o\}$. Remembering that $w_{iv}(P_3) = \varepsilon w_{gc}(P_3) > 0$ we conclude $[P_3, P_4] \cap \partial\mathcal{F}_{iv}^o \neq \varnothing$. We note $X_{iv}^1 \triangleq (x^3, y_{iv})$ with $y_{iv} \triangleq \max\{y/(x^3, y) \in [P_3, P_4] \cap \partial\mathcal{F}_{iv}^o\}$.

3.4 Positive invariance

Let X_0 be a point on the side of the rectangle. We want to prove that the trajectory $\phi(\cdot, X_0) = (\phi_x, \phi_y)^t$ starting at X_0 remains inside $\mathcal{P}$ for $t > 0$. We assume that trajectories are uniquely defined, this will be proven at Section 4.

Using Lemma 2 at points where F_2 is not C^1 Let $X_0 \in [P_1, P_2]$. F_1 vanishes at P_1, so F_1 being C^1 and F_2 only continuous around X_0 will complete the list of hypothesis needed to apply Lemma 2. F_2 is continuous by definition and F_1 is C^1, because $\forall X_0 \in [P_1, P_2]$

$$w_{iv}(X_0) \leq w_{iv}(P_1) = \varepsilon w_{gc}(P_1) > 0$$

Therefore checking condition (5) of Lemma 2 will prove that the trajectory starting at X_0 goes inside ($\mathcal{P}$). If $X_0 \in]P_1, P_2]$ the condition rewrites $-\dot{y}(P_1)\dot{x}(X_0) > 0$. As $-w_{iv}$ is decreasing w.r.t. y, $\dot{x}(X_0) < 0$. Adding that $\dot{y}(P_1) > 0$ ensures that the condition holds. If $X_0 = P_1$ the condition rewrites $-\dot{y}(P_1)\ddot{x}(X_0) > 0$. As $\ddot{x}(X_0) = -\partial_y w_{iv}(X_0)\dot{y}(X_0) < 0$ this condition holds. Following along the same lines it is easy to check that Lemma 2 can be applied at every point of $\partial\mathcal{P}$ except X_{iv}^1 and $[P_4, P_1]$. At these points the C^1 condition is not verified. Notice also that at each vertex two conditions have to be verified, one for each side.

Using Corollary 1 at points where F_1 and F_2 are only C^0 When X_0 is an element of $X_{iv}^1 \cup]X_{iv}^2, P_1]$ none of F coordinates vanish, therefore we can simply use the fact that F is continuous to apply Corollary 1. So for $X_0 = X_{iv}^1$ the condition is $-\dot{x}_2(P_3)\dot{x}_1(X_0) > 0$ which is easily checked. At $X_0 \in]X_{iv}^2, P_1]$ the condition is $\dot{x}_1(P_4)\dot{x}_2(X_0) > 0$.

A proof by contradiction when $X_0 \in [P_4, X_{iv}^2]$ Neither Lemma 2 (F_2 is not C^1) nor Corollary 1 ($\dot{y}(X_0) = 0$) can be used here. Yet, we can prove that a solution starting at X_0 cannot go below $y = y^{pg}$. Assume that there exists t_2 such that $\phi_y(t_2) < x_2^{pg}$, define t_1 such that

$$\begin{cases} \forall t \in]t_1, t_2],\ \phi_y(t) < x_2^{pg} \\ \phi_y(t_1) = x_2^{pg} \end{cases} \tag{6}$$

Refering to the mean value theorem $\phi_y(t_2) = \phi_y(t_1) + (t_2 - t_1)\phi_y'(t_c)$ with $t_c \in [t_1, t_2]$. $\phi_y'(t_c) = 0$ implies $\phi_y(t_2) = \phi_y(t_1)$ which contradicts (6). Finally, as the trajectory starting at $X_0 \in \partial\mathcal{P}$ satisfies condition (3), $\mathcal{P}$ defines a positive invariant set.

4. EXISTENCE AND UNIQUENESS OF THE TRAJECTORIES

The first hypothesis required by the Poincaré-Bendixon theorem is the existence and forward uniqueness of the solutions. Existence of a solution of (1) starting at $X_0 \in \mathcal{X} \times \mathcal{Y}$ follows from the continuity of F. Uniqueness of a solution of (1) starting at $X_0 \in (\mathcal{X} \times \mathcal{Y})/(\partial\mathcal{F}_{iv}^o \cup \partial\mathcal{F}_{pg}^o)$ follows from the differentiable continuity of F around X_0.

4.1 Decoupling

Consider $X_0 \in [P_4, X_{iv}^2[\subset \partial\mathcal{F}_{pg}^o$. w_{iv} is null at P_1 and increasing with respect to x, so it cancels over $[P_4, X_{iv}^2]$. In a neighborhood of any point of this segment the system is decoupled. At this point the system writes

$$\begin{cases} \dot{x}(X_0) = \varepsilon w_{gc}(x_0) \\ \dot{y}(X_0) = -\mu w_{pg}(y_0) \end{cases}$$

Both right hand sides are decreasing functions because w_{pg} is increasing and w_{gc} is decreasing. Thus the solution starting at X_0 is unique (see (Brauer and Nohel, 1989)).

Let X_0 be $\in \partial\mathcal{F}_{iv}^o$, such that $F(X_0) \cdot \nabla\tau_{iv}(X_0) < 0$. Let ϕ be a solution starting at X_0. F being continuous and bounded in a neighborhood of X_0, we can define $T > 0$ such that $\forall t < T$, $X_0\phi(t) \cdot \nabla\tau_{iv}(X_0) > 0$. Therefore the solutions of (1) are the solutions of the decoupled system

$$\begin{cases} \dot{x} = \varepsilon w_{gc}(x) \\ \dot{y} = -\mu w_{pg}(y) \end{cases}$$

Each equation has a unique solution, so there exists a unique solution starting at X_0.

4.2 Transversality argument

Let $X_0 \in \{X \in \partial\mathcal{F}_{iv}^o / F(X) \cdot \nabla\tau_{iv}(X) > 0\} \cup [X_{iv}^2, P_1]$. Rewriting dynamics (1) in the (y, z) coordinates, with $z = \tau_{iv}(x, y)$, yields

$$\begin{cases} \dot{z} = F(\xi(y,z),y) \cdot \nabla \tau_{iv}(\xi(y,z),y) \\ \qquad\qquad \dot{y} = g_{iv}(z) - \mu w_{pg}(y) \end{cases} \tag{7}$$

where ξ is a C^2 function defined from the implicit function theorem applied to $z = \tau_{iv}(\xi(y,z),y)$. The decoupling argument does not hold anymore, but we can use the transversality property at 0, $\dot{z}$ is strictly positive, therefore $\exists \alpha^-, \alpha^+, T \in \mathbb{R}^+ \backslash \{0\}$ such that $\forall t \in [0,T]$

$$z_0 + \alpha^- t \leq z(t) \leq z_0 + \alpha^+ t \tag{8}$$

When $y_0 = \underline{y}$ and $z_0 \neq 0$, $\dot{y}(0)$ is strictly positive which allow us to define $\beta^-, \beta^+, T \in \mathbb{R}^+ \backslash \{0\}$

$$y_0 + \beta^- t \leq y(t) \leq y_0 + \beta^+ t \tag{9}$$

Now consider two distinct solutions (y_1, z_1) and (y_2, z_2), let $e_y \triangleq y_2 - y_1$ and $e_z \triangleq z_2 - z_1$. The key of the proof is to use equation (8) to define an upper-bound to $|e| = |(e_y, e_z)|$. From (8) and (9) we deduce that $\forall t \in]0,T]$ $y(t) > y_0$ and $z(t) > 0$. Therefore the solution of (7) starting at that point is unique. In the case of $(y_0, z_0) = (\underline{y}, 0)$ this property still holds. The two solutions $(\underline{y}_1, z_1)$ and (y_2, z_2) cannot split but at $t = 0$. Furthermore we define T' such that e_y, e_z and their derivatives remain positive over $]0, T']$. The dynamics rewrites as Equation (10). We replace the C^1 functions $\partial_x \tau_{iv}$, $\partial_y \tau_{iv}$ and w_{gc} by their first order expansion around X_0 in the first equation of (10)

$$\dot{z} = A - Bg_{iv}(z) - C\mu w_{pg}(y) + Dz + Ey + R(y,z) \tag{11}$$

With $A > 0$, $C > 0$ and

$$\lim_{(y,z) \to (y_0,0)} \frac{R(y,z)}{|(y,z) - (y_0,0)|} = 0 \tag{12}$$

Using the mean value theorem, we can define $(y_c, y_c', y_c'') \in [y_1, y_2]$ and $(z_c, z_c', z_c'') \in [z_1, z_2]$ such that the dynamics of e is

$$\begin{cases} \dot{e}_y = -\mu w_{pg}'(y_c) e_y + g_{iv}'(z_c) e_z \\ \dot{e}_z = (-C\mu w_{pg}'(y_c') + E + \partial_y R(y_c'', z_2)) e_y \\ \qquad + (-B g_{iv}'(z_c') + D + \partial_z R(y_1, z_c'')) e_z \end{cases} \tag{13}$$

Recalling (12) one can define T', k and k' such that over $]0, T']$

$$\dot{e}_z \leq (-C\mu w_{pg}'(y_c') + kE) e_y + (-B g_{iv}'(z_c') + k'D) e_z$$

To define the upper-bound of (13), we recall the transversality argument. g_{iv}' being monotonous we deduce

$$\begin{cases} 0 \leq \dot{e}_y \leq g_{iv}'(z_0 + \alpha^{\pm} t) e_z \\ 0 \leq \dot{e}_z \leq kE e_y + (-B g_{iv}'(z_0 + \alpha^{\pm} t) + k'D) e_z \end{cases} \tag{14}$$

Notice that for $z_0 > 0$ we do not need the linear bounds of (8) to derive a proper upper-bound in (14). Yet, for $z_0 = 0$ the upper-bound goes to infinity, therefore we use that $\dot{z}(0)$ is not zero. Remark also that this kind of hypothesis is not required for $\dot{y}$. Integrating between s and t ($t < \min(t', t'')$ and $s > 0$) gives

$$e(t) \leq \int_s^t A(u) e(u) du + e(s)$$

$$\text{with } A(t) = \begin{pmatrix} 0 & g_{iv}'(z_0 + \alpha^{\pm} t) \\ kE & (-B g_{iv}'(z_0 + \alpha^{\pm} t) + k'D) \end{pmatrix}$$

Using $|A| = \sum_{i,j=1}^{2} |a_{ij}|$ we deduce

$$|e(t)| \leq \int_s^t |A(u)| |e(u)| du + |e(s)|$$

Therefore the Gronwall inequality theorem((Brauer and Nohel, 1989)) yields

$$|e(t)| \leq |e(s)| \exp\left(\int_s^t |A(u)| du\right) \tag{15}$$

As the exponential term is bounded, the limit of the right-hand side of equation (15) is also 0 when s goes to 0 which concludes the proof.

4.3 Non transverse case

Define X_0 such that $X_0 \in \partial \mathcal{F}_{iv}^o$ and $F(X_0) \cdot \nabla \tau_{iv}(X_0) = 0$. The initial conditions of equation (7) become $\dot{z}(0) = z(0) = 0$, $\dot{y}(0) < 0$ and $y(0) > y_{pg}$. In inequality (8), $\dot{z}(0) = 0$ yields $\alpha^{\pm} = 0$. The upper-bound $|A(u)|$ goes to infinity as u goes to zero. System (14) does not give further result. Yet, using $y \sim y_0 + \dot{y}(0)t$, Equation (11) yields

$$\dot{z} \sim Kt - Bg_{iv}(z)$$

with

$$K = (E - C\mu w_{pg}'(y_0)) \tag{16}$$

The role of assumption (H9) appears here as a substitute to the transversality property of Section 4.2. It implies that when the field is tangent to the switching curve there exists a non vanishing higher order forcing term (which actually arises from the coupling of the y dynamics onto the z dynamics). Using L'Hospital's rule we find that Kt is the predominant term. Thus, for a given K, the solutions are positive or negative exclusively. Therefore, if $K < 0$ we use the decoupling argument to conclude to uniqueness. If $K > 0$ we use $z \sim Kt^2/2$ instead. As $t \mapsto g_{iv}'(t^2)$ is integrable around 0 the exponential term of the right-hand side of Equation (15) is bounded, therefore letting s go to zero yields $e(t) = 0$.

4.4 Conclusion

Away from $\partial \mathcal{F}_{iv}^o \cup \partial \mathcal{F}_{pg}^o$ uniqueness follows from the differentiable continuity of F. Points at which the field points toward the $\tau_{iv} < 0$ zone were studied in Section 4.1 where a decoupling argument

$$\begin{cases} \dot{z} = \partial_x \tau_{iv}(\xi(y,z),y)(\varepsilon w_{gc}(\xi(y,z)) - g_{iv}(z)) + \partial_y \tau_{iv}(\xi(y,z),y)(g_{iv}(z) - \mu w_{pg}(y)) \\ \dot{y} = g_{iv}(z) - \mu w_{pg}(y) \end{cases} \tag{10}$$

was used. Otherwise, when available, transversality was used (see Section 4.2). Finally, the case of a field tangential to $\partial\mathcal{F}^o_{iv}$ was addressed in Section 4.3. All cases being addressed, uniqueness is proven.

5. A CASE STUDY

While appearing as a limit case of our result (see (H2)), square roots are often used for valve modelling. Uniqueness proof follows along the exact same lines except for the final points addressed in Section 4.3. Instructively, an alternative study leads to the conclusion. Let $\mathcal{X} = \mathcal{Y} \triangleq [5/4 - \sqrt{13/8}, 5/4]$, $\varepsilon = 0.1$ and $\mu = 2$. Let

$$\begin{aligned} w_{gc}(x,y) &\triangleq \sqrt{2-x} \\ \tau_{iv}(x,y) &\triangleq 13/8 - (x-5/4)^2 - (y-5/4)^2 \\ \tau_{pg}(x,y) &\triangleq y^{\frac{3}{2}} \end{aligned}$$

with $g_{iv} = g_{pg} \triangleq \sqrt{\max(0,\cdot)}$. Equilibrium points are unstable with positive real part complex conjugate poles. Hypothesis (H1), (H2), (H3), (H4) are verified. Let us check hypothesis (H5), (H6), (H7) and (H8) (with $y_{pg} = 0$)

$(H5)$ $\forall x \in \mathcal{X}$, $\dot{y}(x, \frac{5}{4}) = \sqrt{\frac{13}{8} - \left(x - \frac{5}{4}\right)^2} - 2\left(\frac{5}{4}\right)^{\frac{3}{4}} < 0$

$(H6)$ $\dot{x}(5/4, 0) = 0.1\sqrt{3}/2 - 1/4 < 0$

$(H7)$ $\forall x \in \mathcal{X}$, $\tau_{iv}\left(x, \frac{5}{4} - \sqrt{\frac{13}{8}}\right) = -(x - \frac{5}{4})^2 \leq 0$

$(H8)$ $\forall y \in \mathcal{Y}$, $\tau_{iv}\left(\frac{5}{4} - \sqrt{\frac{13}{8}}, y\right) = -(y - \frac{5}{4})^2 \leq 0$

These hypothesis are also verified. Yet, $\alpha = 1/2$, thus we substitute Section 4.3 with the following study. Around $X_0 = (y_0, 0)$ where the field is tangent to $\partial\mathcal{F}^o_{iv}$ we have, $y \sim y_0 + \dot{y}(0)t$ ($\dot{y}(0) < 0$). Equation (11) now yields

$$\dot{z} \sim -B\sqrt{z} + Kt$$

With $B = -1.93$ and $K = (E - C\mu 3/4 y_0^{-1/4})\dot{y}(0) = 0.503$. Using L'Hospital's rule we compute: $z(t) \sim at^2$, with $a = 1.38$. As $|e(s)| = \circ(s^2)$, Equation (12) becomes

$$|e(t)| \leq \circ(1) e^{b(t-s) + (2 - \frac{1-B}{2\sqrt{a}}) \ln \frac{t}{s}}$$

As $2 - (1-B)/(2\sqrt{a}) = 0.757$, letting s go to 0 implies that $e(t) = 0$. Uniqueness is proven. Figure 3 shows the construction of the positive invariant set and the limit cycle.

Acknowledgments We would like to thank D. E. Chang and L. Praly for their helpful insights regarding properties of this dynamical system.

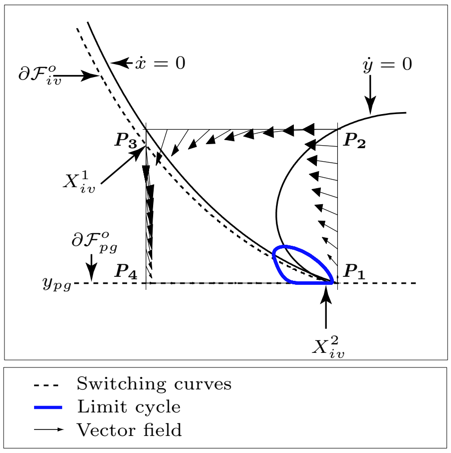


Fig. 3. Limit cycle and positive invariant set for the sample problem.

REFERENCES

Brauer, F. and J.A. Nohel (1989). *The qualitative theory of ordinary differential equations. An introduction.* Dover Publications, INC., New York.

Brown, K. E. (1973). *Gas Lift Theory and Practice.* Petroleum publishing CO., Tulsa, Oklahoma.

Eikrem, G.O., L. Imsland and B. Foss (2003). Stabilization of gas-lifted wells based on state estimation. In: *Preprints of the ADCHEM 2003, International Symposium on Advanced Control of Chemical Processes.*

Hiskens, I.A. (2001). Stability of limit cycles in hybrid systems. In: *Proc. of the 34th Hawaii International Conf. on System Sciences.*

Imsland, L.S. (2002). Topics in Nonlinear Control - Ouput Feedback Stabilization and Control of Positive Systems. PhD thesis. Norwegian University of Science and Technology, Department of Engineering and Cybernetics.

Jansen, B., M. Dalsmo K. Havre, L. Nøkleberg, V. Kritiansen and P. Lemetayer (1999). Automatic control of unstable gas-lifted wells. *SPE Annual technical Confererence and Exhibition.*

Lemetayer, P. and P. M. Miret (1991). Tool of the 90's to optimize gas-lift efficiency in the Gonelle field, Gabon. *SPE Annual technical Confererence and Exhibition.*

Miller, R.K. and A.N. Michel (1982). *Ordinary differential equations.* Academic Press.

Simić, S.N., K.H. Johansson, J. Lygeros and S. Sastry (2002). Hybrid limit cycles and hybrid Poincaré-Bendixon. In: *Proc. of the 15th IFAC World Congress.*

Standard Handbook of Petroleum and Natural Gas Engineering (1996). 6 ed. Gulf Professional Publishing.

REAL-TIME NONLINEAR INDIVIDUAL CYLINDER AIR FUEL RATIO OBSERVER ON A DIESEL ENGINE TEST BENCH

Jonathan Chauvin [*] **Philippe Moulin** [**]
Gilles Corde [**] **Nicolas Petit** [*] **Pierre Rouchon** [*]

[*] *Centre Automatique et Systèmes, École des Mines de Paris,*
60, bd St Michel, 75272 Paris, France
chauvin@cas.ensmp.fr
[**] *Institut Français du Pétrole, 1 et 4 Avenue de Bois Préau, 92852 Rueil Malmaison, France*

Abstract: We propose an estimator of the individual cylinder air fuel ratios in a turbocharged Diesel Engine using as only sensor the single air fuel ratio sensor placed downstream the turbine. The observer consists of a nonlinear filter designed on a physics-based time-varying model for the engine dynamics. Convergence is proven, using a Lyapounov function. Performance is studied through simulations and test bench experiments on a 4 cylinder engine.

Keywords: Engine Control, Observers, Air Fuel Ratio, Individual Cylinder Observer

1. INTRODUCTION

Performance and environmental requirements impose advanced control strategies for automotive applications. In this context, controlling the combustion represents a key challenge (Guzzella and Amstutz, 1998; Kiencke and Nielsen, 2000). Several tentative solutions are combustion torque control and estimation (see for example (Guezennec and Gyan, 1999), (Chauvin *et al.*, 2004*a*) and (Chauvin *et al.*, 2004*b*)), Air Fuel Ratio control and estimation (see (Grizzle *et al.*, 1991) and (Moulin *et al.*, 2004)),.... One important step is the control of the *individual* Air Fuel Ratio (AFR) which is a good representation of the torque produced by the engine. It results from various inputs such as injected quantities and timing, exhaust gas recirculation (EGR) rate. Classically, overall AFR can be directly controlled with the injection system (Grizzle *et al.*, 1991). In this approach, all cylinders share the same closed loop input signal based on the single AFR sensor. Ideally, all the cylinders would have the same AFR as they have the same injection setpoint. Unfortunately, due to inherent flaws of the injection system (pressure waves, mechanical tolerances, ...), the total fuel mass injected in each cylinder is very difficult to predict. Individual cylinder control has been addressed using individual cylinder AFR sensor in (Berggren and Perkovic, 1996). In practice, cost and reliability of multiple AFR sensor may prevent them from reaching commercial products lines.

For forthcoming HCCI engines (see (Kahrstedt *et al.*, 2003; Hultqvist *et al.*, 2001; Chiang and Stefanopoulou, 2004; Rausen *et al.*, 2004) for example) and regeneration filters, even slight unbalance between the cylinders can have dramatic consequences and induce important noise, possible stall and higher emissions. Individual cylinder control is needed. In this context individual cylinder AFR estimation can give crucial information to get the HCCI running better.

The contribution of this paper is the design of a real-time observer for the individual cylinder AFR using the reliable and available AFR sensor placed downstream the turbine as only measurement.

In previous works (see (Fantini and Burq, 2003) and (Carnevale and Hadji, 1998)), the methods used to reconstruct the AFR of each cylinder from the UEGO (Universal Exhaust Gas Oxygen) sensor measurement are based on the permutation dynamics at the TDC (Top-Dead Center) sample angle and a gain identification technique. We propose here a higher frequency approach (6^o sample angle instead of 90^o (TDC)). We design an observer on the balance model of the exhaust and use a nonlinear observer to solve the problem. A key problem in practice is the real-time implementation on an embedded system. Compared to Kalman observers, an interesting feature of our approach is its low computational cost which makes it tractable on a typical MPC555 based embedded card system, such as found on actual test benches.

