Interior penalty approximation for optimal control problems. Optimality conditions in stochastic optimal control theory.

Francisco J. Silva

THÈSE DE DOCTORAT

présentée par

Francisco J. Silva

pour obtenir le grade de Docteur de l'Ecole Polytechnique

Spécialité: Mathématiques Appliquées

Interior penalty approximation for optimal control problems. Optimality conditions in stochastic optimal control theory.

Thèse présentée le 29 novembre 2010 devant le jury composé de :

Frédéric Bonnans	Directeur de thèse
Jean-Pierre Raymond	Rapporteur
Agnès Sulem	Examinatrice
Nizar Touzi	Examinateur
Stefan Ulbrich	Rapporteur
Jiongmin Yong	Rapporteur

A mi padre.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Frédéric Bonnans pour m'avoir accueilli dans son equipe COMMANDS de L'INRIA Saclay. Je lui en suis très reconnaissant pour la grande qualité de son encadrement durant toutes ces années. Il a toujours eté disponible pour répondre à mes questions et il m'a toujours encouragé à assister à de nombreuses conférences, écoles et séminaires. Je lui exprime mon respect le plus profond.

J'exprime également toute ma reconnaissance à Jérôme Bolte, avec qui j'ai eu l'opportunité de travailler au début de ma thèse. Je lui remercie son soutien constant tout au long de mon travail et j'avoue que travailler avec lui m'a permi d'enrichir considérablement mes connaissances en mathématiques.

Je remercie aussi très vivement Felipe Alvarez, qui m'a encouragé à postuler à ce projet. Lors des mes années de thèse j'ai eu la chance de faire des séjours dans mon cher pays le Chili. Je tiens à remercier Felipe Álvarez et Alejandro Jofré pour leur agréable accueil à l'Université du Chili.

Je souhaite exprimer toute ma gratitude à Jean-Pierre Raymond, Stefan Ulbrich et Jiongmin Yong pour avoir accepté la tâche de rapporteur. J'adresse un grand merci pour leur lecture très soigneuse de ce manuscrit et leur remarques très intéressant.

Je suis très reconnaissant à Agnès Sulem et Nizar Touzi d'avoir accepté de faire partie de mon jury. Leur travaux en commande optimale stochastique sont des références incontournables.

Mes remerciement vont également à tous le membres du CMAP. J'ai pu bénéficier d'une ambiance de travail très enrichissante à tout point de vue. Je voudrais exprimer ma gratitude à tout particulièrement à Wallis Fillipi pour sa grande disponibilité et son aide permanente pendant toutes ces années.

Finalement, je remercie aussi tous mes "camarades" du laboratoire qui ont partagé avec moi leur vie quotidienne. Un grand merci notamment aux doctorants de mon bureau pour l'atmosphère très amicale: María Soledad Aronna, Florent Barret, Zhihao Cen, Camille Coron, Khalil Dayri, Sylvie Detournay, Laurent Duvernet, Clément Fabre et Émilie Fabre.

Résumé

Cette thèse est divisée en deux parties. Dans la première partie on s'intéresse aux problèmes de commande optimale *déterministes* et on étudie des approximations intérieures pour deux problèmes modèles avec des contraintes de non-négativité sur la commande. Le premier modèle est un problème de commande optimale dont la fonction de coût est quadratique et dont la dynamique est régie par une équation différentielle ordinaire. Pour une classe générale de fonctions de pénalité intérieure, on montre comment calculer le terme principal du développement ponctuel de l'état et de l'état adjoint. Notre argument principal se fonde sur le fait suivant: si la commande optimale pour le problème initial satisfait les conditions de complémentarité stricte pour le Hamiltonien sauf en un nombre fini d'instants, les estimations pour le problème de commande optimale pénalisé peuvent être obtenues à partir des estimations pour un problème stationnaire associé. Nos résultats fournissent plusieurs types de mesures de qualité de l'approximation pour la technique de pénalisation: estimations des erreurs de la commande pour les normes L^s (s dans $[1, +\infty]$), estimations des erreurs pour l'état et l'état adjoint dans les espaces de Sobolev $W^{1,s}$ (s dans $[1, +\infty)$) et aussi estimations de erreurs pour la fonction valeur. Pour la norme L^1 et la pénalisation logarithmique, les résultats optimaux sont donnés. Dans ce cas-là on obtient des erreurs pour la trajectoire centrale du problème pénalisé de l'ordre $O(\varepsilon|\log\varepsilon|)$.

Le second modèle est le problème de commande optimale d'une équation semi-linéaire elliptique avec conditions de Dirichlet homogène au bord, la commande étant distribuée sur le domaine et positive. L'approche est la même que pour le premier modèle, c'est-à-dire que l'on considère une famille de problèmes pénalisés par $\varepsilon > 0$, dont la solution définit une trajectoire centrale qui converge vers la solution du problème initial. De cette manière, on peut étendre les résultats, obtenus dans le cadre d'équations différentielles, au contrôle optimal d'équations elliptiques semi-linéaires.

Dans la deuxième partie on s'intéresse aux problèmes de commande optimale *stochastiques*. Dans un premier temps, on considère un problème linéaire quadratique stochastique avec des contraintes de non-negativité sur la commande et on étend les estimations d'erreur pour l'approximation par pénalisation logarithmique. La preuve s'appuie sur le principe de Pontriaguine stochastique et un argument de dualité.

Ensuite, on considère un problème de commande stochastique général avec des contraintes convexes sur la commande. L'approche dite *variationnelle* nous permet d'obtenir un développement au premier et au second ordre pour l'état et la fonction de coût, autour d'un minimum local. Avec

ces développements on peut montrer des conditions genérales d'optimalité de premier ordre et, sous une hypothèse géométrique sur l'ensemble des contraintes, des conditions nécessaires du second ordre sont aussi établies.

Abstract

This thesis is divided in two parts. In the first one we consider *deterministic* optimal control problems and we study interior approximations for two model problems with non-negativity constraints. The first model is a quadratic optimal control problem governed by a nonautonomous affine ordinary differential equation. We provide a first-order expansion for the penalized state an adjoint state (around the corresponding state and adjoint state of the original problem), for a general class of penalty functions. Our main argument relies on the following fact: if the optimal control satisfies strict complementarity conditions for its Hamiltonian, except for a set of times with null Lebesgue measure, the functional estimates of the penalized optimal control problem can be derived from the estimates of a related finite dimensional problem. Our results provide three types of measure to analyze the penalization technique: error estimates of the control for L^s norms (s in $[1, +\infty]$), error estimates of the state and the adjoint state in Sobolev spaces $W^{1,s}$ (s in $[1, +\infty)$) and also error estimates for the value function. The sharpest results are given for the L^1 norm and a logarithmic penalty, establishing an error estimate for the central path of order $O(\varepsilon|\log\varepsilon|)$ where $\varepsilon > 0$ is the (small) penalty parameter.

The second model we study is the optimal control problem of a semilinear elliptic PDE with a Dirichlet boundary condition, where the control variable is distributed over the domain and is constrained to be non-negative. Following the same approach as in the first model, we consider an associated family of penalized problems, parametrized by $\varepsilon > 0$, whose solutions define a central path converging to the solution of the original one. In this fashion, we are able to extend the results obtained in the ODE framework to the case of semilinear elliptic PDE constraints.

In the second part of the thesis we consider *stochastic* optimal control problems. We begin with the study of a stochastic linear quadratic problem with non-negativity control constraints and we extend the error estimates for the approximation by logarithmic penalization. The proof is based is the stochastic Pontryagin's principle and a duality argument.

Next, we deal with a general stochastic optimal control problem with convex control constraints. Using the *variational* approach, we are able to obtain first and second-order expansions for the state and cost function, around a local minimum. This analysis allows us to prove general first order necessary condition and, under a geometrical assumption over the constraint set, second-order necessary conditions are also established.

Contents

Part I

General introduction

In this part of the thesis we review some elementary concepts of both deterministic and stochastic optimal control problems with control constraints. After giving the necessary elements of the theory we will expose the main results obtained. Let us start with the study of deterministic optimal control problems.

0.1 Deterministic optimal control

An optimal control problem of ordinary differential equations (ODE) with control constraints can be written in the following form:

$$\begin{array}{ll} \inf_{(y,u)\in\mathcal{V}\times\mathcal{Y}} \int_0^T \ell(t,y(t),u(t))\mathrm{d}t + \phi(T,y(T)) & \\ \text{s.t.} \qquad \dot{y}(t) = f(t,y(t),u(t)) \ \text{ for } t\in[0,T]; \quad y(0)=y_0, & (\mathcal{DCP})_0 \\ \qquad\quad u\in\mathcal{U}. & \end{array}$$

In the notation above, $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ represents the running cost, $\phi : \mathbb{R}^n \to \mathbb{R}$ the final cost and $y(t) \in \mathbb{R}^n$ represents the state variable controlled by $u(t) \in \mathbb{R}^m$ through the dynamics $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$. If f is affine with respect to u we may take as control space $\mathcal{V} = L^2([0,T];\mathbb{R}^m)$ and as state space $\mathcal{Y} = W^{1,2}([0,T];\mathbb{R}^n)$. Otherwise, we take $\mathcal{V} = L^\infty([0,T];\mathbb{R}^m)$ and $\mathcal{Y} = W^{1,\infty}([0,T];\mathbb{R}^n)$. The control variable is constrained to belong to $\mathcal{U} \subseteq \mathcal{V}$. Note that this framework includes global constraints, e.g. $\mathcal{U} := \{v \in \mathcal{V} \ / \ ||v||_2 \leq 1\}$, as well as local constraints, e.g. the so-called box constraints $\mathcal{U} := \{v \in \mathcal{V} \ / \ a \leq v(t) \leq b, \text{ for a.a. } t \in [0,T]\}$. In the first part of this thesis we will focus our attention to the case of box constraints. In order to simplify the analysis, we will restrict ourselves to the case of non negativity constraints. In the second part of the thesis we will determine first and second-order optimality condition for the stochastic version of $(\mathcal{DCP})_0$ and we will work with a more general constraint set $\mathcal{U}$.

Thus, in what follows we assume that

$$\mathcal{U} := \{v \in \mathcal{V} \ / \ v(t) \geq 0, \text{ for a.a. } t \in [0,T]\}\,. \tag{1}$$

Since, the active set (i.e. the set of times where the optimal control is 0) is a priori not known, numerical difficulties appear in the implementation of any direct algorithm. One way to tackle this problem is to extend the natural ideas of interior-point methods for nonlinear programming problems. More precisely, we consider a family of perturbed optimal control problems satisfying that their solutions are strictly positive (and thus they can be computed efficiently), and we expect to obtain some good convergence properties for

the procedure. As an example, for the logarithmic-penalty case, a natural approximation of $(\mathcal{DCP})_0$ is the following problem

$$\begin{array}{ll} \inf_{(y,u)\in\mathcal{V}\times\mathcal{Y}} \int_0^T [\ell(y(t),u(t)) - \varepsilon \log u(t)]\,\mathrm{d}t + \phi(y(T)) & \\ \text{s.t.} \quad \dot{y}(t) = f(y(t),u(t)) \;\text{ for } t\in[0,T]; \;\; y(0)=y_0, & (\mathcal{DCP})_\varepsilon \\ \quad u\in\mathcal{U}. & \end{array}$$

The convergence of the solutions of $(\mathcal{DCP})_\varepsilon$ to the solution of $(\mathcal{DCP})_0$, as $\varepsilon \downarrow 0$, is shown in [22], but no error estimates are obtained. As we will see, these estimates can be obtained as a by-product of the qualitative properties of the central path (defined in section 0.1.2.1), which are strongly related to their finite-dimensional counterparts, recalled in the next section.

0.1.1 A brief review of interior point methods for quadratic programming

Consider the following finite dimensional optimization problem

$$\mathrm{Min}_{x\in\mathbb{R}^n} \;\; \tfrac{1}{2}x^\top Rx + c^\top x; \;\; Ax = b,\; x\geq 0, \qquad (\mathcal{QP})_0$$

where $R\in\mathbb{R}^{n\times n}$ is a positive-semidefinite matrix, $c\in\mathbb{R}^n$ and $b\in\mathbb{R}^p$. We say that the problem is *linear* if $R=0$. If $(\mathcal{QP})_0$ has at least one solution x_0, there there exists $(s_0,\lambda_0)\in\mathbb{R}^n_+\times\mathbb{R}^p$ such that $z_0 := (x_0,s_0,\lambda_0)$ solves

$$\left\{\begin{array}{l} x^\top s = 0, \\ Ax = b, \;\; c + Rx + A^\top\lambda = s, \\ x\geq 0, \quad s\geq 0. \end{array}\right. \qquad (2)$$

In view of this property, from now on we refer to z_0 as a solution of $(\mathcal{QP})_0$.

Now, consider a parameterized family of problems that penalize the non negativity constraint of $(\mathcal{QP})_0$. That is, for every $\varepsilon>0$, define the problem $(\mathcal{QP})_\varepsilon$ as

$$\mathrm{Min}_{x\in\mathbb{R}^n} \;\; \tfrac{1}{2}x^\top Rx + c^\top x - \varepsilon\sum_{i=1}^p \log x_i; \;\; Ax = b. \qquad (\mathcal{QP})_\varepsilon.$$

It is possible to prove that if $(\mathcal{QP})_0$ has a solution x_0 then, for ε small enough, problem $(\mathcal{QP})_\varepsilon$ has a solution x_ε. Moreover, there exists $(s_\varepsilon,\lambda_\varepsilon)\in\mathbb{R}^n_+\times\mathbb{R}^p$ such that $z_\varepsilon := (x_\varepsilon,s_\varepsilon,\lambda_\varepsilon)$ solves

$$\left\{\begin{array}{l} x^\top s = \varepsilon, \\ Ax = b, \;\; c + Rx + A^\top\lambda = s, \\ x\geq 0, \quad s\geq 0. \end{array}\right. \qquad (3)$$

Thus, we refer to z_ε as a solution of $(\mathcal{QP})_\varepsilon$. The application $\varepsilon \to z_\varepsilon$ is called *central path* and it is well known that as $\varepsilon \downarrow 0$, we have $z_\varepsilon \to z_0$. Moreover, qualitative properties of the central path (error estimates of its slope) are related with the following notion of strict complementarity:

Definition 1 *We say that the solution z_0 of $(\mathcal{QP})_0$ is strictly complementary if $x_0 + s_0 > 0$.*

In the linear case ($R = 0$), if the set of solutions of $(\mathcal{QP}_0)$ is nonempty, there exists at least one strictly complementary solution and the central path converges to one solution of this kind (see [82]). In the strictly convex quadratic case ($R \succ 0$), the problem $(\mathcal{QP})_0$ has a unique solution z_0 and $z_\varepsilon \to z_0$. In addition, if z_0 is strictly complementary, then $||z_\varepsilon - z_0|| = O(\varepsilon)$. If strict complementarity does not hold, $||z_\varepsilon - z_0|| = O(\sqrt{\varepsilon})$ - see [92]. Let us give a trivial example where we see the importance of strict complementarity for the speed of convergence of the central path.

Example 1 *Consider the problem*

$$Min_{x\in\mathbb{R}} \ \tfrac{1}{2}x^2; \quad x \geq 0,$$

which has as unique solution $x_0 = 0$. The penalized version of the above problem is

$$Min_{x\in\mathbb{R}} \ \tfrac{1}{2}x^2 - \varepsilon \log x,$$

which has as unique solution $x_\varepsilon = \sqrt{\varepsilon}$, and thus $|x_\varepsilon - x_0| = \sqrt{\varepsilon}$. One can easily verify that x_0 is not strictly complementary. On the other hand, the problem

$$Min_{x\in\mathbb{R}} \ \tfrac{1}{2}x^2; \quad x \geq 1,$$

has a unique solution $x_0 = 1$. In this case strict complementarity is satisfied and a simple computation shows that the solution x_ε of the penalized problem satisfies $|x_\varepsilon - x_0| = O(\varepsilon)$.

These properties of the central path allow us to justify theoretically the use of several types of interior point algorithms for $(\mathcal{QP})_0$. For example, for a fixed ε the penalized problem can be solved by applying Newton's method. Then, ε is decreased and the mentioned method is re-initialized taking as the starting point the approximate solution of the previous problem. Thus, *a priori* this point must belong to the convergence region of the new Newton's algorithm. There are several variations of this general principle, for detailed expositions and complexity analysis we refer the reader to the books [21, 82, 91] and

references therein. Finally, let us mention that these methods are studied for more general settings, as general convex problems with self concordant barrier functions [74], linear monotone complementarity problems [21] and semidefinite programming [82], etc.

0.1.2 Presentation of our main results

In this section we apply the barrier-method ideas to the optimal control of an ordinary differential equation (ODE) and to the optimal control of a semilinear elliptic partial differential equation (PDE). In both cases a parameterized family of penalized problems is considered, for which optimality conditions are derived. The main idea is to eliminate the control variable from the resulting equations and to apply a variation of the implicit function theorem to the reduced optimality system.

The main tool will be the following theorem and its corollary, which is a variant of the surjective mapping theorem of Graves [49].

Theorem 2 (Restoration Theorem) *Let X and Y be Banach spaces, E a metric space and $F : U \subset X \times E \to Y$ a continuous mapping on a nonempty open set U. Let $(\hat{x}, \varepsilon_0) \in U$ be such that $F(\hat{x}, \varepsilon_0) = 0$. Assume that there exists a surjective linear continuous mapping $A : X \to Y$, with bounded right inverse B, and a function $c : \mathbb{R}_+ \to \mathbb{R}_+$ with $c(\beta') \downarrow 0$ when $\beta' \downarrow 0$, such that: if $\beta > 0$ satisfies $c(\beta)||B|| < 1$ and $\varepsilon \in B(\varepsilon_0, \beta)$, then*

$$\|F(x', \varepsilon) - F(x, \varepsilon) - A(x' - x)\| \leq c(\beta)\|x' - x\|, \text{ for all } (x, x') \in \overline{B}(\hat{x}, \beta) \times \overline{B}(\hat{x}, \beta). \tag{4}$$

Under the assumptions above, for all (x, ε) close enough to $(\hat{x}, \varepsilon_0)$, there exists $\bar{x}$ such that $F(\bar{x}, \varepsilon) = 0$ and the following estimate holds:

$$\|\bar{x} - x\| \leq \frac{||B||}{1 - c(\beta)||B||}\|F(x, \varepsilon)\|. \tag{5}$$

Corollary 3 *Suppose that the assumptions of Theorem 2 hold and denote by B a bounded right inverse of A. Then, for ε close to ε_0, there exists x_ε in a neighborhood of $\hat{x}$ such that $F(x_\varepsilon, \varepsilon) = 0$ and*

$$x_\varepsilon = \hat{x} - BF(\hat{x}, \varepsilon) + r(\varepsilon), \tag{6}$$

where the remainder $r(\varepsilon)$ satisfies

$$\|r(\varepsilon)\| \leq c(\beta)(1 - c(\beta)\|B\|)^{-1}\|B\|^2\|F(\hat{x}, \varepsilon)\|. \tag{7}$$

For the proof of the above results, we refer the reader to the appendix of Chapter 1.

0.1.2.1 Optimal control of ODEs

In this section we present the main results obtained in Chapter 1, which are the subject of report [2]. For the sake of clarity, we study a simplified version of the general linear quadratic problem analyzed in Chapter 1. We consider the problem $(\mathcal{DCP})_0$ with

$$\begin{array}{l} \ell(t,y,u) := \frac{1}{2}|u|^2 + \frac{1}{2}\, C(t)\,|y-\bar{y}(t)|^2\,, \\ \phi(T,y) \quad := \frac{1}{2} M\,|y-\bar{y}(T)|^2\,, \\ f(t,y,u) := A(t)y + u, \end{array} \tag{8}$$

and $\mathcal{U}$ given by (1) with $\mathcal{V} = L^2([0,T];\mathbb{R})$. In the notation above, $C \in \mathcal{C}^0([0,T])$ with $C(t) \geq 0$, $M \geq 0$, $A \in \mathcal{C}^0([0,T])$ and $\bar{y} \in \mathcal{C}^0([0,T])$ is a reference state function.

For every $\varepsilon > 0$ define $(\mathcal{DCP})_\varepsilon$, the logarithmic penalized version of $(\mathcal{DCP})_0$, by

$$\begin{array}{ll} \inf_{(y,u)\in\mathcal{Y}\times\mathcal{V}} \int_0^T \ell_\varepsilon(t,y(t),u(t))\mathrm{d}t + \phi(y(T)) & \\ \text{s.t.} \qquad \dot{y}(t) = f(t,y(t),u(t)) \;\; \text{for } t\in[0,T]; \;\; y(0) = y_0, & (\mathcal{DCP})_\varepsilon \\ \qquad\quad u \in \mathcal{U}, & \end{array}$$

where $\ell_\varepsilon(t,y,u) := \ell(t,y,u) - \varepsilon \log u$. For notational convenience we also set $\ell_0(t,y,u) = \ell(t,y,u)$. Classical arguments yield that for every $\varepsilon \in [0,\infty)$ problem $(\mathcal{DCP})_\varepsilon$ has a unique solution, denoted by $(y_\varepsilon, u_\varepsilon)$. Moreover, it can be shown [22] that there exists $c > 0$ such that for every $\varepsilon > 0$ we have that $u_\varepsilon(t) \geq c\varepsilon$ for a.a. $t \in [0,T]$.

For $\varepsilon \in [0,\infty)$, define the Hamiltonian $H_\varepsilon : [0,T]\times\mathbb{R}\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}$ by

$$H_\varepsilon(t,y,p,u) := \ell_\varepsilon(t,y,u) + pf(t,y,u). \tag{9}$$

The Pontryagin minimum principle (cf. [77]) yields the existence of $p_\varepsilon \in W^{1,2}([0,T];\mathbb{R})$ such that

$$\dot{y}_\varepsilon(t) = A(t)y_\varepsilon(t) + u_\varepsilon(t) \;\; \text{for a.a. } t\in[0,T], \tag{10}$$

$$-\dot{p}_\varepsilon(t) = A(t)p_\varepsilon(t) + C(t)[y_\varepsilon(t) - \bar{y}(t)] \;\; \text{for a.a. } t\in[0,T], \tag{11}$$

$$y_\varepsilon(0) = y_0, \quad p_\varepsilon(T) = M[y_\varepsilon(T) - \bar{y}(T)], \tag{12}$$

$$u_\varepsilon(t) = \operatorname{argmin}\{H_\varepsilon(t,y_\varepsilon(t),p_\varepsilon(t),v) : \; v \geq 0\} \;\; \text{for a.a. } t\in[0,T]. \tag{13}$$

Our aim is to establish the relations between $(y_\varepsilon, p_\varepsilon, u_\varepsilon)$, the so-called *the central path*, and (y_0, p_0, u_0), for $\varepsilon > 0$ small enough. The first step is to use (13) in order to eliminate u_ε in the system (10)-(12). In fact, condition (13) yields that for a.a. $t \in [0,T]$, $u_\varepsilon(t) := \varphi_\varepsilon(-p_\varepsilon(t))$, where

$$\varphi_\varepsilon(x) := \begin{cases} \frac{1}{2}\left(x + \sqrt{x^2+4\varepsilon}\right) & \text{if } \varepsilon > 0, \\ \max\{x,0\} & \text{if } \varepsilon = 0. \end{cases} \tag{14}$$

Thus, for every $\varepsilon \in [0, \infty)$, optimality conditions (10)-(12) are equivalent to

$$\begin{array}{rcl} \dot{y}_\varepsilon(t) & = & A(t)y_\varepsilon(t) + \varphi_\varepsilon(-p_\varepsilon(t)), \\ -\dot{p}_\varepsilon(t) & = & A(t)p_\varepsilon(t) + C(t)[y_\varepsilon(t) - \bar{y}(t)], \\ y_\varepsilon(0) & = & y_0, \quad p_\varepsilon(T) = M[y_\varepsilon(T) - \bar{y}(T)]. \end{array} \tag{15}$$

The forward backward system (15) induces the definition of the mapping:

$$F : W^{1,1}([0,T];\mathbb{R}) \times W^{1,1}([0,T];\mathbb{R}) \times \mathbb{R}_+ \to L^1([0,T];\mathbb{R}) \times \mathbb{R} \times L^1([0,T];\mathbb{R}) \times \mathbb{R}$$

defined by

$$F(y,p,\varepsilon)(\cdot) := \begin{pmatrix} \dot{y}(\cdot) - A(\cdot)y(\cdot) - \varphi_\varepsilon\left(-p(\cdot)\right) \\ y(0) - y_0 \\ \dot{p}(\cdot) + A(\cdot)p(\cdot) + C(\cdot)(y(\cdot) - \bar{y}(\cdot)) \\ p(T) - M[y(T) - \bar{y}(T)] \end{pmatrix}. \tag{16}$$

In order to obtain a first order expansion of $(y_\varepsilon, p_\varepsilon)$ around (y_0, p_0) the first idea that comes to mind, as in the classical sensitivity analysis, is to apply the implicit function theorem to the mapping F at $(y_0, p_0, 0)$. Unfortunately, it is shown in Chapter 1 that this theorem is not applicable since, in general, $D_\varepsilon F(y_0, p_0, 0)$ does not exist. As an alternative we use the restoration theorem (theorem 2) and its corollary (corollary 3), to obtain the desired asymptotic expansion and the associated error estimates for the central path. It is seen that the strict differentiability hypothesis (4), which in our case is with respect to (y, p) at $(y_0, p_0, 0)$, is strongly related with the concept of strict complementary for the solution of a finite-dimensional problem, exposed in subsection 0.1.1. In fact, let us assume the

Strict complementarity assumption: *There exists a subset* T_{sing} *of* $[0,T]$ *with meas*$(T_{\text{sing}}) = 0$*, such that for each* t *in* $[0,T] \setminus T_{\text{sing}}$ *the point* $u_0(t)$ *satisfies the strict complementarity conditions for the minimization problem*

$$\min \left\{ H_0(t, y_0(t), p_0(t), w) : w \in \mathbb{R}_+ \right\}.$$

The assumption above can be reformulated in the following geometrical form: Except for a null Lebesgue set the curve $p_0(t)$ does not intersect the x-axis, i.e. the function $t \in [0,T] \to \frac{\mathrm{d}}{\mathrm{d}t}\varphi_0(-p_0(t))$ is a.s. well defined.

Under this hypothesis we can apply theorem 2 and prove our main results. The first one concerns the error estimates for the central path, and it says that the error bounds can be calculated from the error bounds of the analogous finite dimensional problems (which, in the case of the logarithmic penalty, are of order $\sqrt{\varepsilon}$).

Theorem 4 (Error estimates for interior penalty) *Under the strict complementarity assumption , for ε small enough we have that:*
(i) *The error estimates for $u_\varepsilon, y_\varepsilon$ and p_ε are given by*

$$||u_\varepsilon - u_0||_\infty + ||p_\varepsilon - p_0||_{1,\infty} + ||y_\varepsilon - y_0||_{1,\infty} = O(\sqrt{\varepsilon})$$

with in addition $u_\varepsilon \to u_0$ in $W^{1,1}$.
(ii) *In addition, let us assume that $\{t \in [0,T]\ ;\ p_0(t) = 0\}$ is finite and that the following implication holds:*

$$p_0(t_0) = 0 \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} p_0(t_0) \neq 0. \tag{17}$$

Then

$$||u_\varepsilon - u_0||_1 + ||p_\varepsilon - p_0||_{1,1} + ||y_\varepsilon - y_0||_{1,1} = O(\varepsilon |\log \varepsilon|). \tag{18}$$

Now, we state our second main result which yields the asymptotic expansion of $(y_\varepsilon, p_\varepsilon)$ around (y_0, p_0) in $W^{1,1}([0,T];\mathbb{R})$.

Theorem 5 (Asymptotic expansion) *Suppose that the strict complementarity assumption* (1.53) *holds, then for ε small enough,*

$$\begin{pmatrix} y_\varepsilon \\ p_\varepsilon \end{pmatrix} = \begin{pmatrix} y_0 \\ p_0 \end{pmatrix} - D_{(y,p)}F(y_0,p_0,0)^{-1}F(y_0,p_0,\varepsilon) + r(\varepsilon),$$

where

$$r(\varepsilon) = o(||F(y_0,p_0,\varepsilon)||_1).$$

Moreover, the first term of the expansion $-D_{(y,p)}F(y_0,p_0,0)^{-1}F(y_0,p_0,\varepsilon)$ is the unique solution of

$$\begin{cases} \mathrm{Min}\ \frac{1}{2}\displaystyle\int_0^T \left(|v(t)|^2 + C(t)|\sigma(t)|^2\right) \mathrm{d}t + \frac{1}{2}M|\sigma(T)|^2, \\ s.t. \\ \dot{\sigma}(t) = A(t)\sigma(t) + v(t) + \left[\varphi_\varepsilon(-p_0(t)) - \varphi_0(-p_0(t))\right], \\ \sigma(0) = 0, \quad v(t) = 0 \quad \text{if}\ \ p_0(t) \geq 0. \end{cases}$$

Finally, let us mention that theorems 4 and 5 are proved in Chapter 1 for a general linear quadratic problem and for a general class of penalty functions. Of course, the error bounds obtained there depend on the chosen penalty function. The main technical difficulty appears when the control is coupled in the cost function by a non diagonal matrix $R(t)$.

0.1.2.2 Optimal control of PDEs

The study presented here is the subject of the report [25], which extends the results of the previous section to the optimal control problem of a semilinear PDE, under non negativity constraints over the control. For $u \in L^s(\Omega)$ ($s \in [2, \infty]$) denote by $y_u \in W^{2,s}(\Omega)$ the unique solution of

$$\left\{ \begin{array}{rcll} -\Delta y(x) + \phi(y(x)) & = & f(x) + u(x) & \text{for} \quad x \in \Omega, \\ y(x) & = & 0 & \text{for} \quad x \in \partial\Omega, \end{array} \right. \tag{19}$$

where Ω is a bounded open set of $\mathbb{R}^n$ with C^2 boundary, $f \in L^s(\Omega)$ and ϕ is a $\mathcal{C}^2$ Lipschitz nondecreasing real valued function over $\mathbb{R}$. For $s > n/2$ ($s = 2$ if $n \leq 3$), let us define $J_0 : L^s(\Omega) \to \mathbb{R}$ by

$$J_0(u) := \tfrac{1}{2} \int_\Omega (y_u(x) - \bar{y}(x))^2 dx + \tfrac{1}{2} N \int_\Omega u(x)^2 \mathrm{d}x. \tag{20}$$

We are interested in the following optimization problem

$$\text{Min } J_0(u) \quad \text{subject to } u \in \mathcal{U}_+^s. \tag{$\mathcal{CP}_0^s$}$$

where

$$\mathcal{U}_+^s := \{ v \in L^s(\Omega) \ / \ v(x) \geq 0, \text{ for a.a. } x \in \Omega \}.$$

Since ϕ can be nonlinear, problem ($\mathcal{CP}_0^s$) is a non-convex one. Nevertheless, it can be shown (corollary 51) that ($\mathcal{CP}_0^s$) has at least one solution. Our main results will depend heavily on a second-order sufficient condition at a local minimum of ($\mathcal{CP}_0^s$). Lemma 6.27 in [24] yields that $J_0 : L^s(\Omega) \to \mathbb{R}$ is $\mathcal{C}^2$ if $s > n/2$ ($s = 2$ if $n \leq 3$). That is the main reason for considering $L^s(\Omega)$ as control space, rather than the standard space $L^2(\Omega)$.

For every $u \in L^s(\Omega)$ define the adjoint state $p_u \in W^{2,s}(\Omega)$, as the unique solution of

$$\left\{ \begin{array}{rcll} -\Delta p(x) + \phi'(y_u(x))p(x) & = & y_u(x) - \bar{y}(x) & \text{for} \quad x \in \Omega, \\ p(x) & = & 0 & \text{for} \quad x \in \partial\Omega. \end{array} \right. \tag{21}$$

Let $u_0 \in \mathcal{U}_+^s$ be a local solution of ($\mathcal{CP}_0^s$) and denote respectively by y_0 and p_0 its associated state and adjoint state. Applying classical techniques (see [55, 67, 73]) we obtain that (recall (14))

$$u_0(x) = \varphi_0(-p_0(x)) \text{ for a.a. } x \in \Omega.$$

Now, let us suppose that u_0 is locally unique in the $L^s(\Omega)$ ball $\bar{B}_s(u_0, b)$ and, for $\varepsilon > 0$, consider the following logarithmic penalized version of ($\mathcal{CP}_0^s$)

$$\text{Min } J_\varepsilon(u) := J_0(u) - \varepsilon \int_\Omega \log(u(x)) \mathrm{d}x \quad \text{s. t. } u \in \mathcal{U}_+^s \cap \bar{B}_s(u_0, b) \quad (\mathcal{CP}_\varepsilon^{b,s}),$$

As for $(\mathcal{CP}_0^s)$, problem $(\mathcal{CP}_\varepsilon^{b,s})$ has at least one solution. Note that the application

$$u \in L^s(\Omega) \to -\int_\Omega \log(u(x))\mathrm{d}x \in \mathbb{R} \cup \{+\infty\}$$

is not continuous, hence not differentiable. Thus it is not immediate to write optimality conditions for $(\mathcal{CP}_\varepsilon^{b,s})$. However, using an $L^1(\Omega)$ contraction principle (lemma 54), we get that, as $\varepsilon \downarrow 0$, the solutions u_ε of $(\mathcal{CP}_\varepsilon^{b,s})$ converge to u_0 in $L^s(\Omega)$. In addition, there exists $c, K > 0$ such that for ε small enough

$$c\varepsilon \leq u_\varepsilon(x) \leq K \quad \text{for a.a. } x \in \Omega. \tag{22}$$

The estimates (22) imply that u_ε solves

$$\text{Min } J_\varepsilon(u) \quad \text{subject to } u \in \mathcal{U}_+^s \cap \bar{B}_s(u_0, b_0) \cap L^\infty(\Omega)$$

and the application $u \in L^\infty(\Omega) \to -\int_\Omega \log(u(x))\mathrm{d}x \in \mathbb{R} \cup \{+\infty\}$ is differentiable at u_ε, which allows us to write first order optimality conditions. In fact, denoting respectively by y_ε and p_ε the state and adjoint state associated to u_ε, we have that (recall (14))

$$u_\varepsilon(x) = \varphi_\varepsilon(-p_\varepsilon(x)) \text{ for a.a. } x \in \Omega.$$

Therefore, it is natural to define the map $F : W^{1,s} \times W^{1,s} \times \mathbb{R}_+ \to L^s(\Omega) \times L^s(\Omega)$ by

$$F(y, p, \varepsilon)(\cdot) := \begin{pmatrix} \Delta y(\cdot) + \varphi_\varepsilon(-N^{-1}p(\cdot)) + f(\cdot) - \phi(y(\cdot)) \\ \Delta p(\cdot) + y(\cdot) - \bar{y}(\cdot) - \phi'(y(\cdot))p(\cdot) \end{pmatrix}. \tag{23}$$

Let us assume the following hypothesis

(H1) For the adjoint state p_0, associated to any local solution u_0 of $(\mathcal{CP}_0^s)$, it holds that

$$\text{meas}\left(\{x \in \Omega \ / \ p_0(x) = 0\}\right) = 0.$$

(H2) At any local solution u_0 of $(\mathcal{CP}_0^s)$, the following second-order condition holds

$$D^2 J_0(u_0)(h, h) > 0 \quad \text{for all } h \in C(u_0) \setminus \{0\} \tag{24}$$

where $C(u_0) := T_{\mathcal{U}}(u_0) \cap DJ(u_0)^\perp$ is the usual *critical cone* at u_0.

Asummptions **(H1)**, **(H2)** imply that the hypothesis of theorem 2 are satisfied at $(y_0, p_0, 0)$. More precisely, assumption **(H1)** allows to prove (4), while **(H2)** yields the surjectivity assumption of the operator A.

Now we can state our main results:

Theorem 6 *Let u_0 be a solution of $(\mathcal{CP}_0^s)$, suppose that ϕ is C^2 and that **(H1)**, **(H2)** hold. Denote respectively by y_0 and p_0 the state and adjoint state associated to u_0. Then there are $\bar{b} > 0$ and $\bar{\varepsilon} > 0$ such that for $\varepsilon \in [0, \bar{\varepsilon}]$ problem $(\mathcal{CP}_\varepsilon^{\bar{b},s})$ has a unique solution u_ε. In addition, denoting by y_ε and p_ε the associated state and adjoint state for u_ε, the following expansion around (y_0, p_0) holds*

$$\begin{pmatrix} y_\varepsilon \\ p_\varepsilon \end{pmatrix} = \begin{pmatrix} y_0 \\ p_0 \end{pmatrix} + D_{(y,p)}F(y_0, p_0, 0)^{-1}F(y_0, p_0, \varepsilon) + r(\varepsilon), \tag{25}$$

where $r(\varepsilon) = o(||F(y_0, p_0, \varepsilon)||_s)$. Moreover, $D_{(y,p)}F(y_0, p_0, 0)^{-1}F(y_0, p_0, \varepsilon)$ is characterized as being the unique solution of

$$\begin{cases} \text{Min} \displaystyle\int_\Omega \left[\tfrac{1}{2}Nv^2 + \tfrac{1}{2}\left(1 - p_0\phi''(y_0)\right)z^2\right] \mathrm{d}x, \\ s.t. \\ -\Delta z(x) + \phi'(y_u(x))z(x) = v + \varphi_\varepsilon(q_0) - \varphi_0(q_0) \quad \text{for } x \in \Omega, \\ z(x) = 0 \quad \text{for } x \in \partial\Omega, \quad v(x) = 0 \quad \text{if } u_0(x) = 0. \end{cases}$$

Theorem 7 *Suppose that the assumptions of theorem 6 hold. Let $\bar{b} > 0$ be such that $(\mathcal{CP}_\varepsilon^{\bar{b},s})$ has a unique solution u_ε for $\varepsilon > 0$ small enough. Then:*
(i) *We have*

$$||u_\varepsilon - u_0||_\infty + ||p_\varepsilon - p_0||_{2,s} + ||y_\varepsilon - y_0||_{2,s} \;=\; O(\sqrt{\varepsilon}). \tag{26}$$

(ii) *If in addition $n \leq 3$ (hence $s = 2$), there exist $m \in \mathbb{N}$, positive real numbers $\alpha > 0$, $0 < \bar{\delta} < 1$ and a finite collection of closed C^2 curves $(C_i)_{1 \leq i \leq m}$ such that:*

- *The singular set $\{x \in \Omega \ / \ p_0(x) = 0\}$ can be expressed as*

$$\{x \in \Omega \ / \ p_0(x) = 0\} = \bigcup_{i=1}^m C_i. \tag{27}$$

- *For all $i \in \{1, ..., m\}$, defining $C_i^{\bar{\delta}} := \{x \in \Omega;\ dist(x, C_i) \leq \bar{\delta}\}$, it holds that:*

$$|p_0(x)| \geq \alpha\, dist(x, C_i) \quad \text{for all } x \in C_i^{\bar{\delta}}. \tag{28}$$

Then

$$||u_\varepsilon - u_0||_2 + ||p_\varepsilon - p_0||_{2,2} + ||y_\varepsilon - y_0||_{2,2} = O(\varepsilon^{\frac{3}{4}}). \tag{29}$$

We conclude this section remarking that the above results are generalized in Chapter 2 for a large class of penalty functions.

0.2 Stochastic optimal control

Let $T > 0$ and consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, on which a d-dimensional ($d \in \mathbb{N}^*$) Brownian motion $W(\cdot)$ is defined with $\mathbb{F} = \{\mathcal{F}_t\}_{0 \leq t \leq T}$ being its natural filtration, augmented by all $\mathbb{P}$-null sets in $\mathcal{F}$. Consider the following controlled stochastic differential equation (SDE)

$$\begin{array}{rcl} \mathrm{d}y(t) & = & f(y(t), u(t))\mathrm{d}t + \sigma(y(t), u(t))\mathrm{d}W(t), \quad \text{for } s \in (t, T) \\ y(t) & = & x, \end{array} \tag{30}$$

where $x \in \mathbb{R}^n$ and $0 \leq t < T$. In the notation above, $y(t)$ represents the state variable, controlled by $u \in \mathcal{U}[0, T]$, where

$$\mathcal{U}[0, T] := \{u : [0, T] \times \Omega \to U \ / \ u \text{ is prog. measurable}\}$$

for some subset $U \subseteq \mathbb{R}^m$. We say that u is admissible if $u \in \mathcal{U}[0, T]$ and the SDE (30) has a unique solution y_u^x. The set of admissible process is denoted by $\mathcal{U}_{ad}$. For a fixed $x_0 \in \mathbb{R}^n$, we are interested in problem $V(0, x_0)$ defined as

$$V(0, x_0) := \mathrm{Inf}_{u \in \mathcal{U}_{ad}} \mathbb{E}\left(\int_0^T \ell(y_u^{x_0}(t), u(t))\mathrm{d}t + \phi(y_u^{x_0}(T)) \right),$$

where ℓ and ϕ are the running and final cost, respectively. Standard assumptions are supposed to hold for the functions that define the dynamics and the cost.

0.2.1 A review of the global approach

We begin by briefly reviewing the global approach (for a detailed exposition we refer the reader to the excellent books [45, 76, 93]). It consists in to embed the problem $V(0, x_0)$ into a family of problems, parameterized by $(t, x) \in [0, T] \times \mathbb{R}^n$, defined by

$$V(t, x) := \mathrm{Inf}_{u \in \mathcal{U}_{ad}} \mathbb{E}\left(\int_t^T \ell(y_u^x(t), u(t))\mathrm{d}t + \phi(y_u^x(T)) \right).$$

If $V \in C^{1,2}([0, T] \times \mathbb{R}^n)$ then, it is proved, using the dynamic programming principle, that V is a solution of the following second-order PDE:

$$\begin{array}{rcl} \frac{\partial V}{\partial t}(t, x) + \mathcal{H}\left(x, V(t, x), DV(t, x), D^2V(t, x)\right) & = & 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n \\ V(T, x) & = & \phi(x), \quad x \in \mathbb{R}^n. \end{array} \tag{31}$$

where $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times n} \to \mathbb{R}$ is defined by

$$\mathcal{H}(x, r, p, A) := \inf_{u\in U} \left\{ \ell(x, u) + p^\top f(x, u) + \tfrac{1}{2}\mathrm{Tr}\left[\sigma\sigma(x, u)^\top A\right] \right\}.$$

Unfortunately, only continuity results hold *a priori* for V. Nevertheless, it can be shown that V is the unique solution of (31) in the weak sense of *viscosity* solutions (see [37]). In this thesis we will not deal with the latter approach, which has been widely studied theoretically and numerically in the recent years. In fact, we will analyze the stochastic optimal control problem from a variational point of view, which we review in the next section.

0.2.2 A review of the variational approach

We offer here only a brief review of the variational approach. For a complete exposition we refer the reader to [10], [93, Chapter 3] and the references therein. In this approach we work directly with problem $V(0, x_0)$ and, for simplicity, we suppose that the admissible controls belong to a Banach space. This fact allow us to use general optimization techniques in order to establish optimality conditions. More precisely, consider the spaces

$$\begin{aligned} L^2_{\mathcal{F}} &:= \{u : [0, T] \times \Omega \to \mathbb{R}^m \ / \ u \ \text{ is prog. measurable and } ||u||_2 < \infty\}, \\ L^{2,\infty}_{\mathcal{F}} &:= \{y : [0, T] \times \Omega \to \mathbb{R}^n \ / \ y \ \text{ is prog. measurable and } ||y||_{2,\infty} < \infty\}, \end{aligned}$$

where

$$||u||_2^2 := \mathbb{E}\left(\int_0^T |u(t)|^2 \mathrm{d}t\right), \quad ||y||_{2,\infty}^2 := \mathbb{E}\left(\sup_{t\in[0,T]} |y(t)|^2\right).$$

It is well known that if f, σ have linear growth, then for every $u \in L^2_{\mathcal{F}}$ equation (30) admits a unique solution $y_u \in L^{2,\infty}_{\mathcal{F}}$ and the there exists $C > 0$ such that

$$||y_u||_{2,\infty}^2 \leq C\left(|x_0|^2 + ||f(0, u(\cdot))||_2^2 + ||\sigma(0, u(\cdot))||_2^2\right). \tag{32}$$

Therefore, it is natural to assume that

$$\mathcal{U}_{ad} = \left\{u \in L^2_{\mathcal{F}} \ / \ u(t, \omega) \in U \ \text{ for a.a. } (t, \omega) \in [0, T] \times \Omega\right\}. \tag{33}$$

Since x_0 is fixed, we will write $y_u = y_u^{x_0}$. Thus, problem $V(0, x_0)$ can be expressed in the following way

$$\text{Inf } J(u) := \mathbb{E}\left(\int_t^T \ell(y_u(t), u(t))\mathrm{d}t + \phi(y_u(T))\right) \text{ s.t. } u \in \mathcal{U}_{ad}. \tag{$\mathcal{SP}$)$_0$}$$

The existence problem for $(\mathcal{SP})_0$ is a difficult task, which has been analyzed by several researchers. Let us cite the works [7, 38, 41, 44, 60] and the survey [28]. From now on we assume that a solution of $(\mathcal{SP})_0$ exists. The variational approach consists in to study the effects of perturbations of a local minimum on the cost function J. In a very general framework, first order conditions can be established. The procedure is the natural extension of the analysis in the deterministic case. In fact, let $\bar{u}$ be a solution and set $\bar{y} := y_{\bar{u}}$. Consider the following backward stochastic differential equation (BSDE), with variables (p, q),

$$\begin{aligned} \mathrm{d}p(t) &= -\left[\ell_y(t)^\top + f_y(t)^\top p(t) + \sum_{i=1}^{m} \sigma_y^i(t)^\top q^i(t)\right] \mathrm{d}t + q(t)\mathrm{d}W(t), \\ p(T) &= D_y\phi(\bar{y}(T))^\top, \end{aligned} \tag{34}$$

where

$$\ell_y(t) := D_y\ell(\bar{y}(t), \bar{u}(t)); \quad f_y(t) := D_y f(\bar{y}(t), \bar{u}(t)).$$

Under standard assumptions (see [8, 18]), the above equation admits a unique adapted solution $(\bar{p}, \bar{q}) \in L^{2,\infty}_{\mathcal{F}} \times (L^2_{\mathcal{F}})^d$ called the adjoint state associated to $\bar{u}$. Moreover, there exists $C' > 0$ such that

$$||\bar{p}||^2_{2,\infty} + \sum_{i=1}^{d} ||\bar{q}^i||^2_2 \leq C' \left[\mathbb{E}\left(|D_y\phi(\bar{y}(T))|^2\right) + ||\ell_y(\cdot)||^2_2\right]. \tag{35}$$

The Hamiltonian H of the problem is defined as

$$H(y, p, q, u) := \ell(y, u) + p \cdot f(y, u) + \sum_{i=1}^{d} q^i \cdot \sigma^i(y, u). \tag{36}$$

When $\sigma_u \equiv 0$ then by perturbing $\bar{u}$ with the so-called *needle (or spike) variations* (see [77]), it can be shown that the optimal control $\bar{u}$ satisfies the following Pontryagin principle (see [8, 9, 15, 16, 18, 53, 61, 62, 63] for related works)

$$\bar{u}(t, \omega) \in \mathrm{argmin}_{v \in U} H(\bar{y}(t, \omega), \bar{p}(t, \omega), \bar{q}(t, \omega), v) \quad \text{for a.a. } (t, \omega). \tag{37}$$

Also, by introducing a generalized Hamiltonian and adding a second pair of adjoint variables, the previous condition (37) has been generalized, to the case when σ can depend on u by Peng in [75].

0.2.3 Presentation of our main results

We begin by extending the logarithmic barrier method of chapter 1 to the case of a stochastic LQ problem. Even if we do not obtain an asymptotic expansion for the state and adjoint state, we are able to prove the convergence for the central path together with some error estimates. Such estimates are the natural extensions of those obtained in chapter 1 in the deterministic framework.

Next, we deal with a general stochastic optimal control problem with convex constraints but not necessarily of local type. Indeed, using the variational approach we are able to derive first and second order optimality conditions for a local solution. They are the natural extensions of well know results in the deterministic case.

