Pilotage dynamique de l'énergie du bâtiment par commande optimale sous contraintes utilisant la pénalisation intérieure

Paul Malisani

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

Ecole doctorale n°432: Sciences des Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École Nationale Supérieure des Mines de Paris

Spécialité "Mathématique et Automatique"

présentée et soutenue publiquement par

Paul MALISANI

le 21 septembre 2012

Pilotage dynamique de l'énergie du bâtiment par commande optimale sous contraintes utilisant la pénalisation intérieure

Directeur de thèse: **Nicolas PETIT**
Maître de thèse: **François CHAPLAIS**

Jury

M. Knut GRAICHEN, Professeur, Inst. of Meas., Control and Microtech., Univ. Ulm	Rapporteur, Pres.
M. Emmanuel TRELAT, Professeur, Laboratoire J.-L.Lions, Univ. Paris 6	Rapporteur
Mme. Anne-Sophie COINCE, Ingénieur-chercheur, Dept. EnerBat, EDF R&D	Examinatrice
M. Pierre MARTINON, Chargé de recherche, INRIA, Ecole Polytechnique	Examinateur
M. Bruno PEUPORTIER, Maître de recherche, C.E.P., MINES-ParisTech	Examinateur
M. Nicolas PETIT, Professeur, C.A.S., MINES-ParisTech	Examinateur
M. François CHAPLAIS, Ingénieur de recherche, C.A.S., MINES-ParisTech	Examinateur

MINES ParisTech
Centre Automatique et Systèmes
60 Boulevard Saint-Michel 75272 Paris Cedex 06

THÈSE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

Ecole doctorale n°432: Sciences des Métiers de l'Ingénieur

ParisTech Doctorate

T H E S I S

to obtain

Doctor's degree from the Ecole Nationale Supérieure des Mines de Paris

Speciality "Mathematics and Control"

defended in public by

Paul MALISANI

September, 21th, 2012

Dynamic control of energy in buildings using constrained optimal control by interior penalty

Advisor: **Nicolas PETIT**
Co-advisor: **François CHAPLAIS**

Committee

Mr. Knut GRAICHEN, Professeur, Inst. of Meas., Control and Microtech., Univ. Ulm	Referee, Pres.
Mr. Emmanuel TRELAT, Professeur, Laboratoire J.-L.Lions, Univ. Paris 6	Referee
Mrs. Anne-Sophie COINCE, Ingénieur-chercheur, Dept. EnerBat, EDF R&D	Examiner
Mr. Pierre MARTINON, Chargé de recherche, INRIA, Ecole Polytechnique	Examiner
Mr. Bruno PEUPORTIER, Maître de recherche, C.E.P., MINES-ParisTech	Examiner
Mr. Nicolas PETIT, Professeur, C.A.S., MINES-ParisTech	Examiner
Mr. François CHAPLAIS, Ingénieur de recherche, C.A.S., MINES-ParisTech	Examiner

THÈSE

MINES ParisTech
Centre Automatique et Systèmes
60 Boulevard Saint-Michel 75272 Paris Cedex 06

REMERCIEMENTS

Je voudrais tout d'abord remercier très sincèrement mon maître de thèse M. François Chaplais de m'avoir fait confiance pour cette thèse et de m'avoir toujours poussé à être plus précis et rigoureux.

Je remercie ensuite mon directeur de thèse M. Nicolas Petit, dont l'enthousiasme contagieux et les solides conseils scientifiques et humains ont toujours été une source inépuisable de motivation.

Je souhaite remercier Emmanuel Trélat ainsi que Knut Graichen qui ont accepté d'être les rapporteurs de cette thèse. Je remercie également MM. Pierre Martinon, Bruno Peuportier et Mme Anne-Sophie Coince qui m'ont fait l'honneur de participer au jury de ma soutenance.

Je remercie Laurent Praly pour ses commentaires sur mes présentations qui ont notablement amélioré la clarté de mes exposés. Je remercie aussi l'ensemble du Centre Automatique et Systèmes pour le fantastique environnement de travail que ce centre offre aux jeunes chercheurs. Je remercie mes camarades docteurs et doctorants: Al-Kassem, Caroline, Delphine, Eric, Hadis, Mathieu, Nadège, Pierre-Jean, Stéphane D., Stéphane T., Thomas, Zaki.

Je remercie également mes amis Julien S., Florent, Guillaume, Nicolas, Maximilien, Pacôme, David, Julien F. et Marc-Alexandre sur lesquels je peux toujours compter.

Je remercie ma famille et particulièrement mes parents Serge et Gabrielle pour m'avoir toujours encouragé et conseillé, leur force de travail restera toujours un exemple à suivre.

Enfin, je dédie ma thèse à Lyse pour m'avoir soutenu et supporté sans relâche durant ces trois années. Sa constante joie de vivre rend merveilleuse la vie à ses côtés.

Contents

Chapter 1

Introduction (version française)

Contents

1.1 Contexte: l'optimisation énergétique des bâtiments d'habitation individuelle

Malgré les nombreuses politiques mises en place au niveau européen depuis plus de 30 ans en faveur des économies d'énergies, la croissance démographique et l'augmentation de la part des énergies renouvelables intermittentes dans le mélange (mix) électrique menacent aujourd'hui l'équilibre entre la production et la consommation d'électricité. De manière critique, les périodes de pointe de consommation sont difficiles à gérer. Dans ce contexte, les gestionnaires d'énergie (producteurs, distributeurs, en lien avec les instances gouvernementales) utilisent systématiquement l'augmentation des moyens de production et des capacités d'acheminement (réseau) pour maintenir cet équilibre. Or, les moyens de production de pointe présentent des inconvénients: *i)* un risque à l'investissement plus élevé que les autres types de centrales électriques [HC11] et *ii)* un coût important et de lourdes émissions de CO_2,

puisqu'ils utilisent des énergies fossiles. Une alternative à cette solution consiste à développer et à mettre en oeuvre des programmes de *gestion de la demande*.

Le principe est le suivant: lisser le profil temporel de la demande, en décalant la consommation des heures de pointe vers les heures creuses. Ce décalage peut être obtenu par une incitation des consommateurs (via un retour d'information au client ou une utilisation de signal-prix sur l'électricité), ou encore en prenant le contrôle à distance des appareils. Les secteurs industriels et tertiaires sont naturellement des cibles de choix pour cette stratégie de lissage car les puissances consommées par chaque client sont très importantes. En retour, les clients concernés peuvent être rétribués par une compensation financière de la réduction (voire l'arrêt) de leur activité. Ce schéma est connu comme "Demand side biding" [AES08].

Une autre solution est de se tourner vers le secteur des particuliers, et le secteur résidentiel notamment. De manière intéressante, la gestion de la demande dans le secteur résidentiel n'est pas encore très développée et, pourtant, ce secteur est fortement consommateur. A titre d'exemple, il représente environ 29 % [BA07] de la consommation d'électricité totale en Europe. Ce manque de développement est dû à la multitude des clients individuels de ce secteur, qui sont plus difficiles à gérer que les "grands comptes" des secteurs industriels et tertiaire précédemment cités. Récemment, avec le développement des technologies de communication et des réseaux intelligents ("smart grid" [KR10, TYO+12]), on a vu apparaître plusieurs projets de recherche et d'expérimentation visant à tester des implémentations à large échelle de la gestion de la demande dans le secteur résidentiel [Fro07, PWMK07]. Ces programmes sont connus sous le nom de "Demand Response".

C'est précisément dans le cadre de ces programmes que s'inscrivent les travaux de cette thèse. Pour mettre au point des stratégies de "Demand Response" pour le secteur résidentiel, une analyse quantitative de l'aptitude des usages électriques domestiques à servir le gestionnaire d'énergie est nécessaire. Cette évaluation, que nous allons faire grâce à des outils de commande optimale sous contraintes d'état et de commande, est à réaliser dans un contexte d'emploi (scénarios) et en respectant les souhaits opérationnels suivants (comme décrit dans [DS11]):

- *Utilisation de la production renouvelable électrique locale au niveau de chaque habitat.*
- *Introduction de tarifs innovants* représentatifs de l'état d'utilisation du réseau électrique par exemple le "Real-time-pricing" ou le "Critical peak pricing"[1].
- *Contrôle optimisé du fonctionnement individuel des équipements* de façon à limiter la puissance utilisée pendant certaines périodes[2].

Parmi les systèmes pouvant servir à atteindre les objectifs de la gestion active de la demande, plusieurs sont des systèmes dynamiques (ballon d'eau chaude, chauffage

[1] Ces types de tarif doivent encourager le lissage de la courbe de charge totale ou des pointes.

[2] Par exemple, le chauffage des ballons d'eau chaude sanitaire asservis heures pleines/heures creuses.

à accumulation, batterie, inertie du bâtiment...) ou agissent sur un système dynamique (chauffage à effet Joule, pompes à chaleur...). On trouvera en Annexe A, une brève description technologique de ces systèmes.

1.2 Commande optimale: un outil pour la quantification des gisements potentiels de flexibilité

La quantification de la flexibilité apportée par chaque système (convecteurs, panneaux rayonnants, chauffage à accumulation, pompe à chaleur, eau chaude sanitaire à accumulation, notamment) dans le but de répondre à un des objectifs de la gestion de la demande consiste à répondre à la question suivante:

> "Quel est le service *maximal* que peut rendre un *système* pour atteindre un des *objectifs* de la gestion de la demande sous *contraintes* de maintenir le confort?"

Le confort est une notion qui couvre le confort thermique, le fonctionnement des appareils, etc. La réponse à cette question peut être apportée assez naturellement en résolvant un problème de commande optimale sous contraintes d'état et de commande. On donne ci-dessous trois exemples, chacun répondant à un des objectifs de la gestion active de la demande.

1.2.1 Exemple 1: déphasage de la production photovoltaïque

Considérons un particulier dont l'habitat est équipé de panneaux photovoltaïques et d'une batterie permettant de stocker tout ou partie de la puissance produite par les panneaux solaires $PV(t)$ qui est variable en fonction de l'heure. L'énergie est intégralement vendue sur le réseau au prix du marché de l'électricité. Une question naturelle est: "quel est le revenu maximal que peut apporter le déphasage (stockage pour revente ultérieure) de la production photovoltaïque?". Pour y répondre, considérons une année de référence pour laquelle la courbe de prix de l'électricité sur le marché prix(t) est entièrement connu. On peut chercher à comparer le revenu obtenu par le client quand il revend sa production directement

$$\int_0^T \text{prix}(t) PV(t) dt \tag{1.1}$$

avec la quantité suivante:

$$\max_{u_1,u_2} \int_0^T \text{prix}(t) \Big[PV(t) - (1-u_1(t)) \max\{P_{\max}, PV(t)\} + r\left(x(t), u_2(t)\right) u_2(t) \Big] dt \tag{1.2}$$

où $P_{\max}$ est la puissance de charge maximale de la batterie, $r\left(x(t), u_2(t)\right)$ est le rendement de la batterie, $x(t)$ son état de charge, u_1 le pourcentage de puissance photovoltaïque stockée dans la batterie, u_2 la puissance de décharge, u_1 et u_2 sont les commandes du système. Cette optimisation se fait sous la contrainte de dynamique suivante:

$$\dot{x}(t) = f(x(t), u_1(t), u_2(t))$$

En pratique, la capacité de la batterie et la puissance maximale de décharge étant bornées, l'optimisation doit se faire sous les contraintes suivantes

$$\begin{aligned} x(t) &\in [0, C_{\max}], \quad \forall t \in [0, T] \\ u_1(t) &\in [0, 1], \quad \forall t \in [0, T] \\ u_2(t) &\in [0, P_{\max}], \quad \forall t \in [0, T] \end{aligned}$$

avec, une contrainte sur la batterie $x(0) = x(T)$ exprimant le fait qu'à la fin de la période considérée la batterie est chargée comme au début de la période. En comparant la valeur des deux intégrales (1.1) et (1.2), on obtient le gain *maximal* sur un an que peut apporter le déphasage d'une partie de la production photovoltaïque au propriétaire de l'installation. Un tel exemple est traité, en utilisant les méthodes proposées dans cette thèse, en Annexe B.1 sur un horizon temporel $[0, T]$ d'une semaine.

1.2.2 Exemple 2: efficacité d'un tarif innovant

Un des moyens de juger de l'efficacité d'un tarif sur la "demand response" est de regarder quelle est la gestion optimale (vis-à-vis du tarif) d'énergie par un particulier ayant des besoins de chauffage (convecteurs électriques) et de préparation d'eau chaude sanitaire (ballon d'Eau Chaude Sanitaire (ECS) par accumulation). L'optimisation vise ici à minimiser la facture (hors abonnement). On désigne $\text{prix}(t)$ le tarif vu par le client. Le problème posé consiste à résoudre:

$$\min_{u_1, u_2} \left[\int_0^T \text{prix}(t) \left(u_1(t) + u_2(t) \right) dt \right]$$

où u_1 et u_2 sont les commandes, correspondant aux consommation de chauffage et de préparation d'eau chaude, respectivement, sous les contraintes dynamiques induites par l'inertie thermique du bâtiment et du ballon d'eau chaude

$$\begin{aligned} \dot{x}_{\text{bat}}(t) &= f(x_{\text{bat}}(t), u_1(t)) \\ \dot{x}_{\text{ecs}}(t) &= g(x_{\text{ecs}}(t), u_2(t)) \end{aligned}$$

Le chauffage devant garantir le confort des habitants, et en notant $h(x_{\text{bat}})$ la température à l'intérieur du bâtiment, le problème d'optimisation se fait sous la contrainte:

$$h(x_{\text{bat}}(t)) \in [T^{\text{bat}}_{\min}(t), T^{\text{bat}}_{\max}(t)], \quad \forall t \in [0, T]$$

La température de la couche supérieure du ballon d'eau chaude $l(x_{\text{ecs}})$ est également soumise à des contraintes de fonctionnement :

$$l(x_{\text{ecs}}(t)) \in [T^{\text{ecs}}_{\min}(t), T^{\text{ecs}}_{\max}(t)], \quad \forall t \in [0, T]$$

Enfin, les appels de puissance étant limités par les émetteurs, on a:

$$\begin{aligned} u_1(t) &\in \left[0, P^{\text{chauf}}_{\max}\right], \quad \forall t \in [0, T] \\ u_2(t) &\in \left[0, P^{\text{ecs}}_{\max}\right], \quad \forall t \in [0, T] \end{aligned}$$

Un tel exemple est traité, en utilisant les méthodes proposées dans cette thèse, en Annexe B.2 sur sur un horizon temporel $[0, T]$ d'une semaine.

1.2.3 Exemple 3: faisabilité d'un effacement du chauffage en période de pointe

Pour caractériser la faisabilité de l'effacement d'un particulier en période de pointe tout en maintenant un certain confort dans l'habitation (en utilisant l'inertie thermique du bâtiment), on peut résoudre le problème suivant

$$\min_u \left[\int_{\text{pointe}} u(t)dt \right]$$

où u, la commande, est la consommation électrique pour le chauffage, sous les contraintes

$$\begin{aligned} \dot{x}_{\text{bat}}(t) &= f(x_{\text{bat}}(t), u(t)) \\ h(x_{\text{bat}}(t)) &\in \left[T^{\text{bat}}_{\min}(t), T^{\text{bat}}_{\max}(t) \right] \\ u(t) &\in \left[0, P^{\text{chauf}}_{\max}(t) \right] \end{aligned}$$

pour tout $t \in [0, T]$. La valeur optimale du critère indique la faisabilité des effacements : si la valeur est nulle cela signifie qu'il est possible de ne consommer aucune énergie pendant les périodes de pointe, il est donc possible de réaliser un effacement total de la charge en période de pointe. Si la valeur n'est pas nulle, alors il n'est pas possible de ramener la consommation en heures de pointe à zéro tout en respectant les contraintes, il n'est donc pas possible de réaliser un effacement complet.

Dans le cas où il est possible de réaliser un effacement complet, on peut chercher à trouver la stratégie (permettant d'effacer la consommation de pointe) la plus économe. On cherche alors à résoudre

$$\min_u \left[\int_0^T u(t)dt \right]$$

sous les contraintes

$$\begin{aligned} \dot{x}_{\text{bat}}(t) &= f(x_{\text{bat}}(t), u(t)) \\ h(x_{\text{bat}}(t)) &\in \left[T^{\text{bat}}_{\min}(t), T^{\text{bat}}_{\max}(t) \right] \\ u(t) &\in \left[0, P^{\text{chauf}}_{\max} \right], \quad \text{hors pointe} \\ u(t) &= 0 \quad \text{sinon} \end{aligned}$$

Ces quelques problèmes sont des exemples introductifs simples aux problèmes de commande optimale qu'on souhaite pouvoir traiter de manière plus générale.

Nous exposons maintenant les principes et méthodes de la commande optimale avant d'annoncer les contributions de cette thèse.

1.3 Introduction à la théorie de la commande optimale (sans contraintes d'état)

Dans cette thèse, et comme l'ont illustré les exemples précédents, on va s'intéresser au problème général de commande optimale s'écrivant sous la forme suivante

$$\min_{u\in\mathcal{U}} \int_0^T \ell(x(t),u(t))dt \tag{1.3}$$

où ℓ est une fonction à valeur réelle régulière de ses arguments, sous la contrainte de dynamique suivante:

$$\dot{x}^u(t) = f(x^u(t),u(t)) \quad ; \quad x(0)=x_0 \tag{1.4}$$

correspondant à une représentation sous forme d'état d'un système dynamique. Nous laissons *pour l'instant* de côté les contraintes d'état[3]. Sur l'horizon de temps $[0,T]$, T fixé sans perte de généralité, on peut agir sur l'état du système x à travers la variable de commande u, qu'on peut choisir dans un ensemble $\mathcal{U}$, que nous précisons ici sans perte de généralité, sous-ensemble restreint de L^∞

$$\mathcal{U} = \{u\in L^\infty([0,T],\mathbb{R}^m) \text{ t.q. } u(t)\in\mathcal{C} \text{ p.p.t. } t\in[0,T]\}$$

avec $\mathcal{C}$ un ensemble convexe fermé borné d'intérieur non vide de $\mathbb{R}^m$. Le problème de commande optimale consiste à trouver la commande u et l'état associé x^u solution de l'équation différentielle (1.4) minimisant le critère intégral (1.3). De nombreuses extensions et variantes sont possibles: temps final libre, coût final, définition d'une cible, saut de dynamique, etc., voir [BH69, HT11].

1.3.1 Caractérisation des solutions

Pour résoudre un tel problème de commande optimale, deux grandes approches sont usuellement considérées : le principe du minimum de Pontryaguine et le principe de programmation dynamique de Bellman. Pour présenter ces théories, introduisons d'abord l'Hamiltonien $H:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^n\mapsto\mathbb{R}$,

$$H(x(t),u(t),p(t)) \triangleq \ell(x(t),u(t)) + p(t)^t f(x(t),u(t))$$

1.3.1.1 Conditions nécessaires d'optimalité: principe du minimum de Pontryaguine (PMP)

Le principe du minimum [PBGM62, Tré08] donne une condition nécessaire d'optimalité. Si $(u,x)\in\mathcal{U}\times W^{1,\infty}([0,T],\mathbb{R}^n)$[4] est une solution optimale du

[3] Le lecteur pourra trouver un exposé d'extensions, que nous n'utiliserons pas ici, aux techniques présentées ci-dessous aux cas avec contraintes d'état [HSV95]

[4] avec, classiquement [Ada75], $W^{1,\infty}([0,T],\mathbb{R}^n) \triangleq \{x\in L^\infty([0,T],\mathbb{R}^n) \text{ t.q. } \dot{x}\in L^\infty([0,T],\mathbb{R}^n)\}$

problème (1.3) alors il existe $p \in W^{1,\infty}([0,T],\mathbb{R}^n)$ appelé état adjoint, tel que, presque partout sur $[0,T]$, on a:

$$\begin{aligned} \dot{x}^u(t) &= f(x^u(t),u(t)) \\ x(0) &= x_0 \\ \dot{p}(t) &= -\frac{\partial}{\partial x}H(x(t),u(t),p(t)) \\ p(T) &= 0 \\ u(t) &\in \arg\min_{v\in\mathcal{C}} H(x(t),v,p(t)) \end{aligned}$$

1.3.1.2 Programmation dynamique

La deuxième approche est basée sur le principe de programmation dynamique de Bellman [Bel57] et est née au début des années 60. La fonction valeur $\mathcal{J}$ du problème définie par:

$$\mathcal{J}(\xi,t) \triangleq \inf_{(u,x)} \left\{ \int_t^T \ell(x(s),u(s))ds \; : \; \dot{x}(s) = f(x(s),u(s)) \text{ p.p. } s\in[t,T],\, x(t)=\xi,\, u(s)\in\mathcal{C} \right\}$$

est solution d'une équation aux dérivées partielles non linéaire appelée équation de Hamilton-Jacobi-Bellman (HJB)

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial t}(\xi,t) &+ \inf_{v\in\mathcal{C}} H(v,\xi,\frac{\partial \mathcal{J}}{\partial \xi}(\xi,t)) = 0 \qquad (\xi,t)\in\mathbb{R}^n\times(0,T) \\ \mathcal{J}(\xi,T) &= 0 \end{aligned}$$

Cette condition d'optimalité présente l'avantage d'être nécessaire et suffisante. Cependant, pour des raisons de temps et calcul et d'encombrement de mémoire, la méthode de programmation dynamique ne permet pas, en général, de calculer des solutions optimales sur des horizons de temps importants avec une dimension d'état supérieure à 3 [Bel57]. Néanmoins, de nombreux travaux ont apporté des réponses sur des cas de dimension supérieure, notamment dans le domaine spatial [ABZ12].

1.3.2 Méthodes numériques de résolution

1.3.2.1 Méthodes directes

Les méthodes directes, qui sont très couramment utilisées, utilisent une discrétisation des équations du problème pour le ramener à un problème de programmation non linéaire (NLP), c'est-à-dire à un problème d'optimisation non linéaire en dimension finie. Cette approche a par exemple été utilisée avec succès dans les références suivantes [BCM98, BMDP02, Bha06, CP05, JLW03, HP87, KM04, LS99, PMM01, RF04, Vic98, Wri93, YGFDD05].

L'avantage des méthodes directes est qu'elles sont faciles à implémenter et réputées relativement robustes à l'initialisation. Elles ont, en général, la complexité des algorithmes de résolution de la NLP qu'elles emploient, souvent $O(N^3)$ où N est la dimension du problème discrétisé. Elles sont capables de traiter des problèmes avec un grand nombre de variables d'état. Cependant leur précision est en général limitée par la précision de la discrétisation.

1.3.2.2 Méthodes indirectes

Méthodes de tir Les méthodes de tir exploitent les conditions d'optimalité données par le PMP. Sous certaines hypothèses, voir en particulier [AS04], le PMP permet d'exprimer la commande comme une fonction de l'état et de l'état adjoint:

$$u(t) = \Gamma(x(t), p(t)) \quad t \in [0, T]$$

Les conditions nécessaires d'optimalité se résument alors à résoudre les $2n$ équations différentielles sur x et p formant un problème aux deux bouts puisqu'on a une condition initiale sur x et une condition finale sur p. L'idée de l'algorithme de tir est d'introduire une inconnue, la valeur initiale de l'état adjoint p_0, et de considérer la fonction de tir qui à p_0 associe la condition finale $p(T)$, où (x, p) est solution du problème de Cauchy sur [0,T].

$$\begin{aligned} \dot{x}(t) &= f(x(t), \Gamma(x(t), p(t))), & x(0) &= x_0 \\ \dot{p}(t) &= -\frac{\partial}{\partial x} H(x(t), \Gamma(x(t), p(t)), p(t)), & p(0) &= p_0 \end{aligned}$$

On considère que les conditions de stationnarité sont atteintes quand $p(T) = 0$. On se ramène donc par cette méthode à chercher un zéro d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$, ce qui peut se réaliser, par exemple, avec une méthode de Newton [BH69].

La convergence de la méthode nécessite d'avoir une bonne initialisation de la condition initiale de l'état adjoint p_0, ce qui est parfois difficile à obtenir en pratique. De plus, pour un problème avec contraintes d'état tel que ceux que nous allons étudier, une connaissance a priori de la structure de la trajectoire optimale est requise [BH69, Her08][5]. Cependant, cette méthode possède l'avantage d'être extrêmement précise et d'avoir un coût numérique faible. Ces méthodes ont été étudiées par exemple dans [AMR88, Her08, RS72] et notamment utilisées pour des problèmes complexes nécessitant une forte précision dans les références suivantes [BFLT03, CHT11, Tré03].

Méthodes de collocation du problème indirect Les méthodes de collocation du problème indirect reposent elles aussi sur les conditions d'optimalité du PMP, mais au lieu d'intégrer directement un problème aux conditions initiales, on y discrétise les solutions des équations différentielles selon un schéma de différence finie (Euler ou Runge-Kutta par exemple) en N points de maillage. La résolution

[5] C'est une hypothèse assez difficile à réaliser ne pratique sans étude théorique des extrémales

du problème aux deux bouts consiste alors à trouver le zéro d'une fonction de $\mathbb{R}^{2nN}$ dans $\mathbb{R}^{2nN}$ correspondant aux valeurs de l'état $x(t)$ et de l'état adjoint $p(t)$ aux N instants de discrétisation. Comme expliqué dans [AMR88], les méthodes de collocation intègrent souvent des techniques de raffinement de maillage permettant de résoudre les problèmes avec une grande précision, au détriment de la taille (et donc d'un accroissement de la complexité numérique) du problème d'optimisation. Ces méthodes constituent un intermédiaire entre les méthodes directes simples à mettre en oeuvre, mais pouvant être peu précises et nécessitant une grande puissance de calcul, et les méthodes de tir difficiles à mettre en oeuvre mais très précises et peu gourmandes numériquement.

1.4 Contraintes d'état et méthodes de points intérieurs

Historiquement, les méthodes de points intérieurs ont été introduites dans le cadre de l'optimisation en dimension finie sous contraintes[6] par Fiacco et MacCormick [FM68] à la fin des années 60. Ces méthodes ont connu un important succès dans le milieu des années 80 grâce aux travaux de Karmarkar [Kar08] où ce dernier a montré que, sur des problèmes de programmation linéaire (LP), son algorithme de points intérieurs est 50 fois plus rapide que la méthode du simplexe. Nous présentons brièvement l'idée de cette classe de méthodes, avant d'en exposer la généralisation pour les problèmes de commande optimale qui nous intéressent dans cette thèse.

1.4.1 Points intérieurs en dimension finie

Dans le cadre d'un problème d'optimisation en dimension finie

$$\min_{x\in\mathbb{R}^n} f(x)$$

sous les contraintes $g_i(x) \leq 0$, $i = 1 \dots q$, les méthodes de points intérieurs consistent à résoudre une suite de problèmes, indexée par une suite de paramètres positifs (ε_n) décroissante vers zéro, de la forme

$$\min_{x\in\mathbb{R}^n} \left[f(x) + \varepsilon_n \sum_{i=1}^{q} \gamma \circ g_i(x) \right]$$

où $\gamma : \mathbb{R}^- \mapsto \mathbb{R}^+$ est une fonction de pénalisation. Sous certaines hypothèses, notamment réalisées en programmation quadratique, la suite de solutions optimales ainsi obtenue $(x^*_{\varepsilon_n})$ converge vers la solution du problème original à mesure que la suite (ε_n) tend vers zéro [FM68, NW99]. Ces méthodes de points intérieurs sont très attractives car leur résolution consiste en la résolution d'une suite de problèmes sans contraintes. Dans le cadre de l'optimisation en dimension finie, l'analyse et le choix des fonctions de pénalisation (ainsi que le choix de la suite (ε_n)) ont permis d'aboutir

[6] sous la dénomination "sequential unconstrained minimization techniques" or SUMT voir [BV04]

à des algorithmes de résolution extrêmement performants qui ont été intégrés dans des logiciels d'optimisation comme KNITRO [BNW06], OOQP [Wri04], IPOPT [WB06]. Nous invitons le lecteur intéressé à consulter [FGW02] pour un panorama très complet des méthodes de points intérieurs depuis la fin des années 60 jusqu'aux contribution les plus récentes [Gon12].

1.4.2 L'extension des points intérieurs à la commande optimale

Pour pouvoir traiter des exemples plus généraux, notamment ceux évoqués en §1.2, le problème de commande optimale (1.3)-(1.4) doit être soumis à un certain nombre de contraintes du type $g(x(t)) \leq 0$ pour tout $t \in [0, T]$. La résolution du problème de commande optimale sous ces contraintes est l'objet de cette thèse.

Un tel problème de commande optimale (1.3) et (1.4) sous les contraintes $g_i(x(t)) \leq 0,\ \forall t \in [0, T],\ i = 1 \dots q$ a été traité pour la première fois par une méthode de points intérieurs par Lasdon, Waren et Rice [LWR67]. Dans leur article, les auteurs proposent de résoudre une suite de problèmes de commande optimale pénalisés indexée par $\varepsilon > 0$

$$\min_{u \in U} \left[\int_0^T \ell(x(t), u(t)) + \varepsilon \sum_{i=1}^{q} \frac{1}{g_i(x(t))} dt \right]$$

avec

$$U = \{u \in L^\infty([0, T], \mathbb{R}^m) \text{ t.q. } g_i(x(t)) \leq 0,\ \forall t \in [0, T],\ i = 1 \dots q\}$$

sous la contrainte de dynamique décrite par l'équation (1.4). Dans cet article, les auteurs étendent les résultats obtenus en dimension finie par Fiacco et MacCormick aux problèmes de commande optimale.

Cette approche a notamment inspiré les travaux [GKPC10, GP08b, GP09, GPK08] où, en complément, on utilise des changements de variable sur l'état et la commande. D'autres choix de fonction de pénalisation (notamment logarithmique) ont également été considérés [HS06].

L'intérêt de ces méthodes de points intérieurs est que les solutions sont toutes caractérisées par les conditions simples de stationnarité (sans contraintes) du PMP tels qu'exposées au §1.3.1.1. Sous certaines hypothèses, on peut montrer, dans les différents cas, la convergence de la méthode en terme de critère de coût et, sous l'hypothèse supplémentaire de convexité forte du coût par rapport à la commande, la convergence presque partout de l'état et de la commande.

En général, les résultats obtenus reposent sur l'hypothèse (toujours supposée comme réalisée) de *l'intériorité des solutions optimales des problèmes pénalisés*. Cette intériorité est un point clé pour garantir la convergence des méthodes de point intérieurs en commande optimale. C'est l'hypothèse qui garantit que: *i)* les solutions des problèmes pénalisés sont caractérisées par les simples conditions de stationnarité sans contraintes du PMP. *ii)* on ne crée pas de solutions parasites en

dehors des contraintes.

Cette propriété d'intériorité en dimension infinie a été étudiée par Bonnans et Guilbaud [BG03] dans le cadre de contraintes cubiques sur la commande, i.e. pour les problèmes du type

$$\min_{u\in U}\int_0^T \ell(x,u)dt$$

où $U=\{u\in L^\infty([0,T],\mathbb{R}^m)\,\text{t.q.}\,a_i\leq u_i(t)\leq b_i,\ \text{p.p.t.}\,t\in[0,T],\ i=1,\dots,m\}$. Les auteurs montrent que le problème de commande optimale pénalisé suivant

$$\min_{u\in U}\int_0^T \ell(x,u)-\varepsilon\sum_{i=1}^m \log(u_i(t)-a_i)+\log(b_i-u_i(t))dt$$

est tel que, pour tout $\varepsilon>0$, les solutions optimales du problème u^*_ε sont strictement intérieures aux contraintes:

$$a_i<u^*_{i,\varepsilon}(t)<b_i,\ \text{p.p.t.}\,t\in[0,T],\ i=1\cdots m$$

C'est le résultat le plus avancé, à notre connaissance, sur cette question. Nous cherchons dans cette thèse à obtenir un résultat semblable dans le cas des contraintes d'état.

1.5 Contributions de cette thèse

Les contributions de cette thèse sont de 2 ordres:

1. **Contribution méthodologique:** dans cette thèse le résultat d'intériorité est étendu aux contraintes d'état $g_i(x(t))\leq 0$, pour presque tout t, $i=1\dots q$ et aux contraintes de commande de la forme $u(t)\in\mathcal{C}$, $\forall t$ où $\mathcal{C}$ est un ensemble convexe fermé borné. De plus, en reprenant les fonctions de saturations introduites dans [GP09] nous montrons que la résolution du problème original peut se ramener à la résolution d'une suite de problèmes de commande optimale totalement non contraints dont les solutions sont caractérisées par les conditions de stationnarité du PMP.

2. **Applications:** nous considérons deux cas d'optimisation énergétique de bâtiments d'habitation individuelle et utilisons les outils méthodologiques que nous avons développés pour quantifier le gain de flexibilité que les différentes techniques d'isolation peuvent apporter dans le contexte présenté de "Demand Response".

Chapter 2

Introduction

Contents

2.1 Context: energy optimization in individual housing buildings

Despite numerous policies in favor of energy savings in place at European level for more than 30 years, demographic growth and the increasing part of renewable energies in the electrical mix both threaten the balance between production and consumption of electricity. Consumption peak periods are critically difficult to handle. In this context, energy managers (producers, distributors in connection with governmental bodies) systematically increase production facilities and distribution capacities (electric grid) to maintain this balance. But peak production facilities have drawbacks: *i)* an investment risk higher than other types of electric power plant [HC11] and *ii)* an important financial cost and high CO_2 emissions, since they use fossil energies. An alternative to this solution consists in developing and implementing smart programs of energy management.

The principle is as follows: smoothing the temporal profile of the demand by shifting the energy consumption from peak periods to off-peak periods. This shifting

can be achieved by an incitation of consumers (by a feedback to the customer or by using a pricing-signal on electricity), or by remotely controlling some devices. Industrial and residential sectors are appealing for this smoothing strategy because each customer consumes a large amount of power. In return, concerned customers could be retributed by a compensation for the reduction (or stopping) of their activity. This scheme is known as "Demand side biding" [AES08].

Another solution is to turn to the residential sector. Interestingly, demand management in residential sector is not very developed even though this sector represents an important energy consumption. For example, it represents around 29% [BA07] of total electrical consumption in Europe. This lack of development is due to the multitude of individual customers of the sector ; customers who are more difficult to manage than the ≪ big groups ≫ from the industrial and tertiary sectors. Recently, with the development of communication technologies and smart grid [KR10, TYO+12], many research and experimentation projects have been launched to test large-scale demand response implementations in the residential sector [Fro07, PWMK07]. These programs are commonly referred to as "Demand Response".

The works of this thesis are part of these programs. In order to develop demand response strategies for the residential sector, it is necessary to carry out a quantitative analysis of the ability of domestic electrical uses to serve the energy provider/distributor. This evaluation, that we will perform thanks to state and input constrained optimal control tools, is to be carried out in the following scenarios and operational objectives (as described in [DS11]):

- *Use of the local renewable electricity production at the individual housing level.*
- *Introduction of innovative pricing* representing the state of use of the electric grid, for instance the "Real-time-pricing" or the "Critical peak pricing"[1].
- *Optimized control of the individual operation of equipment* in order to limit the power used at certain times[2].

Among the systems that can be used to achieve the objectives of active demand response, several are dynamical systems (hot water boiler, storage heater, battery, building inertia,...) or work on a dynamical system (convector heater, heat pumps...). A brief technological description can be found in Appendix A.

2.2 Optimal control: a tool for the quantification of potential flexibility assets

The quantification of the flexibility provided by each system (convector heaters, radiative heaters, storage heaters, heat pumps, hot water boilers with storage tank)

[1] These types of pricing must encourage the smoothing of the load curve or of the peaks.
[2] For example, domestic hot water heating functioning only at night.

in order to achieve one of the objectives of active demand response boils down to answering to the following question:

"What is the *maximum* contribution of a *system* to achieve one of the *objectives* of active demand response under comfort *constraints*?"

The term comfort can refer to thermal comfort or to the operation of devices etc. The answer to that question can be found quite naturally by solving a state and input constrained optimal control problem. Three examples are given below, each one meeting one of the objectives of active demand response.

2.2.1 Example 1: Photovoltaic production shifting

Let us take the example of an individual whose housing is equipped with photovoltaic panels and a battery that can store all or part of the power produced by the solar panels $PV(t)$, which varies with time. The energy is totally sold on the grid at the electricity market price. The problem we want to solve is : "what is the maximum income that can be drawn from the shifting (storage for later resale) of the photovoltaic production?" In order to solve it, let us consider a reference year for which the electricity price on the market price(t) is totally known. The problem thus consists in comparing the income the customer gets when s/he resells this production without storage

$$\int_0^T \text{price}(t) P_{\text{pv}}(t) dt \tag{2.1}$$

with the following cost

$$\max_{u_1,u_2} \int_0^T \text{price}(t) \Big[PV(t) - (1-u_1(t)) \max\{P_{\max}, PV(t)\} + r\left(x(t), u_2(t)\right) u_2(t) \Big] dt \tag{2.2}$$

where $P_{\max}$ is the maximum power of charge of the battery, $r\left(x(t), u_2(t)\right)$ is the battery efficiency, $x(t)$ its state of charge, u_1 the percentage of the photovoltaic power stored in the battery, u_2 the battery discharge power, u_1 and u_2 are the controls of the system. This optimization is carried out under the following dynamical constraint

$$\dot{x}(t) = f(x(t), u_1(t), u_2(t))$$

Since the battery capacity and the maximal power of discharge are limited, the optimization must be performed under the following constraints

$$\begin{aligned} x(t) &\in [0, C_{\max}], \quad \forall t \in [0,T] \\ u_1(t) &\in [0,1], \quad \forall t \in [0,T] \\ u_2(t) &\in [0, P_{\max}], \quad \forall t \in [0,T] \end{aligned}$$

with a constraint on the battery $x(0) = x(T)$ that is to say that the final state of charge of the battery is equal to the initial one. When we compare the value

of the two integrals (2.1) and (2.2), we find the *maximal* gain on a year that can be brought by the shifting of part of the photovoltaic production of the owner of the device. This example is solved using the methods proposed in this thesis in Appendix B.1 over one week.