We use a physics-based model underlying the role of periodic input flows (gas flows from the cylinders into the exhaust manifold). A nonlinear observer is designed and validated both experimentally (on a four cylinder turbocharged diesel test bench presented in (Moulin *et al.*, 2004)) and theoretically (convergence is proven).

The paper is organized as follows. In Section 2, we present the exhaust modelling and the individual cylinder AFR model. In Section 3, we propose a nonlinear individual cylinder AFR observer. Simulation and experimental results are presented in Section 4. Future directions are given in Section 5.

2. EXHAUST MODELLING

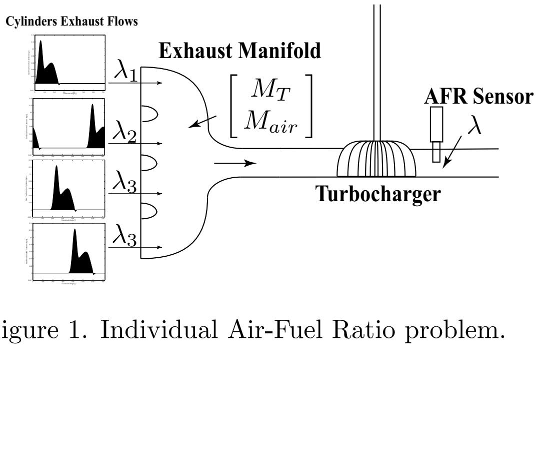


Figure 1. Individual Air-Fuel Ratio problem.

Figure 1 shows the flow sheet of the individual AFR from the cylinders outlet down to the turbine, where the sensor is located at. From the cylinders to the AFR sensor (located downstream the turbine) the gases travel through the exhaust pipes, the exhaust manifold and the turbocharger. All these components have an influence on the gas pressure, temperature, and composition in the exhaust manifold. In an "ideal system", the gases would move at a constant speed, without mixing. The global dynamics of the system are therefore nonlinear and depend on the operating conditions of the engine (engine speed, load, EGR). Our approach is to focus on a macroscopic balance model involving experimentally derived nonlinear functions.

2.1 Mass Balance in the exhaust manifold

Let M_T and M_{air} be the total mass of gas and the mass of fresh air in the exhaust manifold, respectively. Let λ_i be the *Air Fuel Ratio in cylinder i*. The measurements are

- P: the pressure in the exhaust manifold. This measure is not always available on a vehicle and can be given by the open loop estimate of the total mass.
- λ: the Air Fuel Ratio is equivalent to the following definition $\lambda \triangleq 1 - \frac{M_{air}}{M_T}$ with no EGR.

In the crank angle time α, on an operating point the mass balances write

$$\left\{ \begin{aligned} N_e \frac{dM_T}{d\alpha} &= \sum_{i=1}^{n_{cyl}} d_i(\alpha) - d_T(M_T) \\ N_e \frac{dM_{air}}{d\alpha} &= \sum_{i=1}^{n_{cyl}} (1-\lambda_i) d_i(\alpha) - \frac{M_{air}}{M_T} d_T(M_T) \\ N_e \frac{d\lambda_i}{d\alpha} &= 0 \;\; \forall i \in [\ 1,\ n_{cyl}\ ] \end{aligned} \right. \tag{1}$$

where

- N_e is the engine speed.
- d_i is the gas mass flow from cylinder i into the exhaust manifold.
- d_T is the gas mass flow through the turbine. d_T is a function of the total mass M_T and can be factorized as $d_T(M_T) = p(M_T)M_T$ with p a function of the total mass M_T.
- n_{cyl} the number of cylinders, 4 in our case.

The d_i functions are modelled through interpolation of a large number of available data. For sake of simplicity these 4π-periodic functions are approximated using a neural network, with three inputs : engine speed, intake pressure, and crank angle. For a given operating point (engine speed, load), the flow from the cylinders are equal up to 180^o shift. The mass flow through the turbine,

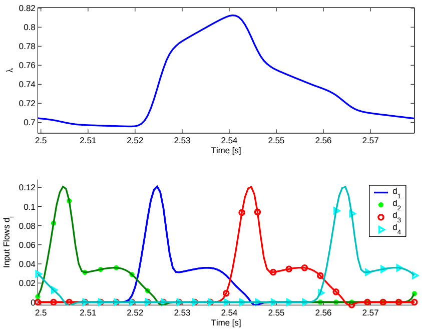


Figure 2. AFR oscillation over 1 engine cycle during a +20% offset on cylinder 1. Top: AFR. Bottom: Cylinders Input Flows.

usually given by a 2D look up table, is modelled as a flow through a restriction (Heywood, 1988) with a variable section depending on the pressure ratio and the turbocharger speed (see (Jensen *et al.*, 1991) and (Moraal and Kolmanovsky, 1999)). Moreover, corrections are made to take the turbine upstream pressure and temperature into account. The composition of the flow through the turbine is the same as in the exhaust manifold.

2.2 Reference Model

Let $x = \begin{bmatrix} M_T & M_{air} & \lambda_1 & \dots & \lambda_{n_{cyl}} \end{bmatrix}^T \in \mathbb{R}^{n_{cyl}+2}$ be the state and $y = \begin{bmatrix} P & \lambda \end{bmatrix}^T \in \mathbb{R}^2$ the measurements. System (1) rewrites as the following 4π-periodic (w.r.t. α) nonlinear model

$$\frac{dx}{d\alpha} = f(x, \alpha), \quad y = h(x) \tag{2}$$

Where

$$\begin{aligned} f_1(x,\alpha) &= \frac{1}{N_e}\left(\sum_{i=1}^{n_{cyl}} d_i(\alpha) - x_1 p(x_1)\right) \\ f_2(x,\alpha) &= \frac{1}{N_e}\left(\sum_{i=1}^{n_{cyl}} (1 - x_{i+2}) d_i(\alpha) - x_2 p(x_1)\right) \\ f_i(x,\alpha) &= 0 \ \ \forall i \in [\ 3,\ n_{cyl}+2\ ] \end{aligned}$$

and

$$h_1(x) = \gamma_T x_1, \quad h_2(x) = 1 - \frac{x_2}{x_1}$$

where γ_T is a positive constant arising from the ideal gas law (temperature is assumed constant around a given operating point).

3. NONLINEAR INDIVIDUAL CYLINDER AFR OBSERVER

3.1 Observer Definition

We consider the following time-varying observer

$$\begin{cases} \dfrac{d\hat{x}_1}{d\alpha} &= f_1(\dfrac{1}{\gamma_T}y_1, \alpha) + \dfrac{L_1}{N_e}(\dfrac{y_1}{\gamma_T} - \hat{x}_1) \\ \dfrac{d\hat{x}_2}{d\alpha} &= f_2(\dfrac{1}{\gamma_T}y_1, (1-y_2)\dfrac{1}{\gamma_T}y_1, \hat{x}_{2+i}, \alpha) \\ & \quad + \dfrac{L_2}{N_e}((1-y_2)\dfrac{1}{\gamma_T}y_1 - \hat{x}_2) \\ \dfrac{d\hat{x}_{2+i}}{d\alpha} &= -\dfrac{L_\lambda}{N_e} d_i(\alpha)\big((1-y_2)\dfrac{1}{\gamma_T}y_1 - \hat{x}_2\big) \end{cases} \tag{3}$$

where the last equation hold for all i in[1, n_{cyl}], and where $(L_1, L_2, L_\lambda) \in (\mathbb{R}^+)^3$. To prove convergence of the observer rate $\hat{x}$, described by System (3), to the state x of the reference System (2), we exhibit a Lyapounov function and use LaSalle's theorem to conclude to the convergence of the observer.
Let $\tilde{x} = x - \hat{x}$. The error dynamics write

$$\begin{cases} \dfrac{d\tilde{x}_1}{d\alpha} &= -\dfrac{1}{N_e} L_1 \tilde{x}_1 \\ \dfrac{d\tilde{x}_2}{d\alpha} &= -\dfrac{1}{N_e}\left(\displaystyle\sum_{i=1}^{n_{cyl}} \tilde{\lambda}_i d_i(\alpha) + L_2 \tilde{x}_2\right) \\ \dfrac{d\tilde{x}_{2+i}}{d\alpha} &= \dfrac{1}{N_e} L_\lambda d_i(\alpha) \tilde{x}_2, \quad \forall i \in [\ 1,\ n_{cyl}\ ] \end{cases} \tag{4}$$

3.2 Lyapounov function candidate

We consider the following Lyapounov function candidate

$$V(\tilde{x}) = \frac{N_e}{2}\left(\frac{1}{L_1}\tilde{x}_1^2 + \frac{1}{L_2}\tilde{x}_2^2 + \frac{1}{L_2 L_\lambda}\sum_{i=1}^{n_{cyl}} \tilde{x}_{2+i}^2\right) \tag{5}$$

On the one hand, $V(\tilde{x}) > 0$ for $\tilde{x} \in \mathbb{R}^{n_{cyl}+2} \setminus \{0\}$ and $V(0) = 0$. Then the following computation yield next lemma.

$$\begin{aligned} \frac{dV}{d\alpha}(\tilde{x}) &= -\tilde{x}_1^2 - \frac{1}{L_2}\sum_{i=1}^{n_{cyl}} \tilde{x}_{2+i} d_i(\alpha) \tilde{x}_2 - \tilde{x}_2^2 \\ & \quad + \frac{1}{L_2}\sum_{i=1}^{n_{cyl}} d_i(\alpha) \tilde{x}_2 \tilde{x}_{2+i} \\ &= -\tilde{x}_1^2 - \tilde{x}_2^2 \leq 0 \end{aligned}$$

Lemma 1. The function V defined by (5) is a Lyapounov function for the error-state System (4).

3.3 Application of LaSalle's theorem

Let $\Omega_r = \{\tilde{x}_f \in \mathbb{R}^{n_{cyl}+2} / V(\tilde{x}_f) < r\} \subset \mathbb{R}^{n_{cyl}+2}$. Ω_r is a compact set positively invariant with respect to the error dynamics because $\frac{dV}{d\alpha} \leq 0$. Therefore V is a continuously differentiable function such that $\frac{dV}{d\alpha}(\tilde{x}_f) \leq 0$ in Ω_r. Let I_f be the largest invariant set in $\{\tilde{x}_f \in \Omega_r / \frac{dV}{d\alpha}(\tilde{x}_f) = 0\}$. From LaSalle's theorem (see (Khalil, 1992) Theorem 4.4), every solution starting in Ω_r approaches I_f as $\alpha \to \infty$.

3.4 Characterization of the invariant set I_f

We first characterize $\{\tilde{x}_f \in \Omega_r / \frac{dV}{d\alpha}(\tilde{x}_f) = 0\}$ and then I_f.

$$x_0 \in \{\tilde{x}_f \in \Omega_r / \frac{dV}{d\alpha}(\tilde{x}_f) = 0\}$$
$$\Leftrightarrow -\tilde{x}_{1_f}^2 - \tilde{x}_{2_f}^2 = 0 \Leftrightarrow \begin{cases} \tilde{x}_{1_f} &= 0 \\ \tilde{x}_{2_f} &= 0 \end{cases}$$

Thus

$$\{\tilde{x}_f \in \Omega_r \ / \ \frac{dV}{d\alpha}(\tilde{x}_f) = 0\}$$
$$= \{ \left[0 \ 0 \ \tilde{\lambda}_{1,0} \ \ldots \ \tilde{\lambda}_{n_{cyl},0} \right]^T \in \mathbb{R}^{n_{cyl}+2} \}$$

From LaSalle's theorem, I_f is the largest invariant set in $\{\tilde{x}_f \in \Omega_r / \frac{dV}{d\alpha}(\tilde{x}_f) = 0\}$. I_f writes

$$I_f = \{ \left[0 \ 0 \ \tilde{\lambda}_{1,0} \ \ldots \ \tilde{\lambda}_{n_{cyl},0} \right]^T \in \mathbb{R}^{n_{cyl}+2} /$$
$$\forall \alpha \in [0, \ 4\pi] \ \sum_{i=1}^{n_{cyl}} \tilde{\lambda}_{i,0} d_i(\ \alpha) = 0\}$$

The functions family $\{d_i\}_{i=1 \ldots n_{cyl}}$ is a linearly independent family of the set $\mathcal{C}^0([0, \ 4\pi], \mathbb{R})$. Therefore the set I_f is reduced to $\{0\}$. The observation error is asymptotically stable and the following results hold.

Lemma 2. The largest set in

$$\Omega_r = \{\tilde{x}_f \in \mathbb{R}^{n_{cyl}+2} / V(\tilde{x}_f) < r\} \subset \mathbb{R}^{n_{cyl}+2}$$

invariant by the dynamics of the system (4) where the function V is defined in (5) is the null space.

Proposition 1. The observer defined in equation (3) converges toward the reference model (2).

4. SIMULATION AND EXPERIMENTAL RESULTS

4.1 Tests setup

The estimator described above is tested in simulation, on a high frequency engine model developed in AMESim (IMAGINE, 2004). This includes a complete combustion model, balance ODEs, thermal transfer laws, gas mixing laws,... On both simulation and experimental testbed, we apply an injection duration timing trajectory to introduce unbalance.It produces offsets in injection which lead to AFR disturbances. More precisely the injection steps have an effect on the average level of the measured AFR and introduce oscillations of the overall AFR signal as represented in Figure 2. These oscillations are the direct consequences of the individual AFR difference. During cylinder 1 exhaust phase, the AFR increases in the manifold, and then decreases while the other cylinders exhaust phases occur. The magnitude of the oscillations is related to the amount of the AFR difference between the cylinders and the gas mass in the manifold (and thus to its volume). The oscillation is then propagated to the turbine, and to the UEGO sensor, where it is filtered. This is the information that we exploit in the nonlinear observer (3).

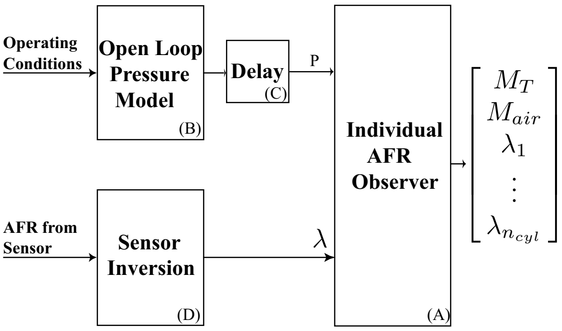


Figure 3. Observer Scheme as used in the test bench.

4.2 Simulation Results and Comments

Figure 4 presents results from simulation on the trajectory reference. The results are both quantitatively and qualitatively accurate. We reproduce well the evolution of the AFR. In practice the convergence is achieved within 4 engine cycles. The real bottleneck is the sensor noise and the quality of its model. In our results a simple first order model was used and seems relevant for this application. Yet, as rpm increases, better approximations would be useful. These results are encouraging for control purposes.

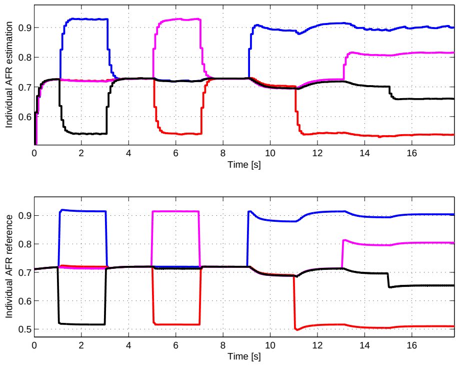


Figure 4. Trajectory on simulation at 1500 rpm and 800μs using the trajectory injection offset. Top: Reconstructed with (3). Bottom: Actual values from simulation

4.3 From simulation to experimentation

On the test bench we use the proposed observer following the scheme in Figure 3. Block (A) is the

Table 1. Experimental results.

N_e	IMEP	$100\,\|\lambda_*\|_\infty$	$100\,\|\lambda_*\|_{\text{mean}}$
1500	3	9.28	3.93
1500	6	3.61	1.47
1500	9	4.68	1.71
2000	6	7.29	2.27
2000	9	5.25	1.63
2500	3	7.16	3.72

implementation of observer (3). Several practical issues are considered.

Open Loop Pressure Model Depending on the vehicle, we may not have a exhaust pressure sensor. This sensor can be expected for HCCI vehicle only. In experimentation, we consider not having this sensor and give to the estimator an open loop value. This value is given by the open loop balance with the input flows (d_i) and output flow d_T as described previously in Section 2.1. This model is implemented in Block (B) in Figure 3.

Delay The lags due to the transport of the gas along the engine exhaust (pipes and volumes), and the dead time of the sensor are not represented by the model described above in System (2). However, all the delayed values used the same delay, the delays can be lumped in a single delay for the complete exhaust system, and the model can be inverted as it is. The global delay can be identified inline on the first offset. This estimated value is then kept as a constant for a given setpoint on the (engine speed, load) map. This estimation is implemented in Block (C) in Figure 3.

AFR Sensor Inversion The AFR sensor has a low-pass transfer function. Quantification noise is filtered by a very high frequency low-pass filter. The dynamics can be approximated by a first order filter. In order to robustly invert this dynamics, we apply an observer based on an adaptive Fourier Decomposition (Block (D) in Figure 3).

4.4 Experimental Results and Comments

We applied the same injection duration timing trajectories at the test bench. The test bench used for validation is a 4 cylinders DI engine with a Variable Geometry Turbocharger (VGT) (see (Moulin *et al.*, 2004)). We can see the results in Figures 5 and 6. These represent the nonlinear estimation of the individual cylinder AFR around two engine setpoints. The same parameters were kept from simulation to experimentation.

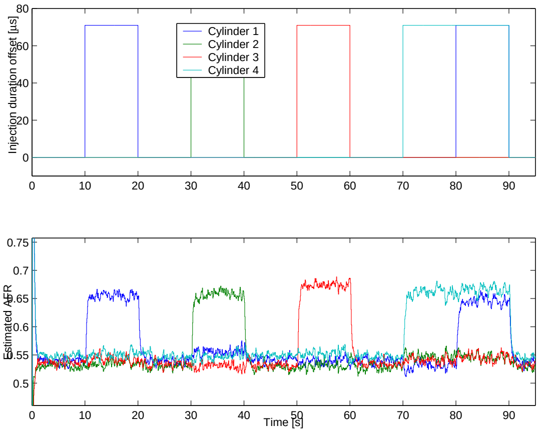


Figure 5. Test bench (2000 rpm and 6 bar). Top: Injection Duration Offsets. Bottom: Individual Estimated AFR

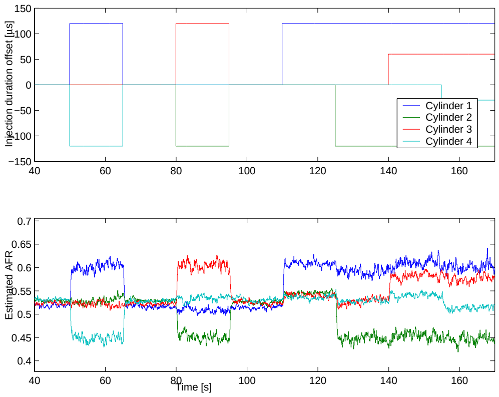


Figure 6. Test bench (1500 rpm and 6 bar). Top: Injection Duration Offsets. Bottom: Individual Estimated AFR

Further tests were conducted. Numerical values are reported in Table 1. They quantify the results of our observer for several setpoints (Engine Speed (rpm), IMEP (bar)). The reference AFR are not directly available but we can correlate them to the torque produced by each cylinder (reconstructed from the experimental individual in-cylinder pressure sensors). These correlated values, noted λ_{ref}, serve as a reference for the AFR in each cylinder. For this study we define two norms which represent (around steady states) the maximum and the mean value of the relative absolute errors around steady state.

- $\|\lambda_*\|_\infty \triangleq \max_{i,\alpha}\left|\frac{\hat{\lambda}_i-\lambda_{i,ref}}{\lambda_{i,ref}}\right|$
- $\|\lambda_*\|_{\text{mean}} \triangleq \text{mean}_{i,\alpha}\left|\frac{\hat{\lambda}_i-\lambda_{i,ref}}{\lambda_{i,ref}}\right|$

In all test bench cases, we were able to predict the individual cylinder AFR well. Further, we can easily detect the AFR unbalance and have a good estimation of the peaks of the AFR disturbances, the magnitude of the individual AFR offsets are satisfactory.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The work presented in this paper reports the development and implementation of a individual cylinder AFR estimator. It reconstructs the AFR of each cylinder from a measurement made by a *single* sensor located downstream the turbine. The availability of such an estimator giving reliable information can lead to improvements on diesel engines in terms of combustion control, noise, and pollutant emissions. This information will be used to control the unbalance between the cylinders. Indeed by controlling the individual injection quantity (which is the relevant control for such unbalance observation) a simple PI controller will lead to the balance of the individual AFR. In the context of combustion real-time control, this observer is a handy tool. It could be used in a closed loop controller of the fuel injectors. This is the long term goal of our work. Moreover this observer is easily transposed to various engine speeds and loads. Its dynamics are expressed in angular time scale and do not require any model for the combustion process. Theoretically, the gains do not need to be updated when the setpoint is changed. However we need to integrate the exhaust gas recirculation flow (EGR) for the HCCI purpose. We are currently investigating this point in an exhaustive test campaign on the test bench.

Acknowledgments: The authors would like to thank Jérôme Vauchel for his high contribution on the experimental part.

REFERENCES

Berggren, P. and A. Perkovic (1996). Cylinder individual lambda feedback control in an SI engine. Master's thesis. Linköpings Universitet.

Carnevale, C. and M. Hadji (1998). Cylinder to cylinder AFR control with an asymmetrical exhaust manifold in a GDI system. In: *Proc. of SAE Conference*. number 981064.

Chauvin, J., G. Corde, P. Moulin, M. Castagné, N. Petit and P. Rouchon (2004*a*). Real-time combustion torque estimation on a Diesel engine test bench using an adaptive Fourier basis decomposition. In: *IEEE Proc. of 2004 Conference in Decision and Control*.

Chauvin, J., G. Corde, P. Moulin, M. Castagné, N. Petit and P. Rouchon (2004*b*). Real-time combustion torque estimation on a Diesel engine test bench using time-varying Kalman filtering. In: *IEEE Proc. of 2004 Conference in Decision and Control*.