0.2.3.1 Error estimates for a penalized stochastic LQ problem

In this section we consider an important instance of $(\mathcal{SP})_0$, which is the case of a control constrained stochastic LQ problem. The analysis presented here are the subject of report [26]. In order to illustrate the result in a simple manner, we consider a very particular convex LQ problem. For a more general convex LQ problem we refer the reader to chapter 3. We suppose here that $m = n = d = 1$ and that the data of $(\mathcal{SP})_0$ is

$$\begin{array}{l} \ell(y,u) = \frac{1}{2}\left(u^2 + y^2\right), \quad \phi(y) = \frac{1}{2}y^2, \\ f(y,u) = y + u, \quad \sigma(y,u) = y + u, \quad x_0 \in \mathbb{R} \end{array}$$

and

$$\mathcal{U}_{ad} := \left\{ u \in L^2_{\mathcal{F}} \ / \ u(t,\omega) \geq 0 \ \text{ for a.a } (t,\omega) \in [0,T] \times \Omega \right\}.$$

Since the cost function is strongly convex and continuous, problem $(\mathcal{SP})_0$ admits a unique solution u_0. We denote respectively by $y_0 := y_{u_0}$ and $(p_0, q_0) := (p_{u_0}, q_{u_0})$ for the state and the adjoint state associated to u_0. The stochastic Pontryagin minimum principle (SPMP) (37) implies that

$$u_0(t,\omega) = \phi_0(-p_0(t,\omega) - q_0(t,\omega)) \quad \text{for a.a. } (t,\omega) \in [0,T] \times \Omega,$$

where we recall that ϕ_0 is defined in (14).

As in section 0.1.2.1, for $\varepsilon > 0$ we define problem $(\mathcal{SP})_\varepsilon$ by modifying the cost ℓ of $(\mathcal{SP})_0$ by

$$\ell_\varepsilon(t,y,u) = \ell(t,y,u) - \varepsilon \log u.$$

It can be checked that the new cost function is strongly convex and lower semicontinuous. Thus, problem $(\mathcal{SP})_\varepsilon$ admits a unique solution u_ε. We denote respectively by $y_\varepsilon := y_{u_\varepsilon}$ and $(p_\varepsilon, q_\varepsilon) := (p_{u_\varepsilon}, q_{u\varepsilon})$ the corresponding

state and adjoint state. Recalling the definition of ϕ_ε in (14), the SPMP yields that

$$u_\varepsilon(t,\omega) = \phi_\varepsilon(-p_\varepsilon(t,\omega) - q_\varepsilon(t,\omega)) \quad \text{for a.a. } (t,\omega) \in [0,T]\times\Omega.$$

Moreover, with the help of the SPMP again in can be proved that (see chapter 3 for details)

Proposition 8 *There exist $C'' > 0$ such that*

$$u_\varepsilon(t,\omega) \geq \frac{C''\varepsilon}{1+|p_\varepsilon(t,\omega)|+|q_\varepsilon(t,\omega)|} \quad \textit{for a.a. } (t,\omega) \in [0,T]\times\Omega.$$

Proposition above and a duality argument yield the following error estimate for the cost function.

Proposition 9 *For every $\varepsilon > 0$, it holds that*

$$J(u_\varepsilon) - J(u_0) \leq T\varepsilon.$$

Sketch of proof. Consider the Lagrangian $\mathcal{L} : L^2_{\mathcal{F}} \times L^2_{\mathcal{F}} \to \mathbb{R}$ defined as

$$\mathcal{L}(u,\lambda) := J_0(u) - \langle \lambda, u\rangle_2.$$

The dual function $d : \mathcal{U}_{ad} \to \mathbb{R}$ is given by $d(\lambda) := \inf_{u\in L^2_{\mathcal{F}}} \mathcal{L}(u,\lambda)$. Proposition 8 and estimate (35) imply that $1/u_\varepsilon \in \mathcal{U}_{ad}$. The SPMP, in its sufficient form for the convex case (see [31, Theorem 3.2]), implies that

$$d\left(\varepsilon\frac{1}{u^\varepsilon}\right) = J_0(u_\varepsilon) - \varepsilon T.$$

Therefore, by weak duality

$$J_0(u_\varepsilon) - \varepsilon T \leq \max_{\lambda\in\mathcal{U}_{ad}} \min_{u\in L^2_{\mathcal{F}}} \mathcal{L}(u,\lambda) \leq \min_{u\in L^2_{\mathcal{F}}} \max_{\lambda\in\mathcal{U}_{ad}} \mathcal{L}(u,\lambda) = \min_{u\in\mathcal{U}_{ad}} J_0(u) = J_0(u_0).$$

□

The strong convexity of $J(\cdot)$ and estimates (32), (35), easily yield

Theorem 10 *For every $\varepsilon > 0$, the following estimates hold*

$$\begin{aligned} ||u_\varepsilon - u_0||_2^2 + ||y_\varepsilon - y_0||_{2,\infty}^2 &= O(\varepsilon) \\ ||p_\varepsilon - p_0||_{2,\infty}^2 + ||q_\varepsilon - q_0||_2^2 &= O(\varepsilon) \end{aligned}$$

0.2.3.2 Optimality conditions in stochastic optimal control theory

The results presented here are studied in report [27]. In this section we consider the following stochastic optimal control problem

$$\begin{array}{l} \text{Min } J(u) := \mathbb{E}\left[\int_0^T \ell(t, y_u(t), u(t))\mathrm{d}t + \phi(y_u(T))\right] \\ \text{subject to} \qquad u \in \mathcal{U}. \end{array} \tag{$\mathcal{SP}$}$$

In the notation above $\mathcal{U} \subseteq L^2_{\mathcal{F}}$ is a nonempty closed, convex set and y_u is the unique solution of the following SDE

$$\begin{array}{rcl} \mathrm{d}y(t) & = & f(t, y(t), u(t))\mathrm{d}t + \sigma(t, y(t), u(t))\mathrm{d}W(t), \\ y(0) & = & x_0. \end{array} \tag{38}$$

Precise assumptions over the data of $(\mathcal{SP})$ are specified in Chapter 4. Let us notice that the structure of $(\mathcal{SP})$ differs slightly to that of $(\mathcal{SP})_0$, in the sense that in the former the control variable belongs to a Banach space and it is constrained to be in a general closed, convex set of $L^2_{\mathcal{F}}$. This framework contains in particular the case of convex global and local constraints.

In this work we present first and second-order necessary conditions for a local optimum $\bar{u}$ of $(\mathcal{SP})$. The main idea is to analyze the behavior of J under perturbations of $\bar{u}$ in $L^\infty_{\mathcal{F}}$, defined as

$$L^\infty_{\mathcal{F}} := \{v : [0,T] \times \Omega \to \mathbb{R}^m \ / \ v \ \text{ is prog. measurable and } ||v||_\infty < \infty\},$$

where

$$||v||_\infty := \text{ess sup}\{|v(t,\omega)|, \ (t,\omega) \in [0,T] \times \Omega\}.$$

Thus, in some sense, the perturbations considered in this work are more regular than the solution itself. From now on we fix a local solution $\bar{u}$ and we denote by $\bar{y}$ its associated state. As before, $(\bar{p}, \bar{q})$ is defined as the unique solution of (34). We set (recall (36)) $H_u(t) := H_u(t, \bar{y}(t), \bar{u}(t), \bar{p}(t), \bar{q}(t))$ and define $\Upsilon_1 : L^\infty_{\mathcal{F}} \to \mathbb{R}$ as

$$\Upsilon_1(v) := \mathbb{E}\left(\int_0^T H_u(t) \ v(t)\mathrm{d}t\right). \tag{39}$$

Using a generalization of estimate (32) and some technical computations (that take into account a first order linearization of the state), we obtain:

Proposition 11 *Let $v \in L^\infty_{\mathcal{F}}$. Then, the following first order expansion of J around $\bar{u}$ holds*

$$J(\bar{u} + v) = J(\bar{u}) + \Upsilon_1(v) + r_1(v)$$

where $\Upsilon_1(v) = O(||v||_2)$ and $r_1(v) = O(||v||^2_\infty)$.

The radial and tangent cone to $\mathcal{U}$ at $\bar{u}$ are defined respectively by

$$\begin{array}{rcl} \mathcal{R}_{\mathcal{U}}(\bar{u}) & := & \{v \in L^2_{\mathcal{F}} \ ; \ \exists\ \sigma > 0 \quad \text{such that} \quad [\bar{u}, \bar{u} + \sigma v] \subseteq \mathcal{U}\}, \\ T_{\mathcal{U}}(\bar{u}) & := & \{v \in L^2_{\mathcal{F}} \ ; \ \exists\ u(\sigma) = \bar{u} + \sigma v + o(\sigma) \in \mathcal{U},\ \sigma \geq 0,\ ||o(\sigma)/\sigma||_2 \to 0\}. \end{array}$$

For a subset $A \subseteq L^2_{\mathcal{F}}$ we write $\mathrm{adh}_2(A)$ for the adherence of A in $L^2_{\mathcal{F}}$. It is well known, since $\mathcal{U}$ is closed and convex, that $T_{\mathcal{U}}(\bar{u}) = \mathrm{adh}_2(\mathcal{R}_{\mathcal{U}}(\bar{u}))$. Let us assume that for every $u \in \mathcal{U}$

$$T_{\mathcal{U}}(u) = \mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(u) \cap L^\infty_{\mathcal{F}}\right). \tag{40}$$

Remark 12 *Assumption* (40) *is satisfied, for example, by constraint sets* $\mathcal{U}$ *which are stable under some truncation processes.*

Estimate $\Upsilon_1(v) = O(||v||_2)$ in proposition 11 implies that the linear form Υ_1 can be extended continuously to $L^2_{\mathcal{F}}$. Henceforth, proposition 11 the following first order necessary condition holds

Proposition 13 *Assume that* (40) *holds and let* $\bar{u}$ *be a local solution of* $(\mathcal{SP})$. *Then*

$$\Upsilon_1(v) \geq 0 \quad \textit{for all } v \in T_{\mathcal{U}}(\bar{u}). \tag{41}$$

In order to obtain second-order necessary conditions, a second-order linearization of the state variable, detailed in Chapter 3, is considered. In our main results we will need that at least one of the following assumptions holds:

(A1) It holds that $\sigma_{uu} \equiv 0$ and the following maps are Lipschitz

$$(u, y) \in \mathbb{R}^m \times \mathbb{R}^n \to \ell(u, y) \in \mathbb{R}, \quad y \in \mathbb{R}^n \to \phi(y) \in \mathbb{R}.$$

(A2) It holds that the following maps are affine

$$(u, y) \in \mathbb{R}^m \times \mathbb{R}^n \to f(u, y) \in \mathbb{R}^n, \quad (u, y) \in \mathbb{R}^m \times \mathbb{R}^n \to \sigma(u, y) \in \mathbb{R}^{n \times d}.$$

Let us set $H_{(y,u)^2}(t) = H_{(y,u)^2}(t, \bar{y}(t), \bar{u}(t), \bar{p}(t), \bar{q}(t))$ and define $\Upsilon_2 : L^\infty_{\mathcal{F}} \to \mathbb{R}$ by

$$\Upsilon_2(v) := \mathbb{E}\left(\int_0^T H_{(y,u)^2}(t)(v(t), y_1(t))^2 \mathrm{d}t + \phi_{yy}(\bar{y}(T))(y_1(T))^2\right),$$

where $y_1 = y_1(v)$ is defined as the unique solution of the following SDE

$$\begin{array}{rcl} \mathrm{d}y_1(t) & = & Df(t)(y_1(t), v(t))\mathrm{d}t + D\sigma(t)(y_1(t), v(t))\mathrm{d}W(t), \\ y_1(0) & = & 0. \end{array} \tag{42}$$

In the above expression $Df(t) := Df((t, \bar{y}(t), \bar{u}(t)))$, similarly notation hold for $D\sigma$. Again, technical computations yield the following second-order expansion for J around $\bar{u}$ (see corollary 97).

Proposition 14 *Assume that either* ***(A1)*** *or* ***(A2)*** *holds. Then, the following expansion holds:*

$$J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + \tfrac{1}{2}\Upsilon_2(v) + r_2(v) \quad \textit{for all } v \in L^\infty_{\mathcal{F}}. \tag{43}$$

where $\Upsilon_1(v) = O(||v||_2)$, $\Upsilon_2(v) = O(||v||_2^2)$ *and* $r_2(v) = O(||v||_\infty ||v||_2^2)$.

Using this expansion, second-order necessary conditions can be obtained under a generalization of assumption (40), to the second-order case, and assuming that $\mathcal{U}$ is *polyhedric*. For a precise statement of this result, we refer the reader to theorem 106 and corollary 109 in Chapter 3. However, for the sake of completeness let us state second-order necessary conditions in the scalar box constraint case, i.e. when

$$\mathcal{U} = \left\{ v \in L^2_{\mathcal{F}} \ / \ a \leq v(t,\omega) \leq b, \quad \text{for a.a. } (t,\omega) \in [0,T]\times\Omega \right\}. \tag{44}$$

Proposition 15 *Let* $\bar{u}$ *be a local solution of* $(\mathcal{SP})$ *where* $\mathcal{U}$ *is defined in* (44). *Suppose that either* ***(A1)*** *or* ***(A2)*** *holds. Then, the following second-order necessary conditions hold at* $\bar{u}$*:*

$$\Upsilon_2(v) \geq 0, \ \textit{for all } v \in C(\bar{u}),$$

where $C(\bar{u}) = \left\{ v \in T_{\mathcal{U}}(\bar{u}) \ / \ H_u(t)v(t,\omega) = 0, \ \textit{if} \ \ u(t,\omega) \in \{a,b\} \right\}$.

Finally, let us mention that proposition 14 directly implies (see proposition 110) a second-order sufficient condition for the unconstrained case, i.e. when $\mathcal{U} = L^2_{\mathcal{F}}$. However, for the constrained case only very partial results are obtained. The main difficulty lies in the fact that the application $u \in L^2_{\mathcal{F}} \to y_u(T) \in L^2(\Omega)$ is not weakly continuous. This fact is proved with two counterexamples (even in the case when $\sigma_u \equiv 0$) in section 3.5. Thus, the interesting question of characterizing Υ_2 in order to obtain a non-gap second-order sufficient condition remains open.

Part II

Asymptotic expansions for interior penalty solutions of control constrained problems

Chapter 1

Optimal control of a linear differential equation

Contents

1.1 Introduction

For finite dimensional optimization problems interior-point methods are recognized as being presently among the most efficient algorithms. For detailed expositions of the theory and recent developments see, for instance, [46, 74, 91] and references therein. In particular, path-following algorithms based on the logarithmic penalty are very popular by virtue of their well-known convergence properties (see [21, Part IV] and [48]).

Penalty and interior-point methods are especially well-suited for optimal control problems. A possible procedure is indeed as follows: fix a small penalty parameter, write the optimality conditions of the resulting unconstrained problem, discretize the system and apply a procedure for solving nonlinear equations. This discretization can be analyzed and evaluated with a good precision, allowing to design efficient grid refinement algorithms [11, 23]. On the other hand the system of equations corresponding to optimality conditions has a Jacobian with a band structure and can be, for instance, efficiently solved using QR factorization algorithm (see [11]). The corresponding approach has been applied to real-world aerospace optimization problems (see [12]).

When the dynamics are described by an ordinary differential equation, interior-point methods have been investigated by several authors (see e.g. [58, 64, 85, 86, 90]). Some convergence results are discussed in [22] and [85]. The latter uses a primal-dual interior point method, based on the Fisher-Burmeister complementarity function, and obtains an $O(\sqrt{\varepsilon})$ error estimate for the L^∞ norm and linear convergence of a short-step path-following method, where $\varepsilon > 0$ is the approximation parameter.

For the PDE framework see [13, 14, 79, 87, 88]. In [87] a control reduced method is developed and error estimates of $O(\sqrt{\varepsilon})$ for the L^∞ norm are obtained. Superlinear convergence is established in [79]. See also [84] for a L^s-analysis ($s \in [2, +\infty[$) where global linear and local superlinear convergence are studied.

In this work we consider a rather general linear-quadratic optimal control problem where the dynamics are described by a non autonomous affine differential equation, while nonnegativity restrictions are imposed on the control. These restrictions are penalized with a general barrier function. For this kind of problems the theoretical result obtained in [85] is not applicable (at least because of the non-boundedness of the constraint set). Let us remark that, even in a more general setting, numerical methods for optimal control problems are analyzed in [50, 51], in which a family of perturbed problems is studied and it is proved that their solutions converge to the solution of the original problem. In addition, error estimates are provided by means of a gen-

eralized implicit function theorem. Nevertheless, for interior point methods the cost function is perturbed by adding a parametrized barrier function. As we will see in section 1.4, this type of perturbation is not regular in the sense that the implicit function theorem approach is not applicable. Instead, in our case error estimates are obtained using a so-called Restoration Theorem (see Appendix) whose applicability depends on a rather general assumption: as time elapses the control of the initial problem satisfies strict complementarity conditions with respect to its Hamiltonian (except eventually on a set of times with null Lebesgue measure). Within this framework error estimates of the state, adjoint state, control and value function are derived from some associated *stationary* problems. These estimates depend on the regularity of the underlying dynamics: they involve either L^s norms or Sobolev norms (see Theorem 30).

In the particular case of the logarithmic penalty, one recovers the $O(\sqrt{\varepsilon})$ bound for the control error in the L^∞ norm and, under a transversality assumption, a bound of order $O(\varepsilon|\log\varepsilon|)$ for the L^1 norm. This is a sharp estimate in view of the example solved in [3].

On the other hand, asymptotic expansions of the state and adjoint state are obtained. This result together with the strict complementarity assumption provide a deeper understanding of the interplay between the variations of the optimal control and its junction points (times where the set of active constraints changes).

The paper is organized as follows: Section 1.2 is devoted to the problem statement and the description of its penalized versions; standard results revolving around these aspects are recalled. In Section 1.3 some associated stationary problems are described into depth, this allows in Section 1.4 to establish our main results. The last Section provides illustrative applications and a thorough study of the logarithmic penalty case for which optimal bounds are given.

The Restoration Theorem is an important tool of the present paper, it was provided in [3] and its proof is reproduced in the Appendix.

1.2 Problem statement and preliminary results

The space $\mathbb{R}^m$ ($m \in \mathbb{N}^*$) is endowed with its standard Euclidean norm denoted by $|\cdot|$. The ith coordinate of a vector x is denoted by x^i. We set $\mathbb{R}^m_+ := \{x \in \mathbb{R}^m : x^i \geq 0\}$, and $\mathbb{R}^m_{++} := \{x \in \mathbb{R}^m : x^i > 0\}$. As usual, the vector $\mathbf{1} \in \mathbb{R}^m$ is defined by $(\mathbf{1})^i = 1$ for all $i \in \{1, ..., m\}$.

Fix $T > 0$ and set $\mathcal{U} := L^2([0,T];\mathbb{R}^m)$, $\mathcal{U}_+ := L^2([0,T];\mathbb{R}^m_+)$. Given $n \in \mathbb{N}$ and $s \in [1,\infty]$, set $L^s := L^s([0,T];\mathbb{R}^n)$ and define the Sobolev space

by $W^{1,s} := \{y \in L^s;\, \dot{y} \in L^s\}$, where $\dot{y}$ is the derivative of y in the weak sense ([1]). The standard norms of these spaces are denoted by $\|\cdot\|_s$ and $\|\cdot\|_{1,s}$ respectively. Denote respectively by $\mathcal{S}^m$, $\mathcal{S}^m_+$ and $\mathcal{S}^m_{++}$ the sets of symmetric, symmetric positive semidefinite and symmetric positive definite matrices of order m. For $S \in \mathcal{S}^m$, let $\lambda_{min}(S)$ (resp. $\lambda_{max}(S)$) denote the smallest (resp. largest) eigenvalue of S.

Let m, n be two positive integers. Consider the following controlled state equation

$$\dot{y}(t) = A(t)y(t) + B(t)u(t) + \psi(t), \;\; t \in (0,T); \quad y(0) = x_0, \tag{1.1}$$

with data $T > 0$, $A \in \mathcal{C}^0([0,T];\mathbb{R}^{n\times n})$, $B \in \mathcal{C}^0([0,T];\mathbb{R}^{n\times m})$, $x_0 \in \mathbb{R}^n$ and $\psi \in L^1$. For any control $u \in \mathcal{U}$, equation (1.1) has a unique solution in $W^{1,1}$ denoted by y_u and called *the state associated with* u.

It is well known that the mapping $u \mapsto y_u$ is linear continuous from $\mathcal{U}$ into $W^{1,1}$. In fact, this follows easily by Gronwall's lemma which implies that:

$$||y_u - y_v||_\infty = O\left(||u - v||_1\right) \quad \text{for all } u, v \in \mathcal{U}. \tag{1.2}$$

1.2.1 Main problem

Let $R \in \mathcal{C}^0([0,T];\mathcal{S}^m_{++})$, $C \in \mathcal{C}^0([0,T];\mathcal{S}^n_+)$, $\varphi \in L^1$, and $M \in \mathcal{S}^m_+$. Consider the function g defined by

$$\mathbb{R}^m \times \mathbb{R}^n \times [0,T] \ni (u,y,t) \;\mapsto g(u,y,t) := \tfrac{1}{2}u^\top R(t)u + \tfrac{1}{2}y^\top C(t)y + \varphi(t)^\top y,$$

and the cost function $J_0 : \mathcal{U} \to \mathbb{R}$ defined by

$$J_0(u) := \int_0^T g(u(t), y_u(t) - \bar{y}(t), t)\,\mathrm{d}t + \tfrac{1}{2}[y_u(T) - \bar{y}(T)]^\top M[y_u(T) - \bar{y}(T)], \tag{1.3}$$

where $\bar{y} \in \mathcal{C}^0([0,T];\mathbb{R}^n)$ is a reference state function. Under our assumptions an elementary argument shows that J_0 is strongly convex and continuous.

Let us consider the following linear-quadratic optimal control problem:

$$\text{Min } J_0(u) \text{ subject to } u \in \mathcal{U}_+. \tag{$\mathcal{CP}_0$}$$

Classical arguments (see e.g. [30, 57]) imply that J_0 has a unique minimum u_0 over $\mathcal{U}_+$. For notational convenience we set $y_0 := y_{u_0}$.

For $(u,y,p,t) \in \mathbb{R}^m_+ \times \mathbb{R}^n \times \mathbb{R}^n \times [0,T]$, the classical Hamiltonian for $(\mathcal{CP}_0)$ is defined by

$$H_0(u,y,p,t) := g(u, y - \bar{y}(t), t) + p^\top[A(t)y + B(t)u + \psi(t)].$$

[1] We recall that every element of $W^{1,s}$ is continuous.

The first-order necessary optimality conditions for $(\mathcal{CP}_0)$ give the existence of $p_0 \in W^{1,1}$ such that

$$\begin{aligned}
\dot{y}_0(t) &= A(t)y_0(t) + B(t)u_0(t) + \psi(t) \quad \text{for a.a. } t \in [0,T], && (1.4)\\
-\dot{p}_0(t) &= A(t)^\top p_0(t) + C(t)[y_0(t) - \bar{y}(t)] + \varphi(t) \quad \text{for a.a. } t \in [0,T], && (1.5)\\
y_0(0) &= x_0, \quad p_0(T) = M[y_0(T) - \bar{y}(T)], && (1.6)\\
u_0(t) &\in \operatorname{argmin}\{H_0(w, y_0(t), p_0(t), t) \ : \ w \geq 0\} \quad \text{for a.a. } t \in [0,T]. && (1.7)
\end{aligned}$$

For $(R,z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$, let us denote by $\pi_0(R,z) \in \mathbb{R}^m_+$ the unique solution of

$$\operatorname{Min} \tfrac{1}{2}(x-z)^\top R(x-z), \quad \text{s.t.} \quad x \in \mathbb{R}^m_+. \qquad (\mathcal{P}_0^{R,z})$$

Indeed, the mapping $z \to \pi_0(R,z)$ is the projection of z onto $\mathbb{R}^m_+$ with respect to the norm induced by the scalar product $\langle x,y\rangle_R := \langle Rx, y\rangle$. For all t in $[0,T]$, the Hamiltonian can be rewritten as

$$\begin{aligned}
H_0(u,y,p,t) &= g\left(u + R(t)^{-1}B(t)^\top p, y - \bar{y}(t), t\right) + p^\top[A(t)y + \psi(t)] \\
&\quad -\tfrac{1}{2}p^\top B(t)\, R(t)^{-1}B(t)^\top p.
\end{aligned} \qquad (1.8)$$

Thus, by using (1.7), the optimal control may be expressed as

$$u_0(t) = \pi_0(R(t), -R(t)^{-1}B(t)^\top p_0(t)) \quad \text{for a.a. } t \in [0,T]. \qquad (1.9)$$

1.2.2 Penalized problems

Let us introduce interior penalty approximations of $(\mathcal{CP}_0)$. Let $\mathcal{L}$ be the class of barrier functions on $\mathbb{R}^m_+$ of the form $L(x) = \sum_{i=1}^m \ell(x^i)$, where ℓ is a convex function whose domain is either $\mathbb{R}_+$ or $\mathbb{R}_{++}$, and which satisfies: ℓ is $\mathcal{C}^\infty$ on $\mathbb{R}_{++}$ and

$$\text{(I)} \quad \lim_{r \downarrow 0} \ell'(r) = -\infty; \quad \text{(II)} \quad \lim_{r \downarrow 0} \frac{\ell''(r)}{\ell'(r)} = -\infty. \qquad (1.10)$$

Remark 16 Standard examples of functions satisfying these properties are:

(i) [*Logarithmic penalty*] $\ell(r) = -\log r$, for all $r \in (0,\infty)$ and $\ell(0) = +\infty$.
(ii) [*Entropy penalty*] $\ell(r) = r\log r$, for all $r \in (0,\infty)$ and $\ell(0) = 0$.
(iii) [*Negative power penalty*] For $p > 0$, $\ell(r) = r^{-p}$, for all $r \in (0,\infty)$ and $\ell(0) = +\infty$.
(iv) [*Power penalty*] For $p \in (0,1)$, $\ell(r) = -r^p$, for all $r \in [0,\infty)$.

Note that, for $L \in \mathcal{L}$ and $u \in \mathcal{U}$, the integral $\int_0^T L(u(t))\mathrm{d}t$ belongs to $\mathbb{R} \cup \{+\infty\}$, since L, being convex over $\mathbb{R}^n$ with a nonempty domain, is bounded from below by an affine function. Let us define $\widehat{L} : \mathcal{U} \mapsto \mathbb{R} \cup \{+\infty\}$ by

$$\widehat{L}(u) := \int_0^T L(u(t))\,\mathrm{d}t. \tag{1.11}$$

Lemma 17 *The convex function $\widehat{L}$ is lower semicontinuous (l.s.c.).*

Proof. Let $\bar{u} \in \mathcal{U}_+$ and suppose that $\widehat{L}$ is not lower semicontinuous at $\bar{u}$. Consider a sequence of functions u_n in $\mathcal{U}_+$ converging to $\bar{u}$ such that $\widehat{L}(\bar{u}) > \lim_{n\to\infty} \widehat{L}(u_n)$. Extracting a subsequence if necessary, we can assume that u_n converges almost surely to $\bar{u}$. Since L is convex there exists $a \in \mathbb{R}^m$, $b \in \mathbb{R}$ such that $L(u_n) \geq a^\top u_n + b$. Applying Fatou's lemma to the nonnegative sequence $L(u_n) - a^\top u_n - b$ and using the fact that L is lower semicontinuous we obtain

$$\lim_{n\to\infty} \widehat{L}(u_n) \geq \int_0^T \liminf_{n\to\infty} L(u_n(t))\mathrm{d}t \geq \int_0^T L(\bar{u}(t))\mathrm{d}t = \widehat{L}(\bar{u}),$$

which yields the desired contradiction. ■

For $\varepsilon > 0$, the perturbed cost function $J_\varepsilon : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$J_\varepsilon(u) := J_0(u) + \varepsilon \widehat{L}(u).$$

The *penalized problem* is defined by

$$\text{Min } J_\varepsilon(u) \text{ subject to } u \in \mathcal{U}_+. \tag{$\mathcal{CP}_\varepsilon$}$$

Since J_0 is strongly convex continuous and $\widehat{L}$ is convex, Lemma 17 implies that J_ε is strongly convex l.s.c. function. As before, classical arguments yield that J_ε has a unique minimum u_ε over $\mathcal{U}_+$. Next, we prove that u_ε is uniformly positive over $[0, T]$. First, we set

$$y_\varepsilon := y_{u_\varepsilon}.$$

Proposition 18 *For any $\bar{\varepsilon} > 0$ it holds that:*
(i) *There exist strictly positive constants $K_0 = K_0(\bar{\varepsilon})$, $K_1 = K_1(\bar{\varepsilon})$ such that:*

$$||u_\varepsilon||_2 \leq K_0, \quad ||y_\varepsilon||_\infty \leq K_1, \quad \text{for all } \varepsilon \in (0, \bar{\varepsilon}). \tag{1.12}$$

(ii) *If $\bar{\varepsilon}$ is sufficiently small, there exists a constant $K_2 = K_2(\bar{\varepsilon}) > 0$ such that for all $i \in \{1, ..., m\}$*

$$u_\varepsilon^i(t) \geq \tfrac{1}{2}(\ell')^{-1}\left(-\frac{2K_2}{\varepsilon}\right) \quad \text{for a.a. } t \in [0, T] \text{ and } \varepsilon \in (0, \bar{\varepsilon}). \tag{1.13}$$

Proof. (i) Let us define $\mathbb{1}$ as the constant mapping $\mathbb{1}(t) := \mathbf{1}$ for all $t \in [0,T]$. Since u_ε is the solution of $(\mathcal{CP}_\varepsilon)$, for all $\varepsilon \in (0,\bar{\varepsilon})$ we have that

$$J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(\mathbb{1}) = J_0(\mathbb{1}) + \varepsilon T L(\mathbf{1}) \leq J_0(\mathbb{1}) + \bar{\varepsilon} T \max\{0, L(\mathbf{1})\}. \tag{1.14}$$

Now, the continuity of $\lambda_{min}(R(\cdot))$ implies that $\underline{\lambda}(R) := \min_{t\in[0,T]} \lambda_{min}(R(t)) > 0$. Let $y \to a^\top y + b$ be an affine minorant of L. We have that

$$J_\varepsilon(u_\varepsilon) = J_0(u_\varepsilon) + \varepsilon \widehat{L}(u) \geq \tfrac{1}{2}\underline{\lambda}(R)||u_\varepsilon||_2^2 - ||\varphi||_1||y_\varepsilon - \bar{y}||_\infty + \int_0^T a^\top u_\varepsilon(t)\mathrm{d}t + bT.$$

Estimate (1.2) and the Cauchy-Schwarz inequality yield the existence of $C_1 > 0$ and $C_2 \in \mathbb{R}$ (both constants independent of ε) such that

$$J_0(u_\varepsilon) + \varepsilon \widehat{L}(u) \geq \tfrac{1}{2}\underline{\lambda}(R)||u_\varepsilon||_2^2 - C_1||u_\varepsilon||_2 + C_2. \tag{1.15}$$

Thus, completing the square in the r.h.s. of (1.15), the first inequality in (1.12) follows from (1.14), while the second one follows from (1.2) and the fact that u_ε is bounded in $\mathcal{U}$.
(ii) We argue along the lines of [22] (where the logarithmic penalty is considered) to extend the result for the class $\mathcal{L}$. With no loss of generality, we suppose that $m = 1$. By (1.10) (I) there exists $0 < \bar{\zeta} < 1$ such that ℓ is decreasing on $[0,\bar{\zeta}]$. For $\zeta \in (0,\bar{\zeta})$ set

$$I_\zeta := \{t \in [0,T];\ u_\varepsilon(t) \leq \zeta/2\},$$

and define

$$u_\varepsilon^\zeta(t) := \begin{cases} \zeta & \text{if } t \in I_\zeta \\ u_\varepsilon(t) & \text{otherwise} \end{cases} \quad ; \quad y_\varepsilon^\zeta(t) := y_{u_\varepsilon^\zeta}(t) \quad \text{for a.a. } t \in [0,T].$$

Now,

$$J_\varepsilon(u_\varepsilon^\zeta) - J_\varepsilon(u_\varepsilon) = \delta J_\varepsilon^1 + \delta J_\varepsilon^2 + \delta J_\varepsilon^3, \tag{1.16}$$

where

$$\begin{array}{lcl} \delta J_\varepsilon^1 & := & \int_0^T \left\{\frac{1}{2}R(t)[u_\varepsilon^\zeta(t) - u_\varepsilon(t)][u_\varepsilon^\zeta(t) + u_\varepsilon(t)] + \varphi(t)^\top[y_\varepsilon^\zeta(t) - y_\varepsilon(t)]\right\}\mathrm{d}t, \\ \delta J_\varepsilon^2 & := & \frac{1}{2}\int_0^T [y_\varepsilon^\zeta(t) - y_\varepsilon(t)]^\top C(t)[y_\varepsilon^\zeta(t) + y_\varepsilon(t) - 2\bar{y}(t)]\ \mathrm{d}t, \\ \delta J_\varepsilon^3 & := & \frac{1}{2}[y_\varepsilon^\zeta(T) - y_\varepsilon(T)]^\top M[y_\varepsilon^\zeta(T) + y_\varepsilon(T) - 2\bar{y}(T)]. \end{array}$$

Note that

$$\delta J_\varepsilon^1 = \tfrac{1}{2}\int_{I_\zeta} R(t)[u_\varepsilon^\zeta(t) - u_\varepsilon(t)][u_\varepsilon^\zeta(t) + u_\varepsilon(t)]\mathrm{d}t + \int_0^T \varphi(t)^\top[y_\varepsilon^\zeta(t) - y_\varepsilon(t)]\mathrm{d}t.$$

Since $\varphi \in L^1$ and $\zeta \in (0,1)$ we obtain, with (1.2) and the definition of u_ε^ζ, the existence of $C_3 > 0$ (independent of ε and ζ) such that

$$\delta J_\varepsilon^1 \leq \frac{3}{4}\zeta ||R||_\infty ||u_\varepsilon^\zeta - u_\varepsilon||_1 + C_3 ||u_\varepsilon^\zeta - u_\varepsilon||_1 \leq C_4 ||u_\varepsilon^\zeta - u_\varepsilon||_1,$$

where $C_4 := 3/4||R||_\infty + C_3$. In view of (i) the functions y_ε are uniformly bounded in L^∞ for $\varepsilon \in (0, \bar{\varepsilon})$. Together with the fact that $\bar{y} \in L^\infty$ and $\zeta < 1$, estimate (1.2) yields that $y_\varepsilon^\zeta + y_\varepsilon - 2\bar{y} = y_\varepsilon^\zeta - y_\varepsilon + 2(y_\varepsilon - \bar{y})$ is uniformly bounded for $\varepsilon \in (0, \bar{\varepsilon})$. Thus, since the matrix C is bounded, we obtain with estimate (1.2) the existence of $C_5 > 0$ (independent of ε and ζ) such that $\delta J_\varepsilon^2 \leq C_5 ||u_\varepsilon^\zeta - u_\varepsilon||_1$. Analogously, we have the existence of $C_6 > 0$ (independent of ε and ζ) such that $\delta J_\varepsilon^3 \leq C_6 ||u_\varepsilon^\zeta - u_\varepsilon||_1$. By (1.16), the definition of I_ζ and u_ε^ζ, we have the existence of $C_7 > 0$ (independent of ε and ζ) such that

$$J_0(u_\varepsilon^\zeta) - J_0(u_\varepsilon) \leq (C_4 + C_5 + C_6)||u_\varepsilon^\zeta - u_\varepsilon||_1 = C_7 \zeta \mathrm{meas}(I_\zeta).$$

Hence,

$$J_\varepsilon(u_\varepsilon^\zeta) - J_\varepsilon(u_\varepsilon) \leq C_7 \zeta \mathrm{meas}(I_\zeta) + \varepsilon \int_{I_\zeta} \left[\ell(u_\varepsilon^\zeta(t)) - \ell(u_\varepsilon(t))\right] \mathrm{d}t.$$

Using the convexity of ℓ and that $\ell'(\zeta) \leq 0$, we find that for a.a. $t \in I_\zeta$

$$\ell(u_\varepsilon^\zeta(t)) - \ell(u_\varepsilon(t)) \leq \ell'(u_\varepsilon^\zeta(t))[u_\varepsilon^\zeta(t) - u_\varepsilon(t)] \leq \tfrac{1}{2}\ell'(\zeta)\zeta.$$

This in turn implies that

$$J_\varepsilon(u_\varepsilon^\zeta) - J_\varepsilon(u_\varepsilon) \leq \zeta \mathrm{meas}(I_\zeta)\left(C_7 + \tfrac{1}{2}\varepsilon\ell'(\zeta)\right) \quad \text{for all } \zeta \in (0, \bar{\zeta}). \tag{1.17}$$

Shrinking $\bar{\zeta}$ if necessary, assumptions (1.10)(I), (II) show that ℓ' defines a bijection from $(0, \bar{\zeta})$ to $(-\infty, \ell'(\bar{\zeta}))$. This implies the existence of $K_2 = K_2(\bar{\varepsilon}) > C_7$ such that equation $K_2 + \frac{1}{2}\varepsilon\ell'(\zeta) = 0$ has a unique solution in $(0, \bar{\zeta})$ given by $\zeta(\varepsilon) := (\ell')^{-1}\left(-\frac{2K_2}{\varepsilon}\right)$. Equation (1.17) yields $0 \leq J_\varepsilon(u_\varepsilon^\zeta) - J_\varepsilon(u_\varepsilon) \leq \zeta \mathrm{meas}(I_\zeta)\left(K_2 + \frac{1}{2}\varepsilon\ell'(\zeta)\right)$ for all $\zeta \in (0, \bar{\zeta})$. Since ℓ' is stricly decreasing in $(0, \bar{\zeta})$, we have that $K_2 + \frac{1}{2}\varepsilon\ell'(\zeta) < 0$ for all $0 < \zeta < \zeta(\varepsilon)$, hence $\mathrm{meas}(I_\zeta) = 0$ for all $0 < \zeta < \zeta(\varepsilon)$. Thus $2u_\varepsilon(t) > \zeta$ for a.a. $t \in [0, T]$ and the result follows by letting $\zeta \uparrow \zeta(\varepsilon)$. ■

Remark 19 a) When $\ell(r) = -\log r$ estimate (1.13) reduces to the estimate $u_\varepsilon(t) \geq c\varepsilon$ $(c > 0)$ obtained in [22].
b) The fact that u_ε is uniformly positive over $[0, T]$ has important consequences from the numerical point of view. The reason is that if in the discretization of the penalized optimal control problem the optimal solution is strictly feasible (no active constraint), then efficient unconstrained solvers can be used to compute its solution (see [11, 23]).

For $(u, y, p, t) \in \mathbb{R}^m_+ \times \mathbb{R}^n \times \mathbb{R}^n \times [0, T]$ and $\varepsilon > 0$, the Hamiltonian H_ε for the problem $(\mathcal{CP}_\varepsilon)$ is defined by

$$H_\varepsilon(u, y, p, t) := H_0(u, y, p, t) + \varepsilon L(u),$$

where we recall that H_0, defined in (1.8), is the Hamiltonian associated to the original problem $(\mathcal{CP}_0)$.

The first-order necessary conditions for $(\mathcal{CP}_\varepsilon)$ ensure the existence of $p_\varepsilon \in W^{1,1}$ such that

$$\begin{aligned}
\dot{y}_\varepsilon(t) &= A(t)y_\varepsilon(t) + B(t)u_\varepsilon(t) + \psi(t) \quad \text{for a.a. } t \in [0, T], && (1.18)\\
-\dot{p}_\varepsilon(t) &= A(t)^\top p_\varepsilon(t) + C(t)[y_\varepsilon(t) - \bar{y}(t)] + \varphi(t) \quad \text{for } t \in [0, T], && (1.19)\\
y_\varepsilon(0) &= x_0, \quad p_\varepsilon(T) = M[y_\varepsilon(T) - \bar{y}(T)], && (1.20)\\
0 &= D_u H_\varepsilon(u_\varepsilon(t), y_\varepsilon(t), p_\varepsilon(t), t) \quad \text{for a.a. } t \in [0, T]. && (1.21)
\end{aligned}$$

Condition (1.21) yields that u_ε is the unique solution in $\mathcal{U}_{++}$ of

$$R(t)u_\varepsilon(t) + \varepsilon \nabla L(u_\varepsilon(t)) = -B(t)^\top p_\varepsilon(t) \quad \text{for a.a. } t \in [0, T]. \tag{1.22}$$

For $(R, z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$ and $\varepsilon > 0$, we denote by $\pi_\varepsilon(R, z)$ the unique solution of

$$\text{Min } \tfrac{1}{2}(x - z)^\top R(x - z) + \varepsilon L(x), \quad \text{s.t.} \quad x \in \mathbb{R}^m_+. \tag{$\mathcal{P}^{R,z}_\varepsilon$}$$

Equation (1.22) yields that

$$u_\varepsilon(t) = \pi_\varepsilon(R(t), -R^{-1}B(t)^\top p_\varepsilon(t)) \quad \text{for a.a. } t \in [0, T]. \tag{1.23}$$

Note that $(\mathcal{P}^{R,z}_\varepsilon)$ is the penalized version of the finite dimensional problem $(\mathcal{P}^{R,z}_0)$. Expressions (1.9) and (1.23) suggest that in order to study the relation between u_ε (solution of $(\mathcal{CP}_\varepsilon)$) and u_0 (solution of $(\mathcal{CP}_0)$) it will be useful to present a detailed analysis of the analogous problems $(\mathcal{P}^{R,z}_\varepsilon)$ and $(\mathcal{P}^{R,z}_0)$ in the finite dimensional setting.

1.3 Interior penalty analysis in the finite dimensional setting

Given $(R, z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$ recall that $\pi_0(R, z)$ is defined as the unique minimum of $f^{R,z}_0(x) := \frac{1}{2}(x - z)^\top R(x - z)$ over $\mathbb{R}^m_+$. Standard results of convex analysis ensures that $z \to \pi_0(R, z)$ is nonexpansive with respect to the norm induced by R. Also, given $L \in \mathcal{L}$ and $\varepsilon > 0$, recall that $\pi_\varepsilon(R, z)$ is defined as the unique minimum of $f^{R,z}_\varepsilon(x) := \frac{1}{2}(x - z)^\top R(x - z) + \varepsilon L(x)$ over $\mathbb{R}^m_+$. By a classical argument, it is easy to see that $\pi_\varepsilon(R, z)$ actually belongs to $\mathbb{R}^m_{++}$.

1.3.1 Convergence properties of the approximate projectors

This section provides several topological and asymptotic results for the family of approximated projection mappings π_ε.

Lemma 20 (boundedness) *Let $K \subseteq \mathcal{S}^m_{++} \times \mathbb{R}^m$ be a compact set. Then for every $\bar{\varepsilon} > 0$, there is a constant $C_1 = C_1(K, \bar{\varepsilon})$ such that*

$$|\pi_\varepsilon(R,z)| \leq C_1 \quad \textit{for all } \varepsilon \in (0,\bar{\varepsilon}) \textit{ and } (R,z) \in K. \tag{1.24}$$

Proof. We argue along the lines of Proposition 18(i). Let $\varepsilon \in (0,\bar{\varepsilon})$ and $y \mapsto a^\top y + b$ be an affine minorant of L. We have

$$\tfrac{1}{2}(\pi_\varepsilon(R,z)-z)^\top R(\pi_\varepsilon(R,z)-z)+\varepsilon(a^\top\pi_\varepsilon(R,z)+b) \leq f^{R,z}_\varepsilon(\pi_\varepsilon(R,z)) \leq f^{R,z}_\varepsilon(\mathbf{1}),$$

Since $f^{R,z}_\varepsilon(\mathbf{1}) \leq \max\{f^{R,z}_0(\mathbf{1}), f^{R,z}_{\bar{\varepsilon}}(\mathbf{1})\}$, we obtain

$$\frac{\lambda_{min}(R)}{2}|\pi_\varepsilon(R,z)-z|^2 + \varepsilon(a^\top\pi_\varepsilon(R,z)+b) \leq \sup_{(R',z')\in K} \max\{f^{R',z'}_{\bar{\varepsilon}}(\mathbf{1}), f^{R',z'}_0(\mathbf{1})\}$$

which is a finite number. The conclusion follows. ■

Proposition 21 (Pointwise convergence) *Let $(R,z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$, then*

$$\lim_{\varepsilon \downarrow 0} \ \pi_\varepsilon(R,z) = \pi_0(R,z).$$

Proof. Since (R,z) is fixed, we omit it in the notation. Let $y \mapsto a^\top y + b$ be an affine minorant of L and c be a lower bound of $y \to |y|^2+(a^\top y+b)$. For all $v \in \mathbb{R}^m_{++}$ we have that $\frac{1}{2}(\pi_\varepsilon-z)^\top R(\pi_\varepsilon-z)+\varepsilon(a^\top\pi_\varepsilon+b) \leq f^{R,z}_\varepsilon(\pi_\varepsilon) \leq f^{R,z}_\varepsilon(v)$, thus

$$\tfrac{1}{2}(\pi_\varepsilon - z)^\top R(\pi_\varepsilon - z) + \varepsilon c - \varepsilon|\pi_\varepsilon|^2 \leq f^{R,z}_\varepsilon(v), \quad \text{for all } v \in \mathbb{R}^m_{++}. \tag{1.25}$$

Lemma 20 (for the particular case $K = \{(R,z)\}$) implies that π_ε has a cluster point π_0 when $\varepsilon \downarrow 0$. Passing to the limit in (1.25) yields $f^{R,z}_0(\pi_0) \leq f^{R,z}_0(v)$ for all $v \in \mathbb{R}^m_{++}$ and thus for all $v \in \mathbb{R}^m_+$. Hence $\pi_0 \in \operatorname{argmin}(\mathcal{P}^{R,z}_\varepsilon)$ and since this property holds for every cluster point of the sequence π_ε the conclusion follows by using the fact that $(\mathcal{P}^{R,z}_0)$ has as unique solution $\pi_0(R,z)$. ■

In order to investigate further the converge properties of π_ε, it is useful to write down the first-order condition for problems $(\mathcal{P}^{R,z}_0)$ and $(\mathcal{P}^{R,z}_\varepsilon)$.The first-order condition for $(\mathcal{P}^{R,z}_0)$ writes

$$\begin{array}{l} R(\pi_0(R,z)-z) - \mu(R,z) = 0 \\ \mu(R,z) \geq 0 \quad ; \quad \pi_0(R,z) \geq 0; \quad \mu^i(R,z)\pi^i_0(R,z) = 0 \quad \text{for all } i \in \{1,...,m\}, \end{array} \tag{1.26}$$

where $\mu(R,z)$ is the Lagrange multiplier of the problem. On the other hand, the first-order condition for $(\mathcal{P}_\varepsilon^{R,z})$ shows that $\pi_\varepsilon(R,z)$ is the unique solution in $\mathbb{R}^m_{++}$ of

$$R(\pi_\varepsilon(R,z)-z)+\varepsilon\nabla L(\pi_\varepsilon(R,z))=0. \tag{1.27}$$

Proposition 21 asserts that for each $z\in\mathbb{R}^m$ and $R\in\mathcal{S}^m_{++}$ the vector $\pi_\varepsilon(R,z)$ converges to $\pi_0(R,z)$. Actually uniform convergence holds over each compact subset of $\mathcal{S}^m_{++}\times\mathbb{R}^m$. Let us first state a preliminary lemma.

