EFFICIENT DUALITY-BASED NUMERICAL METHODS FOR SPARSE PARABOLIC OPTIMAL CONTROL PROBLEMS

CHEN BO

(M.Sc., XMU, China; B.Sc., HIT, China)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
2018

Supervisor:
Professor Sun De Feng, Main Supervisor
Professor Bao Wei Zhu, Co-Supervisor

Examiners:
Professor Zhao Gong Yun
Professor Toh Kim Chuan
Professor Michael Ulbrich, Technische Universität München

To my parents

DECLARATION

I hereby declare that the thesis is my original work and it has
been written by me in its entirety. I have duly
acknowledged all the sources of information which
have been used in the thesis.

This thesis has also not been submitted for any degree
in any university previously.

Chen Bo

Chen Bo

June 7, 2018

Acknowledgements

I would like to express my deepest gratitude to my supervisor Professor Sun Defeng for his professional guidance during these past four and a half years. From the conic programming class, I have learned a lot of knowledge of convex optimization from him. I have also benefited intellectually and mentally from his insightful and patient guidance in both research and life. Moreover, I am very grateful for his financial support for my fifth years study.

My sincere thanks also goes to Professor Bao Weizhu. Professor Bao Weizhu acted as my co-supervisor and has helped me a lot on the part of partial differential equation. I have also learned a lot about numerical analysis and finite element method from him.

I am greatly indebted to Professor Toh Kim Chuan for his help on the numerical implementation for the algorithms. It is very considerate to use his research grant for extending my research assistant in NUS.

I also would like to convey my thanks to all the members in our optimization group. I have learned a lot from the weekly seminar. And I have benefitted a lot from the discussions with them and suggestions from them. In particular, I want to thank Dr Song Xiaoliang for many great suggestions on the research and my thesis

writing. It is a very pleasant experience to cooperate with him.

As always, I owe my deepest gratitude to my parents for their constant and unconditional love and support throughout my life. Last but not least, I am also deeply indebted to my wife, Cai Shujun. Without her understanding, tolerance, encouragement and love, I would have been nowhere.

Contents

Summary

In this thesis, we apply efficient algorithms for solving sparse parabolic optimal control problems (SPOCPs) and exploit convergence of those algorithms.

The objective function of SPOCPs contains both the L^1-norm and the indicator function of a box constraint. It is a nonsmooth optimization problem and hence semismooth Newton method is preferred. Semismooth Newton method, though enjoys the superlinear convergence rate, has several drawbacks for this type of problem.

The first drawback is that the approximate discretization of L^1-norm may introduce extra error. In order to avoid the extra error, we look into the dual problem, which is an unconstrained multi-block optimization problem. In this approach, the L^1-norm is transformed into the indicator function of a unit ball. We give an error estimate for the our new discretization model, that is $\|u_{h,\tau} - u^*\|_{L^2(\Omega_T)} = \mathcal{O}(h + \sqrt{\tau})$. To solve it efficiently, we apply the symmetric Gauss-Seidel (sGS) based inexact majorized accelerated block coordinate descent method (sGS-imABCD) for the new discretization model and the inexact majorized accelerated block coordinate descent method (imABCD) for the conventional discretization problem, both of which have a $\mathcal{O}(1/k^2)$ computation complexity of optimal value. Based on the convergence of optimal value, we exploit the convergence of the primal variable and the first order optimality conditions. Later we also prove the uniformly mesh-independence of the exact majorized ABCD method for solving the conventional discretization problem. This forms the first part of my thesis.

We know that to obtain the fast superliner convergence rate, one needs a good enough initial point for semismooth Newton method and also needs to solve the semismooth Newton equation up to a very high accuracy. Obviously, it is hard to judge whether the initial point is good enough or not. Moreover, when the L^2 regularization parameter α is very small or when α is zero, the situation is often very bad and the requirement of accuracy to solve the semismooth Newton method is very strict. From the convergence analysis of the majorized ABCD method, we know that the convergence rate is sensitive to the parameter α. Hence, in the second part, we aim to find another efficient method to deal with the case when α is very small or when α is zero. Still, we focus on the dual problem and introduce the augmented Lagrangian method, which has a fast linear convergence rate. And for the subproblem, we apply the semismooth Newton method, which will attain the stopping criteria very quickly. We call it semismooth Newton augmented Lagrangian (SSNAL) method. In the thesis, we illustrate the details of implementation for the SSNAL method and prove the convergence of both primal and dual variables. For the case $\alpha > 0$, we prove the uniformly mesh-independence of the method, and its robustness to the parameter α. For the case $\alpha = 0$, the SSNAL method still works very efficiently.