Chiang, C.J. and A.G. Stefanopoulou (2004). Steady-state multiplicity and stability of thermal equilibria in Homogeneous Charge Compression Ignition (HCCI) engines. In: *IEEE Proc. of 2004 Conference in Decision and Control*.

Fantini, J. and J.F. Burq (2003). Exhaust-intake manifold model for estimation of individual cylinder air fuel ratio and diagnostic of sensor-injector. In: *Proc. of SAE Conference*. number 2003-01-1059.

Grizzle, J., K. Dobbins and J. Cook (1991). Individual cylinder air-fuel ratio control with a single EGO sensor. *Proc. in the IEEE Transactions on Vehicular Technology* **40**(1), 357–381.

Guezennec, Y. and P. Gyan (1999). A novel approach to real-time estimation of the individual cylinder pressure for S.I. engine control. In: *Proc. of SAE Conference*.

Guzzella, L. and A. Amstutz (1998). Control of Diesel engines. *Proc. in the IEEE Control Systems Magazine* **18**, 53–71.

Heywood, J.B. (1988). *Internal Combustion Engine Fundamentals*. McGraw-Hill, Inc.

Hultqvist, A., U. Engdar, B. Johansson and J. Klingmann (2001). Reacting boundary layers in a homogeneous charge compression ignition (HCCI) engine. In: *Proc. of SAE Conference*. number 2001-01-1032.

IMAGINE (2004). *AMESim user manual, http://www.amesim.com/*.

Jensen, J.P., A.F. Kristensen, S.C. Sorensen, N. Houbak and E. Hendricks (1991). Mean value modeling of a small turbocharged diesel engine. In: *Proc. of SAE Conference*. number 910070.

Kahrstedt, J., K. Behnk, A. Sommer and T. Wormbs (2003). Combustion processes to meet future emission standards. In: *MTZ*. pp. 1417–1423.

Khalil, H.K. (1992). *Nonlinear Systems*. Prentice-Hall, Inc.

Kiencke, U. and L. Nielsen (2000). *Automotive Control Systems For Engine, Driveline, and Vehicle*. SAE Internationnal.

Moraal, P. and I. Kolmanovsky (1999). Turbocharger modeling for automotive control applications. In: *Proc. of SAE Conference*. number 1999-01-0908.

Moulin, P., G. Corde, J. Chauvin and M. Castagné (2004). Cylinder individual AFR estimation based on a physical model and using kalman filters. In: *Proc. of the FISITA World Automotive Congress 2004*. number F2004V279.

Rausen, D.J., A.G. Stefanopoulou, J-M. Kang, J.A. Eng and T-W. Kuo (2004). A mean-value model for control of Homogeneous Charge Compression Ignition (HCCI) engines. In: *IEEE Proc. of 2004 American Control Conference*.

43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

WeA09.3

Real-Time Combustion Torque Estimation on a Diesel Engine Test Bench Using Time-Varying Kalman Filtering

Jonathan Chauvin, Gilles Corde, Philippe Moulin, Michel Castagné, Nicolas Petit and Pierre Rouchon

Abstract— **We propose an estimator of the combustion torque on a Diesel Engine using as only sensor the easily available instantaneous crankshaft angle speed. The observer consists in a Kalman filter designed on a physics-based time-varying model for the engine dynamics. Convergence is proven, using results from the literature by establishing the uniform controllability and observability properties of this periodic system. A test bench and development environment is presented. Performance is studied through simulations and real test bench experiments.**

I. Introduction

Performance and environmental requirements impose advance control strategies for automotive applications. In this context, controlling the combustion represents a key challenge. A first step is the control of the combustion torque which characterizes the performance of the engine and is the result of various inputs such as injection quantity and timing, EGR (exhaust gas recirculation) rate ....

Ideally this torque could be measured using fast pressure sensors in each cylinder. Unfortunately their cost and reliability prevent them from reaching commercial products lines. As a consequence an interesting problem is the design of a real-time observer for the combustion torque using the reliable and available instantaneous crankshaft angle speed as only measurement.

Combustion torque determination by the measurement of the crankshaft angle speed has been addressed previously in the literature. Most of the proposed solutions have their foundations on a Direct or Indirect Fourier Transform of a black box model (see [10], [11], [7]). Other focus on a stochastic approach (see [8]) but the problem of real-time estimation is not addressed. Other approach such as mean indicated torque are also proposed (see [13] and [14] for example). Solving this first problem opens the door to more exciting applications such as misfiring detection ([1] and [15]) and combustion analysis.

For the design of a combustion torque observer, we use a physics-based model underlying the role of time-varying inertia. A Kalman filter observer is designed and validated both experimentally (on the presented test bench) and theoretically (proof of the convergence). It is computationally tractable on a typical XPC Target (or DSpace system) embedded system capable of handling a 500 μs sampling time.

The contribution is as follows. In the Section II, we explain the engine dynamics. We describe the combustion torque observer design in Section III. In Section IV, we describe the experimental setup. Simulation and experimental results are presented in Section V. Future directions are given in Section VI.

II. Crankshaft dynamics

A. Continuous time dynamics

In this part, we briefly describe the dynamics of the system stressing out the role of the combustion torque, T_{comb}, also referred as the indicated torque. Following [12], the torque balance on the crankshaft can be written

$$T_{comb} - T_{mass} - T^*_{load} = 0 \qquad (1)$$

where $T^*_{load} = T_{load} + T_{fric}$ is referred as "the extended load torque" and T_{load} and T_{fric} are known. The mass torque T_{mass} is the derivative of the kinetic energy E_{mass} of the moving masses in the engine as described in Figure 1.

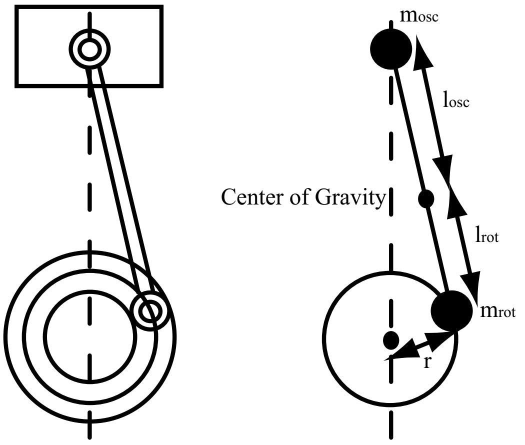


Fig. 1. Mass Model.

$$E_{mass} = \int_0^{2\pi} T_{mass} d\alpha = \frac{1}{2} J(\alpha)\dot{\alpha}^2$$

The mass torque T_{mass} can be expressed as

$$\begin{aligned}\frac{dE_{mass}}{dt} &= T_{mass}\dot{\alpha} \\ &= (J\ddot{\alpha} + \frac{1}{2}\frac{dJ}{d\alpha}\dot{\alpha}^2)\dot{\alpha}\end{aligned}$$

J. Chauvin (corresponding author) is a PhD Candidate in Mathematics and Control, Centre Automatique et Systèmes, École des Mines de Paris, 60, bd St Michel, 75272 Paris, France chauvin@cas.ensmp.fr

G. Corde, P. Moulin and M. Castagné are with the Department of Engine Control in Institut Français du Pétrole, 1 et 4 Avenue de Bois Préau, 92852 Rueil Malmaison, France

N. Petit and P. Rouchon are with the Centre Automatique et Systèmes, École des Mines de Paris, 60, bd St Michel, 75272 Paris, France with

$$\begin{cases} J(\alpha) & = & m_{rot}r^2 + m_{osc}\sum_{j=1}^{4}(\frac{ds_j}{d\alpha})^2 \\ f = \frac{1}{2}\frac{dJ}{d\alpha} & = & m_{osc}\sum_{j=1}^{4}\frac{ds_j}{d\alpha}\frac{d^2s_j}{d\alpha^2} \end{cases}$$

the computation of the various elements of J are described in [6] and are usually perfectly known for a particular engine. $J(\alpha)$ and $\frac{dJ}{d\alpha}(\alpha)$ are periodic functions in α over an engine cycle.

B. Discrete Time-varying Linear Approximation

The torque balance writes

$$J(\alpha)\ddot{\alpha} = T_{comb}(\alpha) - T^*_{load}(\alpha) - f(\alpha)\dot{\alpha}^2$$

We can reformulate this equation as

$$\dot{\alpha}\frac{d\dot{\alpha}}{d\alpha} = \frac{1}{J(\alpha)}(T_{comb}(\alpha) - T^*_{load}(\alpha) - f(\alpha)\dot{\alpha}^2) \quad (2)$$

Using a first order approximation on the right hand-side of the previous equation, we can break the dependence on time and on the crankshaft angle and only save a dependence on the square of the crankshaft angle speed.

$$\dot{\alpha}^2(n+1) - \dot{\alpha}^2(n) \approx \frac{2\Delta\alpha}{J(n)}(T_{comb}(n) - T^*_{load}(n) - f(n)\dot{\alpha}^2(n))$$

In practice an angular path $\Delta\alpha = 6^o$ is used. Using the square of the crankshaft angle speed $\dot{\alpha}^2$ as the first state variable x_1, we get the linear equation

$$x_1(n+1) = \left(1 - \frac{2\Delta\alpha}{J(n)}f(n)\right)x_1(n) + \frac{2\Delta\alpha}{J(n)}x_2(n) \quad (3)$$

where

$$\begin{cases} x_1(n) = \dot{\alpha}^2(n) \\ x_2(n) = T_{comb}(n) - T^*_{load}(n) \end{cases}$$

This formulation of the problem as a two dimensional linear time-varying system suggests that classical methods for combustion torque estimation (x_2) can be used.

C. Mass torque as a filter

The combustion torque generates the movement of the crankshaft. The oscillations of the combustion torque and of the load torque decrease when the engine is accelerating. This oscillation can be described by a low-pass $h(z)$ filter excited by a white noise u(z) as in [9].

$$x_2(z) = h(z)u(z) \quad (4)$$

In the following, $x_2(n)$, is a colored noise.

III. Combustion Torque estimation using Kalman filtering

As stated in Equation (3), the crankshaft is subject to torque excitations created by the combustion process in each cylinder (T_{comb}) which is a highly varying signal (due to combustion cycles and their imperfections). The resulting angular speed has a slowly varying component and a fast varying one resulting from the combustion process.

A colored white noise can be a good representation for the combustion torque. x_2 can be modelled in the z-transform domain as the product of a filter $h(z)$ and a white noise $u(z)$

$$x_2(z) = h(z)u(z)$$

where $h(z)$ is :

$$h(z) = \frac{b_0 + b_1z^{-1} + \cdots + b_pz^{-p}}{1 + a_1z^{-1} + \cdots + a_qz^{-q}} \quad (5)$$

This filter is chosen stable, the roots are $\{\lambda_i\}_{i\in\{1,\ldots,q\}}$.

A. Reference model

Gathering past values of x_2 over $[k-q+1,\ k]$, we obtain a time-varying linear system.

$$\begin{cases} x_{k+1} & = & A_kx_k + B_ku_k \\ y_k & = & C_kx_k + w_k \end{cases} \quad (6)$$

with the state

$$x_k = \begin{pmatrix} \dot{\alpha}^2(k) \\ T_{comb}(k) - T^*_{load}(k) \\ \ldots \\ T_{comb}(k-q+1) - T^*_{load}(k-q+1) \end{pmatrix} \in \mathbb{R}^{q+1}$$

The matrices A_k, B_k and C_k are

$$A_k = \begin{bmatrix} 1 - \frac{2\Delta\alpha}{J(k)}f(k) & v_k \\ 0 & M \end{bmatrix} \in \mathcal{M}_{q+1,q+1}(\mathbb{R}) \quad (7)$$

$$B_k = \begin{bmatrix} 0 & 0 & 0 \\ b_0 & \ldots & b_p \\ 0 & 0 & 0 \\ \ldots & \ldots & \ldots \\ 0 & 0 & 0 \end{bmatrix} \in \mathcal{M}_{q+1,p+1}(\mathbb{R}) \quad (8)$$

$$C_k = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} \in \mathcal{M}_{1,q+1}(\mathbb{R}) \quad (9)$$

with

$$v_k = \begin{bmatrix} \frac{2\Delta\alpha}{J(k)} & 0 & \ldots & 0 \end{bmatrix} \in \mathcal{M}_{1,q}(\mathbb{R}) \quad (10)$$

and

$$M = \begin{bmatrix} -a_1 & -a_2 & -a_3 & \ldots & -a_q \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \in \mathcal{M}_{q,q}(\mathbb{R})$$

Finally, u_k is a white noise. This system is $N = 120$-periodic (since the angle dynamics (2) is 4π periodic and the angle sample is $\pi/30$).

B. *Time-varying prediction algorithm*

We use a time-varying Kalman predictor for the combustion torque. For purpose we introduce the system

$$\hat{x}_{k+1/k} = A_k \hat{x}_{k/k-1} + L_k(y_k - C_k \hat{x}_{k/k-1}) \tag{11}$$

with the initial condition

$$x_{0/-1} = m_0$$

where L_k is the Kalman gain matrix

$$L_k = A_k P_k C_k^T (C_k P_k C_k^T + R_k)^{-1} \tag{12}$$

In this last expression, the covariance error $P_k = cov(x_k - \hat{x}_{k/k-1})$ is recursively computed through

$$\begin{array}{rcl} P_{k+1} & = & A_k P_k A_k^T + B_k Q_k B_k^T \\ & & -A_k P_k C_k^T (C_k P_k C_k^T + R_k)^{-1} C_k P_k A_k^T \end{array} \tag{13}$$

with $P_0 = cov(x_0)$. At last Q_k and R_k are matrices to be chosen.

C. *Convergence*

In the general time-varying case, there is no proof of the convergence of the Kalman observer algorithm. Nevertheless linear periodic systems have received a lot of attention for the last twenty years. The Discrete Periodic Riccati Equation (DPRE) properties are used to extend the Kalman filter to periodic systems. A key challenge is to prove the existence and uniqueness of a Symmetric Periodic Positive Solution (SPPS). In short, Bittanti et *al.* exposes sufficient conditions to prove convergence of the estimator. Theses properties are the detectability and stabilizability of the system. To check theses, the Gramian is a handy tool. Since B_* (i.e. the set of all B_k for $k \in \mathbb{N}$) and C_* are constant matrices, the criteria of detectability (resp. stabilizability) is equivalent to the observability (resp. controllability) criteria. The next subsection focuses on checking these last properties, through controllability and observability Gramians. We prove the positiveness of both Gramians and conclude using theorem 1.

Reference model properties

1) A_k's eigenvalues:
To check stability, we investigate A_k's eigenvalues. All the A_k matrices are block upper-triangular, so

$$eig(A_k) = \{1 - \frac{2\Delta\alpha}{J(k)} f(k) \ , \ \lambda_1 \ , \ \ldots \ , \ \lambda_p\}$$

Both J and $f(k) = \frac{1}{2}\frac{dJ}{d\alpha}(k)$ are periodic while $\frac{2}{J(k)} f(k) = \frac{d \log(J)}{d\alpha}(k)$ is periodic with a 0 mean value. The system is thus unstable when $f(k) > 0$ which occurs half of the time along the engine cycle.

2) Stability of A_*:
The properties of each A_k do not allow us to conclude stability of $A_* = \{A_k\}_{k\in\mathbb{N}}$ as a set. It is a common result that A_* is asymptotically stable if and only if the characteristic multipliers are included in the unitary circle (see [2]). To compute these multipliers we compute by induction the transition matrices
$\forall (k_1, k_2) \in \mathbb{N}^2 \ \ k_2 \geq k_1$

$$\Phi(k_2, k_1) = \left[\begin{array}{cc} \pi_{k_2,k_1} & \phi_{k_2,k_1} \\ 0 & M^{k_2-k_1} \end{array} \right]$$

with

$$\phi_{k_2,k_1} = \left\{ \begin{array}{cl} 0 & \text{if} \quad k_2 = k_1 \\ \sum_{j=k_1}^{k_2-1} (\pi_{k_2,j+1} v_j M^{j-k_1}) & \text{if} \quad k_2 > k_1 \end{array} \right. \tag{14}$$

and

$$\pi_{k_2,k_1} = \left\{ \begin{array}{cl} 1 & \text{if} \quad k_2 = k_1 \\ \prod_{i=k_1}^{k_2-1} (1 - \frac{2\Delta\alpha}{J(i)} f(i)) & \text{if} \quad k_2 > k_1 \end{array} \right.$$

We finally have

$$eig(\Phi(N+1, 1)) = \{\pi_{N+1,1} \ , \ \lambda_1^N \ , \ \ldots \ , \ \lambda_p^N\}$$

The analytical expression of $J(n)$ allows us to state the $\frac{N}{2}$-periodicity of $J(n)$ and $\frac{d}{d\alpha}(\frac{1}{J})(n)$. Note that this last expression is also symmetric with respect of the $n \mapsto -n$ mapping. Thus

$$\forall k \in \{1, \frac{N}{2}\} \quad \frac{2\Delta\alpha}{J(k)} f(k) + \frac{2\Delta\alpha}{J(N-k)} f(N-k) = 0$$

thus

$$\prod_{i=1}^{N} (1 - \frac{2\Delta\alpha}{J(i)} f(i)) = \prod_{i=1}^{\frac{N}{2}} (1 - (\frac{2\Delta\alpha}{J(i)} f(i))^2) < 1$$

finally $eig(\Phi(N+1, 1)) \subset \mathcal{D}_{0,1}$. Stability of the system is proven. The following result holds

Lemma 1: The system $x_k = A_k x_k + B_k u_k$, $y_k = C_k x_k + w_k$ where A_k, B_k and C_k are given by Equations (7), (8) and (9) is asymptotically stable.

3) Controllability:
To show the controllability of the system, we compute the controllability Gramian W_c over an interval $[k_0, k_0+k]$ and check its uniform positiveness over k. Since the system is N-periodic, we just have to check positiveness over $k \in [1, \ N]$.

$$W_c(k_f, k_0) = \sum_{i=k_0}^{k_f-1} \Phi_{k_f,i+1} B_i B_i^T \Phi_{k_f,i+1}^T$$

Let us look wether $W_c(k_0 + nN, k_0)$ is positive definite with $n = q+1$ the size of A. Let $V_c(k, i)$ denote the second column of $\Phi(k, i)$ as given in Equation (14). On the other hand

$$B_i^T \Phi(k_f, i+1)^T = \left[\begin{array}{c} b_0 V_c(k_f, i+1)^T \\ b_1 V_c(k_f, i+1)^T \\ \ldots \\ b_p V_c(k_f, i+1)^T \end{array} \right]$$

So

$$W_c(k_0+nN,k_0)>0 \Leftrightarrow \bigcap_{i=k_0+1}^{k_0+nN} Ker(V_c(k_0+nN,i)^T)=\{0\}$$

Let

$$\mathcal{V}_c(k_2)=\left[\begin{array}{ccc} V_c(k_2,k_2) & \dots & V_c(k_2,k_2-(q-1)) \end{array}\right]$$

Using $b=\left[\begin{array}{cccc}1 & 0 & \dots & 0\end{array}\right]^T \in \mathcal{M}_{q,1}(\mathbb{R})$ we note

$$\begin{aligned}\mathbf{M} &= \left[\begin{array}{ccccc} b & Mb & M^2b & \dots & M^{q-1}b \end{array}\right] \\ &= \left[\begin{array}{cccc} 1 & * & * & * \\ 0 & 1 & * & * \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{array}\right]\end{aligned}$$

And we realize that

$$\mathcal{V}_c(k_2)=\left[\begin{array}{c} * \quad * \\ \mathbf{M} \end{array}\right]$$

The preceding matrix has rank q as $\mathbf{M}$ has.

$$\bigcap_{i=k_0}^{k_0+nN-1} Ker(V_c(k_0+nN-1,i)) \subset Ker(\mathcal{V}_c(k_0+nN))=\{0\}$$

Controllability is thus proven and the following result holds

Lemma 2: The system $x_k = A_k x_k + B_k u_k$, $y_k = C_k x_k + w_k$ where A_k, B_k and C_k are given by Equations (7), (8) and (9) is controllable.

4) Observability:

We now compute the observability Gramian W_o over an interval $[k_0, k_0+k]$ and check its uniform positiveness over k. Again, since the system is periodic, we just have to check positiveness of W_o over $k \in [1,\ N]$. The observability Gramian over $[k_0, k_f]$ is defined by

$$W_o(k_f,k_0)=\sum_{i=k_0}^{k_f} \Phi_{k_f,i}^T C_i^T C_i \Phi_{k_f,i}$$

To check wether $W_o(k_0+nN,k_0)$ is positive definite, we pose

$$\begin{aligned}V_o(k,i) &= C_i\Phi_{k,i} \\ &= \left[\begin{array}{cc} \pi_{k_f,i} & \phi_{k_f,i} \end{array}\right]\end{aligned}$$

We have

$$W_o(k_0+nN,k_0)>0 \Leftrightarrow \bigcap_{i=k_0}^{k_0+nN} Ker(V_o(k_0+nN,i))=\{0\}$$

As before, we pose

$$\mathcal{V}_o(k_2)=\left[\begin{array}{c} V_o(k_2,k_2) \\ V_o(k_2,k_2-1) \\ \dots \\ V_o(k_2,k_2-(q-1)) \end{array}\right] \tag{15}$$

We note $L_1^{(j)}$ the first line of M^j. Due to the analytic expression of v_j as defined in (10) we notice that ϕ_{k_2,k_1} is a linear combination of the elements of $\{L_1^{(j)}\}_{j=0,\dots,k_2-k_1}$. This yields

$$rank(\left[\begin{array}{c} \phi_{k_2,k_2} \\ \phi_{k_2,k_2-1} \\ \dots \\ \phi_{k_2,k_2-(q-1)} \end{array}\right])=rank(\mathbf{L}) \tag{16}$$

with

$$\mathbf{L}=\left[\begin{array}{c} L_1^{(0)} \\ L_1^{(1)} \\ \dots \\ L_1^{(q-1)} \end{array}\right]$$

Yet $|det(\mathbf{L})|=|det(M)|^{q-1}$. So $\mathbf{L}$ is a full rank matrix and so is $\mathcal{V}_o(k_2)$.

$$\bigcap_{i=k_0}^{k_0+nN} Ker(V_o(k_0+nN,i)) \subset Ker(\mathcal{V}_o(k_0+nN))=\{0\}$$

Observability is proven and the following result holds

Lemma 3: The system $x_k = A_k x_k + B_k u_k$, $y_k = C_k x_k + w_k$ where A_k, B_k and C_k are given by Equations (7), (8) and (9) is observable.