Lemma 22 (Equicontinuity) *Let $R\in\mathcal{S}^m_{++}$ and set $\kappa(R):=||R||/\lambda_{min}(R)$ for its condition number. Then for all $\varepsilon\geq 0$*

$$|\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)|\leq\kappa(R)|y-x|,\quad \textit{for all}\quad x,y\in\mathbb{R}^m. \tag{1.28}$$

Proof. Equation (1.27) yields

$$R\left[\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)\right]+\varepsilon\left[\nabla L\left(\pi_\varepsilon(R,y)\right)-\nabla L\left(\pi_\varepsilon(R,x)\right)\right]=R(y-x). \tag{1.29}$$

Multiplying the above equation by $\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)$ and using the monotonicity of ∇L, we obtain

$$\left[\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)\right]^\top R\left[\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)\right]\ \leq\ (x-y)^\top R\left[\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)\right].$$

Whence $\lambda_{min}(R)|\pi_\varepsilon(R,y)-\pi_\varepsilon(R,x)|^2\leq||R||\ |x-y|\ |\pi_\varepsilon(R,x)-\pi_\varepsilon(R,y)|$, and the conclusion follows. ■

Proposition 23 (First order derivatives and uniform convergence)
(i) *The function $(\varepsilon,R,z)\in\mathbb{R}_{++}\times\mathcal{S}^m_{++}\times\mathbb{R}^m\mapsto\pi_\varepsilon(R,z)\in\mathbb{R}^m$ is of class $\mathcal{C}^\infty$.*
(ii) *Let $K_1\subseteq S^m_{++}$ be a compact set. For every $\varepsilon>0$ the partial derivative $D_z\pi_\varepsilon(\cdot,\cdot)$ is bounded, uniformly in ε, over $K_1\times\mathbb{R}^m$ and is given by*

$$D_z\pi_\varepsilon(R,z)=\left(I+\varepsilon R^{-1}\nabla^2L(\pi_\varepsilon(R,z)\right)^{-1}\quad \textit{for all } (R,z)\in S^m_{++}\times\mathbb{R}^m. \tag{1.30}$$

(iii) *Let $\varepsilon_0>0$ be fixed. Then, for $\varepsilon\in(0,\varepsilon_0)$, the partial derivative $D_R\pi_\varepsilon(\cdot,\cdot)$ is bounded over compact subsets of $\mathcal{S}^m_{++}\times\mathbb{R}^m$ uniformly in ε and is characterized by*

$$D_R\pi_\varepsilon(R,z)V=D_z\pi_\varepsilon(R,z)R^{-1}V\left(z-\pi_\varepsilon(R,z)\right)\quad \textit{for all}\ \ V\in\mathcal{S}^m. \tag{1.31}$$

(iv) *The function π_ε converges to π_0 uniformly on each compact subset of $\mathcal{S}^m_{++}\times\mathbb{R}^m$.*
(v) *The function $(\varepsilon,R,z)\mapsto\pi_\varepsilon(R,z)$ is continuous on $\mathbb{R}_+\times\mathcal{S}^m_{++}\times\mathbb{R}^m$.*

Proof. (i) It follows from the implicit function theorem applied to (1.27).
(ii) Since the condition number κ is a continuous function, the uniform boundedness of $D_z\pi_\varepsilon(\cdot,\cdot)$ over $K_1 \times \mathbb{R}^m$ is a consequence of Lemma 22, while equation (1.30) is obtained by differentiating (1.27) with respect to z.
(iii) Formula (1.31) follows from the differentiation of (1.27) with respect to R. The first assertion is then deduced from (ii) and Lemma 20.
(iv) Items (ii) and (iii) imply that the family $(\pi_\varepsilon)_{\varepsilon>0}$ is equicontinuous. The result follows then from Proposition 21.
(v) Let $(\bar{R},\bar{z}) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$. The continuity of $\pi_\varepsilon(R,z)$ for $\varepsilon > 0$ is a consequence of the implicit function theorem. Consider now the case $\varepsilon = 0$. For (R',z'), $(R,z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$ we have $|\pi_\varepsilon(R',z') - \pi_0(R,z)| \leq |\pi_\varepsilon(R',z') - \pi_0(R',z')| + |\pi_0(R',z') - \pi_0(R,z)|$. By using (iv) and the fact that π_0 is continuous the result follows readily. ■

1.3.2 Stratification results and strict complementarity reformulations

In this subsection we will characterize the differentiability domain of the projection mapping $\pi_0(R,\cdot)$. In fact, we will construct 2^m nonempty disjoint subsets of $\mathbb{R}^m$ having the property that the restriction of $\pi_0(R,\cdot)$ to each one of these regions is a linear projection into an appropriate subspace of $\mathbb{R}^m$.

In order to motivate the definitions given below, let us consider the case $R = I$. Optimality conditions (1.26) yield that for every $z \in \mathbb{R}$, we have $(\pi_0(I,z))_i = \max\{0, z_i\}$ for all $i \in \{1,...,m\}$. Therefore, $\pi_0(I,\cdot)$ is differentiable at z if and only if $z_i \neq 0$ for all $i \in \{1,...,m\}$. This fact is strongly related with the so-called *strict complementarity* nature of the solution $\pi_0(I,z)$ as we will see later.

For $R \in \mathcal{S}^m_{++}$ and $z \in \mathbb{R}^m$ consider the following partition of $\{1,...,m\}$

$$\left.\begin{array}{rcl} I^+(R,z) & := & \{i \in \{1,...,m\} \ : \ \pi_0^i(R,z) > 0\}, \\ I^a(R,z) & := & \{i \in \{1,...,m\} \ : \ \pi_0^i(R,z) = 0, \ \ \mu^i(R,z) > 0\}, \\ I^0(R,z) & := & \{i \in \{1,...,m\} \ : \ \pi_0^i(R,z) = 0, \ \ \mu^i(R,z) = 0\}. \end{array}\right\} \quad (1.32)$$

Definition 24 *We say that strict complementarity holds for the ith-coordinate of $\pi_0(R,z)$ if $i \notin I^0(R,z)$. If strict complementarity holds for every coordinate of $\pi_0(R,z)$ (i.e. $I^0(R,z) = \emptyset$) we say that strict complementary holds at $\pi_0(R,z)$.*

Thus, partition (1.32) describes the subsets of coordinates of $\pi_0(R,z)$ of inactive constraints, active constraints satisfying strict complementary, and active constraints where strict complementarity does not hold. In our example, i.e. when $R = I$, the first equation in (1.26) yields $\pi_0(I,z) = z +$

$\mu(I,z)$. This implies that strict complementarity holds for the i-coordinate of $\pi_0(I,z)$ if and only if $z_i \neq 0$. Therefore, we have that $\pi_0(I,\cdot)$ is differentiable at z if and only if strict complementarity holds at $\pi_0(I,z)$.

Our aim now is to extend the above analysis for a general $R \in \mathcal{S}^m_{++}$. The first equation in conditions (1.26) yields

$$z = \pi_0(R,z) - R^{-1}\mu(R,z). \tag{1.33}$$

Equation (1.33) can be interpreted in the following way: the vector z can be "recovered" from $\pi_0(R,z)$ and $\mu(R,z)$. Note that if strict complementarity holds at $\pi_0(R,z)$ then $\pi_0(R,z)$ and $\mu(R,z)$ belong to supplementary subspaces of $\mathbb{R}^m$. More precisely, given a subset Σ of $\{1,...,m\}$, define

$$Q_i := \begin{cases} \{0\} & \text{if } i \in \Sigma, \\ \mathbb{R} & \text{if } i \in \Sigma^c \end{cases} \quad \text{and} \quad Q_\Sigma := \textstyle\prod_{i=1}^m Q_i. \tag{1.34}$$

Thus, if strict complementarity holds at $\pi_0(R,z)$, then $\pi_0(R,z) \in Q_\Sigma$ and $\mu_0(R,z) \in Q_{\Sigma^c}$ with $\Sigma = I^a(R,z)$. Now, since every $z' \in \mathbb{R}^m$ can be written uniquely as $z' = z'_\Sigma + z'_{\Sigma^c}$ with $z'_\Sigma \in Q_\Sigma$ and $z'_{\Sigma^c} \in Q_{\Sigma^c}$, the discussion above suggest to define a linear mapping

$$h_\Sigma : \mathbb{R}^m \to \mathbb{R}^m, \quad h_\Sigma(z') = z'_\Sigma - R^{-1}z'_{\Sigma^c}. \tag{1.35}$$

Hence, if strict complementarity holds at $\pi_0(R,z)$, equation (1.33) can be rewritten as

$$z = h_\Sigma(z') \quad \text{where } \Sigma = I^a(R,z) \quad \text{and } z' = \pi_0(R,z) + \mu(R,z) \in \mathbb{R}^m_{++}.$$

This fact suggests that strict complementarity should hold at $\pi_0(R,z)$ for every $z \in D(R)$, where

$$D(R) := \bigcup_{\Sigma \subseteq \{1,...,m\}} D_\Sigma(R) \quad \text{and } D_\Sigma(R) := h_\Sigma(\mathbb{R}^m_{++}) \quad \text{for } \Sigma \in \{1,...,m\}. \tag{1.36}$$

The last assertion is actually proved in Lemma 25 as well as the differentiability of $\pi_0(R,\cdot)$ over $D(R)$. Conversely, we will also show that strict complementarity at $\pi_0(R,z)$ and differentiability of $\pi_0(R,\cdot)$ at z do not hold for every for $z \in \text{sing}(R)$, where

$$\text{sing}(R) := D(R)^c. \tag{1.37}$$

In order to illustrate the concepts introduced above let us consider the following example.

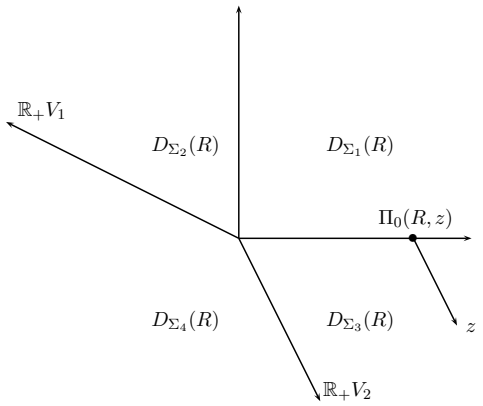

Figure 1.1: The regions $D_{\Sigma_i}(R)$ where $i = 1, ..., 4$.

Example: Here $m = 2$ and $R,\ R^{-1}$ are given by :

$$R = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}; \quad R^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{1.38}$$

Set $\Sigma_1 := \emptyset$, $\Sigma_2 := \{1\}$, $\Sigma_3 := \{2\}$ and $\Sigma_4 := \{1, 2\}$. The singular region $\text{sing}(R)$ is given by

$$\text{sing}(R) = \mathbb{R}_+ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \bigcup \mathbb{R}_+ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \bigcup \mathbb{R}_+V_1 \bigcup \mathbb{R}_+V_2,$$

where V_1, V_2 denote respectevely the first and second column of $-R^{-1}$. The regions $D_{\Sigma_i}(R)$ for $i = 1, ..., 4$ are displayed in Figure 1.1. It is also shown how a general vector z of $D_{\Sigma_3}(R)$ is projected.

Lemma 25 (Differentiability and singular sets) *Let $\Sigma \subseteq \{1, ..., m\}$. We have:*
(i) *The mapping h_Σ is bijective and linear. Thus, $D_\Sigma(R)$ is a nonempty open convex subset of $\mathbb{R}^m$.*
(ii) *For every z' in $\mathbb{R}^m$, the linear projection of $h_\Sigma(z')$ on the subspace Q_Σ (with respect to the metric induced by R) is z'_Σ.*
(iii) *The restriction of the mapping $z \to \pi_0(R, z)$ to $D_\Sigma(R)$ is the projection on the subspace Q_Σ with respect to the metric induced by R. Thus, $\pi_0(R, \cdot)$ is smooth on $D(R)$.*
(iv) *It holds that*

$$D_\Sigma(R) = \left\{ z \in \mathbb{R}^m \ : \ I^+(R, z) = \Sigma^c,\ I^a(R, z) = \Sigma,\ I^0(R, z) = \emptyset \right\}, \tag{1.39}$$

and strict complementarity does not hold at $\pi_0(R,z)$ *iff* $z \in \mathrm{sing}(R)$.
(v) *Let* Σ_1, Σ_2 *be subsets of* $\{1,...,m\}$ *with* $\Sigma_1 \neq \Sigma_2$. *Then,* $D_{\Sigma_1}(R) \cap D_{\Sigma_2}(R) = \emptyset$.
(vi) *For every* $\bar{z} \in \mathrm{sing}(R)$ *there exist subsets of* $\{1,...,m\}$ Σ_1, Σ_2 *with* $\Sigma_1 \neq \Sigma_2$ *and* $z_n \in D_{\Sigma_1}(R)$, $z'_n \in D_{\Sigma_2}(R)$ *such that* $\bar{z} = \lim_{n\uparrow\infty} z_n = \lim_{n\uparrow\infty} z'_n$. *Consequently,* $\pi_0(R,\cdot)$ *is not differentiable over* $\mathrm{sing}(R)$.

Proof. (i) Assume that $z'_\Sigma - R^{-1}z'_{\Sigma^c} = 0$. Multiplying by z'_{Σ^c} we get $(z'_{\Sigma^c})^\top R^{-1} z'_{\Sigma^c} = 0$ and so $z'_{\Sigma^c} = z'_\Sigma = 0$. The second assertion follows directly since h_Σ^{-1} exists and is continuous.
(ii) Since Q_Σ is a subspace of $\mathbb{R}^m$, a point p_Σ is the projection of $h_\Sigma(z')$ with respect to the metric induced by R iff $p_\Sigma \in Q_\Sigma$ and

$$\langle R(h_\Sigma(z') - p_\Sigma), q_\Sigma \rangle = 0 \quad \text{for all } q_\Sigma \in Q_\Sigma. \tag{1.40}$$

It can be easily verified that $p_\Sigma = z'_\Sigma$ solves (1.40). The conclusion follows.
(iii) Let $z' \in \mathbb{R}^m_{++}$. The projection $\pi_0(R, h_\Sigma(z'))$ is characterized by the existence of $\mu(R, h_\Sigma(z')) \in \mathbb{R}^m$ such that

$$\begin{array}{l} R\left[\pi_0(R,h_\Sigma(z')) - h_\Sigma(z')\right] - \mu(R,h_\Sigma(z')) = 0 \\ \mu(R,h_\Sigma(z')) \geq 0 \ ; \ \pi_0(R,h_\Sigma(z')) \geq 0; \\ \mu^i(R,h_\Sigma(z'))\pi_0^i(R,z') = 0 \text{ for all } i \in \{1,...,m\}. \end{array} \tag{1.41}$$

Since the optimality system above has as unique solution $\pi_0(R,h_\Sigma(z')) = z'_\Sigma$ and $\mu(R,h_\Sigma(z')) = z'_{\Sigma^c}$, the result follows by (i) and (ii).
(iv) First we prove (1.39). Let $z' \in \mathbb{R}^m_{++}$, then, as in (ii), $\pi_0(R,h_\Sigma(z')) = z'_\Sigma$ and $\mu(R,h_\Sigma(z')) = z'_{\Sigma^c}$. Whence $h_\Sigma(z')$ belongs to the right hand side of (1.39). Conversely, suppose that z belongs to the right hand side of (1.39). Since, by (1.26),

$$z = \pi_0(R,z) - R^{-1}\mu(R,z) = h_\Sigma(z'),$$

with $z' = \pi_0(R,z) + \mu(R,z) \in \mathbb{R}^m_{++}$, it holds that $z \in D_\Sigma(R)$. Thus (1.39) is proved.
The second assertion is straightforward by definition of $\mathrm{sing}(R)$ and (1.39).
(v) It follows directly from characterization (1.39) of $D_\Sigma(R)$.
(vi) Let $\Sigma_1 := I^a(R,\bar{z}) \cup I^0(R,\bar{z})$ and $z_n = \pi_0(R,\bar{z}) - R^{-1}\mu_n$ where $\mu_n^i = 1/n$ if $i \in I^0(R,\bar{z})$ and $\mu_n^i = \mu^i(R,\bar{z})$ otherwise. Clearly, $z_n \in D_{\Sigma_1}(R)$ and $\bar{z} = \lim_{n\uparrow\infty} z_n$. On the other hand, let us consider $\Sigma_2 := I^a(R,\bar{z})$ and $\xi_n = \pi_n - R^{-1}\mu(R,\bar{z})$ with $\pi_n^i = \pi_0^i(R,\bar{z}) + 1/n$ if $i \in I^0(R,\bar{z})$ and $\pi_n^i = \pi_0^i(R,\bar{z})$ otherwise. Thus, $\xi_n \in D_{\Sigma_2}(R)$ and $\bar{z} = \lim_{n\uparrow\infty} \xi_n$. Assertion (ii) implies that the derivatives of $\pi_0(R,\cdot)$ over $D_{\Sigma_1}(R)$ and $D_{\Sigma_2}(R)$ are respectively the

linear projections (with respect to the metric induced by R) into Q_{Σ_1} and Q_{Σ_2}. The conclusion follows using that $\Sigma_1 \neq \Sigma_2$ and hence $Q_{\Sigma_1} \neq Q_{\Sigma_2}$. ■

In view of Lemma 25, the three statements below are equivalent:

$$\begin{cases} \text{- There exists } \Sigma \subseteq \{1, ..., m\} \text{ such that } z \in D_\Sigma(R), \\ \text{- The mapping } \pi_0(R, \cdot) \text{ is differentiable at } z, \\ \text{- Strict complementarity holds at } \pi_0(R, z). \end{cases}$$

Now we turn our attention to the convergence of the derivatives of π_ε. Let $\bar{R} \in \mathcal{S}^m_{++}$ and $\bar{z} \in D(\bar{R})$. Note that since $\bar{z} \in D(\bar{R})$ it follows that $I^0(\bar{z}, \bar{R}) = \emptyset$. Define $\bar{I}^+ := I^+(\bar{R}, \bar{z})$, $\bar{I}^a := I^a(\bar{R}, \bar{z})$ and consider a compact neighborhood $\mathcal{V}$ of $(\bar{R}, \bar{z})$ in $\mathcal{S}^m_{++} \times \mathbb{R}^m$ satisfying

$$I^+(R', z') = \bar{I}^+, \quad I^a(R', z') = \bar{I}^a \quad \text{for all } (R', z') \in \mathcal{V}. \tag{1.42}$$

Lemma 26 *Using the notation introduced above:*
(i) *There exists $C_\mathcal{V} > 0$ such that, for ε small enough,*

$$\ell''\left(\pi^i_\varepsilon(R, z)\right) \leq C_\mathcal{V} \quad \textit{for all } i \in \bar{I}^+ \textit{ and } (R, z) \in \mathcal{V}. \tag{1.43}$$

(ii) *For every $j \in \bar{I}^a$, the function $-\varepsilon\ell'\left(\pi^j_\varepsilon(\cdot, \cdot)\right)$ converges uniformly in $\mathcal{V}$ to $\mu^j(\cdot, \cdot)$, which is a strictly positive function in $\mathcal{V}$.*

Proof. Let $(R, z) \in \mathcal{V}$. By definition $\pi^i_0(R, z) > 0$ for all $i \in \bar{I}^+$. Hence, assertion (i) follows from the continuity of ℓ'' and Proposition 23(iv). The first equation in conditions (1.26) together with equation (1.27) yield

$$R[\pi_0(R, z) - \pi_\varepsilon(R, z)] = \varepsilon\nabla L(\pi_\varepsilon(R, z)) + \mu(R, z). \tag{1.44}$$

Therefore, assertion (ii) follows from Proposition 23(iv). ■

For $(R, z) \in \mathcal{V}$ the indices $\bar{I}^+$ and $\bar{I}^a$ induce a partition of the underlying matrix R, defined as follows:

Definition 27 For $(R, z) \in \mathcal{V}$ define the matrices

$$\begin{array}{llll} R_{++} := (R_{i,j}) & \text{for } (i,j) \in \bar{I}^+ \times \bar{I}^+, & R_{+a} := (R_{i,j}) & \text{for } (i,j) \in \bar{I}^+ \times \bar{I}^a, \\ R_{a+} := (R_{i,j}) & \text{for } (i,j) \in \bar{I}^a \times \bar{I}^+, & R_{aa} := (R_{i,j}) & \text{for } (i,j) \in \bar{I}^a \times \bar{I}^a. \end{array}$$

The vectors z^+ and z^a are respectively obtained by removing all the coordinates of z except for those in $\bar{I}^+$ and $\bar{I}^a$.

Proposition 28 *Let $\bar{R} \in \mathcal{S}^m_{++}$ and $\bar{z} \in D(\bar{R})$ and let $\mathcal{V}$ be a compact neighborhood of $(\bar{R}, \bar{z})$ in $\mathcal{S}^m_{++} \times \mathbb{R}^m$ satisfying (1.42). Then:*
(i) *The function $D_z\pi_\varepsilon(\cdot,\cdot)$ converges to $D_z\pi_0(\cdot,\cdot)$, uniformly in $\mathcal{V}$.*
(ii) *The function $D_R\pi_\varepsilon(\cdot,\cdot)$ converges to $D_R\pi_0(\cdot,\cdot)$, uniformly in $\mathcal{V}$. In addition,*

$$D_R\pi_0(R,z)V = D_z\pi_0(R,z)R^{-1}V\left(z - \pi_0(R,z)\right) \quad \textit{for all } V \in \mathcal{S}^m. \quad (1.45)$$

(iii) *The mapping $(\varepsilon, R, z) \mapsto D_{(R,z)}\pi_\varepsilon(R,z)$ is continuous in $(\bar{\varepsilon}, \bar{R}, \bar{z})$ for every $\bar{\varepsilon} \geq 0$.*

Proof. In the sequel, for $(R,z) \in \mathcal{V}$ the coordinates of R and z are partitioned according to Definition 27. Since $I^a(\cdot,\cdot) = \bar{I}$ is constant in $\mathcal{V}$, for $(R,z) \in \mathcal{V}$ we have that $\pi^a_0(R,z) = 0$. Consequently, we obtain that $D_z\pi^a_0(R,z) = 0$. On the other hand, complementarity conditions in (1.26) imply that $\mu^+(R,z) = 0$. Thus, the first equation in conditions (1.26) yields that $0 = \left(R\left[\pi_0(R,z) - z\right]\right)^+$. Therefore, we obtain that $\pi^+_0(R,z) = R^{-1}_{++}(Rz)^+$ and as a result

$$D_z\pi^+_0(\bar{R},\bar{z})w = \bar{R}^{-1}_{++}(\bar{R}w)^+ \quad \text{for all } w \in \mathbb{R}^m. \quad (1.46)$$

Now, suppose that $|w| = 1$ and set

$$v_\varepsilon(R,z) := D_z\pi_\varepsilon(R,z)w, \quad \text{for all } \varepsilon > 0 \text{ and } (R,z) \in \mathcal{V}.$$

Equation (1.30) yields

$$Rv_\varepsilon(R,z) + \varepsilon\nabla^2 L(\pi_\varepsilon(R,z))v_\varepsilon(R,z) = Rw. \quad (1.47)$$

Denote by $\mathrm{diag}_a[\nabla^2 L(\pi_\varepsilon(R,z))]$ the diagonal matrix with diagonal $\ell''(\pi^a_\varepsilon(R,z))$, where ℓ'' is applied componentwise. Lemma 26(i) implies that

$$\begin{aligned} R_{++}v^+_\varepsilon(R,z) + R_{+a}v^a_\varepsilon(R,z) + O(\varepsilon) &= (Rw)^+, \\ R_{a+}v^+_\varepsilon(R,z) + R_{aa}v^a_\varepsilon(R,z) + \varepsilon\mathrm{diag}_a[\nabla^2 L(\pi_\varepsilon(R,z))]v^a_\varepsilon(R,z) &= (Rw)^a, \end{aligned} \quad (1.48)$$

where the $O(\varepsilon)$ is uniformly in $\mathcal{V}$. In particular,

$$v^+_\varepsilon(R,z) = R^{-1}_{++}(Rw)^+ - R^{-1}_{++}R_{+a}v^a_\varepsilon(R,z) + O(\varepsilon). \quad (1.49)$$

Let us set $\widehat{R}^+ := R_{aa} - R_{a+}R^{-1}_{++}R_{+a}$ and $A_\varepsilon(R,z) := \widehat{R}^+ + \varepsilon\mathrm{diag}_a[\nabla^2 L(\pi_\varepsilon(R,z))]$. Note that $\widehat{R}^+ \in \mathcal{S}^m_{++}$ is the *Schur complement* of R_{++} in R (see for example [94]). Substituting the expression of $v^+_\varepsilon(R,z)$ given in (1.49) in the second equation of (1.48) yields

$$A_\varepsilon(R,z)v^a_\varepsilon(R,z) = (Rw)^a - R_{a+}R^{-1}_{++}(Rw)^+ + O(\varepsilon).$$

On the other hand, since $\lambda_{min}(A_\varepsilon(R,z)) = \inf\{v^\top A_\varepsilon(R,z)v \ ; |v| = 1\}$, we have that

$$\lambda_{min}(A_\varepsilon(R,z)) \geq \lambda_{min}(\widehat{R}^+) + \min_{i\in\bar{I}^a} \varepsilon\ell'\left(\pi^i_\varepsilon(R,z)\right)\frac{\ell''\left(\pi^i_\varepsilon(R,z)\right)}{\ell'\left(\pi^i_\varepsilon(R,z)\right)}. \tag{1.50}$$

Assumption (1.10)(II), Lemma 26(ii) and (1.50) imply that $||A^{-1}_\varepsilon(R,z)|| \mapsto 0$ uniformly in $\mathcal{V}$. Thus, we obtain that $v^a_\varepsilon(R,z) \to 0 = D_z\pi^a_0(R,z)w$ uniformly in $|w| = 1$ and $(R,z) \in \mathcal{V}$. Finally, equation (1.49) yields that $v^+_\varepsilon(R,z) \to R^{-1}_{++}(Rw)^+$, also uniformly in $|w| = 1$ and $(R,z) \in \mathcal{V}$. Thus, the conclusion follows from (1.46).

(ii) By assertion (i) and Proposition 23 (iii), (iv), we have that

$$D_R\pi_\varepsilon(R,z) \to D_z\pi_0(R,z)R^{-1}V\left(z - \pi_0(R,z)\right) \quad \text{uniformly for } (R,z) \in \mathcal{V}.$$

Therefore, we have that $D\pi_\varepsilon(\cdot,\cdot)$ converges locally uniformly and since $\pi_\varepsilon(\cdot,\cdot)$ converges to $\pi_0(\cdot,\cdot)$ uniformly in $\mathcal{V}$, we conclude (cf. [32] Theorem 3.6.1) that $D\pi_\varepsilon(\cdot,\cdot) \to D\pi_0(\cdot,\cdot)$, from which the result follows.

(iii) Follows in a manner analogous to that in the proof of Proposition 23(v).

■

We end this section with an elementary lemma that gives a geometrical meaning to the assumptiom of strict complementarity (see Theorems 30 and 35 in the next section).

Lemma 29 (Strict complementarity reformulation) *Consider the problem*

$$\min\left\{\tfrac{1}{2}x^\top Rx + c^\top x + d : x \in \mathbb{R}^n_+\right\},$$

where R, c, d belong respectively to $\mathcal{S}^m_{++}$, $\mathbb{R}^m$ and $\mathbb{R}$. The optimal solution of this problem satisfies the strict complementarity conditions if and only if $-R^{-1}c \notin \mathrm{sing}(R)$.

Proof. We have $\frac{1}{2}x^\top Rx + c^\top x + d = \frac{1}{2}(x + R^{-1}c)^\top R(x + R^{-1}c) - \frac{1}{2}c^\top R^{-1}c + d$. Thus, the solution of the above problem is $\pi_0(R, -R^{-1}c)$ and the result follows by Lemma 25 (iv). ■

1.4 Main results

The notation are those of the previous section. Let $\varepsilon \in [0,\infty)$, recall that by equations (1.9) and (1.23) the solution u_ε of $(\mathcal{CP}_\varepsilon)$ is given by

$$u_\varepsilon(t) = \pi_\varepsilon(R(t), -R(t)^{-1}B(t)^\top p_\varepsilon(t)) \quad \text{for a.a. } t \in [0,T]. \tag{1.51}$$

Note that the curve $(y_\varepsilon, p_\varepsilon)$ belong to $W^{1,s} \times W^{1,s}$ and hence the optimal control u_ε is *continuous*. Consequently the optimal control u_ε satisfies

$$u_\varepsilon(t) = \mathrm{argmin}\left\{H_\varepsilon(w, y_\varepsilon(t), p_\varepsilon(t), t) : w \geq 0\right\} \quad \textit{for all } t \in [0,T].$$

1.4.1 Error estimates for interior penalties

Let us now introduce our main assumption.

Strict complementarity assumption: *There exists a subset* $T_{\rm sing}$ *of* $[0,T]$ *with meas*$(T_{\rm sing})=0$*, such that for each* t *in* $[0,T]\backslash T_{\rm sing}$ *the point* $u_0(t)$ *satisfies the strict complementarity conditions for the minimization problem*

$$\min\left\{H_0(w,y_0(t),p_0(t),t) : w\in\mathbb{R}^n_+\right\}.$$

This assumption can be reformulated in an alternative form. Note first that for almost all t, the control $u_0(t)$ actually solves the following (simplified) quadratic problem: $\min\left\{v^\top R(t)v + p_0(t)^\top B(t)v : v\in\mathbb{R}^m_+\right\}$. As in Lemma 29, define

$$q_0(t) := -R(t)^{-1}B(t)^\top p_0(t), \tag{1.52}$$

where p_0 is the adjoint state for problem $(\mathcal{CP}_0)$. In view of Lemma 29, the strict complementarity assumption above exactly amounts to

$$\text{meas}\{t\in[0,T] \ : \ q_0(t)\in \text{sing}(R(t))\} = 0. \tag{1.53}$$

$\mathbf{W^{1,\infty}}$ assumption: We shall say that $W^{1,\infty}$ *assumption holds* if:

$$\left\{\begin{array}{l} R\in W^{1,\infty}([0,T];\mathcal{S}^m_{++}), \quad C\in W^{1,\infty}([0,T];\mathcal{S}^n_+).\\ A\in W^{1,\infty}([0,T];\mathbb{R}^{n\times n}),\ B\in W^{1,\infty}([0,T];\mathbb{R}^{n\times m}).\end{array}\right. \tag{1.54}$$

Clearly, under this assumption, $u_\varepsilon\in W^{1,\infty}$ for all $\varepsilon\geq 0$.

For $\varepsilon\geq 0$ define $\Pi_\varepsilon : W^{1,s}\to L^s$ by

$$\Pi_\varepsilon(w)(t) := \pi_\varepsilon(R(t),w(t)). \tag{1.55}$$

In view of Proposition 23 this function is well defined. For each fixed t, the quantity $|\Pi_\varepsilon(w)(t)-\Pi_0(w)(t)|$ therefore measures the error estimate of the penalty method for the finite dimensional problem

$$\min\left\{(x-w(t))^\top R(t)(x-w(t)) : x\in\mathbb{R}^m_+\right\}.$$

The following result shows that these finite dimensional error bounds can be used to recover the error bounds for the penalized optimal control problem $(\mathcal{CP}_\varepsilon)$.

Theorem 30 (Error estimates for interior penalty) *Let* s *be in* $[1,+\infty)$ *and suppose that* ψ *and* φ *belong to* L^s*. Assume further that the strict complementarity assumption* (1.53) *and the* $W^{1,\infty}$ *assumption* (1.54) *hold. Then,*

for ε close to 0 we have that:
(i) *For $1 \leq s' \leq s$, the error estimates for $u_\varepsilon, y_\varepsilon$ and p_ε are given by*

$$||u_\varepsilon - u_0||_{s'} + ||y_\varepsilon - y_0||_{1,s'} + ||p_\varepsilon - p_0||_{1,s'} = O\left(||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_{s'}\right), \quad (1.56)$$

with in addition $u_\varepsilon \to u_0$ in $W^{1,s}$.
(ii) *The error bound for the control with respect to the supremum norm is given by*

$$||u_\varepsilon - u_0||_\infty = O\left(||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_\infty\right). \quad (1.57)$$

(iii) *The error estimate for the cost is given by*

$$|J_0(u_\varepsilon) - J_0(u_0)| = O\left(||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_1\right). \quad (1.58)$$

Remark 31 *Note that the quality of the approximation in* (i) *depends on the regularity of φ and ψ. Since $s \geq 1$, we always have that φ and ψ belong to L^1 and estimate*

$$||u_\varepsilon - u_0||_1 + ||y_\varepsilon - y_0||_{1,1} + ||p_\varepsilon - p_0||_{1,1} = O\left(||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_1\right) \quad (1.59)$$

always holds. On the other hand, if φ and ψ belong to L^∞ we have, for all $s \in [1, \infty)$,

$$||u_\varepsilon - u_0||_s + ||y_\varepsilon - y_0||_{1,s} + ||p_\varepsilon - p_0||_{1,s} = O\left(||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_s\right). \quad (1.60)$$

From now on, we assume that the hypotheses of Theorem 30 hold. For the proof of that result, we begin by introducing the map

$$F : W^{1,s} \times W^{1,s} \times \mathbb{R}_+ \to L^s \times \mathbb{R}^n \times L^s \times \mathbb{R}^n$$

defined by

$$F(y, p, \varepsilon)(\cdot) := \begin{pmatrix} \dot{y}(\cdot) - A(\cdot)y(\cdot) - B(\cdot)\pi_\varepsilon\left(R(\cdot), -R(\cdot)^{-1}B(\cdot)^\top p(\cdot)\right) - \psi(\cdot) \\ y(0) - x_0 \\ \dot{p}(\cdot) + A(\cdot)^\top p(\cdot) + C(\cdot)(y(\cdot) - \bar{y}(\cdot)) + \varphi(\cdot) \\ p(T) - M[y(T) - \bar{y}(T)] \end{pmatrix}. \quad (1.61)$$

The optimality system of problem $(\mathcal{CP}_\varepsilon)$ may be therefore expressed as

$$F(y_\varepsilon, p_\varepsilon, \varepsilon) = 0 \quad \text{for every } \varepsilon \geq 0. \quad (1.62)$$

Remark 32 In general, F is not differentiable at $(y_0, p_0, 0)$. Indeed, take $m = n = 1$, $R(t) \equiv 1$, $B(t) \equiv 1$, $L(x) = -\log x$. In this case, for $p_0 \in W^{1,s}$ and $\varepsilon \geq 0$, it holds that $\pi_\varepsilon(1, p_0) = \varphi_\varepsilon(-p_0)$ where

$$\varphi_\varepsilon(x) := \tfrac{1}{2}\left(x + \sqrt{x^2 + 4\varepsilon}\right). \quad (1.63)$$

For every $t \in [0,T]$ it holds that

$$\lim_{\varepsilon \downarrow 0} \frac{\pi_\varepsilon(1,p_0(t)) - \pi_0(1,p_0(t))}{\varepsilon} = \begin{cases} \frac{1}{|p_0(t)|} & \text{if } p_0(t) \neq 0 \\ +\infty & \text{if } p_0(t) = 0 \end{cases} \tag{1.64}$$

and generally, this limit does not belong to L^s.

In view of the above remark, a direct application of the Implicit Function Theorem to (1.62) is not possible. Instead, we will use the so-called Restoration Theorem (see [3] and the Appendix), which is a variant of the standard Surjective Mapping Theorem of Graves (see [49]). In the following two lemmas we show that, under very general conditions, the assumptions of the Restoration Theorem are fulfilled.

Lemma 33 (Strict uniform differentiability) *Let $s \in [1,+\infty[$ and $\widehat{w} \in W^{1,s}$ be such that*

$$meas\{t \in [0,T] \ : \ \widehat{w}(t) \in \mathrm{sing}(R(t))\} = 0, \tag{1.65}$$

where the set $\mathrm{sing}(R)$ *is defined in* (1.37). *Then :*
(i) *For every $\varepsilon > 0$, $w \in W^{1,s}$, the function Π_ε is differentiable at w and for every $h \in W^{1,s}$ we have that*

$$(D\Pi_\varepsilon(w)h)(t) = D_z\pi_\varepsilon(R(t),w(t))h(t), \ \textit{for a.a. } t \in (0,T).$$

(ii) *The function Π_0 is differentiable at $\widehat{w} \in W^{1,s}$ and for every $h \in W^{1,s}$*

$$(D\Pi_0(\widehat{w})h)(t) = D_z\pi_0(R(t),\widehat{w}(t))h(t), \quad \textit{for a.a. } t \in (0,T).$$

(iii) *There exists a nondecreasing function $c : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{\beta \downarrow 0} c(\beta) = 0$ such that: For any $w', w \in W^{1,s}$ with $||w' - \widehat{w}||_{1,s} \leq \beta$, $||w - \widehat{w}||_{1,s} \leq \beta$ and $\varepsilon \in [0,\beta]$ we have*

$$||\Pi_\varepsilon(w') - \Pi_\varepsilon(w) - D\Pi_0(\widehat{w})(w'-w)||_s \leq c(\beta)||w'-w||_{1,s}. \tag{1.66}$$

Proof. (i) Follows directly from the implicit function theorem.
(ii) For $h \in W^{1,s}$ and $t \in [0,T]$ denote

$$\vartheta(h)(t) := |\pi_0(R(t),\widehat{w}(t)+h(t)) - \pi_0(R(t),\widehat{w}(t)) - D_z\pi_0(R(t),\widehat{w}(t))h(t)|^s.$$

We have

$$\frac{1}{||h||^s_{1,s}}||\Pi_0(\widehat{w}+h) - \Pi_0(\widehat{w}) - D_z\pi_0(R(\cdot),\widehat{w})h||^s_s = \frac{1}{||h||^s_{1,s}} \int_0^T \vartheta(h)(t)\mathrm{d}t.$$

Since $W^{1,s}$ is continuously embedded in L^∞, there exists $c_s > 0$ such that

$$|h(t)| \le ||h||_\infty \le c_s ||h||_{1,s} \quad \text{for a.a. } t \in [0,T]. \tag{1.67}$$

It follows that

$$\frac{1}{||h||_{1,s}^s} ||\Pi_0(\widehat{w}+h) - \Pi_0(\widehat{w}) - D_z\pi_0(R(t),\widehat{w})h||_s^s \le c_s \int_0^T \frac{\vartheta(h)(t)}{|h(t)|^s} \mathrm{d}t.$$

By using Lemma 22 with $\varepsilon = 0$, it follows that $\vartheta(h)(t)/|h(t)|^s$ is uniformly bounded for $||h||_{1,s} \le 1$ and $t \in [0,T]$. Also, by Lemma 25, $\pi_0(R(t),\cdot)$ is differentiable at $\widehat{w}(t)$ iff $\widehat{w}(t) \notin \mathrm{sing}(R(t))$. Thus, in view of hypothesis (1.65),

$$\frac{\vartheta(h)(t)}{|h(t)|^s} \to 0 \quad \text{for a.a. } t \in [0,T],$$

and the result follows by Lebesgue's dominated convergence theorem.
(iii) Let us first observe that

$$||\Pi_\varepsilon(w') - \Pi_\varepsilon(w) - D\Pi_0(\widehat{w})(w'-w)||_s =$$

$$\left|\left|\left(\int_0^1 [D\Pi_\varepsilon(w+\tau(w'-w)) - D\Pi_0(\widehat{w})]\,\mathrm{d}\tau\right)(w'-w)\right|\right|_s$$

$$\le \sup_{z \in B_{1,s}(\widehat{w},\beta)} ||D\Pi_\varepsilon(z) - D\Pi_0(\widehat{w})||_{W^{1,s}\to L^s} \, ||w'-w||_{1,s},$$

where $B_{1,s}(\widehat{w},\beta)$ denotes the ball in $W^{1,s}$ of center $\widehat{w}$ and radius β and $||\cdot||_{W^{1,s}\to L^s}$ denotes the standard norm of the space of linear bounded functions from $W^{1,s}$ to L^s. Let $h \in W^{1,s}$ with $||h||_{1,s} \le 1$. For every $z \in B_{1,s}(\widehat{w},\beta)$ we have that

$$||D\Pi_\varepsilon(z)h - D\Pi_0(\widehat{w})h||_s^s \le \|h\|_\infty^s \int_0^T |D\pi_\varepsilon(R(t),z(t)) - D\pi_0(R(t),\widehat{w}(t))|^s \mathrm{d}t$$

and thus, in view of (1.67) and that $||h||_{1,s} = 1$,

$$\sup_{z \in B_{1,s}(\widehat{w},\beta)} ||D\Pi_\varepsilon(z) - D\Pi_0(\widehat{w})||_{W^{1,s}\to L^s} \le c(\beta),$$

where $c(\beta)$ is defined by

$$c(\beta) := c_s \left(\int_0^T \sup_{\varepsilon\in[0,\beta]} \sup_{z \in B(\widehat{w}(t),\beta)} |D_z\pi_\varepsilon(R(t),z) - D_z\pi_0(R(t),\widehat{w}(t))|^s \mathrm{d}t \right)^{\frac{1}{s}}.$$

In light of Proposition 23 (ii), Proposition 28 (iii), assumption (1.65) and Lebesgue's dominated convergence theorem, we conclude that $c(\beta) \downarrow 0$ as $\beta \downarrow 0$. ■

The following result establishes the surjectivity of the derivative of F at $(y_0, p_0, 0)$ (where F is defined in (1.61)): this fact is central for the application of the restoration theorem (see Theorem 43). Define

$$\Sigma(t) := \{1, ..., m\} \setminus I^+(R(t), q_0(t)), \quad \text{for all } t \in [0, T] \tag{1.68}$$

and recall that for all $\Sigma \subseteq \{1, ..., m\}$ the linear subspace Q_Σ was defined in (1.34).

Lemma 34 (Surjectivity of F) *Consider problems $(\mathcal{CP}_0)$ and $(\mathcal{CP}_\varepsilon)$ of Section 2. If the strict complementarity assumption (1.53) holds, then the function F is differentiable with respect to (y, p) at $(y_0, p_0, 0)$ and the linear application $D_{(y,p)}F(y_0, p_0, 0)$ is an isomorphism. In addition, for every $(\delta_1, \delta_2, \delta_3, \delta_4) \in L^s \times \mathbb{R}^n \times L^s \times \mathbb{R}^m$, the curve*

$$D_{(y,p)}F(y_0, p_0, 0)^{-1}(\delta_1, \delta_2, \delta_3, \delta_4)$$

is the unique solution of the reduced optimality system of

$$\left\{ \begin{array}{ll} \text{Min } \frac{1}{2} \displaystyle\int_0^T \left(v(t)^\top R(t) v(t) + \sigma(t)^\top C(t) \sigma(t) - \delta_3 \cdot \sigma(t) \right) \mathrm{d}t & \\ \qquad\qquad + \frac{1}{2} (\sigma(T) + M^{-1}\delta_4)^\top M (\sigma(T) + M^{-1}\delta_4), & \\ s.t. \quad \dot{\sigma}(t) = A(t)\sigma(t) + B(t)v(t) + \delta_1(t), & \\ \qquad \sigma(0) = \delta_2, \quad v(t) \in Q_{\Sigma(t)}. & \end{array} \right. \tag{$\mathcal{P}_{\delta_1,\delta_2,\delta_3,\delta_4}$}$$

Proof. The differentiability property of F is a direct consequence of Lemma 33 (ii). Now, for σ and ς in $W^{1,s}$ we have

$$D_{(y,p)}F(y_0, p_0, 0)(\sigma, \varsigma)\ (\cdot) = \left(\begin{array}{c} D_{(y,p)}F^1(y_0, p_0, 0)(\sigma, \varsigma) \\ \sigma(0) \\ \dot{\varsigma}(\cdot) + A(\cdot)^\top \varsigma(\cdot) + C(\cdot)\sigma(\cdot) \\ \varsigma(T) - M\sigma(T) \end{array} \right),$$

where

$$D_{(y,p)}F^1(y_0, p_0, 0)(\sigma, \varsigma) = \dot{\sigma}(\cdot) - A(\cdot)\sigma(\cdot) + B(\cdot) D_z \pi_0 \left(R(\cdot), q_0(\cdot)\right) R(\cdot)^{-1} B(\cdot)^\top \varsigma(\cdot).$$

Let $\delta_1 \in L^s$, $\delta_2 \in \mathbb{R}^n$, $\delta_3 \in L^s$, $\delta_4 \in \mathbb{R}^n$ and consider the system of equations

$$\begin{array}{rcl} \dot{\sigma}(t) - A(t)\sigma(t) + B(t) D_z \pi_0 \left(R(t), q_0(t)\right) R(t)^{-1} B(t)^\top \varsigma(t) & = & \delta_1(t), \\ \dot{\varsigma}(t) + A(t)^\top \varsigma(t) + C(t)\sigma(t) & = & \delta_3(t), \\ \varsigma(T) - M\sigma(T) = \delta_4 \quad ; & & \sigma_0 = \delta_2. \end{array} \tag{1.69}$$

Note that, by Lemma 25(iii), the vector

$$D_z\pi_0\left(R(t),q_0(t)\right)\left[-R(t)^{-1}B(t)^\top\varsigma(t)\right]$$

is the projection of $-R(t)^{-1}B(t)^\top\varsigma(t)$, with respect to the metric induced by $R(t)$, into $Q_{\Sigma(t)}$. Using this fact it is routine to verify that equations (1.69) are the reduced first-order optimality conditions of $(\mathcal{P}_{\delta_1,\delta_2,\delta_3,\delta_4})$. Arguments similar to those already used for the problem $(\mathcal{CP}_0)$ show that $(\mathcal{P}_{\delta_1,\delta_2,\delta_3,\delta_4})$ has a unique solution, which concludes the proof. ■

Now we are in a position to give a proof of Theorem 30.

Proof of Theorem 30. Since L^s is continuously embedded in $L^{s'}$ it suffices to prove the result for $s'=s$. First, for $\varepsilon>0$ let us define

$$q_\varepsilon(t):=-R(t)^{-1}B(t)^\top p_\varepsilon(t)\quad\text{for all } t\in[0,T]. \tag{1.70}$$

(i) Let us first note that

$$\begin{aligned} F(y_0,p_0,\varepsilon)(t) &= F(y_0,p_0,\varepsilon)(t)-F(y_0,p_0,0)(t),\\ &= \left(-B(t)\left[\pi_\varepsilon(R(t),q_0(t))-\pi_0(R(t),q_0(t))\right],0,0,0\right)^\top. \end{aligned} \tag{1.71}$$

In view of Lemma 33 and Lemma 34 the mapping F defined in (1.61)(page 50), satisfies the assumptions of the Restoration Theorem (see the Appendix). Therefore, by (1.71) and the definition (1.55) of Π_ε,

$$||y_\varepsilon-y_0||_{1,s}+||p_\varepsilon-p_0||_{1,s}=O\left(||\Pi_\varepsilon(q_0)-\Pi_0(q_0)||_s\right)$$

On the other hand, for every $t\in[0,T]$ we have

$$\begin{aligned} |u_\varepsilon(t)-u_0(t)| &= |\pi_\varepsilon(R(t),q_\varepsilon(t))-\pi_0(R(t),q_0(t))|\\ &\le |\pi_\varepsilon(R(t),q_\varepsilon(t))-\pi_\varepsilon(R(t),q_0(t))|\\ &\quad +|\pi_\varepsilon(R(t),q_0(t))-\pi_0(R(t),q_0(t))| \end{aligned}$$

Therefore, Lemma 22 implies that

$$|u_\varepsilon(t)-u_0(t)|\le\kappa(R(t))|q_\varepsilon(t)-q_0(t)|+|\pi_\varepsilon(R(t),q_0(t))-\pi_0(R(t),q_0(t))| \tag{1.72}$$

and the first assertion follows by taking the L^s norm.