2.2.2 Example 2: Efficiency of an innovative electricity pricing

One of the ways to gauge the efficiency of the pricing applied to the "demand response" is to see what is the optimal management for this pricing on an individual customer who has needs of heating (convector heaters) and of hot water preparation (hot water boilers (HWB) with storage tank). The optimization process aims here at reducing the electricity bill (without subscription). We note price(t) the pricing seen by the customer. The problem consists in solving

$$\min_{u_1,u_2} \left[\int_0^T \text{price}(t) \left(u_1(t) + u_2(t) \right) dt \right]$$

where u_1 and u_2 are the controls, that respectively correspond to the heating and HWB preparation consumption, under the dynamical constraints induced by the thermal inertia of the building and the hot water boilers

$$\begin{aligned} \dot{x}_{\text{bui}}(t) &= f(x_{\text{bui}}(t), u_1(t)) \\ \dot{x}_{\text{hwb}}(t) &= g(x_{\text{hwb}}(t), u_2(t)) \end{aligned}$$

When the heating is supposed to maintain the inhabitants comfort, and noting $h(x_{\text{bui}})$ the temperature inside the building, the optimization problem is made under the following constraint:

$$h(x_{\text{bui}}(t)) \in [T^{\text{bui}}_{\min}(t), T^{\text{bui}}_{\max}(t)], \quad \forall t \in [0, T]$$

The temperature of the superior layer of the hot water boiler $l(x_{\text{hwb}})$ is also subject to functioning constraint

$$l(x_{\text{hwb}}(t)) \in [T^{\text{hwb}}_{\min}(t), T^{\text{hwb}}_{\max}(t)], \quad \forall t \in [0, T]$$

Finally, with the electric demands limited by the emitters, we have

$$\begin{aligned} u_1(t) &\in \left[0, P^{\text{heat}}_{\max}\right], \quad \forall t \in [0, T] \\ u_2(t) &\in \left[0, P^{\text{hwb}}_{\max}\right], \quad \forall t \in [0, T] \end{aligned}$$

Such an example is solved using the methods proposed in this thesis in Appendix B.2 over one week.

2.2.3 Example 3: feasibility of a load shifting from peak period to off-peak period

In order to characterize the feasibility of a load shifting during a peak period (using the thermal inertia of the building) while preserving some comfort in the housing, we can solve the following problem

$$\min_u \left[\int_{\text{peak}} u(t)dt \right]$$

where the control u is the heating electricity consumption, under the following constraints

$$\begin{aligned} \dot{x}_{\text{bui}}(t) &= f(x_{\text{bui}}(t), u(t)) \\ h(x_{\text{bui}}(t)) &\in \left[T^{\text{bui}}_{\text{min}}(t), T^{\text{bui}}_{\text{max}}(t) \right] \\ u(t) &\in \left[0, P^{\text{heat}}_{\text{max}}(t) \right] \end{aligned}$$

for all $t \in [0, T]$. The optimal value of the criterion indicates the feasibility of the load shiftings: if the criterion has a zero value, it means that it is possible not to consume energy at all during the peak periods. It is therefore possible to perform a load shifting during a peak period. If the criterion does not take a zero value, then it is not possible to bring the consumption in peak period to zero while satisfying the constraints. It is therefore not possible to achieve a complete load shifting.

In the case where it is possible to achieve a complete load shifting, it is then necessary to make sure that the best strategy is used to achieve a complete shifting of the load, and to solve the following problem

$$\min_u \left[\int_0^T u(t)dt \right]$$

under the constraints

$$\begin{aligned} \dot{x}_{\text{bui}}(t) &= f(x_{\text{bui}}(t), u(t)) \\ h(x_{\text{bui}}(t)) &\in \left[T^{\text{bui}}_{\text{min}}, T^{\text{bui}}_{\text{max}} \right] \\ u(t) &\in \left[0, P^{\text{heat}}_{\text{max}} \right], \quad \text{off-peak} \\ u(t) &= 0 \quad \text{otherwise} \end{aligned}$$

These problems are introductive examples to optimal control problems that we want to solve in a more general way.

We will now expose the principles and methods of the optimal control before presenting the contributions of this thesis.

2.3 Introduction to optimal control theory (without state constraints)

In this thesis, and as illustrated through examples above, we will look at the general optimal control problem that can be written as follows

$$\min_{u\in\mathcal{U}} \int_0^T \ell(x(t),u(t))dt \tag{2.3}$$

where ℓ is a smooth real-valued function of its arguments, under the following dynamical constraints

$$\dot{x}^u(t) = f(x^u(t),u(t)) \quad ; \quad x(0)=x_0 \tag{2.4}$$

corresponding to a state space representation of a dynamical system. We will look upon state constraints *later on*[3]. On the considered time horizon T, which is assumed to be fixed without loss of generality, one can act on the system state through the control variable u, which can be chosen in a set $\mathcal{U}$ (a subset of L^∞ without loss of generality)

$$\mathcal{U} = \{u\in L^\infty([0,T],\mathbb{R}^m) \text{ s.t. } u(t)\in\mathcal{C} \text{ a.e. } t\in[0,T]\}$$

with $\mathcal{C}$ a bounded closed convex set of $\mathbb{R}^m$ with non empty interior. The optimal control problem consists in finding the control u and its associated state x^u solution of (2.4) minimizing the integral cost (2.3). Numerous extensions are possible: free final time, final cost, definition of a target, jumps in the dynamics, see [BH69, HT11].

2.3.1 Characterization of the solutions

To solve this optimal control problem two main approaches are usually considered: the Pontryaguine minimum principle (PMP) and the dynamic programming principle of Bellman. To present these theories let us first introduce the Hamiltonian $H:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^n\mapsto\mathbb{R}$,

$$H(x(t),u(t),p(t)) \triangleq \ell(x(t),u(t)) + p(t)^t f(x(t),u(t))$$

2.3.1.1 Necessary conditions of optimality: Pontryaguine minimum principle

The minimum principle [PBGM62, Tré08] states necessary conditions of optimality. If $(u,x)\in\mathcal{U}\times W^{1,\infty}([0,T],\mathbb{R}^n)$[4] is an optimal solution of problem (2.3) then there

[3] The reader can find an extended survey (which will not be used here) to techniques presented above in the presence of state constraints [HSV95].

[4] with, classically [Ada75], $W^{1,\infty}([0,T],\mathbb{R}^n) \triangleq \{x\in L^\infty([0,T],\mathbb{R}^n) \text{ s.t. } \dot{x}\in L^\infty([0,T],\mathbb{R}^n)\}$

exists $p \in W^{1,\infty}([0,T],\mathbb{R}^n)$ called adjoint state, such that almost everywhere on $[0,T]$, we have

$$\begin{aligned} \dot{x}^u(t) &= f(x^u(t),u(t)) \\ x(0) &= x_0 \\ \dot{p}(t) &= -\frac{\partial}{\partial x}H(x(t),u(t),p(t)) \\ p(T) &= 0 \\ u(t) &\in \arg\min_{v\in\mathcal{C}} H(x(t),v,p(t)) \end{aligned}$$

2.3.1.2 Dynamic programming

The second approach is based on the principle of dynamic programming of Bellman [Bel57] which has emerged in the beginning of the 60s. The value function $\mathcal{J}$ of the problem defined by

$$\mathcal{J}(\xi,t) \triangleq \inf_{(u,x)} \left\{ \int_t^T \ell(x(s),u(s))ds \; : \; \dot{x}(s) = f(x(s),u(s)) \text{ a.e. } s\in[t,T],\; x(t)=\xi,\; u(s)\in\mathcal{C} \right\}$$

is a solution of a non-linear partial derivative equation named Hamilton-Jacobi-Bellman equation (HJB)

$$\begin{aligned} \frac{\partial\mathcal{J}}{\partial t}(\xi,t) &+ \inf_{v\in\mathcal{C}} H(v,\xi,\frac{\partial\mathcal{J}}{\partial\xi}(\xi,t)) = 0 \quad (\xi,t)\in\mathbb{R}^n\times(0,T) \\ \mathcal{J}(\xi,T) &= 0 \end{aligned}$$

This optimality condition is necessary and sufficient. However, problems of computation time and of memory allocation prevent, in general, the method to compute optimal solutions of problems with long time horizon and a state dimension superior to 3 [Bel57]. Nevertheless, numerous work have been performed on extended cases, particularly in the spatial domain [ABZ12].

2.3.2 Solving numerical methods

2.3.2.1 Direct methods

These widely used methods consist in a discretization of the problem equations yielding a non-linear programming problem (NLP), that is to say a finite dimensional non-linear optimization problem. This approach has been successfully used in the following references [BCM98, BMDP02, Bha06, CP05, JLW03, HP87, KM04, LS99, PMM01, RF04, Vic98, Wri93, YGFDD05].

The main advantage of direct methods is that they are easy to implement and are relatively robust to a poor initialization. In general, these methods have the complexity of the used NLP algorithm, mostly $O(N^3)$ where N is size of the time

discretized problem. These methods make it possible to solve problems with a large number of state variables. However, their precision is in general limited by the precision of the discretization.

2.3.2.2 Indirect methods

Shooting methods Shooting methods rely on the optimality conditions from the PMP. Under certain assumptions, see [AS04], the PMP allows one to write the control as a function of the state and of the adjoint state

$$u(t) = \Gamma(x(t), p(t)) \quad t \in [0, T]$$

Stationnarity conditions are considered as satisfied when $p(T) = 0$. Necessary conditions of optimality consist in solving the $2n$ differential equations on x and p forming a two point boundary value problem since we have an initial condition on x and a final condition on p. The idea of this algorithm is to consider the initial condition of the adjoint state p_0 as an unknown variable and to consider the shooting function which associates the final condition $p(T)$ to p_0, where (x, p) is solution of the Cauchy problem on $[0, T]$

$$\begin{aligned} \dot{x}(t) &= f(x(t), \Gamma(x(t), p(t))), & x(0) = x_0 \\ \dot{p}(t) &= -\frac{\partial}{\partial x} H(x(t), \Gamma(x(t), p(t)), p(t)), & p(0) = p_0 \end{aligned}$$

This method consists in finding the zero of a function from $\mathbb{R}^n$ into $\mathbb{R}^n$, which can be achieved using for example a Newton method [BH69].

The convergence of the method requires a good initial guess of the initial condition p_0 of the adjoint state, which in practice can be difficult to achieve. Moreover for a state constrained optimal control problem like the ones we are about to study, an *a priori* knowledge of the structure of the trajectory is required [BH69, Her08][5]. However, this method is extremely precise and implies a low numerical cost. These methods have been studied in [AMR88, Her08, RS72] and used for complex problems requiring a high precision in the following references [BFLT03, CHT11, Tré03].

Collocation methods of indirect problems Collocation methods also rely on optimality conditions from the PMP, but instead of directly integrate an initial condition problem the differential equations are discretized in N mesh points using finite element methods such as (Euler or Runge-Kutta for example). Solving the two point boundary value problem consists in finding the zero of a function from $\mathbb{R}^{2nN}$ to $\mathbb{R}^{2nN}$ corresponding to the values of the state $x(t)$ and the adjoint state $p(t)$ at mesh points.

As explained in [AMR88], mesh refinement techniques are embedded in collocation methods allowing a high precision solving. But this precision is achieved at the expense of an increase of the optimization problem dimension (and of its

[5] In practice this assumption is difficult to satisfy without a theoretical study of the extremals

numerical complexity). These methods are a trade-off between direct methods easy to implement but which can have a low precision and require large computational expenses and shooting methods difficult to implement but extremely precise and do not require large computational expenses.

2.4 State constraints and interior point methods

Historically, interior point methods have been introduced for finite dimensional constrained optimization[6] by Fiacco and MacCormick [FM68] during the late 60s. These methods have been very successful in the middle of the 80s thanks to Karmarkar's work [Kar08] where he has shown that on linear programming problems (LP) his interior point algorithm is 50 times faster than the simplex method. We briefly describe the general idea of these methods before exposing their generalization to optimal control problems which are studied in this thesis.

2.4.1 Interior point in finite dimensional optimization

In the framework of finite dimensional optimization

$$\min_{x\in\mathbb{R}^n} f(x)$$

under the following constraints $g_i(x) \leq 0$, $i = 1 \dots q$, interior point methods consist in solving a sequence of problems, indexed by a sequence of positive parameter (ε_n) decreasing to zero of the form

$$\min_{x\in\mathbb{R}^n} \left[f(x) + \varepsilon_n \sum_{i=1}^{q} \gamma \circ g_i(x) \right]$$

where $\gamma : \mathbb{R}^- \mapsto \mathbb{R}^+$ is a penalty function. Under certain assumptions, especially satisfied in quadratic programming, the obtained sequence of optimal solutions $(x^*_{\varepsilon_n})$ converges to the original problem solution as the sequence (ε_n) tends to zero [FM68, NW99]. These interior point methods are very appealing from the programming point of view since their solving consists in solving a sequence of unconstrained problems. In the framework of finite dimensional optimization, the analysis and the choice of penalty functions (and the choice of the sequence (ε_n)) have led to astonishing solving algorithms which have been implemented in optimization softwares such as KNITRO [BNW06], OOQP [Wri04], IPOPT [WB06]. We refer the interested reader to [FGW02] for a full survey of interior point methods from the 60s to most recent contributions [Gon12].

2.4.2 Extension of interior point methods to optimal control

In order to solve more general examples, especially those described in §2.2, the optimal control problem (2.3)-(2.4) must take into account some constraints under

[6] Under the denomination "Sequential Unconstrained Minimization Techniques" or SUMT [BV04].

the form $g(x(t)) \leq 0$ for all $t \in [0,T]$. Solving these problems is the subject of this thesis.

Such a problem of optimal control (2.3) and (2.4) under the constraints $g_i(x(t)) \leq 0$, $\forall t \in [0,T]$, $i = 1 \dots q$ using interior point methods has been first addressed by Lasdon, Waren and Rice [LWR67]. In their article, the authors propose to solve the following optimal control problems sequence

$$\min_{u \in U} \left[\int_0^T \ell(x(t), u(t)) + \varepsilon \sum_{i=1}^{q} \frac{1}{g_i(x(t))} dt \right]$$

with

$$U = \{u \in L^\infty([0,T], \mathbb{R}^m) \text{ s.t. } g_i(x(t)) \leq 0,\ \forall t \in [0,T],\ i = 1 \dots q\}$$

under the dynamical constraint given in equation (2.4). In this article the authors generalize the results obtained in finite dimensional optimization by Fiacco and MacCormick.

This approach has been continued in [GKPC10, GP08b, GP09, GPK08], where, in addition, changes of variables of the state and the control are used. Alternative choices of penalty functions (especially logarithmic functions) have also been used [HS06].

The advantage of these interior point methods is that the solutions are characterized by the simple conditions of stationarity (without constraints) of the PMP as described in §2.3.1.1. Under certain assumptions, the convergence of the cost can be proven and under assumption of strong convexity of the cost with respect to the control, the convergence almost everywhere of the control and the state can also be proven.

In general, these results rely on the assumption (always considered as satisfied) of the *interiority of the optimal solutions of the penalized problems*. This interiority is a key point to guarantee the convergence of interior point methods in optimal control. This is the assumption which induces that *i)* the solutions of the penalized problems are characterized by the simple unconstrained stationarity conditions of the PMP, *ii)* no parasite solutions which do not satisfy the constraints are found.

The interiority property in infinite dimensional optimization has been addressed by Bonnans and Guilbaud [BG03] in the case of cubic constraints on the control, i.e. for problems under the form

$$\min_{u \in U} \int_0^T \ell(x,u) dt$$

where $U = \{u \in L^\infty([0,T], \mathbb{R}^m) \text{ s.t. } a_i \leq u_i(t) \leq b_i, \text{ a.e. } t \in [0,T],\ i = 1, \dots, m\}$. The authors prove that the following penalized optimal control problem

$$\min_{u \in U} \int_0^T \ell(x,u) - \varepsilon \sum_{i=1}^{m} \log(u_i(t) - a_i) + \log(b_i - u_i(t)) dt$$

is such that, for all $\varepsilon > 0$, optimal solutions of the problem u^*_ε are strictly interior to the constraints

$$a_i < u^*_{i,\varepsilon}(t) < b_i, \text{ a.e. } t \in [0, T], \ i = 1 \cdots m$$

This is the most advanced result, to our knowledge, on this question. One of the purposes of this thesis is to obtain a similar result when dealing with state constraints.

2.5 Thesis contributions

This thesis contributions are twofold

1. **Methodological contribution:** in this thesis the result of interiority is extended to state constraints $g_i(x(t)) \leq 0$, $\forall t$, $i = 1 \dots q$ and to control constraints under the form $u(t) \in \mathcal{C}$, for almost every t where $\mathcal{C}$ is a bounded closed convex set. Moreover, using saturation functions developed in [GP09] we prove that solving the original problem can be achieved through the solving of a sequence of fully unconstrained optimal control problems whose solutions are readily characterized by the simple stationarity conditions of the PMP.

2. **Applications:** we consider two cases of energy management optimization in individual housing and use the methodological tools that we have developed to quantify the flexibility provided by different techniques of insulation in the presented "Demand Response" context.

Part I

Contribution méthodologique pour la commande optimale Methodological contribution to optimal control

Chapter 3

A constructive penalty design for non-linear state and input constrained optimal control problems

Contents

3.1 Notations, assumptions and problems statements

3.1.1 Constrained optimal control problem (COCP) and notations

The following Constrained Optimal Control Problem (COCP) is studied

$$\min_{u\in\mathcal{U}\cap\mathcal{X}}\left[J(x^u,u)=\int_0^T \ell(x^u,u)dt\right] \tag{3.1}$$

for the dynamics

$$\dot{x}^u(t)=f(x^u(t),u(t)),\qquad x(0)=x_0 \tag{3.2}$$

where $\ell:\mathbb{R}^n\times\mathbb{R}^m\mapsto\mathbb{R}$ is a locally Lipschitz function of its arguments and continuously differentiable with respect to u, $x^u(t)\in\mathbb{R}^n$ and $u(t)\in\mathbb{R}^m$ are the state and the control which satisfy (MIMO) non-linear dynamics described in equation (3.2). The set $\mathcal{U}\cap\mathcal{X}$ is defined by *control and state constraints* that we detail below. A solution u^* of (3.1) is defined as a global minimizer of the cost function over $\mathcal{U}\cap\mathcal{X}$.

3.1.1.1 Control constraints

The control $u:\mathbb{R}\mapsto\mathbb{R}^m$ is constrained to belong to the set

$$\mathcal{U}\triangleq\{u\in L^\infty([0,T],\mathbb{R}^m)\text{ s.t. }u(t)\in\mathcal{C}\text{ a.e. }t\in[0,T]\}$$

where the set $\mathcal{C}$ satisfies the following assumption

Assumption 1 *$\mathcal{C}$ is a bounded closed convex subset of $\mathbb{R}^m$ which has a nonempty interior which contains 0. Moreover, it is assumed that $\partial\mathcal{C}$ the boundary of $\mathcal{C}$ is continuously differentiable.*

3.1.1.2 State constraints

The state $x^u:\mathbb{R}\mapsto\mathbb{R}^n$ is subjected to satisfy a set of inequalities

$$g_i(x^u(t))\leq 0\ \ ,\ \ i=1\dots q\ \ ,\ \ \forall t\in[0,T]$$

where the g_i are continuously differentiable functions $\mathbb{R}^n\mapsto\mathbb{R}$. They serve to define

$$X_{\text{ad}}\triangleq\{x\in\mathbb{R}^n\text{ s.t. }g_i(x)\leq 0, i=1,\dots,q\}$$

We make the following assumption[1] on X_{ad}:

Assumption 2 *The interior set of X_{ad} is the set noted $\overset{\circ}{X}_{\text{ad}}$ such that*

$$\overset{\circ}{X}_{\text{ad}}\triangleq\{x\in\mathbb{R}^n\ s.t.\ g_i(x)<0\quad i=1\dots q\}$$

To implement interior point methods, we shall naturally make the following assumption

Assumption 3 *The initial condition x_0 of equation (3.2) belongs to $\overset{\circ}{X}_{\text{ad}}$.*

[1] This assumption is not trivial: consider for instance $q=1$ with g a continuously differentiable function from $\mathbb{R}$ to $\mathbb{R}$ such that $g(x)<0$ for $x<0$ and $g(x)=0$ for $x\geq 0$. Then $X_{\text{ad}}=\mathbb{R}$, while the set $g(x)<0$ is $(-\infty,0)$, which is not the interior of X_{ad}.

3.1.1.3 Set of admissible controls

Now, we can properly define the set $\mathcal{X}$ in (3.1) by

$$\mathcal{X} \triangleq \left\{ u \in L^\infty([0,T],\mathbb{R}^m) \text{ s.t. } x^u(t) \in X^{\text{ad}} \text{ for all } t \in [0,T] \right\}$$

Before defining the penalized problem (3.5) that will be used in our interior point method, we shall elaborate on the problem settings and introduce some related concepts and one further assumption.

3.1.1.4 Assumptions on the dynamics and consequences on the state

Assumption 4 *f is continuously differentiable. Moreover, there exists a positive constant D such that*

$$\|f(x,u)\| \leq D(1+\|x\|), \ \forall x \in \mathbb{R}^n, \forall u \in \mathcal{C} \tag{3.3}$$

This is verified by linear dynamics $\dot{x} = Ax + Bu$, for instance.

Under Assumption 4, we classically derive the following proposition

Proposition 1 *For all $u \in \mathcal{U}$, the maximal solution x^u of the dynamics (3.2) is defined on $[0,T]$ and x^u is bounded by a constant that depends only on x_0 and D. Moreover, the following mappings*

$$\begin{aligned} L^\infty([0,T],\mathbb{R}^m) \ni u &\mapsto x^u \in C^0([0,T],\mathbb{R}^n) \\ L^1([0,T],\mathbb{R}^m) \ni u &\mapsto x^u \in C^0([0,T],\mathbb{R}^n) \end{aligned}$$

are Lipschitz.

Proof: Consider x^u the maximal solution of (3.2). The use of the Gronwall lemma ([Kha02] p. 651) for equation (3.3) shows that x^u is bounded on its interval of definition. Since f is continuously differentiable, the boundedness of $u \in \mathcal{U}$ and of x^u implies that the derivatives of f are bounded when $u \in \mathcal{U}$. Consider now two controls u and v in $\mathcal{U}$. Using the Gronwall lemma on $x^u - x^v$ shows that its dynamics is sublinear with respect to $x^u - x^v$ and $u - v$ with a zero initial condition, which proves the regularity of x^u with respect to u, both in the L^1 and L^∞ norms. ■

3.1.2 Gauge functions of convex sets

Classically [Sch78], one can associate a gauge function $G_{\mathcal{C}}$ to any convex set $\mathcal{C}$. Under some mild assumptions, the gauge acts almost like a norm and reveals handy in our problem formulation. Conveniently, the fact that a vector u belongs to the interior, boundary or exterior of $\mathcal{C}$ boils down to comparing $G_{\mathcal{C}}(u)$ to 1. For this reason, in our methodology, the gauge is used as an argument of the penalty function referring to the convex set $\mathcal{C}$.

Definition 1 (Schwartz [Sch78]) *The gauge function defined by $\mathcal{C}$ is the mapping $G_{\mathcal{C}} : \mathbb{R}^m \mapsto \mathbb{R}^+$ defined by*

$$G_{\mathcal{C}}(u) = \inf \{\lambda \geq 0 \ s.t. \ u \in \lambda \mathcal{C}\}$$

In our context, the gauge function of $\mathcal{C}$ satisfies the following properties:

Proposition 2 *Under Assumption 1 on the set $\mathcal{C}$, the gauge function $G_{\mathcal{C}}$ has the following properties*

a) $G_{\mathcal{C}}(u)$ is a well defined non-negative real for all u

b) There exists $0 < N < M$ such that

$$\frac{\|u\|}{M} \leq G_{\mathcal{C}}(u) \leq \frac{\|u\|}{N} \quad \forall u \in \mathbb{R}^m \tag{3.4}$$

In particular, $G_{\mathcal{C}}(u) = 0$ implies $u = 0$

c) The gauge is positively homogeneous, i.e. $G_{\mathcal{C}}(\lambda u) = \lambda G_{\mathcal{C}}(u)$ for all $\lambda \geq 0$

d) $G_{\mathcal{C}}$ is a strictly convex function which is locally bounded; as a consequence, it is continuous

e) $G_{\mathcal{C}}$ has a directional derivative in the sense of Dini[2] at $u = 0$ along direction d and its value is $G_{\mathcal{C}}(d)$

f) If Assumption 1 holds, $G_{\mathcal{C}}$ is differentiable on $\mathbb{R}^m \setminus \{0\}$

*g) **[main result for later discussions]** $G_{\mathcal{C}}(u) < 1$ if and only if u belongs to the interior of $\mathcal{C}$; $G_{\mathcal{C}}(u) = 1$ if and only if u belongs to the boundary $\partial\mathcal{C}$ of $\mathcal{C}$; $G_{\mathcal{C}}(u) > 1$ if and only if u belongs to the exterior of $\mathcal{C}$*

Proof: See Appendix C.1. ■

3.1.2.1 Differentiability issues and control decomposition

In cases of practical interest, it may happen that the gauge function may be non-differentiable, because the boundary of the convex set itself may be non-differentiable. A simple example is the case where $\mathcal{C}$ is the cube defined by $\max_i |u_i| \leq 1$. This may turn troublesome in algorithms where the differentiability of the cost is required (e.g. descent methods), because our penalties (which will be added to the original cost function) will involve gauge functions.

This is why we shall now introduce a more general formalism which (as we shall see later) will encompass the case where the boundary of $\mathcal{C}$ is differentiable as a whole, the case of a cubic convex, and a whole range of intermediate cases. This reformulation will allow a more general definition of the control penalties used later in the penalized control problem (3.5); in particular, these penalties will be differentiable.

[2] The Dini derivative of a function f at point $x \in \mathbb{R}^n$ along the direction $d \in \mathbb{R}^n$ is defined as the limit (when it exists) of $\frac{f(x+hd)-f(x)}{h}$ when h tends to 0 with positive values.

Assumption 5 *The control u is considered to belong to the cartesian product $\mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_p}$, $p \geq 1$, with $\sum m_i = m$ and written under the form $u = (u_1, \ldots, u_p)$. The constraints on u are expressed by $u_i \in \mathcal{C}_i$, where each $\mathcal{C}_i$, $i = 1 \cdots p$ satisfies Assumption 1 and has a continuously differentiable boundary*[3].

Note that any control satisfying Assumption 5 satisfies Assumption 1; conversely, any control set satisfying Assumption 1 *where $\mathcal{C}$ has a differentiable boundary $\partial\mathcal{C}$* satisfies Assumption 5 with $p = 1$.

In Assumption 5, $\mathcal{C}$ stands for the convex defined by the cartesian product of the $\mathcal{C}_i$. The control u belongs to the interior of $\mathcal{C}$ if and only if all of the u_i belong to the interior of $\mathcal{C}_i$, or equivalently (see Proposition 2) if $G_{\mathcal{C}_i}(u_i) < 1$, $\forall i = 1 \ldots p$.

Example: This settings allows one to consider the case where $\mathcal{C}$ is $\{u \in \mathbb{R}^3 \text{ s.t. } u_1^2 + u_2^2 \leq 1 \, , |u_3| \leq 1\}$. The boundary of $\mathcal{C}$ is not differentiable, yet $u \in \mathcal{C}$ can be rewritten as $u = (u_1, u_2)$, where u_1 belongs to an appropriate Euclidian disk, and u_2 belongs to an appropriate segment of $\mathbb{R}$. Conveniently, the formalism used in Assumption 5 includes the case where the convex $\mathcal{C}$ is a hypercube, or alternatively, where $\mathcal{C}$ has a differentiable boundary.

We can now proceed with the presentation of the penalty method that will be instrumental in the implementation of an interior point method.

3.1.3 Presentation of a penalized problem (POCP)

3.1.3.1 Introduction of the penalty functions

Following the approach of interior methods in their application to optimal control [BG03], we introduce two penalty functions

$$\begin{aligned} \gamma_g : \quad (-\infty, +\infty) \quad &\to [0, +\infty) \\ \gamma_u : \quad [0, 1] \quad &\to [0, +\infty) \end{aligned}$$

for which we make the following assumptions

Assumption 6 *We assume that*

- $\begin{cases} \gamma_g(x) = 0 & \text{if} \quad x \geq 0 \\ \gamma_g(x) \geq 0 & \text{if} \quad x < 0 \end{cases}$
- *for $x < 0$, γ_g is continuously differentiable, convex, and increasing*
- $\lim_{x \uparrow 0} \gamma_g(x) = +\infty$
- *γ_u is continuously differentiable, strictly convex, and non-decreasing*
- $\lim_{u \uparrow 1} \gamma_u(u) = +\infty$
- *$\gamma_u(0) = 0$; γ_u is right continuously differentiable at $u = 0$ with $\gamma_u'(0) = 0$.*

[3] this makes sense only for the indices i for which $m_i > 1$.

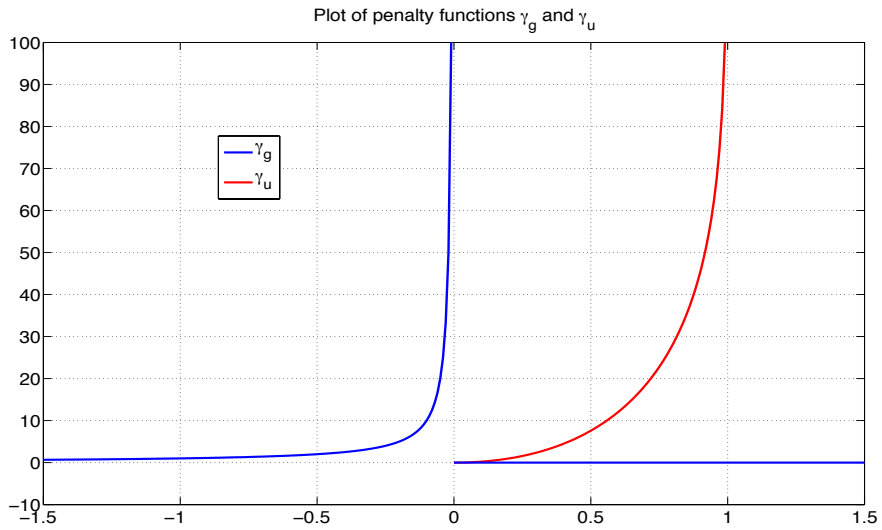

Figure 3.1: Plot of penalty functions γ_g and γ_u.

- $\gamma_u'(u)$ *is right Lipschitz at* $u = 0$

In addition, we shall consider that $\gamma_u(1) = +\infty$.

Adding such penalties to the cost function (3.1) may yield infinite values for the integral (3.5) below. Naturally, controls that lead to an infinite cost will not be considered as optimal.

Typical graphes of γ_g and γ_u are presented in Figure 3.1. In the following section, we combine the penalty functions γ_g and γ_u with the formulation of the constraints to define a penalized optimal control problem (POCP). Interestingly, the usage of $+\infty$ will not lead to indeterminations, as γ_u will be summed with lower-bounded quantities, and there will be no product of γ_u with zero.

3.1.3.2 Definition of a (first) penalized problem

For a given parameter $\varepsilon > 0$, consider the following POCP

$$\min_{u \in \mathcal{U}} \left[K(u,\varepsilon) = \int_0^T \ell(x^u, u) + \varepsilon \left(\sum_{i=1}^{q} \gamma_g \circ g_i(x^u) + \sum_{i=1}^{p} \gamma_u \circ G_{\mathcal{C}_i}(u_i) \right) dt \right] \tag{3.5}$$

under the dynamics (3.2). Observe that now u is constrained to belong to $\mathcal{U}$, which means that the state constraints have disappeared from the formulation (as will be shown, these state constraints are automatically managed by the introduction of the penalties), but the control constraints have not been relaxed (compare with (3.1)). In §3.3 we shall also remove the control constraints, once we have proved that optimal solutions to the penalized problems are interior, in §3.2.2.

3.1.3.3 Properties of the control penalty

We now use the properties of the control sets $\mathcal{C}_i$ and of the penalty function γ_u to exhibit important properties on the POCP (3.5).

Proposition 3 (Differentiability) *For $i = 1, \dots, p$, the application $\gamma_u \circ G_{\mathcal{C}_i}$ is continuously differentiable on the interior of $\mathcal{C}_i$. As a consequence the integrand in the penalized cost (3.5) is continuously differentiable with respect to the control u in the interior of $\mathcal{C}$.*

Proof: From Proposition 2, we know that $G_{\mathcal{C}_i}$ is continuously differentiable on $\mathbb{R}^{m_i} \setminus \{0\}$ because the boundary $\partial\mathcal{C}_i$ is continuously differentiable. On the other hand, γ_u is continuously differentiable on $[0, 1)$; hence $\gamma_u \circ G_{\mathcal{C}_i}$ is continuously differentiable on the interior of $\mathcal{C}_i$ minus the origin.

Since $G_{\mathcal{C}}$ has bounded derivatives at $u = 0$ in the sense of Dini, and since $\gamma_u'(G_{\mathcal{C}}(0)) = \gamma_u'(0) = 0$, we conclude that $\gamma_u \circ G_{\mathcal{C}}$ has a zero derivative at the origin. Moreover, γ_u' being Lipschitz (with constant K) in a neighborhood of 0, one has $|\gamma_u' \circ G_{\mathcal{C}}(u)| \leq K|G_{\mathcal{C}}(u)|$. We derive that the limit of the derivative of $\gamma_u \circ G_{\mathcal{C}}(u)$ is 0 when u tends to 0. This concludes the proof. ■

Convexity

Proposition 4 (Convexity) *For $i = 1, \dots, p$, the penalty $\gamma_u \circ G_{\mathcal{C}_i}$ is convex. As a consequence, if ℓ is convex with respect to u, the integrand in the penalized cost (3.5) is convex with respect to u.*

Proof: We have seen that $G_{\mathcal{C}_i}$ is convex; since γ_u is convex, and since it is non-decreasing, then $\gamma_u \circ G_{\mathcal{C}_i}$ is convex. ■

3.2 Interiority of the optimal constrained variables of the POCP

The objective of this section is to exhibit sufficient conditions on the penalty functions such that any optimal solution $u^* \in \mathcal{U}$ of POCP (3.5) (satisfying the input constraint) actually belongs to $\mathcal{U} \cap \mathcal{X}$ and, as a consequence, is admissible for COCP (3.1) (i.e. it satisfies both input and state constraints).

This section is organized as follows. In §3.2.1, we exhibit a sufficient condition on the *state* penalty γ_g under which any optimal solution of POCP (3.5) strictly satisfies the state constraints, i.e. $g_i(x(t)) < 0$, $\forall i$ $\forall t \in [0, T]$. According to Assumption 2, this is equivalent to the fact that the state remains at all times in $\overset{\circ}{X}_{\text{ad}}$. In §3.2.2, we exhibit an additional sufficient condition on the *control* penalty γ_u under which any optimal solution of POCP (3.5) strictly satisfies the input constraints, specifically the essential supremum for $t \in [0, T]$ of the gauge functions is strictly smaller than 1. In §3.2.3, our first main result (Theorem 1) constructively proves the existence of penalties that satisfy the conditions of §3.2.1 and §3.2.2. As

a consequence, if the penalties satisfy these two sufficient conditions, then any optimal solution of POCP (3.5) is interior.

We shall use the following notations:

Definition 2

$$\begin{aligned}\mathcal{X}^{\text{strict}} &= \{u \in L^\infty([0,T],\mathbb{R}^m) \ s.t. \ x^u(t) \in \overset{\circ}{X}_{ad} \ \forall t \in [0,T]\} \\ \mathcal{U}^{\text{strict}} &= \left\{u \in \mathcal{U} \ s.t. \ \max_i \operatorname{ess\,sup}_t G_{\mathcal{C}_i}(u_i) < 1\right\}\end{aligned}$$

3.2.1 Interiority of the optimal states for the penalized problem

In this section, we exhibit a sufficient condition on the state penalty γ_g ensuring that any optimal solution of POCP (3.5) belongs to $\mathcal{X}^{\text{strict}}$ and hence is admissible for COCP (3.1).

Definition 3 (Proximity to a constraint) *For any constraint g_i, we define the proximity to the constraint by*

$$\alpha \mapsto \mu_{g_i}(u,\alpha) = \text{meas}\left(\{t \in [0,T] \ s.t. \ 0 > g_i(x^u(t)) \geq -\alpha\}\right) \tag{3.6}$$

where meas(.) is the Lebesgue measure of its argument.

Proposition 5 *If, for all $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$, the penalty function γ_g satisfies*

$$\lim_{\alpha \downarrow 0} \gamma_g(-\alpha)\mu_{g_i}(u,\alpha) = +\infty \tag{3.7}$$

then, $\forall \varepsilon > 0$, $\forall u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$

$$K(u,\varepsilon) = +\infty$$

It follows that the penalized cost $K(u,\varepsilon)$ is finite only if $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$.

Proof: Let $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$, then there exists an index i such that $\max_{t\in[0,T]} g_i(x(t)) \geq 0$. Since $\gamma_g(x) = 0$ when $x \geq 0$, we have

$$\mathcal{I}_i \triangleq \int_0^T \gamma_g(g_i(x(t)))dt = \int_{0>g_i(x(t))} \gamma_g(g_i(x(t)))dt$$

Moreover, since $\gamma_g \geq 0$, we have, for $\alpha > 0$

$$\mathcal{I}_i \geq \int_{0>g_i(x(t))\geq -\alpha} \gamma_g(g_i(x(t)))dt \triangleq \mathcal{J}_i(\alpha)$$

The state penalty satisfies $\gamma_g \geq 0$ on $(-\infty, 0)$, thus $\mathcal{J}_i(\alpha)$ is a non-decreasing positive continuous function of $\alpha > 0$ which satisfies

$$\inf_{\alpha>0} \mathcal{J}_i(\alpha) = \lim_{\alpha\downarrow 0} \mathcal{J}_i(\alpha) \triangleq \mathcal{J}_i(0^+)$$

Since γ_g is increasing and since the Lebesgue measure is right continuous [KF99]

$$\begin{aligned}\mathcal{J}(0^+) = \lim_{\alpha\downarrow 0} \int_{0>g_i(x(t))\geq -\alpha} \gamma_g(g_i(x(t)))dt &\geq \lim_{\alpha\downarrow 0} \int_{0>g_i(x(t))\geq -\alpha} \gamma_g(-\alpha)dt \\ &= \lim_{\alpha\downarrow 0} \gamma_g(-\alpha)\mu_{g_i}(u,\alpha)\end{aligned}$$

with $\mu_{g_i}(u,\alpha)$ the Lebesgue measure defined in (3.6). If (3.7) holds, then $\mathcal{J}_i(0^+) = +\infty$ which implies that $\mathcal{I}_i = +\infty$. From Proposition 1, we know that x^u is uniformly bounded in sup norm for $u \in \mathcal{U}$, and, as a consequence, $|\int_0^T \ell(x^u,u)dt|$ is bounded for all $u \in \mathcal{U}$. Moreover, $\sum_{i\leq p}\int_0^T \gamma_u(G_{\mathcal{C}_i}(u_i))dt \geq 0$. As a summary, $K(u,\varepsilon)$ is the sum of lower-bounded terms and of $\varepsilon\,\mathcal{I}_i$, with $\mathcal{I}_i = +\infty$. This proves that the cost $K(u,\varepsilon)$ is infinite for every $\varepsilon > 0$. ■

Since the measure $\mu_{g_i}(u,\alpha)$ which appears in equation (3.7) involves the control u, it is handy to give a lower bound of it when u spans $\mathcal{U}\setminus\mathcal{X}^{\text{strict}}$. This bound will be used, in §3.2.3, in the explicit construction of penalty functions. This bound is given by the following result.