5) Riccati equation for discrete-time periodic systems:

We now focus on the properties of the DPRE described by (13) adapting the results of Theorem 1. The weight matrices R_k and Q_k previously defined are supposed to be constant symmetric definite positive matrices. We have

$$R_k=\tilde{R}\tilde{R}^T \text{ and } Q_k=\tilde{Q}\tilde{Q}^T$$

where $\tilde{R}$ and $\tilde{Q}$ are symmetric definite positive matrices. Let

$$\hat{B}_k=B_k\tilde{Q}_k \text{ and } \hat{C}_k=\tilde{R}_k^{-1}C_k$$

Equation (13) becomes

$$\begin{aligned}P_{k+1} &= A_kP_kA_k^T+\hat{B}_k\hat{B}_k^T \\ &\quad -A_kP_k\hat{C}_k^T(\hat{C}_kP_k\hat{C}_k^T+I)^{-1}\hat{C}_kP_kA_k^T\end{aligned} \tag{17}$$

As a result, the previous simple change of coordinates yields Theorem 1 formulation.

6) Conclusion on time-varying Kalman filter convergence:

Whatever the choice of the filter h in Equation (5) which defines the combustion model, we proved that the system is stable while each matrix A_k is not. Moreover, we proved the controllability and the observability of the reference system. We finally get on Bittanti et *al*'s conditions (the observability (resp. controllability) condition is invariant by multiplication of C_* (resp. B_*) by a definite positive matrix) with more general weighting matrices R_k and Q_k as used in Equation (13). All these steps lead to the convergence of the observer.

Proposition 1: With R_k and Q_k constant symmetric definite positive matrices, the Kalman filter state defined in Equations (11,12,13) converges towards the reference model state (6) whatever the choice of the combustion model (5).

IV. Experimental Setup for Control Design

In this paper we deal with a 4-cylinder Diesel engine. For this work we have at hand a Diesel test bench. For simulation purposes this reference system is approximated using a Chmela combustion model [5] (nondimensional combustion model that relies on the concept of mixing controlled combustion avoiding the detailed description of the individual mixture formation and fuel oxidation process) coded in Simulink.

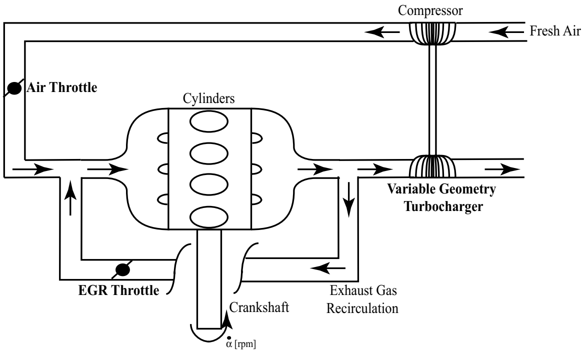


Fig. 2. Engine Scheme.

In our work, we try to restrict most of the design and tuning work to the simulation environment. This reduces the costly work on the engine test bench.

The same code is kept and implemented from the simulation environment to the embedded control system. This HiL (Hardware in the Loop) platform is easily transferred to a fast prototyping system. Typically 1 second of engine simulation is computed in 30 seconds on a 1 GHz Pentium based computer.

V. Simulation and Experimental Results

A. Filter choice

In the following, $x_2(n)$, is a colored noise.

$$h(z) = \frac{(1 - e^{-\delta\Delta\alpha})^2}{(z - e^{-\delta\Delta\alpha})^2} \tag{18}$$

In the discrete-time domain, the state variable $x_2(n)$ can be expressed as

$$\begin{aligned} x_2(n+2) - 2x_2(n+1)e^{-\delta\Delta\alpha} + x_2(n)e^{-2\delta\Delta\alpha} \\ = (1 - e^{-\delta\Delta\alpha})^2 u(n) \end{aligned}$$

B. Results and Comments

In the next figures, we have the comparison of the performance of the observer presented in [4], and the observer presented here. This last observer relies on a pole-placement for a extended state space model of the engine that assumes $\dot{x}_2 = 0$. Though giving qualitatively interesting results it suffers from a lag and a lack of accuracy. T_{mass} is estimated through our observer, then T_{comb} is computed by adding T_{mass} and T^*_{load} according to (1).

1) Simulation results:
We present a simulation corresponding to the following set point:

- Engine Speed : 1000 rpm
- BMEP (Brake Mean Effective Pressure) : 5 bar

To simulate the unbalance, we introduce offsets in the mass injected in each cylinder.

- Cylinder 1: 10% of the reference mass
- Cylinder 2: 0% of the reference mass
- Cylinder 3: 0% of the reference mass
- Cylinder 4: -20% of the reference mass

In Figure 4 the set point is a low engine speed and a low load. This point is very interesting because it represents where the driver feels internal loads and vibrations most. Correcting the unbalance at this points increases the driver's comfort.

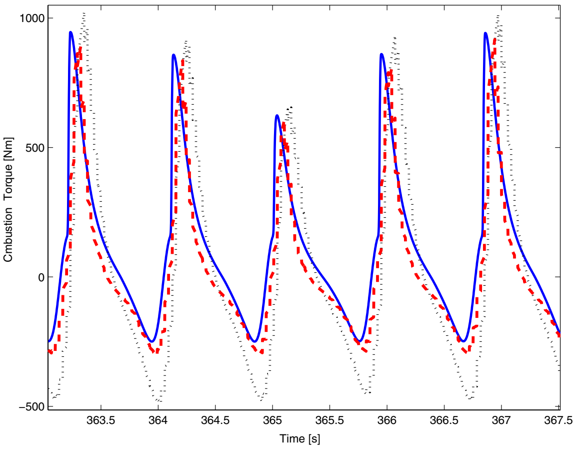


Fig. 4. Combustion torque (1000 rpm, 5 bar). bold (blue) : reference combustion torque, dashed (red) : combustion torque estimated by the time-varying filter, dotted (black) : combustion torque estimated by pole placement as in [4]. Notice the good match between the bold and the dashed signals

2) Experimental Results:
Figures 5 and 6 display the result of the estimator on experimental data. We reconstruct the combustion torque from the bench with the in-cylinder pressure and we test the observer on the flywheel velocity measurement. The set point is different from the simulation one to check robustness.

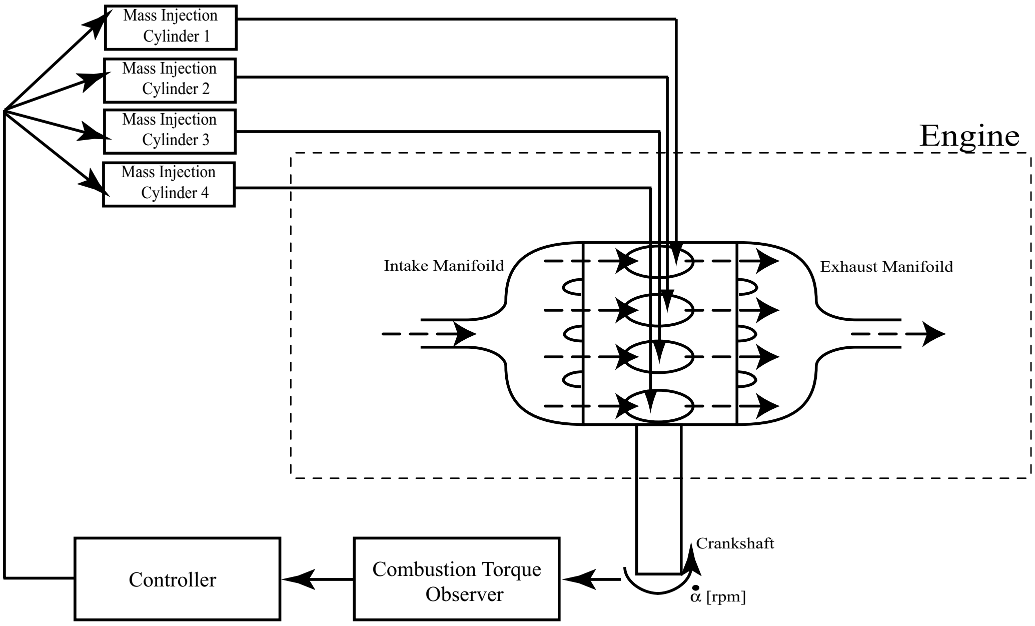

Mass Injection Cylinder 1
Mass Injection Cylinder 2
Mass Injection Cylinder 3
Mass Injection Cylinder 4
Engine
Intake Manifoild
Exhaust Manifoild
Crankshaft
$\dot{\alpha}$ [rpm]
Controller
Combustion Torque Observer

Fig. 3. Global Scheme.

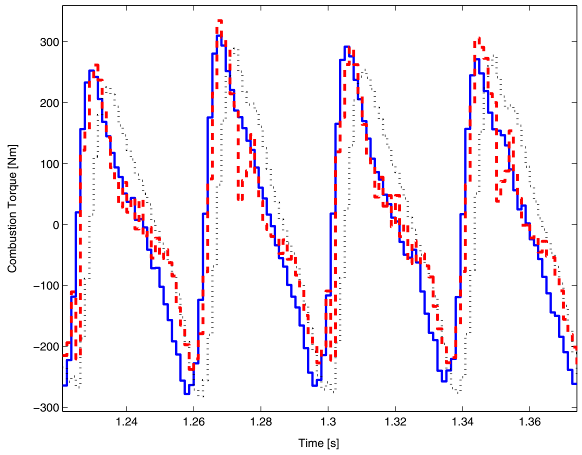


Fig. 5. Combustion torque on the test bench (800 rpm, 2 bar). bold (blue) : reference combustion torque, dashed (red) : combustion torque estimated by the time-varying filter, dotted (black) : combustion torque estimated by pole placement as in [4]. Notice the good match between the bold and the dashed signals

3) Comments:
Today these results are very satisfactory. An exhaustive testing campaign is underway to evaluate the Kalman filter design under various set points (engine speed and load). The predictor gives better results than the one presented in [4]. In both simulation and test bench cases, we are able to predict the combustion torque dynamics well. Further, we can easily detect the torque unbalance and have a good estimation of the peaks of the combustion torque.

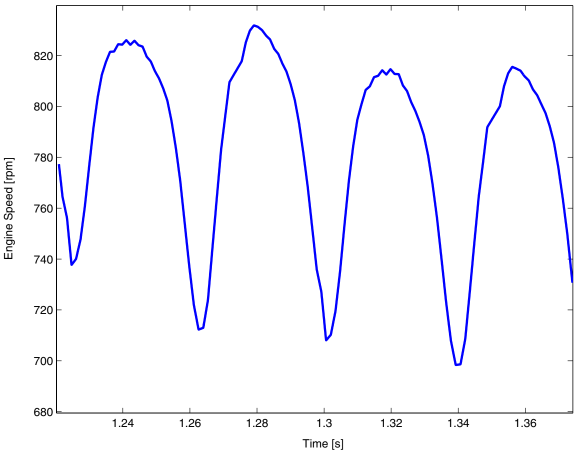


Fig. 6. Engine Speed $[rpm]$ on the test bench used as input of our Kalman filter

VI. Conclusion and Future Directions

The results of the presented time-varying observer are good. As is, a drawback of our approach is that extensive computations have to be done inline (namely matrix Equations (11), (12) and (13)). Nevertheless, we know that the covariance matrix P_k converges to a periodic solution (see Theorem 3). Moreover the $\{P_k\}_{k\in\mathbb{N}}$ matrices converge towards a periodic solution $\{\bar{P}_k\}_{k=1..N}$. These asymptotic solutions can be computed off-line and used as a gain-scheduling observer.

A numerical study (see results in Table I performed on a Matlab environment with a 1.7 GHz Pentium M (compiled code)) shows the computational effort required

order	2	4	6	8
CPU-time (TV)	0.286	1.03	2.70	5.33
CPU-time (APS)	0.0252	0.0250	0.0336	0.0280

TABLE I

CPU-TIMES ARE GIVEN IN MS FOR A SINGLE FILTER UPDATE. TV: TIME VARYING EXACT RICCATI SOLUTION. APS : ASYMPTOTIC PERIODIC SOLUTION

for various order filters modelling combustion (h filter in Equation (5)). It appears that the preceding substitution of the actual solutions with their asymptotic periodic values has a significant impact on the CPU load, while providing similarly good results.

Finally, we believe that a Kalman filter is a good tool to solve the combustion torque estimation problem for Diesel engines. Its computational demand and efficiency are well balanced. We plan to report further test bench results when an exhaustive test campaign is performed, including EGR 4-cylinders and HCCI combustion mode engines. Note also that tests on a 6-cylinders are scheduled. In this last problem, we have to focus on the overlapping phenomenon of the cylinders torques, that is not present in the 4-cylinders setup.

APPENDIX

Three main theorems are exposed in [3]. They allow to conclude on the convergence of the Kalman predictor in the linear periodic case.

Theorem 1 (Bittanti et al. *[3].):* [Predictor Convergence]
With the above notations, consider the optimal Kalman gain

$$L_k = A_k P_k \hat{C}_k^T (\hat{C}_k P_k \hat{C}_k^T + I)^{-1}$$

associated with *any* semi-definite solution P of (17). If $(A_*, \hat{B}_*)$ is stabilizable and $(A_*, \hat{C}_*)$ detectable, then the corresponding closed-loop matrix $\hat{A}_* = A_* - L_* \hat{C}_*$ is exponentially stable

Theorem 2 (Bittanti et al. *[3].):* [Existence and Uniqueness of a SPPS]
There exists a unique SPPS solution $\bar{P}_*$ of the DPRE and the corresponding closed-loop matrix $\hat{A}_* = A_* - L_* \hat{C}_*$ is asymptotically stable iff $(A_*, \hat{B}_*)$ is detectable and $(A_*, \hat{C}_*)$ reachable.

Theorem 3 (Bittanti et al. *[3].):* [Convergence toward SPPS]
Suppose that $(A_*, \hat{B}_*)$ is stabilizable and $(A_*, \hat{C}_*)$ detectable. Then every symmetric and positive semi-definite solution of the DPRE converges to the unique SPPS solution.

REFERENCES

[1] J. Ball, J. Bowe, C. Stone, and P. McFadden, "Torque estimation and misfire detection using block angular acceleration," in *Proc. of SAE Conference*, 2000.

[2] S. Bittanti, *Time Series and Linear Systems*. Springer-Verlag, 1986.

[3] S. Bittanti, P. Colaneri, and G. De Nicolao, "The difference periodic riccati equation for the periodic prediction problem," *Proc. in the IEEE Transactions on Automatic Control*, vol. 33, no. 8, Aug. 1988.

[4] J. Chauvin, G. Corde, P. Moulin, M. Castagné, N. Petit, and P. Rouchon, "Observer design for torque balancing on a di engine," in *Proc. of SAE Conference*, 2004.

[5] F. Chmela and G. Orthaber, "Rate of heat release prediction for direct injection diesel engines based on purely mixing controlled combustion," in *Proc. of SAE Conference*, no. 1999-01-0186, 1999.

[6] H. Fehrenbach, "Model-based combustion pressure computation through crankshaft angular acceleration analysis," *Proceedings of* 22^{nd} *International Symposium on Automotive Technology*, vol. I, 1990.

[7] S. Ginoux and J. Champoussin, "Engine torque determination by crankangle measurements: State of art, future prospects," in *Proc. of SAE Conference*, no. 970532, 1997.

[8] P. Gyan, S. Ginoux, J. Champoussin, and Y. Guezennec, "Crankangle based torque estimation: Mechanistic/stochastic," in *Proc. of SAE Conference*, 2000.

[9] M. Henn, "On-board-diagnose der verbrennung von ottomotoren," Ph.D. dissertation, Universität Karlsruhe, 1995.

[10] L. Jianqiu, Y. Minggao, Z. Ming, and L. Xihao, "Advanced torque estimation and control algorithm of diesel engines," in *Proc. of SAE Conference*, 2002.

[11] ——, "Individual cylinder control of diesel engines," in *Proc. of SAE Conference*, no. 2002-01-0199, 2002.

[12] U. Kiencke and L. Nielsen, *Automotive Control Systems For Engine, Driveline, and Vehicle*, Springer, Ed. SAE Internationnal, 2000.

[13] G. Rizzoni, "Estimate of indicated torque from crankshaft speed fluctuations: A model for the dynamics of the IC engine," vol. 38, pp. 169–179, 1989.

[14] G. Rizzoni and F. Connolly, "Estimate of IC engine torque from measurment of crankshaft angular position," in *Proc. of SAE Conference*, 1993.

[15] J. Williams, "An overview of misfiring cylinder engine diagnostic techniques based on crankshaft angular velocity measurements," in *Proc. of SAE Conference*, 1996.

PI CONTROLLERS PERFORMANCES FOR A PROCESS MODEL WITH VARYING DELAY

J. Barraud*, Y. Creff†, N. Petit‡

*Institut Français du Pétrole, École des Mines de Paris, France, julien.barraud@ifp.fr
†Institut Français du Pétrole, France, yann.creff@ifp.fr
‡École des Mines de Paris, France, nicolas.petit@ensmp.fr

Keywords: Process Control; Delay systems; Varying delay; PI controllers; Smith Predictor.

Abstract

Varying delay systems represent a serious challenge in many facets of process control. A frequent issue that arises in practice is introduced by transportation delays in fixed lengths pipes at speed which varies with setpoints. Many classic control techniques can be used to deal with constant delays systems but they do not specifically address this structural delay variability. In this paper we present a process model (Diesel Hydrodesulfurization) that features this delay variability and explore robustness properties of a wide panel of PI controllers. A conclusion is that the recent method proposed by Tavakoli and Fleming compares favorably with all others, including Smith predictors, when the delay variation is not known.

1 Introduction

In spite of all of the advances in process control over the 50 last years, the PI controller is still the most commonly encountered controller in the process industry. Though PI controllers can address delays in the systems dynamics, one of the serious practical limitations of this SISO controller is reached when dealing with time-varying delays. This situation can be problematic when dealing with transportation delays in fixed lengths pipes at speed which varies with setpoints. Indeed, these systems are ubiquitous in refineries, blending networks, and other systems that imply not negligible transport phenomena.

In a first attempt to solve this problem we explore the robustness properties of a wide panel of PI controllers including the newly proposed controller by Tavakoli and Fleming [7].

After briefly presenting the tuning methods for the PI controllers under consideration (and their key properties), we compare the obtained performances on a simplified hydrodesulfurization process model we use as test case.

2 PI controllers tuning rules

We denote the process model and controller transfer functions:

$$G(s) = \frac{Ke^{-\delta s}}{\tau s + 1}\ , \quad G_c(s) = K_c\left(1 + \frac{1}{sT_i}\right) \tag{1}$$

Tavakoli-Fleming tuning rule (TF) In [7] the authors proposed an optimal method based on a dimensional analysis and numerical optimisation techniques, for the tuning of the PI controllers for first order plus dead time systems (FOPDT). This dimensional analysis leads to relations:

$$KK_c = \ g_1\left(\frac{\delta}{\tau}\right), \quad \frac{T_i}{\delta} = \ g_2\left(\frac{\delta}{\tau}\right) \tag{2}$$

Functions g_1 and g_2 in (2) are determined for a step change in the setpoint so that the integral of the absolute error is minimized. To ensure closed loop robustness, two constraints guarantee a minimum gain margin of 6 dB and a minimum phase margin of 60°. Then genetic algorithms are used to find the best values for each $\frac{\delta}{\tau}$. Eventually functions g_1 and g_2 are determined using curve-fitting techniques:

$$\begin{aligned} KK_c &= 0.4849\,\frac{\tau}{\delta} + 0.3047 \\ \frac{T_i}{\tau} &= 0.4262\,\frac{\delta}{\tau} + 0.9581 \end{aligned} \tag{3}$$

Frequency-response method by Ziegler and Nichols (ZN) This design is based on the knowledge of the *ultimate gain* K_u and *ultimate period* T_u, two parameters that characterize the process dynamics [9]. K_u et T_u can be determined by a relay feedback as shown in [1]. Ziegler and Nichols then studied on a simple real process with a proportional controller, both the effect of disturbance and the effect of load change. Their conclusion was that a good compromise between large offset and large amplitude decay ratio was to choose the tuning giving an amplitude decay ratio of 0.25. An experience of load change is used again to find the best

response with a PI controller where the gain controller is $K_c = 0.45K_u$. The best response was given by an integral time $T_i = T_u/1.2$.
This method gives good results when the dead-time is short. When there is a large dead-time, the closed loop keeps robust but parameters of the controllers are de-tuned, the response is then very loose.

Cohen and Coon tuning formula (CC) Cohen and Coon presented in [2] a method to determine the adjustable parameters for a desired degree of stability.
The tuning is obtained with a theoretical study of a FOPDT system with a dimensionless equation. Harmonics in response after a Heaviside step are neglected and the amplitude ratio of the fundamental is set to 0.25. The integral time is determined with the objective of a 0.25 amplitude ratio and a compromise between a minimum control area and a maximum stability.
The Cohen-Coon method has small gain margin and phase margin when the process dead-time is short. This problem decreases when the dead-time of the process increases, this is why the (CC) tuning design is often used with processes that presents a large dead-time.

Refinements of the Ziegler-Nichols tuning formula (RZN) The design was proposed by Hang, Åstrom and Ho in 1991 [3]. Their tuning formula comes from a dimensional analysis where the dimensionless variables used are the scaled process gain $\kappa = KK_u$ and the scaled dead-time $\Delta = \frac{\delta}{\tau}$. A step response with 10% overshoot and 3% undershoot is required and defines the tuning rule.

Smith predictor (Smith) In 1957, Smith presented a control scheme for single-input single-output systems, which has the potential of improving the control of loops with dead-time (see [5] for example). It is known that Smith predictor gives good results when the model is correctly identified.
The Smith predictor can be seen as four blocks: the internal controller, the process, the process model and the process model without delay. The internal controller can be a PI controller. An open loop control is first obtained, based upon an undelayed prediction, the controller being tuned from the model without delay. Feedback action is provided through the (possibly filtered) difference between the prediction (including the delay) and the real measurement, that is added to the setpoint.