Let us prove the second assertion. Since the convergence of u_ε to u_0 in L^s is already established, it suffices to prove the convergence in L^s of the derivatives. For almost all $t\in[0,T]$, we have that

$$|\dot u_\varepsilon(t)-\dot u_0(t)|\le|\Delta_1(t)|+|\Delta_2(t)|$$

where

$$\begin{aligned}\Delta_1(t) &:= [D_R\pi_\varepsilon(R(t),q_\varepsilon(t)) - D_R\pi_0(R(t),q_0(t))]\,\dot{R}(t)\\ \Delta_2(t) &:= D_z\pi_\varepsilon(R(t),q_\varepsilon(t))\dot{q}_\varepsilon(t) - D_z\pi_0(R(t),q_0(t))\dot{q}_0(t).\end{aligned}$$

The convergence of Δ_1 to 0 in L^s follows from Proposition 28 (ii) and Lebesgue dominated convergence theorem. As for Δ_2, let us first rewrite $\Delta_2(t)$ as

$$D_z\pi_\varepsilon(R(t),q_\varepsilon(t))\,[\dot{q}_\varepsilon(t)-\dot{q}_0(t)]+D_z\pi_\varepsilon(R(t),q_\varepsilon(t))\dot{q}_0(t)-D_z\pi_0(R(t),q_0(t))\dot{q}_0(t)$$

and apply Proposition 28 (i) and Lebesgue theorem.
(ii) Equation (1.72) implies that

$$||u_\varepsilon-u_0||_\infty \le \sup_{t\in[0,T]}\kappa(R(t))||q_\varepsilon-q_0||_\infty + ||\Pi_\varepsilon(q_0)-\Pi_0(q_0)||_\infty. \qquad (1.73)$$

From (i) we obtain that

$$\begin{aligned}||q_\varepsilon-q_0||_\infty = O\left(||p_\varepsilon-p_0||_{1,s}\right) &= O\left(||\Pi_\varepsilon(q_0)-\Pi_0(q_0)||_s\right)\\ &= O\left(||\Pi_\varepsilon(q_0)-\Pi_0(q_0)||_\infty\right),\end{aligned}$$

which concludes the proof in view of (1.73).
(iii) As in the proof of Proposition 18(ii) we have that $|J_0(u_\varepsilon)-J_0(u_0)| = O\left(||u_\varepsilon-u_0||_1\right)$. The result follows by taking $s'=1$ in (i). Thus the proof of Theorem 30 is complete. □

1.4.2 Asymptotic expansion

Now we present our second result, which is based on Corollary 45 of the Restoration Theorem (see the Appendix). This provides asymptotic expansions for the state and the adjoint state of the penalized problems around the state and adjoint state of the original problem.

Theorem 35 (Asymptotic expansion) *Assume that ψ and φ belong to L^s where $s\in[1,+\infty)$. Suppose that the strict complementarity assumption* (1.53) *holds. Then*

$$\begin{pmatrix} y_\varepsilon \\ p_\varepsilon \end{pmatrix} = \begin{pmatrix} y_0 \\ p_0 \end{pmatrix} - D_{(y,p)}F(y_0,p_0,0)^{-1}F(y_0,p_0,\varepsilon) + r(\varepsilon),$$

where

$$r(\varepsilon) = o(||F(y_0,p_0,\varepsilon)||_s).$$

Moreover the first term of the expansion $-D_{(y,p)}F(y_0,p_0,0)^{-1}F(y_0,p_0,\varepsilon)$ *is the unique solution to*

$$\begin{cases} \text{Min } \frac{1}{2}\displaystyle\int_0^T \left(v(t)^\top R(t)v(t) + \sigma(t)^\top C(t)\sigma(t)\right) \mathrm{d}t + \frac{1}{2}\sigma(t)^\top M\sigma(t), \\ s.t. \\ \dot{\sigma}(t) = A(t)\sigma(t) + B(t)v(t) + B(t)\left[\pi_\varepsilon(R(t),q_0(t)) - \pi_0(R(t),q_0(t))\right], \\ \sigma(0) = 0, \quad v(t) \in Q_{\Sigma(t)}. \end{cases}$$

Proof. Since for every $t \in [0,T]$

$$F(y_0,p_0,\varepsilon)(t) = \left(-B(t)\left[\pi_\varepsilon(R(t),q_0(t)) - \pi_0(R(t),q_0(t))\right], 0, 0, 0\right)^\top,$$

the result follows directly from Corollary 45 (see the Appendix), taking $\varepsilon = \beta$, and Lemma 34 taking $\delta_1 = B(t)\left[\pi_\varepsilon(R(t),q_0(t)) - \pi_0(R(t),q_0(t))\right]$, $\delta_2 = 0$, $\delta_3 = 0$ and $\delta_4 = 0$. ■

1.5 Examples

As the following examples show, Theorem 30 can be used to reduce the estimate of error bounds of an optimal control problem to standard computations used in mathematical programming.

1.5.1 Decoupled case: $R(t) \equiv I$

Since R is no longer a variable, we simply write $\pi_\varepsilon(z)$ for $\pi_\varepsilon(R,z)$. In this case one has

$$D\pi_\varepsilon(z) = \left(I + \varepsilon\nabla^2 L(\pi_\varepsilon(z))\right)^{-1} \succ 0. \tag{1.74}$$

Since $(\pi_0(z))_i = \max\{0, z_i\}$ for all $i \in \{1,...,m\}$, we have

$$I^+(I,z) = \{i \in \{1,...,m\} \ : \ z_i > 0\} \quad ; \quad \begin{array}{l} I^a(I,z) = \{i \in \{1,...,m\} \ : \ z_i < 0\}; \\ I^0(I,z) = \{i \in \{1,...,m\} \ : \ z_i = 0\}. \end{array}$$

Clearly $D\pi_\varepsilon(z)$ is a positive-definite diagonal matrix. Therefore, for every $i \in \{1,...,m\}$ the function $(\pi_\varepsilon)_i$ is nondecreasing with respect to z_i and constant with respect to z_j for $j \neq i$. This implies that

$$|\pi^i_\varepsilon(z) - \pi^i_0(z)| = |\pi^i_\varepsilon(z)| \leq |\pi^i_\varepsilon(0)| \quad \text{for all } z \in \mathbb{R}^m, \ i \in I^a(z) \cup I^0(z). \tag{1.75}$$

On the other hand, equations (1.27) and (1.26) give

$$\pi^+_\varepsilon(z) + \varepsilon\nabla L(\pi^+_\varepsilon(z)) = z^+ \quad ; \quad \pi^+_0(z) = z^+$$

and so $D(\pi_\varepsilon^+ - \pi_0^+)(z) = -\varepsilon\nabla^2 L(\pi_\varepsilon^+(z))D\pi_\varepsilon^+(z) \leq 0$. Therefore

$$|\pi_\varepsilon^+(z) - \pi_0^+(z)| \leq |\pi_\varepsilon^+(0) - \pi_0^+(0)| = |\pi_\varepsilon^+(0)|. \tag{1.76}$$

Finally, Theorem 30 (ii) together with equations (1.75) and (1.76) imply that

$$||u_\varepsilon - u_0||_\infty = O\left(|\pi_\varepsilon(0)|\right). \tag{1.77}$$

Let us now compute $|\pi_\varepsilon(0)|$ for some specific barriers.

1.5.1.1 Negative power penalty

For the negative power penalty $\ell(x) = x^{-p}$, (with $p > 0$), we obtain $\pi_\varepsilon(0) - p\varepsilon/\pi_\varepsilon(0)^{p+1} = 0$ by taking $z = 0$ in (1.27). Therefore $\pi_\varepsilon(0) = p^{\frac{1}{2+p}}\varepsilon^{\frac{1}{2+p}}\mathbf{1}$. Conclude with (1.77) that

$$||u_\varepsilon - u_0||_\infty = O(\varepsilon^{\frac{1}{2+p}}). \tag{1.78}$$

The next example shows that the logarithmic barrier provides a smaller L^∞ error bound, and even more importantly, a considerably better and sharper bound for the L^1 norm.

1.5.1.2 Logarithmic penalty

The logarithmic penalty corresponds to the choice $\ell(x) = -\log x$. By taking $z = 0$ in (1.27), we get $\pi_\varepsilon(0) - \varepsilon/\pi_\varepsilon(0) = 0$. Therefore $\pi_\varepsilon(0) = \sqrt{\varepsilon}\mathbf{1}$ and, thus (1.77) yields

$$||u_\varepsilon - u_0||_\infty = O(\sqrt{\varepsilon}).$$

Our aim now is to obtain a sharp estimate in L^1 for $u_\varepsilon - u_0$. Note that from (1.27)

$$\pi_\varepsilon^i(z) = \tfrac{1}{2}\left(z^i + \sqrt{(z^i)^2 + 4\varepsilon}\right) = \phi_\varepsilon(z^i) \quad \text{for all } z \in \mathbb{R}^m, \quad i \in \{1, ..., m\}, \tag{1.79}$$

where ϕ_ε is defined as in (1.63). The family $(\phi_\varepsilon)_{0 \leq \varepsilon < \infty}$ enjoy several properties which can be easily established by the reader.

Lemma 36 *For every $\varepsilon > 0$:*
(i) *The function $s \mapsto \phi_\varepsilon(s) - \phi_0(s)$ is even, increasing in $(-\infty, 0)$ (and decreasing in $(0, +\infty)$).*
(ii) *A primitive of ϕ_ε is given by*

$$\Psi_\varepsilon(x) = \frac{1}{4}x^2 + \frac{1}{4}x\sqrt{x^2 + 4\varepsilon} + \varepsilon \log\left(x + \sqrt{x^2 + 4\varepsilon}\right). \tag{1.80}$$

(iii) *For every $s > 0$ and $x \in \mathbb{R}$, $\phi_{s\varepsilon}(sx) = \sqrt{s}\phi_\varepsilon(\sqrt{s}x)$.*

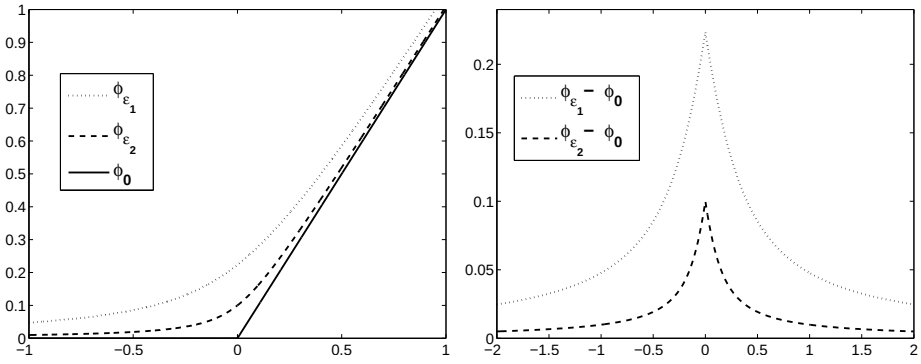

Figure 1.2: Left: ϕ_{ε_1}, ϕ_{ε_2} and ϕ_0. Right: $\phi_{\varepsilon_1} - \phi_0$, $\phi_{\varepsilon_2} - \phi_0$, for $\varepsilon_1 = 0.005$, $\varepsilon_2 = 0.001$.

The following lemma is fundamental for the error estimate in the L^1 norm.

Lemma 37 *Let $q \in C([0,T])$. Assume that $Z(q) := \{t \in [0,T] \ : \ q(t) = 0\}$ is finite and that for every $s_0 \in Z(q)$ the curve q is differentiable at s_0 with $\dfrac{\mathrm{d}q}{\mathrm{d}t}(s_0) \neq 0$. Then*

$$\int_0^T [\phi_\varepsilon(q(t)) - \phi_0(q(t))]\,\mathrm{d}t = O(\varepsilon|\log\varepsilon|). \tag{1.81}$$

Proof. With no loss of generality, let us assume that $Z(q) = \{s_0\}$ and that $\dfrac{\mathrm{d}q}{\mathrm{d}t}(s_0) > 0$. We have $\int_0^T [\phi_\varepsilon(q(t)) - \phi_0(q(t))]\,\mathrm{d}s = A_1 + B_1$, where

$$A_1 = \int_{\{t\ :q(t)>0\}} [\phi_\varepsilon(q(t)) - \phi_0(q(t))]\,\mathrm{d}s \ \ \& \ \ B_1 = \int_{\{t\ :q(t)<0\}} [\phi_\varepsilon(q(t)) - \phi_0(q(t))]\,\mathrm{d}s.$$

Since $\phi_\varepsilon - \phi_0$ is even, it suffices to obtain an estimate for A_1. Note that $\{t \ : q(t) > 0\} = (s_0, T]$ since we are assuming that $Z(q) = \{s_0\}$. Since $\dfrac{dq}{\mathrm{d}t}(s_0) > 0$, there exists $a > 0$ such that $q(s) \geq a(s - s_0) > 0$ for all $s \in [s_0, T]$. On the other hand, by Lemma 36 (i) the function $s \to \phi_\varepsilon(s) - \phi_0(s)$ is decreasing in $]0, +\infty[$ and so

$$A_1 \leq \int_{s_0}^T [\phi_\varepsilon(a(s-s_0)) - \phi_0(a(s-s_0))]\,\mathrm{d}s = \frac{1}{a}\int_0^c (\phi_\varepsilon(s) - s)\,\mathrm{d}s$$

where $c := a(T - s_0)$. By Lemma 36 (ii)

$$\begin{array}{rcl} \int_0^c (\phi_\varepsilon(s) - s)\,\mathrm{d}s & = & -\frac{c^2}{4} + \frac{c}{4}\sqrt{c^2+4\varepsilon} + \varepsilon\log\left(c^2 + \sqrt{c^2+4\varepsilon}\right) - \varepsilon\log\sqrt{4\varepsilon} \\ & \leq & \frac{c}{4}\left(\frac{4\varepsilon}{c+\sqrt{c^2+4\varepsilon}}\right) + O(\varepsilon|\log\varepsilon|) = O(\varepsilon|\log\varepsilon|). \end{array}$$

■

By combining Theorem 30 and Lemma 37, one obtains:

Theorem 38 *Assume that φ and ψ belong to L^s. Consider problems $(\mathcal{CP}_0)$ and $(\mathcal{CP}_\varepsilon)$, with $R(t) \equiv I$ and $\ell(r) = -\log(r)$. Suppose that the strict complementarity conditions* (1.53) *and $W^{1,\infty}$ assumption* (1.54) *hold. Then:*
(i) *We have that*

$$\begin{aligned} ||u_\varepsilon - u_0||_\infty + ||p_\varepsilon - p_0||_{1,\infty} + ||y_\varepsilon - y_0||_{1,\infty} &= O(\sqrt{\varepsilon}), \\ |J_0(u_\varepsilon) - J_0(u_0)| &= O(\sqrt{\varepsilon}). \end{aligned}$$

(ii) *In addition, let us assume that $\{t \in [0,T]\ ;\ q_0(t) \in Sing(I)\}$ is finite and that the following implication holds:*

$$(q_0(t_0))^k = 0 \Rightarrow B \text{ is differentiable at } t_0 \text{ and } \frac{\mathrm{d}}{\mathrm{d}t}(q_0(t_0))^k \neq 0. \qquad (1.82)$$

Then

$$\begin{aligned} ||u_\varepsilon - u_0||_1 + ||p_\varepsilon - p_0||_{1,1} + ||y_\varepsilon - y_0||_{1,1} &= O(\varepsilon|\log\varepsilon|), \\ |J_0(u_\varepsilon) - J_0(u_0)| &= O(\varepsilon|\log\varepsilon|). \end{aligned}$$

Remark 39 The exact computations performed in [3] for a specific problem show that the first bound provided in (ii) is optimal.

1.5.2 Coupled case: $R(t) \succ 0$

Recall that $u_0(t) = \pi_0(R(t), q_0(t))$ for all $t \in [0,T]$. Roughly speaking our hypothesis is that:

– $q_0(t)$ meets the singular region $\mathrm{sing}(R(t))$ a finite numbers of times,

– when the singular region is met at most one inactive (active) constraint can become active (inactive).

This assumption allows, after a localization argument, to bound $|\pi_\varepsilon - \pi_0|$ in terms of $|\phi_\varepsilon - \phi_0|$ (see Subsection 5.1).

Consider again problems $(\mathcal{P}_0^{R,z})$ and $(\mathcal{P}_\varepsilon^{R,z})$ as defined in Section 2. We say that $z \in \mathrm{sing}(R)$ is a *singular point* if $I^0(R,z) \neq \emptyset$. If in addition $I^0(R,z)$ is a singleton we will say that z is a *simple singular point.*

Let $\bar{R} \in \mathcal{S}^m_{++}$ and $k \in \{1,...,m\}$. Consider a simple singular point $\bar{z} \in \mathrm{sing}(\bar{R})$ such that $I^0(\bar{R},\bar{z}) = \{k\}$. Now we proceed to the study of $|\pi_\varepsilon(\cdot,\cdot) - \pi_0(\cdot,\cdot)|$ around $(\bar{R},\bar{z})$. Let $K_1 \times K_2 \subseteq \mathcal{S}^m_{++} \times \mathbb{R}^m$ be a compact neighborhood of $(\bar{R},\bar{z})$ satisfying:

$$\forall (R,z) \in K_1 \times K_2, \quad (z \in K_2 \cap \mathrm{sing}(R) \ \Rightarrow\ I^0(R,z) = \{k\}).$$

In other words the singular points in K_1 are all simple and the active constraint with null multiplier is the same for all of them . The coordinates of $(R,z) \in K_1 \times K_2$ are partitioned according to $I^+(\bar{R},\bar{z})$, $I^a(\bar{R},\bar{z})$ and $I^0(\bar{R},\bar{z}) = \{k\}$. For all $(R,z) \in K_1 \times K_2$, let us define

$$r^k(R,z) := (Rz)^k - R_{k+}R_{++}^{-1}(Rz)^+, \tag{1.83}$$

$$\widehat{R}^k := R_{kk} - R_{k+}R_{++}^{-1}R_{+k}. \tag{1.84}$$

Lemma 40 *Using the notation introduced above, for all $(R,z) \in K_1 \times K_2$ we have:*

$$|\pi_\varepsilon^+(R,z) - \pi_0^+(R,z)| \le C_+ \left[\phi_\varepsilon\left(\frac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right) - \phi_0\left(\frac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right)\right] + O(\varepsilon), \tag{1.85}$$

$$|\pi_\varepsilon^a(R,z) - \pi_0^a(R,z)| = |\pi_\varepsilon^a(R,z)| = O(\varepsilon), \tag{1.86}$$

$$|\pi_\varepsilon^k(R,z) - \pi_0^k(R,z)| = C_k \left[\phi_\varepsilon\left(\frac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right) - \phi_0\left(\frac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right)\right] + O(\varepsilon), \tag{1.87}$$

where

$$C_+ := \frac{||R_{++}^{-1}||\ ||R_{+k}||}{\sqrt{\widehat{R}^k}}, \qquad C_k := \frac{1}{\sqrt{\widehat{R}^k}},$$

and the bounds $O(\varepsilon)$ are uniform on $K_1 \times K_2$.

Proof. Let $(R,z) \in K_1 \times K_2$. Estimate (1.86) is a direct consequence of Lemma 26(ii) using that

$$\log'(\Pi_\varepsilon^i(R,z)) = \frac{1}{\Pi_\varepsilon^i(R,z)} \quad \text{for all } i \in \{1,...,m\}.$$

In view of (1.86) and optimality system (1.27), we have that

$$\begin{aligned} R_{++}\pi_\varepsilon^+(R,z) + R_{+k}\pi_\varepsilon^k(R,z) - \frac{\varepsilon}{\pi_k^+(R,z)} &= (Rz)^+ + O(\varepsilon) \\ R_{k+}\pi_\varepsilon^+(R,z) + R_{kk}\pi_\varepsilon^k(R,z) - \frac{\varepsilon}{\pi_\varepsilon^k(R,z)} &= (Rz)^k + O(\varepsilon), \end{aligned} \tag{1.88}$$

where the bounds $O(\varepsilon)$ are uniform on $K_1 \times K_2$ and correspond to the coordinates in $I^a(\bar{R},\bar{z})$. From the first equation in (1.88) we obtain

$$\pi_\varepsilon^+(R,z) = R_{++}^{-1}\left((Rz)^+ - R_{+k}\pi_\varepsilon^k(R,z)\right) + O(\varepsilon). \tag{1.89}$$

Substituting $\pi_\varepsilon^+(R,z)$ in the second equation of (1.88), we find

$$\widehat{R}^k\pi_\varepsilon^k(R,z) - \frac{\varepsilon}{\pi_\varepsilon^k(R,z)} = r^k(R,z) + O(\varepsilon), \tag{1.90}$$

which is a scalar equation in $\pi_\varepsilon^k(R,z)$. Lemma 36(iii) yields

$$\pi_\varepsilon^k(R,z) = \tfrac{1}{\sqrt{\widehat{R}^k}}\phi_\varepsilon\left(\tfrac{r^k(R,z)}{\sqrt{\widehat{R}^k}} + O(\varepsilon)\right) = \tfrac{1}{\sqrt{\widehat{R}^k}}\phi_\varepsilon\left(\tfrac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right) + O(\varepsilon).$$

Letting $\varepsilon \downarrow 0$ we obtain that $\pi_0^k(R,z) = \tfrac{1}{\sqrt{\widehat{R}^k}}\phi_0\left(\tfrac{r^k(R,z)}{\sqrt{\widehat{R}^k}}\right)$, from which estimate (1.87) follows. Finally, letting $\varepsilon \downarrow 0$ in equation (1.89) yields

$$\pi_0^+(R,z) = R_{++}^{-1}\left((Rz)^+ - R_{+k}\pi_0^k(R,z)\right). \tag{1.91}$$

Thus, estimate (1.85) follows by subtracting equations (1.91), (1.89) and using estimate (1.87). ■

Now we can extend Theorem 38 for the coupled case:

Theorem 41 *Let φ and ψ belong to L^s. Consider problems $(\mathcal{CP}_0)$ and $(\mathcal{CP}_\varepsilon)$ with $\ell(x) = -\log(x)$. Suppose that the strict complementarity conditions (1.53) and $W^{1,\infty}$ assumption (1.54) hold. Also, we assume that*

$$q_0(t_0) \in \mathrm{sing}(R(t_0)) \Rightarrow \quad q_0(t_0) \ \textit{ is a simple singular point.} \tag{1.92}$$

Under these assumptions we have that:
(i) *The following estimates hold:*

$$\begin{aligned} ||u_\varepsilon - u_0||_\infty + ||p_\varepsilon - p_0||_{1,\infty} + ||y_\varepsilon - y_0||_{1,\infty} &= O(\sqrt{\varepsilon}), \\ |J_0(u_\varepsilon) - J_0(u_0)| &= O(\sqrt{\varepsilon}). \end{aligned}$$

(ii) *In addition, let us assume that $\{t \in [0,T]\ ;\ q_0(t) \in \mathrm{sing}(R(t))\}$ is finite and that R, B are differentiable. Suppose that the following implication holds:*

$$I^0(R(t_0), q_0(t_0)) = \{k\} \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} r^k(R(t_0), q_0(t_0)) \neq 0. \tag{1.93}$$

Then, the following estimates hold:

$$\begin{aligned} ||u_\varepsilon - u_0||_1 + ||p_\varepsilon - p_0||_{1,1} + ||y_\varepsilon - y_0||_{1,1} &= O(\varepsilon|\log\varepsilon|), \\ |J_0(u_\varepsilon) - J_0(u_0)| &= O(\varepsilon|\log\varepsilon|). \end{aligned}$$

Remark 42 If $t_0 \in [0,T]$ is such that $I^0(R(t_0), q_0(t_0)) = \{k\}$ then by letting $\varepsilon \downarrow 0$ in (1.90) we see that $r^k(R(t_0), q_0(t_0)) = 0$. Thus assumption (1.93) is an extension of the coupled case (see (1.82)).

Proof. Item (i) is a direct consequence of Theorem 30 and Lemma 40, while item (ii) follows from Theorem 30, Lemma 37 and Lemma 40. ■

1.6 Conclusions

Interior point methods for control constrained optimal control problems have been shown to be very efficient from the practical point of view (see the references given at the introduction), specially when the constraints are penalized with the logarithmic barrier. In this work, for a linear quadratic problem with nonnegativity constraint on the control, we have obtained an explicit expansion for the state and adjoint state, of the penalized problems, around the state and the adjoint state of the main problem. Since the standard implicit function theorem is not applicable to the system of equations associated with the parameterized optimality conditions (see (1.61)), the main results (see Theorems 30 and 35) rely on the Restoration Theorem (see the Appendix), which is a variation of the standard Surjective Mapping Theorem of Graves [49]. The main difficulty in the verification of the assumptions of Theorem 43 comes from the fact that the controls are *coupled* in the cost function through a positive-definite matrix. To overcome this difficulty the thorough analysis of the associated finite dimensional problems (see section 1.3) seems to be unavoidable. It is important to emphasize that the error estimates obtained in Theorem 30, in the different Sobolev norms and for a general class of penalty functions, are derived from a similar analysis in a finite dimensional space. In particular, we obtain (see section 1.5) sharp estimates for the important case of the logarithmic penalty.

An extension of the results of this paper to the case of the optimal control problem of a semilinear elliptic partial differential equation, has been obtained in [2]. As open interesting problems we can mention the computation of the complexity of the method when a self-concordant barrier is considered (in the spirit of [74]), the generalization of the results obtained in this article to the case of state constraints and to the case when the cost and the dynamics are general nonlinear mappings.

Acknowledgements. The authors are indebted to the anonymous referees and specially to the associated editor for various helpful comments that helped to improve the original manuscrip.

Appendix: Restoration Theorem

This material is taken from [3]. Recall that if X and Y are Banach spaces and $A : X \to Y$ is a surjective linear continuous mapping then, by the open mapping theorem, there exists a bounded right inverse of A, which we denote by B, i.e. a (possibly nonlinear) mapping $B : Y \to X$ such that $ABy = y$

for all $y \in Y$, and

$$\|B\| := \sup\{\|By\| : \|y\| \mid y \in Y,\ y \neq 0\} \tag{1.94}$$

is finite.

Theorem 43 (Restoration Theorem) *Let X and Y be Banach spaces, E a metric space and $F : U \subset X \times E \to Y$ a continuous mapping on a nonempty open set U. Let $(\hat{x}, \varepsilon_0) \in U$ be such that $F(\hat{x}, \varepsilon_0) = 0$. Assume that there exists a surjective linear continuous mapping $A : X \to Y$, with bounded right inverse B, and a function $c : \mathbb{R}_+ \to \mathbb{R}_+$ with $c(\beta') \downarrow 0$ when $\beta' \downarrow 0$, such that: if $\beta > 0$ satisfies $c(\beta)||B|| < 1$ and $\varepsilon \in B(\varepsilon_0, \beta)$, then*

$$\|F(x', \varepsilon) - F(x, \varepsilon) - A(x' - x)\| \leq c(\beta)\|x' - x\|, \text{ for all } (x, x') \in \overline{B}(\hat{x}, \beta) \times \overline{B}(\hat{x}, \beta). \tag{1.95}$$

Under the assumptions above, for all (x, ε) close enough to $(\hat{x}, \varepsilon_0)$, there exists $\bar{x}$ such that $F(\bar{x}, \varepsilon) = 0$ and the following estimate holds:

$$\|\bar{x} - x\| \leq \frac{||B||}{1 - c(\beta)||B||}\|F(x, \varepsilon)\|. \tag{1.96}$$

Proof. Let $\rho_0 > 0$ and take $x \in B(\hat{x}, \rho_0)$, $\varepsilon \in \overline{B}(\varepsilon_0, \rho_0)$. By taking $\rho_0 > 0$ small enough, as F is continuous, we may assume that

$$\rho_0 + (1 - L_\beta)^{-1}\|B[F(x, \varepsilon)]\| \leq \beta. \tag{1.97}$$

Let $\{x_n\}$, $n \in \mathbb{N}$, be the sequence defined by $x_0 = x$ and the (modified Newton like) step

$$x_{n+1} = x_n - BF(x_n, \varepsilon). \tag{1.98}$$

Then

$$\|x_{n+1} - x_n\| = \|B[F(x_n, \varepsilon)]\| \leq \|B\|\,\|F(x_n, \varepsilon)\|. \tag{1.99}$$

Relation (1.98) implies

$$F(x_n, \varepsilon) + A(x_{n+1} - x_n) = 0. \tag{1.100}$$

We show by induction that $\{x_n\}$ remains in $\overline{B}(\hat{x}, \beta)$. By (1.97), this is true if $n = 0$. For $n = 1$, we have with (1.99) and (1.97)

$$\|x_1 - \hat{x}\| \leq \|x_1 - x_0\| + \|x_0 - \hat{x}\| \leq \|B[F(x_0, \varepsilon)]\| + \rho_0 \leq \beta.$$

Then if $x_i \in \overline{B}(\hat{x}, \beta)$, for all $1 \leq i \leq n$, (1.95) and (1.100) imply

$$\|F(x_n, \varepsilon)\| \leq c(\beta)\|x_n - x_{n-1}\|. \tag{1.101}$$

Combining with (1.99), we get

$$\|x_{n+1} - x_n\| \leq L_\beta \|x_n - x_{n-1}\| \leq \cdots \leq (L_\beta)^n \|x_1 - x_0\|, \tag{1.102}$$

and hence, with (1.97),

$$\|x_{n+1} - x_0\| \leq (1 - L_\beta)^{-1} \|x_1 - x_0\| \leq (1 - L_\beta)^{-1} \|B[F(x_0, \varepsilon)]\| \leq \beta - \rho_0.$$

Since $\|x_0 - \hat{x}\| < \rho_0$, we deduce that $x_{n+1} \in B(\hat{x}, \beta)$, and hence, the sequence $\{x_n\}$ remains in $B(\hat{x}, \beta)$. With (1.101) and (1.102), we obtain that x_n converges to some $\bar{x}$ such that $F(\bar{x}, \varepsilon) = 0$ and $\|\bar{x} - x_0\| \leq (1 - L_\beta)^{-1} \|B\| \|F(x_0, \varepsilon)\|$, which proves (1.96) with constant η given by

$$\eta = (1 - L_\beta)^{-1} \|B\|. \tag{1.103}$$

□ ■

Remark 44 *The proof of Theorem 43 shows that the assumption that (x, ε) is "close enough to $(\hat{x}, \varepsilon_0)$" can be formulated as: "$x \in B(\hat{x}, \rho_0)$ and $\varepsilon \in \overline{B}(\varepsilon_0, \rho_0)$, where ρ_0 is such that the following inequality holds*

$$\rho_0 + (1 - c(\beta)||B||)^{-1} \|B[F(x, \varepsilon)]\| \leq \beta." \tag{1.104}$$

Now we state an interesting corollary of the Restoration Theorem which is a key tool in the proof of Theorem 35. Its short proof is taken from [3] and is reproduced here for the reader convenience.

Corollary 45 *Assume that the assumptions of Theorem 43 hold and denote by B a bounded right inverse of A. Then, for ε close to ε_0, there exists x_ε in a neighborhood of $\hat{x}$ such that $F(x_\varepsilon, \varepsilon) = 0$ and*

$$x_\varepsilon = \hat{x} - BF(\hat{x}, \varepsilon) + r(\varepsilon), \tag{1.105}$$

where the remainder $r(\varepsilon)$ satisfies

$$\|r(\varepsilon)\| \leq c(\beta)(1 - c(\beta)\|B\|)^{-1} \|B\|^2 \|F(\hat{x}, \varepsilon)\|. \tag{1.106}$$

Proof. Let $\hat{x}(\varepsilon) := \hat{x} - BF(\hat{x}, \varepsilon)$. We have that $F(\hat{x}, \varepsilon) + A(\hat{x}(\varepsilon) - \hat{x}) = 0$ and $\|\hat{x}(\varepsilon) - \hat{x}\| \leq \|B\| \|F(\hat{x}, \varepsilon)\|$. In view of (1.95), $\|F(\hat{x}(\varepsilon), \varepsilon)\| \leq c(\beta)\|B\| \|F(\hat{x}, \varepsilon)\|$. We conclude with Theorem 43. ■

Chapter 2

Optimal control of a semilinear elliptic partial differential equation

Contents

2.1 Introduction

Optimal control of control constrained PDEs is a very rich subject from the theoretical and applied point of view. For an overview of the theory we refer the reader to the classic book [67] and the more recent monographs [43, 65, 55, 73]. Sensitivity analysis as well as second-order conditions have been established in [19, 34, 80].

Numerical methods for these types of problems have been an very active subject of research and we can distinguish two main approaches that are usually referred as direct and indirect methods. Direct methods are those based on the *discretize and then optimize* approach, which means that the infinite dimensional problem is transformed into a finite dimensional one with a very large dimension. Then standard methods of nonlinear programming optimization are used to solve the discretized problem, see for example [4, 5, 33, 40, 69, 68]. In contrast, indirect methods are based on the *optimize and then discretize* approach where optimality conditions are obtained for the infinite dimensional problem and the resulting variational inequalities are discretized, see for example [54, 83, 84].

Interior point methods are among the most popular methods in the indirect approach. They have been investigated, even in the state constraint case [78], extensively in [13, 14, 79, 87, 88]. Specifically, in [79], for box constraints over the control, the optimal solution u_0, with associated state y_0, can be expressed pointwisely as a projection of a linear function of the adjoint state p_0. This enables to avoid the explicit discretization of the control and leads to a very efficient implementation of the method. From the theoretical point of view, the method consists in introducing a family of penalized problems parametrized by $\varepsilon > 0$ whose solution u_ε are *strictly* feasible and studying the convergence of the central path defined by $(y_\varepsilon, p_\varepsilon)$, the state and adjoint state associated with u_ε, towards (y_0, p_0).

Motivated by these works, we consider the optimal control of a semilinear PDE where the control is distributed over the domain Ω and is constrained to be nonnegative. Associated with any isolated solution u_0 we consider a family of localized penalized problems parametrized by $\varepsilon > 0$. We study in detail the relationship between the solution u_ε of the penalized problem and u_0. Our approach is the same that in [2], which was studied in the ODE framework, and consists in obtaining an asymptotic expansion for state y_ε and the adjoint state p_ε, which are associated to u_ε, around the state y_0 and adjoint state p_0, which are associated to u_0. In this sense, our approach is complementary to that in [79] where the slope of the central path, defined by $(y_\varepsilon, p_\varepsilon)$, is integrated in order to obtain error bounds. Under very general hypothesis we can show that $(y_\varepsilon, p_\varepsilon)$ can be expressed as (y_0, p_0) plus a *principal* term

which is characterized as being the state and adjoint state associated to the solution of a *tangent* optimization problem. This fact enable us to obtain, as a corollary, precise error bounds for the central path in various Sobolev norms and for a rather general class of penalty functions.

The paper is organized as follows: In Section 2, after introducing the necessary notations, we state the problem as well as its penalized versions. Regularity results are specified and convergence of the central path is obtained, which allows us to write the solution of the penalized problem in term of its associated adjoint state. This fact will be crucial for Section 3, since the optimality system for the penalized problem can be written in terms of $(y_\varepsilon, p_\varepsilon)$ only. Then we show, by means of a Restoration theorem as in [2] and under very general conditions, that is possible to obtain the desired asymptotic expansion of the central path around (y_0, p_0). We finalize Section 3 by obtaining that error bounds for the infinite dimensional problem, in various norms, can be obtained from its finite dimensional counterparts, generalizing the result of [2]. In particular, for the logarithmic penalty, we recover in Section 4 an error for the control of $O(\sqrt{\varepsilon})$ in the L^∞ norm and under more restrictive hypothesis we improve this bound in the L^2 norm to $O(\varepsilon^{3/4})$. Similar results are obtained for the error of the central path $(y_\varepsilon, p_\varepsilon)$ in the H^2 norm.

2.2 Problem statement and preliminary results

Consider the following semilinear elliptic equation

$$\left\{ \begin{array}{rcll} -\Delta y(x) + \phi(y(x)) & = & g(x) & \text{for} \quad x \in \Omega, \\ y(x) & = & 0 & \text{for} \quad x \in \partial\Omega, \end{array} \right. \tag{2.1}$$

where Ω is a bounded open set of $\mathbb{R}^n$ with C^2 boundary, $g \in L^2(\Omega)$ and ϕ is a nondecreasing real valued function over $\mathbb{R}$, Lipschitz with associated constant L_ϕ and continuously differentiable. Given $s \in [2, \infty]$, denote by $||\cdot||_s$ the standard norm in $L^s(\Omega)$. For $m \in \mathbb{N}$ set

$$W^{m,s}(\Omega) := \{y \in L^s(\Omega) \ ; \ D^\alpha y \in L^s(\Omega) \text{ for } \alpha \text{ such that } |\alpha| \leq m\},$$

where $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}^n$, $|\alpha| := \alpha_1 + ... + \alpha_n$ and

$$D^\alpha := \frac{\partial^{\alpha_1 + ... + \alpha_n}}{\partial x_1^{\alpha_1} ... \partial x_n^{\alpha_n}}$$

represents a derivative operator in the distribution sense. As usual, for $s = 2$ we will write $H^m(\Omega) := W^{m,2}(\Omega)$. It is well know that $W^{m,s}(\Omega)$ endowed

with the norm

$$||y||_{m,s} := \sum_{0\leq|\alpha|\leq m} ||D^\alpha y||_s$$

is a Banach space and $H^m(\Omega)$ endowed with the norm

$$||y||_{m,2} := \left(\sum_{0\leq|\alpha|\leq m} ||D^\alpha y||_2^2 \right)^{1/2}$$

is a Hilbert space. We also denote $W_0^{m,s}(\Omega)$, which will be written as $H_0^m(\Omega)$ when $s = 2$, the space defined as the closure of $\mathcal{D}(\Omega)$ in $W^{m,s}(\Omega)$, where $\mathcal{D}(\Omega)$ denotes the set of C^∞ functions with compact support in Ω. For the reader convenience we recall the following Sobolev embeddings (cf. [1], [42], [47])

$$W^{m,s}(\Omega) \subseteq \begin{cases} L^{q_1}(\Omega) & \text{with } \frac{1}{q_1} = \frac{1}{s} - \frac{m}{n} \quad \text{if } s < \frac{n}{m} \\ L^q(\Omega) & \text{with } q \in [1, +\infty) \quad \text{if } s = \frac{n}{m} \\ C^{m-[\frac{n}{s}]-1,\gamma(n,s)}(\Omega) & \qquad\qquad\quad \text{if } s > \frac{n}{m} \end{cases} \tag{2.2}$$

where $\gamma(n, s)$ is defined as

$$\gamma(n,s) = \begin{cases} [\frac{n}{s}] + 1 - \frac{n}{s}, & \text{if } \frac{n}{s} \notin \mathbb{Z} \\ \text{any positive number} < 1 & \text{if } \frac{n}{s} \in \mathbb{Z} \end{cases} \tag{2.3}$$

and $C^{m-[\frac{n}{s}]-1,\gamma(n,s)}(\Omega)$ denotes the Holder space with exponents $m - [\frac{n}{s}] - 1$ and $\gamma(n, s)$ (for the definition see [42] p. 240). In this work we will use repeatedly the fact that $W^{2,s}(\Omega) \subseteq C(\Omega)$ when $s > n/2$ ($s = 2$ if $n \leq 3$). This is equivalent to the existence of a constant c_s such that

$$||y||_\infty \leq c_s ||y||_{2,s}. \tag{2.4}$$

An space that will play an important role is $\mathcal{Y}^s := W^{2,s}(\Omega) \cap W_0^{1,s}(\Omega)$ which endowed with the norm $||\cdot||_{2,s}$ is a Banach space.

In the following $s \in [2, \infty)$ will be fixed and we will assume, without loss of generality, that $\phi(0) = 0$. We collect in the next proposition some properties of the PDE (2.1) (see for example [19, 29]).

Proposition 46 *If $g \in L^s(\Omega)$ the following holds:*
(i) *The semilinear equation (2.1) has a unique solution $y_g \in \mathcal{Y}^s$ and there exists a constant $\bar{c}_s > 0$ such that*

$$||y_g||_{2,s} \leq \bar{c}_s ||g||_s. \tag{2.5}$$

(ii) *The mapping $g \to y_g$ is continuous from $L^s(\Omega)$ into $\mathcal{Y}^s$, both spaces endowed with the weak topology.*

Proof. (i) Equation (2.1) can be interpreted as the optimality system, in the weak sense, of the variational problem

$$\operatorname*{Min}_{y} \int_\Omega \left\{ \tfrac{1}{2}|\nabla y(x)|^2 + \Phi(y(x)) - g(x)y(x) \right\} \mathrm{d}x \quad \text{subject to } y \in H_0^1(\Omega), \tag{2.6}$$

where $\Phi : [0, +\infty) \to \mathbb{R}$ is defined by $\Phi(t) := \int_0^t \phi(t)$. Since $|\Phi(t)| \leq \frac{1}{2} L_\phi t^2$, the convex mapping $y \in H_0^1(\Omega) \to \int_\Omega \Phi(y(x))\mathrm{d}x \in \mathbb{R}$ is bounded over the bounded sets and whence is continuous. In addition, the cost function is strongly convex and continuous and thus problem (2.6) has a unique solution $y_g \in H_0^1(\Omega)$. Multiplying equation (2.1) by y_g and using Green's formula yields

$$\int_\Omega \left\{ |\nabla y_g(x)|^2 + \phi(y_g(x))y_g(x) \right\} \mathrm{d}x = \int_\Omega g y_g(x) \mathrm{d}x.$$

Since $\phi(y_g)y_g \geq 0$, by the Cauchy-Schwarz and Poincaré inequalities we obtain that

$$||y_g||_{1,2} \leq ||g||_2. \tag{2.7}$$

On the other hand, since $\phi(0) = 0$, it holds that $||\phi(y_g)||_r \leq L_\phi ||y_g||_r$ for all $r \in [1, +\infty)$. Hence, in view of (2.7), an standard boostraping argument yields the existence of $a_s > 0$ such that $||y_g||_s \leq a_s ||g||_s$. Thus $||\Delta y_g||_s \leq (L_\phi a_s + 1)||g||_s$, from which (2.5) follows.
(ii) Let $(g_k)_{k\in\mathbb{N}}$ converge weakly to $\bar{g}$. Then the sequence g_k is bounded in $L^s(\Omega)$ and consequently, by (2.7), the associated states $y_k := y_{g_k}$ are bounded in $\mathcal{Y}^s$. Thus, extracting a subsequence if necessary, y_k converges weakly in $\mathcal{Y}^s$ to some $\bar{y}$ and hence strongly in $L^s(\Omega)$. This implies, since ϕ is Lipschitz, that $\phi(y_k) \to \phi(\bar{y})$ strongly in $L^s(\Omega)$. Passing to the weak limit in $L^s(\Omega)$ in equation (2.1) yields that $\bar{y} = y_g$ from which the conclusion follows. ■

Denote respectively by $\mathbb{R}_+$ and $\mathbb{R}_{++}$ the subsets of nonnegative and positive real numbers. Also, set $\mathcal{U}_+^s := L^s(\Omega; \mathbb{R}_+)$.

Suppose that $g = f + u$, where $f \in L^s(\Omega)$ and $u \in L^2(\Omega)$. By proposition 46 we have that $u \in L^2(\Omega) \to y_{f+u} \in \mathcal{Y}^2$ is well defined. In the following f will be a fixed function and, in order to simplify the notation, we will write y_u for the unique solution in $\mathcal{Y}^2$ of

$$\left\{ \begin{array}{rcll} -\Delta y(x) + \phi(y(x)) & = & f(x) + u(x) & \text{for} \quad x \in \Omega, \\ y(x) & = & 0 & \text{for} \quad x \in \partial\Omega. \end{array} \right. \tag{2.8}$$

Let us define the *cost function* $J_0 : L^2(\Omega) \to \mathbb{R}_+$ by

$$J_0(u) := \tfrac{1}{2} \int_\Omega (y_u(x) - \bar{y}(x))^2 dx + \tfrac{1}{2} N \int_\Omega u(x)^2 \mathrm{d}x, \tag{2.9}$$

where $N > 0$ and $\bar{y} \in L^\infty(\Omega)$ is a reference state function. It holds that:

Corollary 47 *The function $J_0 : L^s(\Omega) \to \mathbb{R}$ is w.l.s.c. (weakly lower semi-continuous).*

Proof. We have that $J_0(\cdot) = \frac{1}{2}||\cdot||_2^2 + \frac{1}{2}N||y_{(\cdot)} - \bar{y}||_2^2$. The map $u \in L^s(\Omega) \to ||u||_2^2$ is convex and continuous therefore is w.l.s.c. In view of proposition 46(ii), and since the inclusion from $W^{2,s}(\Omega)$ into $L^2(\Omega)$ is compact, the function $u \in L^s(\Omega) \mapsto \|y_u - \bar{y}\|_2^2$ is weakly continuous. The result follows. ■

Now, consider the following optimal control problem

$$\text{Min } J_0(u) \quad \text{subject to } u \in \mathcal{U}_+^s. \tag{$\mathcal{CP}_0^s$}$$

By constrast to the case when (2.8) is linear in y (for example when $\phi \equiv 0$), problem $(\mathcal{CP}_0^2)$ is not necessarily convex. Thus, the classical argument to show the existence and uniqueness of a solution of $(\mathcal{CP}_0^2)$ does not apply. Instead, we have the following existence result.

Proposition 48 *Problem $(\mathcal{CP}_0^2)$ has (at least) one solution.*

Proof. Any minimizing sequence u_k for $(\mathcal{CP}_0^2)$ is bounded in $L^2(\Omega)$. Therefore, extracting a subsequence if necessary, we may suppose that it weakly converges to some $u_0 \in L^2(\Omega)$. Since $\mathcal{U}_+^2$ is weakly closed, we have that $u_0 \in \mathcal{U}_+^2$ and, in view of corollary 47 (with $s = 2$), it is a solution of $(\mathcal{CP}_0^2)$. ■

As usual in optimal control theory, it will be convenient to write the derivative of J_0 in terms of an adjoint state. For every $u \in L^2(\Omega)$ the adjoint equation associated with u is defined by

$$\begin{cases} -\Delta p(x) + \phi'(y_u(x))p(x) = y_u(x) - \bar{y}(x) & \text{for} \quad x \in \Omega, \\ p(x) = 0 & \text{for} \quad x \in \partial\Omega. \end{cases} \tag{2.10}$$

It holds that (see [24] lemma 6.18):

Lemma 49 *Let $u \in L^2(\Omega)$. Then the adjoint equation has a unique solution $p_u \in H_0^1(\Omega)$, called the adjoint state associated with u. In addition, the function J_0 is of class C^1 and*

$$DJ_0(u) = p_u + Nu. \tag{2.11}$$

Remark 50 Note that equation (2.10) and the Sobolev embeddings (2.2) imply that $p_u \in \mathcal{Y}^q$ where

$$q = \begin{cases} \frac{2n}{n-4} & \text{if } n > 4, \\ \text{any real number in } [2, \infty) & \text{if } n \leq 4. \end{cases}$$

Now, let u_0 be a solution of $(\mathcal{CP}_0^2)$. In what follows we will write $y_0 := y_{u_0}$ and $p_0 := p_{u_0}$. The first-order condition for the optimality of u_0 is given by

$$DJ_0(u_0)(v - u_0) \geq 0, \quad \text{for all } v \in \mathcal{U}_+^2. \tag{2.12}$$

Expressions (2.11) and (2.12) easily yield that

$$u_0 = P_{\mathcal{U}_+^2}(-N^{-1}p_0), \tag{2.13}$$

where $P_{\mathcal{U}_+^s}$ denotes the orthogonal projection in $L^2(\Omega)$ onto $\mathcal{U}_+^2$. This in turn implies that the following *punctual* relation holds

$$u_0(x) = \pi_0(-N^{-1}p_0(x)) \quad \text{for a.a.} \quad x \in \Omega, \tag{2.14}$$

where for $a \in \mathbb{R}$ we denote $\pi_0(a) := \max\{0, a\}$.