Proposition 6 *Define $-\alpha_0 = \max_i(g_i(x_0))$; one has $\alpha_0 > 0$ because the initial condition is interior (Assumptions 2 and 3). Then, there exists a constant $\Gamma < +\infty$ such that for all $\alpha \in [0,\alpha_0]$, for all $u \in \mathcal{U}\setminus\mathcal{X}^{\text{strict}}$ the measure $\mu_{g_i}(u,\alpha)$ defined in equation (3.6) is lower-bounded as follows*

$$\mu_{g_i}(u,\alpha) \geq \frac{\alpha}{\Gamma}$$

Proof: The proof is given in Appendix C.2 together with the expression of Γ. ■

Using Assumption 3 together with Propositions 5 and 6, one finally obtains

Lemma 1 *If the state penalty γ_g is such that*

$$\lim_{\alpha\downarrow 0} \alpha\gamma_g(-\alpha) = +\infty \tag{3.8}$$

then, any local optimal solution u^* of POCP (3.5) is admissible for COCP (3.1) because

$$u^* \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$$

Then any local optimal control u^* for problem (3.5) yields a trajectory x^{u^*} with values in $\overset{\circ}{X}_{ad}$.

Proof: If equation (3.8) holds, then we derive from Proposition 6 that (3.7) holds for $u \in \mathcal{U}\setminus\mathcal{X}^{\text{strict}}$. From Proposition 5, we derive that $K(u,\varepsilon) < +\infty$ for $u \in \mathcal{U}$ only if $u \in \mathcal{U}\cap\mathcal{X}^{\text{strict}}$. This holds, in particular, for any local optimal control of POCP (3.5). ■

3.2.2 Interiority of the optimal constrained control

In this section, we assume that the state penalty satisfies condition (3.8) from Lemma 1. In particular, any optimal control for the penalized problem belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. Then, we exhibit a sufficient condition on the control penalty γ_u such that any optimal solution u^* of POCP (3.5) belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. In particular $u_i(t)$ belongs to the interior of $\mathcal{C}_i$ for almost every t.

3.2.2.1 Construction of an interior control v

In what follows, we shall use the following result:

Proposition 7 *For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, there exists $\alpha > 0$ such that, for all $v \in \mathcal{U}^{\text{strict}}$ satisfying $\|u - v\|_{L^\infty} \leq \alpha$, we have*

$$v \in \mathcal{X}^{\text{strict}}$$

Proof: Let $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$ and note $-2\beta_0 = \max_{t\in[0,T],i=1,\dots,q} g_i(x^u(t))$. Since $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, we have $\beta_0 > 0$. From Proposition 1, and the continuity of the function g, there exists $\alpha_N > 0$ and $\Lambda > 0$ such that for all $v \in \mathcal{U}^{\text{strict}}$

$$\max_i \|u_i - v_i\|_{L^\infty} \leq \alpha_N \Rightarrow \max_i \| \, g_i(x^u) - g_i(x^v) \, \|_{L^\infty} \leq \Lambda\alpha_N$$

Setting $\alpha = \beta_0/\Lambda$, one has $\max_i \max_{t\in[0,T]} g_i(x^v(t)) \leq -\beta_0 < 0$. Therefore, $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. This concludes the proof. ■

We now proceed to the construction of a control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ which will be used in Proposition 9.

Definition 4 (Desaturated control) *For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, for all $\alpha > 0$, we define a desaturated control $v(u, \alpha) = (v_1 \cdots v_p)$ as follows*

$$v_i(t) = \begin{cases} u_i(t) & \text{if } G_{\mathcal{C}_i}(u_i(t)) < 1 - \alpha \\ (1 - 2\alpha)u_i(t) & \text{otherwise} \end{cases} \tag{3.9}$$

Proposition 8 *For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, there exists $\alpha > 0$ such that the modified control v from Definition 4 satisfies*

$$v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$$

Proof: We shall use the following definitions, inspired by Definition 3

$$\begin{aligned} E_u(\alpha) &\triangleq \{t \in [0,T] \text{ s.t. } \exists i \leq p \text{ s.t. } G_{\mathcal{C}_i}(u_i(t)) \geq 1 - \alpha\} \\ \mu_u(\alpha) &\triangleq \text{meas}(E_u(\alpha)) \end{aligned} \tag{3.10}$$

First, let us prove that $v \in \mathcal{U}^{\text{strict}}$. Assume that $\mu_u(\alpha) = 0$; in this case, for all i, $G_{\mathcal{C}_i}(u_i(t)) < 1 - \alpha$ almost everywhere. Therefore, $u \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. Using

equation (3.9) yields $v = u \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$.
Now, let us assume that $\mu_u(\alpha) > 0$. In this case, for all i,

$$\begin{aligned} G_{\mathcal{C}_i}(v_i(t)) &< 1-\alpha \quad \text{a.e. } t \in [0,T] \setminus E_u(\alpha) \\ G_{\mathcal{C}_i}(v_i(t)) &\leq 1-2\alpha \quad \forall t \in E_u(\alpha) \end{aligned}$$

For all i, $u_i(t) \in \mathcal{C}_i$ almost everywhere, and, since $1-2\alpha \in (0,1)$, then $v_i(t) \in \text{int}(\mathcal{C}_i)$ almost everywhere, therefore $v \in \mathcal{U}^{\text{strict}}$.

We now prove that $v \in \mathcal{X}^{\text{strict}}$. Let M_i be the radius of a ball that contains $\mathcal{C}_i$, using equation (3.9) we have $\|u_i(t) - v_i(t)\| \leq 2\alpha\|u_i(t)\| \leq 2\alpha M_i$. From Proposition 7, there exists $\alpha^+ > 0$ such that if $\| u - v \|_{L^\infty} \leq \alpha^+$ then $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. For all $\alpha \in (0, \min\{1/2, \min_i \frac{\alpha^+}{2M_i}\})$ we have

$$\|u_i(t) - v_i(t)\| \leq 2\alpha\|u_i(t)\| \leq \alpha^+, \quad i = 1 \cdots p$$

Therefore $v \in \mathcal{X}^{\text{strict}}$. Thus, $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. This concludes the proof. ■

3.2.2.2 Condition guaranteeing the strict interiority of the optimal control

To prove that any optimal control belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$, it is enough to find a condition on the penalties such that for any $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$, the modified control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ from Definition 4 satisfies

$$K(v,\varepsilon) < K(u,\varepsilon)$$

This fact contradicts the optimality of every point of $(\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$.
The following result gives an upper estimate on the difference $K(v,\varepsilon) - K(u,\varepsilon)$. This estimate is the sum of three terms, representing respectively

(i) the integral variation of the original cost (3.1)

(ii) the integral variation of the state penalties $\varepsilon \sum_{i \leq q} \gamma_g \circ g_i$

(iii) the integral variation of the input penalty $\varepsilon \sum_{i \leq p} \gamma_u \circ G_{\mathcal{C}_i}$

In §3.2.3 we give constructive conditions on the penalties that make this upper bound strictly negative when $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$.

Proposition 9 *For any control $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$, considering the modified control v from equation (3.9), for any $\varepsilon > 0$ one has*

$$K(v,\varepsilon) - K(u,\varepsilon) \leq \alpha \left[U_\ell + U_g(\varepsilon) - \varepsilon\gamma_u'(1-3\alpha)\right] \mu_u(\alpha) \tag{3.11}$$

where $\mu_u(\alpha)$ is defined by (3.10), U_l is a constant parameter and $U_g(\varepsilon)$ only depends linearly on ε.

Proof: See Appendix C.3. ■

Finally, using (3.11), the following result holds.

Lemma 2 *If an optimal control u^* for POCP (3.5) belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$, and if*

$$\lim_{\alpha \downarrow 0} \gamma_u'(1-\alpha) = +\infty \tag{3.12}$$

then,

$$u^* \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$$

Proof: Note that the construction of γ_u in Assumption 6 makes (3.12) always satisfied. Now, remember that if, for some $\alpha > 0$, $\mu_{u^*}(\alpha) = 0$, then $u^* \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. We shall now assume that $\mu_{u^*}(\alpha) > 0$ for $\alpha > 0$ in a neighborhood of 0. If u^* does not belong to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$, then, using (3.12), for α small enough one can build a control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ such that $K(v,\varepsilon) < K(u,\varepsilon)$ because of (3.11); this contradicts the assumed optimality of u^* and concludes the proof. ■

3.2.3 First main result

We are now ready to state our first main result.

Theorem 1 (Existence of penalties providing interior optima) *Under Assumptions 2, 3, 4, 5, there exists penalty functions $\gamma_g(.)$ and $\gamma_u(.)$ such that any optimal solution u^* of POCP (3.5) belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. A constructive choice is to use penalties that satisfy Assumption 6, and which satisfy the conditions of Lemmas 1 and 2 (equations (3.8) and (3.12), respectively)*

For example, a suitable choice of penalties is:

$$\begin{aligned} \gamma_g(x) &= (-x)^{-n_g} && \text{for } x < 0 && (3.13)\\ \gamma_g(x) &= 0 && \text{for } x \geq 0 && \\ \gamma_u(u) &= -u\log(1-u) && \text{for } u \in [0,1) && (3.14)\\ \gamma_u(1) &= +\infty && && \end{aligned}$$

with $n_g > 1$.

Proof: The existence is proven constructively by showing that (3.13) and (3.14) are suitable penalties. This is done by checking that Lemmas 1 and 2 hold in the present case. Prior to this, we first easily check that both penalties satisfy Assumption 6.

The penalty (3.13) is such that equation (3.8) is satisfied, then Lemma 1 holds. Therefore, any optimal solution of POCP (3.5) belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. From Lemma 2, we know that any optimal control must actually belong to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ if $\gamma_u'(1-\alpha)$ tends to $+\infty$ when $\alpha > 0$ tends to 0. This concludes the proof. Finally, let us compute $\gamma_u'(u)$ for the choice (3.14)

$$\gamma_u'(u) = -\log(1-u) + \frac{u}{1-u}$$

Hence

$$\gamma_u'(1-\alpha) = -\log(\alpha) + \frac{1-\alpha}{\alpha}$$

which tends indeed to $+\infty$ when $\alpha > 0$ tends to 0. ■

3.3 Removing the control constraints

At this point, we have proven that, provided that suitable penalty functions are chosen, the optimal solutions of the penalized problem are interior and thus satisfy the stationarity conditions of the PMP. Nevertheless, the control constraint has not been completely relaxed at this point since the solution the POCP (3.5) are sought in the set $\mathcal{U}$. Our ultimate purpose being to solve completely unconstrained optimal control problems, we generalize the saturation function approach (introduced by Graichen et al [Gra06, GP08a, GP09, GPK08]) in §3.3.1. In §3.3.2, we study this change of variable on the control. Then, we introduce, in §3.3.3 (3.20), a new POCP that incorporates this change of variables. Next, we show that it is equivalent to the POCP (3.5). Further on, this convenient reformulation allows us, in §3.5.1, to propose a simple solving algorithm.

3.3.1 Saturation functions for convex sets

Following Graichen et al [Gra06, GP08a, GP09, GPK08], one can use saturation functions to represent some constraints (on control or state variables) in an optimal control problem. Saturation functions [Gra06] typically map $\mathbb{R}$ into the open interval $(-1,+1)$. One commonly considered saturation function is tanh(.). For example, if a variable z is such that $|z|<1$ then it can be written as $z=\tanh(\xi)$, $\xi\in\mathbb{R}$. This approach is readily generalized to dimensions higher than 1 when the constraint set has a cubic shape, e.g. $|z_1|<1, |z_2|<1, \ldots, |z_m|<1$ for some m. Simply, saturation functions are used for each coordinate.

In order to generalize saturation functions to general smooth convex sets it is handy to first consider the mapping $\psi:\mathbb{R}^m\mapsto B^m_{\|.\|}(0,1)$ such that

$$\psi(\nu)\triangleq\begin{cases}0 \text{ if } \nu=0\\ \tanh(\|\nu\|)\dfrac{\nu}{\|\nu\|} \text{ otherwise}\end{cases} \tag{3.15}$$

where $B^m_{\|.\|}(0,1)$ is the open unit ball of $\mathbb{R}^m$ for the norm $\|.\|$, e.g. the Euclidian norm. This mapping is a homeomorphism[4] and is differentiable on $\mathbb{R}^m\setminus\{0\}$. The next proposition states the generalization[5]. This generalization, formally represented by function ϕ in equation (3.16), will be used in §3.3 to deal with constraints on the control.

Proposition 10 (Generalized saturation functions) *Let $\mathcal{C}\subset\mathbb{R}^m$ be a convex set satisfying Assumption 1. The function $\phi:\mathbb{R}^m\mapsto\text{int}(\mathcal{C})$ defined by*

$$\phi(\nu)\triangleq\begin{cases}0 \text{ if } \nu=0\\ \dfrac{\tanh^2(\|\nu\|)}{G_{\mathcal{C}}(\psi(\nu))}\dfrac{\nu}{\|\nu\|} \quad \text{otherwise}\end{cases} \tag{3.16}$$

[4] whose inverse is $\psi^{-1}(u)\triangleq \text{atanh}(\| u \|)\frac{u}{\|u\|}$

[5] it is indeed a generalization, as we recover the usual saturation function from [Gra06] when the convex is an interval of $\mathbb{R}$.

where ψ is defined in (3.15), is a homeorphism. Moreover, this mapping is differentiable on $\mathrm{int}(\mathcal{C}) \setminus \{0\}$. *Its inverse is the function* $\sigma : \mathrm{int}(\mathcal{C}) \mapsto \mathbb{R}^m$ *defined by*

$$\sigma(u) \triangleq \begin{cases} 0 \text{ if } u = 0 \\ \mathrm{atanh}(G_{\mathcal{C}}(u)) \dfrac{u}{\|u\|} \text{ otherwise} \end{cases} \tag{3.17}$$

Proof: See Appendix C.4. Notations are illustrated in Fig. 3.2. ■

Proposition 10 implies that: if u belongs to $\mathrm{int}(\mathcal{C})$, then there exists $\nu \in \mathbb{R}^m$ such that $u = \phi(\nu)$ and the correspondence is one-to-one.

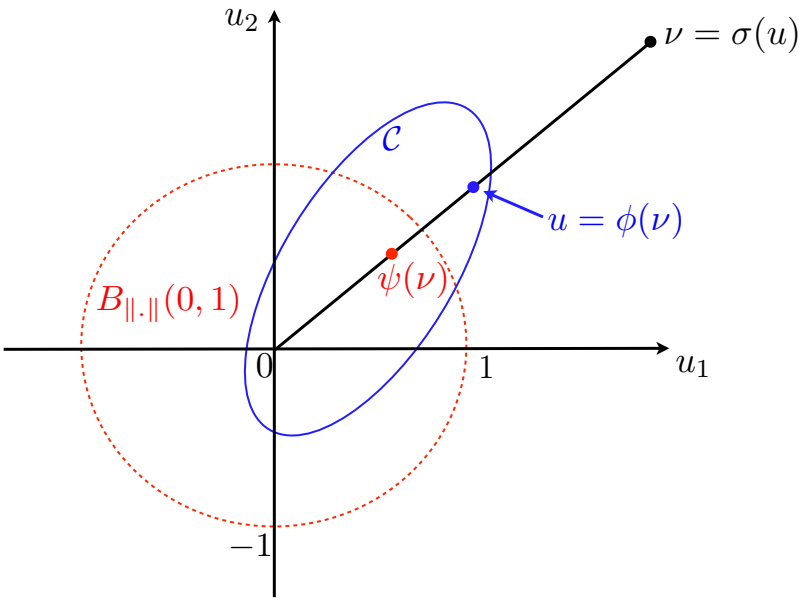

Figure 3.2: Example of generalized saturation function. On this figure, we note $\partial\mathcal{C}$ the boundary of the set $\mathcal{C}$ (the ellipse shaped set). If u belongs to $\mathrm{int}(\mathcal{C})$, then there exists $\nu \in \mathbb{R}^m$ such that $u = \phi(\nu)$ where ϕ is defined in (3.16). The correspondence is one-to-one. We say that ϕ is the saturation function associated to $\mathcal{C}$.

3.3.2 Correspondence of control sets

Let

$$\mathcal{L} \triangleq \prod_{i=1}^{p} L^{\infty}([0,T], \mathbb{R}^{m_i})$$

For each convex $\mathcal{C}_i$, define with (3.16)-(3.17) the related functions ϕ_i (3.16) and $\sigma_i = \phi_i^{-1}$ defined in equation (3.17).

Proposition 11 *We have*

$$\mathcal{L} = \left\{(\sigma_1(u_1), \dots, \sigma_p(u_p)), u \in \mathcal{U}^{\text{strict}}\right\} \tag{3.18}$$

and

$$\mathcal{U}^{\text{strict}} = \{(\phi_1(\nu_1), \dots, \phi_p(\nu_p)), \nu \in \mathcal{L}\} \tag{3.19}$$

Proof: We recall that

$$\mathcal{U}^{\text{strict}} = \left\{u = (u_1, \dots, u_p) \text{ s.t. } \operatorname*{ess\,sup}_t \max_i G_{\mathcal{C}_i}(u_i(t)) < 1\right\}$$

For $u \in \mathcal{U}^{\text{strict}}$, we shall note $G(u) = \operatorname{ess\,sup}_t \max_i G_{\mathcal{C}_i}(u_i(t))$ which is strictly smaller than 1. Hence, for $u \in \mathcal{U}^{\text{strict}}$, the $\sigma_i(u_i)$ are well defined, and $\|\sigma_i(u_i)(t)\| \leq \operatorname{atanh}(G(u)) < \infty$. This proves that the right hand-side of (3.18) is included in $\mathcal{L}$.

Conversely, let $\nu \in \mathcal{L}$ and define $u_i = \phi_i(\nu_i)$. We have $G_{\mathcal{C}_i}(u_i) = \| \psi(\nu)\| = \tanh(\|\nu_i\|) \leq \tanh(\|\nu\|_{L^\infty}) < 1$ and, hence, $u \in \mathcal{U}^{\text{strict}}$. Since $\sigma_i \circ \phi_i(\nu_i) = \nu_i$, this proves that $\mathcal{L}$ is included in the right-hand side of (3.18), yielding the desired equality (3.18).

The proof of (3.19) goes along the same lines and is simply omitted here. ■

3.3.3 Penalized problem (final version)

Finally, we define a last penalized optimal control problem

$$\min_{\nu \in \mathcal{L}} \left[P(\nu, \varepsilon) = \int_0^T \ell(x^{\phi(\nu)}, \phi(\nu)) + \varepsilon \left[\sum_{i \leq q} \gamma_g \circ g_i(x^{\phi(\nu)}) + \sum_{i \leq p} \gamma_u \circ G_{\mathcal{C}_i} \circ \phi_i(\nu_i)\right] dt\right] \tag{3.20}$$

where the penalty functions are given by equations (3.13)-(3.14), and make the following assumption

Assumption 7 *The (unconstrained) penalized problem (3.5) has at least one optimal solution.*

3.3.4 Second main result

We have the following equivalence theorem between problems (3.5) and (3.20), which is our second main result

Theorem 2 *Under the assumptions of Theorem 1 and (existence) Assumption 7, for any $\varepsilon > 0$ POCP (3.5) and POCP (3.20) are equivalent in the sense that*

$$\arg\min_{u \in \mathcal{U}} K(u, \varepsilon) = \phi\left(\arg\min_{\nu \in \mathcal{L}} P(\nu, \varepsilon)\right)$$

where $\phi(\nu)$ denotes $(\phi_1(\nu_1), \dots, \phi_p(\nu_p))$.

As a consequence, one can solve the POCP (3.5) which is constrained by $u \in \mathcal{U}$ by solving instead the unconstrained POCP (3.20), and then apply the operator ϕ to obtain an optimal solution for (3.5).

Proof: Let us consider $u^* \in \mathcal{U}$ a minimizer of $K(.,\varepsilon)$, which exists by Assumption 7. We have

$$K(u^*,\varepsilon) \leq K(u,\varepsilon), \quad \forall u \in \mathcal{U}^{\text{strict}}$$

Define $\nu^* = \sigma(u^*)$ and $\nu = \sigma(u)$. This definition is valid because both controls belong to $\mathcal{U}^{\text{strict}}$. Then, $u^* = \phi(\nu^*)$ and $u = \phi(\nu)$. Therefore,

$$K(\phi(\nu^*),\varepsilon) \leq K(\phi(\nu),\varepsilon)$$

or, equivalently,

$$P(\nu^*,\varepsilon) \leq P(\nu,\varepsilon)$$

From Proposition 11, we know that $\sigma(u)$ spans $\mathcal{L}$ when u spans $\mathcal{U}^{\text{strict}}$. Therefore, ν^* is optimal for POCP (3.20); this proves, incidentally, the existence of a solution to POCP (3.20). Since $u^* = \phi(\nu^*)$, this proves

$$\arg\min_{u\in\mathcal{U}} K(u,\varepsilon) \subset \phi\left(\arg\min_{\nu\in\mathcal{L}} P(\nu,\varepsilon)\right)$$

Now, let us consider $\nu^* \in \mathcal{L}$ a minimizer of $P(.,\varepsilon)$ (which has been proven to exist). From Proposition 11, $u^* \triangleq \phi(\nu^*) \in \mathcal{U}^{\text{strict}}$. We have

$$P(\nu^*,\varepsilon) \leq P(\nu,\varepsilon), \quad \forall \nu \in \mathcal{L}$$

From Proposition 11, this implies

$$P(\sigma(u^*),\varepsilon) \leq P(\sigma(u),\varepsilon), \quad \forall u \in \mathcal{U}^{\text{strict}}$$

i.e.

$$K(u^*,\varepsilon) \leq K(u,\varepsilon), \quad \forall u \in \mathcal{U}^{\text{strict}} \tag{3.21}$$

From Theorem 1, we know that any optimal control for $K(u,\varepsilon), u \in \mathcal{U}$ must belong to $\mathcal{U}^{\text{strict}}$. Therefore, we can substitute one of these optimal controls in place of u in (3.21); which proves that $u^* = \sigma(\nu^*)$ is optimal for POCP (3.5). Therefore,

$$\arg\min_{u\in\mathcal{U}} K(u,\varepsilon) \supset \phi\left(\arg\min_{\nu\in\mathcal{L}} P(\nu,\varepsilon)\right)$$

Finally, we have

$$\arg\min_{u\in\mathcal{U}} K(u,\varepsilon) = \phi\left(\arg\min_{\nu\in\mathcal{L}} P(\nu,\varepsilon)\right)$$

This concludes the proof. ■

3.4 Convergence of the interior point method

3.4.1 Well-posedness of the interior point method

To exploit interior point methods, it is usually considered that, in numerical implementation, the sequence of POCPs should start with relatively large value of ε (typically 1 for a suitable scaled COCP). Then, ε is decreased and a previous solution serves to initialize the solving method of the next POCP. Naturally the question of convergence of this process arises. In the following, we give sufficient conditions on the optimal control problem such that the interior points methods is well-posed in a sense defined below, which is of interest for the convergence when the penalty parameter ε is decreased.

Definition 1 (Well-posedness) *If the following condition is satisfied, the COCP (3.1) is said to be well-posed for interior point methods.*

$$\mathcal{U} \cap \mathcal{X} = \mathrm{clos}(\mathcal{U} \cap \mathcal{X}^{\mathrm{strict}})$$

where the closure is taken in the L^∞ sense.

In the following, we assume that the COCP (3.1) is well-posed for interior point methods.

3.4.2 Results on convergence

The following proof of convergence follows the exact same line as [GP09, LWR67] but one does not need to formulate the assumption on the interiority of the optimal solution anymore because it has been established in Theorem 1. First, let us note

$$\begin{aligned} \bar{J}(\nu) &= \int_0^T \ell\left(x^{\phi(\nu)}(t), \phi(\nu(t))\right) dt \\ \Gamma(\nu) &= \int_0^T \sum_{i\leq q} \gamma_g \circ g(x^{\phi(\nu)}(t)) + \sum_{i\leq p} \gamma_u \circ G_{\mathcal{C}_i} \circ \phi_i(\nu_i(t)) dt \end{aligned}$$

which gives : $P(\nu, \varepsilon) = \bar{J}(\nu) + \varepsilon\Gamma(\nu)$

Lemma 3 *Let ν_{k+1} and ν_k be the optimal controls of (3.20) for $0 < \varepsilon_{k+1} < \varepsilon_k$. Then, the following inequalities hold for the cost functional (3.20):*

$$\begin{aligned} \bar{J}(\nu_{k+1}) &\leq \bar{J}(\nu_k) \\ \Gamma(\nu_{k+1}) &\geq \Gamma(\nu_k) \\ P(\nu_{k+1}, \varepsilon_{k+1}) &\leq P(\nu_k, \varepsilon_k) \end{aligned}$$

Proof: See [GP09, LWR67]. ∎

The following theorem concerns the convergence of the cost $P(\nu_k, \varepsilon_k)$ using the results of Lemma 3.

Theorem 3 *Let (ε_k) be a decreasing sequence of positive penalty parameters with $\lim_{k\to\infty}\varepsilon_k = 0$. Then, $P(\nu_k, \varepsilon_k)$ converges to the optimal cost J^* of COCP (3.1)*

$$\lim_{k\to\infty} P(\nu_k, \varepsilon_k) = J^*$$

with

$$\lim_{k\to\infty} \bar{J}(\nu_k) = J^*, \quad \lim_{k\to\infty} \varepsilon_k \Gamma(\nu_k) = 0$$

Proof: See [GP09, LWR67]. ■

To prove the convergence of the states, we require the following assumption

Assumption 8 *The cost functional $J(u)$ of COCP (3.1) satisfies the strong convexity property:*

$$D \parallel u - v \parallel_{L^2}^2 \leq J(u) + J(v) - 2J\left(\frac{u+v}{2}\right) \ \forall u, v \in \mathcal{U} \cap \mathcal{X}$$

for some $D > 0$. Moreover, the optimal control u^* of problem (3.1) is assumed to be unique.

Theorem 4 *If Assumption 8 holds, the input $u_k = \phi(\nu_k)$ as well as $x^{u_k} = x^{\phi(\nu_k)}$ solutions of POCP (3.20) converge to the optimal trajectory (u^*) of problem (3.1) in the following sense*

$$\lim_{k\to\infty} \parallel u_k - u^* \parallel_{L^2} = 0, \quad \lim_{k\to\infty} \parallel x^{u_k} - x^{u^*} \parallel_{L^\infty} = 0$$

Proof: See [GP09, LWR67]. ■

3.5 Solving algorithms

The purpose of the main results of this chapter, i.e. Theorems 1 and 2 respectively, is to allow one to solve, for a decreasing sequence of $\varepsilon_k > 0$ that tends to 0, a sequence of simple (unconstrained) POCPs (Problem (3.20)) instead of POCP (3.5) because they are equivalent. In §3.4 it has been recalled that, under classic strong convexity assumptions, the sequence of states and controls converge (in relevant topologies) to the optimal solutions of COCP (3.1). Thus, using suitable penalties and saturation functions, one can solve the original COCP (3.1) by solving a sequence of unconstrained problems.

3.5.1 Indirect method

The indirect method proposed here is based on the solving of the unconstrained optimality conditions of PMP by using collocation. Each POCP (3.20) penalized

by ε belonging to a sequence (ε_n) can be solved using the calculus of variations. Define the Hamiltonian of the penalized problem (3.20) as follows

$$H_\varepsilon(x^{\phi(\nu)},\nu,p) \triangleq \ell(x^{\phi(\nu)},\phi(\nu)) + \varepsilon\left[\sum_{i\leq q}\gamma_g\circ g_i(x^{\phi(\nu)}) + \sum_{i\leq p}\gamma_u\circ G_{\mathcal{C}_i}\circ\phi_i(\nu_i)\right] + p^t f(x^{\phi(\nu)},\phi(\nu))$$

where $p\in\mathbb{R}^n$ is the adjoint state solution of $\frac{dp}{dt} = -\frac{\partial H_\varepsilon}{\partial x^{\phi(\nu)}}$, $p(T)=0$ and where the penalty functions are chosen according to Theorem 1. Now, defining a positive decreasing sequence, one can approach the solution of COCP (3.1).

- Step 1: Initialize the continuous functions $x^{\phi(\nu)}(t)$ and $p(t)$ such that the initial values satisfy $g_i(x^{\phi(\nu)}(t)) < 0$ for all $t\in[0,T]$, and set $\varepsilon=\varepsilon_0$. Note that $x^{\phi(\nu)}(t)$ and $p(t)$ need not satisfy any differential equation at this stage, even if it is better if they do.
- Step 2: Solve for each time $\frac{\partial H_\varepsilon}{\partial\nu}=0$, and note ν^*_ε the solution.
- Step 3: Solve the $2n$ differential equations $\frac{dx^{\phi(\nu)}}{dt} = f(x^{\phi(\nu)},\phi(\nu^*_\varepsilon))$ and $\frac{dp}{dt} = -\frac{\partial H_\varepsilon}{\partial x^{\phi(\nu)}}(x^{\phi(\nu)},\nu^*_\varepsilon,p)$ forming a two point boundary values problem using e.g. bvp5c or bvp4c (see [SKR00]), with the following boundary constraints $x^{\phi(\nu)}(0)=x_0$ and $p(T)=0$.
- Step 4: Decrease ε, initialize $x^{\phi(\nu)}(t)$ and $p(t)$ with the solutions found at Step 3 and restart at Step 2.

3.5.2 Direct method

Solving the problem with direct methods does not involve any adjoint vector p and does not rely on the calculus of variations. To compute the solution using these methods, let us first introduce the augmented system as follows

$$\begin{cases} \dot{x}^{\phi(\nu)} &= f(x^{\phi(\nu)},\phi(\nu)) \\ \dot{z} &= \ell(x^{\phi(\nu)},\phi(\nu)) + \varepsilon\left[\sum_{i\leq q}\gamma_g\circ g_i(x^{\phi(\nu)}) + \sum_{i\leq p}\gamma_u\circ G_{\mathcal{C}_i}\circ\phi_i(\nu_i)\right] \end{cases}$$

with the following initial conditions: $x^{\phi(\nu)}(0)=x_0$ and $z(0)=0$. Direct methods consist in transforming an infinite dimensional optimization problem into a finite dimension one. To do so, a time discretization is chosen which defines the mapping

$$x^{\phi(\nu)}(t), z(t), \nu(t) \mapsto x^{\phi(\nu)}[0\dots N], z[0\dots N], \nu[0\dots N]$$

where $[0\cdots N]$ are indexes for mesh points spread over $[0,T]$. This time discretization relies on numerical schemes such as Euler, Gauss, Runge-Kunta. Several numerical methods are fully described in [But08]. This time discretization transforms the dynamical constraints into $(N+1)\times(n+1)$ equality constraints, where $N+1$ is the number of collocation points and n the dimension of the state vector $x^{\phi(\nu)}(t)$. The direct method algorithm is thus the following:

- Step 1: Initialize $x^{\phi(\nu)}[0 \dots N]$, $z[0 \dots N]$, $\nu[0 \dots N]$ such that the state constraints are satisfied at each collocation point and set $\varepsilon = \varepsilon_0$.
- Step 2: Solve the following optimization problem

$$\min_{x^{\phi(\nu)}[0 \cdots N], z[0 \cdots N], \nu[0 \cdots N]} z[N]$$

  under the $(N+1) \times (n+1)$ equality constraints corresponding to the time discretization of the dynamical constraints.
- Step 3: Decrease ε, initialize $x^{\phi(\nu)}[0 \dots N], z[0 \dots N], \nu[0 \dots N]$ with the values computed at Step 2 and start over at Step 2.

Chapter 4

Illustrative numerical examples

Contents

In this chapter we propose a series of examples to illustrate the methodology exposed in the previous chapter.

4.1 Toy Problem

As a first example to illustrate the methodology, we consider a simple double integrator with a position and control constraints

$$\begin{cases} \min & \displaystyle\int_0^5 x^2(t)dt \\ \ddot{x} = u \\ |u| \leq 1 \\ x \geq 0.2 \\ x(0) = 1 \\ \dot{x}(0) = 0 \end{cases}$$

4.1.1 Indirect method

By a direct application of (POCP) formulation (3.20), we now consider the following POCP, $\varepsilon > 0$

$$\min_{\nu} \int_0^5 x^2 + \varepsilon\big(\gamma_g(x - 0.2) + \gamma_u \circ \phi(\nu)\big)dt$$

where $\gamma_g(x) \triangleq (x)^{-n_g}$, $n_g > 1$, according to (3.13) and γ_u and ϕ are now defined. The stationarity condition $H_\nu = 0$ yields the equation $\gamma_u' \circ \phi(\nu) = -\frac{\mu}{\varepsilon}$ obtained by

differentiating the Hamiltonian of our POCP

$$H \triangleq x^2 + \varepsilon\big(\gamma_g(x-0.2) + \gamma_u \circ \phi(\nu)\big) + p_1\dot{x} + p_2\phi(\nu)$$

Consistently with equation (3.16) where we set $\| \, . \, \| = G_{\mathcal{C}}(.) = |.|$, we have

$$\phi(\nu) \triangleq \tanh(\nu)$$

Then, consistently with Lemma 2, a sufficient condition to guarantee that Theorem 2 holds is that $\lim_{\nu\to\pm\infty} \gamma_u' \circ \phi(\nu) = \pm\infty$. This last condition leaves us with a vast choice of functions for γ_u and ϕ. Conveniently, in order to easily solve the $H_\nu = 0$ condition for the unknown ν, we choose

$$\gamma_u' \circ \phi(\nu) = \sinh(\nu)$$

Then, one simply has to formulate the two-point boundary value problem for OCP

$$\begin{cases} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \phi\left(-\text{asinh}\left(\frac{p_2}{\varepsilon}\right)\right) \\ \dot{p_1} &= -2x_1 + \varepsilon\gamma_g'(x_1 - 0.2) \\ \dot{p_2} &= -p_1 \\ x_1(0) &= 1 \\ x_2(0) &= 0 \\ p_1(5) &= 0 \\ p_2(5) &= 0 \end{cases}$$

For every $\varepsilon > 0$, according to Theorem 2, the obtained solution gives the solution ν_ε^* which is such that $\phi(\nu_\varepsilon^*) = u_\varepsilon^*$ where $u_\varepsilon^* = -\tanh\left(\text{asinh}\left(\frac{p_2}{\varepsilon}\right)\right)$ is solution for POCP (3.5) which is interior. To solve this problem, the sequence (ε_n) is logarithmically decreasing from 1 to 10^{-7}. The code uses one tuning parameter $n_g > 1$ which is set to $n_g = 1.1$ in this example. The final solution is obtained on a (not equally distributed) mesh of 192 points automatically generated by the two-point boundary value problem solved by bvp5c. The histories of optimal state $x_\varepsilon^*(t)$ for $\varepsilon = 1 \ldots 10^{-7}$ is displayed on Figure 4.1, the corresponding histories of optimal control is displayed on Figure 4.2, and the histories of the first adjoint states $p_{1,\varepsilon}^*(t)$ is displayed on Figure 4.3. The script file implemented in Matlab, is available at `http://cas.ensmp.fr/~petit/code_optimisation_PM/`.

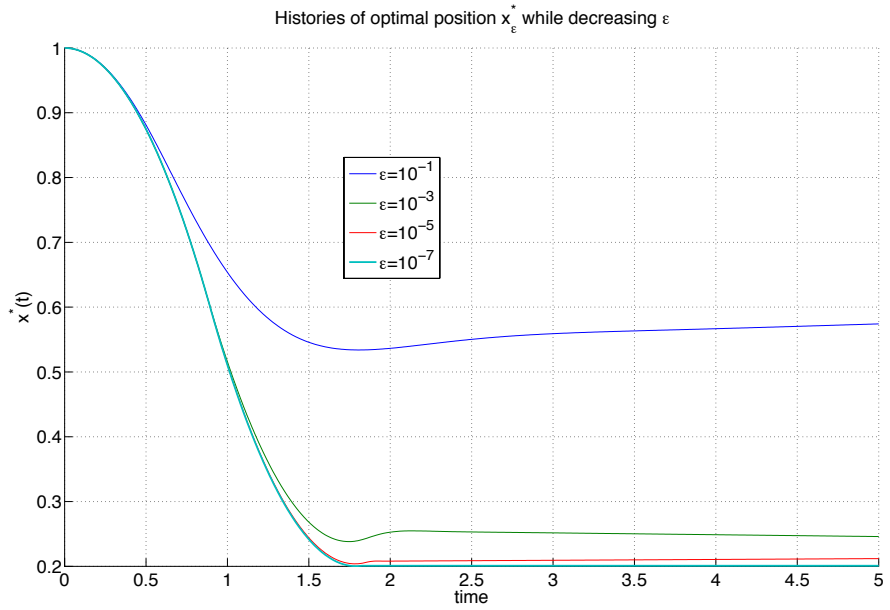

Figure 4.1: Optimal state $x^*_\varepsilon(t)$ for $\varepsilon = 10^{-7}$ (indirect method).

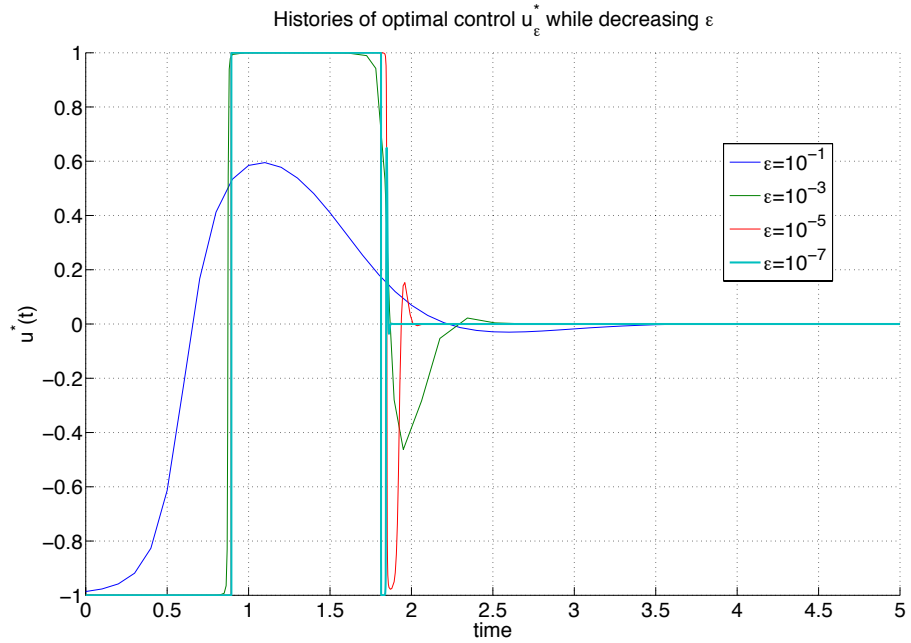

Figure 4.2: Optimal control $u^*_\varepsilon(t) = \tanh(\nu^*_\varepsilon(t))$ for $\varepsilon = 10^{-7}$ (indirect method).