3 Process model and varying delay

Diesel Hydrodesulfurization Hydrodesulfurization is a process met in all refineries for various fluids. Here, we are looking at the desulfurization of an intermediate cut that enters the composition of diesel fuels.
For a real process, the feed to be desulfurized is mixed with a gas (essentially hydrogen). This mixture is preheated against the reactor outlet, then heated in a furnace, and is processed through the reactor. Downstream, the mixture is cooled and flashed. The gas phase is treated and then partially recycled: combined with an hydrogen make-up, it constitutes the gas to be mixed with the feedstock. The liquid phase is splitted, then cooled before being sent to the diesel pool for blending.
The operating plan we are using is the following: sulfur in the desulfurized product must be controlled at 50 ppm weight. The feed flowrate (straight run diesel, about 300 ppm weight sulfur) is equal to 200 t/h. The feed flowrate and composition change. The reactor inlet temperature is used to compensate for these disturbances. From a control point of view, the output is the sulfur concentration of the desulfurized product, the input is the reactor inlet temperature.
Some simplifying assumptions are made

- The reactor inlet temperature can be given arbitrary values instantaneously. This is not a very strong assumption: for real processes, this temperature is easily and quickly controlled by a regulatory controller acting upon the fuel flowrate (fuel to be burnt in the furnace).
- Light components are instantaneously and totally removed from the liquid in the separator located downstream the reactor. No heavy component is withdrawn in the vapor.
- The splitter is seen as a simple mixing drum.
- The ratio between feed and gas (recycle+make up) is kept constant.
- The composition of the gas mixed with the liquid feed is constant. Otherwise stated, we do not consider the variations of the hydrogen fraction in the gas, that are due to the recycle.

These assumptions allow us to limit the usage of energy balances to the reactor part. They do not oversimplify the problem, so that the conclusions we give on a model are valid for a real process. Releasing the last two assumptions would not lead to qualitatively different results. The simplified model keeps the two main characteristics we wanted to isolate for the tests: besides nonlinearities providing a variable gain, transportation through piping gives us a variable delay. Figure 1 shows the behavior of the outlet reactor and the outlet drum in open-loop when the feed flowrate varies. We denote, especially that the delay is varying from 15 min to 25 min.

Reaction We present a simplified diagram of an hydrodesulfurization unit on Figure 2.
The reaction is of the form

$$2A + B \rightarrow 2C + 2D \tag{4}$$

with $A = RSH$, $B = H_2$, $C = R$ and $D = H_2S$.

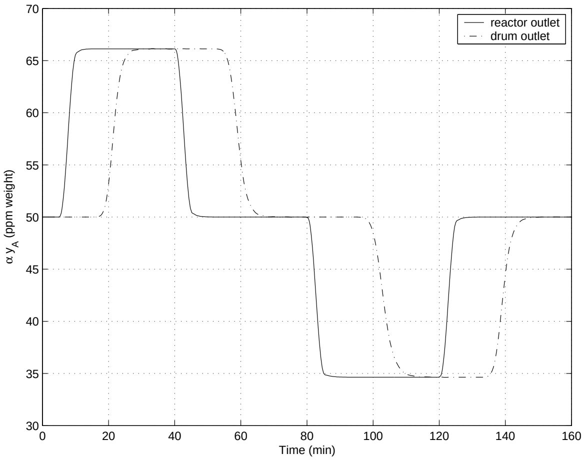


Figure 1: Weight fraction of RSH. Open-loop test.

Balance equations The model of the reactor is a plug-flow model with diffusion of energy and matter. We assume that the pressure profile inside the reactor is constant. The state of the model is only the molar fraction of the two reactants and the temperature inside the reactor. Molar fractions and energy balances are given by

$$\begin{aligned}
\frac{\partial x_A}{\partial t} &= v_{mol}\left(-\frac{F}{\Omega}\frac{\partial x_A}{\partial z} + r(.)\,(2 - x_A)\right) + D\frac{\partial^2 x_A}{\partial z^2} \\
\frac{\partial x_B}{\partial t} &= v_{mol}\left(-\frac{F}{\Omega}\frac{\partial x_B}{\partial z} + r(.)\,(1 - x_B)\right) + D\frac{\partial^2 x_B}{\partial z^2} \\
\tau_T\frac{\partial T}{\partial t} &= -\frac{FC_P}{\Omega}\frac{\partial T}{\partial z} + \Delta H\, r(.) + D_T\tau_T\frac{\partial^2 T}{\partial z^2}
\end{aligned} \tag{5}$$

where $T(0,t)$ is the control and $x_A(0,t)$ and $x_B(0,t)$ are constants (that can be used as disturbances). The term τ_T stands for a ρC_p-like term taking into account the fluid, the catalyst and the metal of the reactor. We assume that the separation downstream the reactor is perfect and modelled with the algebraic equations

$$y_j = \frac{x_j}{x_A + x_C} \quad \forall\, j \in \{A, C\} \tag{6}$$

We assume further that piping between the outlet of the separator and the inlet of the drum generates a 15 minutes delay when the feed flowrate is constant at the reference value and the mass fraction of A is stabilized at 50 ppm weight. The model is a transport equation:

$$\frac{\partial y_A}{\partial t} = -\frac{F^{drum} v^t_{mol}}{\Omega^P}\frac{\partial y_A}{\partial z} \text{ (from separator to drum)} \tag{7}$$

As there is no reaction in the drum, the model we propose is a simple mixer:

$$\frac{dy_A}{dt} = -\frac{F^{drum}}{N^{drum}}\left(y_A^{in} - y_A\right) \text{ (in the drum)} \tag{8}$$

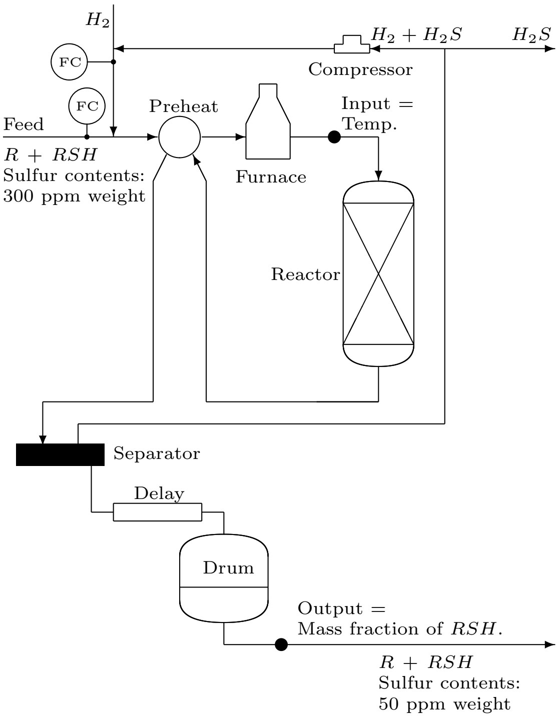


Figure 2: Simplified diagram of an hydrodesulfurization unit.

The kinetics of the reaction is denoted

$$r(.) = r\,(T, x_A, x_B) = k\exp\left(-\frac{E_a}{RT}\right)x_A\, x_B \tag{9}$$

To simulate this process model, we use a classical 1D-discretization scheme for equations (5) and (7) with 30 elements for the reactor.

4 Control model and simulations

We choose to identify the process as a first order plus dead-time system using **ISIAC**, the identification software of Institut Français du Pétrole [8]. The control model thus obtained is valid around the operating point:

$$\begin{aligned}
y_r &= \alpha\, 50 \text{ ppm molar} \\
u_r &= 623 \text{ K}
\end{aligned}$$

where α is a constant used to convert weight fractions to molar fractions. The linear input-output model is noted:

$$\dot{y}(t) = -\frac{1}{\tau}(y(t) - y_r) + \frac{K}{\tau}(u(t-\delta) - u_r) \tag{10}$$

where **ISIAC** identification gives:

$$\begin{aligned}
K &= -2.17 \text{ ppm.K}^{-1} \\
\tau &= 2.5 \text{ min} \\
\delta &= 15.7 \text{ min}
\end{aligned}$$

Symb.	Quantity	Unit
D	Diffusion coef. for matter	$m^2.s^{-1}$
D_T	Diffusion coef. for temp.	$m^2.s^{-1}$
E_a	Activation energy	$J.mol^{-1}$
F	Molar flow at z	$mol.min^{-1}$
F^{drum}	Molar flow inside the drum	$mol.min^{-1}$
k	Rate constant	$mol.m^3.s^{-1}$
K	Static gain	$ppm.K^{-1}$
K_c	Controller gain	$K.ppm^{-1}$
R	Gas constant	$J.K^{-1}.mol^{-1}$
t	Time	min
T	Temperature (Temp.)	K
T_i	Integral time	min
v_{mol}	Molar volume in the reactor	$m^3.mol^{-1}$
v^t_{mol}	Molar volume in the pipe	$m^3.mol^{-1}$
x_A	Molar fraction of A	
x_B	Molar fraction of B	
y_A	Molar fraction of A after sep.	
y_B	Molar fraction of B after sep.	
z	Length unit	m
α	ppm weight → molar	
δ	Delay	min
ΔH	Reaction enthalpy	$J.mol^{-1}$
Ω	Reactor's section	m^2
Ω^P	Pipe section	m^{-1}
τ	Time constant	min
τ_T	Pseudo time constant	$J.K^{-1}.m^{-3}$

Table 1: Nomenclature.

The limit gain and limit period are obtained with relay controller on the process:

$$\begin{aligned} K_u &= -0.594 \, \text{K.ppm}^{-1} \\ T_u &= 38 \, \text{min} \end{aligned}$$

Robustness with delay changes The varying delay is due to the varying feed flowrate, this leads us to test robustness by introducing changes in the flowrate F at the inlet of the reactor. F_{ref} is the feed flowrate which has been used for the model identification. The simulation involves five steps:
Step 1: when $t \in [0, 5]$, $F = F_{ref}$. **Step** 2: when $t \in [10, 145]$, $F = 1.2F_{ref}$. **Step** 3: when $t \in [150, 295]$, $F = F_{ref}$. **Step** 4: when $t \in [300, 445]$, $F = 0.8F_{ref}$. **Step** 5: when $t \in [450, 600]$, $F = F_{ref}$.
Figure 3 shows the sulfur mass fraction at the drum outlet. During **Step 2**, as the flowrate is more important, the dead-time decreases. All the controllers make the output converge towards the reference. ZN tunings gives the worst result. Good responses can be achieved by three different PI controllers. The two first ones, respectively TF and RZN lead to similar responses while the CC method, although different, converges as fast as the later ones. The Smith predictor response is faster than the PI responses. Dead-time identification errors create small oscillations on the output. The magnitude of oscillations increases with delay identification error, if this error becomes too large, the Smith predictor destabilizes the output. When the delay identification error is known to be large, Smith can be used with de-tuned controllers and with an important filter time constant. The response thus obtained is worse than the response given by the best PI controller.
Step 3 emphasizes the superiority of the Smith predictor when the delay is accurately identified. Indeed, the Smith predictor brings the output at setpoint very quickly. After it, the three best PI controllers are the same than those in step 2.
During **Step 4**, as the flowrate is less important, the dead-time increases. All the controllers make the output converge towards the reference and the three best PI controllers are the same than those in step 2. The Smith predictor response keeps stable but the response oscillates around the setpoint.
During **Step 5**, the feed flowrate is equal to the reference feed flowrate, conclusions are the same than in the step 3. The Smith predictor gives better behavior than PI controllers. The three best PI controllers are the same than those in step 2.

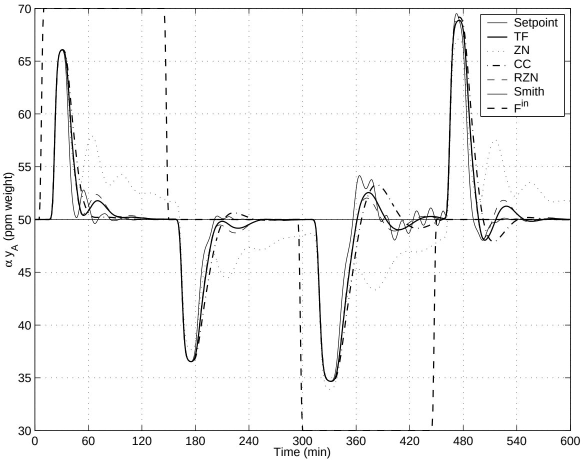


Figure 3: Weight fraction of RSH at the outlet of the drum.

Tracking We propose in this section a tracking example without varying delay where the reference changes four times within ten hours. The simulation results are presented on figure 4. At the beginning, the process is initialized on an equilibrium point with a setpoint $y_r = \alpha 50$ ppm molar. The simulation involves five steps:
Step 1: when $t \in [0, 5]$, $y_r = \alpha 50$ ppm molar. **Step** 2: when $t \in [10, 145]$, $y_r = \alpha 60$ ppm molar. **Step** 3: when $t \in [150, 295]$, $y_r = \alpha 50$ ppm molar. **Step** 4: when $t \in [300, 445]$, $y_r = \alpha 40$ ppm molar. **Step** 5: when $t \in [450, 600]$, $y_r = \alpha 50$ ppm molar.

The Smith predictor response is faster than the others for the fourth steps, and in spite of the small first overshoot, its rise time and its settling time are the shortest. The RZN and TF responses have a similar behavior. The rise time and the settling time keep fast, although slower than the Smith response. The CC response has the lowest rise time, and the response follows a sizeable single overshoot to converge towards the setpoint in the same settling time than RZN and TF.

Results are very similar to the disturbances rejection case. Again the TF tuned PI controller behaves well when compared to others. Only the Smith predictor can perform about the same.

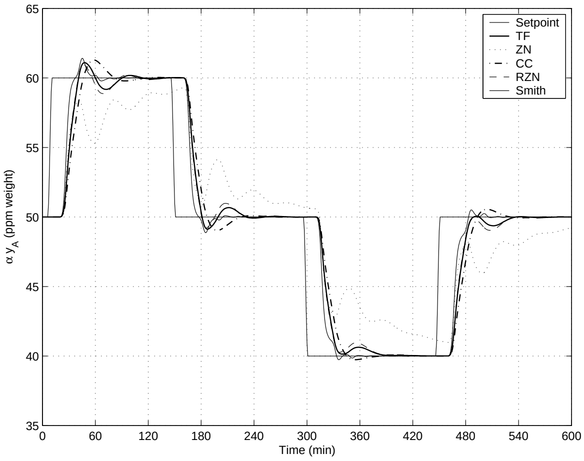


Figure 4: Weight fraction of *RSH* at the outlet of the drum. Tracking test.

5 Conclusion

The obtained results illustrate the behaviors of the process model with some different PI controllers and with the Smith predictor.

If the dead-time is accurately identified, Smith predictor can be tuned so that the obtained closed-loop response is fast. Indeed, the Smith predictor can give faster convergence than the best of PI controllers. Nevertheless, with this type of extreme tuning, a small dead-time mismatch can make the output diverge. Usually, the Smith predictor tunings will be loosen in order to avoid any divergence, at the expense of suboptimality when the delay does not vary.

When the delay is not well known, the most interesting response is obtained with the Tavakoli and Fleming (TF) PI tunings. In this situation, Smith predictor tunings must be detuned significantly, which leads to a response less effective than the TF PI tunings one.

As a conclusion, the TF PI tuning rules seem to be a good choice because of its higher stability compared to the Smith predictor in case of dead-time disturbance. Although easily implemented and effective, this controller is however not optimal when the model is accurate. As noticed before, the Smith predictor is sensitive to dead-time mismatch, and if the dead-time is varying significantly with time, the dynamic performance of the Smith predictor can be damaged. However, if an on-line dead-time estimation is applied, the Smith predictor could then be used easily with large improvement. Our current work focuses on such adaptive Smith-like predictors. Other PID tuning rules such as Lee et al. [4] and Skogestad [6] may be conceivable as well but a fair comparison would require that the D term is also considered.

References

[1] K. J. Åström and T. Hägglund. *PID Controllers: Theory, Design, and Tuning.* Instrument Society of America, (1995).

[2] G. H. Cohen and G. A. Coon. Theoretical consideration of retarded control. *Trans. A.S.M.E.*, Vol. 75(No. 1):pp. 827–834, (1953).

[3] C. C. Hang, K. J. Åström, and W. K. Ho. Refinements of the Ziegler-Nichols tuning formulas. *IEE Proceeding-D*, Vol. 138(No. 2):pp. 111–118, (1991).

[4] Y. Lee, S. Park, M. Lee, and C. Brosilow. PID controller for desired closed-loop responses for SI/SO systems. *AIChE Journal*, Vol. 44(No. 1):pp. 106–115, (2002).

[5] W. S. Levine. *The Control Handbook.* CRC Press and IEEE Press, (1996).

[6] S. Skogestad. Probably the best simple PID tuning rules in the world. *AIChE Annual Meeting*, page pp. 276h, (2001).

[7] S. Tavakoli and P. Fleming. Optimal tuning of PI controllers for first order plus dead time/long dead time models using dimensional analysis. *Proc. of the 7th European Control Conf.*, (2003).

[8] P. Tona and D. Humeau. Efficient system identification for model predictive control with the ISIAC software. Submitted to the 1st International Conference on Informatics in Control (ICINCO) 2004.

[9] J. G. Ziegler and N. B. Nichols. Optimum settings for automatic controllers. *Trans. A.S.M.E.*, Vol. 64:pp. 759–765, (1942). Available from www.driedger.ca.

Collocation and inversion for a reentry optimal control problem

Tobias NECKEL [1] Christophe TALBOT [2] Nicolas PETIT [3]

1 - École Polytechnique, 91128 Palaiseau Cedex France
Tobias.Neckel@cnes.fr
2 - Centre National d'Études Spatiales, CNES - DDA/SDT/SP Evry, Rond Point de l'Espace 91023 EVRY Cedex France
Christophe.Talbot@cnes.fr
3 - Centre Automatique et Systèmes, École Nationale Supérieure des Mines de Paris 60, bd Saint-Michel 75272, Paris Cedex 06, France
petit@cas.ensmp.fr

Abstract

The purpose of this article is to provide the reader with an overview of an inversion based methodology applied to a shuttle atmospheric reentry problem. The proposed method originates in the search for computationally efficient trajectory optimization as an enabling technology for versatile real-time trajectory generation. The technique is based on the nonlinear control theory notion of inversion and flatness. This point of view allows to map the system dynamics, objective, and constraints to a lower dimensional space. The optimization problem is then solved in the lower dimensional space. Eventually the optimal states and inputs are recovered from the inverse mapping.

1 Introduction

The purpose of this article is to provide the reader with an overview of an inversion based methodology applied to a shuttle atmospheric reentry problem. This problem has a 6 states, 2 controls nonlinear dynamics with terminal and initial constraints and a terminal cost function. Aerodynamics models (linear for lift and quadratic for drag) are considered. Gravity and air density are modelled according to the classic non rotating spherical earth potential and exponential models.

The proposed method originates in the search for computationally efficient trajectory optimization as an enabling technology for versatile real-time trajectory generation. Trajectory generation of unmanned aerial vehicles is an example where the tools of real-time trajectory optimization can be extremely useful. In [9, 13, 12], this new technique was presented and used to solve such problems. In [11] this methodology was applied to formation flight of micro-satellites under J2 gravitational effect. Following the same ideas the real time trajectory generation of a planar missile was addressed [10] with similar drag and lift models.

The technique is based on the nonlinear control theory notion of inversion [7] and flatness [3, 4]. This point of view allows to map the system dynamics, objective, and constraints to a lower dimensional space. The optimization problem is then solved in the lower dimensional space. Eventually the optimal states and inputs are recovered from the inverse mapping.

The example treated in this report has interesting features. First it is more complex in terms of dimensionality and nonlinearities than the previously cited examples. Second the dynamics are not flat. In other words it is not possible to fully invert the system dynamics. This particular situation deserves a careful treatment of the parametrization of the states variables. Numerical results are given, and a comparison with existing techniques for this example [1] is given. In short, the proposed approach appears tracktable, but could be improved further by paying more attention to the choice of the nonlinear programming solver and the finite dimensional representation that are used.

2 Background information

In this section we present the general framework of inversion-based collocation methods for numerical solution to optimal control problems. Most of this material can be found in [13]. We address the simple single-input case which is by far the most easy and emphasizes the role of inversion.

2.1 Optimal Control Problem

Consider the single input nonlinear control system

$$\begin{aligned} &\dot{x} = f(x) + g(x)u, \\ &\mathbb{R} \ni t \mapsto x \in \mathbb{R}^n, \mathbb{R} \ni t \mapsto u \in \mathbb{R} \end{aligned} \tag{1}$$

where all vector fields and functions are smooth functions. It is desired to find a trajectory of (1) $[t_0, t_f] \ni t \mapsto (x, u)(t) \in \mathbb{R}^{n+1}$ that minimizes the cost

$$\begin{aligned} J(x,u) =& \phi_f(x(t_f), u(t_f)) + \phi_0(x(t_0), u(t_0)) \\ &+ \int_{t_0}^{t_f} L(x(t), u(t))dt, \end{aligned}$$

where L is a nonlinear function, subject to a vector of initial, final, and trajectory constraints

$$\begin{aligned} &lb_0 \leq \psi_0(x(t_0), u(t_0)) \leq ub_0, \\ &lb_f \leq \psi_f(x(t_f), u(t_f)) \leq ub_f, \\ &lb_t \leq S(x,u) \leq ub_t, \end{aligned} \tag{2}$$

respectively. For conciseness, we will refer to this optimal control problem as

$$\begin{cases} \min\limits_{(x,u)} J(x,u) \\ \text{subject to} \\ \dot{x} = f(x) + g(x)u, \\ lb \leq c(x,u) \leq ub. \end{cases} \tag{3}$$

2.2 Different approaches

2.2.1 Classical collocation

One numerical approach to solve this optimal control problem is the direct collocation method outlined by Hargraves and Paris in [6]. The idea behind this approach is to transform the optimal control problem into a nonlinear programming problem. This is accomplished using a time mesh

$$t_0 = t_1 < t_2 < \ldots < t_N = t_f \tag{4}$$

and approximating the state x and the control input u as piecewise polynomials $\hat{x}$ and $\hat{u}$, respectively. Cubic polynomial may be chosen for the states and a linear polynomial for the control on each interval represents a good choice. Collocation is then used at the midpoint of each interval to satisfy Equation (1). Let $\hat{x}(x(t_1)^T, ..., x(t_N)^T)$ and $\hat{u}(u(t_1), ..., u(t_N))$ denote the approximations to x and u, respectively, depending on $(x(t_1)^T, ..., x(t_N)^T) \in \mathbb{R}^{nN}$ and $(u(t_1), ..., u(t_N)) \in \mathbb{R}^N$ corresponding to the value of x and u at the grid points. Then one solves the following finite dimension approximation of the original control problem (3)

$$\begin{cases} \min\limits_{y \in \mathbb{R}^M} F(y) = J(\hat{x}(y), \hat{u}(y)) \\ \text{subject to} \\ \dot{\bar{x}} - f(\hat{x}(y), \hat{u}(y)) = 0, \quad lb \leq c(\hat{x}(y), \hat{u}(y)) \leq ub, \\ \qquad \forall t = \dfrac{t_j + t_{j+1}}{2} \quad j = 1, \ldots, N-1 \end{cases} \tag{5}$$

where $y = (x(t_1)^T, u(t_1), \ldots, x(t_N)^T, u(t_N))$, and $M = \dim y = (n+1)N$.