Expression (2.14) allows us, by a bootstrapping argument and using the Sobolev embeddings, to specify the regularity of (y_0, p_0). In fact, proposition 48 implies the following corollary:

Corollary 51 *Problem $(\mathcal{CP}_0^s)$ has (at least) one solution and it holds that:*

$$\begin{array}{rl} y_0 \in \left\{ \begin{array}{ll} L^{q_1}(\Omega) & \text{with } q_1 = \frac{ns}{n-2s} \quad \text{if } s < \frac{n}{2}, \\ L^q(\Omega) & \text{with } q \in [1, +\infty) \quad \text{if } s = \frac{n}{2}, \\ C^{1-[\frac{n}{s}], \gamma(n,s)}(\Omega) & \quad \text{if } s > \frac{n}{2}. \end{array} \right. \\ p_0 \in \left\{ \begin{array}{ll} L^{q_2}(\Omega) & \text{with } q_2 = \frac{ns}{n-4s} \quad \text{if } s < \frac{n}{4}, \\ L^q(\Omega) & \text{with } q \in [1, +\infty) \quad \text{if } s = \frac{n}{4}, \\ C^{1-[\frac{n-2s}{s}], \gamma(n,q_1)}(\Omega) & \quad \text{if } s > \frac{n}{4}. \end{array} \right. \end{array} \tag{2.15}$$

Proof. Let u_0 be a solution of $(\mathcal{CP}_0^2)$. Replacing expression (2.14) into equations (2.8) and (2.10) yields that y_0 and p_0 satisfy

$$\left\{ \begin{array}{rcll} -\Delta y(x) + \phi(y(x)) & = & f(x) + \pi_0(-N^{-1}p_0(x)) & \text{for} \quad x \in \Omega. \\ -\Delta p(x) + \phi'(y_u(x))p(x) & = & y_u(x) - \bar{y}(x) & \text{for} \quad x \in \Omega \\ y(x) = p(x) & = & 0 & \text{for} \quad x \in \partial\Omega \end{array} \right. \tag{2.16}$$

An standard boostraping argument in equations (2.16) implies that $p_0 \in L^{q_2}(\Omega)$ where $q_2 = \frac{ns}{n-4s}$. Since $q_2 > s$, expression (2.14) yields that $u_0 \in L^s(\Omega)$ and therefore solves $(\mathcal{CP}_0^s)$. Regularity results (2.15) follow by (2.2), using that $f + u_0 \in L^s(\Omega)$. ■

Next we consider a *localized* penalized version of $(\mathcal{CP}_0^s)$. Since we could have several solutions of $(\mathcal{CP}_0^s)$, the idea is to localize the problem around

an strict solution (if there is any). Let ℓ be a convex function with domain either $\mathbb{R}_+$ or $\mathbb{R}_{++}$, which is C^2 on the interior of its domain, and satisfies:

$$\begin{array}{l}\text{(i) } \lim_{t\downarrow 0}\ell'(t)=-\infty; \quad \text{(ii) } \lim_{t\downarrow 0}\dfrac{\ell''(t)}{\ell'(t)}=-\infty;\\ \text{(iii) There exist } \alpha\geq 0 \text{ such that } |\ell'(t)|\leq \alpha t \quad \forall\ t\geq 1.\end{array} \tag{2.17}$$

Remark 52 Standard examples of functions satisfying these properties are:

$$\ell(t)=-\log t; \quad \ell(t)=t^{-p}, \quad p>0; \quad \ell(t)=-t^p, \quad p\in(0,1); \quad \ell(t)=t\log t.$$

Let u_0 be a solution of $(\mathcal{CP}_0^s)$. For $b,\varepsilon>0$ the localized *penalized* problem is defined as

$$\text{Min } J_\varepsilon(u):=J_0(u)+\varepsilon\int_\Omega \ell(u(x))\mathrm{d}x \quad \text{subject to } u\in\mathcal{U}_+^s\cap\bar{B}_s(u_0,b) \tag{$\mathcal{CP}_\varepsilon^{b,s}$}$$

where $\bar{B}_s(u_0,b)$ denotes the closed ball in $L^s(\Omega)$ centered at u_0 of radius b. Note that ℓ, being a convex function, is bounded by below by some affine function and thus J_ε takes values in $\mathbb{R}\cup\{+\infty\}$.

Lemma 53 *The function $J_\varepsilon : L^s(\Omega)\to\mathbb{R}$ is w.l.s.c. and problem $(\mathcal{CP}_\varepsilon^{b,s})$ has (at least) one solution.*

Proof. By corollary 47, the function J_0 is w.l.s.c. Adapting the argument of proposition 1 in [2] (which is based in Fatou's lemma), we obtain that $u\in L^s(\Omega)\to\int_\Omega\ell(u(x))\mathrm{d}x$ is convex l.s.c. and hence convex w.l.s.c. which yields the first assertion. Let $u_n\in\mathcal{U}_+^s\cap\bar{B}_s(u_0,b)$ be a minimizing sequence for J_ε. Since $L^s(\Omega)$ is a reflexive Banach space, extracting a subsequence if necessary, there exists $u_\varepsilon\in L^s(\Omega)$ such that $u_n\to u_\varepsilon$ weakly. Clearly, u_ε is feasible for $(\mathcal{CP}_\varepsilon^{b,s})$ and since J_0 is w.l.s.c. it is a solution of the problem. ■

We give here an elementary argument, for the semilinear case, to prove a well known contraction principle which is a corollary of Stampacchia's results (see [81]).

Lemma 54 *There exists a constant $C_1>0$ such that for every $u_1,u_2\in L^s(\Omega)$ we have*

$$||y_{u_1}-y_{u_2}||_1\leq C_1||u_1-u_2||_1. \tag{2.18}$$

Proof. Set $z = y_{u_1} - y_{u_2}$ and $h = u_1 - u_2$. Clearly z satisfies

$$\left\{ \begin{array}{rcll} -\Delta z(x) + \psi_{u_1,u_2}(x) z(x) & = & h(x) & \text{for} \quad x \in \Omega, \\ z(x) & = & 0 & \text{for} \quad x \in \partial\Omega, \end{array} \right. \tag{2.19}$$

where

$$\psi_{u_1,u_2}(x) := \left\{ \begin{array}{ll} \dfrac{\phi(y_{u_2}(x)) - \phi(y_{u_1}(x))}{(y_{u_2} - y_{u_1})(x)}, & \text{if } y_{u_2}(x) \neq y_{u_1}(x), \\ \phi'(y_{u_1})(x), & \text{otherwise.} \end{array} \right. \tag{2.20}$$

Evidently $0 \leq \psi_{u_1,u_2}(x) \leq L_\phi$ for all $x \in \Omega$. Now, let v_z be the unique solution of

$$\left\{ \begin{array}{rcll} -\Delta v_z(x) + \psi_{u_1,u_2}(x) v_z(x) & = & \mathrm{sgn}(z(x)) & \text{for} \quad x \in \Omega \\ v_z(x) & = & 0 & \text{for} \quad x \in \partial\Omega \end{array} \right. \tag{2.21}$$

Multiplying by v_z the first equation in (2.19) and using Green's formula yields that

$$\int_\Omega |z(x)| \mathrm{d}x = \int_\Omega h(x) v_z(x) \mathrm{d}x. \tag{2.22}$$

On the other hand, by the maximum principle for elliptic equations (see for example [30, proposition IX.29]) it holds that $-v_1 \leq v_z \leq v_1$ where $v_1 \geq 0$ solves

$$\left\{ \begin{array}{rcll} -\Delta v_1(x) + \psi_{u_1,u_2}(x) v_1(x) & = & 1 \quad \text{for} & x \in \Omega \\ v_1(x) & = & 0 \quad \text{for} & x \in \partial\Omega. \end{array} \right. \tag{2.23}$$

Using that $\psi \geq 0$ and the maximum principle again, we see that $v_1 \leq v_2$ where v_2 solves

$$\left\{ \begin{array}{rcll} -\Delta v_2(x) & = & 1 \quad \text{for} & x \in \Omega \\ v_2(x) & = & 0 \quad \text{for} & x \in \partial\Omega. \end{array} \right. \tag{2.24}$$

Since v_2 is bounded in $L^\infty(\Omega)$ the result follows from (2.22). ■

The following result yields that the solutions of the penalized problem are bounded in $L^\infty(\Omega)$ by a constant which is independent of ε.

Proposition 55 *Suppose that $s > n/2$ ($s = 2$ if $n \leq 3$) and let u_ε be a solution of $(\mathcal{CP}_\varepsilon^{b,s})$. If ε is small enough, there exists a constant K_ℓ (independent of ε) such that*

$$u_\varepsilon(x) \leq K_\ell \quad \textit{for a.a. } x \in \Omega. \tag{2.25}$$

Proof. For $K > 2||u_0||_\infty$ set

$$\overline{\Omega}_K := \{x \in \Omega;\ u_\varepsilon(x) \geq K\}$$

and

$$u_\varepsilon^K(x) := \begin{cases} K/2 & \text{if } x \in \overline{\Omega}_K \\ u_\varepsilon(x) & \text{otherwise} \end{cases} \quad ; \quad y_\varepsilon^K(x) := y_{u_\varepsilon^K}(x) \quad \text{for a.a. } x \in \Omega. \tag{2.26}$$

Note that u_ε^K is feasible. For all $u \in L^s(\Omega)$ we have (omitting the function arguments in the integral)

$$J_0(u) - J_0(u_\varepsilon) = \tfrac{1}{2} \int_\Omega \{(u + u_\varepsilon)(u - u_\varepsilon) + (y_u + y_\varepsilon - 2\bar{y})(y_u - y_\varepsilon)\} \, \mathrm{d}x. \tag{2.27}$$

Taking $u = u_\varepsilon^K$ in (2.27) we see that, since $s > n/2$ ($s = 2$ if $n \leq 3$) and $u_\varepsilon \in \bar{B}_s(u_0, b)$, proposition 46(i) implies that $y_\varepsilon^K + y_\varepsilon - 2\bar{y}$ is uniformly bounded by a constant independent of ε and K. In addition, by the very definition of $\overline{\Omega}_K$ and u_ε^K, it holds that

$$\left(u_\varepsilon + u_\varepsilon^K\right)\left(u_\varepsilon - u_\varepsilon^K\right) \geq \frac{3}{2} K (u_\varepsilon - u_\varepsilon^K) \mathbf{1}_{\overline{\Omega}_K} \geq 0$$

where $\mathbf{1}_{\overline{\Omega}_K}$ is the indicator function of $\overline{\Omega}_K$. Therefore, in view of lemma 54, we have the existence of $K_2 > 0$ such that

$$J_0(u_\varepsilon) - J_0(u_\varepsilon^K) \geq \left(\frac{3}{4} K + K_2\right) K \, meas(\overline{\Omega}_K). \tag{2.28}$$

Using the convexity of ℓ, we obtain that

$$J_\varepsilon(u_\varepsilon) - J_\varepsilon(u_\varepsilon^K) \geq K \, meas(\overline{\Omega}_K) \left(\frac{3}{4} K + K_2 + \tfrac{1}{2} \varepsilon \ell'(\tfrac{1}{2} K)\right). \tag{2.29}$$

On the other hand, hypothesis (2.17)(iii) implies, for ε small enough, the existence of K_ℓ (independent of ε) such that $\frac{3}{4} K_\ell + K_2 + \frac{1}{2} \varepsilon \ell'(\frac{1}{2} K_\ell) > 0$. Therefore $meas(\overline{\Omega}_{K_\ell}) = 0$ from which the conclusion follows. ■

Let us give an elementary lemma that will be useful in the convergence proof of the central path to the optimal solution (proposition 57). First, define $\bar{F} : \mathcal{Y}^s \times \mathcal{Y}^s \to L^s(\Omega)$ by

$$\bar{F}(y, p) := -\Delta p + \phi'(y) p - y + \bar{y} \tag{2.30}$$

and for every $y \in \mathcal{Y}^s$ denote by $p[y]$ the unique solution of $\bar{F}(y, p) = 0$. It holds that:

Lemma 56 *Suppose that ϕ is C^2 and that $s > n/2$ ($s = 2$ if $n \leq 3$). Then*
(i) *The function $\bar{F}$ is C^1.*
(ii) *The mapping $y \in \mathcal{Y}^s \to p[y] \in \mathcal{Y}^s$ is C^1.*
(iii) *The mapping $u \in L^s(\Omega) \to y_u \in \mathcal{Y}^s$ is C^2.*

Proof. In order to prove (i) it is enough to note that $\phi'(y)p$ is C^1 since ϕ is C^2 and $s > n/2$ ($s = 2$ if $n \leq 3$). Assertions (ii) and (iii) follow directly by the implicit function theorem. ■

For the solutions u_ε of the penalized problems we will write $y_\varepsilon := y_{u_\varepsilon}$ for the state functions and $p_\varepsilon := p_{u_\varepsilon}$ for the adjoint state functions. Now we can state the convergence result.

Proposition 57 *Assume that $s > n/2$ ($s = 2$ if $n \leq 3$) and suppose that there exists $b_0 > 0$ such that u_0 is the unique minimum of $(\mathcal{CP}_0^s)$ in $\bar{B}_s(u_0, b_0)$. Then*
(i) *The controls u_ε, solutions of $(\mathcal{CP}_\varepsilon^{b_0,s})$, strongly converge to u_0 in $L^s(\Omega)$ as $\varepsilon \downarrow 0$.*
(ii) *It holds that $J_\varepsilon(u_\varepsilon) \to J_0(u_0)$ and that $J_0(u_\varepsilon) \downarrow J_0(u_0)$.*
(iii) *The states y_ε converge to y_0 in $\mathcal{Y}^s$ and the adjoint states p_ε converge to p_0 in $\mathcal{Y}^s$.*

Proof. Since u_ε is bounded in $L^2(\Omega)$, extracting a subsequence if necessary, it converges weakly to some $\bar{u}$. Similary, since $J_0(u_\varepsilon)$ is bounded in $\mathbb{R}$, we can assume, extracting a subsequence again, that there exists $\bar{J} \geq 0$ such that $J_0(u_\varepsilon)$ converges to $\bar{J}$.

In view of the optimality of u_ε, for every $\eta > 0$ such that $u_0 + \eta$ is feasible for $(\mathcal{CP}_\varepsilon^{b_0,s})$, we have that

$$J_\varepsilon(u_\varepsilon) \leq J_0(u_0 + \eta) + \varepsilon \int_\Omega \ell(u_0(x) + \eta) \mathrm{d}x.$$

Letting first $\varepsilon \downarrow 0$ and then $\eta \downarrow 0$ yields

$$\overline{\lim}_{\varepsilon \downarrow 0} J_\varepsilon(u_\varepsilon) \leq J_0(u_0). \tag{2.31}$$

On the other hand, because of the convexity of ℓ, there exist some β_1 and β_2 such $\ell(x) \geq \beta_1 x + \beta_2$ for all $x \in \mathbb{R}_+$. Thus

$$J_\varepsilon(u_\varepsilon) \geq J_0(u_\varepsilon) + \varepsilon \int_\Omega (\beta_1 u_\varepsilon(x) + \beta_2) \, \mathrm{d}x. \tag{2.32}$$

Using (2.31), (2.32) and the fact that J_0 is w.l.s.c. yields that

$$J_0(u_0) \geq \overline{\lim}_{\varepsilon \downarrow 0} J_\varepsilon(u_\varepsilon) \geq \underline{\lim}_{\varepsilon \downarrow 0} J_\varepsilon(u_\varepsilon) \geq \bar{J} \geq J_0(\bar{u}). \tag{2.33}$$

Since u_0 is the unique minimum of $(\mathcal{CP}_0^s)$ in $\bar{B}_s(u_0, b_0)$, it holds that $\bar{u} = u_0$ and hence (ii) is established.

In order to prove (i) it suffices to note that thanks to proposition 46 (ii) the states y_ε converge strongly in $L^2(\Omega)$ to y_0. Therefore, since $J_0(u_\varepsilon) \to J_0(u_0)$ we have that $||u_\varepsilon||_2 \to ||u_0||_2$. Together with the weak convergence in $L^2(\Omega)$ of u_ε to u_0, we obtain the strong convergence in $L^2(\Omega)$. The convergence in $L^s(\Omega)$ follows directly from the convergence in $L^2(\Omega)$ and the fact that u_ε is uniformly bounded in $L^\infty(\Omega)$ by proposition 55. Finally (iii) is a direct consequence of lemma 56. ■

Remark 58 Note that, under the hypothesis of the theorem above, the convergence in $L^s(\Omega)$ of u_ε to u_0 implies that for ε small enough the constraint $u_\varepsilon \in \bar{B}_s(u_0, b)$ is inactive.

Now we obtain lower bounds for u_ε.

Proposition 59 *Under the hypothesis of proposition 57 there exists a constant $K_1 > 0$ such that for $\varepsilon > 0$ small enough*

$$\ell'(2u_\varepsilon(x)) \geq -\frac{2K_1}{\varepsilon} \quad \textit{for a.a. } x \in \Omega. \tag{2.34}$$

Proof. By (2.17)(i) there exists $\zeta > 0$ such that ℓ is decreasing on $(0, \zeta]$. Set

$$\underline{\Omega}^\zeta := \{x \in \Omega;\ u_\varepsilon(x) \leq \zeta/2\}$$

and

$$u_\varepsilon^\zeta(x) := \begin{cases} \zeta & \text{if } x \in \underline{\Omega}^\zeta \\ u_\varepsilon(x) & \text{otherwise} \end{cases} \quad ; \quad y_\varepsilon^\zeta(x) := y_{u_\varepsilon^\zeta}(x) \quad \text{for a.a. } x \in \Omega. \tag{2.35}$$

Note that, by remark 58, u_ε^ζ is feasible for ζ small enough. In addition,

$$0 \leq \left(u_\varepsilon^\zeta + u_\varepsilon\right)\left(u_\varepsilon^\zeta - u_\varepsilon\right) \leq \frac{3}{2}\zeta(u_\varepsilon^\zeta - u_\varepsilon)\mathbf{1}_{\underline{\Omega}^\zeta}.$$

Thus, taking $u = u_\varepsilon^\zeta$ in (2.27) and reasoning as in the proof of proposition 55, we obtain the existence of $K_1' > 0$ such that

$$J_\varepsilon(u_\varepsilon^\zeta) - J_\varepsilon(u_\varepsilon) \leq K_1'\zeta \text{meas}(\underline{\Omega}^\zeta) + \varepsilon \int_{\underline{\Omega}^\zeta} \left(\ell(u_\varepsilon^\zeta(x)) - \ell(u_\varepsilon(x))\right) \mathrm{d}x.$$

By the mean value theorem and the convexity of ℓ, which implies that ℓ' is increasing, we find that

$$\ell(u_\varepsilon^\zeta(x)) - \ell(u_\varepsilon(x)) \leq \tfrac{1}{2}\ell'(\zeta)\zeta$$

for a.a. $x \in \underline{\Omega}^{\zeta}$. This in turn implies that

$$J_{\varepsilon}(u_{\varepsilon}^{\zeta}) - J_{\varepsilon}(u_{\varepsilon}) \leq \zeta \mathrm{meas}(\underline{\Omega}^{\zeta}) \left(K_1' + \tfrac{1}{2}\varepsilon \ell'(\zeta)\right). \tag{2.36}$$

Therefore, by the optimality of u_{ε}, if $\mathrm{meas}(\underline{\Omega}^{\zeta}) > 0$ we have that $K_1' \geq -\frac{1}{2}\varepsilon\ell'(\zeta)$. By choosing ζ' such that $K_1' < -\frac{1}{2}\varepsilon\ell'(\zeta')$ we obtain that for a.a. $x \in \underline{\Omega}^{\zeta}$

$$\ell'(2u_{\varepsilon}(x)) \geq \ell'(\zeta').$$

Relation (2.34) follows by letting $\ell'(\zeta') \uparrow -2K_1\varepsilon$. ■

Remark 60 For the examples given in remark 52 inequality (2.34) yields
(i) If $\ell(t) = -\log t$ then there exists $C_1 > 0$ such that $u_{\varepsilon}(x) \geq C_1\varepsilon$ for a.a. $x \in \Omega$.
(ii) If $\ell(t) = t\log t$ then there exists $C_2, C_3 > 0$ such that $u_{\varepsilon}(x) \geq C_2 \exp(-C_3/\varepsilon)$ for a.a. $x \in \Omega$.
(iii) If $\ell(t) = t^{-p}$ with $p > 0$ then there exists $C_4 > 0$ such that $u_{\varepsilon}(x) \geq C_4\varepsilon^{1/(p+1)}$ for a.a. $x \in \Omega$.
(iv) If $\ell(t) = -t^p$ with $p \in (0,1)$ then there exists $C_5 > 0$ such that $u_{\varepsilon}(x) \geq C_5\varepsilon^{1/(1-p)}$ for a.a. $x \in \Omega$.

Note that $u \in L^s(\Omega) \to \int_{\Omega} \ell(u(x))\mathrm{d}x$ is, in general, not continuous and whence not differentiable. This implies that we cannot write directly the first-order condition for the optimality of u_{ε}. However, we can avoid this difficulty by noting that, in view of propositions 55 and 59, $u \in L^{\infty}(\Omega) \to \int_{\Omega} \ell(u(x))\mathrm{d}x$ is differentiable at any solution of $(\mathcal{CP}_{\varepsilon}^{b_0,s})$.

Proposition 61 *Under the hypothesis of proposition 57, for $\varepsilon > 0$ small enough it holds that*

$$u_{\varepsilon}(x) = \pi_{\varepsilon}(-N^{-1}p_{\varepsilon}(x)) \quad \textit{for a.a. } x \in \Omega, \tag{2.37}$$

where for every $z \in \mathbb{R}$, $\pi_{\varepsilon}(z)$ is the unique solution of

$$\mathrm{Min}\ \tfrac{1}{2}(x-z)^2 + \varepsilon\ell(x), \quad s.t. \quad x \in \mathbb{R}_{++}. \tag{$\mathcal{P}_{\varepsilon,z}$}$$

Proof. By proposition 55 it holds that $u_{\varepsilon} \in L^{\infty}(\Omega)$. Hence, it is a local solution of

$$\mathrm{Min}\ \ J_{\varepsilon}(u) \quad \text{subject to } u \in \mathcal{U}_+^s \cap \bar{B}_s(u_0, b_0) \cap L^{\infty}(\Omega).$$

Proposition 59 implies that $J_\varepsilon : L^\infty(\Omega) \to \mathbb{R}$ is differentiable. Therefore, writing the first-order condition for the above problem and noting remark 58, we have

$$DJ_0(u_\varepsilon)h + \varepsilon \int_\Omega \ell'(u_\varepsilon(x))h(x)\mathrm{d}x \;\; = \;\; 0 \quad \text{for all } h \in L^\infty(\Omega),$$

which implies that

$$Nu_\varepsilon(x) + p_\varepsilon(x) + \varepsilon\ell'(u_\varepsilon) \;\; = \;\; 0 \quad \text{for a.a. } x \in \Omega. \tag{2.38}$$

The conclusion follows noting that for $x \in \Omega$, equation (2.38) is the first-order optimality condition of $(\mathcal{P}_{\varepsilon,z})$ with $z = -N^{-1}p_\varepsilon(x)$. ■

Remark 62 *Note that for every $z \in \mathbb{R}$ the function $\pi_\varepsilon(z)$ corresponds to the interior penalty approximation of $\pi_0(z)$.*

We collect in the following lemma, some useful properties of the family $\{\pi_\varepsilon\}_{\varepsilon\geq 0}$ whose proof can be found in [2] Section 3 for a more general case.

Lemma 63 *The family of functions $\{\pi_\varepsilon\}_{\varepsilon\geq 0}$ satisfies*
(i) *There exist c_π, independent of 'ε', such that for all $z_1, z_2 \in \mathbb{R}$,*

$$|\pi_\varepsilon(z_1) - \pi_\varepsilon(z_2)| \leq c_\pi|z_1 - z_2|. \tag{2.39}$$

(ii) *As $\varepsilon \downarrow 0$ the sequence π_ε converges to π_0 uniformly on each compact set of $\mathbb{R}$.*
(iii) *The function $(\varepsilon, z) \to D_z\pi_\varepsilon(z)$ is continuous in $(\bar\varepsilon, \bar z)$ for every $\bar\varepsilon \geq 0$ and $\bar z \neq 0$.*
(iv) *The continuous function $\pi_\varepsilon - \pi_0$ is increasing in $(-\infty, 0)$ and decreasing in $(0, \infty)$. Henceforth,*

$$\sup_{z\in\mathbb{R}} |\pi_\varepsilon(z) - \pi_0(z)| = |\pi_\varepsilon(0) - \pi_0(0)| = |\pi_\varepsilon(0)|.$$

(v) *For each compact set $K \subseteq \mathbb{R}$ not containing 0, it holds that:*

$$\sup_{z\in K} |\pi_\varepsilon(z) - \pi_0(z)| = O(\varepsilon).$$

Remark 64 Hypothesis (ii) in (2.17) is used to prove (iii) in the lemma above.

2.3 Main results

As before, we consider $f \in L^s(\Omega)$ and for the rest of the article we assume that $s > \frac{1}{2}n$ ($s = 2$ if $n \leq 3$). Let u_0 be a solution of $(\mathcal{CP}_0^s)$ and y_0, p_0 its associated state and costate, respectively. Analogously, for $\varepsilon > 0$, $b > 0$ let u_ε be a solution of $(\mathcal{CP}_\varepsilon^{b,s})$ and denote, as in the previous section, by y_ε and p_ε its associated state and costate, respectively. Consider the mapping $F : \mathcal{Y}^s \times \mathcal{Y}^s \times \mathbb{R}_+ \to L^s(\Omega) \times L^s(\Omega)$ defined by

$$F(y, p, \varepsilon) := \begin{pmatrix} \Delta y + \Pi_\varepsilon(-N^{-1}p) + f - \phi(y) \\ \Delta p + y - \bar{y} - \phi'(y)p \end{pmatrix}. \tag{2.40}$$

In view of (2.14), proposition 57 and (2.37) we see that if u_0 is a local strict solution of $(\mathcal{CP}_0^s)$ then for b and $\varepsilon \geq 0$ small enough

$$F(y_\varepsilon, p_\varepsilon, \varepsilon) = 0.$$

Motivated by this fact, our objective is to obtain an "asymptotic expansion" for $(y_\varepsilon, p_\varepsilon)$ around (y_0, p_0). As in the ODE case (see [2]), the mapping F is, in general, not differentiable at $(y_0, p_0, 0)$. In fact, it can be easily seen that $D_\varepsilon F(y_0, p_0, 0)$ does not always exists. Therefore, we cannot apply the standard implicit function theorem in order to obtain such expansion. We will overcome this difficulty in the same way as in [2], i.e. by using the following restoration theorem, whose proof can be found in the Appendix of [2].

Theorem 65 (Restoration theorem) *Let X and Y be Banach spaces, E a metric space and $F : U \subset X \times E \to Y$ a continuous mapping on an open set U. Let $(\hat{x}, \varepsilon_0) \in U$ be such that $F(\hat{x}, \varepsilon_0) = 0$. Assume that there exists a surjective linear continuous mapping $A : X \to Y$ and a function $c : \mathbb{R}_+ \to \mathbb{R}_+$ with $c(\beta) \downarrow 0$ when $\beta \downarrow 0$ such that, if $x \in \overline{B}(\hat{x}, \beta)$, $x' \in \overline{B}(\hat{x}, \beta)$ and $\varepsilon \in B(\varepsilon_0, \beta)$, then*

$$\|F(x', \varepsilon) - F(x, \varepsilon) - A(x' - x)\| \leq c(\beta)\|x' - x\|. \tag{2.41}$$

Then, denoting by B a bounded right inverse of A, for ε close to ε_0, $F(\cdot, \varepsilon)$ has, in a neighborhood of $\hat{x}$, a zero denoted by x_ε such that the following expansion holds

$$x_\varepsilon = \hat{x} - BF(\hat{x}, \varepsilon) + r(\varepsilon) \quad \text{with } ||r(\varepsilon)|| = o\left(\|F(\hat{x}, \varepsilon)\|\right). \tag{2.42}$$

Remark 66 *Note that hypothesis* (2.41) *implies that if A is invertible and β is such that $c(\beta)||A^{-1}||_{Y\to X} < 1$ (where $||\cdot||_{Y\to X}$ denotes the standard norm for the space of bounded linear functionals from Y to X) then for all $\varepsilon \in B(\varepsilon_0,\beta)$ the mapping $F(\cdot,\varepsilon)$ is injective in $\bar{B}(\hat{x},\beta)$. In particular, for $\varepsilon \in B(\varepsilon_0,\beta)$ there exists a unique $x_\varepsilon \in \bar{B}(\hat{x},\beta)$ such that $F(x_\varepsilon,\varepsilon)=0$.*

In order to verify that F, defined in (2.40), satisfies the hypothesis of theorem 65 we need the following lemmas.

Lemma 67 *Let $f:\mathbb{R}\to\mathbb{R}$ be a Lipschitz function and denote by $\mathcal{A}(f)$ the set of points were f is not differentiable. For $s\in[1,\infty)$ set $\bar{f}: L^\infty(\Omega)\to L^s(\Omega)$ defined by*

$$\bar{f}[w](x) := f(w(x)). \tag{2.43}$$

Then $\bar{f}$ is Fréchet differentiable at every $\bar{w}\in L^\infty(\Omega)$ satisfying that

$$meas\,\{x\in\Omega\ ;\ \bar{w}(x)\in\mathcal{A}(f)\}=0 \tag{2.44}$$

and $\left(D\bar{f}[\bar{w}]h\right)(x) = f'(\bar{w}(x))h(x)$ for all $h\in L^\infty(\Omega)$.

Proof. Let $\theta: L^\infty(\Omega)\to\mathbb{R}_+$ defined by

$$\theta(h) := \frac{||\bar{f}(\bar{w}+h)-\bar{f}(\bar{w})-f'(\bar{w}(\cdot))h||_s^s}{||h||_\infty^s}. \tag{2.45}$$

We have to show that $\theta(h)\to 0$ as $h\to 0$. In fact we have

$$0\le\theta(h)\le\int_\Omega \frac{|f(\bar{w}(x)+h(x))-f(\bar{w}(x))-f'(\bar{w}(x))h(x)|^s}{|h(x)|^s}\mathrm{d}x \tag{2.46}$$

and the result follows by the dominated convergence theorem using the fact that f is Lipschitz. ■

For $w\in\mathcal{Y}^s$ set

$$Sing(w) := \{x\in\bar{\Omega}\ ;\ w(x)=0\} \tag{2.47}$$

and for every $\varepsilon\ge 0$ define $\Pi_\varepsilon:\mathcal{Y}^s\to L^s(\Omega)$ by $(\Pi_\varepsilon(w))(x) := \pi_\varepsilon(w(x))$ for a.a. $x\in\Omega$. Lemmas 63 and 67 allows us to prove the following result.

Lemma 68 *Let $\widehat{w}\in\mathcal{Y}^s$ and suppose $meas(Sing(\widehat{w}))=0$. Then*
(i) *For every $\varepsilon>0$, $w\in\mathcal{Y}^s$, the function Π_ε is differentiable at w and for every $h\in\mathcal{Y}^s$*

$$(D\Pi_\varepsilon(w)h)(x) = \pi_\varepsilon'(w(x))h(x), \quad \text{for a.a. } x\in\Omega.$$

(ii) *The function* Π_0 *is differentiable at* $\widehat{w}$ *and for every* $h \in \mathcal{Y}^s$

$$(D\Pi_0(\widehat{w})h)(x) = \pi_0'(\widehat{w}(x))h(x), \quad \text{for a.a. } x \in \Omega.$$

(iii) *There exist a nondecreasing function* $c : \mathbb{R}_+ \to \mathbb{R}_+$ *with* $\lim_{\beta \downarrow 0} c(\beta) = 0$ *such that for any* $w', w \in \mathcal{Y}^s$ *with* $||w' - \widehat{w}||_{2,s} \leq \beta, \quad ||w - \widehat{w}||_{2,s} \leq \beta$ *and* $\varepsilon \in [0, \beta]$ *we have*

$$||\Pi_\varepsilon(w') - \Pi_\varepsilon(w) - D\Pi_0(\widehat{w})(w' - w)||_s \leq c(\beta)||w' - w||_{2,s}. \tag{2.48}$$

Proof. (i) Since, for $\varepsilon > 0$, π_ε is $\mathcal{C}^1$ it holds that Π_ε, viewed as mapping from $L^\infty(\Omega)$ into $L^\infty(\Omega)$, is also $\mathcal{C}^1$. Therefore, noting that $s > n/2$ ($s = 2$ if $n \leq 3$), the result easily follows.
(ii) Consequence of lemma 67 using that $\mathcal{Y}^s \subseteq L^\infty(\Omega)$.
(iii) Note that

$$||\Pi_\varepsilon(w') - \Pi_\varepsilon(w) - D\Pi_0(\widehat{w})(w' - w)||_s =$$

$$\left|\left|\left(\int_0^1 \{D\Pi_\varepsilon(w + s(w' - w)) - D\Pi_0(\widehat{w})\}\, \mathrm{d}s\right)(w' - w)\right|\right|_s$$
$$\leq \sup_{z \in B_{2,s}(\widehat{w},\beta)} ||D\Pi_\varepsilon(z) - D\Pi_0(\widehat{w})||_{\mathcal{Y}^s \to L^s(\Omega)}\, ||w' - w||_{2,s}\,.$$

where $B_{2,s}(\widehat{w}, \beta)$ denotes the ball in $W^{2,s}(\Omega)$ of center $\widehat{w}$ and radius β and $||\cdot||_{\mathcal{Y}^s \to L^s(\Omega)}$ denotes the standard norm for the space of linear bounded functions from $\mathcal{Y}^s$ to $L^s(\Omega)$. Let $h \in \mathcal{Y}^s$ with $||h||_{2,s} \leq 1$. Since $s > n/2$ ($s = 2$ if $n \leq 3$), we have

$$||D\Pi_\varepsilon(z)h - D\Pi_0(\widehat{w})h||_s^s \leq c_s^s \left(\int_\Omega |\pi_\varepsilon'(z(x)) - \pi_0'(\widehat{w}(x))|^s \mathrm{d}x\right)$$

with c_s being defined in (2.4). Thus,

$$||\Pi_\varepsilon(w') - \Pi_\varepsilon(w) - D\Pi_0(\widehat{w})(w' - w)||_s \leq c(\beta)\, ||w' - w||_{2,s}$$

where $c(\beta)$ is the nondecreasing function defined by

$$c(\beta) := c_s \left(\int_\Omega \sup_{\varepsilon \in [0,\beta]} \sup_{z \in B(\widehat{w}(x),\beta)} |\pi_\varepsilon'(z(x)) - \pi_0'(\widehat{w}(x))|^s \mathrm{d}x\right)^{\frac{1}{s}}.$$

Since $meas(Sing(\widehat{w})) = 0$, lemma 63 (i) and (iii) yields that $c(\beta) \downarrow 0$ as $\beta \downarrow 0$ by the dominated convergence theorem. ■

In order to establish our main result we will have to impose a second-order sufficient condition at any solution of $(\mathcal{CP}_0^s)$. First let us study the following abstract setting:

Consider a nonempty closed and convex set $K \subseteq L^2(\Omega)$ and define $K_s := K \cap L^s(\Omega)$. We will establish some second-order sufficient conditions for the problem

$$\text{Min } J_0(u) \text{ subject to } u \in K_s \ . \qquad (\mathcal{AP})$$

Let $\bar{u} \in K$. The radial, tangent, normal cones to K at $\bar{u}$ and the critical cone in $L^2(\Omega)$ at $\bar{u}$ are defined respectively by

$$\begin{array}{rcl} \mathcal{R}_K(\bar{u}) & := & \{h \in L^2(\Omega) \ ; \ \exists \ \sigma > 0; \ \bar{u} + \sigma h \in K\}, \\ T_K(\bar{u}) & := & \{h \in L^2(\Omega) \ ; \ \exists \ u(\sigma) = \bar{u} + \sigma h + o_2(\sigma) \in K, \ \sigma \geq 0, \}, \\ N_K(\bar{u}) & := & \{h^* \in L^2(\Omega) \ ; \ \langle h^*, u - \bar{u} \rangle \leq 0, \ \forall \ u \in K\}, \\ C(\bar{u}) & := & \{h \in T_K(u) \text{ and } DJ_0(\bar{u})h \leq 0\}. \end{array} \qquad (2.49)$$

In the definition of $T_K(\bar{u})$ the function o_2 is such that $||o_2(\sigma)/\sigma||_2 \to 0$. If $\bar{u} \in K_s$ we define analogously the radial, tangent and normal cones to K_s at $\bar{u}$ and the critical cone in $L^s(\Omega)$ at $\bar{u}$ by replacing $L^2(\Omega)$ by $L^s(\Omega)$ and K by K_s in (2.49). We denote them by $\mathcal{R}_{K_s}, T_{K_s}(\bar{u}), N_{K_s}(\bar{u})$ and $C_s(\bar{u})$ respectively.

We say that J_0 satisfies the local quadratic growth condition at $\bar{u}$ if there exists $\alpha > 0$ and a neighborhood $\mathcal{V}_s$ of $\bar{u}$ in $L^s(\Omega)$ such that

$$J_0(u) \geq J_0(\bar{u}) + \alpha ||u - \bar{u}||_2^2 + o(||u - \bar{u}||_2^2) \quad \text{for all } u \in K_s \cap \mathcal{V}_s. \qquad (2.50)$$

The following notion of polyhedricity will be required (see [52, 71]). The set K_s is said to be *polyhedric* in $L^s(\Omega)$ at $u \in K_s$ if for all $u^* \in N_{K_s}(u)$ (sets of normal of K_s at u), the set $\mathcal{R}_{K_s}(u) \cap (u^*)^\perp$ is dense in $T_{K_s}(u) \cap (u^*)^\perp$ with respect to the $L^s(\Omega)$ norm. If K_s is polyhedric in $L^s(\Omega)$ at each $u \in K_s$ we say that K_s is *s-polyhedric*.

For various types of optimization problems (see [24]), positivity of the second derivative of the cost function over the critical cone at a point u can be related to the quadratic growth condition at u. This is usually referred as a no gap second-order sufficient condition which under some hypothesis will be satisfied in our problem.

If ϕ is C^2 then, since $s > n/2$ ($s = 2$ if $n \leq 3$), the function $J_0 : L^s(\Omega) \to \mathbb{R}$ is C^2 (see [24, lemma 6.27]) and for all $u, v \in L^s(\Omega)$ we have

$$D^2 J_0(u)(v, v) = \int_\Omega \left\{ N v(x)^2 + (1 - p_u(x)\phi''(y_u(x)))\, z_v(x)^2 \right\} \mathrm{d}x, \qquad (2.51)$$

where z_v is the unique solution of the linearized state equation

$$\left\{ \begin{array}{rcll} -\Delta z(x) + \phi'(y_u(x))z(x) & = & v(x) & \text{for} \quad x \in \Omega, \\ z(x) & = & 0 & \text{for} \quad x \in \partial\Omega. \end{array} \right. \qquad (2.52)$$

In addition, it is proved that the quadratic form $D^2J_0(u)$ has a unique continuous extension over $L^2(\Omega) \times L^2(\Omega)$ and this extension is a Legendre form, which means that it is sequentially w.l.s.c. and that if h_k converges weakly to h in $L^2(\Omega)$ and $D^2J_0(u)(h_k, h_k) \to D^2J_0(u)(h, h)$ then h_k converges strongly to h in $L^2(\Omega)$.

The theorem below, which concerns to second-order sufficient conditions for $(\mathcal{AP})$, is proved in [24, theorem 6.31].

Theorem 69 *Consider problem $(\mathcal{AP})$ and let $\bar{u} \in K_s$. If K_s is s-polyhedric and $C_s(\bar{u})$ is dense in $C(\bar{u})$, then the quadratic growth condition* (2.50)*, the second-order condition*

$$\exists\ \alpha > 0, \textit{such that } D^2J_0(\bar{u})(h,h) \geq \alpha ||h||_2^2 \quad \textit{for all } h \in C(\bar{u}) \tag{2.53}$$

and the punctual relation

$$D^2J_0(\bar{u})(h,h) > 0 \quad \textit{for all } h \in C(\bar{u}) \setminus \{0\} \tag{2.54}$$

are equivalent.

When $K = \mathcal{U}_+^2$ and $u \in K$ it is easy to verify that

$$\begin{array}{rcl} T_K(u) & := & \{v \in L^2(\Omega)\ ;\ v(x) \geq 0 \ \text{ if } u(x) = 0 \text{ for a.a. } x \in \Omega\} \\ N_K(u) & := & \{v \in (L^2(\Omega))^*\ ;\ v(x) \leq 0 \text{ and } v(x) = 0 \text{ if } u(x) > 0\}. \end{array} \tag{2.55}$$

If $u \in K_s$ the correspondig expressions for $T_{K_s}(u)$ and $N_{K_s}(u)$ are obtained by replacing $L^2(\Omega)$ by $L^s(\Omega)$ in (2.55). If u_0 is a local solution of $(\mathcal{CP}_0^s)$ and $p_0(x) \neq 0$ for almost all $x \in \Omega$, expression (2.11) yields that

$$C_s(u_0) \ := \ \{v \in L^s(\Omega)\ ;\ v(x) = 0 \ \text{ if } u_0(x) = 0 \text{ for a.a. } x \in \Omega\}. \tag{2.56}$$

Analogously, if u_0 is a solution of $(\mathcal{CP}_0^2)$, the corresponding expression for $C(u_0)$ is obtained by replacing $L^s(\Omega)$ by $L^2(\Omega)$ in (2.56).

Now we give a simple proof of the following well known result (see for example [24, proposition 6.33]) which shows that theorem 69 can be applied in our case $(K_s = \mathcal{U}_+^s)$.

Lemma 70 *Suppose that $K_s = \mathcal{U}_+^s$, then*
(i) The set K_s is s-polyhedric.
(ii) If u_0 is a local solution of $(\mathcal{CP}_0^s)$, then $C_s(u_0)$ is dense in $C(u_0)$.

Proof. (i) Let $u \in \mathcal{U}_+^s$ and $u^* \in N_{\mathcal{U}_+^s}(u)$. For $h \in T_{\mathcal{U}_+^s}(u) \cap (u^*)^\perp$ and $k \in \mathbb{N}$ let $h_k \in L^\infty(\Omega)$ be defined as

$$h_k(x) := \begin{cases} 0 & \text{if } 0 < u(x) \leq 1/k \\ \max\{-k, \min\{h(x), k\}\} & \text{otherwise.} \end{cases} \tag{2.57}$$

It is easy to check that $h_k \in \mathcal{R}_{\mathcal{U}_+^s} \cap (u^*)^\perp$ and $h_k \to h$ in $L^s(\Omega)$ by the dominated convergence theorem.
(ii) Given $h \in C(u_0)$ the sequence h_k defined in (2.57) belongs to $C_s(u_0)$ and converges in $L^2(\Omega)$ to h by the dominated convergence theorem. ■

To obtain our main result we will assume two hypothesis. The first one allows to ensure that hypothesis (2.41) holds at $(y_0, p_0, 0)$ for the mapping F defined in (2.40). The second one will imply that the set of solutions of $(\mathcal{CP}_0^s)$ is isolated and that $D_{(y,p)}F(y_0, p_0, 0)$ is an isomorphism (see lemma 72). We consider the following hypothesis:

(H1) For the adjoint state p_0, associated to any local solution u_0 of $(\mathcal{CP}_0^s)$, it holds that

$$meas(Sing(p_0)) = 0.$$

(H2) At any local solution u_0 of $(\mathcal{CP}_0^s)$, condition (2.53) holds.

Remark 71 *Suppose that **(H1)** does not hold. Then, the $W^{2,s}$ regularity of p_0 implies that $-\Delta p_0 = 0$ in $Sing(p_0)$ (see [30] page 195). Therefore, by equations* (2.8) *and* (2.10),

$$-\Delta\bar{y}(x) + \phi(\bar{y}(x)) \;\; = \;\; f(x) \quad for \quad x \in Sing(p_0)$$

which yields a compatibility condition between the data $\bar{y}$ and f.

Lemma 72 *Let u_0 be a solution of $(\mathcal{CP}_0^s)$, suppose that ϕ is C^2 and that **(H1)**, **(H2)** hold. Then F (defined in* (2.40)*) is differentiable with respect to (y, p) at $(y_0, p_0, 0)$ and the linear mapping $D_{(y,p)}F(y_0, p_0, 0)$ is an isomorphism.*

In addition, for every $(\delta_1, \delta_2) \in L^s(\Omega) \times L^s(\Omega)$, we have that

$$D_{(y,p)}F(y_0, p_0, 0)^{-1}(\delta_1, \delta_2)$$

is the unique solution of the reduced optimality system of

$$Min \left\{ \int_\Omega \left[\tfrac{1}{2} N v^2 + \tfrac{1}{2} \left(1 - p_0 \phi''(y_0)\right) z_{v+\delta_1}^2 + \delta_2 z_{v+\delta_1} \right] \mathrm{d}x \; ; v \in C(u_0) \right\} \tag{$\mathcal{QP}_{\delta_1,\delta_2}$}$$

where z_v is defined in (2.52).

Proof. In view of assumption **(H1)** and lemma 68, the mapping F is differentiable with respect to (y,p) at $(y_0,p_0,0)$ and

$$D_{(y,p)}F(y_0,p_0,0)(z,q)=\begin{pmatrix}\Delta z-\Pi_0'(-N^{-1}p_0)N^{-1}q-\phi'(y_0)z\\ \Delta q+z-\phi''(y_0)p_0z-\phi'(y_0)q\end{pmatrix}.$$

Let $\delta_1,\delta_2\in L^s(\Omega)$, to find $(z,q)\in\mathcal{Y}^s$ such that $D_{(y,p)}F(y_0,p_0,0)(z,q)=(\delta_1,\delta_2)$ is equivalent to solve in $\mathcal{Y}^s\times\mathcal{Y}^s$ the following system of PDE's

$$\begin{aligned}-\Delta z(x)+\phi'(y_0(x))z(x) &= \delta_1(x)-\frac{\Pi_0'(-N^{-1}p_0(x))q(x)}{N}\\ -\Delta q(x)+\phi''(y_0(x))p_0(x)z(x)+\phi'(y_0(x))q(x) &= \delta_2(x)+z(x)\end{aligned}$$

for all $x\in\Omega$. But these equations are exactly the reduced optimality system for problem $(\mathcal{QP}_{\delta_1,\delta_2})$ which can be written, denoting by $\langle\cdot,\cdot\rangle_{L^2}$ the standard duality product in $L^2(\Omega)$, as

$$\text{Min }\tfrac{1}{2}D^2J_0(u_0)(v,v)+\langle\gamma^*_{\delta_1,\delta_2},v\rangle_{L^2}+\beta^*_{\delta_1,\delta_2}\quad\text{subject to } v\in C(u_0)$$

for some $\gamma^*_{\delta_1,\delta_2}\in L^2(\Omega)$ and

$$\beta^*_{\delta_1,\delta_2}:=\int_\Omega\left[\tfrac{1}{2}\left(1-p_0\phi''(y_0)\right)z_{\delta_1}^2+\delta_2z_{\delta_1}\right]\mathrm{d}x.$$

In fact, since $z_{v+\delta_1}=z_v+z_{\delta_1}$, the cost function of $(\mathcal{QP}_{\delta_1,\delta_2})$ is given by

$$\tfrac{1}{2}D^2J_0(u_0)(v,v)+\int_\Omega\left[\left(1-p_0\phi''(y_0)\right)z_vz_{\delta_1}+\delta_2z_v\right]\mathrm{d}x+\beta^*_{\delta_1,\delta_2}.$$

Since the above integral is a linear form, as a function of v, the existence of $\gamma^*_{\delta_1,\delta_2}$ follows by the Riesz's theorem.

By (H2) this cost function is strongly convex over the closed subspace $C(u_0)$ and therefore has a unique minimum. The $W^{2,s}$ regularity for its associated state and adjoint state follows readily by a boostrapping argument. ■

For every $\varepsilon\geq0$ let us define $q_\varepsilon:=-p_\varepsilon/N$. Now we can state our main result.

Theorem 73 *Let u_0 be a solution of $(\mathcal{CP}_0^s)$, suppose that ϕ is C^2 and that **(H1)**, **(H2)** hold. Denote respectevely by y_0 and p_0 the state and adjoint state associated to u_0. Then there are $\bar{b}>0$ and $\bar{\varepsilon}>0$ such that for $\varepsilon\in[0,\bar{\varepsilon}]$ problem $(\mathcal{CP}_\varepsilon^{\bar{b},s})$ has a unique solution u_ε. In addition, denoting by y_ε and p_ε*

the associated state and adjoint state for u_ε, the following expansion around (y_0, p_0) holds

$$\begin{pmatrix} y_\varepsilon \\ p_\varepsilon \end{pmatrix} = \begin{pmatrix} y_0 \\ p_0 \end{pmatrix} + D_{(y,p)}F(y_0, p_0, 0)^{-1}F(y_0, p_0, \varepsilon) + r(\varepsilon), \tag{2.58}$$

where $r(\varepsilon) = o(||F(y_0, p_0, \varepsilon)||_s)$. Moreover, $D_{(y,p)}F(y_0, p_0, 0)^{-1}F(y_0, p_0, \varepsilon)$ is characterized as being the unique solution of $(\mathcal{QP}_{\delta\Pi(\varepsilon),0})$ where

$$\delta\Pi(\varepsilon) := \Pi_\varepsilon(q_0) - \Pi_0(q_0).$$

Proof. Lemma 68 (ii) implies that hypothesis (2.41) of theorem 65 is satisfied with $A = D_{(y,p)}F(y_0, p_0, 0)$. Lemma 72 yields that A is invertible, whence the first assertion follows from the convergence of $(y_\varepsilon, p_\varepsilon)$ to (y_0, p_0) in $\mathcal{Y}^s \times \mathcal{Y}^s$, established in proposition 57, and remark 66.