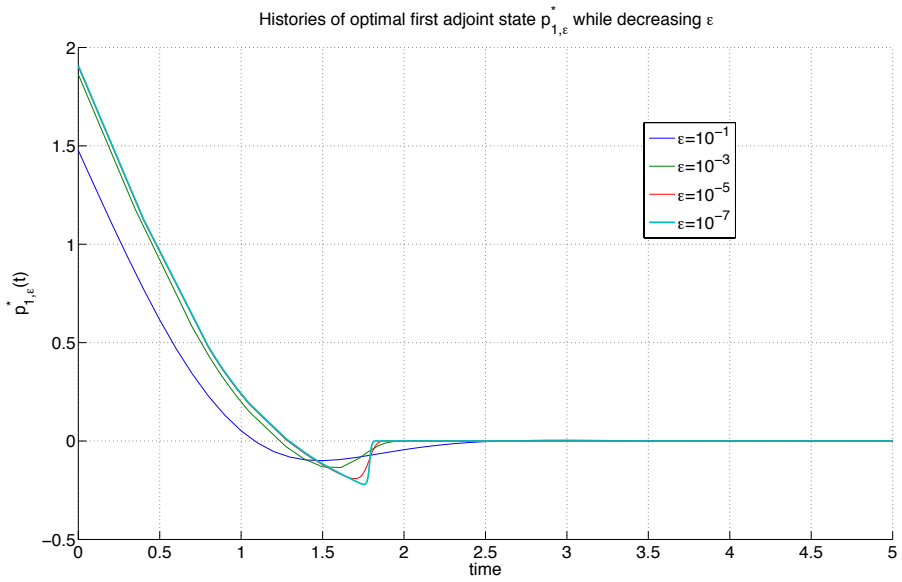

Figure 4.3: Optimal adjoint states $p_1(t)$ for $\varepsilon = 10^{-7}$ (indirect method).

4.1.2 Direct method

As described in Section 3.5.2, direct methods use a discretization scheme to yield a finite dimensional optimization problem. Then, to solve this finite dimensional optimization problem numerous softwares are available. In this Section we use two optimization softwares both based on the IPOPT [WB06] solver.

4.1.2.1 Using BOCOP

We first use the package Scilab/BOCOP [MGB12] (see `http://bocop.org`) by Martinon, Grélat and Bonnans. We choose the explicit 4^{th} order Runge-Kutta discretization scheme provided by this software with 500 equally distributed mesh points. The sequence of $(\varepsilon_n)_{n=1\cdots 40} = 1.5^{-n}$ is considered. The evolution of the optimal cost (3.20) while decreasing ε is displayed on Figure 4.4. The optimal state $x^*(t)$ for $\varepsilon = 1.5^{-40}$ is displayed on Figure 4.5, and the corresponding optimal control $u^*(t)$ is displayed on Figure 4.6.

4.1.2.2 Using AMPL

To solve the direct problem, we use the software AMPL [FGK90] with the solver IPOPT [WB06]. We choose the 3-stage Lobatto IIIa discretization formula (which is also employed in bvp4c) with 500 equally distributed mesh points. The sequence of $(\varepsilon_n)_{n=1\cdots 40} = 1.5^{-n}$ is considered. The optimal state $x^*(t)$ for $\varepsilon = 1.5^{-40}$ is

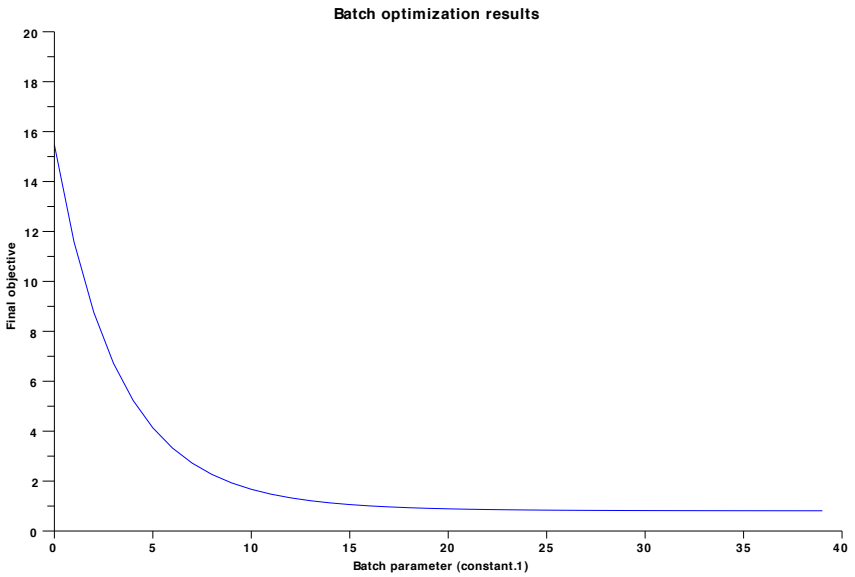

Figure 4.4: Histories of optimal values of the penalized cost (3.20) for decreasing values of ε (direct method with BOCOP).

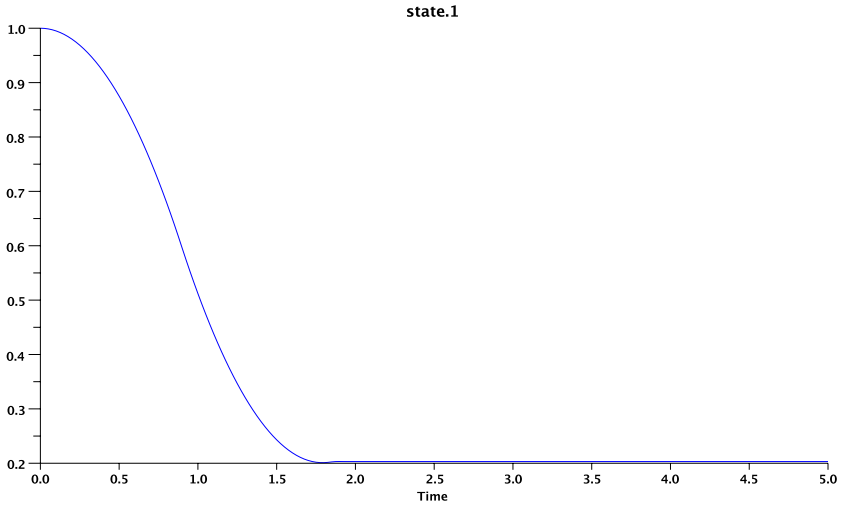

Figure 4.5: Optimal position $x_\varepsilon^*(t)$ for $\varepsilon = 10^{-11}$ (direct method with BOCOP).

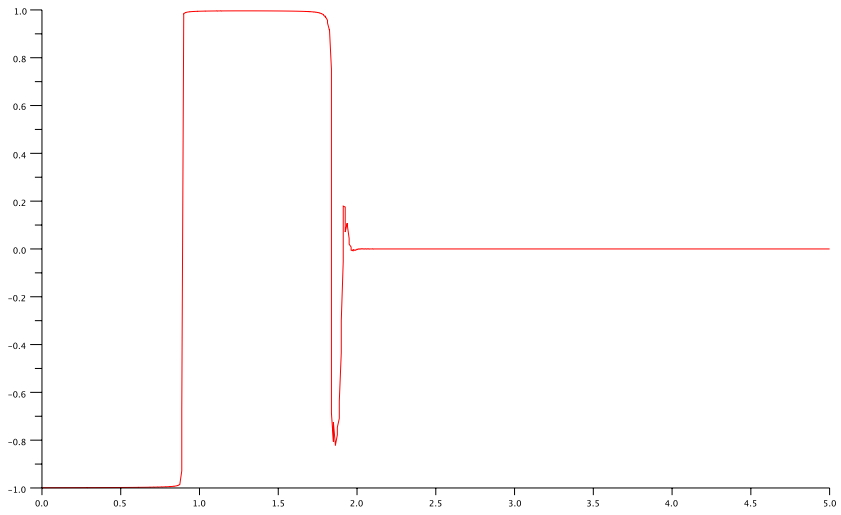

Figure 4.6: Optimal control $u^*_\varepsilon(t) = \tanh(\nu^*_\varepsilon(t))$ for $\varepsilon = 10^{-11}$ (direct method with BOCOP).

displayed on Figure 4.7, and the corresponding optimal control $u^*(t)$ is displayed on Figure 4.8.

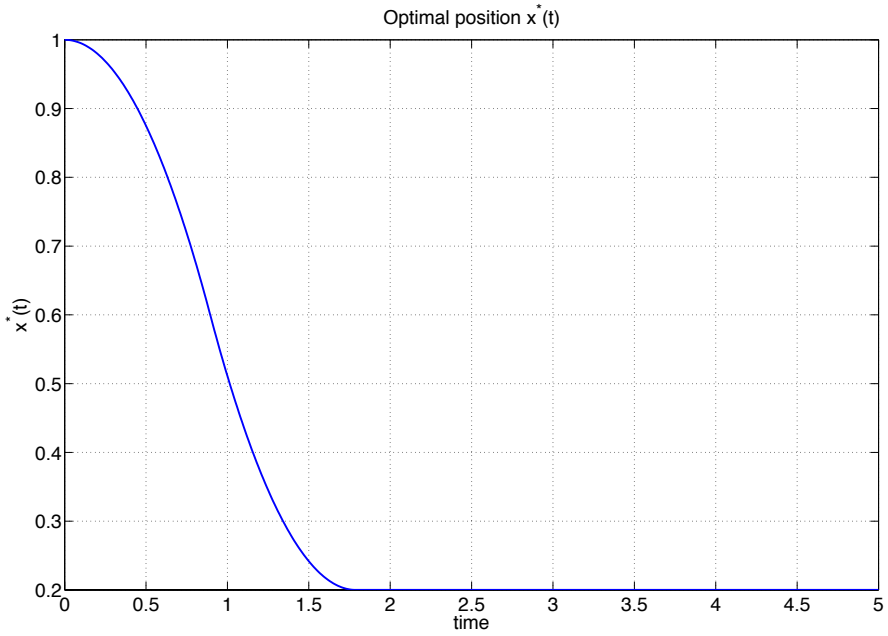

Figure 4.7: Optimal position $x_\varepsilon^*(t)$ for $\varepsilon = 1.5^{-40}$ (direct method with AMPL).

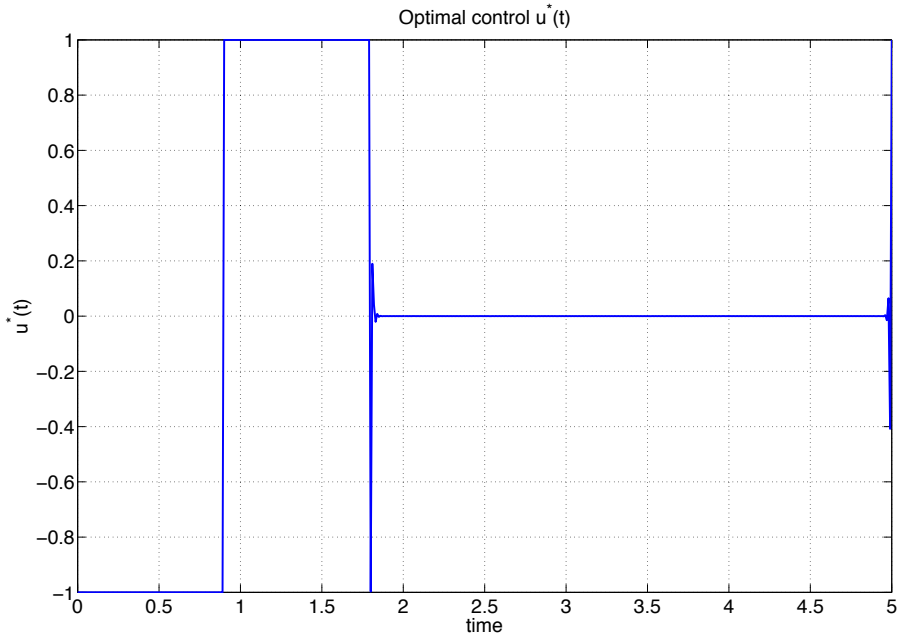

Figure 4.8: Optimal control $u_\varepsilon^*(t)$ for $\varepsilon = 1.5^{-40}$ (direct method with AMPL).

4.2 Goddard's problem

The historical Goddard's problem (first presented in 1919 [God19]) is the maximization of the final altitude of a rocket flying in vertical direction. The problem has become a benchmark in optimal control due to a characteristic singular-constrained arc behavior in connection with a relatively simple model structure, which makes the Goddard's rocket an ideal object of study, see [BMT08, GP08b, Rug06].

4.2.1 Problem statement

4.2.1.1 Model equations

The equations of motion of the rocket are given by the ordinary differential equations

$$\begin{aligned} \dot{h} &= v \\ \dot{v} &= \frac{u - D(h,v)}{m} - \frac{1}{h^2} \\ \dot{m} &= -\frac{u}{c} \end{aligned} \tag{4.1}$$

with h the altitude, v the upward velocity, and m the mass of the rocket. The states h, v, m, the thrust u as the input of the system, and the time t are commonly normalized and dimension–free. The drag function $D(h,v)$ is given by

$$D(h,v) = q(h,v)\frac{C_D A}{m_0 g}$$

as a function of the Earth's gravitational acceleration g and the dynamic pressure

$$q(h,v) = \frac{1}{2}\rho_0 v^2 e^{\beta(1-h)}$$

depending on the altitude h and the velocity v. The constants in the model equations are

$$\begin{aligned} C_D &\text{ drag coefficient}, & \rho_0 &\text{ air density at sea level}, \\ A &\text{ reference area}, & \beta &\text{ density decay rate}, \\ m_0 &\text{ initial mass}, & c &\text{ exhaust velocity} \end{aligned}$$

The following values are taken from [GP08b, Sey94]:

$$\beta = 500,\ c = 0.5,\ \frac{\rho_0 C_D A}{m_0 g} = 620,\ g = 9.81$$

4.2.1.2 Constrained optimal control problem

The optimal control problem is the following:

$$\min_u -h(T)$$

under the dynamics (4.1) and the following state and input constraints

$$\begin{array}{rcl} u(t) & \in & [0; 3.5] \text{ a.e. } t \in [0, T] \\ q(h(t), v(t)) & \leq & 10, \ \forall t \in [0, T] \end{array}$$

where the final time T is a free parameter. First, one can reformulate the problem as a fixed horizon optimal control problem. To do so we make the following change of variable $\tau = \frac{t}{T}$ and obtain the following augmented dynamics

$$\begin{cases} \dot{h} = Tv \\ \dot{v} = T\left[\dfrac{u - D(h,v)}{m} - \dfrac{1}{h^2}\right] \\ \dot{m} = -T\dfrac{u}{c} \\ \dot{T} = 0 \end{cases} \tag{4.2}$$

and the optimal control problem becomes

$$\min_u - \int_0^1 Tv dt$$

under the aforementioned augmented dynamics (4.2) and the following constraints:

$$\begin{cases} u(t) & \in & [0; 3.5] \text{ a.e. } t \in [0, 1] \\ q(h(t), v(t)) & \leq & 10, \ \forall t \in [0, 1] \end{cases} \tag{4.3}$$

In this case, 0 does not belong to the interior of the admissible control set. To overcome this difficulty, we simply use two invertible changes of variables: $\phi : \mathbb{R} \mapsto (-1, 1)$ and $\psi : (-1, 1) \mapsto (u_{\min}, u_{\max})$ defined by

$$\begin{array}{rcl} \phi(\nu) & \triangleq & \tanh\left(\dfrac{2\nu}{u_{\max} - u_{\min}}\right) \\ \psi(z) & \triangleq & \dfrac{u_{\max} - u_{\min}}{2}(z + 1) + u_{\min} \end{array}$$

Using these change of variables we have

$$u \triangleq \psi \circ \phi(\nu)$$

Now, we introduce a control penalty in POCP (3.20) of the form $\gamma_u \circ G_{[-1,1]} \circ \phi(\nu)$ where γ_u remains to be chosen and $G_{[-1,1]}$ is the gauge function of $[-1, 1]$ which is simply $|.|$. According to the formulation in (3.20), the Hamiltonian of the POCP corresponding to this problem is the following ($\varepsilon > 0$):

$$\begin{array}{rcl} H(h, v, m, T, \nu, p_h, p_v, p_m, p_T) & \triangleq & T\Bigg[- v + p_h v + p_v \left[\dfrac{\psi \circ \phi(\nu) - D(h,v)}{m} - \dfrac{1}{h^2}\right] \\ & & -p_m \dfrac{\psi \circ \phi(\nu)}{c}\Bigg] \\ & & +\varepsilon \left[\gamma_u \circ G_{[-1,1]} \circ \phi(\nu) + \gamma_g(q(h, v) - 10)\right] \end{array}$$

The two point boundary value problem consists in solving the followings ODEs

$$\begin{cases} \dot{h} = Tv & \dot{v} = T\left[\frac{\psi\circ\phi(\nu^*)-D(h,v)}{m} - \frac{1}{h^2}\right] & \dot{m} = -T\frac{\psi\circ\phi(\nu^*)}{c} & \dot{T} = 0 \\ \dot{p}_h = -\frac{\partial H}{\partial h} & \dot{p}_v = -\frac{\partial H}{\partial v} & \dot{p}_m = -\frac{\partial H}{\partial m} & \dot{p}_T = -\frac{\partial H}{\partial T} \\ h(0) = 1 & v(0) = 0 & m(0) = 1 & m(1) = 0.6 \\ p_h(1) = 0 & p_v(1) = 0 & p_T(0) = 0 & p_T(1) = 0 \end{cases}$$

where ν^* is solution of $\frac{\partial H}{\partial \nu} = 0$. To compute this solution, first let us consider the function $\gamma_u \circ G_{[-1,1]} \circ \phi(\nu)$ which writes

$$\gamma_u \circ G_{[-1,1]} \circ \phi(\nu) = \gamma_u \circ \phi(\nu)$$

where $\gamma_u : (-1,1) \mapsto \mathbb{R}^+$ is a smooth symmetric function. The control penalty is differentiable with respect to ν and we have

$$\frac{\partial H}{\partial \nu} = T\left[p_v \frac{\phi'(\nu)\psi' \circ \phi(\nu)}{m} - p_m \frac{\phi'(\nu)\psi' \circ \phi(\nu)}{c}\right] + \varepsilon\phi'(\nu)\gamma_u' \circ \phi(\nu)$$

Therefore, ν^* is the solution of

$$0 = T\left[\frac{p_v}{m} - \frac{p_m}{c}\right] + \varepsilon\frac{2}{3.5}\gamma_u' \circ \phi(\nu) \tag{4.4}$$

From Lemma 2 and the symmetry of γ_u, we know that the solution are interior as soon as γ_u' is a bijective increasing mapping from $(-1,1)$ to $\mathbb{R}$. Moreover, $\phi(\nu)$ being an increasing bijective mapping from $\mathbb{R}$ to (-1,1), one can simply choose the following parameterization of the control penalty

$$\gamma_u' \circ \phi(\nu) \triangleq \sinh(\nu)$$

which is a bijective increasing mapping from $\mathbb{R}$ to $\mathbb{R}$. Thanks to this choice, equation (4.4) has an analytical solution

$$\nu^* = \sinh^{-1}\left(-\frac{T}{\varepsilon}\left(\frac{p_v}{m} - \frac{p_m}{c}\right)\right)$$

One can notice that γ_u need not be defined analytically. The problem is initialized with constant values of the variables as follows

$$\begin{array}{cccc} h(t) = 1 & v(t) = 0.2 & m(t) = 1 & T = 0.5 \\ p_h(t) = 0 & p_v(t) = 1 & p_m(t) = 0 & p_T(t) = 0 \end{array}$$

The sequence (ε_n) is initialized with $\varepsilon_0 = 10^{-2}$, the parameter n_g from equation (3.13) is set at $n_g = 1.1$. Moreover, to initialize the problem the state constraint for the first value of ε is $q(h,v) \leq 15$, then, for the rest of the sequence (ε_n) the constraints is set exactly as described in equation (4.3). To solve each two-point boundary value problem of the sequence, we use the MATLAB implementation of collocation code bvp5c. The script file, is available at `http://cas.ensmp.fr/~petit/code_optimisation_PM/`.

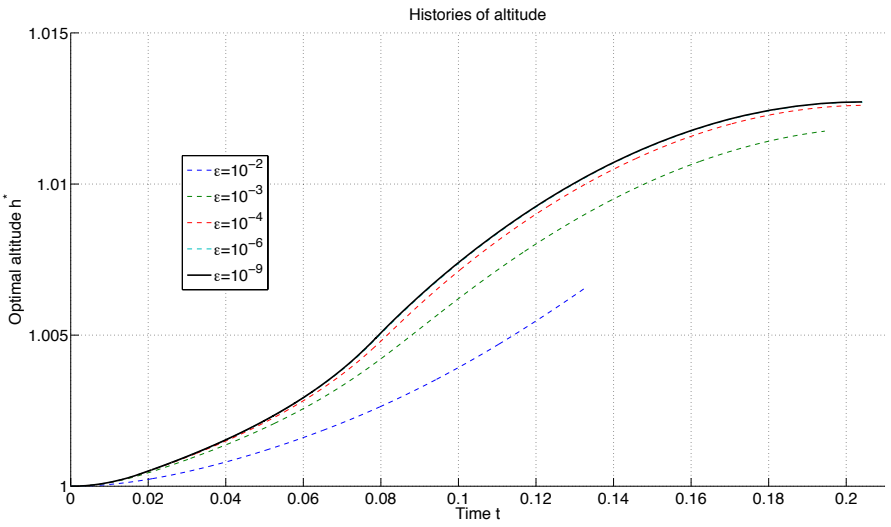

Figure 4.9: Histories of optimal altitude for decreasing values of ε.

4.2.2 Results

In Figures 4.9 to 4.13, histories of state variables, thrust and state constraint are given for decreasing values of the parameter ε. One can see that these solutions are very similar to those reported in [GP08b]. Moreover, the optimal final time and the optimal value of the criterion are the following:

$$T = 0.20405546 \quad ; \quad h(T) = 1.01271747$$

4.3 A multivariable Linear Quadratic Problem

4.3.1 Problem statement

Consider the following optimal control problem

$$J = \int_0^T \frac{1}{2} \left(u_1^2(t) + u_2^2(t) + x^2(t) + y^2(t) \right) dt$$

with T fixed, for the following dynamics

$$\ddot{x}(t) = u_1(t) \quad ; \quad \ddot{y}(t) = u_2(t)$$

having the following initial conditions

$$x(0) = y(0) = 5 \quad ; \quad \dot{x}(0) = \dot{y}(0) = 0$$

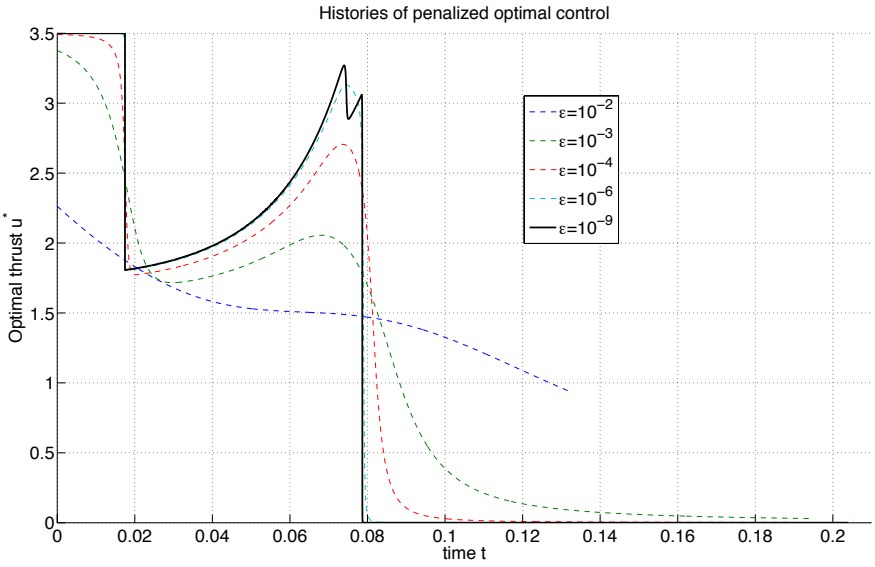

Figure 4.10: Histories of optimal thrust for decreasing values of ε.

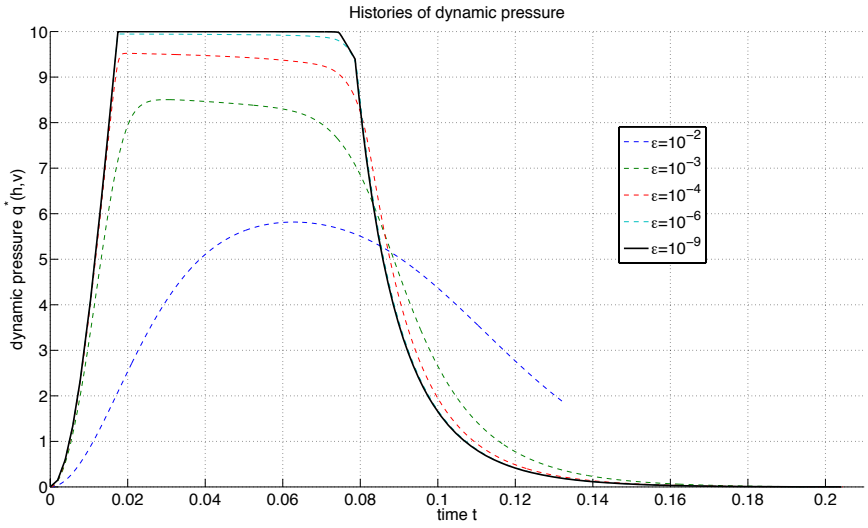

Figure 4.11: Histories of optimal dynamic pressure $q(h^*_\varepsilon, v^*_\varepsilon)$ for decreasing values of ε.

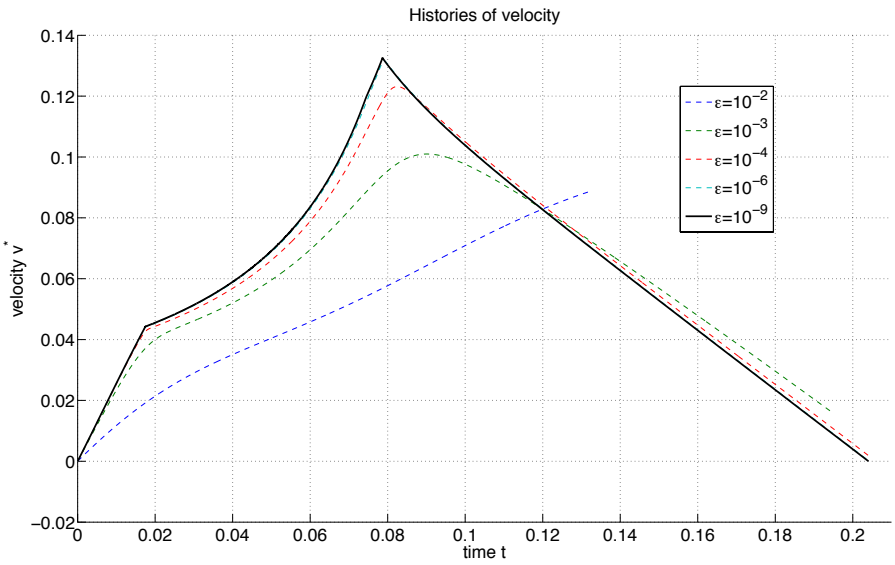

Figure 4.12: Histories of optimal velocity for decreasing values of ε.

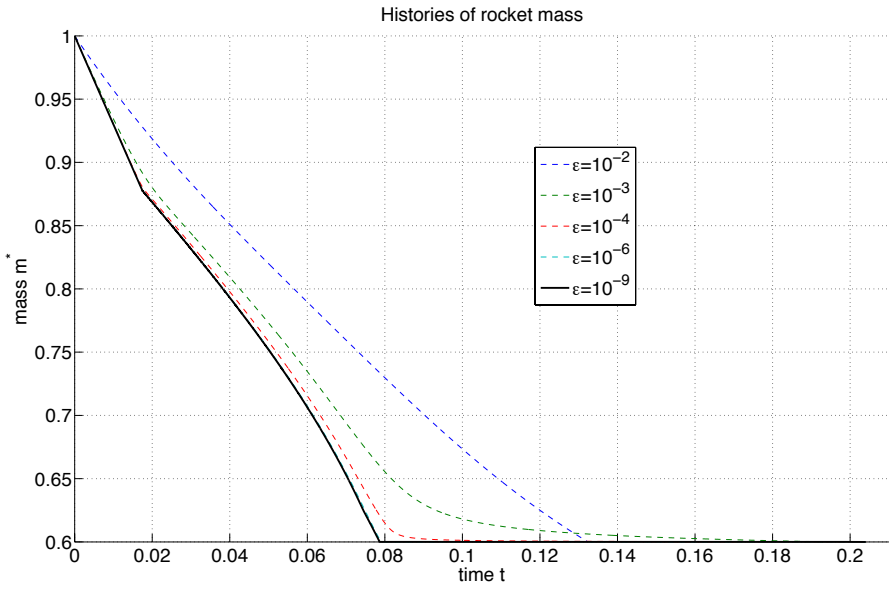

Figure 4.13: Histories of optimal mass for decreasing values of ε.

under the following path constraints:

$$\begin{aligned} 0 &\geq g_1(x(t)) \triangleq x(t) - 2 - 1.5\,[\sin(0.2t) + \sin(0.2\pi t)] \\ 0 &\geq g_2(y(t)) \triangleq y(t) - 1.5 - 1.5\,[\sin(0.2t + 5.5) + \sin(0.2\pi t + 5.5)] \\ 1 &\geq \sqrt{u_1^2(t) + u_2^2(t)} \end{aligned}$$

The input constraint, which requires that the vector $(u_1(t), u_2(t))$ belongs to the unit ball, couples the two variables. We use our generalization of saturation function of §3.3.1, and introduce the vector change of variables (3.16) where $G_{\mathcal{C}} \triangleq \|.\|$ is the Euclidian norm

$$\begin{aligned} u_1(t) &\triangleq \phi_1(\nu) = \frac{\tanh(\| \nu(t) \|)}{\| \nu(t) \|}\nu_1(t) \\ u_2(t) &\triangleq \phi_2(\nu) = \frac{\tanh(\| \nu(t) \|)}{\| \nu(t) \|}\nu_2(t) \end{aligned}$$

with $\nu = (\nu_1, \nu_2)^t$. Using this vector change of variables in the POCP formulation (3.20), we obtain the following Hamiltonian, presented in Theorem 1

$$\begin{aligned} H(x, \dot{x}, y, \dot{y}, \nu, p) \triangleq\ & \tanh(\| \nu \|)^2 + x^2 + y^2 + p_1\dot{x} + p_2\phi_1(\nu) + p_3\dot{y} + p_4\phi_2(\nu) \\ & + \varepsilon\big(\gamma_g \circ g_1(x) + \gamma_g \circ g_2(x) - \tanh(\| \nu \|)\log(1 - \tanh(\| \nu \|))\big) \end{aligned}$$

with $\gamma_g(x) = (-x)^{-n_g}$, $n_g > 1$, and $\gamma_u \circ G_{\mathcal{C}} \circ \phi(\nu) = -\tanh(\| \nu \|)\log(1 - \tanh(\| \nu \|))$ because

$$\begin{aligned} \gamma_u(u) &= -u\log(1 - u) \\ G_{\mathcal{C}}(u) &= \| u \| \\ \phi(\nu) &= \tanh(\| \nu \|)\frac{\nu}{\| \nu \|} \end{aligned}$$

The adjoint vector p satisfies the following differential equations

$$\left\{ \begin{array}{rcl} \dot{p}_1 &=& -\varepsilon\gamma_g' \circ g_1 \\ \dot{p}_2 &=& -p_1 \\ \dot{p}_3 &=& -\varepsilon\gamma_g' \circ g_2 \\ \dot{p}_4 &=& -p_3 \end{array} \right.$$

with the following boundary conditions, $T = 14$, $p_i(T) = 0$, $i = 1 \ldots 4$. The optimal unconstrained control ν^* satisfies the following algebraic equations

$$\frac{\partial H}{\partial \nu}(x, \dot{x}, y, \dot{y}, \nu^*, p) = 0$$

To solve this problem we use a self-developed collocation code[1] for two-point boundary value problems of differential and algebraic equations (index 1). This collocation code uses a 3-stage Lobatto IIIa formula (also employed in bvp4c) for the differential variables (x^ϕ, p) and a simple interpolation of order 1 for the algebraic variable (ν). Collocation method equations are solved using the software IPOPT [WB06].

[1] which is available for internal use only at EDF R&D

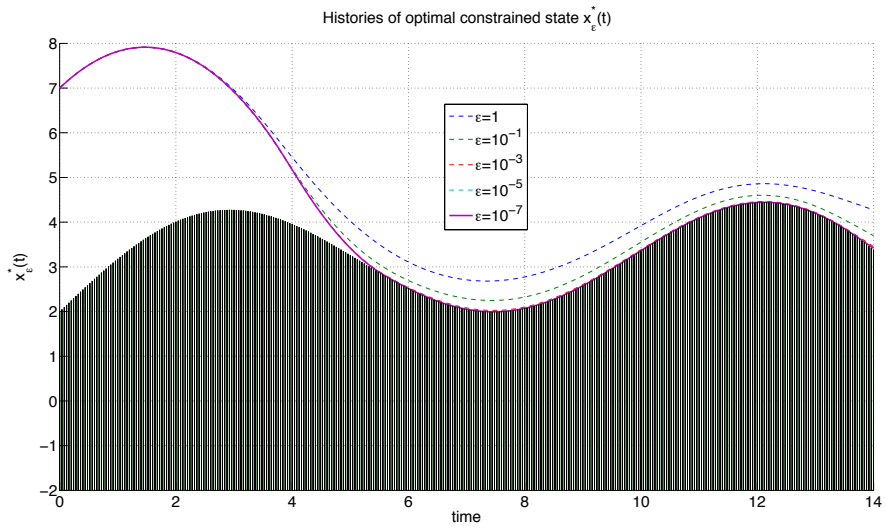

Figure 4.14: Histories of optimal $x_\varepsilon^*(t)$ for decreasing values of ε. The dark domain is forbidden.

4.3.2 Results

From Figures 4.14 to 4.20, histories of constrained state variables, controls and control constraint are given for decreasing values of the parameter ε. The optimal value of the criterion is the following

$$J^* = 252.2082$$

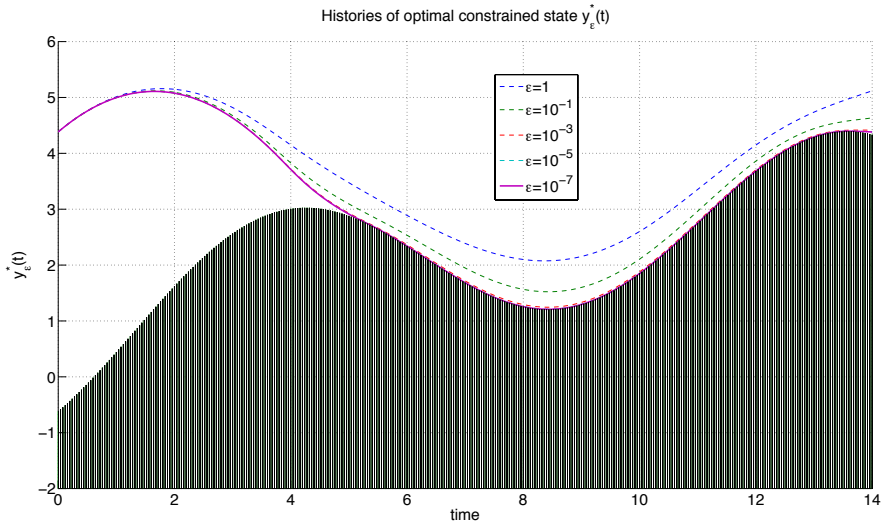

Figure 4.15: Histories of optimal $y^*_\varepsilon(t)$ for decreasing values of ε. The dark domain is forbidden.

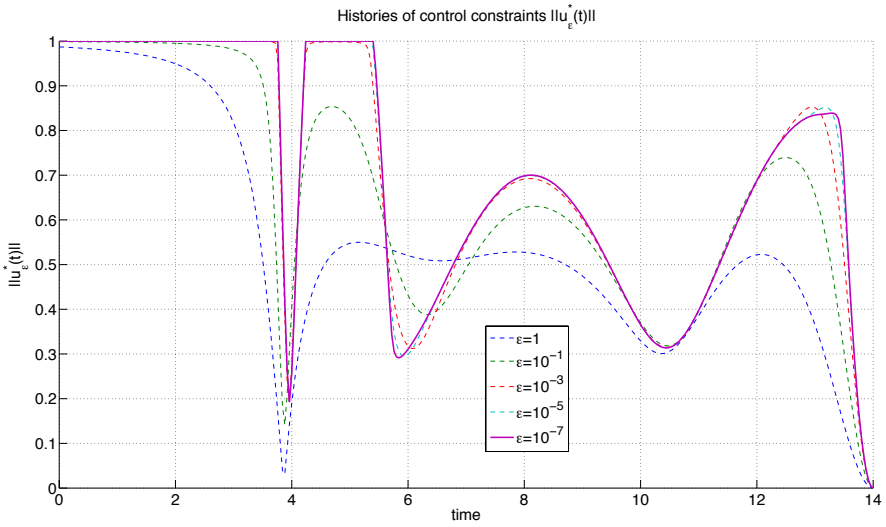

Figure 4.16: Histories of optimal $\| u^*_\varepsilon \|$ for decreasing values of ε. The norm is required to remain inferior to 1 for all times.

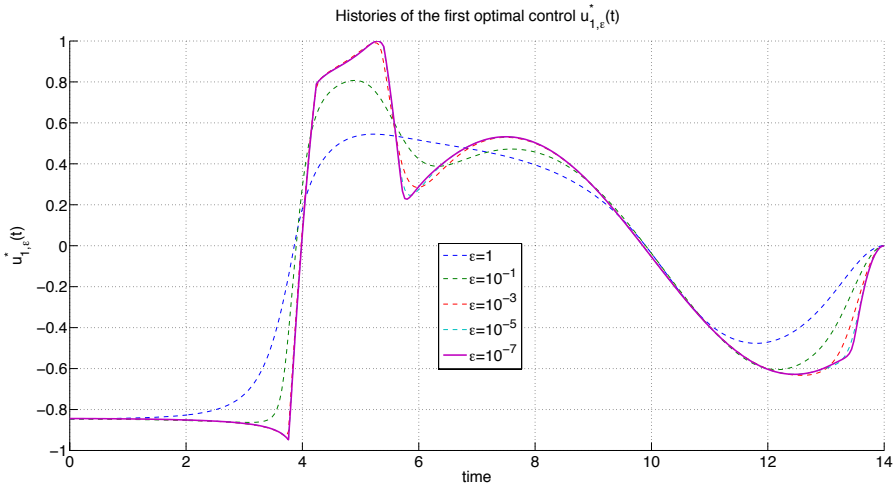

Figure 4.17: Histories of the first optimal control $u^*_{1,\varepsilon}(t)$ for decreasing values of ε.