2.2.2 Inverse dynamic optimization

In [15] Seywald suggested an improvement to the previous method (see also [2] page 362 for an overview of this method). Following this work, one first solves a subset of system dynamics in (3) for the the control in terms of combinations of the state and its time derivative. Then one substitutes for the control in the remaining system dynamics and constraints. Next all the time derivatives $\dot{x}_i$ are approximated by the finite difference approximations

$$\dot{\bar{x}}(t_i) = \frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i}$$

to get

$$\left.\begin{array}{l} p(\dot{\bar{x}}(t_i), x(t_i)) = 0 \\ q(\dot{\bar{x}}(t_i), x(t_i)) \leq 0 \end{array}\right\} \quad i = 0, ..., N-1.$$

The optimal control problem is turned into

$$\begin{cases} \min\limits_{y \in \mathbb{R}^M} F(y) \\ \text{subject to} \\ p(\dot{\bar{x}}(t_i), x(t_i)) = 0 \\ q(\dot{\bar{x}}(t_i), x(t_i)) \leq 0 \end{cases} \tag{6}$$

where $y = (x(t_1)^T, \ldots, x(t_N)^T)$, and $M = \dim y = nN$. As with the Hargraves and Paris method, this parameterization of the optimal control problem (3) can be solved using nonlinear programming.

The dimensionality of this discretized problem is lower than the dimensionality of the Hargraves and Paris method, where both the states and the input are the unknowns. This induces substantial improvement in numerical implementation (see again [15] for an implementation of the Goddard problem).

2.2.3 Proposed Numerical Approach

In fact, it is usually possible to reduce the dimension of the problem further. Given an output, it is generally possible to parameterize the control and a part of the state in terms of this output and its time derivatives. In contrast to the previous approach, one must use more than one derivative of this output for this purpose.

When the whole state and the input can be parameterized with one output, one says that the system is flat [3]. When the parameterization is only partial, the dimension of the subspace spanned by the output and its derivatives is given by r the *relative degree* of this output.

Definition 1 ([7]) A single input single output system

$$\begin{cases} \dot{x} = f(x) + g(x)u \\ y = h(x) \end{cases} \tag{7}$$

is said to have *relative degree* r at point x_0 if $L_g L_f^k h(x) = 0$, in a neighborhood of x_0, and for all $k < r - 1$ $L_g L_f^{r-1} h(x_0) \neq 0$ where $L_f h(x) = \sum_{i=1}^n \frac{\partial h}{\partial x_i} f_i(x)$ is the derivative of h along f.

Roughly speaking, r is the number of times one has to differentiate y before u appears.

Result 1 ([7]) *Suppose the system* (7) *has relative degree r at x^0. Then $r \leq n$. Set*

$$\begin{aligned} \phi_1(x) &= h(x) \\ \phi_2(x) &= L_f h(x) \\ &\vdots \\ \phi_r(x) &= L_f^{r-1} h(x). \end{aligned}$$

If r is strictly less than n, it is always possible to find $n - r$ more functions $\phi_{r+1}(x), ..., \phi_n(x)$ such that the mapping

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_n(x) \end{pmatrix}$$

has a Jacobian matrix which is nonsingular at x^0 and therefore qualifies as a local coordinates transformation in a neighborhood of x^0. The value at x^0 of these additional functions can be fixed arbitrarily. Moreover, it is always possible to choose $\phi_{r+1}(x), ..., \phi_n(x)$ in such a way that $L_g\phi_i(x) = 0$, for all $r+1 \leq i \leq n$ and all x around x^0.

The implication of this result is that there exists a change of coordinates $x \mapsto z = (z_1, z_2, ..., z_n)$ such that the systems equations may be written as

$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = z_3 \\ \quad \vdots \\ \dot{z}_{r-1} = z_r \\ \dot{z}_r = b(z) + a(z)u \\ \dot{z}_{r+1} = q_{r+1}(z) \\ \quad \vdots \\ \dot{z}_n = q_n(z) \end{cases}$$

where $a(z)$ is nonzero for all z in a neighborhood of $z^0 = \phi(x^0)$.

In these new coordinates, any optimal control problem can be solved by a partial collocation, i.e. collocating only $(z_1, z_{r+1}, ..., z_n)$ instead of a full collocation $(z_1, ..., z_r, z_{r+1}, ..., z_n, u)$. Inverting the change of coordinates, the state and the input $(x_1, ..., x_n, u)$ can be expressed in terms of $(z_1, ..., z_1^{(r)}, z_{r+1}, ..., z_n)$. This means that once translated into these new coordinates, the original control problem (3) will involve r successive derivatives of z_1.

It is not realistic to use finite difference approximations as soon as $r > 2$. In this context, it is convenient to represent $(z_1, z_{r+1}, ...z_n)$ as B-splines. B-splines are chosen as basis functions because of their ease of enforcing continuity across knot points and ease of computing their derivatives.

Both equation from the dynamics and the constraints will be enforced at the collocation points. In general, w collocation points are chosen uniformly over the time interval $[t_o, t_f]$, (though optimal knots placements or Gaussian points may also be considered and are numerically important). The problem can be stated as the following nonlinear programming form:

$$\begin{cases} \min\limits_{y \in \mathbb{R}^M} F(y) \\ \text{subject to} \\ \dot{z}_{r+1}(y) - q_{r+1}(z)(y) = 0 \\ \quad \vdots \\ \dot{z}_n(y) - q_n(z)(y) = 0 \text{ for every } w \\ \quad lb \leq c(y) \leq ub \end{cases} \tag{8}$$

where y represents the unknown coefficients of the B-splines. These have to be found using nonlinear programming.

2.2.4 Comparisons

Our approach is a generalization of inverse dynamic optimization. Let us summarize the presented approaches One could write the optimal control problem with:

- "Full collocation" solving problem (5) by collocating $(x, u) = (x_1, ..., x_n, u)$ without any attempt of variable elimination. After collocation the dimension of the unknowns space is $\mathcal{O}(n + 1)$.

- "Inverse dynamic optimization" solving problem (6) by collocating $x = (x_1, ..., x_n)$. Here the input is eliminated from the equation using one derivative of the state. After collocation the dimension of the unknowns space is $\mathcal{O}(n)$.

- "Flatness parametrization" (Maximal inversion), our approach, solving problem (8) in the new coordinates collocating only $(z_1, z_{r+1}, ..., z_n)$. Here we eliminate as many variables as possible and replace them using the first r derivatives of z_1. After collocation, the dimension of the unknowns space is $\mathcal{O}(n - r + 1)$.

2.3 The ruled manifold criterion

When facing a new system dynamics, it would be interesting to know wether these can be fully inverted or not. The single-input case presented before is the exception. Unfortunately, up today, there does not exist any flatness criterion. Nevertheless the following necessary condition can be a handy tool to check wether one may completely invert a system. This necessary condition for a system to be flat is given by the following criterion [14] (see also [8]).

Result 2 ([14]) *Assume the system $\dot{x} = f(x, u)$ is flat. The projection on the p-space of the submanifold $p = f(x, u)$, where x is considered as a parameter, is a ruled manifold for all x.*

Eliminating u from the dynamics $\dot{x} = f(x, u)$ yields a set of equations $F(x, \dot{x}) = 0$ that defines a ruled manifold. In other words for all $(x, p) \in \mathbb{R}^{2n}$ such that $F(x, p) = 0$, there exists a direction $d \in \mathbb{R}^n$, $d \neq 0$ such that

$$\forall \lambda \in \mathbb{R}, F(x, p + \lambda d) = 0.$$

3 The reentry problem

In this section we present the reentry problem. We detail the nonlinear dynamics, the constraints and the cost function. We show that this system is not flat and explain how to parameterize its trajectories using a reduced number of variables and additional constraints. Finally we give a rewriting of the optimal control problem in terms of this reduced number of unknowns.

3.1 Dynamics

As detailed in Betts [1], the motion of the space shuttle are defined by the following set of equations

$$\dot{h} = v \sin\gamma \tag{9}$$

$$\dot{\phi} = \frac{v}{r}\cos\gamma\sin\psi/\cos\theta \tag{10}$$

$$\dot{\theta} = \frac{v}{r}\cos\gamma\cos\psi \tag{11}$$

$$\dot{v} = -\frac{D(\alpha)}{m} - g\sin\gamma \tag{12}$$

$$\dot{\gamma} = \frac{L(\alpha)}{mv}\cos\beta + \cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right) \tag{13}$$

$$\dot{\psi} = \frac{1}{mv\cos\gamma}L(\alpha)\sin\beta + \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta \tag{14}$$

where h denotes the altitude, ϕ the longitude, θ the latitude, v the velocity, γ the flight path, ψ the azimuth. The two control are α the angle of attack and β the bank angle.

3.2 Control objective and constraints

Here our problem is to maximize the final value of the θ variable in a *given time* t_f. The initial conditions are prescribed as

$$\begin{aligned} h(0) &= 260000 \text{ ft}\\ \phi(0) &= 0 \text{ deg}\\ \theta(0) &= 0 \text{ deg}\\ v(0) &= 25600 \text{ ft/sec}\\ \gamma(0) &= -1 \text{ deg}\\ \psi(0) &= 90 \text{ deg}\end{aligned}$$

In the numerical example treated in this report the final time t_f equals 2008.59 s. The study is restricted to the trajectory satisfying

$$\begin{aligned} 0 \le h, -89 \text{ deg} \le \theta \le 89 \text{ deg}\\ 1 \le v, -89 \text{ deg} \le \gamma \le 89 \text{ deg}\\ -90 \text{ deg} \le \alpha \le 90 \text{ deg}, -89 \text{ deg} \le \beta \le 89 \text{ deg}\end{aligned}$$

The final point of the trajectory is defined by the terminal area energy management (TAEM) interface which is defined by the following relations

$$h(t_f) = 80000 \text{ ft}, v(t_f) = 2500 \text{ ft/s}, \gamma(t_f) = -5 \text{ deg}$$

3.3 Physics constants and parameters

We use $\mu =$ 0.14076539e17 as gravitational constant, $Re =$ 20902900 ft as the radius of the Earth, $S = 2690$ ft^2 as the aerodynamic reference surface, $h_{ref} = 23800$ ft and ρ_0 =0.002378 for the following physics parameters

$$g = \mu/r^2 \tag{15}$$

$$\rho = \rho_0 \exp(-(r - Re)/h_{ref}) \tag{16}$$

We use $C_L = a_0 + a_1\alpha$ where α is in deg, a_0 =-0.20704, a_1 =0.029244. Lift is then given by

$$L = \frac{1}{2}C_L S\rho v^2 \tag{17}$$

Also we note $C_D = b_0 + b_1\alpha + b_2\alpha^2$, where b_0 =0.07854, $b_1 =$ -0.61592e-2, $b_2 =$ 0.621408e-3 and use it in

$$D = \frac{1}{2}C_D S\rho v^2 \tag{18}$$

The mass of the shuttle was chosen as

$$m = 6309.44 \text{ lbs}$$

3.4 The system is not flat

We use the ruled manifold criterion presented in section 2.3 to prove that the system is not flat.

Eliminating the control from the reentry dynamics yields an equation $F(x, \dot{x}) = 0$. To get this equation we have to solve for the unknowns α and β in terms of the states and its derivatives.

First one may pick equation (12) to get

$$D(\alpha) = -m\dot{v} - mg\sin(\gamma)$$

Then solve according to the physical model (18) to get

$$\alpha = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_2(b_0 + \frac{2m(\dot{v}+g\sin\gamma)}{\rho S v^2})}}{2b_2} \tag{19}$$

On the other hand it straightforward to solve for β using equation (13), equation (14) and the fact that -89 deg $\le \beta \le 89$ deg. This gives

$$\beta = \arctan\left(\frac{\cos\gamma(\dot{\psi} - \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta)}{\dot{\gamma} - \cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right)}\right) \tag{20}$$

Using these last two relations in the reentry dynamics we get the manifold equation $F(x,p) = 0$, where $p = (p_1, p_2, p_3, p_4, p_5, p_6)^T = \dot{x}$ satisfy

$$p_1 = v\sin\gamma \tag{21}$$

$$p_2 = \frac{v}{r}\cos\gamma\sin\psi/\cos\theta \tag{22}$$

$$p_3 = \frac{v}{r}\cos\gamma\cos\psi \tag{23}$$

and Equation (24) Now let us look for a non-zero direction $d = (d_1, d_2, d_3, d_4, d_5, d_6)^T \in \mathbb{R}^6$ such that at a point (x, p) such that $F(x,p) = 0$, for all $\lambda \in \mathbb{R}$, $F(x, p + \lambda d) = 0$.

The first three equations (21), (22), (23) give

$$\begin{aligned} p_1 + \lambda d_1 &= v\sin\gamma\\ p_2 + \lambda d_2 &= \frac{v}{r}\cos\gamma\sin\psi/\cos\theta\\ p_3 + \lambda d_3 &= \frac{v}{r}\cos\gamma\cos\psi\end{aligned}$$

which give

$$d_1 = 0, d_2 = 0, d_3 = 0$$

Equation (24) gives after using the simplification $\sin(\arctan x) = \frac{x}{\sqrt{1+x^2}}$ Equation (25)

This equation must hold for all $\lambda \in \mathbb{R}$. After taking the square of the last expression, the square root in the last expression involving d_4 is the only one that still contains a

$$\begin{aligned} p_6 =& \frac{\rho S v^2}{2mv\cos\gamma}\left(a_0 + a_1\frac{180}{\pi}\frac{-b_1 \pm \sqrt{b_1^2 - 4b_2(b_0 + \frac{2m(p_4 + g\sin\gamma)}{\rho S v^2})}}{2b_2}\right) \times ... \\ & \sin\left(\arctan\left(\frac{\cos\gamma(p_6 - \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta)}{p_5 - \cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right)}\right)\right) \\ & + \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta \end{aligned} \tag{24}$$

$$\begin{aligned} p_6 + \lambda d_6 =& \frac{\rho S v^2}{2mv\cos\gamma}\left(a_0 + a_1\frac{180}{\pi}\frac{-b_1 \pm \sqrt{b_1^2 - 4b_2(b_0 + \frac{2m(p_4 + \lambda d_4 + g\sin\gamma)}{\rho S v^2})}}{2b_2}\right) \times ... \\ & \frac{\cos\gamma(p_6 + \lambda d_6 - \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta)}{\sqrt{\left(\cos\gamma(p_6 + \lambda d_6 - \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta)\right)^2 + \left(p_5 + \lambda d_5 - \cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right)\right)^2}} \\ & + \frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta \end{aligned} \tag{25}$$

square root terms in λ. It can not be matched to anything else in the expression. Thus, necessarily,

$$d_4 = 0$$

Taking the square of the last equation gives rise to the following second order polynomial in λ

$$\begin{aligned} & \lambda^2(d_5^2 + d_6^2) \\ & + 2\lambda\big(p_5 d_5 - d_5(\cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right))... \\ & + \cos\gamma^2(p_6 d_6 - a_6(\frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta)\big)... \\ & + p_5^2 - 2p_5\cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right) + (\cos\gamma\left(\frac{v}{r} - \frac{g}{v}\right))^2 \\ & + \cos^2\gamma(p_6^2 - 2p_6\frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta... \\ & + \Big(\frac{v}{r\cos\theta}\cos\gamma\sin\psi\sin\theta\Big)) \\ & - c\cos^2\gamma \end{aligned}$$

where

$$\begin{aligned} c =& \frac{\rho S v^2}{2mv\cos\gamma} \\ & \big(a_0 + a_1\frac{180}{\pi} \\ & \quad \frac{-b_1 \pm \sqrt{b_1^2 + 4b_2(b_0 + \frac{2m(p_4 + \lambda d_4 - g\sin\gamma)}{\rho S v^2})}}{2b_2}) \end{aligned}$$

For this polynomial to be identically zero, necessarily we must have

$$d_5 = 0,\ d_6 = 0$$

Thus the candidate vector for a direction of the ruled manifold is $d = 0$. This shows the manifold is not ruled and so the system is not flat.

3.5 Parameterization

Should the system have been flat, we would have been using only 2 quantities (same number as inputs) for the parametrization of all its variables. As we will see in the following, we need 3 quantities instead. We now use

$$\begin{aligned} z_1 &= r = h + Re \\ z_2 &= \theta \\ z_3 &= \phi \end{aligned}$$

where Re is the radius of the Earth. Assuming that around the trajectory -90 deg $< \psi <$ 90 deg, we recover from (10) and (11)

$$\psi = \arctan\left(\frac{\dot{z}_3}{\dot{z}_2}\cos z_2\right) \tag{26}$$

Since -90 deg $< \gamma <$ 90 deg, we get from (9) and (11)

$$\begin{aligned} \gamma &= \arctan\left(\frac{\dot{z}_1}{\dot{z}_2}\frac{\cos\psi}{z_1}\right) \\ &= \arctan\left(\frac{\dot{z}_1}{z_1\sqrt{\dot{z}_2^2 + \dot{z}_3^2\cos^2 z_2}}\right) \end{aligned} \tag{27}$$

and then

$$\begin{aligned} v &= \sqrt{\left(\frac{z_1\dot{z}_2}{\cos\psi}\right)^2 + \dot{z}_1^2} \\ &= \sqrt{\dot{z}_1^2 + z_1^2\left(\dot{z}_2^2 + \dot{z}_3^2\cos^2 z_2\right)} \end{aligned} \tag{28}$$

It is convenient in the sequel to solve for the derivatives $\dot{v}$, $\dot{\gamma}$, $\dot{\psi}$. These quantites can be obtained either by direct differentiation of (26) (27) and (28) as

$$\begin{aligned}\dot\psi(1+\tan^2\psi) &= \frac{d(\tan\psi)}{dt}\\ &= \frac{d}{dt}\left(\frac{\dot z_3}{\dot z_2}\cos z_2\right)\\ &= \frac{\ddot z_3}{\dot z_2}\cos z_2 - \dot z_3\sin z_2 - \frac{\dot z_3\ddot z_2}{\dot z_2^2}\cos z_2\end{aligned}$$

which gives

$$\begin{aligned}\dot\psi = &\left(1+\frac{\dot z_3^2}{\dot z_2^2}\cos^2 z_2\right)^{-1}\\ &\left(\frac{\ddot z_3}{\dot z_2}\cos z_2 - \dot z_3\sin z_2 - \frac{\dot z_3\ddot z_2}{\dot z_2^2}\cos z_2\right) \qquad (29)\end{aligned}$$

and

$$\begin{aligned}\dot v =&\ddot z_1\sin\gamma + \cos\gamma\cos\psi\left(\ddot z_2 z_1 + \dot z_2\dot z_1\right)\\ &+\cos\gamma\sin\psi\times\\ &(\ddot z_3 z_1\cos z_2 + \dot z_3\dot z_1\cos z_2 - \dot z_2\dot z_3 z_1\sin z_2) \qquad (30)\\ \dot\gamma =&\frac{1}{v}\ddot z_1\cos\gamma - \frac{1}{v}\sin\gamma\cos\psi\left(\ddot z_2 z_1 + \dot z_2\dot z_1\right)\\ &-\frac{1}{v}\sin\gamma\sin\psi\times\\ &(\ddot z_3 z_1\cos z_2 + \dot z_3\dot z_1\cos z_2 - \dot z_2\dot z_3 z_1\sin z_2) \qquad (31)\end{aligned}$$

The lift is computed from equations (13) and (14) as

$$\begin{aligned}L =&mv\Big((\dot\psi - v/z_1\cos\gamma\sin\psi\tan z_2)\cos\gamma)^2\\ &+(\dot\gamma-(v^2/z_1-g)\cos\gamma/v)^2\Big)^{1/2}\\ &sign(\dot\gamma-(v^2/z1-g)\cos\gamma/v)\end{aligned}$$

which we note after substitution with equations (26), (27), (28), (30) and (29)

$$L = f_L(z_1,\dot z_1,\ddot z_1,z_2,\dot z_2,\ddot z_2,z_3,\dot z_3,\ddot z_3) \qquad (32)$$

The bank angle can be recomputed from the previous expression and equation (13)

$$\beta = -\arccos((\dot\gamma-(v^2/z_1-g)\cos\gamma/v/m)v/L)$$

which we note after substitution with equations (27), (28) and (31)

$$\beta = f_\beta(z_1,\dot z_1,\ddot z_1,z_2,\dot z_2,z_3,\dot z_3) \qquad (33)$$

Using the linear model for lift (see appendix), we can solve for the angle of attack

$$\alpha = (2L/\rho/v^2/S-a_0)/a_1$$

which we note after substitution with equations (28) and (32) and the air density model for $\rho(z_1)$ given by equation (16)

$$\alpha = f_\alpha(z_1,\dot z_1,\ddot z_1,z_2,\dot z_2,\ddot z_2,z_3,\dot z_3,\ddot z_3) \qquad (34)$$

The drag is then recomputed from the law

$$D = \frac{1}{2}\rho S v^2 C_D$$

3.5.1 Parameterization constraints

The reentry dynamics have the same nonlinear structure as the following simple nonlinear system with 3 states and 2 inputs

$$\begin{aligned}\dot x_1 &= -D(u_1)\\ \dot x_2 &= L(u_1)\cos u_2\\ \dot x_3 &= L(u_1)\sin u_2\end{aligned}$$