Noting that $F(y_0, p_0, \varepsilon) = F(y_0, p_0, \varepsilon) - F(y_0, p_0, 0) = (\delta\Pi(\varepsilon), 0)$, the second assertion follows by theorem 65 and lemma 72 with $\delta_1 = \delta\Pi(\varepsilon)$ and $\delta_2 = 0$. ■

Theorem 73 yields, in particular, the following error bounds.

Corollary 74 (Error bounds) *Under the assumptions of theorem 73 we have*

(i) *The error estimates for $u_\varepsilon, y_\varepsilon$ and p_ε are given by*

$$||u_\varepsilon - u_0||_s + ||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O\left(||\delta\Pi(\varepsilon)||_s\right). \tag{2.59}$$

(ii) *The error bound for the control in the infinity norm is given by*

$$||u_\varepsilon - u_0||_\infty = O\left(||\delta\Pi(\varepsilon)||_\infty\right) = O(\pi_\varepsilon(0)). \tag{2.60}$$

(iii) *The error estimate for the cost is given by*

$$|J_0(u_\varepsilon) - J_0(u_0)| = O\left(||\delta\Pi(\varepsilon)||_s\right). \tag{2.61}$$

Proof. (i) Theorem 65 yields that

$$||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O(||F(y_0, p_0, \varepsilon)||_s) = O(||\delta\Pi(\varepsilon)||_s). \tag{2.62}$$

Therefore, using proposition 63 (i) we obtain that

$$||u_\varepsilon - u_0||_s = ||\Pi_\varepsilon(q_\varepsilon) - \Pi_0(q_0)||_s = O(||q_\varepsilon - q_0||_s) + O(||\delta\Pi(\varepsilon)||_s), \tag{2.63}$$

which combined with (2.62) yields (2.59).
(ii) Clearly, as in (i)

$$||u_\varepsilon - u_0||_\infty = O(||q_\varepsilon - q_0||_\infty) + O(||\delta\Pi(\varepsilon)||_\infty), \tag{2.64}$$

and thus, using that $s > n/2$ ($s = 2$ if $n \leq 3$),

$$||u_\varepsilon - u_0||_\infty = O(||q_\varepsilon - q_0||_{2,s}) + O(||\delta\Pi(\varepsilon)||_\infty).$$

Hence, using the estimation given in (i),

$$||u_\varepsilon - u_0||_\infty = O(||\delta\Pi(\varepsilon)||_s) + O(||\delta\Pi(\varepsilon)||_\infty) = O(||\delta\Pi(\varepsilon)||_\infty),$$

and the result follows from lemma 63(iv).
(iii) We have

$$J_0(u_\varepsilon) - J_0(u_0) = \tfrac{1}{2}\int_\Omega \{(u_\varepsilon + u_0)(u_\varepsilon - u_0) + (y_\varepsilon + y_0 - 2\bar{y})(y_\varepsilon - y_0)\}\,\mathrm{d}x. \tag{2.65}$$

Since $s > n/2$ ($s = 2$ if $n \leq 3$), proposition 59 and lemma 46 (i) imply that $u_\varepsilon + u_0$ and $y_\varepsilon + y_0 - 2\bar{y}$ are uniformly bounded in $L^\infty(\Omega)$. Henceforth lemma 54 implies that

$$J_0(u_\varepsilon) - J_0(u_0) = O(||u_\varepsilon - u_0||_1) = O(||u_\varepsilon - u_0||_s)$$

and the result follows by (i). ■

2.4 Examples

In this section the results of section 3 are applied to the examples given in remark 52. In subsection 4.1 we obtain precise error bounds for the central path. We pay particular attention to the logarithmic barrier in view of its well known properties as a penalty function. In section 4.2 we study the error for the cost function. in what follows we will assume that ϕ is C^2.

2.4.1 Error estimates for the central path

First, note that combining (i) and (ii) of corollary 74 yields

$$||u_\varepsilon - u_0||_\infty + ||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O\left(\pi_\varepsilon(0)\right). \tag{2.66}$$

First order condition for $(\mathcal{P}_{\varepsilon,0})$ implies that $\pi_\varepsilon(0)$ is the unique solution of

$$t + \varepsilon\ell'(t) = 0. \tag{2.67}$$

Thus, particularizing ℓ and using (2.67) will give precise error bounds for the central path.

2.4.1.1 Negative power penalty

If $\ell(t) = \ell_1(t) := t^{-p}$ with $p > 0$, then (2.67) yields that $\pi_\varepsilon(0) = O\left(\varepsilon^{1/(2+p)}\right)$ and thus

$$||u_\varepsilon - u_0||_\infty + ||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O\left(\varepsilon^{1/(2+p)}\right). \tag{2.68}$$

Expression (2.68) implies that for every $p > 0$ the error is worst than $O(\sqrt{\varepsilon})$.

2.4.1.2 Power penalty

When $\ell(t) = \ell_2(t) := -t^p$ with $p \in (0,1)$, equation (2.67) yields that $\pi_\varepsilon(0) = O(\varepsilon^{1/(2-p)})$ and thus

$$||u_\varepsilon - u_0||_\infty + ||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O\left(\varepsilon^{r(p)}\right). \tag{2.69}$$

where $r(p) := 1/(2-p) < 1$. Note that $r(p) \uparrow 1$ as $p \uparrow 1$.

2.4.1.3 Entropy penalty

The case $\ell(t) = \ell_3(t) := t \log t$ will be the one with the smallest error bound. In fact, equation (2.67) implies that $\pi_\varepsilon(0)$ is the unique solution of

$$t + \varepsilon(\log t + 1) = 0. \tag{2.70}$$

Even if we do not have an explicit solution for this equation, the monotony of left hand side of (2.70) can be used in order to obtain a precise estimate for $\pi_\varepsilon(0)$. Indeed, it can be easily seen that for every $k \geq 1$, denoting by

$$\log^k(\cdot) := \log ... \log(\cdot)$$

(there are k logarithms), we have that $\pi_\varepsilon(0) = O(\psi(\varepsilon))$ where

$$\varepsilon \log^k |\log \varepsilon| \leq \psi(\varepsilon) \leq \varepsilon |\log \varepsilon| \quad \text{for } \varepsilon \text{ small enough.}$$

Thus

$$||u_\varepsilon - u_0||_\infty + ||y_\varepsilon - y_0||_{2,s} + ||p_\varepsilon - p_0||_{2,s} = O\left(\psi(\varepsilon)\right). \tag{2.71}$$

2.4.1.4 Logarithmic penalty

It is well known that the case $\ell(t) = \ell_4(t) := -\log t$ is particularly important. Fortunately, $\pi_\varepsilon(z)$ can be computed explicitly for all $z \in \mathbb{R}$. Indeed, first-order condition for $(\mathcal{P}_{\varepsilon,z})$ implies that $\pi_\varepsilon(z)$ is the unique solution of

$$t - z - \varepsilon/z = 0. \tag{2.72}$$

Henceforth, $\pi_\varepsilon(z)$ is given by

$$\pi_\varepsilon(z) = \tfrac{1}{2}\left(x + \sqrt{x^2 + 4\varepsilon}\right). \tag{2.73}$$

If $n \leq 3$ (hence $s = 2$) expression (2.73) will allow us, using corollary 74(i), to compute the error for the control in the L^2 norm (see (2.77)).

Theorem 75 *Suppose that the assumptions of theorem 73 hold. Let $\bar{b} > 0$ be such that $(\mathcal{CP}_\varepsilon^{\bar{b},s})$ has a unique solution u_ε for $\varepsilon > 0$ small enough. Then:*
(i) *We have*

$$||u_\varepsilon - u_0||_\infty + ||p_\varepsilon - p_0||_{2,s} + ||y_\varepsilon - y_0||_{2,s} \quad = \quad O(\sqrt{\varepsilon}). \tag{2.74}$$

(ii) *If in addition $n \leq 3$ (hence $s = 2$), there exist $m \in \mathbb{N}$, positive real numbers $\alpha > 0$, $0 < \bar{\delta} < 1$ and a finite collection of closed C^2 curves $(C_i)_{1\leq i\leq m}$ such that:*

- *The singular set $Sing(p_0)$ can be expressed as*

$$Sing(p_0) = \bigcup_{i=1}^{m} C_i. \tag{2.75}$$

- *For all $i \in \{1, ..., m\}$, defining $C_i^{\bar{\delta}} := \{x \in \Omega;\ dist(x, C_i) \leq \bar{\delta}\}$, it holds that:*

$$|p_0(x)| \geq \alpha\, dist(x, C_i) \quad \text{for all } x \in C_i^{\bar{\delta}}. \tag{2.76}$$

Then

$$||u_\varepsilon - u_0||_2 + ||p_\varepsilon - p_0||_{2,2} + ||y_\varepsilon - y_0||_{2,2} = O(\varepsilon^{\frac{3}{4}}). \tag{2.77}$$

Proof. (i) Follows directly from (2.66) since (2.73) implies that $\pi_\varepsilon(0) = 0$.
(ii) In view of corollary 74(i), with $s = 2$, we will estimate the right hand side of (2.59). For simplicity we assume that $Sing(p_0) = \partial\Omega$ and that $p_0 < 0$ in Ω. We will use an argument based on local mappings. Set

$$Q := \left\{x = (x', x_n) \in \mathbb{R}^{n-1} \times \mathbb{R},\ |x'| < 1, |x_n| < 1\right\}.$$

Since $\partial\Omega$ is C^2 there exists $I \in \mathbb{N}$ and $\{(\omega_i, \phi_i)\}_{0\leq i\leq I}$ such that for every $i \in \{1, ..., I\}$ we have that ω_i is an open set and $\phi_i : \omega_i \to Q$ is a C^2 mapping with a C^2 inverse satisfying that $\overline{\omega_0} \subsetneq \Omega$, $\bar{\Omega} \subseteq \cup_{i=0}^{I}\omega_i$, $\partial\Omega \subseteq \cup_{i=i}^{I}\omega_i$ and

$$\begin{aligned} \phi_i(\omega_i \cap \Omega) &= Q \cap \{x = (x', x_n) \in \mathbb{R}^{n-1} \times \mathbb{R}, x_n > 0\} =: Q^+ \\ \phi_i(\omega_i \cap \partial\Omega) &= Q \cap \{x = (x', x_n) \in \mathbb{R}^{n-1} \times \mathbb{R}, x_n = 0\} =: Q^0. \end{aligned}$$

Clearly $||\Pi_\varepsilon(q_0) - \Pi_0(q_0)||_2^2 \leq \sum_{i=0}^{I} I_i$ where for every $i \in \{1, ..., I\}$

$$I_i := \int_{\Omega \cap\ \omega_i} |\pi_\varepsilon(q_0(x)) - \pi_0(q_0(x))|^2 \mathrm{d}x.$$

Since $\overline{\omega_0} \subsetneq \Omega$, lemma 63 (iv) yields that $I_0 = O(\varepsilon^2)$. Let us now fix $i \in \{1, ..., I\}$ and set $\tau = q_0 \circ \phi_i^{-1}$. By a change of variable we obtain the existence of K_i such that

$$I_i \leq K_i \int_{B_{n-1}} \int_0^1 \left|\pi_\varepsilon(\tau(x', x_n)) - \pi_0(\tau(x', x_n))\right|^2 \mathrm{d}x_n \mathrm{d}x',$$

where B_{n-1} denotes the unit ball in $\mathbb{R}^{n-1}$. Hypothesis (2.76) implies the existence of $\bar{\alpha} > 0$ such that

$$\tau(x', x_n) \geq \bar{\alpha} x_n \quad \text{for all } x_n \in [0, \bar{\delta}]. \tag{2.78}$$

Therefore, using the uniformity with respect to $x' \in B_{n-1}$ in (2.78), we have that

$$\sum_{i=1}^{I} I_i = O\left(\int_0^1 \left|\pi_\varepsilon(\alpha x_n) - \pi_0(\alpha x_n)\right|^2 \mathrm{d}x_n\right).$$

Expression (2.73) yields that

$$\begin{array}{rcl} \int_0^1 \left|\pi_\varepsilon(\alpha x_n) - \pi_0(\alpha x_n)\right|^2 \mathrm{d}x_n & = & \int_0^1 \left(x^2 + 2\varepsilon - x\sqrt{x^2 + 4\varepsilon}\right) \mathrm{d}x \\ & = & \frac{1}{3} + 2\varepsilon - \frac{1}{3}(1 + 4\varepsilon)^{3/2} + \frac{1}{3}(4\varepsilon)^{3/2} \end{array}$$

and noting that $(1 + 4\varepsilon)^{3/2} = 1 + 6\varepsilon + O(\varepsilon^2)$, we obtain the desired result. ■

2.4.2 Error estimate for the cost function

Note that by corollary 74(iii) we have directly that

$$J_0(u_\varepsilon) - J_0(u_0) = O(||u_\varepsilon - u_0||_\infty) \tag{2.79}$$

which is bigger than $O(\varepsilon)$ for the four examples studied in subsection 2.4.1. Now we improve estimate (2.79) for $\ell = \ell_2, \ell_3$ and ℓ_4 by generalizing an argument suggested by Anton Schiela, in a personal communication, for the convex case (for example, when $\phi \equiv 0$) and for the logarithmic barrier.

Theorem 76 *Let $\ell = \ell_2, \ell_3, \ell_4$ (defined in subsection 2.4.1) and suppose that the assumptions of theorem 73 hold. Let $\bar{b} > 0$ be such that $(\mathcal{CP}_\varepsilon^{\bar{b},s})$ has a unique solution for $\varepsilon > 0$ small enough. Then*

$$J_0(u_\varepsilon) - J_0(u_0) = O(\varepsilon) \tag{2.80}$$

Proof. Since J_0 is of class C^2 we have that

$$J_0(u_0) \geq J_0(u_\varepsilon)+DJ_0(u_\varepsilon)(u_\varepsilon-u_0)-O\left(\sup_{z\in[u_\varepsilon,u_0]}||D^2J_0(z)||_{\mathcal{L}(\mathcal{Y}^s,\mathcal{Y}^s)}||u_\varepsilon-u_0||^2_\infty\right) \tag{2.81}$$

where $\mathcal{L}(\mathcal{Y}^s,\mathcal{Y}^s)$ denotes the space of continuous bilinear forms over $\mathcal{Y}^s \times \mathcal{Y}^s$. Expression (2.51) yields that $\sup_{z\in[u_\varepsilon,u_0]}||D^2J_0(z)||_{\mathcal{L}(\mathcal{Y}^s,\mathcal{Y}^s)}$ is uniformly bounded in ε. Therefore by (2.69), (2.71) and (2.74),

$$\sup_{z\in[u_\varepsilon,u_0]}||D^2J_0(z)||_{\mathcal{L}(\mathcal{Y}^s,\mathcal{Y}^s)}||u_\varepsilon-u_0||^2_\infty = O(||u_\varepsilon-u_0||^2_\infty) = O(\varepsilon). \tag{2.82}$$

On the other hand, optimality conditions for $(\mathcal{CP}^{\bar{b},s}_\varepsilon)$ yield that

$$DJ_0(u_\varepsilon) = -\varepsilon\ell'(u_\varepsilon), \tag{2.83}$$

hence, using (2.81) and (2.82), we have that

$$J_0(u_\varepsilon) - J_0(u_0) \leq -\varepsilon\int_\Omega \ell'(u_\varepsilon(x))(u_\varepsilon(x)-u_0(x))\mathrm{d}x + O(\varepsilon). \tag{2.84}$$

Since for $\ell_2(t)$ and $\ell_4(t)$ it holds that $\ell_2', \ell_4' \leq 0$, we obtain that

$$J_0(u_\varepsilon) - J_0(u_0) \leq -\varepsilon\int_\Omega \ell'(u_\varepsilon(x))u_\varepsilon(x)\mathrm{d}x + O(\varepsilon). \tag{2.85}$$

For ℓ_2 inequality (2.85) yields

$$J_0(u_\varepsilon) - J_0(u_0) \leq \varepsilon p\int_\Omega u_\varepsilon(x)^p\mathrm{d}x + O(\varepsilon) = O(\varepsilon),$$

by (2.25). For ℓ_4 inequality (2.85)

$$J_0(u_\varepsilon) - J_0(u_0) \leq -\varepsilon\, meas(\Omega) + O(\varepsilon) = O(\varepsilon).$$

Finally, for ℓ_3 inequality (2.84) implies that $J_0(u_\varepsilon) - J_0(u_0) \leq I_1 + I_2 + O(\varepsilon)$, where

$$\begin{array}{rcl} I_1 & := & -\varepsilon\int_{\{u_\varepsilon(x)\leq e^{-1}\}}\ell'(u_\varepsilon(x))(u_\varepsilon(x)-u_0(x))\mathrm{d}x \quad \text{and} \\ I_2 & := & -\varepsilon\int_{\{u_\varepsilon(x)\geq e^{-1}\}}\ell'(u_\varepsilon(x))(u_\varepsilon(x)-u_0(x))\mathrm{d}x. \end{array}$$

Since $u_\varepsilon\log u_\varepsilon$ is bounded uniformly in ε, we have that

$$I_1 \leq -\varepsilon\int_{\{u_\varepsilon(x)\leq e^{-1}\}}(1+\log u_\varepsilon(x))\,u_\varepsilon(x)\mathrm{d}x = O(\varepsilon)$$

and

$$I_2 = -\varepsilon\int_{\{u_\varepsilon(x)\geq e^{-1}\}}(1+\log u_\varepsilon(x))\,(u_\varepsilon(x)-u_0(x))\,\mathrm{d}x = O(\varepsilon)$$

by (2.25). ■

Part III

Stochastic optimal control theory

Chapter 3

Error estimates for the logarithmic barrier method in linear quadratic stochastic optimal control problems

Contents

3.1 Introduction

The study of stochastic linear quadratic (LQ) optimal control problems is an area of active research. In fact, many problems arising in engineering design and mathematical finance can be modeled as stochastil LQ problems. Let us cite, for example, the portfolio selection problem ([96, 66]) and the contingent claim problem ([59]). The stochastic LQ problem, in a finite time horizon $[0,T]$ and without constraints, can be stated as follows:

$$\text{Minimize } \mathbb{E}\left(\int_0^T \left[u(t)^\top R(t)u(t) + y(t)^\top C(t)y(t)\right] \mathrm{d}t + y(T)^\top M y(T)\right)$$
$$\text{s.t. } \begin{cases} \mathrm{d}y(t) &= [A_0(t)y(t) + B_0(t)u(t)]\,\mathrm{d}t + [A_1(t)y(t) + B_1(t)u(t)]\,\mathrm{d}W(t), \\ y(0) &= x_0 \end{cases}$$

Assuming that $R(t)$ is positive definite, the problem above was extensively investigated in the 1960s and 1970s (see e.g. [89, 70, 16, 17, 39], the surveys in [6] and references therein). In the mid-1990s, using an approach based on a stochastic Riccati equation, Chen-Li-Zhou [35] treated the stochastic LQ problem even when $R(t)$ can be indefinite. See also [36], where the relations between the stochastic LQ problem, the stochastic Pontryagin minimum principle (SPMP) and linear forward-backward stochastic differential equations, are studied.

Even if the unconstrained case is well studied, when control constraints are present the only reference that we know is [56]. In fact, the authors consider a stochastic LQ problem where the control is constrained in a cone. They obtain explicit solutions for the optimal control and the optimal cost via solutions of a system of extended stochastic Riccati equations.

In this work we study a convex stochastic LQ problem involving non-negativity control constraints. We consider a family of *logarithmic penalized* problems, parameterized by $\varepsilon > 0$. This means that the cost function is modified by adding a logarithmic barrier function multiplied by ε, which implies that the solution of the new problem is strictly positive. Our aim is to study the convergence, as $\varepsilon \downarrow 0$, of the solution of the penalized problem to the solution of the *initial* one. In fact, we will obtain error estimates for the cost, control, state and adjoint state in the appropriate spaces. This result extend the classical error estimates obtained by Weiser [85] in the deterministic framework.

The article is organized as follows: In section 3.2 we fix the standard notation and the initial and penalized problems are stated. Using the stochastic Pontryaguin minimum principle (SPMP) (see [8, 9, 15, 16, 18, 75, 31]), first order necessary and sufficient conditions are derived. Our main result is provided in section 3.3, in which we derive the error estimates. The proof use a simple duality argument and an application of the SPMP.

3.2 Problem Statement and Optimality Conditions

Let us first fix some notations. The space $\mathbb{R}^m$ ($m \in \mathbb{N}^*$) is endowed with its standard Euclidean norm denoted by $|\cdot|$. The ith coordinate of a vector x is denoted by x^i. We set $\mathbb{R}^m_+ := \{x \in \mathbb{R}^m : x^i \geq 0\}$, and $\mathbb{R}^m_{++} := \{x \in \mathbb{R}^m : x^i > 0\}$. Let $T > 0$ and consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, on which a d-dimensional ($d \in \mathbb{N}^*$) Brownian motion $W(\cdot)$ is defined with $\mathbb{F} = \{\mathcal{F}_t\}_{0 \leq t \leq T}$ being its natural filtration, augmented by all $\mathbb{P}$-null sets in $\mathcal{F}$. For $\ell \in \mathbb{N}^*$ let us define

$$\begin{array}{rcl} L^2_{\mathcal{F}}\left([0,T];\mathbb{R}^\ell\right) & := & \left\{v : [0,T] \times \Omega \to \mathbb{R}^\ell \ / \ v \text{ is adapted and } ||v||_2 < \infty\right\}, \\ L^{2,\infty}_{\mathcal{F}}\left([0,T];\mathbb{R}^\ell\right) & := & \left\{v : [0,T] \times \Omega \to \mathbb{R}^\ell \ / \ v \text{ is adapted and } ||v||_{2,\infty} < \infty\right\}, \end{array}$$

where we assume that all the mappings are $\mathcal{B}([0,T]) \times \mathcal{F}_T$-measurable and ´

$$||v||_2 := \left[\mathbb{E}\left(\int_0^T |v(t)|^2 \mathrm{d}t\right)\right]^{\frac{1}{2}}, \quad ||v||_{2,\infty} := \left[\mathbb{E}\left(\sup_{t\in[0,T]} |v(t)|^2\right)\right]^{\frac{1}{2}}.$$

It is well known that $\left(L^2_{\mathcal{F}}\left([0,T];\mathbb{R}^\ell\right), \langle\cdot,\cdot\rangle_2\right)$ is a Hilbert space, where

$$\langle u, v\rangle_2 := \sum_{i=1}^{\ell} \mathbb{E}\left(\int_0^T u^i(t)v^i(t)\mathrm{d}t\right). \tag{3.1}$$

Let $x_0 : \Omega \to \mathbb{R}^n$ be $\mathcal{F}_0$ measurable and such that $\mathbb{E}(|x_0|^2) < \infty$. Consider the following affine stochastic differential equation (SDE)

$$\begin{array}{rcl} \mathrm{d}y(t) & = & f(t,\omega,y(t),u(t))\mathrm{d}t + \sum_{i=1}^d \sigma^i(t,\omega,y(t),u(t))\mathrm{d}W(t), \\ y(0) & = & x_0 \in \mathbb{R}. \end{array} \tag{3.2}$$

In the notation above $y(t) \in \mathbb{R}^n$ denotes the state function, which is controlled by $u(t) \in \mathbb{R}^m$, and

$$f : [0,T] \times \Omega \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n, \quad \sigma^i : [0,T] \times \Omega \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n\times d}$$

are defined by

$$\begin{array}{rcl} f(t,\omega,y,u) & := & A_0(t,\omega)y + B_0(t,\omega)u + D_0(t,\omega), \\ \sigma^i(t,\omega,y,u) & := & A_i(t,\omega)y + B_i(t,\omega)u + D_i(t,\omega), \end{array}$$

where, for $i = 0, ..., d$, $A_i : [0,T] \times \Omega \to \mathbb{R}^{n\times n}$, $B_i : [0,T] \times \Omega \to \mathbb{R}^{n\times m}$ and $D_i : [0,T] \times \Omega \to \mathbb{R}^n$. We assume that:

(H1) The random matrices A_i, B_i, D_i are progressively measurable with respect to $\mathbb{F}$ and bounded uniformly in $(t,\omega) \in [0,T]$ by a constant $\bar{D} > 0$.

We take as state and control space, respectively,

$$\mathcal{Y} := L^{2,\infty}_{\mathcal{F}}([0,T];\mathbb{R}^m), \;\; \mathcal{U} := L^2_{\mathcal{F}}([0,T];\mathbb{R}^m). \tag{3.3}$$

It is well known that for every $u \in \mathcal{U}$, equation (3.2) has a unique solution $y_u \in \mathcal{Y}$ and the following estimate hold:

$$||y||^2_{2,\infty} \leq L_1 \left(\mathbb{E}(y_0^2) + ||u||^2_2 + \sum_{i=0}^{d} ||D_i||^2_2 \right), \tag{3.4}$$

for some positive constant L_1. Denote respectively by $\mathcal{S}^m_+$ and $\mathcal{S}^m_{++}$ the sets of symmetric positive semidefinite and symmetric positive definite matrices of order m. Now, let us consider the set

$$\mathcal{U}^+ := \{u \in \mathcal{U} \;/\; u(t,\omega) \geq 0 \;\text{ for a.a. } (t,\omega) \in [0,T] \times \Omega\}, \tag{3.5}$$

and the random matrices $R : [0,T] \times \Omega \to \mathcal{S}^m_{++}$, $C : [0,T] \times \Omega \to \mathcal{S}^n_+$, $M : \Omega \to \mathcal{S}^n_+$. We assume:

(H2) The matrices R, C, M are bounded uniformly in $(t,\omega) \in [0,T]$ by a constant $\bar{C}$. In addition, we assume that R is uniformly positive definite, i.e. there exists $\alpha > 0$ such that for a.a. $(t,\omega) \in [0,T] \times \Omega$

$$v^\top R(t,\omega) v \geq \alpha |v|^2 \;\text{ for all } v \in \mathbb{R}^m. \tag{3.6}$$

3.2.1 The initial problem

Let $\bar{y} \in \mathcal{Y}$ be a reference state function and define $g_0 : [0,T] \times \Omega \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as

$$g_0(t,\omega,y,u) := \tfrac{1}{2} u^\top R(t,\omega) u + [y - \bar{y}(t,\omega)]^\top C(t,\omega)[y - \bar{y}(t,\omega)]. \tag{3.7}$$

The cost function $J_0 : \mathcal{U} \to \mathbb{R}$ is defined as

$$J_0(u) := \mathbb{E}\left(\tfrac{1}{2} \int_0^T g_0(t, y(t), u(t)) \mathrm{d}t + \tfrac{1}{2} y_u(T)^\top M y_u(T) \right). \tag{3.8}$$

We consider the following stochastic optimal control problem:

$$\text{Min } J_0(u) \quad \text{subject to } u \in \mathcal{U}^+. \tag{$\mathcal{CP}$}_0$$

Assumptions **(H1)**, **(H2)** imply that J_0 is a strongly convex continuous function. Since $\mathcal{U}^+$ is closed and convex, we have that $(\mathcal{CP})_0$ has a unique solution u_0. We denote $y_0 := y_{u_0}$ its associated state.

As usual in optimal control theory, optimality conditions can be expressed in terms of a Hamiltonian and an adjoint state. In fact, let

$$H_0 : [0,T] \times \Omega \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times d} \times \mathbb{R}^m \to \mathbb{R}$$

be the *Hamiltonian* of problem $(\mathcal{CP})_0$, defined as

$$H_0(t,\omega,y,p,q,u) := g_0(t,\omega,y,u) + p \cdot f(t,\omega,y,u) + \sum_{i=1}^{d} q^i \cdot \sigma^i(t,\omega,y,u),$$

where q^i denotes the ith column of q. For $u \in \mathcal{U}$ let $(p_u, q_u) \in L^{2,\infty}_{\mathcal{F}}([0,T] \times \mathbb{R}^n) \times L^2_{\mathcal{F}}([0,T] \times \mathbb{R}^{n\times d})$, called the *adjoint state* associated to u, be the unique solution of the following linear backward stochastic differential equation (BSDE)(see [15]) :

$$\begin{array}{rcl} \mathrm{d}p(t) & = & -D_y H_0(t, y_u(t), p(t), q(t), u(t))\mathrm{d}t + q(t)\mathrm{d}W(t), \\ p(T) & = & My_u(T). \end{array} \tag{3.9}$$

It is well known (see e.g. [72, Proposition 3.1]) that there exists $L_2 > 0$, such that

$$||p_u||^2_{2,\infty} + ||q_u||^2_2 \leq L_2 \left(\mathbb{E}(y_u(T)^2) + ||u||^2_2\right). \tag{3.10}$$

Let us set $p_0 := p_{u_0}$ and $q_0 := q_{u_0}$. Since $g_0(t,\omega,y,\cdot)$ is strictly convex, the stochastic Pontryagin minimum principle (SPMP) for linear convex optimal control with random coefficients [31, Theorem 3.2], yields that u_0 is a solution of $(\mathcal{CP})_0$ if and only if for a.a. $(t,\omega) \in [0,T] \times \Omega$,

$$u_0(t,\omega) = \mathrm{argmin}_{w\in\mathbb{R}^m_+} H_0(t,\omega,y_0(t,\omega),p_0(t,\omega),q_0(t,\omega),w). \tag{3.11}$$

A straightforward computation (see [2, Section 2.1]) yields that

$$u_0(t,\omega) = \pi_0\left(R(t,\omega), z_0(t,\omega)\right) \quad \text{for a.a. } (t,\omega) \in [0,T] \times \Omega, \tag{3.12}$$

where

$$z_0(t,\omega) := -R(t,\omega)^{-1}\left[B_0(t,\omega)^\top p_0(t,\omega) + \sum_{i=1}^{d} B_i^\top(t,\omega)^\top q_0^i(t,\omega)\right]$$

and for $(R,z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$ the map $\pi_0(R,z)$ is defined as the unique solution of

$$\mathrm{Min}\ \tfrac{1}{2}(x-z)^\top R(x-z), \quad \text{s.t.} \quad x \in \mathbb{R}^m_+.$$

3.2.2 The penalized problem

For $\varepsilon > 0$ define the function $J_\varepsilon : \mathcal{U}^+ \to \mathbb{R} \cup \{+\infty\}$ by

$$J_\varepsilon(u) := \mathbb{E}\left(\tfrac{1}{2}\int_0^T \left[g_0(t, y_u(t), u(t)) + \varepsilon \hat{L}(u(t))\right] \mathrm{d}t + y_u(T)^\top \tfrac{1}{2} M y_u(T)\right), \tag{3.13}$$

where $\hat{L} : \mathbb{R}^m_+ \to \mathbb{R} \cup \{+\infty\}$ is defined as $\hat{L}(u) := -\sum_{i=1}^m \log u^i$. Let us consider the *penalized* problem

$$\text{Min } J_\varepsilon(u) \quad \text{subject to } u \in \mathcal{U}^+. \tag{$\mathcal{CP}$}_\varepsilon$$

Using the arguments of [2, Lemma 1], we have that

$$u \in \mathcal{U}^+ \to \mathbb{E}\left(\int_0^T \hat{L}(u(t))\mathrm{d}t\right) \in \mathbb{R} \cup \{+\infty\}$$

is convex lower-semicontinuous (l.s.c), hence J_ε is a strongly convex l.s.c. function. Therefore, $(\mathcal{CP})_\varepsilon$ has a unique solution u_ε with associated state $y_\varepsilon := y_{u_\varepsilon}$. The Hamiltonian for $(\mathcal{CP})_\varepsilon$

$$H_\varepsilon : [0,T] \times \Omega \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times d} \times \mathbb{R}^m_+ \to \mathbb{R} \cup \{+\infty\}$$

is defined as

$$H_\varepsilon(t, \omega, y, p, q, u) := H_0(t, \omega, y, p, q, u) + \varepsilon \hat{L}(u).$$

We set $(p_\varepsilon, q_\varepsilon) := (p_{u_\varepsilon}, q_{u_\varepsilon})$ for the unique solution of the following BSDE:

$$\begin{array}{rcl} \mathrm{d}p(t) & = & -D_y H_\varepsilon(t, y_\varepsilon(t), p(t), q(t), u_\varepsilon(t))\mathrm{d}t + q(t)\mathrm{d}W(t), \\ p(T) & = & M y_\varepsilon(T). \end{array} \tag{3.14}$$

As for the initial problem, the SPMP implies that u_ε is the solution of $(\mathcal{CP})_\varepsilon$ if and only if for a.a. $(t,\omega) \in [0,T] \times \Omega$

$$u_\varepsilon(t,\omega) = \mathrm{argmin}_{w \in \mathbb{R}^m_+} H_\varepsilon(t, \omega, y_\varepsilon(t,\omega), p_\varepsilon(t,\omega), q_\varepsilon(t,\omega), w), \tag{3.15}$$

Since $H_\varepsilon(t,\omega,\cdot)$ is convex and differentiable in u, condition (3.15) is satisfied if and only if for a.a. $(t,\omega) \in [0,T] \times \Omega$,

$$D_u H_0(t, \omega, y_\varepsilon(t,\omega), p_\varepsilon(t,\omega), q_\varepsilon(t,\omega), u_\varepsilon(t,\omega)) - \varepsilon \frac{1}{u_\varepsilon(t,\omega)} = 0, \tag{3.16}$$

where $1/u_\varepsilon(t,\omega) \in \mathbb{R}^m$ denotes the vector whose ith component is $1/u^i_\varepsilon(t,\omega)$. Equation (3.16) implies that (see [2, Section 2.2])

$$u_\varepsilon(t,\omega) = \pi_\varepsilon\left(R(t,\omega), z_\varepsilon(t,\omega)\right) \quad \text{for a.a. } (t,\omega) \in [0,T] \times \Omega, \tag{3.17}$$

where

$$z_\varepsilon(t,\omega) := -R(t,\omega)^{-1}\left[B_0(t,\omega)^\top p_\varepsilon(t) + \sum_{i=1}^{d} B_i^\top(t,\omega)^\top q_\varepsilon^i(t)\right]$$

and for $(R,z) \in \mathcal{S}^m_{++} \times \mathbb{R}^m$ the map $\pi_\varepsilon(R,z)$ is defined as the unique solution of

$$\text{Min } \tfrac{1}{2}(x-z)^\top R(x-z) + \varepsilon \hat{L}(x), \quad \text{s.t.} \quad x \in \mathbb{R}^m_+.$$

3.3 Main Result

In this section we provide error estimates for the cost, control, state and adjoint state of the penalized problem. We denote by $1/u_\varepsilon : [0,T]\times\Omega \to \mathbb{R}^m$ the mapping $(1/u_\varepsilon(t,\omega))^i := 1/u^i_\varepsilon(t,\omega)$.

Lemma 77 *For every $\varepsilon > 0$ we have that $1/u_\varepsilon \in \mathcal{U}^+$.*

Proof. The proof is based on (3.15). For notational convenience we assume that $n = m = d = 1$. The proof for the general case can be easily adapted. First, note that integrability problem comes when $u_\varepsilon(t,\omega)$ is small. Thus, fix $K_0 > 0$ and set

$$\Omega_{K_0} := \{(t,\omega) \in [0,T]\times\Omega \ / \ u_\varepsilon(t,\omega) \le K_0\}.$$

Now, let $\eta \in (0, K_0)$ and set

$$\hat{H}_\varepsilon(t,\omega,w) := H_\varepsilon(t,\omega,y_\varepsilon(t,\omega),p_\varepsilon(t,\omega),q_\varepsilon(t,\omega),w).$$

If $u_\varepsilon(t,\omega) \le \eta/2$ we have for a.a. $(t,\omega) \in \Omega_{K_0}$, omitting the (t,ω) argument,

$$\begin{aligned}\hat{H}_\varepsilon(\eta) - \hat{H}_\varepsilon(u_\varepsilon) &= \tfrac{1}{2}R(\eta+u_\varepsilon)(\eta-u_\varepsilon) + [B_0 p_\varepsilon + B_1 q_\varepsilon](\eta-u_\varepsilon) \\ &\quad +\varepsilon[\log(u_\varepsilon) - \log(\eta)]\end{aligned}$$

On the other hand, using that $\log(\cdot)$ is concave,

$$\log(u_\varepsilon) - \log(\eta) \le \frac{1}{\eta}(u_\varepsilon - \eta) \le \frac{1}{\eta}\frac{-\eta}{2} = -\tfrac{1}{2}.$$

Therefore, by optimality of u_ε,

$$0 \le \hat{H}_\varepsilon(\eta) - \hat{H}_\varepsilon(u_\varepsilon) \le \bar{C}K_0\eta + \bar{D}(|p_\varepsilon| + |q_\varepsilon|)\eta - \frac{\varepsilon}{2} \le \eta K_1(1 + |p_\varepsilon| + |q_\varepsilon|) - \tfrac{1}{2}\varepsilon,$$

where $K_1 := \max\left\{\bar{C}K_0, \bar{D}\right\}$. Thus, we conclude that

$$u_\varepsilon \leq \tfrac{1}{2}\eta \quad \Rightarrow \quad \eta \geq \frac{\varepsilon}{2K_1\left(1+|p_\varepsilon|+|q_\varepsilon|\right)}.$$

Henceforth, for a.a. $(t,\omega) \in \Omega_{K_0}$,

$$u_\varepsilon \geq \frac{\varepsilon}{4K_1\left(1+|p_\varepsilon|+|q_\varepsilon|\right)} \quad \text{and thus} \quad \frac{1}{u_\varepsilon} \leq \frac{4K_1\left(1+|p_\varepsilon|+|q_\varepsilon|\right)}{\varepsilon}. \tag{3.18}$$

The result follows from (3.10) using that $u_\varepsilon \in \mathcal{U}$ and that $y_\varepsilon \in \mathcal{Y}$ is almost surely continuous. ■

Remark 78 Estimate (3.18) generalizes [22, Theorem 1] obtained in the deterministic framework. In the deterministic case we have that u_ε is uniformly positive, whereas in our setting we can prove only (3.18).

Consider the Lagrangian $\mathcal{L} : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$, associated to problem $(\mathcal{CP})_0$, defined by

$$\mathcal{L}(u,\lambda) := J_0(u) - \langle \lambda, u\rangle_2, \tag{3.19}$$

where we recall that $\langle\cdot,\cdot\rangle_2$ is defined in (3.1). Define the *dual function* $d : \mathcal{U}^+ \to \mathbb{R}$ by $d(\lambda) := \inf_{u\in\mathcal{U}} \mathcal{L}(u,\lambda)$. We have:

Lemma 79 *For every $\varepsilon > 0$,*

$$d\left(\varepsilon\frac{1}{u^\varepsilon}\right) = J_0(u_\varepsilon) - \varepsilon m T.$$

Proof. Consider the following auxiliary problem

$$\text{Min } J_0(u) - \varepsilon\langle 1/u_\varepsilon, u\rangle_2 \ \text{ subject to } u \in \mathcal{U}. \tag{$\mathcal{CP}$}_{aux}$$

Lemma 77 implies that the above problem is well-defined. Since the cost function is strongly convex and continuous, problem $(\mathcal{CP})_{aux}$ admits a unique solution u_{aux}, with associated state $y_{aux} := y_{u_{aux}}$. The Hamiltonian H_{aux} of problem $(\mathcal{CP})_{aux}$ is defined as

$$H_{aux}(t,\omega,y,p,q,u) = H_0(t,\omega,y,p,q,u) - \varepsilon\sum_{i=1}^{m}\frac{1}{u^i_\varepsilon(t,\omega)}u^i.$$

We let (p_{aux}, q_{aux}) be the unique solution of the following BSDE:

$$\begin{array}{rcl} \mathrm{d}p(t) & = & -D_y H_{aux}(t, y_{aux}(t), p(t), q(t), u_{aux}(t))\mathrm{d}t + q(t)\mathrm{d}W(t), \\ p(T) & = & M y_{u_{aux}}(T). \end{array} \tag{3.20}$$

Define $\hat{H}_{aux} : [0,T] \times \Omega \times \mathbb{R}^m \to \mathbb{R}$ as

$$\hat{H}_{aux}(t,\omega,u) := H_{aux}(t,\omega, y_{u_{aux}}(t,\omega), p_{u_{aux}}(t,\omega), q_{u_{aux}}(t,\omega), u).$$

The SPMP yields that u_{aux} is a solution of $(\mathcal{CP})_{aux}$ if and only if

$$u_{aux}(t,\omega) = \text{argmin}_{w\in\mathbb{R}^m} \hat{H}_{aux}(t,\omega,w). \quad \text{for a.a. } (t,\omega) \in [0,T]\times\Omega. \quad (3.21)$$

Using that $\hat{H}_{aux}(t,\omega,\cdot)$ is convex and differentiable, (3.21) is satisfied if and only if

$$D_u H_0(t,\omega, y_{u_{aux}}(t,\omega), p_{u_{aux}}(t,\omega), q_{u_{aux}}(t,\omega), u) - \varepsilon \frac{1}{u_\varepsilon(t,\omega)} = 0. \qquad (3.22)$$

Therefore, noting that (ommiting the (t,ω) argument)

$$D_y H_{aux}(t,\omega, y_{aux}, p_{aux}, q_{aux}, u_{aux}) = D_y H_\varepsilon(t,\omega, y_{aux}, p_{aux}, q_{aux}, u_{aux}),$$

equations (3.14), (3.16) imply that $(y_\varepsilon, p_\varepsilon, q_\varepsilon, u_\varepsilon)$ satisfies (3.20)-(3.22). Therefore, $u_{aux} = u_\varepsilon$ solves $(\mathcal{CP})_{aux}$. Finally,

$$\text{Min}_{\ u\in\mathcal{U}}\ J_0(u) - \varepsilon\langle 1/u_\varepsilon, u\rangle = J_0(u_\varepsilon) - \varepsilon\langle 1/u_\varepsilon, u_\varepsilon\rangle = J_0(u_\varepsilon) - \varepsilon mT.$$

■

Now, we can prove our main result, which yields error bounds for $(y_\varepsilon, p_\varepsilon, q_\varepsilon, u_\varepsilon)$, usually referred as the *central path*. In particular, we obtain the convergence of $(y_\varepsilon, p_\varepsilon, q_\varepsilon, u_\varepsilon)$ to (y_0, p_0, q_0, u_0) in the appropriate spaces.

Theorem 80 *Assume that* ***(H1)*** *and* ***(H2)*** *hold. Then for every $\varepsilon > 0$, the following estimates hold*

$$J_0(u_\varepsilon) - J_0(u_0) \leq \varepsilon mT \quad (3.23)$$

$$||u_\varepsilon - u_0||_2^2 + ||y_\varepsilon - y_0||_{2,\infty}^2 + ||p_\varepsilon - p_0||_{2,\infty}^2 + ||q_\varepsilon - q_0||_2^2 \leq O(\varepsilon) \quad (3.24)$$

Proof. By lemma 79, we have

$$J_0(u_\varepsilon) - \varepsilon mT \leq \max_{\lambda\in\mathcal{U}^+}\min_{u\in\mathcal{U}} \mathcal{L}(u,\lambda) \leq \min_{u\in\mathcal{U}}\max_{\lambda\in\mathcal{U}^+} \mathcal{L}(u,\lambda) = \min_{u\in\mathcal{U}^+} J_0(u) = J_0(u_0),$$

from which (3.23) follows. The strong convexity of $J_0(\cdot)$ implies that

$$||u_\varepsilon - u_0||_2^2 = O(\varepsilon).$$

Taking $u = u_\varepsilon - u_0$ in (3.4) yields that

$$||y_\varepsilon - y_0||_{2,\infty}^2 = O(\varepsilon).$$

Finally, using the estimates above and that $y_\varepsilon - y_0$ is almost surely continuous, estimate (3.10) implies that

$$||p_\varepsilon - p_0||_{2,\infty}^2 + ||q_\varepsilon - q_0||_2^2 \leq L_2\left(\mathbb{E}\left[(y_\varepsilon(T) - y_0(T))^2\right] + ||u_\varepsilon - u_0||_2^2\right) = O(\varepsilon).$$

■

Chapter 4

First and second order necessary conditions for stochastic optimal control problems

Contents

4.1 Introduction

Because of its wide range of applications (e.g. in mathematical finance), stochastic optimal control theory is a very active research domain. In this work we consider the following type of stochastic optimal control problem

$$\begin{array}{ll} \text{Min } \mathbb{E}\left(\int_0^T \ell(t,y(t),u(t))\mathrm{d}t + \phi(y(T))\right) & \\ \text{s.t.} \quad \mathrm{d}y(t) = f(t,y(t),u(t))\mathrm{d}t + \sigma(t,y(t),u(t))\mathrm{d}W(t) & (\mathcal{SP}) \\ \qquad y(0) = y_0, \quad u(t,\omega) \in U \text{ for a.a. } (t,\omega), & \end{array}$$

where U is a nonempty, closed and convex subset of $\mathbb{R}^m$ and we suppose that the above stochastic differential equation (SDE) is well possed.

As in the case of deterministic optimal control problems, there are two main approaches to study problem $(\mathcal{SP})$. The first one is the global approach, based in the Bellman's dynamic programming principle, which yields that the value function of $(\mathcal{SP})$ is the unique viscocity solution of an associated second order Hamilton-Jacobi-Bellman equation. For a complete account of this point of view, widely used in practical computations, we refer the reader to the books [45, 76, 93]. The second approach is the variational one, which consists in to analyse the local behavior of the value function under small perturbations of a local minimum. Using this technique Kushner [61, 60, 63] Bensoussan [8, 9], Bismut [15, 16, 18] and Haussmann [53] obtained natural extensions of Pontryagin maximum principle to the stochastic case, that were generalized by Peng [75]. Relations between the global and variational approach are studied in [95].

Nevertheless, to the best of our knowledge, nothing has been said about second order optimality conditions. Using the variational technique we are able to obtain first and second order expansions for the cost function, which are expressed in terms of the derivatives of the Hamiltonian of problem $(\mathcal{SP})$. The main tool is a kind of generalization of Gronwall's lemma for the SDEs (proposition 81) obtained by Mou and Yong [72], which allows to expand the cost with respect to directions belonging to a more regular space than the control space. A similar idea was applied in [20] in the context of state constrained optimal control problems. By a density argument, we esablish first order optimality conditions, which include the case of not necessarily local constraints. In addition, under a polyhedricity assumption (see [52, 71]), we obtain second order necessary conditions which are the natural extensions of their deterministic counterparts.

The article is organized as follows: After introducing the standard notations and assumptions in section 4.2, we obtain in section 4.3 first and second order expansions for the state and cost function. In section 4.4, first and second order necessary conditions are proved and explicit results are given for

the case of box constraints. Finally, a discussion about a non gap second order sufficient condition is given in section 4.5.