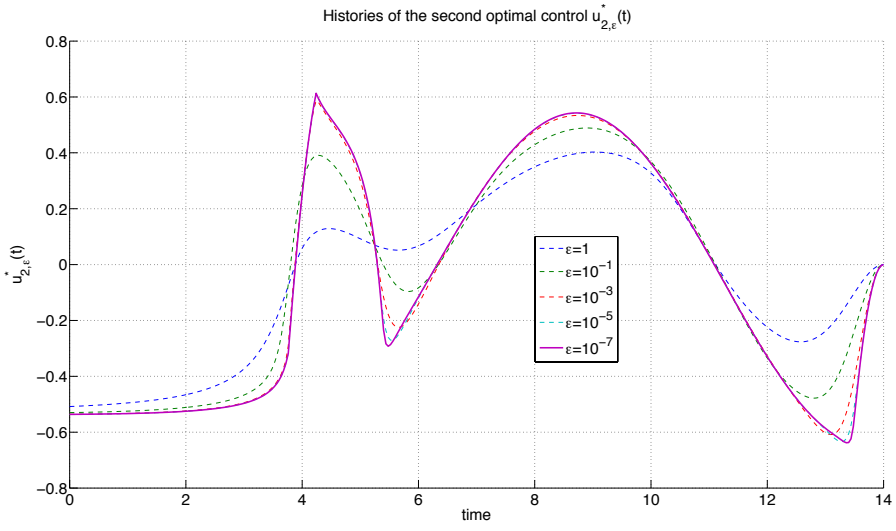

Figure 4.18: Histories of the second optimal control $u^*_{2,\varepsilon}(t)$ for decreasing values of ε.

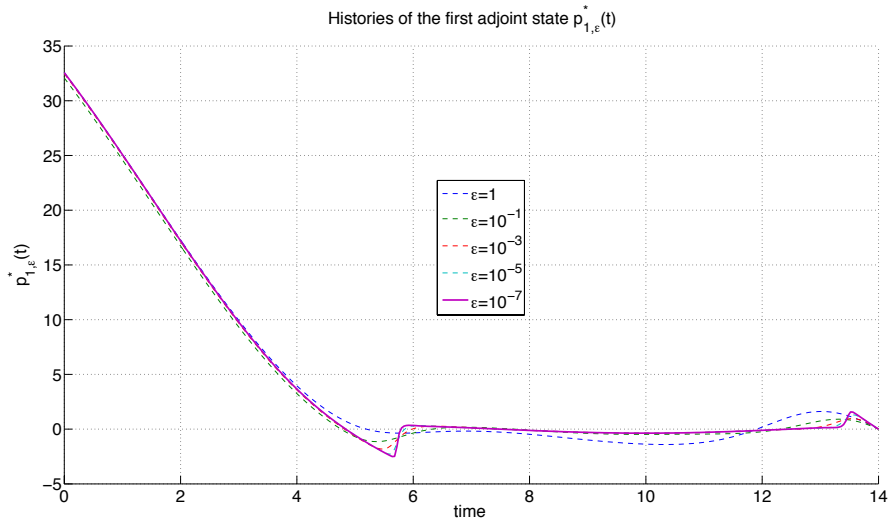

Figure 4.19: Histories of the first adjoint state $p^*_{1,\varepsilon}(t)$ for decreasing values of ε.

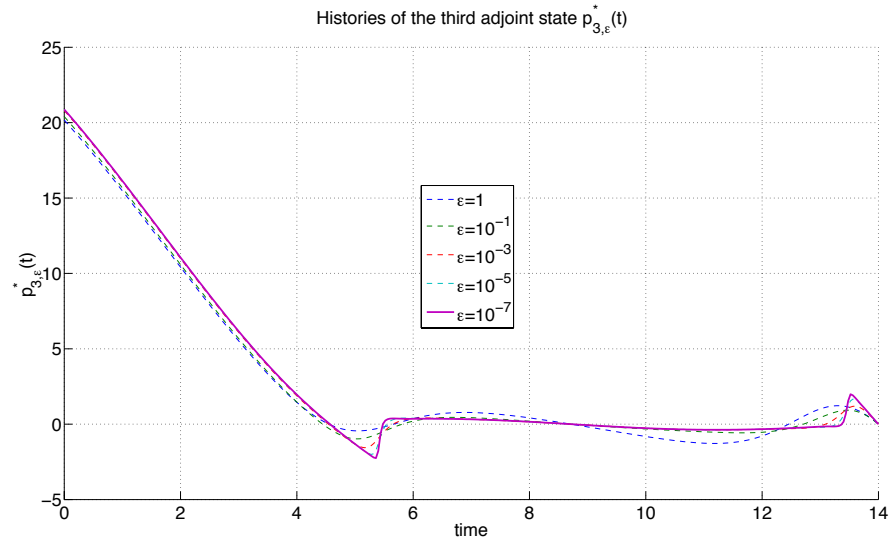

Figure 4.20: Histories of the third adjoint state $p^*_{3,\varepsilon}(t)$ for decreasing values of ε.

Part II

Applications énergétiques
Applications to energy systems

Chapter 5

Investigating the ability of various buildings in handling load shiftings

Contents

5.1 Introduction

As discussed in Chapter 2, significant recent efforts have been targeted at reducing electricity peak-demand. In Europe, these peaks mostly occur in winter time, and are, for the main part, due to heating systems. To guarantee the electric grid stability, numerous studies have focused on the overall load reduction. At the level of individual houses, this reduction can be achieved thanks to a careful architectural design aiming at efficiently capturing and, later, restoring solar gains [BHM77]. Advanced heating control strategies can also be a solution. Such control strategies must account for the occurrence of discount periods of power tariff [KM04] and use the building thermal mass as an asset to shift the building consumption. A

beneficial effect is the created reduction of the peak consumption [Bra90, Che01, HM02, XHBH04].

This chapter follows such an approach and studies the impact of load shifting on five thermal models ranging from poorly to well insulated houses. The method of analysis consists in solving COCPs to accurately compute optimal trajectories following the approach presented in this thesis. Gradually, considering the duration of the load shift as a parameter, one determines the maximum allowable duration of a complete heating load shifting while maintaining an acceptable level of comfort. The results obtained in this study show that the thermal mass of a poorly insulated building is not sufficient to perform load shiftings superior to twenty minutes. Thus, the use of the house inertia as energy storage capacity is shown to be relevant only in the case of sufficiently insulated buildings (which can actually handle load shiftings of several hours). Practical cases of interest are presented.

In §5.2, a description of the considered building is given, together with a description of the discretization scheme yielding a high-order linear model of the system. In §5.3, this model is reduced and constraints are formulated on its input and outputs. In §5.4, the algorithm serving to solve the obtained constrained dynamical optimization is presented. In §5.5, the results on the abilities of the different considered systems are presented together with the maximum bearable duration of daily load shiftings for each model. Finally in §5.6 the conclusion and the perspectives of the study are presented.

5.2 Model of the building

5.2.1 Building description

The building under study is a single-family house. It corresponds to an actual experimental passive house being part of the INCAS platform built in Le Bourget du Lac, France (see Figure 5.1). For our study, five low performance versions of the building are considered. The reference version corresponds to a house built prior to the introduction of the first French thermal regulation (1975). This reference version used to represent 58% of the French stock in 2008. The four other versions correspond to various renovation levels on this reference. In this chapter, they serve to study the beneficial effects of renovation efforts on the peak load management. The house has two floors for a total living area of 89 m^2. 34% of its South facade surface is glazed while the North facade has only two small windows. All the windows are single-glazed. The South facade is also equipped with solar protections for the summer period. The external walls are made of a 30 cm-thick layer of concrete blocks and the floor consists in 20 cm of reinforced concrete. There is no insulation in the building except for the 10 cm of glass-wool in the attic. According to thermal simulation results using the Pléiades+COMFIE software [PS90], the heating load is 253 kWh/(m^2.year) which is typical for such type of house in this area. Comparisons have been performed during the design phase on the passive house version of this building with other simulation tools like Energy Plus and

TRNSYS [BSW09] and have shown similar results.

Figure 5.1: Computer graphics view of the house (west and south facades).

Four different renovations of this building are presented in Table 5.1

Version	Renovation applied	Heating consumption (kWh/m^2/year)
Reference (1st)	none	253
Roof insulation (2nd)	(1) + 30 cm of glass-wool in the attic	246
Triple glazing (3rd)	(2) + Triple glazed windows	215
Insulation of external walls (4th)	(3) + 15 cm of glass-wool in external walls	93
Heat recovery ventilation (HRV) (5th)	(4) + HRV with an efficiency of 0.5 (accounting for air infiltration)	80

Table 5.1: Versions of the considered building throughout renovations.

5.2.2 Thermal model

The building is modeled with a set of spatial zones of homogeneous temperature. For each zone, each wall is divided in fine meshes small enough to consider homogeneous temperature in each mesh point. Two additional mesh points are considered for the air and furniture in the zone, respectively. Eventually, a thermal balance is performed on each mesh within the building. It takes into account:

- P_{cond}: the losses (or gains) by conduction in walls, floor and ceiling
- P_{sol}: the gains due to solar irradiance through the windows
- P_{conv}: the losses (or gains) due to convection at walls surface

- P_{in}: the internal gains due to heating, occupancy and other loads (only for zone air mesh)
- P_{bridges}: heat losses through thermal bridges, not associated to thermal mass
- P_{ventil} : heat losses due to air exchange.

When applied to the air of each zone, the thermal balance equation reads:

$$C_{\text{air}}\dot{T}_{\text{air}} = P_{\text{in}} + P_{\text{cond}} + P_{\text{bridges}} + P_{\text{ventil}} + P_{\text{sol}} + P_{\text{conv}} \tag{5.1}$$

with C_{air} the thermal capacity of the air node (including furniture) and T_{air} the temperature of the mesh. For each zone, repeating equation (5.1) for each mesh point and including an output equation leads to the following continuous linear time-invariant system

$$\begin{cases} C\dot{T}(t) &= AT(t) + EU(t) \\ Y(t) &= JT(t) + GU(t) \end{cases} \tag{5.2}$$

with:

- T mesh temperatures vector
- U driving forces vector (climate parameters, heating, etc.)
- Y output vector (here, temperature of the air nodes)
- C thermal capacity (diagonal) matrix
- A, E, J, G matrices relating the vectors of the dynamics.

For representative simulations, it is important to account for the occupancy of the building, which partly defines P_{in} through the emission of heat by the inhabitants and the appliances. The second part of heat emission in P_{in} is due to the heating system. Another important factor is the weather model. It defines the losses due to heat transfer with the ambient temperature and the gains with solar irradiance. All the data of the house occupancy and weather models are included in the input vector U.

5.3 Model reduction and definition of constraints

5.3.1 Model reduction

The high-order linear model (5.2) is now reduced. In view of solving COCPs over relatively long time horizons, the state dimension (order 33) is too large and discards hope of a fast convergence of the optimization algorithm. Therefore, a reduction method is applied to lower the state dimension. For this task, several methods can be used, among which are singular perturbations [Kha02], and identification methods [MCPF10]. In our case, an efficient method is the balanced truncation

[ZDG96]. Indeed, this truncation consists in removing the state variables which receive the least effort from the input and contribute the least to the variations of the output[1]. Precisely, let us call $\Sigma_h(s)$ (resp. $\Sigma_r(s)$) the Laplace-transform of the high (resp. reduced) order system. The order of reduction is chosen as the minimal order such that

$$\| \Sigma_h(s) - \Sigma_r(s) \|_{\mathcal{H}_\infty} \leq -70 \text{ dB}$$

where the $\mathcal{H}_\infty$ norm is defined in [ZDG96]. In this case, all models are at least third order models, and one of them is a fourth order one.
In Table 5.2, the various time constants of the considered models are reported. One shall notice that these thermal building models clearly have (at least) three well separated time scales [Kha02]. Interestingly, it shall be noticed that the main effects of the renovation is to enlarge the slower time constant.

Building version	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Time constants	8 min	7 min	8 min	9 min	9 min
	13 h	13 h	2 h	13 h	18 h
	95 h	98 h	8 h	160 h	180 h
			91 h		

Table 5.2: Value of the time constants of the five different models.

5.3.2 Model and constraints

5.3.2.1 Model notations

In the following, we use the classical linear state space representation to represent the model:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BP(t) + d(t) \\ T(t) &= Cx(t) \end{aligned}$$

where x is the state of the model, T is the inside temperature, d represents the influence of the outside temperature and the solar fluxes on the heating of the house, and P represents the heating flux on the air node and is the control variable.

5.3.2.2 Constraints

Inside temperature constraints The temperature constraints are 24 hours periodic and are:

- $T \leq 24^{\circ}C$ at all times
- $T \geq 14^{\circ}C$ between 9 a.m. and 5 p.m.

[1] We refer the interested reader to [ZDG96], for mathematical definitions of the considered approximation

- $T \geq 20^{\circ}C$ otherwise.

To simplify the notations, we write these temperature constraints as follows

$$T^{-}(t) \leq T(t) \leq T^{+}(t) \tag{5.3}$$

where $\dot{T}^{-}(t) = \dot{T}^{+}(t) = 0$ almost everywhere.

Control constraints The control constraints are not the same for all systems:

- $0 \leq P \leq 20$ kW for the buildings whose walls have not been insulated
- $0 \leq P \leq 10$ kW for the buildings whose walls have been insulated.

Again, to simplify the notations in the algorithm, we write the control constraints as follows

$$0 \leq P(t) \leq P^{+}(t) \tag{5.4}$$

where $\dot{P}^{+}(t) = 0$ almost everywhere.

Load shifting In the considered scenarios, the load shiftings consist in a daily time period when the heating of the house is not allowed to consume any energy. These shiftings start everyday at 5 p.m.. The objective of this study is to determine the maximum duration of these load shiftings beyond which it becomes impossible to satisfy both (5.3) and (5.4).

5.4 Problem statement and solution method

5.4.1 Method

To characterize the duration of load shifting which allows the inside temperature to satisfy (5.3) while the heating power satisfies (5.4), we solve the corresponding state and input COCP. When no solution can be found, it is deduced that the load shiftings are too long. This property is independent of the temperature control system, and solely stems from the ability of the building to store energy.
To determine the maximum allowable duration of the load shifting, we gradually increase the load shifting period durations until no solution satisfying the constraints (5.3) and (5.4) exists.

5.4.2 Algorithm

To solve the COCP, we use the interior-point methodology proposed in this thesis and summarized in §3.5.1. In this example it is desired to minimize the energy consumption. The criterion is given by the following (to minimize energy consumption):

$$J = \min_{P(t) \in [0, P^{+}(t)]} \int_{0}^{T} P(t) dt$$

with the dynamics and the state constraint $T(t) \in [T^-(t), T^+(t)]$ seen above, and where $T = 7$ days. The change of variables permitting to remove the input constraint is the following

$$P \triangleq \phi(\nu) = P^+ \left(\frac{e^{k\nu}}{1 + e^{k\nu}} \right), \; k > 0 \tag{5.5}$$

The Hamiltonian of (3.20) is then

$$\begin{aligned} H_\varepsilon(x, p, \nu) \;\triangleq\; & \phi(\nu) + p^t \Big(Ax + B\phi(\nu) + d \Big) \\ & + \varepsilon \Big(\gamma_g(Cx - T^-) + \gamma_g(T^+ - Cx) + \gamma_u \circ \phi(\nu) \Big) \end{aligned}$$

In this example, the adjoint vector p satisfies the following differential equation

$$\frac{dp}{dt}(t) = -A^t p(t) - \varepsilon C^t \big(\gamma_g'(Cx(t) - T^-) - \gamma_g'(T^+ - Cx(t)) \big)$$

where γ_g' is the derivative of the following function

$$\gamma_g(x) = \begin{cases} x^{-1.1} & \forall x > 0 \\ 0 & \text{otherwise} \end{cases}$$

This function is chosen accordingly to Theorem 1. The algorithm described in §3.5.1 used in this example is the following:

Step 1: Initialize the functions $x(t)$ and $p(t)$ such that the initial unknown $Cx(t) \in (T^-, T^+)$ for all $t \in [0, T]$, and set $\varepsilon = \varepsilon_0$. Simply, p can be chosen identically equal to zero at first step.

Step 2: Compute $\nu_\varepsilon^* = \sinh^{-1}\left(-\frac{1 + p^t B}{\varepsilon} \right)$[2]. Thus, the optimal solution $P_\varepsilon^* = \phi(\nu_\varepsilon^*)$ is given using equation (5.5) with $k = 1$.

Step 3: Solve the two-point boundary value problem

$$\frac{dx(t)}{dt} = Ax(t) + BP_\varepsilon^*(t) + d(t)$$

and

$$\frac{dp(t)}{dt} = -A^t p(t) - \varepsilon C^t \left[\gamma_g' \left(Cx(t) - T^- \right) - \gamma_g' \left(T^+ - Cx(t) \right) \right]$$

with the following boundary constraints $x(0) = x_0$ and $p(T) = 0$.

Step 4: Decrease ε, initialize $x(t)$ and $p(t)$ with the solutions found at Step 3 and start over at Step 2.

In our case, the sequence (ε_n) has been chosen such that $\varepsilon_n = 10^{-\frac{n}{10}}$ with $n = 0 \ldots 40$.

[2] This is the analytical solution of $\frac{\partial H_\varepsilon}{\partial \nu} = 0$, where we have set $\gamma_u' \circ \phi(\nu) \triangleq \sinh(\nu)$

5.5 Simulations and results

5.5.1 Simulations

The considered optimization takes place in winter over one particularly cold week. The ambient temperature history is reported on Figure 5.2. For each version of the building, indoor temperatures (see Figure 5.3) and energy consumptions over the week have been computed, first without load shiftings and then, with maximal bearable load shiftings (see Figure 5.4).

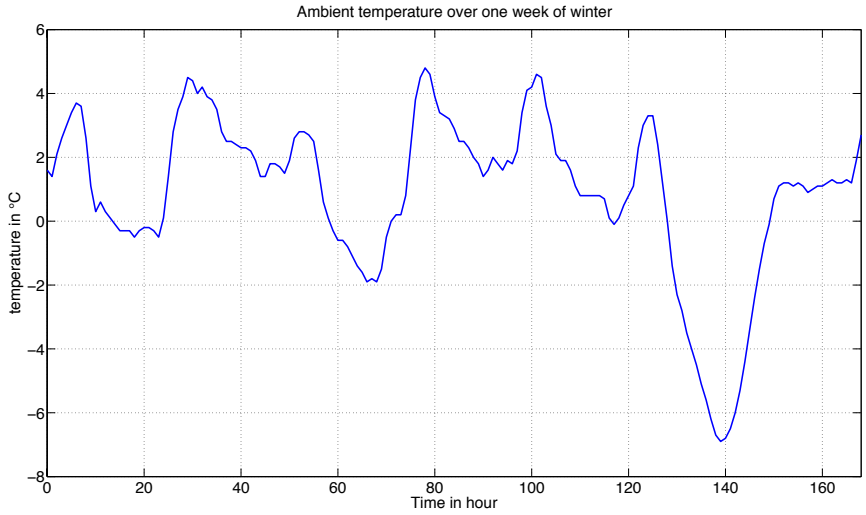

Figure 5.2: Ambient temperature over one week of winter.

5.5.2 Summary of the results

In terms of energy consumption, the first and second versions of the building are quite similar (Fig. 5.4). The adjunction of triple glazed windows (3^{rd} version) induces a significant decrease of energy consumption ($\approx 30\%$). The insulation of the external walls (4^{th} version) clearly induces a further reduction of the energy consumption ($\approx 50\%$). The most effective renovation strategy (in terms of energy consumption) seems to be the increasing of insulation.

Now, we consider the ability in handling load shiftings. Table 5.3 and Figure 5.5 illustrate that the three first versions of the building cannot handle load shifting durations superior to 20 minutes. Interestingly, the adjunction of triple glazed windows (3^{rd} version) does not improve the load shifting ability whereas it is efficient for energy savings. Actually, handling large load shifting periods becomes possible solely when the insulation is sufficiently increased.

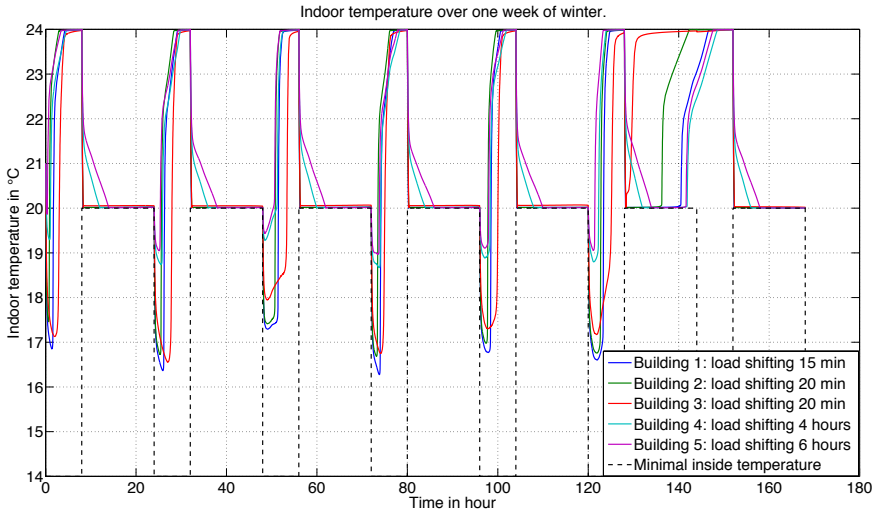

Figure 5.3: Comparison between the optimal indoor temperature for each building (with the maximum bearable load shiftings duration in each case). The behavior of the indoor temperature is different on the last day from the other days because the ambient temperature is particularly cold.

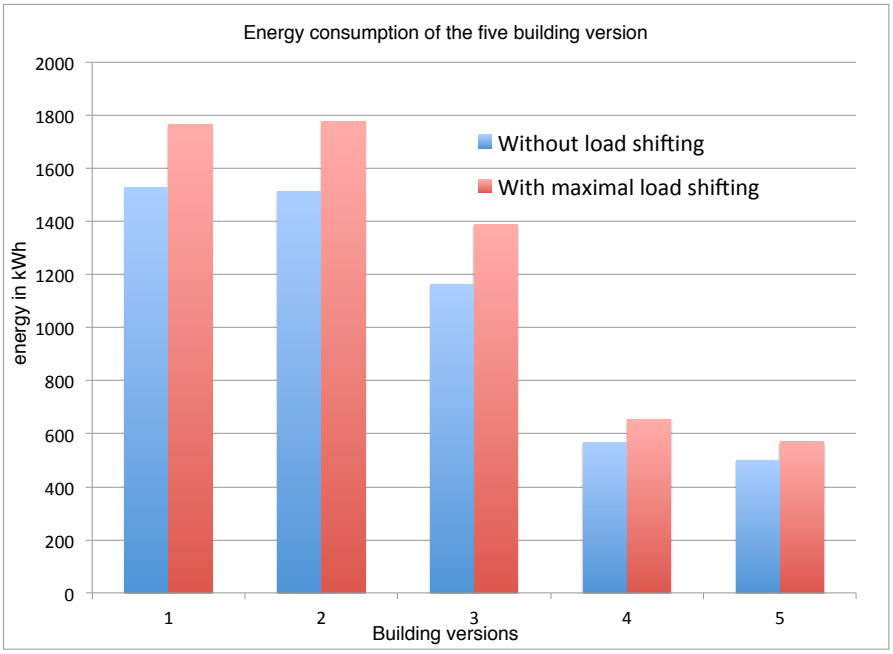

Figure 5.4: Energy consumption over one week for the five versions of the building. For each building the consumed energy is displayed without load shifting and with the maximal bearable one.

Building version	1st	2nd	3rd	4th	5th
Load shifting duration	15 min	20 min	20 min	4 h	6 h

Table 5.3: Value of the maximum load shifting duration for each version of the building.

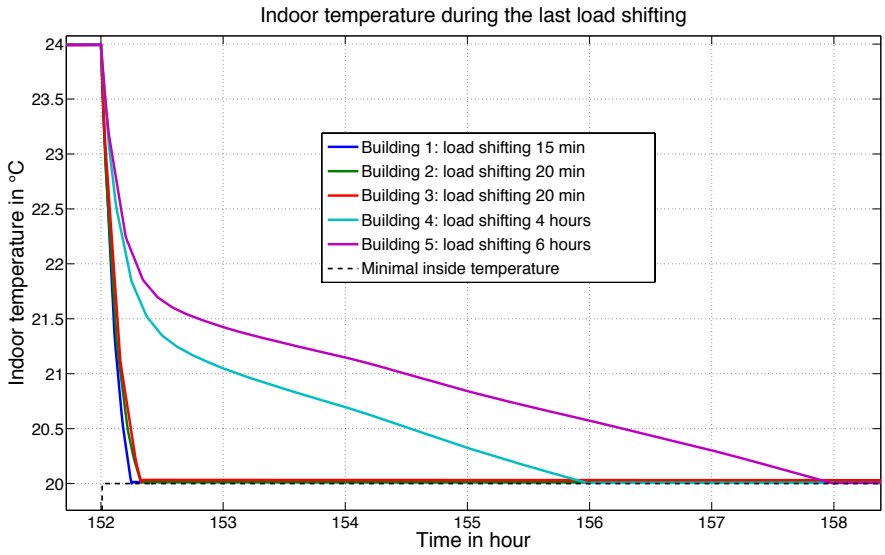

Figure 5.5: Comparison of the optimal indoor temperature during the load shifting of the last day of the week.

5.6 Conclusion

On the methodological side, it appears that solving the discussed COCPs is an effective tool to study properties of the buildings. The existence of feasible trajectories only depends on the characteristics of the buildings. The presented method yields quantitative results even when considering fast time scales phenomenon.
On the applicative side, we have emphasized that a non-insulated residential house cannot handle load shifting durations superior to 20 minutes even if an advanced strategy of regulation is used. To allow these buildings to handle long load shiftings, their thermal mass is not sufficient, the buildings must be insulated enough or have auxiliary energy storage capacity.

Chapter 6

Day to night load shifting for low-consumption buildings

Contents

6.1 Introduction

In Western Europe, peaks in the overall consumption of electricity mostly occur in winter time. At the beginning of the 2000s, a decrease in the smoothed national ambient temperature of 1°C used to induce an increase of the peak consumption of 1000 MW at the national scale of France. Nowadays this effect, called *thermal sensitivity* is estimated around 2300 MW/°C [RTE11]. Simultaneously, due to global Earth warming, more and more attention is being payed on global energy consumption and CO_2 emissions. This has resulted in the emergence of ever more restrictive laws on levels of insulation and consumption of primary energy of new buildings (prime example being RT 2005 see [RT206] and RT 2012 [RT211]). Installing high efficiency electric heaters in such houses allows home owners to have a

reliable and low CO_2 emissions heating device. A way to reduce thermal sensitivity of the electrical peak consumption, while maintaining low CO_2 emissions for heating systems in well insulated buildings, is to use their thermal mass as an asset to shift all or part of the energy consumption [XHBH04, Bra90, Che01, HM02] from day to night time during the whole heating season. This is true for two reasons. First, the peak is naturally smoothed-out by the considered individual load shifting. Second electric heaters are virtually CO_2 emission free, when used at night, because their power comes from nuclear plants.

This chapter uses the methodology of this thesis to evaluate the feasibility of complete load shiftings from day to night time during the whole heating period. It is applied to two well insulated buildings corresponding to two different construction methods (position of the insulation in the walls). Gradually, considering the amount of energy shifted from day to night as a parameter, one can determine the ability of the house to perform load shiftings while maintaining an acceptable level of comfort. The first conclusion of the conducted study is that it is possible to use the thermal mass of well insulated buildings to heat the ambience during night time only (from 10 p.m. to 6 a.m.) while maintaining the comfort. Hence, it is possible to use electric convector heaters in well insulated buildings without increasing the thermal sensitivity of peak consumption.

The second conclusion of this study is that the construction method consisting in putting the concrete core of the house between the insulation and the interior of the building (exterior insulation) is more efficient in performing complete load shiftings (both in term of comfort and energy consumption) than the classic interior insulation technique.

The chapter is organized as follows. In §6.2, a description of the considered building is given, together with the method used to obtain an accurate reduced order linear model. In §6.3, the scenario of optimization is given, i.e. the weather, the occupancy period and the constraints. In §6.4, the various optimization scenarios are presented. In §6.5, the algorithm used to solve the problem is detailed. In §6.6, the quantitative results are presented. Finally §6.7 contains the conclusions along with perspectives of the study.

6.2 Model

6.2.1 Building description

The buildings under study in this chapter are low-consumption single-family detached houses. Two types of building (I and E) are considered. They are built using the same materials but the first one (I) is insulated from the interior whereas the second (E) is insulated from the exterior as described in Table 6.1.

The two houses have the same geometry:

- Total floor area 100.86 m^2 ; Area of the roof: 100.86 m^2
- Area of the southern wall: 25.75 m^2 ; Area of the southern window: 5 m^2

- Area of the western wall: 18.5 m^2 ; Area of the western window: 2 m^2
- Area of the northern wall: 26.75 m^2 ; Area of the northern window: 4 m^2
- Area of the eastern wall: 16.5 m^2 ; Area of the eastern window: 4 m^2

Layers	Building I	Building E
External layer	20 cm of concrete	15 cm of *insulation*
Intermediate layer	15 cm of *insulation*	20 cm of concrete
Interior layer	1 cm of plater	1 cm of plater

Table 6.1: Constitution of the external walls for the two buildings.

6.2.2 Building model

In this study, we consider the temperature of the air node within the buildings to be homogeneous. The building is modeled using the software Dymola [Elm95], resulting in the following high-order linear system (order 42)

$$\begin{aligned} \dot{X}(t) &= AX(t) + B_T T_{\text{amb}}(t) + B_s \Phi_s(t) + B_w \Phi_w(t) + B_n \Phi_n(t) \cdots \\ &\quad + B_e \Phi_e(t) + B_i \Phi_i(t) + B_h P(t) \\ T(t) &= CX(t) \end{aligned}$$

with

- T_{amb}: the ambient temperature
- Φ_s :the solar flux on the southern wall
- Φ_w: the solar flux on the western wall
- Φ_n: the solar flux on the northern wall
- Φ_e: the solar flux on the eastern wall
- Φ_i: the internal gains (occupancy, lights...) on the air node
- P: the heating power on the air node
- T: the indoor temperature.

6.2.3 Model reduction

The high-order model is now reduced with the same method as in §5.3.1. In this case, the order of reduction for the two systems is 5. In Table 6.2, the various time constants of the considered models are reported. One shall notice that these thermal building models clearly have three well-separated time scales [Kha02]. Interestingly, it shall be noticed that the building E has a slowest time constant much bigger than

the other. This phenomenon is due to the fact that the thermal mass of the part of the wall between the air node and the insulation is greater for the building E than for building I.

	Building I	Building E
Time constants	2 min 8 s	1 min 42 s
	27 min 21 s	16 min 23 s
	2 H 36 min	48 min
	10 H 30 min	9 H 30 min
	212 H	357 H

Table 6.2: Value of the time constants of the two reduced models.

In the following, we use the classical linear state space representation for the reduced models

$$\dot{x}(t) = Ax(t) + BP(t) + d(t) \quad (6.1)$$
$$T(t) = Cx(t) \quad (6.2)$$

where x is the state of the reduced model (dimension 5 vector), T is the output, d lumps the influence of the outside temperature, the solar fluxes and the internal gains on the heating of the house, and P is the control variable.

6.3 Scenario of optimization

6.3.1 Weather data and occupancy period

The employed weather data are actual measurements of external temperature, direct and indirect solar fluxes of the year 1991 in the city of Trappes in France. For this study, one is only interested in the heating period which starts on the 1st of November and ends at the end of March. The reason for this is that, with well insulated buildings, it is usually considered that no heating is needed after the 10th of March. The time horizon of optimization is then of 137 days.

This scenario of optimization includes a period of vacancies, between the 25th of December and the 1st of January, where the inhabitants leave the house.

6.3.2 Constraints

6.3.2.1 Indoor temperature constraints

The temperature constraints are the following

- $T \leq 26$ °C between 6 a.m. and 10 p.m.
- $T \leq 23$ °C otherwise
- $T \geq 12$ °C during the one week holiday period

- $T \geq 19$ °C otherwise.

To simplify the notations, we write these temperature constraints as follows:

$$T^-(t) \leq T(t) \leq T^+(t) \tag{6.3}$$

with $\dot{T}^-(t) = \dot{T}^+(t) = 0$ almost everywhere.

6.3.2.2 Control constraint

The control constraint is the following

$$0 \leq P(t) \leq 3 \text{ kW}$$

To simplify the notations in the algorithm, we write the control constraint as follows

$$0 \leq P(t) \leq P^+ \tag{6.4}$$

6.4 Methodology

The objective is to evaluate the efficiency of these low-consumption buildings to shift a certain amount of energy from day to night. We proceed by solving the following state and input COCP

$$\min_P \int_0^{T_f} f(t)P(t)dt$$

where the weight factor $f(t)$ is

$$\begin{array}{lcll} f(t) & = & F_{\text{day}} & \text{between 6 a.m. and 10 p.m.} \\ f(t) & = & F_{\text{night}} & \text{otherwise} \end{array}$$

This formulation aims at minimizing the cost of electricity for the dynamics (6.1)-(6.2) under the constraints (6.3)-(6.4). Naturally, it is expected that the amount of shifted energy will be related to the ratio $F_{\text{day}}/F_{\text{night}}$. For both buildings a collection, indexed by the ratio $F_{\text{day}}/F_{\text{night}}$, of COCPs is solved. The higher this ratio, the more energy will be shifted from the day to the night period.

This problem is solved for both buildings with the following values of the ratio $F_{\text{day}}/F_{\text{night}}$

$$F_{\text{day}}/F_{\text{night}} \in \{1 \ ; \ 1.2 \ ; \ 1.5 \ ; \ 3 \ ; \ 5 \ ; \ 10\}$$

6.5 Algorithm

To solve the collection of COCPs, we use the interior-point algorithm described in Chapter 3. Each COCP in the collection is addressed using a sequence of penalized

unconstrained OCPs. In this example, the change of variables yielding an unconstrained formulation is the same as in equation (5.5) $P = P^+ \frac{e^{k\nu}}{1+e^{k\nu}}$ and therefore the cost to minimize is

$$\min_{\nu} \int_0^{T_f} f(t)\phi(\nu(t))dt$$

with the dynamics and the state constraint $T(t) \in [T^-(t), T^+(t)]$ seen above. In this example, the adjoint vector p satisfies the following differential equation

$$\frac{dp}{dt}(t) = -A^t p(t) - \varepsilon C^t \left[\gamma_g'(Cx(t) - T^-(t)) - \gamma_g'(T^+(t) - Cx(t))\right]$$

where γ_g' is explicitly defined in equation (5.4.2). According to the methodology exposed in §3.5.1, the solving algorithm is the following:

Step 1: Initialize the functions $x(t)$ and $p(t)$ such that the initial unknown $Cx(t) \in (T^-(t), T^+(t))$ for all $t \in [0, T_f]$, and set $\varepsilon = \varepsilon_0$. Simply, p can be chosen identically equal to zero at first step.

Step 2: Compute $\nu_\varepsilon^*(t) = \sinh^{-1}\left(-\frac{f(t)+p^t(t)B}{\varepsilon}\right)$[1]. Thus, the optimal solution $P_\varepsilon^*(t) = \phi(\nu_\varepsilon^*(t))$ is given using equation (5.5) with $k = 1$.

Step 3: Solve the two point boundary value problem

$$\begin{cases} \dfrac{dx(t)}{dt} = Ax(t) + BP_\varepsilon^*(t) + d(t) \\ P_\varepsilon^*(t) = \phi\left(\operatorname{asinh}\left(-\dfrac{f(t) + p^t(t)B}{\varepsilon}\right)\right) \\ \dfrac{dp(t)}{dt} = -A^t p(t) - \varepsilon C^t \left[\gamma_g'\left(Cx(t) - T^-\right) - \gamma_g'\left(T^+ - Cx(t)\right)\right] \end{cases}$$

with the following boundary constraints $x(0) = x_0$ and $p(T_f) = 0$.

Step 4: Decrease ε, initialize $x(t)$ and $p(t)$ with the solutions found at Step 3 and restart at Step 2.

In our case, the sequence (ε_n) has been chosen such that $\varepsilon_n = 10^{-\frac{n}{10}}$ with $n = 0 \cdots 40$.

6.6 Results

6.6.1 Influence of the $F_{\text{day}}/F_{\text{night}}$ ratio on indoor temperature

We now discuss the obtained numerical results. As expected, as the ratio $F_{\text{day}}/F_{\text{night}}$ increases, more energy is shifted from the day period to the night time. To satisfy the indoor temperature constraints, substantial overheatings of the house during

[1] Defining the Hamiltonian H as follows $H(x, \nu, p) = f\phi(\nu) + p^t(Ax + B\phi(\nu) + d) + \varepsilon\left(\gamma_g \circ g(x) + \gamma_u \circ \phi(\nu)\right)$, ν_ε^* is the solution of $\frac{\partial H}{\partial \nu} = 0$ where we chose $\gamma_u' \circ \phi(\nu) \triangleq \sinh(\nu)$.

night time are, unfortunately, necessary. On Figures 6.1 and 6.2, the averaged temperature over one day is displayed for the two buildings under consideration.

First, one can see that the averaged temperature over the heating season at 6 a.m. is all the higher as the ratio $F_{\text{day}}/F_{\text{night}}$ increases. This phenomenon is natural because, to minimize the thermal loss during the day period, the indoor peak of temperature must be achieved at 6 a.m., i.e. at the beginning of the day period.

Moreover, one can see that the mean overheating of the building E is lower than the one of the building I. The averaged temperature over the whole heating season (leaving out holidays) are given in Table 6.3 for the two buildings and for each value of the ratio.

	ratio=1	ratio=1.2	ratio=1.5	ratio=3	ratio=5	ratio=10
Building I	19.31	19.35	19.60	19.93	20.00	20.07
Building E	19.21	19.32	19.46	19.56	19.60	19.62

Table 6.3: Mean temperature over the whole heating season except holidays for the two considered building for each value of the ratio $F_{\text{day}}/F_{\text{night}}$.