In general this system is not flat (e.g. if D and L correspond to drag and lift models). In other words, not any time function $t\mapsto(x_1(t),x_2(t),x_3(t))$ is a trajectory of the system. But the trajectories of the system, i.e. time functions $t\mapsto(x_1(t),x_2(t),x_3(t),u_1(t),u_2(t))$ solution to the dynamics, indeed satisfy

$$\tan u_2 = \left(\frac{\dot x_3}{\dot x_2}\right) \qquad (35)$$

and

$$L = \sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2)$$

These are only necessary conditions. Sufficient extra conditions are that

$$\dot x_1 = -D(L^{-1}(\sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2)))$$

In order to solve equation (35), one has to pick the right determination of the angle. In general it can not be assumed that $u_2\in]-\pi/2,\pi/2[$ (it is the case in our example though). Let us call u_2^* this solution (defined up to π). A suitable value has to be such that

$$\begin{aligned}\dot x_2 &= L(u_1)\cos u_2^*\\ \dot x_3 &= L(u_1)\sin u_2^*\end{aligned}$$

To summarize, the trajectories of the system are of the form

$$\begin{aligned}t\mapsto&(x_1(t),x_2(t),x_3(t),\\ &L^{-1}(\sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2^*),u_2^*)\end{aligned}$$

where x_1, x_2, x_3, u_2^* are any arbitrary function that satisfy

$$\begin{aligned}\dot x_1 &= -D(L^{-1}(\sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2^*)))\\ \dot x_2 &= \sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2^*)\cos u_2^*\\ \dot x_3 &= \sqrt{\dot x_2^2+\dot x_3^2}sign(\dot x_2\cos u_2^*)\sin u_2^*\\ \tan u_2^* &= \left(\frac{\dot x_3}{\dot x_2}\right)\end{aligned}$$

Similarly, in our case the following constraints must hold

1. First the drag and the lift must correspond. In other words, the drag that is computed from the lift must be such that

$$m\dot v + g\sin\gamma + D = 0$$

2. Also the sign that appears in the lift expression has to be taken into account. Two additional constraints have to be satisfied to transform the previous necessary condition in a sufficient condition. It is assumed that $\alpha \in]-\pi/2, \pi/2[$. So u_2^* is uniquely defined by the $\arctan$ function. As a summary, the trajectories have to satisfy

$$\begin{aligned}&(\dot\psi - v/z_1 \cos\gamma \sin\psi \tan z_2)\cos\gamma) \\ &\quad = L\cos\beta/m/v/\cos\ \gamma \\ &(\dot\gamma - (v^2/z_1 - g)\cos\gamma/v) \\ &\quad = L\sin\beta/m/v\end{aligned}$$

3.5.2 Parameterization of the trajectories

The previous relations derived at section 3.5 are necessary conditions. In other words if the time functions

$$t \mapsto (h(t), \phi(t), \theta(t), V(t), \gamma(t), \psi(t), \alpha(t), \beta(t))$$

are solutions of the reentry dynamics then they are of the form

$$\begin{aligned}h &= z_1 - R_e \\ \phi &= z_3 \\ \theta &= z_2 \\ v &= \sqrt{\dot z_1^2 + z_1^2\,(\dot z_2^2 + \dot z_3^2 \cos^2 z_2)} \\ \gamma &= \arctan\left(\frac{\dot z_1}{z_1\sqrt{\dot z_2^2 + \dot z_3^2\cos^2 z_2}}\right) \\ \psi &= \arctan\left(\frac{\dot z_3}{\dot z_2}\cos z_2\right) \\ \alpha &= f_\alpha(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) \\ \beta &= f_\beta(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, z_3, \dot z_3)\end{aligned}$$

Conversely any time function $t \mapsto (h(t), \phi(t), \theta(t), V(t), \gamma(t), \psi(t), \alpha(t), \beta(t))$ computed from the same relations are not solutions to the reentry dynamics. Sufficient extra conditions are that these functions must satisfy the extra conditions

$$m\dot v + g\sin\gamma + \frac{1}{2}C_D\rho S\left(\left(\frac{z_1\dot z_2}{\cos z_3}\right)^2 + \dot z_1^2\right) = 0$$

$$\begin{aligned}&(\dot\psi - v/z_1 \cos\gamma \sin\psi \tan z_2)\cos\gamma) \\ &= L\cos\beta/m/v/\cos\ \gamma \\ &(\dot\gamma - (v^2/z_1 - g)\cos\gamma/v) \\ &= L\sin\beta/m/v\end{aligned}$$

These three relations can be rewritten, after substitution with the necessary conditions (26), (27), (28), (30), (32), (33)

$$F_1(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{36}$$
$$F_2(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{37}$$
$$F_3(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{38}$$

3.6 Rewriting of the optimal control problem

The problem is only to find the best time functions $[0, t_f] \ni t \mapsto (z_1(t), z_2(t), z_3(t))$ so as to maximize $z_2(t_f)$ under the following constraints.

- Initial constraints

$$h(0) = z_1(0) - Re \tag{39}$$
$$\phi(0) = z_3(0) \tag{40}$$
$$\theta(0) = z_2(0) \tag{41}$$
$$v(0) = \sqrt{\dot z_1^2(0) + z_1^2(0)\,(\dot z_2^2(0) + \dot z_3^2(0)\cos^2 z_2(0))} \tag{42}$$
$$\gamma(0) = \arctan\left(\frac{\dot z_1(0)}{z_1(0)\sqrt{\dot z_2^2(0) + \dot z_3^2(0)\cos^2 z_2(0)}}\right) \tag{43}$$
$$\psi(0) = \arctan\left(\frac{\dot z_3(0)}{\dot z_2(0)}\cos z_2(0)\right) \tag{44}$$

- Trajectory constraints (must hold for all $t \in [0, t_f]$)

$$F_1(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{45}$$
$$F_2(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{46}$$
$$F_3(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) = 0 \tag{47}$$

$$\begin{aligned}&0 \le z_1 - Re, -89 \le z_2 \le 89 \\ &1 \le \sqrt{\dot z_1^2 + z_1^2\,(\dot z_2^2 + \dot z_3^2\cos^2 z_2)}, \\ &-89 \le \arctan\left(\frac{\dot z_1}{z_1\sqrt{\dot z_2^2 + \dot z_3^2\cos^2 z_2}}\right) \le 89, \\ &-90 \le f_\alpha(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, \ddot z_2, z_3, \dot z_3, \ddot z_3) \le 90, \\ &-89 \le f_\beta(z_1, \dot z_1, \ddot z_1, z_2, \dot z_2, z_3, \dot z_3) \le 89\end{aligned}$$

- Endpoint constraints

$$h(t_f) = z_1(t_f) - Re \tag{48}$$
$$v(t_f) = \sqrt{\dot z_1^2(t_f) + z_1^2(t_f)\,(\dot z_2^2(t_f)\dot z_3^2(t_f)\cos^2 z_2(t_f))} \tag{49}$$
$$\gamma(t_f) = \arctan\left(\frac{\dot z_1(t_f)}{z_1(t_f)\sqrt{\dot z_2^2 + \dot z_3^2\cos^2 z_2}}\right) \tag{50}$$

4 Numerical results

In this section we give numerical results using the proposed methodology. Details about the initialisation and convergence are given. Accuracy of the method is discussed and comparisons with reference results are given.

$h(t_f)$ (ft)	102600
$v(t_f)$ (ft/sec)	3291.6
$\gamma(t_f)$ (deg)	-3.6479
$\theta(t_f)$ (deg)	31.0802

Figure 1: Initial guess terminal values and cost function value.

$h(t_f)$ (ft)	80182
$v(t_f)$ (ft/sec)	2475.3
$\gamma(t_f)$ (deg)	-5.0179
$\theta(t_f)$ (deg)	33.0656

Figure 2: Terminal values and cost function value after optimisation.

4.1 Numerical setup

4.1.1 Initial guess

The system was initialized with control variables set to α =21 deg for the angle of attack, and $\beta(t) = 75 \times (-1 + t/t_f)$ for the bank angle (in deg). After a careful integration performed with Matlab `ode23`, the corresponding trajectory was found to give the data given in Figure 1.

From these trajectories the unknown coefficients were computed through a least square B-spline approximation. Of course the results depend on the number of coefficients, the order of the B-splines and the multiplicity of their knots and the fitting mesh.

Then we recomputed the control histories from the B-splines representation of the outputs z_1, z_2, z_3 using the formulas given in Section 3.5.

Finally we reintegrated the system dynamics from the same initial condition as before while using the latest control histories. Results are given for a typical case with 40 intervals (44 coefficients) per variable, 60 points mesh, 4^{th} order B-Splines with multiplicity of 3.

$$h^{40\times60}(t_f) - h_{guess}(t_f) = 55.244 \text{ ft} ,$$
$$v^{40\times60}(t_f) - v_{guess}(t_f) = -0.7559 \text{ ft/sec} ,$$
$$\gamma^{40\times60}(t_f) - \gamma_{guess}(t_f) = -0.0266 \text{ deg}$$

Results vary with the number of coefficients and Results are given for a typical case with 100 intervals (104 coefficients) per variable unknown variables, 200 points mesh, 4^{th} order B-Splines with multiplicity of 3.

$$h^{100\times200}(t_f) - h_{guess}(t_f) = -11.1395 \text{ ft} ,$$
$$v^{100\times200}(t_f) - v_{guess}(t_f) = 0.5795 \text{ ft/sec} ,$$
$$\gamma^{100\times200}(t_f) - \gamma_{guess}(t_f) = -0.0216 \text{ deg}$$

In these two cases the mesh was refined around the two boundaries of the domain, to limit the side effects of least square approximation. In fact, a linearly spaced mesh would produce much larger errors. With the 100 intervals and the 200 points linearly spaced mesh the same test gives

$$h^{100\times200l}(t_f) - h_{guess}(t_f) = 141 \text{ ft} ,$$
$$v^{100\times200l}(t_f) - v_{guess}(t_f) = 7.49 \text{ ft/sec} ,$$
$$\gamma^{100\times200l}(t_f) - \gamma_{guess}(t_f) = 0.024 \text{ deg}$$

We were investigating wether the B-Splines were able to provide us with a high degree of accuracy as required for our application. The above numerical investigation suggests that they are well suited provided a sufficiently large number of coefficient is chosen. Also the choice of the mesh matters. In the rest of the report we conduct the tests with a mesh refined around the two ends of the time interval.

4.1.2 Solving the optimal control problem

All the tests were conducted using Matlab 6.5 with the collocation routines from the Splines toolbox and the `fmincon` routine from the Optimisation toolbox.

No analytical gradients were provided, neither for the cost nor for the constraints. This has an impact on the computation times.

Scalings were used for the cost function and the constraints. This helped the nonlinear programming routine to find appropriate search lines. Also nonlinear equality constraints over the time interval (due to the parameterization) were relaxed to help convergence. Eventually the optimisation procedure was restarted once with the previous solution as an initial guess and more stringent values for the relaxation parameter.

We used 40 intervals (44 coefficients per variable) and a 65 nonlinearly spaced points mesh.

With a first run (relaxation parameter set to 1e-4) the obtained solution gave $h(t_f)$ = 79906 ft, $v(t_f)$ =2753.7 ft/sec, $\gamma(t_f)$ =-4.0726 deg, $\theta(t_f)$= 33.5771 deg. This first problem was solved using 104 iterations of `fmincon`, which used 14117 F-count and took approximatively 20 minutes on a Pentium III 1.13 GHz Windows XP based computer.

These results were eventually improved using a new relaxation parameter of 1e-5. Final results are given in Figure 2. The corresponding trajectory is detailed in Figure 3 and Figure 4. This run used 122 iterations of `fmincon`, which used 16703 F-count and took approximatively 50 minutes on the same computer.

5 Conclusions

The numerical results must be compared to the solution given in [1] that gives $\theta(t_f)$ =34.1412 deg, a higher value. The result presented here were obtained by a much different technique. It seems we converged to a different solution. Also it seems that the accuracy could be improved further using more coefficients for the B-splines representation and well adapted meshes. It should be noted that only a simple nonlinear solver was used in this study and that the use of more complex, yet less convenient for implementation, solvers such as `NPSOL` [5] with analytic gradients could help too.

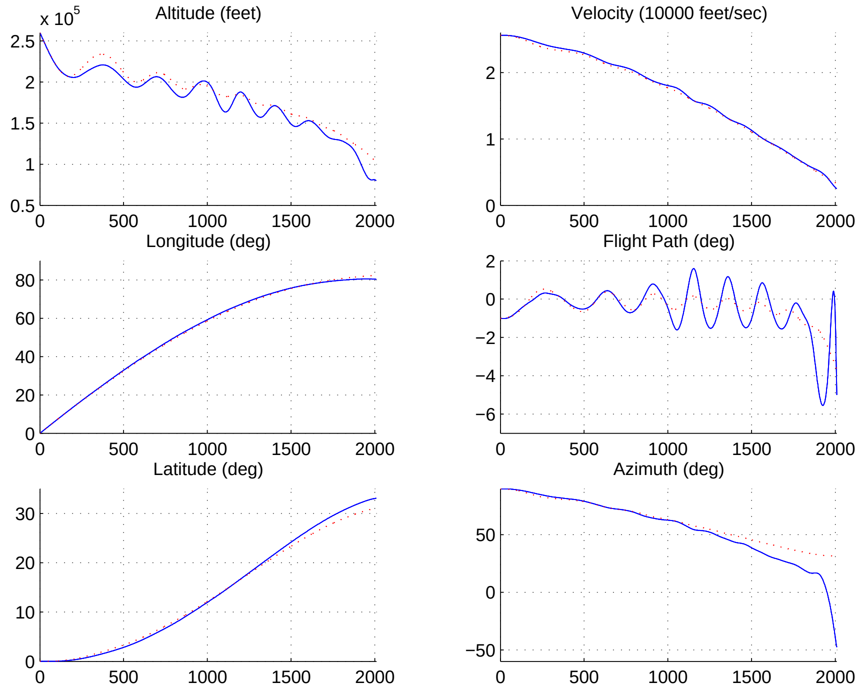


Figure 3: Reentry state variables. Optimal solution (plain) and initialisation (dotted).

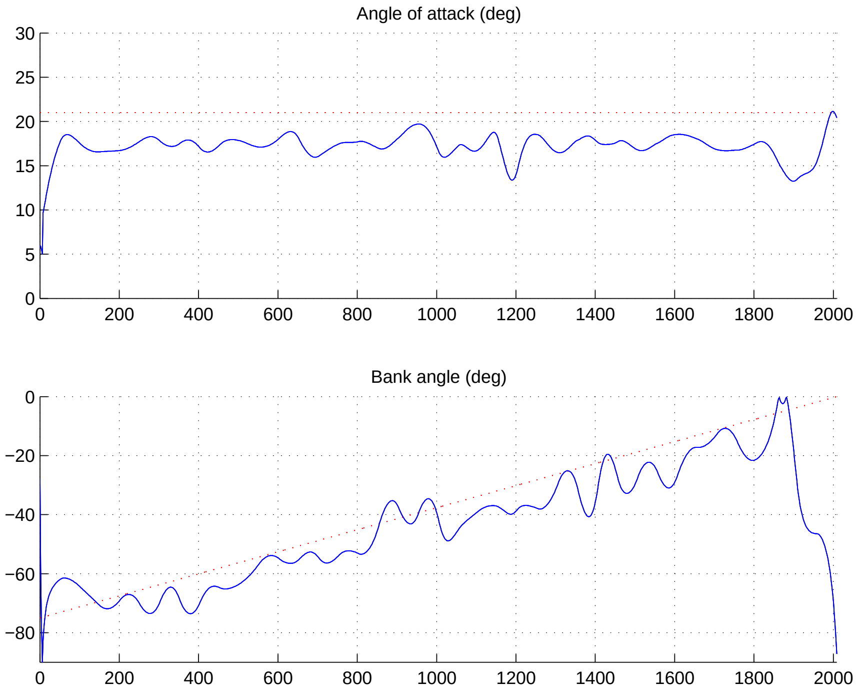


Figure 4: Reentry control variables. Optimal solution (plain) and initialisation (dotted).

Acknowledgement

This work was sponsored by CNES through a ARMINES contract.

References

[1] J. T. Betts, *Practical Methods for optimal control using nonlinear programming*, SIAM, 2001.

[2] A. E. Bryson, *Dynamic optimization*, Addison Wesley, 1999.

[3] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: introductory theory and examples*, Int. J. Control, 61 (1995), pp. 1327–1361.

[4] ——, *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 922–937.

[5] P. Gill, W. Murray, M. Saunders, and M. Wright, *User's Guide for NPSOL 5.0: A Fortran Package for Nonlinear Programming*, Systems Optimization Laboratory, Stanford University, Stanford, CA 94305, 1998.

[6] C. Hargraves and S. Paris, *Direct trajectory optimization using nonlinear programming and collocation*, AIAA J. Guidance and Control, 10 (1987), pp. 338–342.

[7] A. Isidori, *Nonlinear Control Systems*, Springer, New York, 2nd ed., 1989.

[8] P. Martin, R. M. Murray, and P. Rouchon, *Flat systems*, in Proc. of the 4th European Control Conf., Brussels, 1997, pp. 211–264. Plenary lectures and Mini-courses.

[9] M. B. Milam, K. Mushambi, and R. M. Murray, *A new computational approach to real-time trajectory generation for constrained mechanical systems*, in IEEE Conference on Decision and Control, 2000.

[10] M. B. Milam and N. Petit, *Constrained tra jectory generation for a planar missile*, tech. report, California Institute of Technology, Control and Dynamical Systems, 2001.

[11] M. B. Milam, N. Petit, and R. M. Murray, *Constrained trajectory generation for micro-satellite formation flying*, in AIAA Guidance, Navigation and Control Conference, 2001, pp. 328–333.

[12] R. M. Murray, J. Hauser, A. Jadbabaie, M. B. Milam, N. Petit, W. B. Dunbar, and R. Franz, *Software-Enabled Control, Information technology for dynamical systems*, Wiley-Interscience, 2003, ch. Online control customization via optimization-based control, pp. 149–174.

[13] N. Petit, M. B. Milam, and R. M. Murray, *Inversion based constrained trajectory optimization*, in 5th IFAC symposium on nonlinear control systems, 2001.

[14] P. Rouchon, *Necessary condition and genericity of dynamic feedback linearization*, J. Math. Systems Estim. Control, 5 (1995), pp. 345–358.

[15] H. Seywald, *Trajectory optimization based on differential inclusion*, J. Guidance, Control and Dynamics, 17 (1994), pp. 480–487.

A NEW COMPUTATIONAL METHOD FOR OPTIMAL CONTROL OF A CLASS OF CONSTRAINED SYSTEMS GOVERNED BY PARTIAL DIFFERENTIAL EQUATIONS

Nicolas Petit [*] **Mark Milam** [**] **Richard Murray** [**]

[*] *Centre Automatique et Systèmes*
École des Mines de Paris
60, bd St Michel
75272 Paris, France
[**] *Division of Engineering and Applied Science*
Control and Dynamical Systems
California Institute of Technology
Pasadena,CA 91125, USA

Abstract: A computationally efficient technique for the numerical solution of constrained optimal control problems governed by one-dimensional partial differential equations is considered in this paper. This technique utilizes inversion to map the optimal control problem to a lower dimensional space. Results are presented using the Nonlinear Trajectory Generation software package (NTG) showing that real-time implementation may be possible. *Copyright © 2002 IFAC*

Keywords: Optimal Control, Partial Differential Equations, Inversion, Real-time, Computational methods.

1. INTRODUCTION

In the recent years, optimal control problems with systems governed by partial differential equations subject to control and state constraints have been extensively studied. We refer for instance to (Lions, 1971; Bonnans and Casas, 1995; Bergounioux *et al.*, 1998) for necessary optimality conditions for special cases of elliptic problems and to (Maurer and Mittelmann, 2001) for numerical studies. A typical approach to solve these problems is to discretize both the control and the state and use nonlinear programming to solve the resulting optimization problem. In (Maurer and Mittelmann, 2001), this approach was proposed resulting in a large nonlinear programming problem on the order of one thousand variables.

In this paper, we will propose a different methodology. For optimal control of nonlinear ordinary differential equations of the form $\dot{x} = f(x)+g(x)u$, where $\mathbb{R} \ni t \mapsto x \in \mathbb{R}^n$ and $\mathbb{R} \ni t \mapsto u \in \mathbb{R}^m$, we have shown (Milam *et al.*, 2000; Petit *et al.*, 2001) that it is possible and computationally efficient to reduce the dimension of the nonlinear programming problem by using inversion to reduce the number of dynamic constraints, thus eliminating variables, in the problem. Given a particular output, it is generally possible to parameterize a part of the control and a part of the state in terms of this output and its time derivatives. The case of complete parameterization of nonlinear ordinary differential equations is called "flatness" (Fliess *et al.*, 1995; Fliess *et al.*, 1999).

The idea of reducing the dynamic constraints via inversion has been implemented in the Nonlinear Trajectory Generation (NTG) software package (Milam *et al.*, 2000). The outputs of the system are approximated by B-splines and nonlinear programming is used to solve for the coefficients of the B-splines. This software can today be considered as an alternative to the well-established collocation software packages developed using methods described in (Hargraves and Paris, 1987; Seywald, 1994), and (von Stryk and Bulirsch, 1992). Other publications (Milam *et al.*, 2002) deal with the real-time implementation of NTG and thus underlines the importance of the computation-time reduction.

In this paper we propose to extend the "inversion" concept to the field of partial differential equations. In this case the outputs are parameterized by tensor-product B-splines instead of B-splines. B-spline tensor products' partial derivatives can be easily computed, combined and substituted to as many components of the states and the control as possible in both the cost functions and the constraints.

The contribution of our current work is to develop theory and a set of corresponding software tools for the real-time solution of constrained optimal control problems for a class of systems governed by partial differential equations. We think that these set of software tools would be useful in the model predictive and process control communities.

In Section 1 we detail our approach. We apply our proposed methodology to an example from the literature in Section 2. The results show that this methodology is efficient and that solutions of optimal control problems for systems governed by partial differential equations may be computed in real-time using our technique.