4.2 Notations, assumptions and problem statement

Let us first fix some standard notation. For a x in a Euclidean space we will write x^i for its i-th coordinate and $|x|$ for its Euclidean norm. Let $T > 0$ and consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, on which a d-dimensional ($d \in \mathbb{N}^*$) Brownian motion $W(\cdot)$ is defined with $\mathbb{F} = \{\mathcal{F}_t\}_{0 \leq t \leq T}$ being its natural filtration, augmented by all $\mathbb{P}$-null sets in $\mathcal{F}$. Let $(X, ||\cdot||_X)$ be a Banach space and for $\beta \in [1, \infty)$ set

$$\begin{aligned} L^\beta(\Omega; X) &:= \left\{ v : \Omega \to X;\ v \text{ is measurable and } \mathbb{E}\left(||v(\omega)||_X^\beta \right) < \infty \right\}, \\ L^\infty(\Omega; X) &:= \left\{ v : \Omega \to X;\ v \text{ is measurable and ess sup}_{\omega \in \Omega} ||v(\omega)||_X < \infty \right\}. \end{aligned}$$

For $\beta, p \in [1, \infty]$ and $m \in \mathbb{N}$ let us define

$$L_{\mathcal{F}}^{\beta,p} := \left\{ v \in L^\beta\left(\Omega; L^p([0,T]; \mathbb{R}^m)\right);\ (t, \omega) \to v(t, \omega) := v(\omega)(t) \text{ is -adapted} \right\}.$$

We endow these space with the norms

$$||v||_{\beta,p} := \left[\mathbb{E}\left(||v(\omega)||_{L^p([0,T];\mathbb{R}^m)}^\beta \right) \right]^{\frac{1}{\beta}} \text{ and } ||v||_{\infty,p} := \text{ess sup}_{\omega \in \Omega} ||v(\omega)||_{L^p([0,T];\mathbb{R}^m)}.$$

For the sake of clarity, when the context is clear, the statement "for a.a. $t \in [0,T]$, a.s. $\omega \in \Omega$ ($\mathbb{P}$-a.s.)" will be simplified to "for a.a. (t, ω)". We will write $L_{\mathcal{F}}^p := L_{\mathcal{F}}^{p,p}$ and $||\cdot||_p := ||\cdot||_{p,p}$. The spaces $L_{\mathcal{F}}^{\beta,p}$ endowed with the norms $||\cdot||_{\beta,p}$ are Banach spaces and for the specific case $p = 2$ the space $L_{\mathcal{F}}^2$ is a Hilbert space. We will write $\langle \cdot, \cdot \rangle_2$ for the obvious scalar product. Evidently, for $\beta \in [1, \infty]$ and $1 \leq p_1 \leq p \leq p_2 \leq \infty$, there exist positive constants c_{β,p_1}, c_{β,p_2}, $c_{p_1,\beta}$, $c_{p_2,\beta}$ such that

$$c_{\beta,p_1}||v||_{\beta,p_1} \leq ||v||_{\beta,p} \leq c_{\beta,p_2}||v||_{\beta,p_2}, \quad c_{p_1,\beta}||v||_{p_1,\beta} \leq ||v||_{p,\beta} \leq c_{p_2,\beta}||v||_{p_2,\beta}$$

For a function $[0,T] \times \mathbb{R}^n \times \mathbb{R}^m \times \Omega \ni (t, y, u, \omega) \to \psi(t, y, u, \omega) \in \mathbb{R}^n$ which is $\mathcal{C}^2$ with respect to (y, u), set $\psi_y(t, y, u, \omega) := D_y\psi(t, y, u, \omega)$ and $\psi_u(t, y, u, \omega) := D_u\psi(t, y, u, \omega)$. As usual, when the context is clear, we will systematically omit the ω argument in the defined functions. Now let $z \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ be variations associated with y and u respectively. The second derivatives of ψ are written in the following form

$$\psi_{yy}(t, y, u)z^2 := D_{yy}^2\psi(t, y, u)(z, z); \quad \psi_{uu}(t, y, u)v^2 := D_{uu}^2\psi(t, y, u)(v, v);$$
$$\psi_{yu}(t, y, u)zv := D_{yv}^2\psi(t, y, u)(z, v).$$

Consider the maps $f, \sigma^i : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \times \Omega \to \mathbb{R}^n$ $(i = 1, ..., d)$. These maps will define the dynamics for our problem. Let us assume that:

(H1) [Assumptions for the dynamics] The maps $\psi = f, \sigma^i$ satisfy:

(i) The maps are $\mathcal{B}([0,T] \times \mathbb{R}^n \times \mathbb{R}^m) \otimes \mathcal{F}_T$-measurable.

(ii) For all $(y,u) \in \mathbb{R}^n \times \mathbb{R}^m$ the process $[0,T] \ni t \to \psi(t,y,u) \in \mathbb{R}^n$ is $\mathbb{F}$-adapted.

(iii) For almost all $(t,\omega) \in [0,T] \times \Omega$ the mapping $(y,u) \to \psi(t,y,u,\omega)$ is C^3. Moreover, we assume that there exists a constant $L_1 > 0$ such that for almost all (t,ω)

$$\left\{ \begin{array}{c} |\psi(t,y,u,\omega)| \leq L_1 \left(1 + |y| + |u|\right), \\ |\psi_y(t,y,u,\omega)| + |\psi_u(t,y,u,\omega)| \leq L_1, \\ |\psi_{yy}(t,y,u,\omega)| + |\psi_{yu}(t,y,u,\omega)| + |\psi_{uu}(t,y,u,\omega)| \leq L_1 \\ |D^2\psi(t,y,u,\omega) - D^2\psi(t,y',u',\omega)| \leq L_1 \left(|y-y'| + |u-u'|\right). \end{array} \right. \tag{4.1}$$

Let us define $\sigma(t,y,u) := (\sigma^1(t,y,u), ..., \sigma^d(t,y,u)) \in \mathbb{R}^{n \times d}$. For variations $z \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, associated with y and u, set

$$\begin{array}{rcl} \sigma_y(t,y,u)z & := & (\sigma^1_y(t,y,u)z, ..., \sigma^d_y(t,y,u)z), \\ \sigma_{yy}(t,y,u)z^2 & := & (\sigma^1_{yy}(t,y,u)z^2, ..., \sigma^d_{yy}(t,y,u)z^2), \end{array} \tag{4.2}$$

with analogous definitions for $\sigma_u(t,y,u)v$, $\sigma_{yu}(t,y,u)zv$ and $\sigma_{uu}(t,y,u)v^2$.

For every $\beta \geq 1$, let us define the space $\mathcal{Y}^\beta$ as

$$\mathcal{Y}^\beta := \left\{ y \in L^\beta\left(\Omega; C([0,T];\mathbb{R}^n)\right);\ (t,\omega) \to y(t,\omega) := y(\omega)(t) \text{ is } \mathbb{F}\text{-adapted} \right\}.$$

Let $y_0 : \Omega \to \mathbb{R}^n$ be $\mathcal{F}_0$ measurable and such that $\mathbb{E}(|y_0|^2) < \infty$. Under **(H1)**, we have that for every $u \in L^{\beta,2}_{\mathcal{F}}$ the SDE

$$\begin{array}{rcl} \mathrm{d}y(t) & = & f(t,y(t),u(t))\mathrm{d}t + \sigma(t,y(t),u(t))\mathrm{d}W(t), \\ y(0) & = & y_0 \end{array} \tag{4.3}$$

is well posed. In fact (see [72, Proposition 2.1]):

Proposition 81 *Suppose that **(H1)** holds. Then, there exists $C > 0$ such that for every $u \in L^{\beta,2}_{\mathcal{F}}$ $(\beta \geq 1)$ equation (4.3) has a unique solution $y \in \mathcal{Y}^\beta$ with continuous trayectories a.s. and*

$$\mathbb{E}\left(\sup_{t \in [0,T]} |y(t)|^\beta \right) \leq C\mathbb{E}\left(|y_0|^\beta + ||f(\cdot,0,u(\cdot))||^\beta_{\beta,1} + ||\sigma(\cdot,0,u(\cdot))||^\beta_{\beta,2} \right). \tag{4.4}$$

Remark 82 *Note that by the first condition in* (4.1)*, the right hand side of* (4.4) *is finite.*

Now, let us consider maps $\ell : [0,T]\times\mathbb{R}^n\times\mathbb{R}^m\times\Omega \to \mathbb{R}^n$ and $\phi : \mathbb{R}^n\times\Omega \to \mathbb{R}$. These maps will define the cost function of our problem. We assume:

(H2) [Assumptions for the cost maps] It holds that:

(i) The maps ℓ and ϕ are respectively $\mathcal{B}([0,T]\times\mathbb{R}^n\times\mathbb{R}^m)\otimes\mathcal{F}_T$ and $\mathcal{B}(\mathbb{R}^n)\otimes\mathcal{F}_T$ measurables.

(ii) For all $(y,u)\in\mathbb{R}^n\times\mathbb{R}^m$ the process $[0,T]\ni t\to\ell(t,y,u)\in\mathbb{R}^n$ is $\mathbb{F}$-adapted.

(iii) For almost all (t,ω) the maps $(y,u)\to\ell(t,y,u,\omega)$ and $y\to\phi(y,\omega)$ are $\mathcal{C}^2$. In addition, there exists $L_2>0$ such that:

$$\left\{\begin{array}{c} |\ell(t,y,u,\omega)|\le L_2\left(1+|y|+|u|\right)^2,\ |\phi(y,\omega)|\le L_2\left(1+|y|\right)^2, \\ |\ell_y(t,y,u,\omega)|+|\ell_u(t,y,u,\omega)|\le L_2\left(1+|y|+|u|\right), \\ |\ell_{yy}(t,y,u,\omega)|+|\ell_{yu}(t,y,u,\omega)|+|\ell_{uu}(t,y,u,\omega)|\le L_2, \\ |D^2\ell(t,y,u,\omega)-D^2\ell(t,y',u',\omega)|\le L_2\left(|y-y'|+|u-u'|\right), \\ |\phi_y(y,\omega)|\le L_2\left(1+|y|\right) \\ |\phi_{yy}(y,\omega)|\le L_2,\quad |\phi_{yy}(y,\omega)-\phi_{yy}(y',\omega)|\le L_2\left(|y-y'|\right). \end{array}\right. \tag{4.5}$$

Remark 83 *The assumptions above include the important case when the cost function is quadratic in* (y,u)*.*

In some of the results obtained in the sequel it will be useful to strengthen the second and fifth conditions in (4.5). In fact, as we will see in sections 4.3 and 4.4, under the assumption below the results obtained will be the natural extensions of the well know deterministic results.

[Lipschitz cost] There exists $C_\ell, C_\phi > 0$ such that for almost all $(t,\omega)\in[0,T]\times\Omega$ and for all (y,u), $(y',u')\in\mathbb{R}^n\times\mathbb{R}^m$ we have

$$\begin{array}{rcl} |\ell(t,y,u,\omega)-\ell(t,y',u',\omega)| & \le & C_\ell\left(|u-u'|+|y-y'|\right), \\ |\phi(y,\omega)-\phi(y',\omega)| & \le & C_\phi|y-y'|. \end{array} \tag{4.6}$$

For every $u\in L^2_{\mathcal{F}}$ denote by $y_u\in\mathcal{Y}^2$ the solution of (4.3). Let us define the function $J : L^2_{\mathcal{F}}\to\mathbb{R}$ by

$$J(u)=\mathbb{E}\left[\int_0^T \ell(t,y_u(t),u(t))\mathrm{d}t+\phi(y_u(T))\right]. \tag{4.7}$$

Note that, in view of the first condition in (4.5) and estimate (4.4) the function J is well defined. Let $\mathcal{U}$ be a nonempty closed and convex subset of $L^2_{\mathcal{F}}$ and consider the problem

$$\text{Min } J(u) \quad \text{subject to } u\in\mathcal{U}. \tag{$\mathcal{SP}$}$$

4.3 Expansions for the state and cost function

From now on we fix $\bar{u} \in L^2_{\mathcal{F}}([0,T];\mathbb{R}^m)$ and set $\bar{y} := y_{\bar{u}}$. We also suppose that assumptions **(H1)** and **(H2)** hold. For $\psi = f, \sigma$ and $t \in [0,T]$, define

$$\psi_y(t) = \psi_y(t,\bar{y}(t),\bar{u}(t)); \quad \psi_u(t) = \psi_u(t,\bar{y}(t),\bar{u}(t)), \quad \psi_{yu}(t) = \psi_{yu}(t,\bar{y}(t),\bar{u}(t));$$
$$\psi_{yy}(t) = \psi_{yy}(t,\bar{y}(t),\bar{u}(t)); \quad \psi_{uu}(t) = \psi_{uu}(t,\bar{y}(t),\bar{u}(t)).$$

Let $\beta \in [1,\infty]$ and $v \in L^{\beta,2}_{\mathcal{F}}$. We define $y_1[\bar{u}](v) \in \mathcal{Y}^\beta$ as the unique solution of

$$\begin{array}{rcl} \mathrm{d}y_1(t) & = & [f_y(t)y_1(t) + f_u(t)v(t)]\mathrm{d}t + [\sigma_y(t)y_1(t) + \sigma_u(t)v(t)]\mathrm{d}W(t), \\ y_1(0) & = & 0. \end{array} \tag{4.8}$$

The second assumption in (4.1) and proposition 81 yields that the mapping $v \in L^{\beta,2}_{\mathcal{F}} \to y_1[\bar{u}](v) \in \mathcal{Y}^\beta$ is well defined. If the context is clear, for notational convenience we will write $y_1 = y_1[\bar{u}](v)$. Also, let us define $\delta y = \delta y[\bar{u}](v)$ and $d_1 = d_1[\bar{u}](v)$ by

$$\delta y := y_{\bar{u}+v} - \bar{y}, \quad d_1 := \delta y - y_1. \tag{4.9}$$

Our aim now is to obtain a first order expansion of J around $\bar{u}$. For this purpose it will be useful to obtain bounds for y_1, δy and d_1. The main tool for obtaining such bounds is the following corollary of proposition 81, whose proof is straightforward.

Corollary 84 *Let $A_1, A_2 \in L^\infty_{\mathcal{F}}([0,T];\mathbb{R}^{n\times n})$, $B_1^i \in L^{\beta,2}_{\mathcal{F}}([0,T];\mathbb{R}^n)$ and $B_2^i \in L^\infty_{\mathcal{F}}([0,T];\mathbb{R}^{n\times d})$ for $i = 1,2$. Assume that there exists a constant $K > 0$ such that*

$$||B_1^1||_{\beta,1} \leq K||B_2^1||_{\beta,2}, \tag{4.10}$$

Then, omitting time from function arguments, for every $w \in L^{\beta,2}$, the SDE

$$\begin{array}{rcl} \mathrm{d}z & = & [A_1 z + B_1^1 + B_1^2 w]\,\mathrm{d}t + [A_2 z + B_2^1 + B_2^2 w]\,\mathrm{d}W(t) \\ z(0) & = & 0, \end{array} \tag{4.11}$$

has a unique solution in $\mathcal{Y}^\beta$ and the following estimate holds

$$\mathbb{E}\left(\sup_{t\in[0,T]} |z(t)|^\beta\right) = \begin{cases} O\left(\max\left\{||B_2^1||^\beta_{\beta,2}, ||w||^\beta_{\beta,1}\right\}\right) & \text{if } B_2^2 \equiv 0, \\ O\left(\max\left\{||B_2^1||^\beta_{\beta,2}, ||w||^\beta_{\beta,2}\right\}\right) & \text{otherwise.} \end{cases}$$

Remark 85 *Note that the estimates given in corollary 84 are sharp. In fact, suppose that $d = 1$ and let $w \in L^2([0,T];\mathbb{R})$. Consider the process $z(t)$ defined by*

$$z(t) := \int_0^t w(s)\mathrm{d}W(s) \quad \textit{for all } t \in [0,T].$$

We have that $\mathbb{E}\left(\sup_{t\in[0,T]} |z(t)|^\beta\right) \geq \mathbb{E}(|z(T)|^\beta) = ||w||_2^\beta \mathbb{E}(|Z|^\beta)$, where Z is an standard normal random variable. Since, in this specific case, $||w||_{\beta,2}^\beta = ||w||_2^\beta$, the conclusion follows.

Corollary 84 will be the main tool for establishing the following useful estimates:

Lemma 86 *Consider y_1 defined by (4.8) and δy, d_1 defined in (4.9). For every $\beta \geq 1$ and $v \in L_{\mathcal{F}}^{2\beta,4}$, the following estimates hold:*

$$\mathbb{E}\left(\sup_{t\in[0,T]} |\delta y|^\beta\right) = \begin{cases} O(||v||_{\beta,1}^\beta) & \textit{if } \sigma_u \equiv 0, \\ O(||v||_{\beta,2}^\beta) & \textit{otherwise.} \end{cases} \tag{4.12}$$

$$\mathbb{E}\left(\sup_{t\in[0,T]} |y_1|^\beta\right) = \begin{cases} O(||v||_{\beta,1}^\beta) & \textit{if } \sigma_u \equiv 0, \\ O(||v||_{\beta,2}^\beta) & \textit{otherwise.} \end{cases} \tag{4.13}$$

$$\mathbb{E}\left(\sup_{t\in[0,T]} |d_1|^\beta\right) = \begin{cases} O(||v||_{2\beta,2}^{2\beta}) & \textit{if } \sigma_{uu} \equiv 0, \\ O(||v||_{2\beta,4}^{2\beta}) & \textit{otherwise.} \end{cases} \tag{4.14}$$

Proof. For notational convenience we will suppose that $m = n = d = 1$. We have

$$\begin{array}{rcl} \mathrm{d}\delta y(t) & = & \left[\tilde{f}_y(t)\delta y(t) + \tilde{f}_u(t)v(t)\right]\mathrm{d}t + \left[\tilde{\sigma}_y(t)\delta y(t) + \tilde{\sigma}_u(t)v(t)\right]\mathrm{d}W(t), \\ \delta y(0) & = & 0. \end{array} \tag{4.15}$$

where, for $\psi = f, \sigma$,

$$\begin{array}{rcl} \tilde{\psi}_y(t) & := & \int_0^1 \psi_y\left(\bar{y}(t) + \theta\delta y(t), \bar{u}(t) + \theta v(t)\right)\mathrm{d}\theta, \\ \tilde{\psi}_u(t) & := & \int_0^1 \psi_u\left(\bar{y}(t) + \theta\delta y(t), \bar{u}(t) + \theta v(t)\right)\mathrm{d}\theta. \end{array}$$

Using the second assumption in (4.1), estimates (4.12), (4.13) follow from corollary 84 applied to (4.15) and (4.8) respectively.

We next prove (4.14). We have that

$$\begin{array}{rcl} \mathrm{d}d_1(t) & = & \left[\tilde{f}_y(t)\delta y(t) - f_y(t)y_1(t) + \left(\tilde{f}_u(t) - f_u(t)\right)v(t)\right]\mathrm{d}t + \\ & & \left[\tilde{\sigma}_y(t)\delta y(t) - \sigma_y(t)y_1(t) + \left(\tilde{\sigma}_u(t) - \sigma_u(t)\right)v(t)\right]\mathrm{d}W(t), \\ d_1(0) & = & 0. \end{array}$$

For $\psi = f, \sigma$, we have that $\left[\tilde{\psi}_y(t) - \psi_y(t)\right] y_1(t) = O\left([\ |\delta y(t)| + |v(t)|\ ]\, |y_1(t)|\right)$. Also,

$$[\tilde{\sigma}_u(t) - \sigma_u(t)]\, v(t) = \begin{cases} O\left(|\delta y(t)||v(t)|\right) & \text{if } \sigma_{uu} \equiv 0, \\ O\left([\ |\delta y(t)| + |v(t)|\ ]\, |v(t)|\right) & \text{otherwise.} \end{cases}$$

Therefore, the following equation holds for d_1:

$$\begin{aligned} \mathrm{d}d_1(t) &= \left[\tilde{f}_y(t)d_1(t) + O\left([|\delta y(t)| + |v(t)|][|y_1(t)| + |v(t)|]\right)\right]\mathrm{d}t + \\ &\quad [\tilde{\sigma}_y(t)d_1(t) + O\left(D(\delta y, y_1, v)\right)]\,\mathrm{d}W(t), \end{aligned}$$

where

$$D(\delta y(t), y_1(t), v(t)) = \begin{cases} [|\delta y(t)| + |v(t)|]\,[|y_1(t)| + |v(t)|] - |v(t)|^2 & \text{if } \sigma_{uu} \equiv 0, \\ [|\delta y(t)| + |v(t)|]\,[|y_1(t)| + |v(t)|] & \text{otherwise.} \end{cases}$$

By (4.12) and (4.13),

$$\begin{aligned} ||\ |\delta y||y_1|\ ||_{\beta,2}^{\beta} &= \mathbb{E}\left[\left(\int_0^T |\delta y(t)|^2|y_1(t)|^2\mathrm{d}t\right)^{\frac{\beta}{2}}\right] \\ &= O\left[\mathbb{E}\left(\sup|\delta y(t)|^{\beta}|y_1(t)|^{\beta}\right)\right] \\ &= O\left(\left[\mathbb{E}\left(\sup|\delta y(t)|^{2\beta}\right)\right]^{\frac{1}{2}}\left[\mathbb{E}\left(\sup|y_1(t)|^{2\beta}\right)\right]^{\frac{1}{2}}\right) \\ &= O(||v||_{2\beta,2}^{2\beta}). \end{aligned} \tag{4.16}$$

Also, by the Cauchy Schwarz inequality and (4.12), (4.13),

$$\begin{aligned} ||\ |y_1||v|\ ||_{\beta,2}^{\beta} = \mathbb{E}\left[\left(\int_0^T |y_1(t)|^2|v(t)|^2\mathrm{d}t\right)^{\frac{\beta}{2}}\right] &= O(||v||_{2\beta,2}^{2\beta}), \\ ||\ |\delta y||v|\ ||_{\beta,2}^{\beta} = \mathbb{E}\left[\left(\int_0^T |\delta y(t)|^2|v(t)|^2\mathrm{d}t\right)^{\frac{\beta}{2}}\right] &= O(||v||_{2\beta,2}^{2\beta}), \end{aligned}$$

and (4.14) follows by corollary 84, since $||v^2||_{\beta,1}^{\beta} = ||v||_{2\beta,2}^{2\beta}$ and $||v^2||_{\beta,2}^{\beta} = ||v||_{2\beta,4}^{2\beta}$. ■

The estimates obtained in lemma 86 will provide a first order expansion of J around $\bar{u}$. This expansion will be expressed, as usual, in terms of an adjoint state. Let $(\bar{p}, \bar{q}) \in L^2_{\mathcal{F}}([0,T];\mathbb{R}^n) \times (L^2_{\mathcal{F}}([0,T];\mathbb{R}^n))^d$ be the unique solution of the following backward stochastic differential equation (BSDE) (see [8, 18])

$$\begin{aligned} \mathrm{d}p(t) &= -\left[\ell_y(t)^{\top} + f_y(t)^{\top}p(t) + \sum_{i=1}^{m}\sigma_y^i(t)^{\top}q^i(t)\right]\mathrm{d}t + q(t)\mathrm{d}W(t), \\ p(T) &= \phi_y(\bar{y}(T))^{\top}. \end{aligned} \tag{4.17}$$

In the notation above σ^i and q^i denote respectively the ith column of σ and q. The following estimates hold (see [72, Proposition 3.1]):

Proposition 87 *Assume that* ***(H1)***, ***(H2)*** *hold and that* $\bar{u} \in L^{\beta,2}_{\mathcal{F}}$. *Then there exists* $C_q > 0$ *such that*

$$\mathbb{E}\left(\sup_{t\in[0,T]}|\bar{p}(t)|^\beta\right) + \sum_{i=1}^{d}||\bar{q}^i||^\beta_{\beta,2} \leq C_q\left(1+||\bar{u}||^\beta_{\beta,2}\right).$$

Define the Hamiltonian $H : [0,T]\times\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^{n\times d}\to\mathbb{R}$ by

$$H(t,y,u,p,q) := \ell(t,y,u) + p\cdot f(t,y,u) + \sum_{i=1}^{d} q^i\cdot\sigma^i(t,y,u), \tag{4.18}$$

and set $H_u(t) := H_u(t,\bar{y}(t),\bar{u}(t),\bar{p}(t),\bar{q}(t))$. Define $\Upsilon_1 : L^2_{\mathcal{F}}\to\mathbb{R}$ by

$$\Upsilon_1(v) := \mathbb{E}\left(\int_0^T H_u(t)\ v(t)\mathrm{d}t\right). \tag{4.19}$$

In view of proposition 87, with $\beta = 2$, the function Υ_1 is well defined. The following lemma is a consequence of Itô's lemma for multidimensional Itô process (see [93]).

Lemma 88 *Let* Z_1 *and* Z_2 *be* $\mathbb{R}^n$*-valued continuous process satisfying*

$$\begin{cases} \mathrm{d}Z_1(t) &= b_1(t)\mathrm{d}t + \sigma_1(t)\mathrm{d}W(t) \quad \textit{for all} \quad t\in[0,T], \\ \mathrm{d}Z_2(t) &= b_2(t)\mathrm{d}t + \sigma_2(t)\mathrm{d}W(t) \quad \textit{for all} \quad t\in[0,T], \end{cases} \tag{4.20}$$

where b_1, $b_2 \in L^2(\Omega, L^2([0,T],\mathbb{R}^n))$ *and* σ_1, $\sigma_2 \in L^2(\Omega, L^2([0,T],\mathbb{R}^{n\times d}))$ *are* $\mathbb{F}$*-adapted process. Also, let us suppose* $\mathbb{P}$*-a.s. we have that* $Z_1(0) = 0$. *Then*

$$\mathbb{E}\left(Z_1(T)\cdot Z_2(T)\right) = \mathbb{E}\left(\int_0^T\left[Z_1(t)\cdot b_2(t) + Z_2(t)\cdot b_1(t) + \sum_{i=1}^{d}\sigma_1^i(t)\cdot\sigma_2^i(t)\right]\mathrm{d}t\right).$$

Lemma 88 yields the following well known alternative expression for Υ_1.

Lemma 89 *For every* $v\in L^2_{\mathcal{F}}([0,T];\mathbb{R}^m)$ *we have that:*

$$\Upsilon_1(v) = \mathbb{E}\left(\int_0^T\left[\ell_y(t)y_1(t) + \ell_u(t)v(t)\right]\mathrm{d}t + \phi_y(\bar{y}(T))y_1(T)\right). \tag{4.21}$$

Proof. Noting that

$$\phi_y(\bar{y}(T))y_1(T) = \bar{p}(T)^\top y_1(T) - \bar{p}(0)^\top y_1(0),$$

lemma 88, applied to $Z_1 = y_1$ and $Z_2 = \bar{p}$, yields $\mathbb{E}\left(\phi_y(\bar{y}(T))y_1(T)\right) = I_1 + I_2 + I_3$, where

$$\begin{array}{rcl} I_1 & := & -\mathbb{E}\left(\int_0^T y_1(t)^\top \left[\ell_y(t)^\top + f_y(t)^\top \bar{p}(t) + \sum_{i=1}^d \int_0^T \sigma_y^i(t)^\top \bar{q}^i(t)\right] \mathrm{d}t\right), \\ I_2 & := & \mathbb{E}\left(\int_0^T \bar{p}(t)^\top \left[f_y(t)y_1(t) + f_u(t)v(t)\right] \mathrm{d}t\right), \\ I_3 & := & \sum_{i=1}^d \mathbb{E}\left(\int_0^T \bar{q}^i(t)^\top \left[\sigma_y^i(t)y_1(t) + \sigma_u^i(t)v(t)\right] \mathrm{d}t\right). \end{array}$$

Plugging the expressions of I_1, I_2 and I_3 introduced above into the right hand side of (4.21) yields the result. ■

The expression above for Υ_1 allows to obtain a *first order expansion* of J around $\bar{u}$.

Proposition 90 *Assume that* ***(H1)***, ***(H2)*** *hold and let* $v \in L^4_{\mathcal{F}}$. *Then,* $\Upsilon_1(v) = O(||v||_2)$ *and the following expansion holds* $J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + r_1(v)$ *with*

$$r_1(v) = \left\{ \begin{array}{ll} O\left(||v||^2_{4,2}\right) & \textit{if } \sigma_{uu} \equiv 0, \\ O\left(||v||^2_4\right) & \textit{otherwise.} \end{array} \right. \tag{4.22}$$

If in addition (4.6) *holds, then*

$$\Upsilon_1(v) = \left\{ \begin{array}{ll} O(||v||_1) & \textit{if } \sigma_u \equiv 0, \\ O(||v||_{1,2}) & \textit{otherwise,} \end{array} \right. \quad ; \quad r_1(v) = \left\{ \begin{array}{ll} O\left(||v||^2_2\right) & \textit{if } \sigma_{uu} \equiv 0, \\ O\left(||v||^2_{2,4}\right) & \textit{otherwise.} \end{array} \right. \tag{4.23}$$

Proof. Let us denote $\delta J := J(\bar{u}+v) - J(\bar{u})$. By definition

$$\begin{array}{rl} \delta J & = \mathbb{E}\left(\int_0^T \left[\ell(y_{\bar{u}+v}, \bar{u}+v) - \ell(\bar{y}, \bar{u})\right] \mathrm{d}t + \phi(y_{\bar{u}+v}(T)) - \phi(\bar{y}(T))\right) \\ & = \Upsilon_1(v) + r_1(v), \end{array}$$

where $r_1(v) = O\left(z_1(v) + z_2(v)\right)$ and

$$\begin{array}{rcl} z_1(v) & := & \mathbb{E}\left[\int_0^T |\ell_y(t)d_1(t)|\mathrm{d}t + \phi_y(\bar{y}(T))d_1(T)\right], \\ z_2(v) & := & \mathbb{E}\left(\sup_{t\in[0,T]} |\delta y(t)|^2\right) + ||v||_2^2. \end{array}$$

Now, we estimate $\Upsilon_1(v)$. By assumption **(H2)** and the Cauchy Schwartz inequality $\mathbb{E}\left(\int_0^T \ell_u(t)v(t)\mathrm{d}t\right) = O\left(||v||_2\right)$. On the other hand, by (4.13)

$$\begin{array}{rl} \mathbb{E}\left(\int_0^T \ell_y(t)y_1(t)\mathrm{d}t + \phi_y(\bar{y}(T))y_1(T)\right) & = O\left(\left[\mathbb{E}\left(\sup_{t\in[0,T]} |y_1(t)|^2\right)\right]^{\frac{1}{2}}\right) \\ & = O(||v||_2). \end{array}$$

Thus $\Upsilon_1(v) = O(||v||_2)$. If (4.6) holds, then $\mathbb{E}\left(\int_0^T \ell_u(t)v(t)\mathrm{d}t\right) = O\left(||v||_1\right)$, and

$$\mathbb{E}\left(\int_0^T \ell_y(t)y_1(t)\mathrm{d}t + \phi_y(\bar{y}(T))y_1(T)\right) = O\left(\mathbb{E}\left[\sup_{t\in[0,T]}|y_1(t)|\right]\right).$$

Thus, estimates for $\Upsilon_1(v)$ in (4.23) follow from (4.13) with $\beta = 1$. Let us estimate $r_1(v)$. Assumption **(H2)** and (4.12) imply that $z_2(v) = O(||v||_2^2)$. On the other hand, by **(H2)** and the Cauchy Schwarz inequality

$$z_1(v) = O\left(\left[\mathbb{E}\left(\sup_{t\in[0,T]}|d_1(t)|^2\right)\right]^{\frac{1}{2}}\right).$$

Thus (4.22) follows from estimates (4.14) with $\beta = 2$. If in addition (4.6) holds, then $z_1(v) = O\left(\mathbb{E}\left[\sup_{t\in[0,T]}|d_1(t)|\right]\right)$ and the estimates for $r_1(v)$ in (4.23) follows from (4.14) with $\beta = 1$. ■

Remark 91 *The above proof shows that the hypothesis for the perturbation v can be weakened. For example, if (4.6) holds and $\sigma_{uu} = 0$, for all $v \in L^2_{\mathcal{F}}$ we have that $J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + r_1(v)$ with $\Upsilon_1(v) = O(||v||_1)$ and $r_1(v) = O(||v||_2^2)$. Thus, in this case, the function J is differentiable at $\bar{u}$.*

Corollary 92 *Assume that **(H1)**, **(H2)** hold and let $v \in L^\infty_{\mathcal{F}}$. Then, $\Upsilon_1(v) = O(||v||_2)$ and $J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + r_1(v)$ with $r_1(v) = O(||v||_\infty^2)$.*

The *second order* linearization of $u \in L^2_{\mathcal{F}} \mapsto y_u \in \mathcal{Y}^2$ around $\bar{u}$ in the direction $v \in L^\infty_{\mathcal{F}}$ is defined as the unique solution $y_2 = y_2(v)$ of

$$\begin{array}{rcl} \mathrm{d}y_2(t) & = & \left[f_y(t)y_2(t) + \frac{1}{2}f_{yy}(t)y_1(t)^2 + f_{yu}(t)y_1(t)v(t) + \frac{1}{2}f_{uu}(t)v(t)^2\right]\mathrm{d}t \\ & & + \left[\sigma_y(t)y_2(t) + \frac{1}{2}\sigma_{yy}(t)y_1(t)^2 + \sigma_{yu}(t)y_1(t)v(t) + \frac{1}{2}\sigma_{uu}(t)v(t)^2\right]\mathrm{d}W(t); \\ y_2(0) & = & 0. \end{array} \tag{4.24}$$

Note that by the third assumption in (4.1) and proposition 81, we have that y_2 is well defined.

Lemma 93 *Consider y_2 defined in (4.24) and $d_2 := \delta y - y_1 - y_2 = d_1 - y_2$. For every $\beta \geq 1$ and $v \in L^\infty_{\mathcal{F}}$, the following estimates hold:*

$$\mathbb{E}\left(\sup_{t\in[0,T]}|y_2|^\beta\right) = \begin{cases} O(||v||^{2\beta}_{2\beta,2}) & \text{if } \sigma_{uu} \equiv 0, \\ O(||v||^{2\beta}_{2\beta,4}) & \text{otherwise.} \end{cases} \tag{4.25}$$

$$\mathbb{E}\left(\sup_{t\in[0,T]}|d_2|^\beta\right) = \begin{cases} O(||v||^{\beta}_{2\beta,2}||v||^{2\beta}_{4\beta,4}) & \text{if } \sigma_{uuu} \equiv 0, \\ O(||v||^{\beta}_{2\beta,2}||v||^{2\beta}_{4\beta,4} + ||v||^{3\beta}_{3\beta,6}) & \text{otherwise.} \end{cases} \tag{4.26}$$

Proof. As in the proof of lemma 86 we suppose that $m = n = d = 1$. We will use repeatedly that for every $\beta, p, q \in [1, \infty)$, we have

$$|| \, |v|^q ||^\beta_{\beta,p} = ||v||^{q\beta}_{q\beta,qp} \quad \text{for all } v \in L^{q\beta,qp}_{\mathcal{F}}.$$

Proof of (4.25)*:* Recall that, by **(H1)**, for $\psi = f, \sigma$ we assume that ψ_{yy}, ψ_{yu} and ψ_{uu} are bounded. Using (4.13),

$$||y_1^2||^\beta_{\beta,2} = \mathbb{E}\left[\left(\int_0^T |y_1(t)|^4 \mathrm{d}t\right)^{\frac{\beta}{2}}\right] = O\left[\mathbb{E}\left(\sup |y_1(t)|^{2\beta}\right)\right] = O\left(||v||^{2\beta}_{2\beta,2}\right). \tag{4.27}$$

Analogously, the estimates associated with the term $y_1 v$ is of order $||v||^{2\beta}_{2\beta,2}$. Estimate (4.25) follows from corollary 84 since $||v^2||^\beta_{\beta,1} = ||v||^{2\beta}_{2\beta,2}$ and $||v^2||^\beta_{\beta,2} = ||v||^{2\beta}_{2\beta,4}$.

Proof of (4.26): Recall that $d_2 = \delta y - y_1 - y_2$. We have, omitting time from the arguments,

$$\begin{array}{rcl} \mathrm{d}d_2(t) & = & \left[f_y d_2 + \frac{1}{2} f_{yy}([\delta y]^2 - y_1^2) + f_{yu}(\delta y - y_1)v + r_t(f)(\delta y, v)^2\right] \mathrm{d}t + \\ & & \left[\sigma_y(t) d_2 + \frac{1}{2}\sigma_{yy}([\delta y]^2 - y_1^2) + \sigma_{yu}(\delta y - y_1)v + r_t(\sigma)(\delta y, v)^2\right] \mathrm{d}W(t). \end{array}$$

where for $\psi = f, \sigma$ the map $r_t(\psi)$ is defined by

$$r_t(\psi) := \int_0^1 (1-\theta)\left[\psi_{yy}(\bar{y}(t) + \theta\delta y(t), \bar{u}(t) + \theta v(t)) - \psi_{yy}(\bar{y}(t), \bar{u}(t))\right] \mathrm{d}\theta.$$

Thus, since $[\delta y]^2 - y_1^2 = (\delta y + y_1) d_1$ and $D\psi$ is Lipschitz, we obtain

$$\begin{array}{rcl} \mathrm{d}d_2(t) & = & \left[f_y d_2 + O\left(|d_1|\left\{|\delta y| + |y_1|\right\} + |d_1||v| + \alpha_t(f)\right)\right] \mathrm{d}t + \\ & & \left[\sigma_y d_2 + O\left(|d_1|\left\{|\delta y| + |y_1|\right\} + |d_1||v| + \alpha_t(\sigma)\right)\right] \mathrm{d}W(t) \end{array} \tag{4.28}$$

where, for $\psi = f, \sigma$,

$$\alpha_t(\psi) := \left\{ \begin{array}{ll} |\delta y(t)|^3 + |v(t)|^3 & \text{if } \psi_{uuu} \neq 0, \\ |\delta y(t)|^3 + |\delta y(t)|\, |v(t)|^2 & \text{if } \psi_{uuu} \equiv 0. \end{array} \right.$$

Now, let us estimate the terms in the $\mathrm{d}W(t)$ part of (4.28),

$$\begin{array}{rcl} || \, |d_1||\delta y| \, ||^\beta_{\beta,2} & = & \mathbb{E}\left[\left(\int_0^T |d_1(t)|^2 |\delta y(t)|^2 \mathrm{d}t\right)^{\frac{\beta}{2}}\right] = O\left[\mathbb{E}\left(\sup |d_1(t)|^\beta |\delta y(t)|^\beta\right)\right] \\ & = & O(||v||^\beta_{2\beta,2} ||v||^{2\beta}_{4\beta,4}), \end{array}$$

by (4.12) and (4.14). Analogously, estimates for the terms $d_1 y_1$ and $d_1 v$ are of the same order. Let us estimate the terms appearing in $\alpha_\sigma(t)$. Using (4.12),

$$|| \, |\delta y|^3 ||^\beta_{\beta,2} = \mathbb{E}\left[\left(\int_0^T |\delta y(t)|^6 \mathrm{d}t\right)^{\frac{\beta}{2}}\right] = O\left[\mathbb{E}\left(\sup |\delta y(t)|^{3\beta}\right)\right] = O(||v||^{3\beta}_{3\beta,2}). \tag{4.29}$$

By (4.12), we obtain

$$\begin{aligned} || \, |\delta y||v|^2 \, ||^\beta_{\beta,2} &= \mathbb{E}\left[\left(\int_0^T |\delta y(t)|^2 |v(t)|^4 \mathrm{d}t\right)^{\frac{\beta}{2}}\right] \\ &= O\left(\mathbb{E}\left[\sup |\delta y(t)|^\beta \left(\int_0^T |v(t)|^4 \mathrm{d}t\right)^{\frac{\beta}{2}}\right]\right) = O(||v||^\beta_{2\beta,2} ||v||^{2\beta}_{4\beta,4}). \end{aligned}$$

Also, we have that $||v^3||^\beta_{\beta,1} = ||v||^{3\beta}_{3\beta,3}$ and $||v^3||^\beta_{\beta,2} = ||v||^{3\beta}_{3\beta,6}$. By the Cauchy Schwarz inequality,

$$||v||^{3\beta}_{3\beta,3} = \mathbb{E}\left[\left(\int_0^T |v(t)|^3 \mathrm{d}t\right)^\beta\right] \leq \mathbb{E}\left[\left(\int_0^T |v(t)|^2 \mathrm{d}t\right)^{\frac{\beta}{2}} \left(\int_0^T |v(t)|^4 \mathrm{d}t\right)^{\frac{\beta}{2}}\right].$$

Using the Cauchy Schwarz inequality again, we get $||v||^{3\beta}_{3\beta,3} = O(||v||^\beta_{2\beta,2} ||v||^{2\beta}_{4\beta,4})$. Therefore, estimate (4.26) follows from corollary 84. ■

Our aim now is to obtain a second order expansion of J around $\bar{u}$. Let us set $H_{(y,u)^2}(t) = H_{(y,u)^2}(t, \bar{y}(t), \bar{u}(t), \bar{p}(t), \bar{q}(t))$ and define $\Upsilon_2 : L^\infty_{\mathcal{F}} \to \mathbb{R}$ by

$$\Upsilon_2(v) := \mathbb{E}\left(\int_0^T H_{(y,u)^2}(t)(v(t), y_1(t))^2 \mathrm{d}t + \phi_{yy}(\bar{y}(T))(y_1(T))^2\right).$$

As for Υ_1 a useful alternative expression for Υ_2 holds.

Lemma 94 *For every $v \in L^\infty_{\mathcal{F}}$ we have that:*

$$\begin{aligned} \tfrac{1}{2}\Upsilon_2(v) &= \mathbb{E}\left(\int_0^T \left[\ell_y(t) y_2(t) + \tfrac{1}{2}\ell_{(y,u)^2}(t)(y_1(t), v(t))^2\right] \mathrm{d}t\right) \\ &+ \mathbb{E}\left[\phi_y(\bar{y}(T)) y_2(T) + \tfrac{1}{2}\phi_{yy}(\bar{y}(T))(y_1(T))^2\right]. \end{aligned} \tag{4.30}$$

Proof. By definition of y_2 and $\bar{p}$, we have that

$$\phi_y(\bar{y}(T)) y_2(T) = \bar{p}(T) \cdot y_2(T) - \bar{p}(0) \cdot y_2(0).$$

Lemma 88 yields $\mathbb{E}\left(\phi_y(\bar{y}(T)) y_2(T)\right)) = I_1' + I_2' + I_3'$, where

$$\begin{aligned} I_1' &:= -\mathbb{E}\left(\int_0^T y_2(t)^\top \left[\ell_y(t)^\top + f_y(t)^\top \bar{p}(t) + \textstyle\sum_{i=1}^d \int_0^T \sigma^i_y(t)^\top \bar{q}^i(t)\right] \mathrm{d}t\right), \\ I_2' &:= \mathbb{E}\left(\int_0^T \bar{p}(t)^\top \left[f_y(t) y_2(t) + \tfrac{1}{2} f_{(y,u)^2}(t)(y_1(t), v(t))^2\right] \mathrm{d}t\right), \\ I_3' &:= \textstyle\sum_{i=1}^d \mathbb{E}\left(\int_0^T \bar{q}^i(t)^\top \left[\sigma^i_y(t) y_2(t) + \tfrac{1}{2}\sigma^i_{(y,u)^2}(t)(y_1(t), v(t))^2\right] \mathrm{d}t\right). \end{aligned}$$

Plugging the expressions of I_1', I_2' and I_3' introduced above into the right hand side of (4.30) yields the result. ■

Now we are able to obtain a *second order expansion* of J around $\bar{u}$.

Proposition 95 *Assume that* ***(H1)***, ***(H2)*** *hold and let* $v \in L^\infty_{\mathcal{F}}$. *Then,*

$$J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + \tfrac{1}{2}\Upsilon_2(v) + r_2(v), \tag{4.31}$$

and the following estimates hold:

$$\Upsilon_2(v) = \begin{cases} O(||v||^2_{4,2}) & \text{if } \sigma_{uu} \equiv 0, \\ O(||v||^2_4) & \text{otherwise,} \end{cases} \quad r_2(v) = \begin{cases} O(||v||_\infty ||v||^2_{4,2}) & \text{if } \sigma_{uuu} \equiv 0, \\ O(||v||_\infty ||v||^2_4) & \text{otherwise.} \end{cases} \tag{4.32}$$

If in addition (4.6) *holds then*

$$\Upsilon_2(v) = \begin{cases} O(||v||^2_2) & \text{if } \sigma_{uu} \equiv 0, \\ O(||v||^2_{2,4}) & \text{otherwise,} \end{cases} \quad r_2(v) = \begin{cases} O(||v||_\infty ||v||^2_2) & \text{if } \sigma_{uuu} \equiv 0, \\ O(||v||_\infty ||v||^2_{2,4}) & \text{otherwise.} \end{cases} \tag{4.33}$$

Proof. Let us first estimate $\Upsilon_2(v)$ by using its expression obtained in lemma 94 and lemmas 86 and 93. By (4.13) with $\beta = 2$,

$$\mathbb{E}\left(\sup_{t\in[0,T]} |y_1(t)|^2 + \int_0^T |v(t)|^2 \mathrm{d}t \right) = O(||v||^2_2). \tag{4.34}$$

In view of assumption **(H2)** and (4.34) we obtain that

$$\mathbb{E}\left(\int_0^T \ell_{(y,u)^2}(t)(y_1(t), v(t))^2 \mathrm{d}t + \phi_{yy}(\bar{y}(T))(y_1(T))^2 \right) = O(||v||^2_2). \tag{4.35}$$

On the other hand, assumption **(H2)** and the Cauchy Schwartz inequality yield

$$\mathbb{E}\left(\int_0^T \ell_y(t) y_2(t) \mathrm{d}t + \phi_y(\bar{y}(T)) y_2(T) \right) = O\left(\left[\mathbb{E}(\sup_{t\in[0,T]} |y_2|^2) \right]^{\frac{1}{2}} \right), \tag{4.36}$$

and the estimate for $\Upsilon_2(v)$ in (4.32) follows from (4.25). If (4.6) holds, then

$$\mathbb{E}\left(\int_0^T \ell_y(t) y_2(t) \mathrm{d}t + \phi_{yy}(\bar{y}(T)) y_2(T) \right) = O\left(\mathbb{E}\left[\sup_{t\in[0,T]} |y_2| \right] \right), \tag{4.37}$$

and the estimate for $\Upsilon_2(v)$ in (4.33) follows from (4.25).