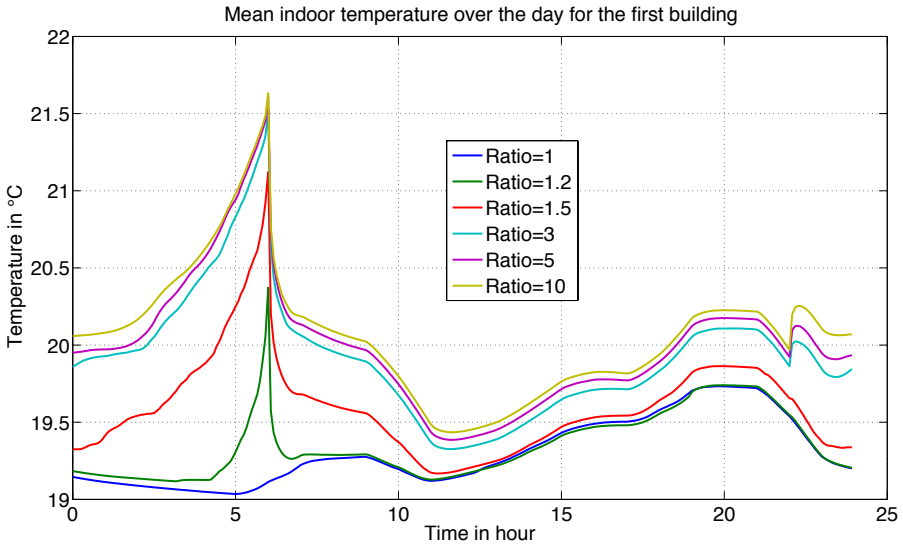

Figure 6.1: For the building I, the average temperature over one day is displayed for various values of the ratio $F_{\text{day}}/F_{\text{night}}$. Temperature at 6 a.m. increases with the ratio.

On Figures 6.3 and 6.4, the time histogram of the indoor temperature is given for both buildings during the night period, i.e. between 10 p.m. and 6 a.m. When minimizing the energy consumption (ratio=1), one can see that the indoor temperature of the building I does not exceed 21.5°C while it does not exceed 20.5°C for

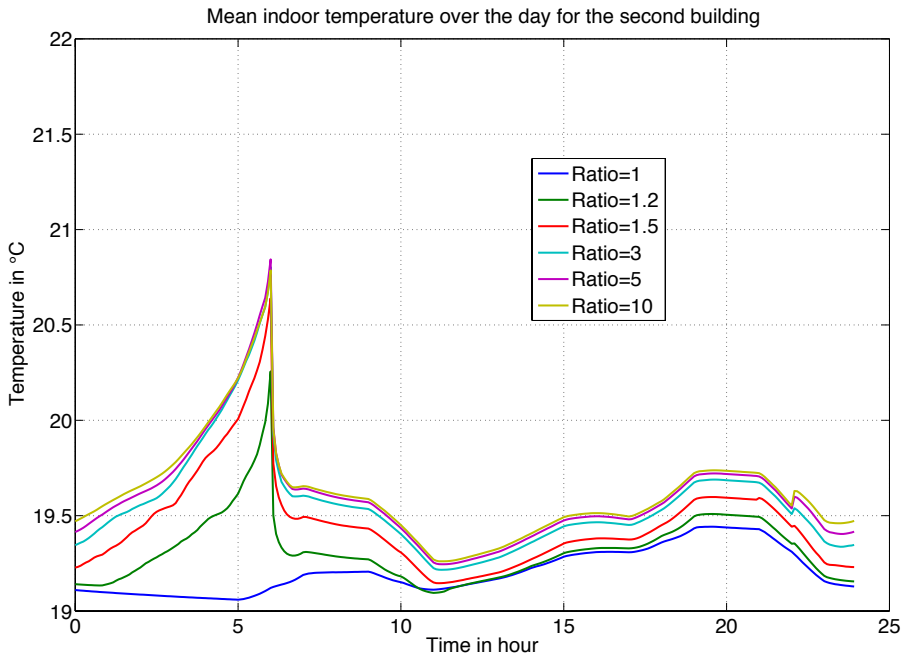

Figure 6.2: For the building E, the average temperature over one day is displayed for various values of the ratio $F_{\text{day}}/F_{\text{night}}$. Temperature at 6 a.m. increases with the ratio.

building E. On Figure 6.3 (building I) one can see that for values of the ratio superior to 1.5 the cumulated time spent with a night overheating superior to 22.5°C is large whereas the cumulated time spent with a night overheating between 20.5°C and 22.5°C is not. This phenomenon does not happen for building E as shown on Figure 6.4. This is probably a striking advantage of building E compared to building I. This confirms that, in order to shift energy from the day period to night period, the building I generates a larger overheating than the building E.

6.6.2 Influence of the $F_{\text{day}}/F_{\text{night}}$ ratio on the heating power

On Figures 6.5 and 6.6, the averaged power is displayed for the two buildings under consideration. First, when minimizing energy consumption (ratio=1) one can see that 47% of the total amount of consumed energy is consumed during the day period in both cases. Accordingly with the discussion in §6.6.1, one can see that, as the ratio $F_{\text{day}}/F_{\text{night}}$ increases, the average power at the end of the night period (between 4 a.m. and 6 a.m.) increases. Moreover, one can see that for a ratio superior or equal to 3, the mean power during the day period is almost equal to zero, i.e. the total load shifting from the day period is almost complete. In addition, as observed in §6.6.1, the average temperature at the end of the night

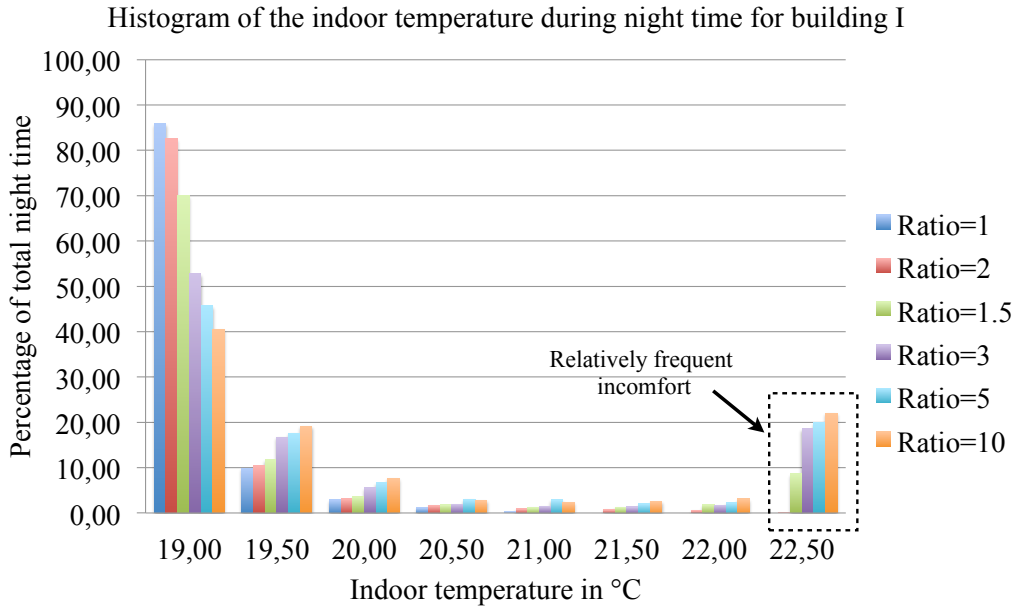

Figure 6.3: For the building I the histogram of the indoor temperature during night time is given for various values of the ratio $F_{\text{day}}/F_{\text{night}}$. Building I is relatively uncomfortable when the load-shifting strategy is employed.

period is higher for the building I than for the building E. Concerning the mean power, the opposite phenomenon appears : the mean power for the building E is higher than for the building I. This phenomenon is due to the upper temperature constraint. Indeed, Figure 6.3 shows that during the night period, the indoor temperature of the building I is close to 23°C during a significant cumulated time. Thus, an increase of the heating power would induce an overheating superior to 23°C which is forbidden. But on Figure 6.4, whatever the ratio, the indoor temperature never gets close to 23°C which allows a larger use of heating power without inducing a forbidden overheating.

6.6.3 Influence of the $F_{\text{day}}/F_{\text{night}}$ ratio on the energy consumption

On Figure 6.7 for the two buildings and each value of the price ratio, the total amount of consumed energy versus the amount of energy consumed during the day period is reported. First, minimizing energy (ratio=1) shows that the building I consumes less energy than the building E (727 kWh vs 758 kWh). This difference is mostly due to the management of the holiday period and a detrimental effect of thermal inertia of building E. Indeed as described in §6.3.2, the minimal temperature constraints drops to 12°C during the holiday period. Thus, to reach 19°C at the end of the holiday period, the restarting of the heating has to be anticipated.

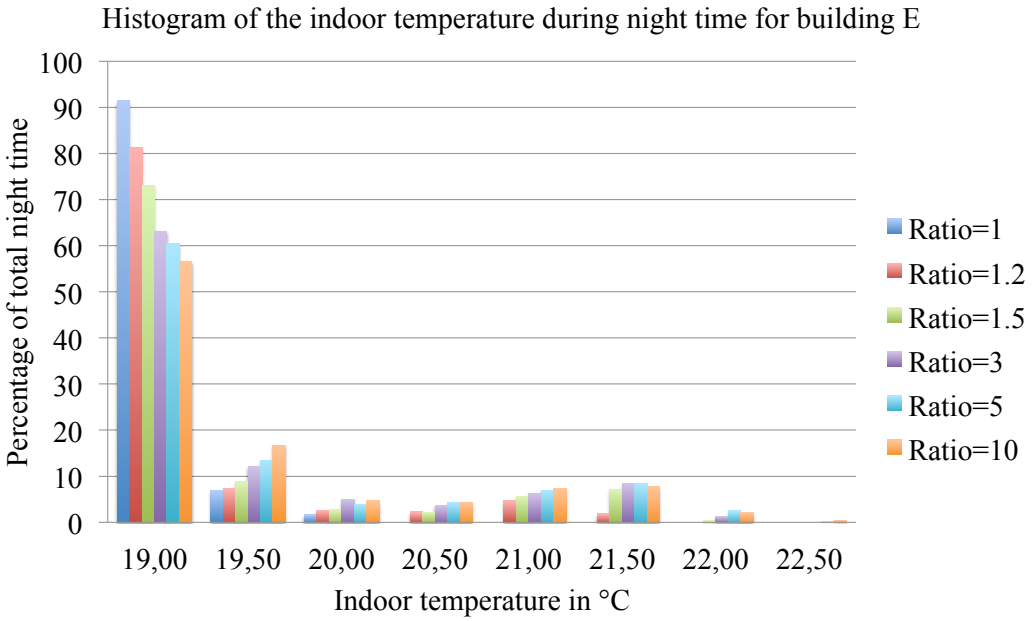

Figure 6.4: For the building E the histogram of the indoor temperature during night time is given for various values of the ratio $F_{\text{day}}/F_{\text{night}}$. Building E is relatively comfortable when the load-shifting strategy is employed.

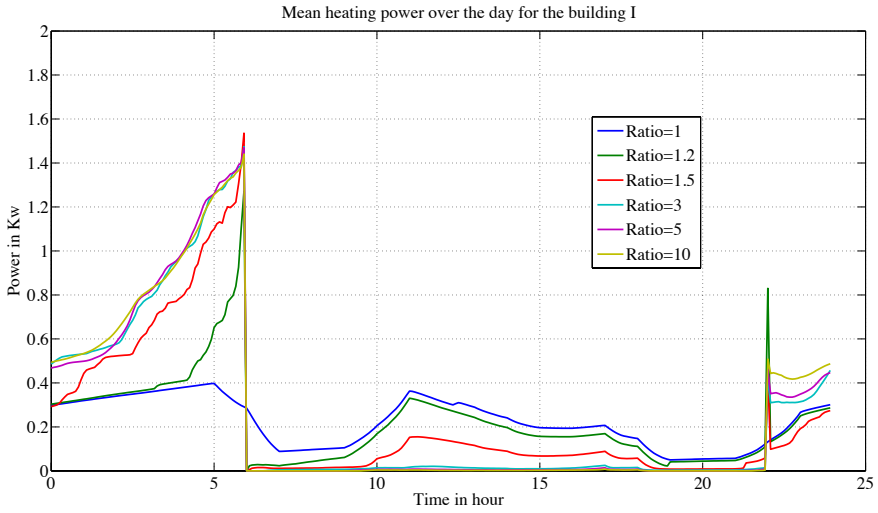

Figure 6.5: Building I. Average heating power over one day for various values of the ratio $F_{\text{day}}/F_{\text{night}}$, using the optimal load shifting from day to night.

Since the building E has a thermal mass higher than the building I, its heating restarting occurs 22 hours before the one of building I, which causes the difference

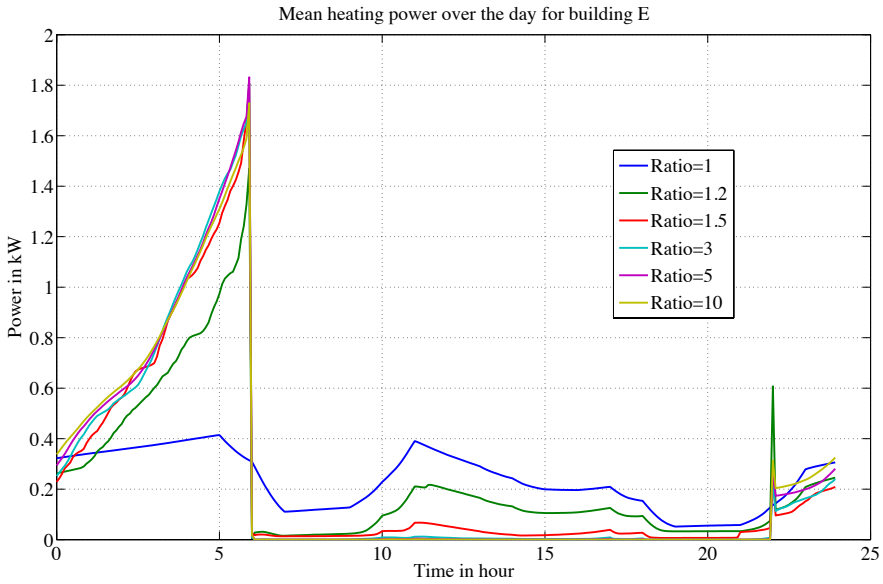

Figure 6.6: Building E. Average heating power over one day is displayed for various values of the ratio $F_{\text{day}}/F_{\text{night}}$, using the optimal load shifting from day to night.

of energy consumption. But, this difference in energy consumption does not exceed 6 kWh between the two buildings when withdrawing the holiday period. Also, as expected, the increase of the ratio $F_{\text{day}}/F_{\text{night}}$ induces an increase of the global energy consumption compared to the reference case where the ratio is equal to 1 because the mean indoor temperature is higher than in the reference case. The percentages of overconsumption are given in Table 6.4 for the two buildings and each value of the ratio.

	ratio=1	ratio=1.2	ratio=1.5	ratio=3	ratio=5	ratio=10
Building I	0%	1.64%	8.27%	19.39%	24.29%	27.24%
Building E	0%	3%	7.45%	13%	14.85%	15.8%

Table 6.4: Percentage of total energy consumption increase for each building for increasing values of $F_{\text{day}}/F_{\text{night}}$ compared to the reference case where $F_{\text{day}}/F_{\text{night}} = 1$.

On Figure 6.8, the percentage of energy shifted from the day period (compared to the reference case where $F_{\text{day}}/F_{\text{night}} = 1$) is displayed as a function of the ratio for the two buildings. First, one can see that, for a value of the ratio $F_{\text{day}}/F_{\text{night}}$ superior to 3, the two buildings shift more than 90% of their consumption from the day to the night period. A value of 10 for the ratio yields a shifting of 98.9% and 99.7% of the energy consumed during the day period by the buildings I and E respectively. This figure highlights that building E is more sensitive to the ratio $F_{\text{day}}/F_{\text{night}}$ than building I. Indeed, using the same value of the ratio, the percentage

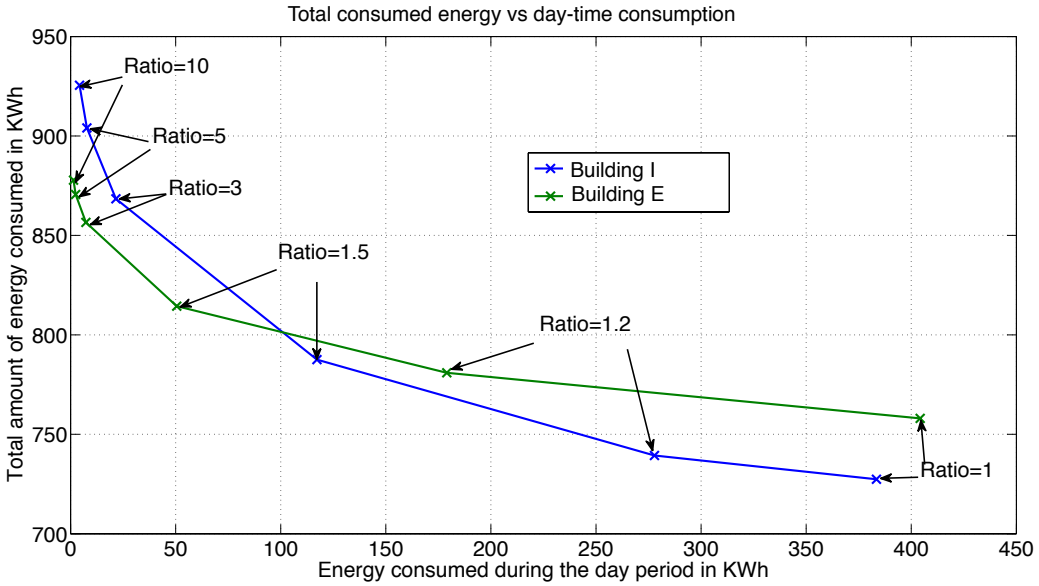

Figure 6.7: Total consumed energy as a function of day time energy consumption. Building E is all the more efficient compared to building I as the ratio $F_{\text{day}}/F_{\text{night}} = 1$ is increased.

of energy shifted from the day period of the building E is always higher than the percentage of the building I.

6.6.4 Efficiency of the heating load shifting from day to night

In the previous section, we have seen that the low-consumption buildings under consideration are able to shift almost all the heating consumption from day to night. To fully characterize this ability in shifting the load to the night period one has to compute the efficiency of these shiftings. First, let us define the following variables

$$\begin{aligned} E^{\alpha}_{n,r} &= \text{Optimal amount of energy consumed during the night period by} \\ &\quad \text{the building } \alpha \text{ with } F_{\text{day}}/F_{\text{night}}\text{=r} \\ E^{\alpha}_{d,r} &= \text{Optimal amount of energy consumed during the day period by} \\ &\quad \text{the building } \alpha \text{ with } F_{\text{day}}/F_{\text{night}}\text{=r} \end{aligned}$$

For the building α with $F_{\text{day}}/F_{\text{night}} = r$, we define the efficiency of the load shifting as follows

$$\eta^{\alpha}_{r} \triangleq \left| \frac{E^{\alpha}_{d,1} - E^{\alpha}_{d,r}}{E^{\alpha}_{n,1} - E^{\alpha}_{n,r}} \right|$$

On Figure 6.9, for each building, the efficiency of the load shifting versus the amount of energy shifted from the day period is displayed. The efficiency to shift almost all

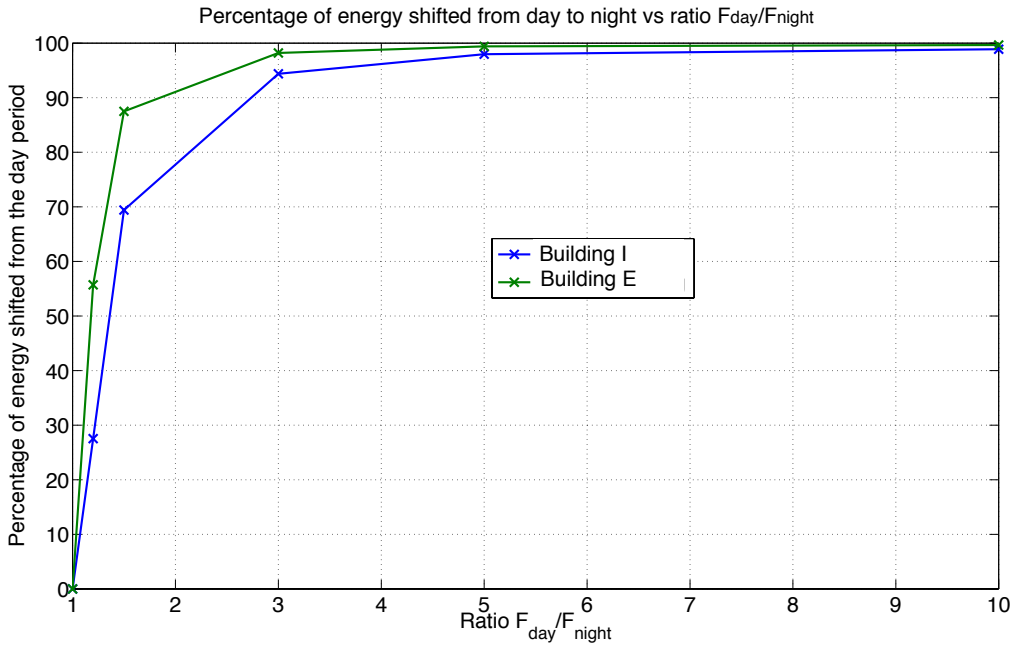

Figure 6.8: Normalized percentage of energy shifted from day to night as a function of the ratio $F_{\text{day}}/F_{\text{night}} = 1$.

the energy consumption to night period is of 66% for the building I and of 77% for the building E. Also, for example, if we wish to work with an efficiency of 85%, it is possible to shift around 50% of the day energy consumption to the night period for the building I and around 90% of this energy for the building E.

It is noticeable that even if the two buildings are efficient in handling long daily load shiftings, the building E is much more efficient than building I. Insulating the concrete core of the building from the exterior increases its thermal mass and therefore its inertia. So, once a set point temperature is reached, the temperature variations around this set-point temperature are really slow. As seen in Section 6.6.3, it might increase the need of anticipation (switching from a set-point to another) but also yields improved flexibility and a higher efficiency.

6.7 Conclusion

In this chapter, we have used our methodology to solve state and input COCPs for low-consumption building heating problems. These type of problems are an efficient way to study the dynamical properties (such as the ability to perform load shiftings) of energetic systems.

We have shown that well insulated buildings can be heated only during night time (10 p.m. to 6 a.m.) while maintaining a certain level of comfort and therefore

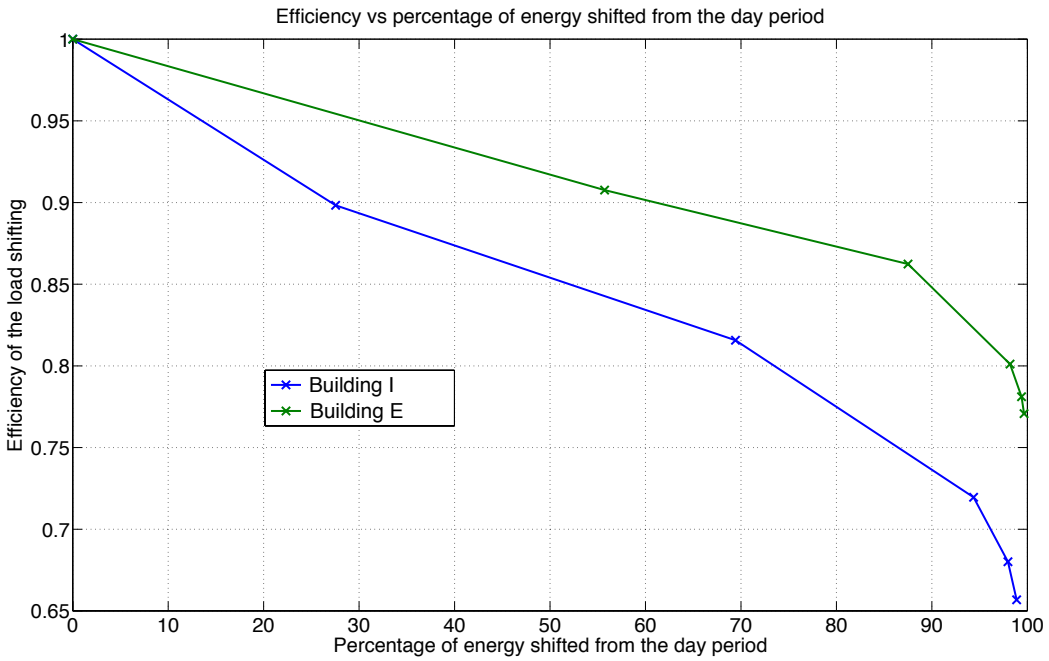

Figure 6.9: Efficiency of the load shifting as a function of the percentage of energy shifted from day to night. Relatively high level of efficiency can be achieved only with building E.

have the ability to reduce the electric thermal sensitivity in a country like France.

In turn, CO_2 consumption of electricity production in France during night time being very low (40 g/MW), this heating strategy makes the electric heating a low CO_2 emitter.

We have also shown that the construction method influences the performances of the load shiftings both in comfort for the inhabitants and in global efficiency. Indeed, insulating the concrete core from the exterior allows the load shiftings to be more comfortable by limiting the need of over-heating and, additionally, significantly increases their global efficiency, e.g. it is possible to shift 90% of energy from the day period to the night time with an efficiency above 85%.

Part III

Bibliographie et annexes
Bibliography and appendices

List of publications

- P. Malisani, F. Chaplais, N. Petit, and D. Feldmann. Thermal building model identification using time-scaled identification methods. *49th IEEE Conference on Decision and Control*, pages 308–315, 2010.

- P. Malisani, B. Favre, S. Thiers, B. Peuportier, F. Chaplais, and N. Petit. Investigating the ability of various buildings in handling load shiftings. *IEEE Conference on Power Engineering and Automation 2011*, pages 393–397, 2011.

- P. Malisani, F. Chaplais, and N. Petit. Design of penalty functions for optimal control of linear dynamical systems under state and input constraints. *50th IEEE Conference on Decision and Control and European Control Conference*, pages 6697 – 6704, 2011.

- P. Malisani, F. Chaplais, and N. Petit. A constructive interior penalty method for non linear optimal control problems with state and input constraints. *IEEE American Control Conference*, pages 2669–2676, 2012.

- P. Malisani, F. Chaplais, and N. Petit. A fully unconstrained interior point algorithm for multivariable state and input constrained optimal control problems. *European Congress on Computational Methods in Applied Sciences and Engineering*, 2012.

Bibliography

[ABZ12] A. Altarovici, O. Bokanowski, and H. Zidani. A general Hamilton-Jacobi framework for nonlinear state-constrained control problems. *Control, Optimisation and Calculus of Variations*, 2012.

[Ada75] R. Adams. *Sobolev spaces*. Academic Press, 1975.

[AES08] M. H. Albadi and E. F. El Saadany. A summary of demand response in electricity markets. *Electric Power Systems Research*, 78:1989–1996, 2008.

[AHU72] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non linear programming*. Stanford University Press, CA, 1972.

[AMR88] U. Ascher, R. Mattheij, and R. Russel. *Numerical solution of boundary value problems for ordinary differential equations*. Prentice Hall, 1988.

[AS04] A. Agrachev and Y. Sachkov. *Control theory from the geometric viewpoint*. Springer, 2004.

[BA07] P. Bertoldi and B. Atanasiu. *Electricity consumption and efficiency trends in the enlarged European union*. Tech. rept. Institute for Environment and Sustainability, 2007.

[BBCI97] J. J. Bézian, P. Barles, F. Claude, and C. Inard. *Les émetteurs de chaleur*. Presses de l'Ecole des Mines, 1997.

[BBW92a] M. Basseville, A. Benveniste, and S. Willsky. Multiscale autoregressive processes, part i: Schur-Levinson parameterizations. *IEEE Transactions on Automatic Control*, 40(8):1915–1934, 1992.

[BBW92b] M. Basseville, A. Benveniste, and S. Willsky. Multiscale autoregressive processes, part ii: Lattice structures for whitening and modelling. *IEEE Transactions on Automatic Control*, 40(8):1935–1954, 1992.

[BCM98] A. Bemporad, A. Casavol, and E. Mosca. A predictive reference governor for constrained control systems. *Computers in Industry*, 36:55–64, 1998.

[Bel57] R. Bellman. *Dynamic programming*. Princeton University Press, 1957.

[BFLT03] B. Bonnard, L. Faubourg, G. Launay, and E. Trélat. Optimal control with state constraints and the space shuttle re-entry problem. *Journal of Dynamical Control Systems*, 9:155–199, 2003.

[BG03] J. F. Bonnans and Th. Guilbaud. Using logarithmic penalties in the shooting algorithm for optimal control problems. *Optimal Control Applications and Methods*, 24:257–278, 2003.

[BH69] A. E. Bryson and Y. C. Ho. *Applied optimal control*. Ginn and Company: Waltham, MA, 1969.

[Bha06] R. Bhattacharya. OPTRAGEN: A Matlab toolbox for optimal trajectory generation. *45th IEEE Conference on Decision and Control*, pages 6832–6836, 2006.

[BHM77] J. D. Balcomb, J. C. Hedstrom, and R. D. McFarland. Simulation analysis of passive solar heated buildings—preliminary results. *Solar Energy*, 19(3):277–282, 1977.

[BMDP02] A. Bemporad, M. Morari, V. Dua, and E. N. Pisikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38:3–20, 2002.

[BMT08] J. F. Bonnans, P. Martinon, and E. Trélat. Singular arcs in the generalized goddard's problem. *Journal of Optimization Theory and Applications*, 139(2):439–461, 2008.

[BNW06] R. H. Byrd, J. Nocedal, and R. A. Waltz. Knitro: An integrated package for nonlinear optimization. In *Large Scale Nonlinear Optimization, 35–59, 2006*, pages 35–59. Springer Verlag, 2006.

[Bra90] J. E. Braun. Reducing energy costs and peak electrical demand through optimal control of building thermal storage. *ASHRAE Transactions*, 96(2):839–84, 1990.

[BSW09] A. Brun, C. Spitz, and E. Wurtz. Analyse du comportement de différents codes de calcul dans le cas de bâtiments à haute efficacité énergétique. *IXème colloque interuniversitaire Franco-québécois sur la thermique des systèmes*, 2009.

[But08] J. C. Butcher. *Numerical methods for ordinary differential equations (Second Edition)*. John Wiley & Sons, 2008.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[CAEA96] F. Chaplais and A. Alaoui El Azher. Two time scaled parameter identification by coordination of local identifiers. *Automatica*, 32(9):1303–1309, 1996.

[Che01] T. Y. Chen. Real-time predictive supervisory operation of building thermal systems with thermal mass. *Energy and Buildings*, 33:141–150, 2001.

[CHT11] M. Cerf, T. Haberkorn, and E. Trélat. Continuation from a flat to a round earth model in the coplanar orbit transfer problem. *Optimal Control Applications and Methods*, 2011.

[CP04] A. Chiuso and G. Picci. On the ill-conditioning of subspace identification with input. *Automatica*, 40:575–589, 2004.

[CP05] D. E. Chang and N. Petit. Toward controlling dielectrophoresis. *International Journal of Robust and Nonlinear Control*, 15:769–784, 2005.

[DS11] D. Da Silva. *Analyse de la flexibilité des usages électriques résidentiels. Application aux usages thermiques.* PhD thesis, Ecole Nationale Supérieure des Mines de Paris, 2011.

[Elm95] H. Elmqvist. *Dymola: modeling language - User's guide.* Dynasim AB, Lund, Sweden, 1995.

[FGK90] R. Fourer, D.M. Gay, and B.W. Kernighan. A modeling language for mathematical programming. *Management Science*, 36:519–554, 1990.

[FGW02] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 4(4), 2002.

[FM68] A. V. Fiacco and G. P. McCormick. *Nonlinear programming: sequential unconstrained minimization techniques.* Wiley : New York, 1968.

[Fro07] R. Fromby. *Task Xi: Time of use pricing and energy use for demand management delivery.* Tech. rept. IEA - international energy agency, 2007.

[FVLA02] G. Fraisse, C. Viardot, O. Lafabrie, and G. Achard. Development of a simplified and accurate building model based on electrical analogy. *Energy and Buildings*, 34:1017–1031, 2002.

[GDP02] M. M. Gouda, S. Danaher, and Underwood C. P. Building thermal model reduction using nonlinear constrained optimization. *Building and Environment*, 37:1017–1031, 2002.

[GKPC10] K. Graichen, A. Kugi, N. Petit, and F. Chaplais. Handling constraints in optimal control with saturation functions and system extension. *Systems and Control Letters*, 59(11):671–679, 2010.

[God19] R. H. Goddard. *A method for reaching extreme altitudes.* Smithsonian Int. Misc. Collections 71, 1919.

[Gon12] J. Gondzio. Interior point methods 25 years later. *European Journal of Operational Research*, 218:587–601, 2012.

[GP08a] K. Graichen and N. Petit. Constructive methods for initialization and handling mixed state-input constraints in optimal control. *Journal Of Guidance, Control, and Dynamics*, 31(5):1334–1343, 2008.

[GP08b] K. Graichen and N. Petit. Solving the Goddard problem with thrust and dynamic pressure constraints using saturation functions. *17th World Congress of The International Federation of Automatic Control*, Proc. of the 2008 IFAC World Congress:14301–14306, 2008.

[GP09] K. Graichen and N. Petit. Incorporating a class of constraints into the dynamics of optimal control problems. *Optimal Control Applications and Methods*, 30:537–561, 2009.

[GPK08] K. Graichen, N. Petit, and A. Kugi. Transformation of optimal control problems with a state constraint avoiding interior boundary conditions. *In Proc. of the 47th IEEE Conference on Decision and Control*, 2008.

[Gra06] K. Graichen. *Feedforward control design for finite-time transition problems of nonlinear systems with input and output constraints.* PhD thesis, Universität Stuttgart, 2006.

[HC11] A. Hoffer and N. Charton. *Marché de capacité : éclairage sur sept paramètres clés du futur mécanisme français.* Tech. rept. E-Cube Strategy consultants, 2011.

[Her08] A. Hermant. *Sur l'algorithme de tir pour les problèmes de commande optimale avec contraintes sur l'état.* PhD thesis, Ecole Polytechnique, 2008.

[HM02] R. Hämäläinen and J. Mäntysaari. Dynamic multi-objective heating optimization. *European Journal of Operational Research*, 142:1–15, 2002.

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

[HS06] J. Hauser and A. Saccon. A barrier function method for the optimization of trajectory functionals with constraints. *45th IEEE Conference on Decision and Control*, pages 864–869, 2006.

[HSV95] R. Hartl, S. Sethi, and R. Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995.

[HT11] T. Haberkorn and E. Trélat. Convergence results for smooth regularizations of hybrid nonlinear optimal control problems. *SIAM Journal on Control and Optimization*, 49(4):1498–1522, 2011.

[JLW03] T. Jockenhövel, T. B. Lorenz, and A. Wächter. Dynamic optimization of the Tennessee Eastman process using the Optcontrolcentre. *Computers and Chemical Engineering*, 27:1513–1531, 2003.

[JM08] M. J. Jiménez and H. Madsen. Models for describing the thermal characteristics of building components. *Building and Environment*, 43:152–162, 2008.

[JMA08] M. J. Jiménez, H. Madsen, and K. K. Andersen. Identification of the main thermal characteristics of building components using MATLAB. *Building and Environment*, 43:170–180, 2008.

[Kai80] T. Kailath. *Linear systems*. Prentice Hall, 1980.

[Kar08] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 2008.

[KF99] A. N. Kolmogorov and S. V. Fomin. *Elements of the theory of functions and functional analysis*. Dover Publications, 1999.

[Kha02] H. Khalil. *Nonlinear systems*. Prentice Hall, 2002.

[KM04] A. Kojima and M. Morari. LQ control for constrained continuous-time systems. *Automatica*, 40:1143–1155, 2004.

[KR10] M. Khosrow and K. Ranjit. A reliability perspective of the smart grid. *IEEE Transactions on Smart Grid*, 1(1):57–64, 2010.

[Lju87] L. Ljung. *System identification, theory for the user*. Prentice-Hall, 1987.

[LK85] D. W. Luse and H. K. Khalil. Frequency domain results for systems with slow and fast dynamics. *IEEE Transactions on Automatic Control*, AC-30(12):1171–1178, 1985.

[LM95] A. Le Mouel. *Contribution à l'étude des caractéristiques dynamiques réduites des systèmes thermiques complexes*. PhD thesis, Ecole Nationale Supérieure des Mines de Paris, 1995.

[LS99] F. Leibfritz and E. W. Sachs. Inexact SQP interior point methods and large scale optimal control problems. *SIAM Journal on Control and Optimization*, 38:272–293, 1999.

[LWR67] L. Lasdon, A. Waren, and R. Rice. An interior penalty method for inequality constrained optimal control problems. *IEEE Transactions on Automatic Control*, 12:388–395, 1967.

[MCPF10] P. Malisani, F. Chaplais, N. Petit, and D. Feldmann. Thermal building model identification using time-scaled identification methods. *49th IEEE Conference on Decision and Control*, pages 308–315, 2010.

[MG86] R. H. Middleton and G. C. Goodwin. Improved finite word length characteristic in digital control using delta operators. *IEEE Transactions on Automatic Control*, 1C-31(1):1015–1021, 1986.

[MGB12] P. Martinon, V. Grélard, and F. Bonnans. *BOCOP v1.03 user guide*, 2012.

[NW99] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[PBGM62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Interscience Publishers John Wiley & Sons, Inc. New York, London, 1962.

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

[PS90] B. Peuportier and I.B. Sommereux. Simulation tool with its expert interface for the thermal design of multizone buildings. *International Journal of Sustainable Energy*, 8(2):109–120, 1990.

[PWMK07] M. A. Piette, D. Watson, N. Motegi, and S Kiliccote. *Automated critical peak pricing field tests: 2006 pilot program description and results*. Tech. rept. Ernest Orlando Lawrence berkeley national laboratory, 2007.

[RF04] I. M. Ross and F. Fahroo. Pseudospectral knotting methods for solving nonsmooth optimal control problems. *AIAA Journal of Guidance, Control and Dynamics*, 27, 2004.

[RS72] S. Roberts and J. Shipman. *Two-point boundary value problems : shooting methods*. American Elsevier, 1972.

[RT206] Décret n° 2006-592 du 24 mai 2006 relatif aux caractéristiques thermiques et à la performance énergétique des constructions., 2006.

[RT211] Arrêté du 11 octobre 2011 relatif aux attestations de prise en compte de la réglementation thermique et de réalisation d'une étude de faisabilité relative aux approvisionnements en énergie pour les bâtiments neufs ou les parties nouvelles de bâtiments, 2011.

[RTE11] RTE. Bilan prévisionnel de l'équilibre offre-demande d'électricité en france, 2011.

[Rug06] Radu Rugescu. Goddard's 85 years optimal ascent problem finally solved. 2006.

[SB93] J. Stoer and R. Burlish. *Introduction to numerical analysis*. Springer-Verlag, 1993.

[Sch78] L. Schwartz. *Analyse Hilbertienne*. Ecole Polytechnique, 1978.

[Sey94] H. Seywald. Trajectory optimization based on differential inclusion. *AIAA Journal of Guidance, Control and Dynamics*, 17:480–487, 1994.