2. PROBLEM FORMULATION AND PROPOSED METHOD OF SOLUTION

2.1 *Optimal Control Problem*

Notationally, we use $\mathbb{N} = \{1, 2, 3, ...\}$ to represent the natural numbers and $\mathbb{R}$ to represent the reals. Let Ω be an open set in $\mathbb{R}^2$ and $\Gamma = \bar{\Omega} - \Omega$ its boundary. We denote $\Omega \ni (t, x) \mapsto \phi(t, x)$ the state of the system, $\Omega \ni (t, x) \mapsto u(t, x)$ the control, with $n = \dim \phi$, $m = \dim u$. Let ξ represent the first $(n_t + 1)(n_x + 1)$ partial derivatives of ϕ, with $n_t \in \mathbb{N}$ and $n_x \in \mathbb{N}$

$$\begin{aligned}\xi \doteq (&\phi, \frac{\partial \phi}{\partial t}, \ldots, \frac{\partial^{n_t} \phi}{\partial t^{n_t}},\\ &\frac{\partial \phi}{\partial x}, \frac{\partial^2 \phi}{\partial t \partial x}, \ldots, \frac{\partial^{n_t+1} \phi}{\partial t^{n_t} \partial x},\\ &\cdots\\ &\frac{\partial^{n_x} \phi}{\partial^{n_x} x}, \frac{\partial^{(n_t-1)n_x+1} \phi}{\partial t \partial^{n_x} x}, \ldots, \frac{\partial^{n_t+n_x} \phi}{\partial t^{n_t} \partial^{n_x} x}).\end{aligned}$$

We consider systems that are governed by partial differential equations of the form

$$f(\xi(t, x)) = Bu(t, x) \tag{1}$$

in Ω, where $B \in \mathbb{R}^{n \times m}$ is a matrix with coefficients in $\mathbb{R}$, $f : \mathbb{R}^{n(n_t+1)(n_x+1)} \rightarrow \mathbb{R}^n$ is a nonlinear function.

We desire to find a trajectory of (1) that minimizes the cost functional

$$\min_{(\phi,u)} J(\phi, u) = \int_\Omega L(\xi(t, x), u(t, x)) dx\, dt \tag{2}$$

subject to the domain constraints

$$lb_\Omega \leq S_\Omega(\xi, u) \leq ub_\Omega \tag{3}$$

on Ω and the boundary constraints

$$lb_\Gamma \leq S_\Gamma(\xi, u) \leq ub_\Gamma \tag{4}$$

on Γ, where $L : \mathbb{R}^{n(n_t+1)(n_x+1)+m} \rightarrow \mathbb{R}$, $S_\Omega : \mathbb{R}^{n(n_t+1)(n_x+1)+m} \rightarrow \mathbb{R}^{n_\Omega}$, $S_\Gamma : \mathbb{R}^{n(n_t+1)(n_x+1)+m} \rightarrow \mathbb{R}^{n_\Gamma}$ are nonlinear functions, $n_\Omega \in \mathbb{N}$, $n_\Gamma \in \mathbb{N}$.

We tacitly assume that there exists such an optimal control and refer to (Lions, 1971; Bonnans and Casas, 1995; Bergounioux *et al.*, 1998; Maurer and Mittelmann, 2001) for discussions concerning this important issue.

2.2 *Proposed Methodology of Solution*

There are three components to the methodology we propose. The first is to determine a parameterization (output) such that Equation (1) can be mapped to a lower dimensional space (output space). Once this is done the cost in Equation (2) and constraints in Equations (3) and (4) can also be mapped to the output space. The second is to parameterize each component of the output in terms of an appropriate tensor product B-spline surface. Finally, sequential quadratic programming is used to solve for the coefficients of the B-splines that minimize the cost subject to the constraints in output space.

In most cases, it is desirable to find and output $\Omega \ni (t, x) \mapsto z(t, x) \in \mathbb{R}^p$, $p \in \mathbb{N}$ and a mapping ψ of the form

$$z = \psi(\xi, u) \tag{5}$$

such that (ξ, u) (and thus ϕ) can be completely determined from z and a finite number of its partial derivatives through Equation (1)

$$
\begin{aligned}
(\xi, u) =&\vartheta(\frac{\partial^s z}{\partial t^s}, \frac{\partial^s z}{\partial t^{s-1}\partial x}, \dots, \frac{\partial^s z}{\partial x^s},\\
&\frac{\partial^{s-1} z}{\partial t^{s-1}}, \dots, \dots, \dots, z).
\end{aligned}
$$

Once the output z is chosen, we look for the optimum in a particular functional space: we parameterize each of its components in terms of tensor product B-spline basis functions defined over Ω. These tensor products are only one of many possible choices for basis functions. They are chosen for their flexibility and ease of enforcing continuity between patches of surface. A complete treatment of these functions can be found in (de Boor, 1978). A pictorial representation of one component of an output from an example optimization problem is given in Figure 1 for which $\Omega = (-2, 2) \times (-3, 2)$.

Each component z^l, $l = 1, \dots, p$ of the output z is written in terms of a finite dimensional B-spline surface as

$$
\begin{aligned}
z^l(t, x) =& \sum_{i=1}^{p_t} \sum_{j=1}^{p_x} B_{i,k_t}(t) B_{j,k_x}(x) C^l_{i,j} \qquad (6)\\
& p_t = l_t(k_t - m_t) + m_t \quad \text{and} \qquad (7)\\
& p_x = l_x(k_x - m_x) + m_x \qquad (8)
\end{aligned}
$$

where $\mathbb{R} \ni t \mapsto B_{i,k_t}(t)$ and $\mathbb{R} \ni x \mapsto B_{j,k_x}(x)$ are the B-spline basis functions given by the recursion formula in (de Boor, 1978). In this case we chose $l_t = 5$ and $l_x = 4$ knot intervals in the t and x directions, respectively. The piecewise polynomials in each of the knot intervals will be of order $k_t = 5$ and $k_x = 6$ in the t and x directions, respectively. Smoothness of the piecewise polynomials will be given by the multiplicities $m_t = 3$ and $m_x = 4$ in the t and x directions, respectively. Note that it is also possible to use different parameters k_t, k_x, m_t, m_x for each component of the output. There is a total of $p_t \times p_x = 156$ total coefficients $C^l_{i,j}$ used to define the component z^l of the output in Figure 1.

The breakpoints are a grid ($nbps_t \times nbps_x$) where the boundary and domain constraints will be enforced. There is a similar notion for the integration points. We chose 21 breakpoints in the t direction and 26 breakpoints in the x direction.

After the output has been parameterized in terms of B-spline surfaces, the coefficients $C^l_{i,j}$ of the B-spline basis functions will be found using sequential quadratic programming. This problem is stated as

$$
\begin{aligned}
&\min_{y \in \mathbb{R}^N_c} F(y) \qquad \text{subject to} \quad lb \leq c(y) \leq ub \qquad (9)\\
&\quad \text{where } y = (C^1_{1,1}, C^1_{1,2}, \dots, C^p_{p_t,p_x})\\
&\quad \text{and } N_c = p_t * p_x * p.
\end{aligned}
$$

$F(y)$ is the discrete approximation in output space to the objective in Equation (2). The number of constraints is

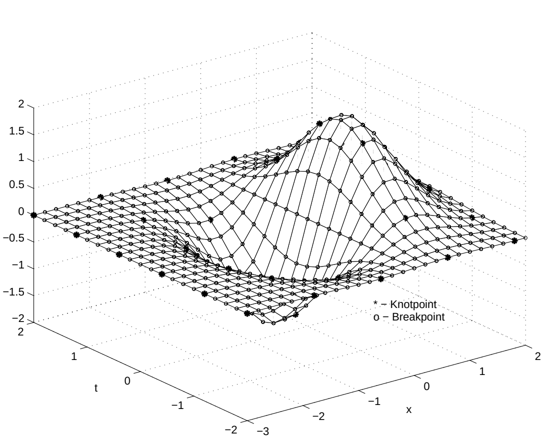


Fig. 1. B-spline Tensor Product Basis Representation

$$
\begin{aligned}
M =& nbps_t * nbps_x * n_\Omega\\
&+ 2 * (nbps_t + nbps_x) * n_\Gamma.
\end{aligned}
$$

The vector $\mathbb{R}^{N_c} \ni y \mapsto c(y) \in \mathbb{R}^M$ contains the constraints mapped to output space from Equations (3) and (4). We will use NPSOL (Gill *et al.*, 1998) as the sequential quadratic programming to solve this new problem.

3. EXAMPLE

We use here one of the example treated in (Maurer and Mittelmann, 2001). It is related to a simplified *Ginzburg-Landau* equation arising in superconductivity.

As before Ω is an open set in $\mathbb{R}^2$ and Γ its boundary. We consider the following nonlinear partial differential with homogeneous Dirichlet boundary condition ($n = \dim \phi = 1$, $m = \dim u = 1$)

$$
\begin{aligned}
-\Delta y - \exp(y) &= u \text{ on } \Omega\\
y &= 0 \text{ on } \Gamma.
\end{aligned}
$$

We look for a control u that minimizes the following cost functional of tracking type

$$
F(y, u) = \frac{1}{2} \|y - y_d\|_{L^2(\Omega)} + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}
$$

where $y_d(t, x) = 1 + 2\,(t(t-1) + x(x-1))$, while satisfying the constraints

$$
\begin{aligned}
y &\leq .185 \text{ on } \Omega\\
1.5 \leq u &\leq 4.5 \text{ on } \Omega.
\end{aligned}
$$

It is clearly possible to parameterize the control using y and its partial derivatives (in this simple case we note $z = y$). Doing so we cast the problem into the following

$$
\begin{aligned}
&\min \frac{1}{2} \|y - y_d\|_{L^2(\Omega)} \\
&\qquad + \frac{\alpha}{2} \|{-\Delta y(x)} - \exp(y)\|_{L^2(\Omega)} \\
&\text{subject to } y = 0 \text{ on } \Gamma \\
&\qquad y \leq .185 \text{ on } \Omega \\
&\qquad 1.5 \leq -\Delta y - \exp(y) \leq 4.5 \text{ on } \Omega.
\end{aligned}
$$

3.1 *Results*

As in (Maurer and Mittelmann, 2001) we choose $\Omega = (0,1) \times (0,1)$, $\alpha = 0.001$. No analytical gradients of the cost and the constraints were provided to NPSOL. Instead, finite difference approximation is used for the gradients. In the future, a function will analytically compute gradients within the NTG software package (it is already the case for ordinary differential equations, not yet for partial differential equations). It is expected to cut down the cpu-time even further (at least by a factor of 2) and increase the accuracy as well.

A set of optimal control and states are plotted in Figure 2. The results of numerical investigations of our approach are detailed in Table 1.

Nomenclature

- N_c: number of coefficients.
- $nbps_t$, $nbps_x$: number of breakpoints in the t and x direction respectively.
- CPU: CPU time (in seconds) on a Pentium-III 733MHZ under Linux Red Hat 6.2 .
- ig: initial guess for coefficients, where 0 means that zeros are used as an initial guess. If the solution from another run with less breakpoints was used for ig then the name of the run is specified, e.g. $ig = (20, 20)$ means the initial guess is the solution to the run with the same degrees and multiplicities but with $(20, 20)$ breakpoints.
- Objective: objective value at the optimum
- k: degree of the polynomials, $k_t = k_x = k$ in this example.
- m: multiplicity of the knotpoints, $m_t = m_x = m$ in this example.
- l: number of intervals, $m_t = m_x = m$ in this example.
- err_u: absolute violation of the constraint on the control.
- err_y: absolute violation of the constraint on the state.

The results in Table 1 show that it is possible to compute fast and with a reasonable accuracy a solution of the optimal control problem. There is also the trade off of a more precise solution for larger computation times. The choice of initial guess also influences computation time.

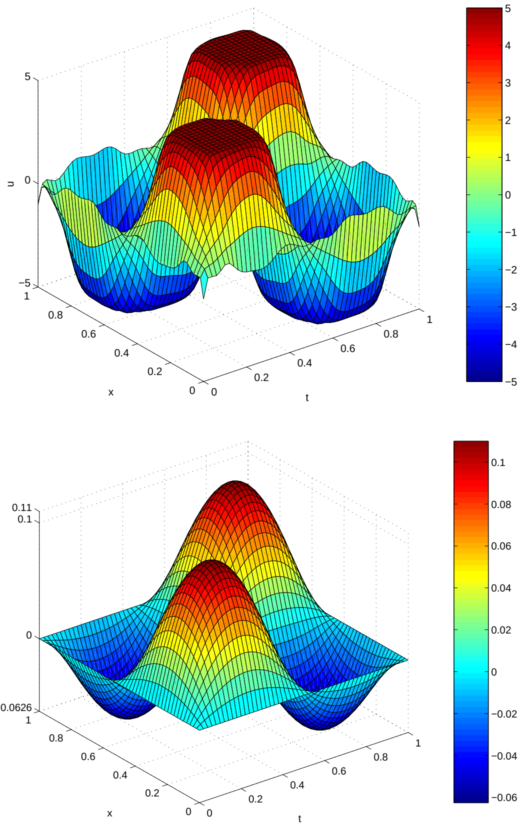


Fig. 2. Example. Optimal control (up) and state (down).

In the numerical experiments presented above, we evaluate afterwards the cost associated with each set of solution coefficients by an adaptative Lobatto quadrature with accuracy to 8 digits. Thus the costs given here are not the costs evaluated by the nonlinear programming solver but more accurate evaluations of them. Similarly, the absolute violations of the contraints err_u and err_y are evaluated afterwards, using a very large number of breakpoints in both directions.

We reproduce in Table 2 some numerical results from (Maurer and Mittelmann, 2001). It is important to notice that the evaluation of the objective in their approach is different. The quantities y and u are evaluated only at the grid points. The cost is evaluated by the nonlinear programming solver and is not as accurate as the results in Table 1. Asymptotically the two approaches seem to converge to the same value that is not known. In terms of computation time, it is to be noted that the results by Maurer and Mittelmann (in Table 2) were obtained on a 450MHz Pentium-II with a different nonlinear programming solver than the one we use. This has to be taken into account when comparing their absolute value.

N_c	$nbps_t$,$nbps_x$	CPU	ig	Objective	k	m	l	err_u	err_y
64	(10,10)	4	0	0.1112665	5	4	4	5.6e-1	2.5e-3
64	(15,15)	7	0	0.1117173	5	4	4	9.8e-2	0
64	(20,20)	18	0	0.1117267	5	4	4	1.6e-1	0
64	(40,40)	62	0	0.1118501	5	4	4	2.2e-2	0
64	(60,60)	176	0	0.1118300	5	4	4	4.4e-3	0
64	(80,80)	312	0	0.1118253	5	4	4	1.2e-3	0
64	(100,100)	560	0	0.1118291	5	4	4	2.7e-3	0
64	(40,40)	40	(20,20)	0.1118501	5	4	4	2.2e-2	0
64	(80,80)	163	(20,20)	0.1118253	5	4	4	1.2e-3	0
64	(80,80)	334	(40,40)	0.1118254	5	4	4	1.1e-3	0
64	(100,100)	261	(20,20)	0.1118291	5	4	4	2.7e-3	0
144	(10,10)	30	0	0.1104154	6	4	4	4.7e-1	1.0e-2
144	(15,15)	61	0	0.1105299	6	4	4	1.6e-1	2.2e-3
144	(20,20)	90	0	0.1105640	6	4	4	9.2e-2	8.0e-5
144	(40,40)	464	0	0.1106407	6	4	4	1.3e-2	1.0e-5
144	(60,60)	998	0	0.1106456	6	4	4	4.0e-2	6.8e-5
144	(80,80)	1674	0	0.1106465	6	4	4	2.2e-3	1.1e-4
144	(100,100)	2670	0	0.1106481	6	4	4	1.4e-3	3.3e-5
144	(80,80)	924	(20,20)	0.1106465	6	4	4	2.2e-3	1.5e-5
400	(80,80)	15810	0	0.1102986	6	4	8	3.2e-3	0
400	(90,90)	4910	(80,80)	0.1102987	6	4	8	2.2e-3	0

Table 1. Numerical results with the NTG approach.

$gridpoints$	CPU	Objective
2401	131	0.110242
9801	2257	0.110263
39601	42644	0.110269

Table 2. Numerical results by Maurer and Mittelmann.

3.2 *Remarks*

It is important to realize that the methodology we propose produces exact solutions. Once the solution coefficients are determined, the control can be exactly evaluated at any desired point without any refinement of the grid by combinations of exact partial derivatives of the output. Some constraints may be slightly violated in between breakpoints. Asymptotically though, as the number of breakpoints increases the violation experimentally goes to zero.

4. CONCLUSION

The idea in this paper is the use of inversion to eliminate variables from the optimal control problem before using a nonlinear programming solver. To do so, partial derivatives of the output (the parameterizing quantities) are needed. In this context tensor product B-Splines are a useful representation. Numerical results suggest that this methodology is efficient and that fast resolution of such problems can be achieved. Real-time implementation on a reasonably fast process seems close at hand. One can consider for instance a tubular polymerization reactor governed by a one dimensional hyperbolic equation with reaction and heat exhange terms (see for instance (Westerterp *et al.*, 1988)), its time scale is typically 5 minutes which is long enough for receding horizon control purpose.

The methodology presented here can be used in various situations including the following problem that we detail to show the generality of our approach. In (Heinkenschloss and Sachs, 1994) the authors expose the following solid-liquid phase transitions control problem. The model consists of two non-linear parabolic equations in Ω subset of $\mathbb{R}^2$.

$$\begin{aligned} T_t + \frac{1}{2}\varphi_t &= kT_{xx} + u \\ \tau\varphi_t &= \xi^2\varphi_{xx} + g(\varphi) + 2T. \end{aligned}$$

In this model the state is $\phi = (\varphi, T) \in \mathbb{R}^2$ (phase function and temperature of the medium, $n = 2$), u is the control ($m = 1$), k, τ, ξ^2 are given parameters, and g is a given nonlinear function. A certain desired phase function φ_d and a temperature u_d are given. An interesting optimal control problem is to minimize the following objective function

$$\begin{aligned} J =& \frac{\alpha}{2}\left\|T - T_d\right\|_{L^2(\Omega)} + \frac{\beta}{2}\left\|\varphi - \varphi_d\right\|_{L^2(\Omega)} \\ &+ \frac{\gamma}{2}\left\|u\right\|_{L^2(\Omega)}. \end{aligned}$$

Inversion can be used in this problem. Both T and u express in terms of the output φ and its partial derivatives

$$
\begin{aligned}
T =& \frac{1}{2}\left(\tau\varphi_t - \xi^2\varphi_{xx} - g(\varphi)\right) \doteq h_1(\varphi, \varphi_t, \varphi_{xx}) \\
u =& \frac{1}{2}\left(\tau\varphi_{tt} - \xi^2\varphi_{txx} - \varphi_t\dot{g}(\varphi)\right) + \frac{1}{2}\varphi_t \\
& - \frac{k}{2}\left(\tau\varphi_{txx} - \xi^2\varphi_{xxxx} - \varphi_{xx}\dot{g}(\varphi) - \varphi_x^2\ddot{g}(\varphi)\right) \\
& \doteq h_2(\varphi, \varphi_t, \varphi_{tt}, \varphi_{txx}, \varphi_x, \varphi_{xx}, \varphi_{xxxx}).
\end{aligned}
$$

After substitution in the cost function, the functional to minimize is

$$
\begin{aligned}
& J'(\varphi) = \\
& \frac{\alpha}{2}\left\|h_1(\varphi, \varphi_t, \varphi_{xx}) - T_d\right\|_{L^2(\Omega)} \\
& + \frac{\beta}{2}\left\|\varphi - \varphi_d\right\|_{L^2(\Omega)} \\
& + \frac{\gamma}{2}\left\|h_2(\varphi, \varphi_t, \varphi_{tt}, \varphi_{txx}, \varphi_x, \varphi_{xx}, \varphi_{xxxx})\right\|_{L^2(\Omega)}.
\end{aligned}
$$

We are currently numerically investigating examples of this kind and believe that for such systems the computation time reduction induced by the use of inversion will be very attractive.

5. REFERENCES

Bergounioux, M., M. Haddou, M. Hintermuller and K. Kunisch (1998). A comparision of interior point methods and a Moreau-Yosida based active set strategy for constrained optimal control problems. Preprint MAPMO 98 - 15 Université d'Orléans.

Bonnans, F. and E. Casas (1995). An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities. *SIAM J. Control & Opt.* **33**(1), 274–298.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag.

Fliess, M., J. Lévine, Ph. Martin and P. Rouchon (1995). Flatness and defect of nonlinear systems: introductory theory and examples. *Int. J. Control* **61**(6), 1327–1361.

Fliess, M., J. Lévine, Ph. Martin and P. Rouchon (1999). A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems. *IEEE AC* **44**, 922–937.

Gill, P., W. Murray, M. Saunders and M. Wright (1998). *User's Guide for NPSOL 5.0: A Fortran Package for Nonlinear Programming.* Systems Optimization Laboratory. Stanford University, Stanford, CA 94305.

Hargraves, C. and S. Paris (1987). Direct trajectory optimization using nonlinear programming and collocation. *AIAA J. Guidance and Control* **10**, 338–342.

Heinkenschloss, M. and E. W. Sachs (1994). *Numerical solution of a constrained control problem for a phase field model.* Vol. 118 of *International Series of Numerical Mathematics.* Birkhäuser.

Lions, J.-L. (1971). *Optimal control of systems governed by partial differential equations.* Springer-Verlag.

Maurer, H. and H. Mittelmann (2001). Optimization techniques for solving elliptic control problems with control and state constraints. Part 2: Distributed control. *Comp. Optim. Applic.* **18**, 141–160.

Milam, M. B., K. Mushambi and R. M. Murray (2000). A new computational approach to real-time trajectory generation for constrained mechanical systems. In: *IEEE Conference on Decision and Control.*

Milam, M. B., R. Franz and R. M. Murray (2002). Real-time constrained trajectory generation applied to a flight control experiment. In: *Proc. of the IFAC World Congress.*

Petit, N., M. B. Milam and R. M. Murray (2001). Inversion based constrained trajectory optimization. In: 5*th IFAC symposium on nonlinear control systems.*

Seywald, H. (1994). Trajectory optimization based on differential inclusion. *J. Guidance, Control and Dynamics* **17**(3), 480–487.

von Stryk, O. and R. Bulirsch (1992). Direct and indirect methods for trajectory optimization. *Annals of Operations Research* **37**, 357–373.

Westerterp, K. R., W. P. M. Van Swaaij and A. A. C. M. Beenackers (1988). *Chemical Reactor Design and Operation.* Wiley, John & Sons, Inc.