Now we proceed to obtain (4.31). As in the proof or proposition 90 we denote $\delta J := J(\bar{u}+v) - J(\bar{u})$. By definition,

$$\delta J = \mathbb{E}\left(\int_0^T \left[\ell(y_{\bar{u}+v}, \bar{u}+v) - \ell(\bar{y},\bar{u})\right] \mathrm{d}t + \phi(y_{\bar{u}+v}(T)) - \phi(\bar{y}(T))\right) = I_1 + I_2,$$

where, omitting the time argument in the integral,

$$\begin{array}{rcl} I_1 & := & \mathbb{E}\left(\int_0^T \left[\ell_y \delta y + \ell_u v + \frac{1}{2}\ell_{(y,u)^2}(\delta y, v)^2 + r_\ell(\delta y, v)^2\right] \mathrm{d}t\right), \\ I_2 & := & \mathbb{E}\left[\phi_y(\bar{y}(T))\delta y(T) + \frac{1}{2}\phi_{yy}(\bar{y}(T))(\delta y(T))^2 + r_\phi(\bar{y}(T))(\delta y(T))^2\right]. \end{array} \tag{4.38}$$

Recalling that $\delta y = y_1 + d_1 = y_1 + y_2 + d_2$, assumption (4.5) in **(H2)** yields

$$\begin{array}{rcl} I_1 & = & \mathbb{E}\left(\int_0^T \ell_y(t)(y_1+y_2) + \ell_u(t)v + \frac{1}{2}D^2\ell(t)(y_1,v)^2 \mathrm{d}t\right) + \mathbb{E}\left(\int_0^T \ell_y d_2 \mathrm{d}t\right) \\ & & + O(z_1(v)), \end{array}$$

where, omitting time from function arguments,

$$z_1(v) := \mathbb{E}\left(\sup\left[|d_1|^2 + |d_1(t)||y_1| + |\delta y|^3\right]\right) + ||v||_1 \mathbb{E}\left(\sup|d_1|\right) + ||v||_3^3.$$

On the other hand,

$$\begin{array}{rcl} I_2 & = & \mathbb{E}\left[\phi_y(\bar{y}(T))\left(y_1(T)+y_2(T)\right) + \frac{1}{2}\phi_{yy}(\bar{y}(T))\left(y_1(T)\right)^2\right] \\ & & + \mathbb{E}\left[\phi_y(\bar{y}(T))d_2(T)\right] + O(z_2(v)), \end{array}$$

where

$$z_2(v) := \mathbb{E}\left(|\delta y(T)|^3 + |y_1(T)||d_1(T)| + |d_1(T)|^2\right).$$

Denoting $z(v) := z_1(v) + z_2(v)$ we get that

$$\begin{array}{rcl} \delta J & = & \mathbb{E}\left(\int_0^T \left[\ell_y(t)(y_1(t)+y_2(t)) + \ell_u(t)v(t) + \frac{1}{2}\ell_{(y,u)^2}(t)(y_1(t),v(t))^2\right] \mathrm{d}t\right) \\ & & + \mathbb{E}\left[\phi_y(\bar{y}(T))(y_1(T)+y_2(T)) + \frac{1}{2}\phi_{yy}(\bar{y}(T))\left(y_1(T)\right)^2\right] + \zeta(v) + z(v), \end{array}$$

where,

$$\zeta(v) := \mathbb{E}\left(\int_0^T \ell_y(t)d_2(t)\mathrm{d}t + \phi_y(\bar{y}(T))d_2(T)\right). \tag{4.39}$$

Therefore, using (4.21) and (4.30), we get (4.31) with $r_2(v) := \zeta(v) + z(v)$. Now, we proceed to estimate $z(v)$. By (4.14) we have that

$$\mathbb{E}\left(\sup_{t\in[0,T]}|d_1(t)|^2\right) = O(||v||_4^4) = O(||v||_\infty^2 ||v||_2^2).$$

Estimates (4.13), (4.14) and the Cauchy Schwartz inequality yield

$$\mathbb{E}\left(\sup_{t\in[0,T]}|d_1(t)||y_1(t)|\right)=O\left(||v||_4^2||v||_2\right)=O\left(||v||_\infty||v||_2^2\right).$$

Analogously, using (4.14), we have

$$\mathbb{E}\left(||v||_1\sup_{t\in[0,T]}|d_1(t)|\right)=O\left(||v||_4^2||v||_{2,1}\right)=O\left(||v||_\infty||v||_2^2\right).$$

Estimate (4.12) yields $\mathbb{E}\left(\sup_{t\in[0,T]}|\delta y(t)|^3\right)=O(||v||_{3,2}^3)$. But

$$||v||_{3,2}^3=\mathbb{E}\left(\left[\int_0^T|v(t)|^2\mathrm{d}t\right]^{\frac{3}{2}}\right)=O\left(||v||_\infty||v||_2^2\right),$$

and$||v||_3^3=O\left(||v||_\infty||v||_2^2\right)$. Thus, $z(v)=O\left(||v||_\infty||v||_2^2\right)$. Finally, let us estimate $\zeta(v)$. Assumption **(H2)** and the Cauchy Schwartz inequality yield that

$$\zeta(v)=O\left(\left[\mathbb{E}\left(\sup_{t\in[0,T]}|d_2(t)|^2\right)\right]^{\frac{1}{2}}\right).$$

Hence, using (4.26) with $\beta=2$,

$$\zeta(v)=\begin{cases}O(||v||_{4,2}||v||_{8,4}^2) & \text{if } \sigma_{uuu}\equiv 0,\\ O(||v||_{4,2}||v||_{8,4}^2+||v||_{6,6}^3) & \text{otherwise.}\end{cases}$$

Since $O(||v||_{4,2}||v||_{8,4}^2)=O(||v||_\infty||v||_{4,2}^2)$ and $O(||v||_{6,6}^3)=O(||v||_\infty||v||_4^2)$, the estimate for $r_2(v)$ in (4.32) follows. If in addition assumption (4.6) holds, then by (4.26) with $\beta=1$,

$$\zeta(v)=O\left(\mathbb{E}\left[\sup_{t\in[0,T]}|d_2(t)|\right]\right)=\begin{cases}O(||v||_2||v||_4^2) & \text{if } \sigma_{uuu}\equiv 0,\\ O(||v||_2||v||_4^2+||v||_{3,6}^3) & \text{otherwise.}\end{cases}$$

Since $O(||v||_2||v||_4^2)=O(||v||_\infty||v||_2^2)$ and $O(||v||_{3,6}^3)=O(||v||_{2,4}^2)$, the estimate for $r_2(v)$ in (4.33) follows. ■

Remark 96 (i) *Since Υ_2 is a quadratic form and, for every $\beta,p\in[1,\infty]$, the space $L_{\mathcal{F}}^\infty$ is dense in $L_{\mathcal{F}}^{\beta,p}$, we have that: If $\Upsilon_2(v)=O(||v||_{\beta,p})$ then Υ_2 admits a unique continuous extension in $L^{\beta,p}$.*
(ii) *The proof of proposition 4.31 shows that the estimates $\Upsilon_2(v)=O(||v||_2^2)$ and $r_2(v)=O(||v||_\infty||v||_2^2)$ also hold in the case when f and σ are affine mappings, since in this case $y_2=d_2=0$.*

The following corollary will allow us to state second order necessary condition with respect to perturbations $v \in L^\infty_{\mathcal{F}}$.

Corollary 97 *Assume that* ***(H1)***, ***(H2)*** *hold and either* (4.6) *holds and* $\sigma_{uu} \equiv 0$, *or* f *and* σ *are affine mappings. Then, the following expansion holds:*

$$J(\bar{u}+v) = J(\bar{u}) + \Upsilon_1(v) + \tfrac{1}{2}\Upsilon_2(v) + r(v) \quad \textit{for all } v \in L^\infty_{\mathcal{F}}, \tag{4.40}$$

where $\Upsilon_1(v) = O(||v||_2)$, $\Upsilon_2(v) = O(||v||_2^2)$ *and* $r(v) = O(||v||_\infty ||v||_2^2)$.

4.4 Necessary optimality conditions

The asymptotic expansions obtained for J in section 4.3 allow us to obtain first and second order necessary conditions at a local optimum $\bar{u} \in L^2_{\mathcal{F}}$ for the control constrained problem $(\mathcal{SP})$. We first obtain first order optimality conditions using the procedure explained at the introduction: According to the regularity of the data of $(\mathcal{SP})$ and the dependence on u of the σ-term, a perturbation in an appropriate space is taken. Then, the results of the previous section yield a positivity condition of Υ_1 over a certain cone which is extended, by a density argument, to a larger one. Similar considerations apply in order to establish second order necessary conditions. Finally, we give a second order sufficient condition for the unconstrained case and we briefly discuss the difficulties arising in the constrained case.

Let us first fix some notations which are standard in optimization theory. Consider a Banach space $(X, ||\cdot||_X)$ and a nonempty closed convex set $C \subseteq X$. For $x, x' \in X$ define the *segment* $[x, x'] := \{x + \lambda(x'-x)\ ; \lambda \in [0,1]\}$. The radial, the tangent and the normal cone to C at $\bar{x}$ are defined respectively by

$$\begin{array}{rcl} \mathcal{R}_C(\bar{x}) & := & \{h \in X\ ;\ \exists\ \sigma > 0 \quad \text{such that} \quad [\bar{x}, \bar{x}+\sigma h] \subseteq C\}, \\ \mathcal{T}_C(\bar{x}) & := & \{h \in X\ ;\ \exists\ x(\sigma) = \bar{x} + \sigma h + o(\sigma) \in C,\ \sigma > 0,\ ||o(\sigma)/\sigma||_X \to 0\}, \\ N_{\mathcal{U}}(\bar{u}) & := & \{h^* \in X^*\ /\ \langle x^*, x\rangle_{X^*,X} \leq 0,\ \text{for all } h \in \mathcal{T}_C(\bar{x})\}\,, \end{array} \tag{4.41}$$

where X^* denotes the dual space of X and $\langle \cdot, \cdot \rangle_{X^*,X}$ is the duality product. Recall that, since C is a closed convex set, the cone $\mathcal{T}_C(\bar{x})$ is the adherence of $\mathcal{R}_C(\bar{x})$ in X.

4.4.1 First order necessary conditions

Consider as in section 4.3 a fixed $\bar{u} \in L^2_{\mathcal{F}}$. For $\beta, p \in [1, \infty]$ and a subset $A \subseteq L^{\beta,p}_{\mathcal{F}}$ we write $\mathrm{adh}_{\beta,p}(A)$ for the adherence of A in $L^{\beta,p}_{\mathcal{F}}$. If $A \subseteq L^{\beta}_{\mathcal{F}}$ we

write $\mathrm{adh}_\beta(A) := \mathrm{adh}_{\beta,\beta}(A)$.

We have the following first order conditions for $(\mathcal{SP})$.

Lemma 98 *Assume that* ***(H1)***, ***(H2)*** *hold and let $\bar{u} \in \mathcal{U}$ be a local solution of $(\mathcal{SP})$. Then:*

$$\Upsilon_1(v) \geq 0 \quad \textit{for all } v \in \begin{cases} \mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{4,2}_{\mathcal{F}}\right) & \textit{if } \sigma_{uu} \equiv 0, \\ \mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{4}_{\mathcal{F}}\right) & \textit{otherwise.} \end{cases} \tag{4.42}$$

If in addition (4.6) *holds then*

$$\Upsilon_1(v) \geq 0 \quad \textit{for all } v \in \begin{cases} \mathrm{adh}_1\left(\mathcal{R}_{\mathcal{U}}(\bar{u})\right) & \textit{if } \sigma_{uu} \equiv 0, \\ \mathrm{adh}_{1,2}\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{2,4}_{\mathcal{F}}\right) & \textit{otherwise.} \end{cases} \tag{4.43}$$

Proof. Let $v \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^4_{\mathcal{F}}$. Proposition 90 implies that, for $\sigma > 0$ small enough, we have

$$0 \leq J(\bar{u} + \sigma v) - J(\bar{u}) = \sigma \Upsilon_1(v) + ||v||_4^2 O(\sigma^2). \tag{4.44}$$

Thus, dividing by σ in (4.44) and letting $\sigma \downarrow 0$, we have that $\Upsilon_1(v) \geq 0$. Analogously, if $\sigma_{uu} = 0$ we have that $\Upsilon_1(v) \geq 0$ for all $v \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{4,2}_{\mathcal{F}}$. Condition (4.42) follows from the fact that, by proposition 90, $v \in L^4_{\mathcal{F}} \to \Upsilon_1(v)$ can be extended continuously to $L^2_{\mathcal{F}}$. The proof of (4.43) follows in the same manner, with the obvious modifications. ■

Note that the results obtained in lemma 98 are rather general, since they include the case of non local constraints. On the other hand, for some constraints the result gives no information. In fact, consider the following example.

Example 2 *Let $u_0 \in L^2_{\mathcal{F}}$ and suppose that $u_0 \notin L^{\beta,p}_{\mathcal{F}}$ for any $\beta, p \in (2, \infty]$. The constraint $\mathcal{U} := \{u = \alpha u_0 \ / \ \textit{for some } \alpha \in [0,1]\}$ is such that, at $\bar{u} = 0$, the radial cone is given by $\mathcal{R}_{\mathcal{U}}(\bar{u}) = \{\lambda u_0 \ / \ \textit{for } \lambda \geq 0\}$, but $\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\beta,p}_{\mathcal{F}} = \{0\}$.*

Thus, we will assume the following assumption over the constraint set $\mathcal{U}$:

(H3) For every $\bar{u} \in \mathcal{U}$ we have that

$$\mathcal{T}_{\mathcal{U}}(\bar{u}) = \mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}}\right). \tag{4.45}$$

We have the following proposition whose proof is straightforward.

Proposition 99 *Asumme that* ***(H1)***, ***(H2)***, ***(H3)*** *hold and let* $\bar{u}$ *be a local solution of* $(\mathcal{SP})$. *Then*

$$\Upsilon_1(v) \geq 0 \quad \textit{for all } v \in \mathcal{T}_{\mathcal{U}}(\bar{u}). \tag{4.46}$$

Remark 100 *Note that if* $J(\cdot)$ *is convex, then* (4.46) *is a sufficient condition for the (global) optimality of* $\bar{u}$.

Clearly, we have that **(H3)** can hold for non local constraints. As an example, it can be checked that (4.45) holds for and $\mathcal{U} = \{u \in L^2_{\mathcal{F}} \ / \ ||u||_2 \leq 1\}$ and $\bar{u} \in \mathcal{U}$. Now we consider the case when $\mathcal{U}$ is defined by *local constraints*. Let $(t,\omega) \in [0,T] \times \Omega \to U(t,\omega) \in \mathcal{P}(\mathbb{R}^m)$ be a $B([0,T]) \times \mathcal{F}_T$ measurable multifunction satisfying that

(i) For all a.a. t the multifunction $U(t,\cdot)$ is $\mathcal{F}_t$-measurable.
(ii) For a.a. (t,ω) we have that $U(t,\omega)$ is a closed convex subset of $\mathbb{R}^m$.
We set

$$\mathcal{U} := \left\{u \in L^2_{\mathcal{F}} \ ; \ u(t,\omega) \in U(t,\omega), \quad \text{a.a. } (t,\omega) \in [0,T] \times \Omega\right\}. \tag{4.47}$$

Lemma 101 *Suppose that* $\bar{u} \in \mathcal{U}$, *where* $\mathcal{U}$ *is given by* (4.47). *Then,*
(i) *Assumption* (4.45) *holds at* $\bar{u}$.
(ii) *The tangent cone is given by*

$$\mathcal{T}_{\mathcal{U}}(\bar{u}) = \left\{v \in L^2_{\mathcal{F}} \ ; \ v(t,\omega) \in \mathcal{T}_{U(t,\omega)}(\bar{u}(t,\omega)) \ \textit{a.a. } (t,\omega) \in [0,T] \times \Omega\right\}. \tag{4.48}$$

Proof. (i) By a diagonal argument, it suffices to prove that for every $v \in \mathcal{R}_{\mathcal{U}}(\bar{u})$ there exists a sequence $v_k \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^\infty_{\mathcal{F}}$ such that $||v_k - v||_2 \to 0$. Indeed, set

$$v_k(t,\omega) := \begin{cases} v(t,\omega) & \text{if } |v(t,\omega)| \leq k \\ 0 & \text{otherwise.} \end{cases} \tag{4.49}$$

The convexity of $U(t,\omega)$ yields that $v_k \in \mathcal{R}_{\mathcal{U}}(\bar{u})$. Also, $v_k(t,\omega) \to v(t,\omega)$ as $k \to \infty$ for a.a. (t,ω). The convergence in $L^2_{\mathcal{F}}$ follows by the dominated convergence theorem.
(ii) Let $v \in \mathcal{T}_{\mathcal{U}}(\bar{u})$. By definition, for σ small enough and a.a. (t,ω)

$$\bar{u}(t,\omega) + \sigma v(t,\omega) + r_\sigma(t,\omega) \in U(t,\omega),$$

where $r_\sigma(\cdot,\cdot)/\sigma \to 0$ in $L^2_{\mathcal{F}}$ as $\sigma \downarrow 0$. Thus, extracting a subsequence if necessary, we have that $r_\sigma(t,\omega)/\sigma \to 0$ for a.a. (t,ω) from which we deduce that $v(t,\omega) \in \mathcal{T}_{U(t,\omega)}(\bar{u}(t,\omega))$. Conversely, let v belongs to the r.h.s. of (4.48) and for $\varepsilon > 0$ set

$$v_\varepsilon := \varepsilon^{-1}\left(P_{\mathcal{U}}(\bar{u} + \varepsilon v) - \bar{u}\right), \tag{4.50}$$

where $P_{\mathcal{U}}(\cdot)$ denotes the orthogonal projection in $L^2_{\mathcal{F}}$ onto $\mathcal{U}$. By definition of $P_{\mathcal{U}}(\cdot)$ we have that $v_\varepsilon \in \mathcal{R}_{\mathcal{U}}(\bar{u})$. For (t,ω) in $[0,T]\times\Omega$ set $P_{U(t,\omega)}(\cdot)$ for the orthogonal projection in $\mathbb{R}^m$ onto $U(t,\omega)$. Definition of v_ε in (4.50) implies that for a.a. (t,ω)

$$v_\varepsilon(t,\omega) := \varepsilon^{-1}\left(P_{\mathcal{U}(t,\omega)}(\bar{u}(t,\omega)+\varepsilon v(t,\omega)) - \bar{u}(t,\omega)\right).$$

Clearly, $v_\varepsilon(t,\omega) \in \mathcal{R}_{U(t,\omega)}(\bar{u}(t,\omega))$ and for a.a. (t,ω) we have $v_\varepsilon(t,\omega) \to v(t,\omega)$. Since $|v_\varepsilon(t,\omega)| \le |v(t,\omega)|$, the dominated convergence theorem implies that $v_\varepsilon \to v$ in $L^2_{\mathcal{F}}$. Using that $v_\varepsilon \in \mathcal{R}_{\mathcal{U}}(\bar{u})$ we obtain that $v \in \mathcal{T}_{\mathcal{U}}(\bar{u})$. ■

Let $a, b \in \overline{\mathbb{R}}^m$ with $-\infty \le a^i < b^i \le +\infty$ for all $i \in \{1,...,m\}$ and define

$$U_{a,b} := \left\{x \in \mathbb{R}^m \ ; \ a^i \le x^i \le b^i\right\}. \tag{4.51}$$

For $u \in L^2_{\mathcal{F}}$ and every index $i \in \{1,...,m\}$, set

$$\begin{array}{rcl} I_{a^i}(u) & := & \{(t,\omega) \in [0,T]\times\Omega \ ; \ u^i(t,\omega) = a^i\}, \\ I_{b^i}(u) & := & \{(t,\omega) \in [0,T]\times\Omega \ ; \ u^i(t,\omega) = b^i\}. \end{array}$$

The following corollary is a direct consequence of proposition 99 and lemma 101.

Corollary 102 *Assume that* ***(H1)***, ***(H2)*** *hold suppose that* $\mathcal{U}$ *is in the form* (4.47). *Let* $\bar{u} \in \mathcal{U}$ *be a local solution of* $(\mathcal{SP})$, *then*

$$H_u(t,\omega)v \ge 0 \quad \textit{for all } v \in \mathcal{T}_{U(t,\omega)}(\bar{u}(t,\omega)). \tag{4.52}$$

In particular, if $U(t,\omega) \equiv U_{a,b}$ *(defined in* (4.51)*), then for every* $i \in \{1,...,m\}$

$$H^i_u(t,\omega) = \begin{cases} \ge 0 & \textit{if } (t,\omega) \in I_{a^i}(\bar{u}), \\ \le 0 & \textit{if } (t,\omega) \in I_{b^i}(\bar{u}), \\ = 0 & \textit{elsewhere.} \end{cases} \tag{4.53}$$

Remark 103 *Since* (4.52) *is equivalent to* (4.46) *when* $\mathcal{U}$ *is in the form* (4.47), *we have that if* $J(\cdot)$ *is convex then* (4.52) *is a sufficient condition for the (global) optimality of* $\bar{u}$.

4.4.2 Second order necessary conditions

In order to obtain second order necessary conditions for $(\mathcal{SP})$ we proceed as in the previous section, i.e. we prove a general result and after, under

some standard assumptions, we yield a more precise characterization for the important case of local constraints. Let us define

$$\Upsilon_1^{\perp} := \left\{ v \in L^2_{\mathcal{F}} \; ; \; \Upsilon_1(v) = 0 \right\}. \tag{4.54}$$

We have the following general second order necessary conditions.

Proposition 104 *Assume that* ***(H1)***, ***(H2)*** *hold and let* $\bar{u} \in \mathcal{U}$ *be a local solution of* $(\mathcal{SP})$. *Then, the following second order necessary condition holds:*

$$\Upsilon_2(v) \geq 0 \;\; \textit{for all } v \in \begin{cases} \mathrm{adh}_{4,2}\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}\right) & \textit{if } \sigma_{uu} \equiv 0, \\ \mathrm{adh}_{4}\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}\right) & \textit{otherwise.} \end{cases} \tag{4.55}$$

If in addition (4.6) *holds, or* f *and* σ *are affine mappings, then*

$$\Upsilon_2(v) \geq 0 \;\; \textit{for all } v \in \begin{cases} \mathrm{adh}_{2}\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}\right) & \textit{if } \sigma_{uu} \equiv 0, \\ \mathrm{adh}_{2,4}\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}\right) & \textit{otherwise.} \end{cases} \tag{4.56}$$

Proof. If $v \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}$, then for σ small enough

$$0 \leq J(\bar{u} + \sigma v) - J(\bar{u}) = \frac{\sigma^2}{2} \Upsilon_2(v) + \sigma^3 O(||v||^3_{\infty}).$$

Dividing the above equation by σ and letting $\sigma \downarrow 0$ yields $\Upsilon_2(v) \geq 0$ and the result follows from remark 96 (i). ■

The critical cone to $\mathcal{U}$ at $\bar{u}$ are defined by

$$C(\bar{u}) \;\; := \;\; \{ v^* \in \mathcal{T}_{\mathcal{U}}(\bar{u}) \; / \; \Upsilon_1(v) \leq 0 \}. \tag{4.57}$$

In order to obtain more precise second order necessary conditions, we suppose standard assumptions in the second order analysis of problems with convex constraints. The first one is a natural extension of **(H3)** to the second order case.

(H4) For every $\bar{u} \in \mathcal{U}$ and $v^* \in N_{\mathcal{U}}(\bar{u})$ (recall (4.41)), we have that

$$\mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap (v^*)^{\perp}\right) = \mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap (v^*)^{\perp}\right). \tag{4.58}$$

For our second assumption, we need the following notion of polyhedricity (see [52, 71]). The set $\mathcal{U}$ is said to be *polyhedric* at $\bar{u} \in \mathcal{U}$ if for all $v^* \in N_{\mathcal{U}}(\bar{u})$, the set $\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap (v^*)^{\perp}$ is dense in $\mathcal{T}_{\mathcal{U}}(\bar{u}) \cap (v^*)^{\perp}$ with respect to the $||\cdot||_2$ norm. If $\mathcal{U}$ is polyhedric at each $u \in \mathcal{U}$ we say that $\mathcal{U}$ is *polyhedric*.

Remark 105 *Note that, if* ***(H1)- (H3)*** *hold, proposition 99 yields that, at a local minimum,* $-\Upsilon_1 \in N_{\mathcal{U}}(\bar{u})$ *and* $C(\bar{u}) = \mathcal{T}_{\mathcal{U}}(\bar{u}) \cap \Upsilon_1^{\perp}$. *Thus, if* $\mathcal{U}$ *is polyhedric and* ***(H4)*** *holds,*

$$\mathrm{adh}_2\left(\mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}\right) = C(\bar{u}) \tag{4.59}$$

We state a second order necessary condition which is a natural extension of the deterministic counterpart.

Theorem 106 *Let* $\bar{u}$ *be a local solution of* $(\mathcal{SP})$ *and assume that*

(i) *Assumptions* ***(H1)-(H4)*** *hold.*

(ii) *Either* (4.6) *holds and* $\sigma_{uu} = 0$ *or* f *and* σ *are affine mappings.*

(iii) *The constraint set* $\mathcal{U}$ *is polyhedric.*

Then, the following second order necessary condition hold at $\bar{u}$*:*

$$\Upsilon_2(v) \geq 0 \quad \textit{for all } v \in C(\bar{u}). \tag{4.60}$$

Proof. As in the proof of proposition 104 we have that $\Upsilon_2(v) \geq 0$ for all $v \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap \Upsilon_1^{\perp}$. The resut follows, by remark 96 (i), since under our assumptions $\Upsilon_2(v) = O(||v||_2^2)$ and (4.59) holds. ■

Now, let us focus our attention in local constraints, i.e. when $\mathcal{U}$ is defined by (4.47).

Lemma 107 *Let* $\mathcal{U}$ *be defined by* (4.47) *and let* $\bar{u} \in \mathcal{U}$. *It holds that*

(i) *The normal cone* $N_{\mathcal{U}}(\bar{u})$ *is given by*

$$N_{\mathcal{U}}(\bar{u}) = \left\{v^* \in L^2_{\mathcal{F}} \ / \ v^*(t,\omega) \in N_{U(t,\omega)}(\bar{u}(t,\omega)), \quad a.a.\ (t,\omega) \in [0,T] \times \Omega\right\}. \tag{4.61}$$

(ii) *For every* $v^* \in N_{\mathcal{U}}(\bar{u})$ *we have that*

$$\mathcal{T}_{\mathcal{U}}(\bar{u}) \cap (v^*)^{\perp} = \{v \in \mathcal{T}_{\mathcal{U}}(\bar{u}) \ / \ v^*(t,\omega) \cdot v(t,\omega) = 0, \quad a.a.\ (t,\omega) \in [0,T] \times \Omega\}. \tag{4.62}$$

Proof. Since (ii) follows directly from (i) and lemma 101 (ii), it is enough to show (i). By lemma 101 (ii), the r.h.s. of (4.61) is included in $N_{\mathcal{U}}(\bar{u})$. To prove the other inclusion, let us argue by contradiction. Let $v^* \in N_{\mathcal{U}}(\bar{u})$ and suppose that it does not belong to the r.h.s. of (4.61). Then we can find a non null measurable set $\mathcal{A} \subseteq \mathcal{B}([0.T]) \otimes \mathcal{F}$ such that for each $(t,\omega) \in \mathcal{A}$ there

exists $v(t,\omega) \in \mathcal{T}_{U(t,\omega)}(\bar{u}(t,\omega))$, which can be taken with $|v(t,\omega)| = 1$, such that $v^*(t,\omega) \cdot v(t,\omega) > \alpha$, for some $\alpha > 0$. Defining $\hat{v} \in L^2_{\mathcal{F}}$ by

$$\hat{v}(t,\omega) := \begin{cases} v(t,\omega) & \text{if } v^*(t,\omega) \cdot v(t,\omega) > \alpha, \\ 0 & \text{otherwhise,} \end{cases}$$

we see that $\hat{v} \in \mathcal{T}_{\mathcal{U}}(\bar{u})$ and $\langle v^*, \hat{v} \rangle_2 > 0$ and thus we obtain a contradiction with the fact that $v^* \in N_{\mathcal{U}}(\bar{u})$. ■

In order to verify the polyhedricity assumption in the case of local constraints, we will need in fact to assume that for a.a. (t,ω) the set $U(t,\omega)$ is a polyhedron. More precisely, let $q \in \mathbb{N}$ and suppose that there exist mappings $\Sigma : [0,T] \times \Omega \to \mathcal{P}(\{1,...,q\})$, $a_i : [0,T] \times \Omega \to \mathbb{R}^m$, $b_i : [0,T] \times \Omega \to \mathbb{R}^m$, where $i \in \{1,...,q\}$, such that Σ, a_i and b_i are $\mathcal{B}([0,T]) \times \mathcal{F}_T$ measurable and for each t we have that $\Sigma(t,\cdot)$, $a_i(t,\cdot)$ and $b_i(t,\cdot)$ are $\mathcal{F}_t$ measurable. We suppose that

$$U(t,\omega) = \{x \in \mathbb{R}^m \ / \ \langle a_i(t,\omega), x \rangle \leq b_i(t,\omega), \text{ for } i \in \Sigma(t,\omega) \}. \tag{4.63}$$

We have

Lemma 108 *The set of local constraints $\mathcal{U}$ defined in* (4.47)*, with $U(t,\omega)$ given by* (4.63)*, is polyhedric and satisfies* (4.58)*.*

Proof. Let $\bar{u} \in \mathcal{U}$ and $v^* \in N_{\mathcal{U}}(\bar{u})$. For $v \in \mathcal{T}_{\mathcal{U}}(\bar{u}) \cap (v^*)^{\perp}$ and $k \geq 0$ set

$$\widehat{v}_k(t,\omega) := \begin{cases} v(t,\omega) & \text{if } |v(t,\omega)| \leq k \text{ and } \bar{u}(t,\omega) + \frac{1}{k} v(t,\omega) \in U(t,\omega), \\ 0 & \text{otherwise.} \end{cases} \tag{4.64}$$

Lemma 107(ii) implies that $\widehat{v}_k \in \mathcal{R}_{\mathcal{U}}(\bar{u}) \cap L^{\infty}_{\mathcal{F}} \cap (v^*)^{\perp}$. On the other hand, since $U(t,\omega)$ is a polyhedron, lemma 101(ii) implies that $v(t,\omega) \in \mathcal{T}_{\mathcal{U}}(\bar{u}(t,\omega)) = \mathcal{R}_{\mathcal{U}}(\bar{u}(t,\omega))$. Thus, as $k \uparrow \infty$, we have that $\widehat{v}_k \to v(t,\omega)$ for a.a. (t,ω). The dominated convergence theorem, yields that $\widehat{v}_k \to v$ in $L^2_{\mathcal{F}}$, hence $\mathcal{U}$ is polyhedric and (4.58) holds. ■

The following corollary is a direct consequence of theorem 106 and lemmas 107, 108.

Corollary 109 *Assume that **(H1)** - **(H2)** hold and let $\bar{u}$ be a local solution of $(\mathcal{SP})$ where $\mathcal{U}$ is defined in* (4.47)*, with $U(t,\omega)$ given by* (4.63)*. Further, suppose that either* (4.6) *holds and $\sigma_{uu} = 0$ or f and σ are affine mappings. Then, the following second order necessary conditions hold at $\bar{u}$:*

$$\Upsilon_2(v) \geq 0, \text{ for all } v \in \mathcal{T}_{\mathcal{U}}(\bar{u}) \text{ such that } H_u(t)v(t,\omega) = 0 \text{ for a.a. } (t,\omega).$$

4.5 On the second order sufficient condition

Let us first consider the unconstrained case, i.e. when $\mathcal{U} = L^2_{\mathcal{F}}$. Note that, in this specific case, **(H3)** is trivially satisfied and for every $\bar{u} \in \mathcal{U}$ it holds that $\mathcal{T}_{\mathcal{U}}(\bar{u}) = L^2_{\mathcal{F}}$. The following proposition is a consequence of corollary 97.

Proposition 110 *Assume that **(H1)**, **(H2)** hold and that $\mathcal{U} = L^2_{\mathcal{F}}$. Further, let us assume that either* (4.6) *holds and $\sigma_{uu} \equiv 0$, or f and σ are affine mappings. Suppose there exist $\alpha > 0$ such that $\bar{u} \in L^2_{\mathcal{F}}$ satisfies:*

$$\Upsilon_1(v) = 0, \text{ and } \Upsilon_2(v) \geq \alpha ||v||_2^2 \quad \text{for all } v \in L^2_{\mathcal{F}}. \tag{4.65}$$

Then, there exists $\delta > 0$ such that for all $v' \in L^\infty_{\mathcal{F}}$ with $||v'||_\infty \leq \delta$, we have

$$J(\bar{u} + v') \geq J(\bar{u}) + \tfrac{1}{2}\alpha ||v'||_2^2. \tag{4.66}$$

Only very partial results are obtained when $\mathcal{U} \neq L^2_{\mathcal{F}}$. Let us recall that a quadratic form $Q : H \to \mathbb{R}$, where H is a Hilbert space, is a Legendre form if it is weakly lower semi continuous (w.l.s.c.) quadratic form over H, such that, if $h_k \to h$ weakly in H and $Q(h_k) \to Q(h)$, then $h_k \to h$ strongly. We have the following proposition, whose proof follows the lines of the parallel deterministic result (see [24, Section 3.3]):

Proposition 111 *Assume that **(H1)**, **(H2)** hold and that that either* (4.6) *holds and $\sigma_{uu} \equiv 0$, or f and σ are affine mappings. Suppose that at $\bar{u} \in \mathcal{U}$, the quadratic form Υ_2 is a Legendre form and there exist $\alpha > 0$ such that*

$$\Upsilon_1(v) = 0, \text{ and } \Upsilon_2(v) \geq \alpha ||v||_2^2 \quad \text{for all } v \in C(\bar{u}). \tag{4.67}$$

Then, there exists $\delta > 0$ such that for all $u \in \mathcal{U}$ with $||u - \bar{u}||_\infty \leq \delta$, we have

$$J(u) \geq J(\bar{u}) + \tfrac{1}{2}\alpha ||u - \bar{u}||_2^2. \tag{4.68}$$

In the deterministic case there is a well known sufficient condition for the associated quadratic form to be a Legendre form, which is based essentialy in the fact that the application $u \in L^2([0,T];\mathbb{R}^m) \to y_1(u)(T) \in \mathbb{R}^n$ is weakly continuous. We show with two examples that $u \in L^2_{\mathcal{F}} \to y_1(u)(T) \in L^2_{\mathcal{F}_T}(\mathbb{R}^n)$ is not weakly continuous.

Example 3 (σ dependent on u) *Let us take $m = n = 1$ and let us consider the dynamics*

$$\mathrm{d}y_1(t) = u(t)\mathrm{d}W(t) \text{ for } t \in [0,T]; \quad y_1(0) = 0.$$

Let u_n be a (deterministic) orthonormal base of $L^2([0,T];\mathbb{R})$ and denote $y_n := y_1(u_n)$. By the dominated convergence theorem it is easy to check that u_n converges weakly to 0 in $L^2_{\mathcal{F}}$, but

$$\mathbb{E}\left[y_n(T)^2\right] = \mathbb{E}\left[\left(\int_0^T u_n(t)\mathrm{d}W(t)\right)^2\right] = \int_0^T u_n^2(t)\mathrm{d}t = 1.$$

Example 4 (σ independent on u) *We take $m = n = 1$ and $T = 2$. Let us consider the dynamics*

$$\mathrm{d}y_1(t) = u(t)\mathrm{d}t \text{ for } t \in [0,T]; \quad y_1(0) = 0.$$

Let ϕ_n be an orthonormal base of the Hilbert space $L^2(\mathbb{R})$ endowed with the scalar product

$$\langle g, h\rangle_* := \int_{-\infty}^{+\infty} g(x)h(x)e^{\frac{-x^2}{2}}\mathrm{d}x,$$

and consider the sequence $u_n \in L^2_{\mathcal{F}}$ defined by $u_n(t) := \phi_n(W(1))\mathbb{I}_{(1,2]}(t)$ and set $y_n := y_1(u_n)$. For every $f \in L^2_{\mathcal{F}}$, we have

$$\begin{aligned}\mathbb{E}\left(\int_0^2 f(t)u_n(t)\ \mathrm{d}t\right) &= \mathbb{E}\left(\phi_n(W(1))\int_1^2 f(t)\mathrm{d}t\right),\\ &= \mathbb{E}\left[\phi_n(W(1))\mathbb{E}\left(\int_1^2 f(t)\mathrm{d}t|W(1)\right)\right] \to 0,\end{aligned}$$

by definition of ϕ_n. Thus u_n converges weakly to 0 in $L^2_{\mathcal{F}}$. On the other hand,

$$\mathbb{E}\left(y_n(T)^2\right) = \mathbb{E}\left(\left[\int_0^2 u_n\mathrm{d}t\right]^2\right) = \mathbb{E}\left(\phi_n(W(1))^2\right) = 1.$$

Bibliography

[1] R.A. Adams. *Sobolev spaces*. Academic Press, New York, 1975.

[2] F. Alvarez, J. Bolte, J.F. Bonnans, and F. Silva. Asymptotic expansions for interior penalty solutions of control constrained linear-quadratic problems. *INRIA Report RR-6863*, 2009.

[3] F. Alvarez, J.F. Bonnans, and J. Laurent-Varin. Asymptotic expansion of the optimal control under logarithmic penalty: worked example and open problems. INRIA Report RR-6170, 2007.

[4] T. Appel, A. Rösch, and G. Winkler. Optimal control in non-convex domains: A priori discretization error estimates. *Calcolo*, 44:137–158, 2007.

[5] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Comp. Optim. Appls.*, 23(2):201–229, 2002.

[6] M. Athens. Special issues on linear-quadratic-gaussian problem. *IEEE Trans. Auto. Control.*, AC-16:527–869, 1971.

[7] V.E. Benes. Existence of optimal control laws. *SIAM J. on Control and Optimization*, 9:446–472, 1971.

[8] A. Bensoussan. *Lectures on stochastic control*. Lecture Notes in Math. Vol. 972, Springer-Verlag, Berlin.

[9] A. Bensoussan. Stochastic maximum principle for distributed parameter system. *J. Franklin Inst.*, 315:387–406, 1983.

[10] Alain Bensoussan. *Perturbation methods in optimal control*. Wiley/Gauthier-Villars Series in Modern Applied Mathematics. John Wiley & Sons Ltd., Chichester, 1988. Translated from the French by C. Tomson.

[11] N. Bérend, J.F. Bonnans, J. Laurent-Varin, M. Haddou, and C. Talbot. Fast linear algebra for multiarc trajectory optimization. In G. Di Pillo and M. Roma, editors, *Large scale nonlinear optimization*, volume 83 of *Nonconvex Optimization and Its Applications*, pages 1–14. Springer, 2006.

[12] N. Bérend, J.F. Bonnans, J. Laurent-Varin, M. Haddou, and C. Talbot. An interior-point approach to trajectory optimization. *AIAA J. of Guidance, Control and Dynamics*, 30(5):1228–1238, 2007.

[13] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of a Moreau-Yosida-based active set strategy and interior point methods for constrained optimal control problems. *SIAM Journal on Optimization*, 11:495–521 (electronic), 2000.

[14] J.T. Betts, S.K. Eldersveld, P.D. Frank, and J.G. Lewis. An interior-point algorithm for large scale optimization. In *Large-scale PDE-constrained optimization (Santa Fe, NM, 2001)*, volume 30 of *Lect. Notes Comput. Sci. Eng.*, pages 184–198. Springer, Berlin, 2003.

[15] J.M. Bismut. *Analyse convexe et probabilités*. PhD thesis, Faculté des Sciences de Paris, 1973.

[16] J.M. Bismut. Théorie probabiliste du contrôle des diffusions. *Mem. Amer. Math. Soc.*, 4:1–130, 1976.

[17] J.M. Bismut. On optimal control of linear stochastic equations with a linear-quadratic criterion. *SIAM J. Control Optim.*, 15:1–4, 1977.

[18] J.M. Bismut. An introductory approach to duality in optimal stochastic control. *SIAM Rev.*, 20:62–78, 1978.

[19] J.F. Bonnans. Second order analysis for control constrained optimal control problems of semilinear elliptic systems. *Applied Math. Optimization*, 38-3:303–325, 1998.

[20] J.F. Bonnans and E. Casas. On the choice of the function spaces for some state-constrained control problems. *Numerical Functional Analysis and Optimization*, 7–4:333–348, 1984.

[21] J.F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization: theoretical and numerical aspects*. Universitext. Springer-Verlag, Berlin, 2006. second edition.

[22] 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.

[23] J.F. Bonnans and J. Laurent-Varin. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. *Numerische Mathematik*, 103(1):1–10, 2006.

[24] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer-Verlag, New York, 2000.

[25] J.F. Bonnans and F. Silva. Asymptotic expansion for the solution of a penalized control constrained semilinear elliptic problem. *INRIA Report RR-7126*, 2009.

[26] J.F. Bonnans and F. Silva. Error estimates for the logarithmic barrier method in stochastic linear quadratic optimal control problems. *INRIA Report RR-7455*, 2010.

[27] J.F. Bonnans and F. Silva. First and second order necessary conditions for stochastic optimal control problems. *INRIA Report RR-7154*, 2010.

[28] V. S. Borkar. Controlled diffusion processes. *Probability Surveys*, 2:213–244, 2005.

[29] H. Brézis. Problèmes unilatéraux. *J. Mathématiques pures et appliquées*, 51:1–168, 1972.

[30] H. Brézis. *Analyse Fonctionnelle. Théorie et Applications*. Collection Mathématiques Appliquées pour la Maîtrise. Paris: Masson, 1983.

[31] A. Cadenillas and I. Karatzas. The stochastic maximum principle for linear convex optimal control with random coefficients. *SIAM J. Control and Optimization*, 33:590–624, 1995.

[32] H. Cartan. *Cours de calcul différentiel*. Hermann, Paris, 1997.

[33] E. Casas. Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems. *Adv. Comp. Math.*, 26:137–153, 2007.

[34] E. Casas, F. Tröltzsch, and A. Unger. Second order sufficient optimality conditions for a nonlinear elliptic boundary control problem. *Zeitschrift für Analysis und ihre Anwendungen*, 15:687–707, 1996.

[35] S. Chen, X. Li, and X.Y. Zhou. Stochastic linear quadratic regulators with indefinite control weight costs. *SIAM J. Control and Optimization*, 36, 1998.

[36] S. Chen and J. Yong. Stochastic linear quadratic optimal control problems. *Appl. Math. Optim.*, 43, 2001.

[37] M.G. Crandall, H. Ishii, and P.-L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. American Mathematical Society (New Series)*, 27:1–67, 1992.

[38] M. Davis. On the existence of optimal policies in stochastic control. *SIAM J. on Control and Optimization*, 11:587–594, 1973.

[39] M. Davis. *Linear Estimation and Stochastic Control*. Chapman and Hall, 1977.

[40] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state constrainted elliptic control problem. *SIAM J. Numer. Anal.*, 45:1937–1953, 2007.

[41] N. ElKaroui, D. Huu Nguyen, and M. Jeanblanc-Picqué. Compactification methods in the control of degenerate diffusions: existence of an optimal control. *Stochastics*, 20:169–219, 1987.

[42] L.C. Evans. *Partial differential equations*. Amer. Math Soc., Providence, RI, 1998. Graduate Studies in Mathematics 19.

[43] H.O. Fattorini. *Infinite dimensional optimization and control theory*. Cambridge University Press, New York, 1998.

[44] W.H. Fleming and M. Nisio. On the existence of optimal stochastic controls. *J. Math. Mech.*, 15:777–794, 1966.

[45] W.H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions*. Springer, New York, 1993.

[46] A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44:525–597 (electronic) (2003), 2002.

[47] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*. Springer Verlag, Berlin, 1983.

[48] Clovis C. Gonzaga. Path-following methods for linear programming. *SIAM Rev.*, 34(2):167–224, 1992.

[49] L.M. Graves. Some mapping theorems. *Duke Mathematical Journal*, 17:111–114, 1950.

[50] W.W. Hager. Multiplier methods for nonlinear optimal control. *SIAM J. on Numerical Analysis*, 27:1061–1080, 1990.

[51] W.W. Hager. Numerical analysis in optimal control. *Internat. Ser. Numer. Math. Birkhäuser, Basel*, 139:83–93, 2002.

[52] A. Haraux. How to differentiate the projection on a convex set in Hilbert space. some applications to variational inequalities. *J. Mathematical Society of Japan*, 29:615–631, 1977.

[53] U.G. Haussmann. General necessary conditions for optimal control of stochastic systems. *Math. Prog. Study*, 6:34–48, 1976.

[54] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. on Optimization*, 17:159–187, 2006.

[55] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Springer, New York, 2008.

[56] Y. Hu and X.Y. Zhou. Constrained stochastic LQ control with random coefficients, and application to mean–variance portfolio selection. *SIAM Journal on Control and Optimization*, 44, 2005.

[57] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer-Verlag, Berlin, 1994.

[58] T. Jockenhövel, L. T. Biegler, and A. Wächter. Dynamic optimization of the tennessee eastman process using the optcontrolcentre. *Computers and Chemical Engineering*, 27:1513–1531, 2003.

[59] M. Kohlmann and X.Y. Zhou. Relationship between backward stochastic differential equations and stochastic control. *SIAM J. Control Optim.*, 38, 200.

[60] H.J. Kushner. On the existence of optimal stochastic controls. *J. Math. Anal. Appl.*, 3:463–474, 1965.

[61] H.J. Kushner. On the stochastic maximum principle: Fixed time of control. *J. Math. Anal. Appl.*, 11:78–92, 1965.

[62] H.J. Kushner. On the stochastic maximum principle with "average" constraints. *SIAM J. on Control and Optimization*, 12:13–26, 1965.

[63] H.J. Kushner. Necessary conditions for continuous parameter stochastic optimization problems. *SIAM J. on Control and Optimization*, 10:550–565, 1972.

[64] 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 (electronic), 1999.

[65] X. Li and J. Yong. *Optimal control theory for infinite dimensional systems*. Birkhäuser, Boston, 1995.

[66] A.E.B. Lim and X.Y. Zhou. Mean-variance portfolio selection with random parameters in a complete market. *Math. Oper. Res.*, 29:101–120, 2002.

[67] J.-L. Lions. *Contrôle optimal de systèmes gouvernés par des equations aux dérivées partielles*. Dunod, Paris, 1968.

[68] H. Maurer. Optimization techniques for solving elliptic control problems with control and state constraint. part 2: Distributed control. *Comp. Optim. Applic.*, 18:141–160, 2001.

[69] H. Maurer and H. D. Mittelmann. Optimization techniques for solving elliptic control problems with control and state constraint. part 1: Boundary control. *Comp. Optim. Applic.*, 16:29–55, 2000.

[70] P.J. McLane. Optimal stochastic control of linear systems with state- and control-dependent disturbances. *IEEE Trans. Auto. Control*, 16:793–798, 1971.

[71] F. Mignot. Contrôle dans les inéquations variationnelles. *J. Functional Analysis*, 22:25–39, 1976.

[72] L. Mou and J. Yong. A variational formula for stochastic controls and some applications. *Pure and Applied Mathematics Quarterly*, 3:539–567, 2007.

[73] P. Neittaanmaki, J.Sprekels, and D. Tiba. *Optimization of elliptic systems*. Springer, New York, 2006.

[74] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[75] S. Peng. A general stochastic maximum principle for optimal control problems. *SIAM J. on Control and Optimization*, 28:966–979, 1990.

[76] H. Pham. *Optimisation et contrôle stochastique appliqués à la finance*, volume 61 of *Mathématiques & Applications*. Springer, Berlin, 2007.

[77] L. Pontryagin, V. Boltyanski, R. Gamkrelidze, and E. Michtchenko. *The Mathematical Theory of Optimal Processes*. Wiley Interscience, New York, 1962.

[78] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. on Optimization.*, 20(2):1002–1031, 2009.

[79] A. Schiela and M.Weiser. Superlinear convergence of the control reduced interior point method for PDE constrained optimization. *Comp.Opt.Appl.*, 39 (3):369–393, 2008.

[80] J. Sokolowski. Sensitivity analysis of control constrained optimal control problems for distributed parameter systems. *SIAM Journal of Control and Optimization*, 25:1542–1556, 1987.

[81] G. Stampacchia. Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. *Ann. Inst. Fourier (Grenoble)*, 15:189–258, 1965.

[82] T. Terlaky, editor. *Interior Point Methods of Mathematical Programming*. Kluwer Academic Publishers, Boston, 1996.

[83] M. Ulbrich and S. Ulbrich. Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds. *SIAM J. Control Optim.*, 38:1938–1984, 2000.

[84] M. Ulbrich and S. Ulbrich. Primal-Dual Interior point methods for PDE-constrained optimization. *Math. Program.*, 117:435–485, 2009.

[85] M. Weiser. Interior point methods in function space. *SIAM J. on Control and Optimization*, 44:1766–1786 (electronic), 2005.

[86] M. Weiser and P. Deuflhard. Inexact central path following algorithms for optimal control problems. *SIAM J. on Control and Optimization*, 46(3):792–815, 2007.

[87] M. Weiser, T. Gänzler, and A. Schiela. A control reduced primal interior point method for a class of control constrained optimal control problems. *Comp. Opt. Appl.*, 41 (1):127–145, 2008.

[88] M. Weiser and A. Schiela. Function space interior point methods for PDE constrained optimization. *PAMM*, 4 (1):43–46, 2004.

[89] W. M. Wonham. On a matrix Riccati equation of stochastic control. *SIAM J. Control*, 6:681–697, 1968.

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

[91] S.J. Wright. *Primal-dual interior-point methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[92] S.J. Wright and D. Orban. Properties of the log-barrier function on degenerate nonlinear programs. 2001.

[93] J. Yong and X.Y. Zhou. *Stochastic controls, Hamiltonian systems and HJB equations*. Springer-Verlag, New York, Berlin, 2000.

[94] F. Zhang. *Matrix Theory*. Springer, first edition, New York, 1999.

[95] X.Y. Zhou. The connection between the maximum principle and dynamic programming in stochastic control. *Stoch. & Stoch. Rep.*, 31:1–13, 1990.

[96] X.Y. Zhou and D. Li. Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Appl. Math. Optim.*, 42:19–33, 2000.