[SKD08] G. Stephanopoulos, O. Karsligil, and M. S. Dyer. Multiscale theory for linear dynamic processes. part 1. foundations. *Computers and Chemical Engineering*, 32:857–884, 2008.

[SKR00] L. Shampine, J. Kierzenka, and M. Reichelt. *Solving boundary value problems for ordinary differential equations in MATLAB with* ***bvp4c***, 2000.

[SZ09] A. Sodja and B. Zupančič. Modelling thermal processes in buildings using an object-oriented approach and modelica. *Simulation Modelling Practice and Theory*, 17:1143–1159, 2009.

[Tré08] E. Trélat. *Contrôle optimal : Théorie et applications*. Vuibert, 2008.

[Tré03] E. Trélat. Optimal control of a space shuttle, and numerical simulations. *Discrete continuous dynamical systems*, pages 842–851, 2003.

[TYO+12] K. Tanaka, A. Yoza, K. Ogimi, A. Yona, and T. Senjyu. Optimal operation of DC smart house system by controllable loads based on smart grid topology. *Renewable Energy*, 39:132–139, 2012.

[Vic98] L. N. Vicente. On interior-point Newton algorithms for discretized optimal control problems with state constraints. *Optimization Methods and Software*, 8:249–275, 1998.

[WB06] A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[Wri93] S. J. Wright. Interior point methods for optimal control of discrete time systems. *Journal of Optimization Theory and Applications*, 77:161–187, 1993.

[Wri04] M. H. Wright. The interior-point revolution in optimization: History, recent developments, and lasting consequences. *Bulletin (New Series) of the American Mathematical Society*, 42:39–56, 2004.

[XHBH04] P. Xu, P. Haves, J.E. Braun, and L. T. Hope. Peak demand reduction from pre-cooling with zone temperature reset in an office building. *Proceedings of the ACEEE 2004 Summer Study on Energy Efficient in Buildings*, 2004.

[YGFDD05] J. Yuz, G. Goodwin, A. Feuer, and J. De Doná. Control of constrained linear systems using fast sampling rates. *Systems and Control Letters*, 54:981–990, 2005.

[ZDG96] K. Zhou, J. C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, 1996.

Appendix A

Technical description of electric appliances playing a role in active demand response

A.1 Electric heating

An electric heating device can be a dynamical system itself, for example, an electric storage heater, or can be considered as a static system but coupled with a dynamical system such as the building.

A.1.1 Convector heaters

The principle of this device consists in letting cool air in at the bottom of the convector to heat it with an electrical resistance. Then by convection, the air comes out from the superior part of the device. Each equipment can efficiently heat 15 to 20 m^2 rooms in which the ceiling is not too high. Because of their important heat emission by convection, these systems lead to a higher stratification[1] compared to other systems, i.e. temperature rises with height up to 1.2°C per meter in old buildings and up to 0.5°C in well insulated buildings [BBCI97, DS11]. Heating powers generally range from 750 to 2000 watts. For those systems, the part of heat emitted by radiation usually reaches 5 to 10% [BBCI97, DS11].

Generally, the dynamics of this equipment is considered as instantaneous (negligible) compared to that of the building in which it is installed and is therefore not taken into account in establishing optimization models.

A.1.2 Radiative heaters

This type of heating is made of a heating unit, which releases an important part of heating, by radiation (approximately 40% according to [BBCI97]). The inertia of these systems (the total mass varies from 7 to 20 kg depending on the model) can be slightly superior to that of electric convectors.

A.1.3 Storage heaters

As opposed to the previous systems, the inertia of this type of systems is not negligible towards the building dynamics. Indeed, since these systems can be made

[1] Gradient of temperature depending on the height.

of refractory bricks of high density or granite crush or even lava, they have a very important thermal capacity.

This heating system allows for a heating storage in thermal form on an intraday horizon. For example, they allow the customer to buy electricity at night, to store it as heat, and to restore it during the day. As indicated in [DS11], two types of storage heaters exist: static storage heaters and dynamical storage heaters.

A.1.3.1 Static storage heaters

These devices store heat using electrical resistances. The heat release is only static, non-controllable.

A.1.3.2 Dynamical storage heaters

For this type of devices, the heat release can be accelerated using a fan that is installed in the inferior part of the system. The fan allows the cool air to circulate inside the storage heater and thus to increase the heat release of the device by convection.

A.1.4 Heat pumps

A heat pump is a thermodynamical device allowing a heat transfer from a cold source to a hot source thanks to a refrigerated device, generally a compression mechanism. This refrigerated device is made of at least the four following elements:

- **Compressor**: the compressor is first going to pump the low-pressure low-temperature refrigerant gas. The mechanical energy provided by the compressor is going to raise the pressure and temperature of the refrigerant gas.

- **Condenser**: the condenser is a heat exchanger in which are circulating both the exterior fluid to heat (air or water) and the refrigerant fluid. The hot gases transmit their heat to the exterior fluid to heat: it is the phase of desuperheating of the high-pressure gases up until condensation. This condensation temperature is superior to the exterior fluid temperature.

- **Expansion device**: the liquid formed into the condenser shifts from high-pressure to low-pressure. This shifting occurs in an expansion valve or in a capillary aperture. During this shifting, a slight formation of gas occurs. This happens with no exchange with the exterior whatsoever: no heat nor mechanical energy.

- **Evaporator**: at a low-pressure, the equilibrium temperature liquid-steam is lower than the temperature of the exterior fluid. The evaporator is a heat exchanger in which circulate both the refrigerant fluid from the expansion valve, and the exterior fluid (air or water) from which heat is derived (air or water). The liquid refrigerant fluid from the expansion valve then starts to boil

in the evaporator by absorbing heat from the exterior fluid. The compressor then sucks up the gas for another cycle.

Air/water heat pumps stand for the majority of models installed in France. This type of heat pump can be installed for heating (some installations exist with and without storage tank, according to the inertia of water loop of the system) or for heating and domestic hot water (with storage tank). They are cheaper than other types of water heat pumps, but their coefficient of performance (COP) is inferior because of lower temperatures from the heat source, and because they need a defrosting system or an auxiliary heating system when the COP is poor (sometimes these systems are also equipped with electric resistance).

A.2 Domestic hot water with storage tank

In the residential sector, domestic hot water (DHW) stands for approximately 13% of electricity consumption. DHW systems with storage tank can be made of a vertical and horizontal cylinder. The water is heated with an electric resistance. This resistance is generally located at the centre of a vertical cylinder, but the cylinder can also possess resistances installed horizontally or even possess several resistances. With the heating, there is a stratification effect (convection), which means that hot water goes up while colder water remains down. The hot water is then removed from the superior part and the cold water comes through the inferior part of the reservoir.

The heat provided can also be produced, in recent models, by a thermodynamical cycle (thermodynamical tank or heat pump), where the heat is no longer provided by a resistance but by an exchanger.

A.3 Electrical storage

The battery is often integrated to a photovoltaic generator. The battery can be used to shift the solar energy production to synchronize it either with the local consumption or with the requirements of the grid. The battery can also be charged from the grid and thus participate to the intraday smoothing of the load curve.

Appendix B

Examples of energy optimization

B.1 PV shifting

B.1.1 Model

$$\begin{aligned} PV(t) &= \text{collected photovoltaic power} \\ r(.) &= \text{efficiency} \end{aligned}$$

As for the battery, here are the following used notations

$$\begin{aligned} C_{\max} = 3.2 &: \text{battery capacity kWh} \\ x(t) &: \text{battery state of charge in kWh} \\ P_{\max} = 2.2 &: \text{maximal power of the battery (charge and discharge) in kW} \\ u_1(t) &: \text{percentage of power directly sold} \\ u_2(t) &: \text{discharge power} \end{aligned}$$

Later on we will use the following variable

$$P_{\text{pv}}(t) \triangleq \min\{P_{\max}, PV(t)\}$$

Thus, $P_{\text{pv}}(t)$ stands for the power that can be potentially stored in the battery. Indeed, the part of the power superior to 2 kW is directly sold on the grid. Now, the dynamical equation of the battery is as follows

$$\dot{x}(t) = 0.95(1 - u_1(t))P_{\text{pv}}(t) - u_2(t) - \frac{x(t)}{100}$$

B.1.2 Optimal control problem

The optimal control problem to solve is

$$\max_{u_1,u_2} \int_0^T \text{price}(t)\Big[PV(t) - (1 - u_1(t))P_{\text{pv}}(t) + 0.95u_2(t)\Big]dt$$

under the constraints

$$\begin{aligned} x(t) &\in [0, C_{\max}] \\ u_1(t) &\in [0, 1] \\ u_2(t) &\in [0, P_{\max}] \\ x(0) &= 2 \\ x(T) &= 2 \end{aligned}$$

B.1.3 Results

Figure B.1 displays the price of electricity per kWh over one week. Figures B.2 and B.3 respectively display the photovoltaic power injected in the battery ($(1-u_1)P_{\text{pv}}$) and the optimal discharge power (u_2). Figure B.4 represents the optimal charge of the battery (x) over one week.

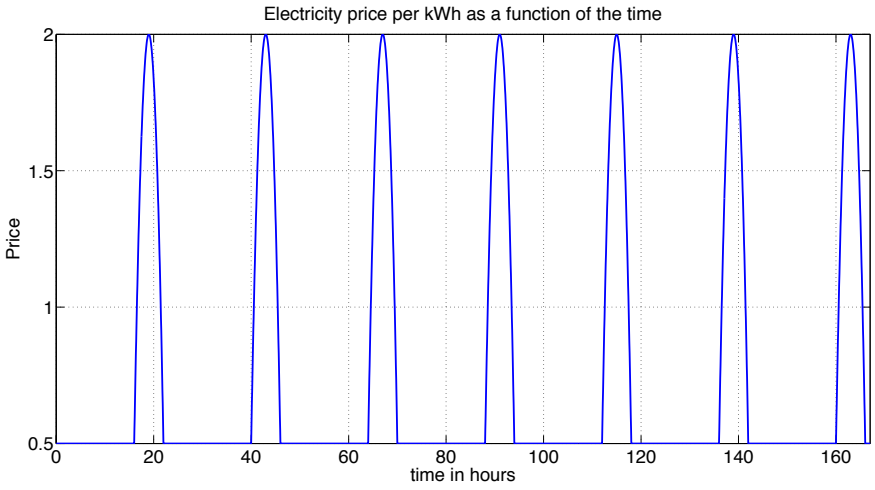

Figure B.1: Price of electricity per kWh over one week.

B.2 Electricity pricing efficiency

B.2.1 Model of the HWB storage tank

B.2.1.1 Modeling

The hot water boiler (HWB) model is a two layer model yielding a state vector of dimension 2

$$x_{\text{hwb}}(t) = \begin{pmatrix} x^1_{\text{hwb}}(t) \\ x^2_{\text{hwb}}(t) \end{pmatrix}$$

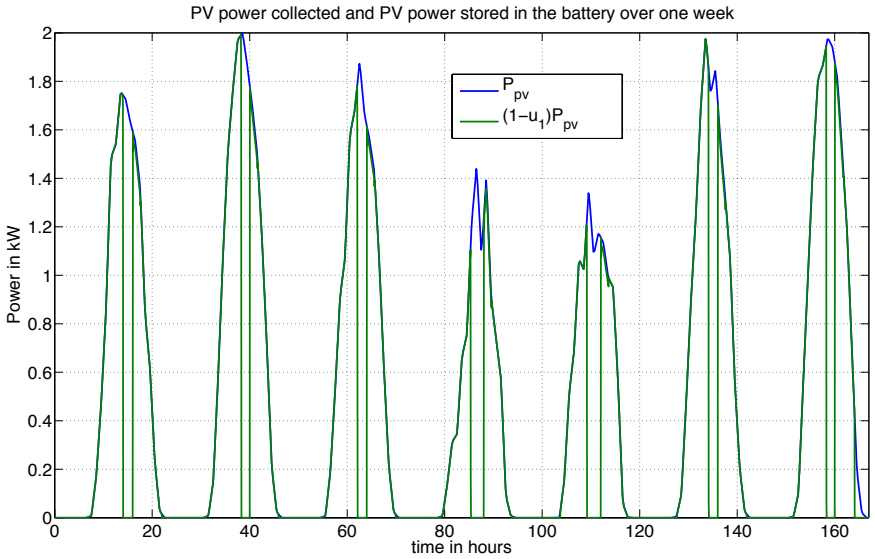

Figure B.2: Collected photovoltaic power (P_{pv}) and photovoltaic power fed into the battery ($(1-u_1)P_{\text{pv}}$) over one week.

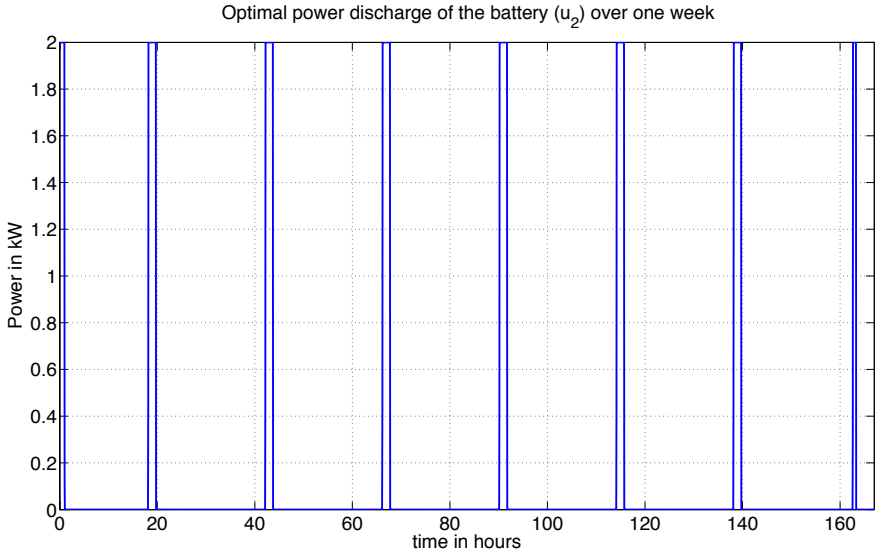

Figure B.3: Optimal power of discharge (u_2) over one week.

where x^1_{hwb} (resp. x^2_{hwb}) is the temperature of the lower layer (resp. high). Finally, we find the following state equation

$$\dot{x}_{\text{hwb}}(t) = A(x_{\text{hwb}}(t), \dot{m}(t))x_{\text{hwb}}(t) + B_{\text{Puis}}u_2(t) + B_{\text{Tef}}(x_{\text{hwb}}(t), \dot{m}(t))T_{\text{wi}} + B_{\text{Tamb}}T_{\text{room}} \tag{B.1}$$

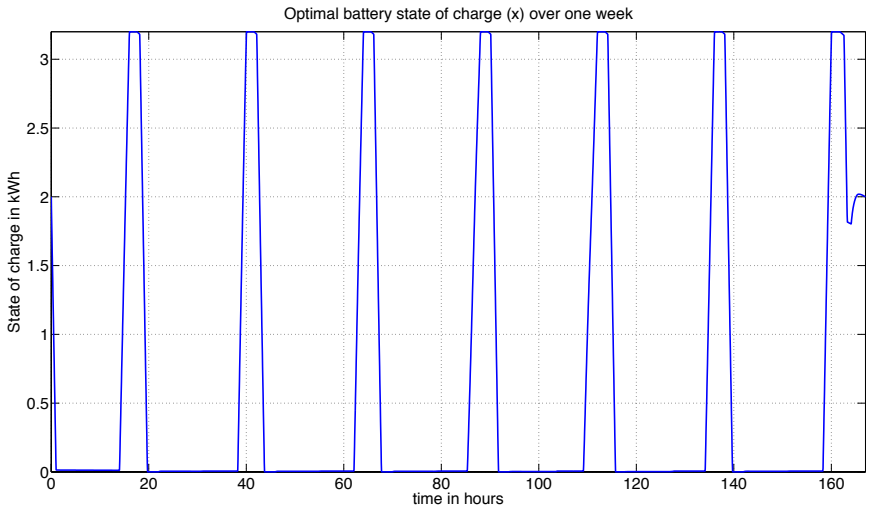

Figure B.4: Optimal state of charge of the battery over one week. The charge values must range from 0 to 3kWh. The conditions at the level of the edges are $x(0) = x(T) = 2$.

where $u_2(t)$ is the electric power consumed by the storage tank, $\dot{m}(t)$ stands for the extraction of hot water from the storage tank, T_{wi} is the temperature of the water input, T_{room} is the temperature of the room where the HWB is stored. These temperatures are chosen as follows

$$T_{\text{wi}} = 12°\text{C}, \qquad T_{\text{room}} = 19°\text{C}$$

The matrices $A(x_{\text{hwb}}(t), \dot{m}(t))$ and $B_{\text{Tef}}(x_{\text{hwb}}(t), \dot{m}(t))$ are matrices whose coefficients are non-linear continuous functions of $x_{\text{hwb}}(t)$ and $\dot{m}(t)$. For confidentiality reasons, we do not give additional informations on this model.

B.2.2 Building model

The model of the building used for this example corresponds to the model n°2 from Table 5.1 after balanced reduction of order 4.

$$\begin{cases} \dot{x}_{\text{bui}} &= Ax_{\text{bui}}(t) + Bu_1(t) + d(t) \\ T_{\text{bui}}(t) &= Cx_{\text{bui}}(t) \end{cases} \tag{B.2}$$

B.2.3 Optimal control problem

The optimal control problem is the following

$$\min_{u_2,u_1} \left[\int_0^T \text{price}(t) \left(u_2(t) + u_1(t)\right) dt \right]$$

with

$$\text{price}(t) \begin{cases} 1.5 \text{ between 6 a.m. and 11 p.m.} \\ 1 \text{ otherwise} \end{cases}$$

Under the dynamical constraints (B.1) and (B.2), under the following control constraints

$$\begin{aligned} u_1 &\in [0, 20\,\text{kW}] \\ u_2 &\in [0, 2.2\,\text{kW}] \end{aligned}$$

and the following state constraints

$$Cx_{\text{bui}} \in \begin{cases} [14, 22] \text{ between 9 a.m. and 4 p.m.} \\ [20, 22] \text{ otherwise} \end{cases}$$

$$x^2_{\text{hwb}} \in \begin{cases} [57, 62] \text{ between 5 a.m. and 6 a.m.} \\ [20, 62] \text{ otherwise} \end{cases}$$

B.2.4 Results

Figures B.5 and B.6 respectively represent the heating optimal power for the building (u_1) and the heating optimal power for HWB (u_2) over one week.
Figures B.7 et B.8 respectively represent the optimal indoor temperature (T_{bui}) and the optimal temperatures of the two layers of the HWB storage tank (x_{hwb}) over one week.

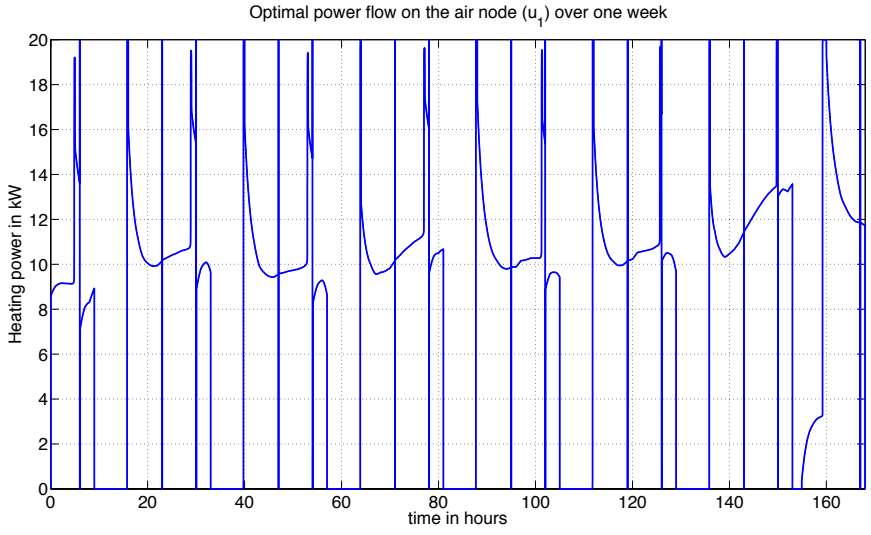

Figure B.5: Optimal heating power on the air node over one week.

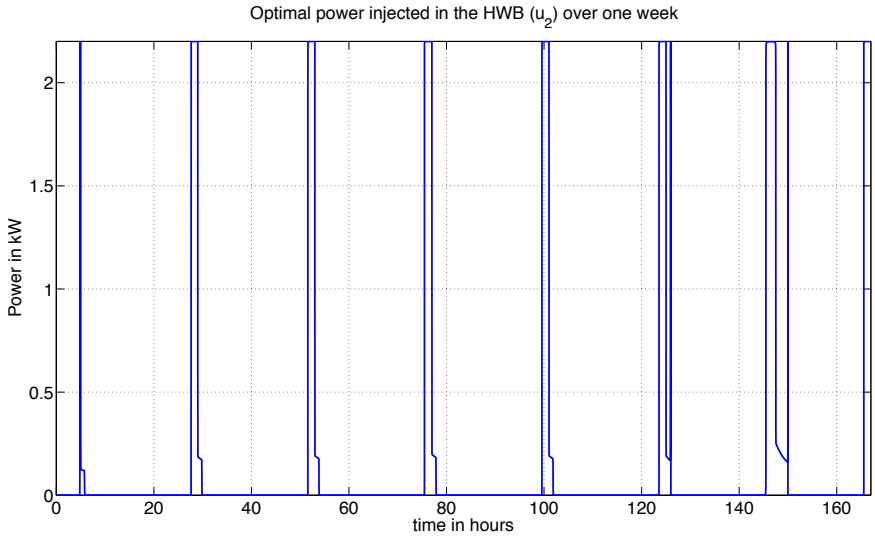

Figure B.6: Optimal heating power of HWB over one week.

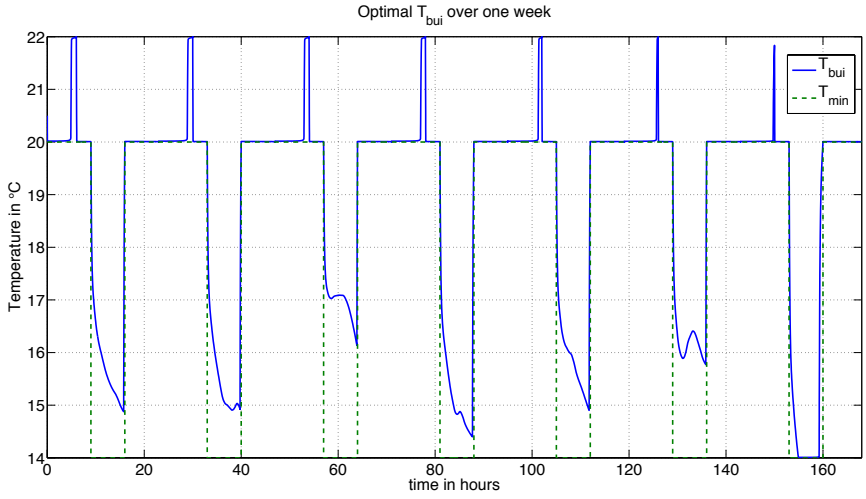

Figure B.7: Optimal indoor temperature over one week.

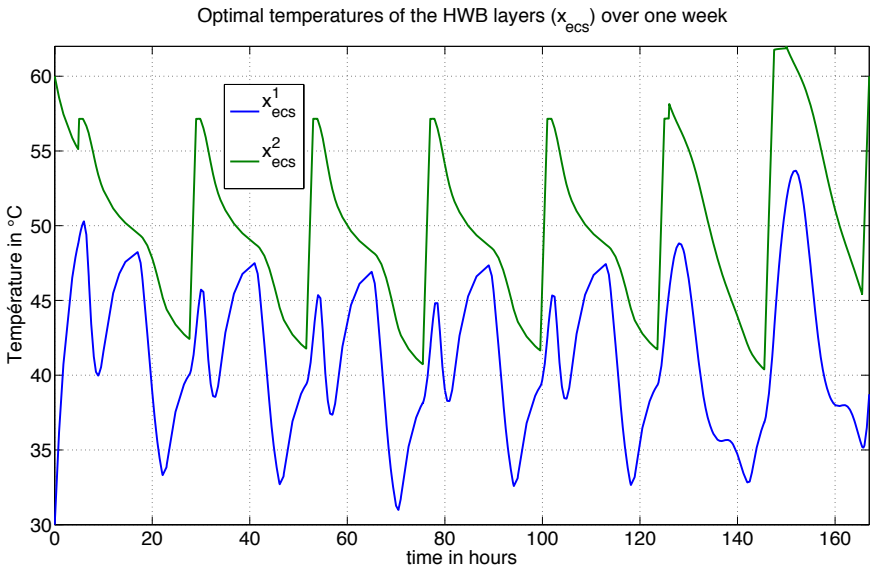

Figure B.8: Optimal temperatures of the two layers of the HWB over one week.

Appendix C

Proofs of some results of Chapter 3

C.1 Proof of Proposition 2

From Assumption 1, there exists two closed ball B_N and B_M such that

$$B_N \subset \mathcal{C} \subset B_M$$

with *strict* inclusions. We define $N > 0$ (resp. $M > 0$) as the radius of the ball B_N (resp. B_M).Now, if $u = 0$, then $G_{\mathcal{C}}(u)$ is well defined and is equal to 0. We now assume that $u \neq 0$. Then

$$N \frac{u}{\|u\|} \in \mathcal{C}$$

because it has norm N; as a consequence $u \in \frac{\|u\|}{N}\mathcal{C}$ which proves that $G_{\mathcal{C}}(u)$ is well defined and upper bounded by $\frac{\|u\|}{N}$. This proves property a) and the right hand side inequality of (3.4).

On the other side, if $u \neq 0$ then

$$M \frac{u}{\|u\|} \notin \mathcal{C}$$

because its norm is M. As a consequence $u \notin \frac{\|u\|}{M}\mathcal{C}$, and $u \notin \lambda\mathcal{C}$ if $\lambda \leq \frac{\|u\|}{M}$. Then, $G_{\mathcal{C}}(u)$ is lower bounded by $\frac{\|u\|}{M}$; this also holds if $u = 0$. This end the proof of property b).

The positive homogeneity of the gauge is trivial; since it is sub-additive [Sch78], it is convex. The continuity comes from the fact that it is convex and lower and upper bounded in the neighborhood of any point. This proves properties c) and d).

The Dini derivative at 0 is obtained by observing that $G_{\mathcal{C}}(0) = 0$ and that $\frac{G_{\mathcal{C}}(hd)}{h} = G_{\mathcal{C}}(d)$ if $h > 0$. We see that there exists a directional derivative at 0 along the direction d if and only if the Dini derivatives along the directions d and $-d$ are equal, which is equivalent to the intersection of $\mathcal{C}$ with the line directed by d being symmetrical with respect to 0. This proves property e). Note that, if this symmetry holds for all directions, then the gauge function is a norm.

Let us prove property f). Since the boundary is continuously differentiable, there exists a continuously differentiable function $\varphi : \mathbb{R}^m \mapsto \mathbb{R}$ such that $\partial\mathcal{C} = \{u \text{ s.t. } \varphi(u) = 0\}$. For all $u \in \mathbb{R}^m \setminus \{0\}$, $\lambda u \in \partial\mathcal{C} \Leftrightarrow g(u, \lambda) \triangleq \varphi(\lambda u) = 0$. In the following, for any $u \in \mathbb{R}^m \setminus \{0\}$, we consider λ such that $g(u, \lambda) = 0$. From

the convexity of $\mathcal{C}$ and since 0 belongs to the interior of $\mathcal{C}$, one has $\frac{\partial g}{\partial \lambda}(u,\lambda) =< \nabla\varphi(\lambda u), u >\neq 0$ for all $u \in \mathbb{R}^m \setminus \{0\}$. Using the implicit function theorem, there exits $(-\alpha,\alpha) \subset \mathbb{R}$ and U a neighborhood of u and a C^1 function $h : U \mapsto (-\alpha,\alpha)$ such that $\forall \mu \in (\lambda-\alpha, \lambda+\alpha)$ and $\forall v \in U$ $g(v,\mu) = 0 \Leftrightarrow \mu = h(v) = G_{\mathcal{C}}(v)$. Therefore $G_{\mathcal{C}}$ is C^1 on $\mathbb{R}^m \setminus \{0\}$. This proves f).

Let us now prove property g). We first verify easily that $u \in \mathcal{C}$ if and only if $G_{\mathcal{C}}(u) \leq 1$ because $\mathcal{C}$ is closed [Sch78]. Moreover, for any $u \neq 0$, the intersection of $\mathcal{C}$ with the half axis directed by u is the segment $\left[0, \frac{u}{G_{\mathcal{C}}(u)}\right]$ because $\mathcal{C}$ is closed and $G_{\mathcal{C}}(u) > 0$ [Sch78]. As a consequence $G_{\mathcal{C}}(u) = 1$ implies that u is in the boundary of $\mathcal{C}$. Conversely, if $G_{\mathcal{C}}(u) = 1 - 2\alpha$ with $\alpha > 0$, since $G_{\mathcal{C}}$ is continuous, there exists a neighborhood V of u where $G_{\mathcal{C}}(u) \leq 1-\alpha$. For all elements $v \in V$, the intersection of $\mathcal{C}$ with the half-axis directed by v contains $\left[0, \frac{v}{1-\alpha}\right]$. This implies the existence of a neighborhood of u that is included in $\mathcal{C}$, and hence that u is interior to $\mathcal{C}$. Similarly, if $G_{\mathcal{C}}(u) > 1$, $u \notin \mathcal{C}$, one shows the existence of a neighborhood V of u and of $\alpha > 0$ such that the intersection of $\mathcal{C}$ with the half-axis directed by $v \in V$ is included in $\left[0, \frac{v}{1+\alpha}\right]$. Therefore, u belongs to the exterior of $\mathcal{C}$. A consequence of all this is that the boundary of $\mathcal{C}$ is exactly defined by $G_{\mathcal{C}}(u) = 1$, its interior by $G_{\mathcal{C}}(u) < 1$, and its exterior by $G_{\mathcal{C}}(u) > 1$. This ends the proof.

C.2 Proof of Proposition 6

The result is trivial if $\alpha = 0$. We now assume $\alpha > 0$. From Proposition 1 and from the continuous differentiability of the g_i, there exists a constant Γ such that, for all $u \in \mathcal{U}$ and any s, t in $[0,T]$

$$|g_i(x^u(t)) - g_i(x^u(s))| \leq \Gamma |t-s| \tag{C.1}$$

Let $\alpha \in (0, \alpha_0]$ and $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$. Then, there exists an index i for which $g_i(x^u)$ reaches 0 in $[0,T]$. Remember that $g_i(x_0) = -\alpha_0 < 0$. Denote by t_2 the first instant at which $g_i(x^u) = 0$ and $t_1 = \max\{s < t_2 \text{ s.t. } g_i(x^u(s)) = -\alpha \in [-\alpha_0, 0)\}$. From equation (C.1), we have

$$\alpha = g_i(x^u(t_2)) - g_i(x^u(t_1)) \leq \Gamma |t_2 - t_1| = \Gamma(t_2 - t_1)$$

As a consequence, we have $(t_2 - t_1) \geq \alpha/\Gamma$. Then, we have

$$-\alpha \leq g_i(x^u(s)) \leq 0 \quad \forall s \in [t_1, t_2]$$

and hence $\mu_{g_i}(u,\alpha) \geq t_2 - t_1 \geq \alpha/\Gamma$. This concludes the proof.

C.3 Proof of Proposition 9

To exhibit an upper bound on the variation of the cost, this variation is split into three additive terms, bounding respectively the variation of the original cost, of the integral of the state penalty, and the integral of the control penalty.

Define $M = \max_i M_i$. From §3.2.2.1, one readily sees that

$$\|u - v\|_{L^1} \leq 2\alpha M \mu_u(\alpha)$$

We now proceed to establish bounds for the various terms.

C.3.1 Upper bound on the variation of the original cost

Here, an upper bound on $|\int_0^T \ell(x^v, v) - \ell(x^u, u)dt|$ is exhibited. It is noted K_ℓ. From Proposition 1, there exist $\Lambda \geq 0$ such that

$$\begin{aligned} K_\ell \leq \Lambda \int_0^T \|x^v - x^u\|_{L^\infty} + \| v(t) - u(t) \| dt &\leq \Lambda [CT + 1] \|v - u\|_{L^1} \\ &\leq \Lambda[CT + 1]2\alpha M \mu_u(\alpha) \end{aligned}$$

Define $U_l = \Lambda(CT + 1)2M$; then

$$K_l \leq U_l \alpha \mu_u(\alpha) \tag{C.2}$$

C.3.2 Upper bound on the variation of the state penalty

Note $K_{\gamma_g} \triangleq \varepsilon \sum_{i=1}^q \int_0^T \gamma_g \circ g_i(x^v) - \gamma_g \circ g_i(x^u)dt$. Because γ_g is increasing, the integrand is positive only when $g_i(x^v(t)) \geq g_i(x^u(t))$. Yet, from the construction of v in (3.9), one has $\max_i g_i(x^v(t)) \leq -\beta_0$ for all $t \in [0, T]$. Using the convexity of γ_g, and the fact that g_i is Lipschitz with constant K_g on X^{ad}, one obtains

$$\begin{aligned} K_{\gamma_g} &\leq \varepsilon \sum_{i=1}^q \int_{g_i(x^v(t)) \geq g_i(x^u(t))} \gamma_g \circ g_i(x^v) - \gamma_g \circ g_i(x^u)dt \\ &\leq \varepsilon \sum_{i=1}^q \int_{g_i(x^v(t)) \geq g_i(x^u(t))} |g_i(x^u(t)) - g_i(x^v(t))| \gamma_g'(g_i(x^v(t)))dt \\ &\leq \varepsilon q \int_0^T K_g \|x^u - x^v\|_\infty \gamma_g'(-\beta_0)dt \\ &\leq \varepsilon q T K_g C \|u - v\|_{L^1} \gamma_g'(-\beta_0) \\ &\leq \varepsilon q T K_g C \gamma_g'(-\beta_0) 2\alpha M \mu_u(\alpha) \end{aligned} \tag{C.3}$$

Define

$$U_g(\varepsilon) = \varepsilon q T K_g C \gamma_g'(-\beta_0) 2M$$

then, we have

$$K_{\gamma_g} \leq U_g(\varepsilon) \alpha \mu_u(\alpha)$$

C.3.3 Upper bound on the variation of the control penalty

There, we aim at getting a negative variation so that, as a whole, the cost is decreased when replacing u by v.

Define

$$K_u \triangleq \varepsilon \sum_{i=1}^{p} \int_0^T \gamma_u(G_{\mathcal{C}_i}(v_i(t))) - \gamma_u(G_{\mathcal{C}_i}(u_i(t)))dt.$$

From the construction of v (3.9), we know that $G_{\mathcal{C}_i}(v_i(t)) \leq G_{\mathcal{C}_i}(u_i(t))$. Since γ_u is non decreasing, this proves that the integral is negative or null. Moreover, since $u_i = v_i$ when $G_{\mathcal{C}_i}(u_i) < 1 - \alpha_i$, we have

$$K_u = \varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha_i} \gamma_u(G_{\mathcal{C}_i}(v_i(t))) - \gamma_u(G_{\mathcal{C}_i}(u_i(t)))dt$$

Using the convexity of γ_u, one has

$$\begin{aligned}
K_u &\leq -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'(G_{\mathcal{C}_i}(v_i(t)))dt \\
&= -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u' \left[(1-2\alpha) G_{\mathcal{C}_i}(u_i(t))\right] dt \\
&\leq -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u' \left[(1-2\alpha)(1-\alpha)\right] dt \\
&\leq -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'(1-3\alpha)dt \\
&\leq -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} 2\alpha \| G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'(1-3\alpha)dt \\
&\leq -\varepsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i) \geq 1-\alpha} 2\alpha(1-\alpha)\gamma_u'(1-3\alpha)dt \\
&= -\varepsilon \sum_{i=1}^{p} \mu_{u_i}(\alpha)\alpha\gamma_u'(1-3\alpha) \\
&\leq -\varepsilon\alpha\gamma_u'(1-3\alpha)\mu_u(\alpha) \qquad \text{(C.4)}
\end{aligned}$$

C.3.4 An upper bound on $K(u_2, \varepsilon) - K(u_1, \varepsilon)$

Gathering equations (C.2,C.3,C.4) we obtain

$$K(v, \varepsilon) - K(u, \varepsilon) \leq \alpha \left[U_\ell + U_g(\varepsilon) - \varepsilon\gamma_u'(1-3\alpha)\right] \mu_u(\alpha)$$

This concludes the proof of Proposition 9. One can see that the variation is negative for α small enough if $\gamma_u'(1-\alpha)$ tends to $+\infty$ when α tends to 0.

C.4 Proof of Proposition 10

Let us define $f : B_{\|.\|}(0,1) \mapsto \text{int}(\mathcal{C})$ as

$$f(\xi) = \begin{cases} 0 \text{ if } \xi = 0 \\ \dfrac{\|\xi\|}{G_{\mathcal{C}}(\xi)}\xi \text{ otherwise} \end{cases}$$

The differentiability of the function f on $\mathbb{R}^m \setminus \{0\}$ stems from the differentiability of both $\|.\|$ and $G_{\mathcal{C}}$. The continuity at 0 stems from (3.4). Its inverse is given by the following function

$$f^{-1}(\xi) = \begin{cases} 0 \text{ if } \xi = 0 \\ G_{\mathcal{C}}(\xi)\dfrac{\xi}{\|\xi\|} \text{ otherwise} \end{cases}$$

Similarly, the differentiability of the function f^{-1} on $\mathbb{R}^m \setminus \{0\}$ stems from the differentiability of both $\|.\|$ and $G_{\mathcal{C}}$. The continuity at 0 stems from (3.4).
Using equation (3.15), the function

$$\begin{aligned} \phi(\nu) \triangleq f \circ \psi(\nu) &= \tanh(\| \nu \|) \left[G_{\mathcal{C}} \circ \psi(\nu)\right]^{-1} \tanh(\| \nu \|) \frac{\nu}{\| \nu \|} \\ &= \tanh^2(\|\nu\|) \left[G_{\mathcal{C}} \circ \psi(\nu)\right]^{-1} \frac{\nu}{\|\nu\|} \end{aligned}$$

maps $\mathbb{R}^m$ into $\text{int}(\mathcal{C})$. This mapping being the composition of two homeomorphism not differentiable only in 0, ϕ is a homeomorphism differentiable everywhere except at 0. The inverse function $\sigma : \text{int}(\mathcal{C}) \mapsto \mathbb{R}^m$ is the following:

$$\sigma(u) \triangleq \psi^{-1} \circ f^{-1}(u) = \text{atanh}(G_{\mathcal{C}}(u)) \frac{u}{\| u \|}$$

This concludes the proof.

Appendix D

Identification of building models

D.1 Introduction

According to ([FVLA02], [GDP02],[JMA08]), low order linear models form a good set of models to describe the general thermal behavior of buildings. But, as has been stressed in [JM08], these models can give quite good results on prediction errors while providing poor estimates of the building physical characteristics. This is a serious problem in the presented context of optimal control (especially under constraints) which requires good estimates of poles, zeros and static gains.

Usually, such bad performances can be the result of a bad conditioning of the identification optimization problem. For the three identification methods presented here, these optimizations are formulated as quadratic problems, and the condition is the conditioning of the excitation matrix (or matrices).[1]It is related to the sensitivity of the solution of $Ax = b$ with respect to variations of A or b.

Ill conditioning of the excitation matrix(ces) can be the result of insufficient frequency content in the input data; it can be also related to near collinearity of the state and future input subspaces [CP04]. However, it has been proved in [CAEA96] that, even for inputs which are rich enough in the frequency domain, the excitation matrix of two time scaled systems (such as low consumption buildings) is asymptotically degenerate as the ratio between the large and small time constants of the system tends to the infinity. Identifying these systems locally in the frequency domain removes these degeneracy problem.

It should be noted that, in the last two or three decades, time scales have been largely associated to wavelet transforms. Wavelets can be used in several ways in dynamical systems identification. The first usage is for data filtering. Indeed, we could use wavelet transforms to separate frequency bands in the data. However, if one sticks to the popular dyadic transforms, one is limited to time scales which are equal to powers of 2. More classical low-pass and high-pass filters are more flexible, and quite sufficient for our purpose. The other usage is to model the system directly in the wavelet domain. Characterization of finite dimensional systems in this domain have been studied in ([BBW92a] , [BBW92b]). Reference [SKD08] covers a similar topic. A limitation is that these processes are hardly (or even not at all) related to classical (rational) Linear Time Invariant (LTI) systems. The most visible reason for this is that the transforms from the time domain to the wavelet

[1]We recall that, for the L^2 norm, the condition number [SB93] of a matrix A is the ratio between the largest and the smallest eigenvalues of $A^T A$.

domain and back are not causal; therefore it seems unlikely that operations in the wavelet domain can be turned into causal operations in the time domain.

The purpose of this Appendix is to compare the performance of a classical ARX identification procedure to two variants of the two time scaled identification (see [CAEA96]), for the purpose of modeling a low consumption building with a second order model. The difference with [CAEA96] is that we are never in the model matching case. The performance is considered both in terms of simulation error with respect to a high order model, and in robustness with respect to data corruption.

This Appendix is organized as follows.

In Section D.2, we describe the plant we wish to identify, and define various data sets that will be used for that purpose. For comparison purposes, we introduce here data sets where each input generates a separate output; actually, we currently have access to the sum of these outputs, that is, the temperature inside the building. It is interesting to consider this possibility because it gives more information on the system, and we wish to evaluate the benefits of having access to that extra information.

In Section D.3, we describe the various model classes within which we will look for a model, and how we parameterize them with a finite set of numbers. This where we introduce two time scaled models. We detail how the parameters of a model class are related to the parameters of another one.

In Section D.4, we define the various optimization problems which, with the parameterizations of D.3 and the data sets of Section D.2, will define how the various parameters used in the model classes are obtained from the data sets. The definition of these optimization problem are important because the plant does not match any model of any class of Section D.3. Indeed, the output data is generated by a LTI system of order 47 (possibly corrupted with noise), whereas we are looking for a model of order 2. Therefore the choice of the optimization problems greatly influences the determination of the system parameters.

In Section D.5 we compare the results obtained in terms of static gains identification, statistical properties of simulation errors, conditioning of the optimization problems and poles and zeros locations. This is done using various data sets, models, and model parameterizations. These results are interpreted in the light of simulation accuracy and robustness with respect to data corruption.

In Section D.6, we conclude on the results and show the substantial efficiency of the time scaled method in terms of simulation errors and robustness of the parameters identification to noises.

D.2 Plant and data

Our desired goal is to obtain a low-order thermal model of a one-area building describing the general behavior of the internal temperature depending on several inputs. We need those models to optimally control the heating of a building under

constraints. We shall use a high order (47^{th}) linear system as the "true" input-output mapping. This high-order model is a spatial discretization of the heat equation in the building.

The inputs and output and listed in Table D.1. The control of this system is a part of the last input, together with human activities. We consider a person to be a constant input of $100W$ and we also know the heat provided by the devices inside the house. For identification purposes, we use inputs which are an average of chronicles

Output	input
Internal temperature	External temperature Solar flux on the floor Solar flux on the walls Heating flux on the air node

Table D.1: Input-output.

over several decades. These data are experimentally measured weather histories sampled with a period of one hour over one year; due to their poor time-resolution it is likely that these signals are not well shaped to perform a good identification (see [CP04]).

The knowledge of the building's geometric shape and its orientation, allows us to generate the input of the system. These preliminary transformations are non-linear, and because a linear model is sought after, one cannot directly use the measured data but the transformed data to perform the identification. These non-linear transformations are described in [SZ09]. The output is then computed by simulation using a LTI model of order 47 which accounts for the three-dimensions geometry of the building.

This data set the *noise free data*. By contrast, we will call *noisy data* the same data set to which we add noise *independently* on each input and output. The noises on each signal are Gaussian white noises of standard deviation equal to one thirtieth of the standard deviation of the signal. Because the signals are not stationary it represents a quite strong noise on the signals. For instance, this represents a standard deviation of .3°C on a temperature measurement, which is a realistic value for a temperature sensor. This signal/noise ratio is consistent with real application.

In addition, we shall use another data set, which we call *separated output* data set. It is obtained by separating (in simulation) the influence of each input within the internal temperature. This gives much more information on the plant behavior. We shall also allow ourselves to corrupt the data with independent noises; in this case, the data set will be called *noisy separated output*.

D.3 Model classes and parameterization

It is well known (see [LM95]) that the system detailed in Section D.2 can be efficiently represented by a second order linear model. This can be done in several manners, which we now discuss.

D.3.1 Classical ARX model

This is the classical LTI model with rational transfer function. The order here is two. We have restricted our study to strictly proper transfer. This model class, together with the chosen parameterization (see equation (D.1)), has been found to represent the best trade-off between robustness and simulation accuracy in numerical results.

The parameterization is given by (see [Lju87])

$$y[k] + a_1 y[k-1] + a_2 y[k-2] = \sum_{i=1}^{4} b_{i1} u_i[k-1] + b_{i2} u_i[k-2] \tag{D.1}$$

with $i = 1, \ldots, 4$, and where the models parameters are $a_1, a_2, b_{11}, \ldots, b_{14}, b_{21}, \ldots, b_{24}$.

D.3.2 Two Time scale transfer

The difference with the previous model class (see equation (D.1)) is the introduction of a parameter $\varepsilon \ll 1$ which represents the ratio between the "slow" and the "fast" time scales. Specifically, the transfer is expressed as

$$T_\varepsilon(s) = T_f(s) T_s(\frac{s}{\varepsilon}) \tag{D.2}$$

T_s and T_f are slow and fast transfer functions independent of ε. Thermal models are known to be two time scale and, then, can be represented by the equation (D.2) (see [LK85]).

For a given ε, the model class is the same as the ARX; however, it suggests a different parameterization and an adequate handling of each time scale. To do so, the following definition is needed:

Definition 5 *We define the fast transfer $\tau_f(s)$ and the slow transfer $\tau_s(s)$ as follows*

$$\lim_{\varepsilon \to 0} T(i\omega) = T_s(i\infty) T_f(i\omega) \stackrel{def}{=} \tau_f(i\omega) \tag{D.3}$$

$$\lim_{\varepsilon \to 0} T(i\varepsilon\omega) = T_s(i\omega) T_f(i0) \stackrel{def}{=} \tau_s(i\omega) \tag{D.4}$$

Observe that, as ε goes to zero, the slow and the fast transfer keep a similar magnitude if and only if the slow transfer is biproper as defined in [Kai80]. As suggested by Definition 5, T_ε behaves like τ_s in the low frequencies and like τ_f in the high frequencies. For a given ε, we can recover T_ε from τ_s and τ_f if the static gain of the fast transfer is equal to the high frequency gain of the slow transfer.

If some knowledge of a frequency that separates the two parts of T_ε in the frequency domain is available, we can design a low-pass pre-filter F_l and a high-pass pre-filter F_h from which the following model class and parameterization are defined:

Definition 2 *The two time scale model class for the filters F_l and F_h are described in transfer form by*

$$F_l y = \tau_s F_l u \tag{D.5}$$
$$F_h y = \tau_f F_h u \tag{D.6}$$

For a given T_ε the orders of τ_s and τ_f are given by definition 5. These two transfers are parameterized linearly as in the ARX class and are subject to the constraint that

$$|\tau_s(i\infty)| = |\tau_f(0)| \tag{D.7}$$

Several observations can be made

- a suitable change of time scale in the differential operator, as suggested by (D.4), makes (D.6) independent of ε.
- for a finite ε, a system with transfer T_ε does *not* satisfy (D.5,D.6). However, there is a one-to-one correspondence between the parameters of T_ε and the parameters of τ_s and τ_f when (D.7) holds. This is essentially similar to the correspondence of the linear parameterization of ARX models and their gain/poles/zeros description.
- if one uses a classic least square method to identify T_ε, the excitation matrix, i.e. the Hessian of the cost, is asymptotically degenerate as ε tends to zero [CAEA96]. Therefore this method is not robust for small ε.
- it has been proven in [CAEA96] that, if one considers the classical L^2 prediction error as cost for the models (D.5,D.6), then its minimum tends to zero when ε tends to zero if (y, u) satisfy $y = T_\varepsilon u$. Further, the limit excitation matrix is non-degenerate.

In the experiments carried-out on the discussed thermal model, it has been observed that the poles given by the ARX identification provide a good indication of the value of the cutting frequency that should used to design the low and high pass pre-filters (see eq. (D.5) and (D.6)). Figure D.1 shows the amplitude Bode plot of the high order model for the heating control, and its value when multiplied by the low-pass and high-pass pre-filters F_l and F_h, respectively used in the following numerical experiments. The filters are Butterworth filters.

D.4 The parametric identification problems

Here, we define optimization problems to perform the identification of the parameters for each model class. Some emphasis is put on the difference between the separated and non separated output data sets.

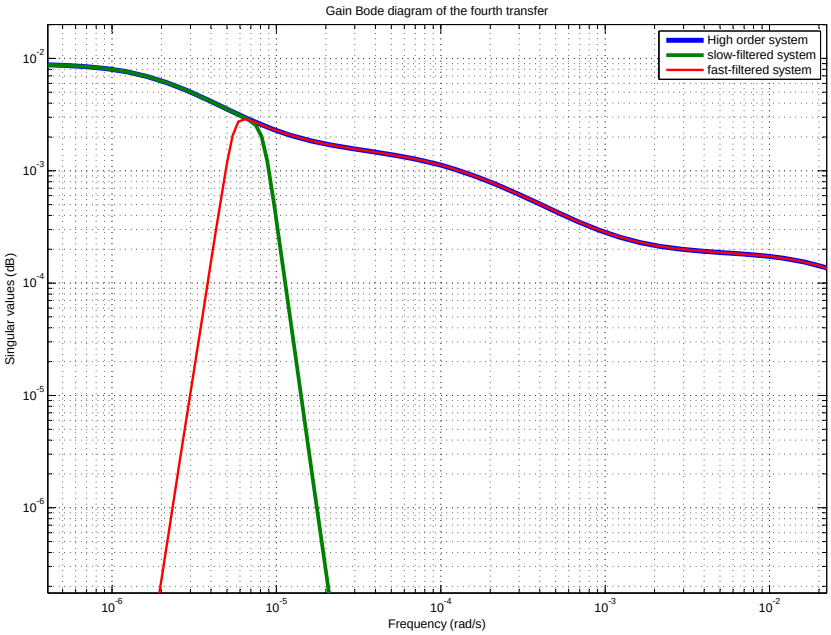

Figure D.1: Global and pre-filtered heating transfers. The right part of the plot, where the slope goes back to -1, is irrelevant to the identification because the data sample rate makes it disappear.

D.4.1 Global ARX model

Using (D.1), we minimize the L^2 norm of the prediction error, as defined by the difference between the two sides of (D.1). In practice, we use MATLAB ARX routine to determine optimal coefficients.

D.4.2 Two time scales identification with a global measurement of the inside temperature

D.4.2.1 Parameterization

The number of poles and zeros of each transfer function has to be set. Since we want a model of order two, we chose a model T_ε with two poles, with one pole in τ_s and one pole in τ_f. The third pole that is visible in Figure D.1 is irrelevant because its time constant is significantly faster than the sampling rate. As in Section D.3.2, the method requires a slow zero. A fast zero could be considered too. This one is visible in Figure D.1. It turns out that, for the data set where the global inside temperature is measured, the best trade-off between robustness and simulation accuracy is achieved by including a fast zero in the fast transfers. Thus the parameterization for the slow and fast models are

$$\begin{aligned} \tau_s(t) &= \frac{1}{s+\alpha}\left(k_1 s + z_1, \quad \cdots, \quad k_4 s + z_4\right) \\ \tau_f(t) &= \frac{1}{\beta s + 1}\left(\rho_1 s + p_1, \quad \cdots, \quad \rho_4 s + p_4\right) \end{aligned} \tag{D.8}$$

D.4.2.2 The identification problem

To perform the identification, we follow the two following steps

- Step 1 : Use of high-pass and low-pass pre-filtering data for which the approximations as given by Definition 5 are as accurate as possible.
- Step 2 : Perform separate identifications of τ_n and τ_s under the constraint that

$$|\tau_{si}(i\infty)| = |\tau_{fi}(i0)| \quad (i = 1 \cdots 4) \tag{D.9}$$

Let $y_s(t)$ and $u_s(t)$ (resp. $y_f(t)$ and $u_f(t) = \begin{pmatrix} u_{s1}(t) & \cdots & u_{s4}(t) \end{pmatrix}$) be the low-pass (resp. high-pass) filtered data, then the corresponding differential equations are given by

$$\frac{d}{dt}y_s(t) + \alpha y_s(t) = \sum_{i=1}^{4} k_i \frac{d}{dt}u_{si}(t) + z_i u_{si}(t) \tag{D.10}$$

$$\beta\frac{d}{dt}y_f(t) + y_f(t) = \sum_{i=1}^{4} \rho_i \frac{d}{dt}u_{fi}(t) + p_i u_{fi}(t) \tag{D.11}$$

Using finite differences we obtain, using usual discrete-time notations,

$$\frac{y_s^{k+1}-y_s^k}{\Delta_s}+\alpha y_s^k = \sum_{i=1}^{4} k_i \frac{u_{si}^{k+1}-u_{si}^k}{\Delta_s}+z_i u_{si}^k \tag{D.12}$$

$$\beta\frac{y_f^{k+1}-y_f^k}{\Delta_f}+y_f^k = \sum_{i=1}^{4} \rho_i \frac{u_{fi}^{k+1}-u_{fi}^k}{\Delta_f}+p_i u_{fi}^k \tag{D.13}$$

where Δ_s and Δ_f are rescaling parameters chosen to improve the conditioning of the problem by adapting the finite difference to the considered time scale (see [MG86]). Note that in (D.12) the sampling rate may be smaller than Δ_s since y_s has been pre-filtered by a low pass filter.

The problem is linear with respect to the parameters so it is convenient to use a least squares method to identify the two transfer matrices. Moreover, this parameterization of the transfer matrix allows to write the constraints linearly with respect to the parameters as shown in (D.14)

$$\nu^T = \nu_s^T - \nu_f^T = \begin{pmatrix} k1-p_1, & \cdots, & k_4-p_4 \end{pmatrix} = 0 \tag{D.14}$$

where the parameters k_i, p_i are appearing in the equations (D.10), (D.11), (D.12) and (D.13).

D.4.2.3 Problem statement

We can now formulate an optimization problem. Given a set of data, the problem is to find the parameters vectors $\theta_s = \begin{pmatrix} k_1 & \cdots & k_4 & z_1 & \cdots & z_4 & \alpha \end{pmatrix}$ and $\theta_f = \begin{pmatrix} p_1 & \cdots & p_4 & \rho_1 & \cdots & \rho_4 & \beta \end{pmatrix}$, corresponding to the parameters from the equations (D.10), (D.11), (D.12) and (D.13), by solving the following problem

$$\begin{array}{ll} \min\limits_{\theta_s,\,\theta_f} & J_s(\theta_s)+J_f(\theta_f) \\ \nu = 0 & \end{array} \tag{D.15}$$

where $J_s(\theta_s)$ (resp. $J_f(\theta_f)$) is the least squares cost of the slow (resp. fast) matrix transfer given by

$$J_s(\theta_s) = \frac{1}{M}\sum_{1}^{M} \hat{\varepsilon}_s^2[k,\theta_s]$$

$$J_f(\theta_f) = \frac{1}{M}\sum_{1}^{M} \hat{\varepsilon}_f^2[k,\theta_f]$$

where

$$\hat{\varepsilon}_s^2[k,\theta_s] = \frac{y_s^{k+1}-y_s^k}{\Delta_s}-\varphi_s[k]\theta_s$$

$$\hat{\varepsilon}_f^2[k,\theta_f] = y_f^k-\varphi_f[k]\theta_f$$

$$\varphi_s[k] = \Big(\frac{u_{s1}^{k+1}-u_{s1}^k}{\Delta_s},\cdots,\frac{u_{s4}^{k+1}-u_{s4}^k}{\Delta_s}\cdots u_{s1}^k,\cdots,u_{s4}^k,-y_s^k\Big)$$

$$\varphi_f[k] = \Big(u_{f1}^k,\cdots,u_{f4}^k,\frac{u_{f1}^{k+1}-u_{f1}^k}{\Delta_f},\cdots,\frac{u_{f4}^{k+1}-u_{f4}^k}{\Delta_f}\cdots -\frac{y_f^{k+1}-y_f^k}{\Delta_f}\Big)$$

In [CAEA96] it has been proved that, if the real transfer is indeed T_ε, the minimum of (D.15) is asymptotically reached (as ε tends to zero) by the parameters corresponding to the slow and fast transfers. Moreover, the Hessians of J_s and J_f are not degenerate when ε tends to zero.

D.4.2.4 Problem solving

While this is not a requirement, we chose to solve problem (D.15) with Uzawa algorithm (see [AHU72]). Its main feature is that, at the minimization stage, each subproblem is very similar to an identification problem on the relevant frequency range, (see [Lju87]), in the sense that the Hessian of the inner optimization problem is a matrix that contains the signals covariance. Moreover, the gradient step of the maximization problem is adapted to each constraint.

D.4.3 Two time scales identification with a separation of the influences of each input

Using a data set which is different from the data set used in the previous section leads to a different tradeoff between accuracy and robustness. Indeed, we have observed that for separated outputs it was best to make some of the fast zeros "vanish" from the parameterization.

D.4.3.1 Parameterization

Even if using the two time scaled method to identify the system allows a clear improvement of the results in terms of simulation errors and parameters identification, as compared to the classical least squares method, an even better identification can be achieved. One explanation is that the system has four inputs and just one output. These inputs are really poorly balanced and some of them do not excite the system in an appropriate frequency range. For instance, the solar fluxes are almost perfectly 24 hours-periodic signals. Therefore, it is difficult to clearly identify the influence of these inputs on the temperature inside the building. That is why in this part we now separate the influence of each input on the temperature. Instead of identifying a transfer matrix we identify four separate transfer functions. This method is referred to as the separated time scaled method.Because we look for a

second order model we have to impose that the four slow (resp. fast) transfer share the same poles.

Observe that, even if the model class may appear similar to the one in Section D.4.2 (once the equality of the poles in the four transfers is duly accounted for), the cost that we will minimize in (D.23) is not the same as in (D.15), because we add four prediction error costs.

In other words, the sum of the excitation matrices of four signals is different from the excitation matrix of the sum of these four signals.

In this case, we have four transfer functions $T_i(s) = T_{si}(s/\varepsilon)T_{f_i}(s)$ $(i = 1 \cdots 4)$. Each transfer can be decomposed into a fast and a slow transfer as mentioned in Definition 5. We now separately identify the four slow (resp. fast) sub-systems in their own time scale under the following constraints :

- for each transfer function the high frequency gain of the slow system must be equal to the static gain of the fast system $(|\tau_{sj}(i\infty)| = |\tau_{fj}(i0)|$ $j = 1 \cdots 4)$
- the fast (resp. slow) sub-systems share the same poles.

To perform the identification, we follow the two following steps

- Step 1 : use of high-pass and low-pass pre-filtering data for which the approximations as given by definition (5) are accurate.
- Step 2 : perform separate identifications of τ_n and τ_s under the constraint that $|\tau_{sj}(i\infty)| = |\tau_{fj}(i0)|$ $(j = 1 \cdots 4)$ and that the transfer functions τ_{sj} (resp. τ_{fj}) share the same poles.

Let $y_{si}(t)$ and $u_{si}(t)$ (resp. $y_{fi}(t)$ and $u_{fi}(t)$) be the low-pass (resp. high-pass) filtered simulations data of the i^{th} transfer function, then the corresponding differential equations are given by a slow subsystem

$$\frac{d}{dt}y_{si}(t) + \alpha_i y_{si}(t) = k_i \frac{d}{dt}u_{si}(t) + z_i u_{si}(t) \quad (i = 1 \cdots 4)$$

and a fast subsystem

$$\begin{aligned} \beta_i \frac{d}{dt}y_{fi}(t) + y_{fi}(t) &= p_i u_{fi}(t) \quad (i = 1 \cdots 3) \\ \beta_4 \frac{d}{dt}y_{f4}(t) + y_{f4}(t) &= \rho_4 \frac{d}{dt}u_{f4}(t) + p_4 u_{f4}(t) \end{aligned} \tag{D.16}$$

This model class has been found to achieve the best trade-off between robustness and simulation accuracy. In particular, deleting the zeros in (D.16) achieves the best trade off between robustness and simulation accuracy.

Using finite differences we have, using the same notations employed in Section D.4.2

$$\frac{y_{si}^{k+1}-y_{si}^{k}}{\Delta_{si}}+\alpha_i y_{si}^{k} = k_i\frac{u_{si}^{k+1}-u_{si}^{k}}{\Delta_{si}}+z_i u_{si}^{k} \quad (i=1\cdots 4) \tag{D.17}$$

$$\beta_i\frac{y_{fi}^{k+1}-y_{fi}^{k}}{\Delta_{fi}}+y_{fi}^{k} = p_i u_{fi}^{k} \quad (i=1\cdots 3) \tag{D.18}$$

$$\beta_4\frac{y_{f4}^{k+1}-y_{f4}^{k}}{\Delta_{f4}}+y_{f4}^{k} = \rho_4\frac{u_{f4}^{k+1}-u_{f4}^{k}}{\Delta_{f4}}+p_4 u_{f4}^{k} \tag{D.19}$$

Once again, the constraints can be expressed linearly with respect to the parameters. Actually, the constraints of the identification problem are :

$$\alpha_i-\alpha_{i+1}=0, \quad i=1\cdots 3 \tag{D.20}$$

$$\beta_i-\beta_{i+1}=0, \quad i=1\cdots 3 \tag{D.21}$$

$$k_i-p_i=0, \quad i=1\cdots 4 \tag{D.22}$$

Thus, the vector of constraints $\nu=\nu_s-\nu_f$ is given by the concatenation of the ten equalities given by (D.20), (D.21) and (D.22).

D.4.3.2 Problem statement

Given a set of data, the problem is to find the four parameters vectors $\theta_{si}=\begin{pmatrix}k_i & z_i & \alpha_i\end{pmatrix}^T$, the three $\theta_{fi}=\begin{pmatrix}p_i & \beta_i\end{pmatrix}^T$ $(i=1\cdots 3)$ and $\theta_{f4}=\begin{pmatrix}p_4 & \rho_4 & \beta_4\end{pmatrix}^T$ by solving the following problem

$$\min_{\substack{\theta_{si},\theta_{fi}\\ \nu=0}} \sum_{i=1}^{4} J_{si}(\theta_{si})+J_{fi}(\theta_{fi}) \tag{D.23}$$

where $J_{si}(\theta_{si})$ (resp. $J_{fi}(\theta_{fi})$) is the least squares cost of the i^{th} slow (resp. fast) transfer function.

$$J_{si}(\theta_{si}) = \frac{1}{M}\sum_{1}^{M}\hat{\varepsilon}_{si}^{2}[k,\theta_{si}]$$

$$J_{fi}(\theta_{fi}) = \frac{1}{M}\sum_{1}^{M}\hat{\varepsilon}_{fi}^{2}[k,\theta_{fi}]$$

where

$$\begin{aligned}
\hat{\varepsilon}^2_{si}[k,\theta_{si}] &= \frac{y^{k+1}_{si}-y^k_{si}}{\Delta_{si}} - \varphi_{si}[k]\theta_{si} \\
\hat{\varepsilon}^2_{fi}[k,\theta_{fi}] &= y^k_{fi} - \varphi_{fi}[k]\theta_{fi} \\
\varphi_{si}[k] &= \left(\frac{u^{k+1}_{si}-u^k_{si}}{\Delta_{si}} \quad u^k_{si} \quad -y^k_{si}\right) \\
\varphi_{fi}[k] &= \left(u^k_{fi} \quad -\frac{y^{k+1}_{fi}-y^k_{fi}}{\Delta_{fi}}\right) (i = 1\cdots 3) \\
\varphi_{f4}[k] &= \left(u^k_{f4} \quad \frac{u^{k+1}_{f4}-u^k_{f4}}{\Delta_{f4}} \quad -\frac{y^{k+1}_{f4}-y^k_{f4}}{\Delta_{f4}}\right)
\end{aligned}$$

D.4.3.3 Problem solving

Here again, we use Uzawa algorithm to solve this problem.

D.5 Numerical results

D.5.1 Conditioning of the problems

To perform a robust parameter identification, the Hessian of the optimization problem has to be well conditioned (see [Lju87]). Yet, a two-time scaled system usually induces bad conditioning (see [CAEA96]). The time scaled identification has been designed to improve the conditioning of the optimization problem. For the classical least squares method there is one conditioning number, while there are two conditioning numbers for the time-scaled method (one for the slow transfer matrix and one for the fast one), and there are eight conditioning numbers for the separated time scaled method (one for each subsystem). The conditioning numbers are given in the Table D.2. We use everywhere data without noise corruption resulting from the high order model (see section D.2).

	least squares identification	Time scaled identification	Separated time scaled
conditioning numbers	$r_{LS} = 2.6/10^{10}$	$r_s = 1.4/10^8$ $r_f = 1.4/10^9$	$r_{s1} = 0.0011$ $r_{s2} = 0.00083$ $r_{s3} = 0.00078$ $r_{s4} = 0.0016$ $r_{f1} = 0.043$ $r_{f2} = 0.013$ $r_{f3} = 0.037$ $r_{f4} = 0.020$

Table D.2: Conditioning numbers.

As one can see it on Table D.2, the separated time scaled method improves the conditioning of the problem. But, we can also see that using a non separated time

scaled method does not improve the conditioning numbers as well as the previous method. Having separated outputs provides extra information on the system as we have virtually three extra sensors.

The bad conditioning of the least squares method (ARX) is highly problematic because the results of the identification are very poor, in simulation results and in parameter identification. Concerning the non separated time scaled method, in the following, we will see that, despite the bad conditioning of the system, this method yields better results in simulation and in parameter identification than the least squares method. On the other side, it will be seen that this method fails to estimate the location of the zeros of the system, particularly when the identification is performed using noisy data.

Finally, we can see that the separation of the transfers allows us to normalize the problem and then to improve the conditioning numbers. As a result, this method is really robust with respect to noises and consistent results[2] in parameter identification are obtained wether noisy or noise free data are used.

D.5.2 Simulation results

D.5.2.1 Simulation protocol

This protocol is decomposed in four steps:

1. Using noise free input data described in Section D.2 and using the high order model, we get the four noise free corresponding outputs.

2. Then, we perform a first identification using the previous data.

3. Further, we add independent noises on the inputs and the outputs collected from the first step. Then, we perform three identifications using this noisy data and the three identification methods.

4. Finally a validation step is performed. We simulate all the models from 2 and 3 using noise free inputs to obtain the global temperature of the building. We compare these temperature to the global output of step 1. The Table D.3 gives some statistical properties of the simulation error between the global temperature from the high order model and the global temperature of each of the six identified models.

D.5.2.2 Results

Figure D.2 shows the errors of simulation between the high order reference model and the three identified systems, the latter being identified using noisy data. These simulations are performed using noise free inputs over 25 days. As can be seen in Figure D.2 the ARX model identified using MATLAB's identification toolbox does not give good results in terms of simulation errors. Moreover, one can see in Figure

[2]This comparison is made with noises which have the same statistical properties.

D.2 that the standard deviation of the simulation error seems to be better with the separation of the influences of the inputs than without.

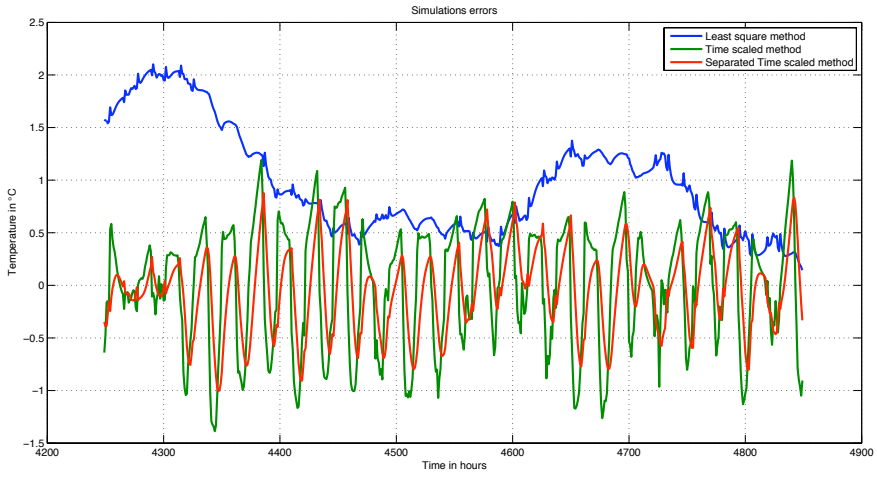

Figure D.2: Comparison of simulations errors using noise free data obtained with the three identified models which have been identified using noisy data.

	Stat. prop.	Least squares identification	Time scaled identification	Separated time scaled
Noise free data	Mean	−0.0989	−0.0024	−0.0088
	std dev	0.93	0.35	0.266
Noisy data	Mean	0.0461	0.0021	−0.0045
	std dev	0.641	0.49	0.34

Table D.3: Comparison of statistical properties of the simulation error with respect to the noise free simulation using the high order model.

Using noise free data The Table D.3 shows some statistical properties of the simulation errors of the three identified systems. Considering the identification using noise free data, one can see that the best results are obtained by the time-scaled methods. Indeed, the statistical properties of the simulation error obtained with the time scaled methods are similar. One can also notice that the worst results are clearly obtained with the ARX model.

Using noisy data Let us focus on the identification using noisy data. one can see that the best results are again achieved with the time-scaled methods. One can also observe that the deterioration of the standard deviation is less important

when we separate the influences of the input than with a global measurement of the temperature. We can also notice that even if the ARX model is still the worst, the addition of noises has clearly improved its performances.

D.5.2.3 Conclusion

One can see that using a time scaled method allows a good improvement of the results in terms of simulation error. The comparison between the least squares method and the time scaled method shows that the results are better using the latter.

D.5.3 Static gains, poles and zeros identification

Concerning poles and zeros, we give, in Tables D.6 ,D.7 and D.8 the corresponding time constants. Those time constants are calculated using the discrete model provided by equations (D.1), (D.12), (D.13), (D.17), (D.18) and (D.19)

D.5.3.1 Static gains identification

Let us see the results of the three identification on the static gains

	High order model	least squares identification	Time scaled identification	Separated time scaled
Gain 1	1	0.903	1.0013	1.0004
Gain 2	0.0088	0.0058	0.0089	0.0088
Gain 3	$6.75 \cdot 10^{-5}$	$5.68 \cdot 10^{-4}$	$5.48 \cdot 10^{-5}$	$6.75 \cdot 10^{-5}$
Gain 4	0.009	0.0125	0.009	0.009

Table D.4: Comparison of the identified static gain using noise free data.

	High order model	least squares identification	Time scaled identification	Separated time scaled
Gain 1	1	0.959	1.0008	1.0007
Gain 2	0.0088	0.006	0.0088	0.0088
Gain 3	$6.75 \cdot 10^{-5}$	$5 \cdot 10^{-4}$	$6.73 \cdot 10^{-5}$	$6.75 \cdot 10^{-5}$
Gain 4	0.009	0.0107	0.009	0.009

Table D.5: Comparison of the identified static gain using noisy data.

As one can see on the Table D.4 and Table D.5, using a time-scaled identification method yields a substantial improvement compared to the classical least squares method. In fact, the classical least squares method never correctly estimates the static gains whereas the time-scaled methods estimate the gains of the transfer matrix adequately.

Moreover, the separation of the influences of each input allows one to reach the same accuracy using noisy or noise free data, whereas the others methods give better results using noisy data.

D.5.3.2 Time constants and zeros location

Time constants identification. Table D.6 reports the identified time constants using noisy and noise free data. We can see that with and without noises the identification of the two time constants of the system are similar using the time-scaled methods, whereas the ARX model provides time constants quite different of the other models. Since the simulation results are better with the models identified by time scaled methods, one can suppose that the time constants are well identified by these methods.

		least squares identification	Time scaled identification	Separated time scaled
Noise free data	Slow	117	143	147
	Fast	1.1	2.5	2.9
noisy data	Slow	108	144	147
	Fast	1.2	2.4	2

Table D.6: Identified time constant in hours.

Zeros time constants Table D.7 and Table D.8 give the zeros time constants of each transfer using respectively noise free and noisy data to perform the identification. As one can see on these tables, the only method yielding a weak dispersion of the identified parameters is the time scaled identification with separation of the influences of the inputs. Moreover, using Moore's method to reduce the system does not keep the two time scaled structure of the system. Indeed, looking at the Bode diagram of the reduced system, one can notice that the two time scaled structure exhibited by both the high order model and the identified one is not preserved by the reduced one.

D.5.3.3 Conclusion

The least squares method does not really identify static gains, time constants and zeros of the system, the time scaled method with global measurement allows us to identify the static gains and the time constants, but shows poor results in the identification of the zeros. Finally, the time-scaled method with separation of each inputs allows to identify all these parameters with robustness to noises.

		least squares identification	Time scaled identification	Separated time scaled
First Transfer	Slow zero	7.8	14.8	24
	Fast zero	None	4.6	None
Second Transfer	Slow zero	2.1	9.9	5.1
	Fast zero	None	0.85	None
Third Transfer	Slow zero	1.65	197†	11.1
	Fast zero	None	0.44	None
Fourth Transfer	Slow zero	9.99	21.3	24.9
	Fast zero	None	0.051†	0.76

Table D.7: Identified zeros location using noise free data. The † symbol means that the zero has been found to be unstable.

		least squares identification	Time scaled identification	Separated time scaled
First Transfer	Slow zero	5.3	15.5	23.2
	Fast zero	None	4.4	None
Second Transfer	Slow zero	2.8	9.9	5.2
	Fast zero	None	1.1	None
Third Transfer	Slow zero	1.9	116†	11.6
	Fast zero	None	0.023†	None
Fourth Transfer	Slow zero	11.2	22.6	24.3
	Fast zero	None	0.019†	0.44

Table D.8: Identified zeros location using noisy data. The † symbol means that the zero has been found to be unstable.

D.6 Conclusion

In this Appendix, it was shown that using a time-scaled identification method (see [CAEA96]) allows a substantial improvement of the model identification compared to a classical least squares method, in terms of prediction error and of parameters sensitivity to measurement noises. It was also emphasized that to clearly identify a time-scaled system it is needed to find a good compromise between simulation error and robustness to noise. The time-scaled method allows a great improvement in the search of this compromise compared to the least squares method.

Nevertheless, it has been observed that the identification of the zeros of the system is too sensitive to the noises using that method. To improve the conditioning of the system the inputs and the output should be normalized. Using a separation of the influences of each input is a solution. Then, it has been shown that separating the influences of the inputs and using a time scaled method can provide a good compromise between identification error and robustness toward noises since the results obtained with or without noises are quite similar.

In summary, this work proposes an efficient method, based on a two time scale models to identify a low order linear model describing the thermal behavior of the system. This efficiency is measured in terms of simulation errors and in terms of robustness of the parameters identification to noises. This is due to the normalization of the two time scale problems, both in magnitude of the signals and in their frequency range. The model obtained by this method can be used in simulation and it can also be used in constrained optimal control since the parameters are well identified.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

Pilotage dynamique de l'énergie du bâtiment par commande optimale sous contraintes utilisant la pénalisation intérieure.

Résumé : Dans cette thèse, nous proposons une méthode de résolution de problèmes de commande optimale non linéaires sous contraintes d'état et de commande. Cette méthode repose sur l'adaptation des méthodes de points intérieurs, utilisées en optimisation de dimension finie, à la commande optimale. Un choix constructif de fonctions de pénalisation intérieure est fourni. On montre que ce choix permet d'approcher la solution d'un problème de commande optimale sous contraintes en résolvant une suite de problèmes de commande optimale sans contraintes dont les solutions sont simplement caractérisées par les conditions de stationnarité du calcul des variations. Deux études dans le domaine de la gestion de l'énergie dans les bâtiments sont ensuite conduites. La première consiste à quantifier la durée maximale d'effacement quotidien du chauffage permettant de maintenir la température intérieure dans une certaine bande de confort, et ce pour différents types de bâtiments classés de mal à bien isolés. La seconde étude se concentre sur les bâtiments basse consommation (BBC) et consiste à quantifier la capacité de ces bâtiments à réaliser des effacements électriques complets du chauffage de 6h00 à 22h00.
Mots clés : Commande optimale, points intérieurs, contraintes d'état et de commande, effacements électriques, bâtiments BBC, isolation extérieure-intérieure.

Dynamic control of energy in buildings using constrained optimal control by interior penalty

Abstract: This thesis exposes a methodology to solve state and input constrained optimal control of non-linear systems by interior penalty methods. A constructive choice for the penalty functions used to implement the interior method is exhibited. It is shown that the methodology allows one to approach the solution of the non-linear optimal control problem using a sequence of unconstrained problems, whose solutions are readily characterized by the simple calculus of variations. Two representatives study of energy management in buildings are conducted using the provided algorithm. The first study consists in quantifying the maximal duration of daily complete load shiftings achievable by several buildings ranging from poorly to well insulated. The second study focuses on low consumption buildings and aim at quantifying the ability of these buildings to perform complete load shiftings of the heating consumption from the day (6 a.m. to 10 p.m.) to the night period.
Keywords: Optimal control, interior point methods, state and input constraints, load shiftings, interior/exterior insulation, low consumption buildings.

THÈSE