In the numerical experiments, we compare our methods with the inexact semi-proximal alternative direction method (isPADMM), and the globalized semismooth Newton method (SSN). Numerical results demonstrate that for α not very small, both the imABCD method and the SSNAL method are very efficient in solving all the problems given, while for small α, our SSNAL method outperforms other methods for all the examples given. And we could see the robustness of the SSNAL method to the parameter α. For the case α being zero, the SSNAL method is even much more efficient than the isPADMM method.

Chapter 1

Introduction

1.1 Sparse parabolic optimal control problems

In the thesis, we focus on designing efficient algorithms for solving sparse parabolic optimal control problems (SPOCPs) as following,

$$\min_{y\in Y,u\in U_{ad}} \quad J(y,u)=\frac{1}{2}\|y-y_d\|^2_{L^2(\Omega_T)}+\frac{\alpha}{2}\|u\|^2_{L^2(\Omega_T)}+\beta\|u\|_{L^1(\Omega_T)} \tag{P}$$

such that the following parabolic equation holds

$$\begin{cases} y_t-\Delta y=u+y_c \ \text{ in } \Omega_T:=\Omega\times(0,T), \\ y(x,0)=0, \forall x\in\Omega, \\ y(x,t)=0, \forall x\in\Gamma, t\in[0,T], \end{cases} \tag{1.1}$$

where

1. Ω is a convex, open and bounded domain with $C^{1,1}$- or polygonal boundary, $\Omega=(0,T)\times\Omega$.

2. $U=L^2(\Omega_T):=L^2((0,T)\times\Omega)$, $y_d\in L^2(\Omega_T)$ is the desired state,

$$\begin{aligned} Y&=W(0,T)\\ &=\{y\in L^2((0,T);H^1_0(\Omega_T))|y_t\in L^2((0,T),H^{-1}(\Omega)),y(x,0)|_{x\in\Omega}=0\}, hspace2cm \end{aligned} \tag{1.2}$$

3. $\alpha \geq 0, \beta > 0$, $U_{ad} = \{u \in U | a \leq u(x,t) \leq b, \text{a.e. } x \in \Omega, t \in [0,T]\}$, with $-\infty < a < 0 < b < \infty$ constant numbers.

The weak formulation of (1.1) is

$$\int_0^T \langle y_t(t), v(t)\rangle_{V^*,V} dt + \int_0^T a(y(t), v(t))dt = \int_0^T \langle u(t) + y_c(t), v(t)\rangle_{L^2(\Omega)} dt, \tag{1.3}$$

for all $v \in L^2(0,T; H_0^1(\Omega))$, with the initial condition

$$y(x,0) = 0, \forall x \in \Omega, \tag{1.4}$$

where $V = H_0^1(\Omega), V^* = H^{-1}(\Omega)$, $a(y(t), v(t)) := \int_\Omega \nabla y(t) \nabla v(t) dx.$

The details can be referred to [45, Definition 1.28].

We choose the dual pairings such that we obtain the Gelfand triples,

$$H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega), \tag{1.5}$$

with continuous and dense embeddings.

Now we can write the weak formulation (1.3) in the equivalent form

$$\mathcal{A}y = \mathcal{B}(u - y_c), \tag{1.6}$$

with

1. $\mathcal{A} \in \mathcal{L}(Y, L^2((0,T); V^*))$ is defined by,

$$\langle \mathcal{A}y, v\rangle_{L^2((0,T);V^*),L^2((0,T);V)} = \int_0^T (\langle y_t(t), v(t)\rangle_{V^*,V} + \langle \nabla y(t), \nabla v(t)\rangle_{L^2(\Omega)})dt, \tag{1.7}$$

for all $v \in L^2((0,T); V)$. Here

$$\begin{aligned} L^2((0,T); V) := \{y : [0,T] \to X \text{ strongly measurable} : \\ \|y\|_{L^2((0,T),V)} := (\int_0^T \|y(t)\|_V^2 dt)^{1/2} < \infty\} \end{aligned} \tag{1.8}$$

2. $\mathcal{B} \in \mathcal{L}(L^2(\Omega_T))$ is defined by

$$\langle \mathcal{B}u, v\rangle_{L^2(\Omega_T)} = \int_0^T \langle u(t), v(t)\rangle_{L^2(\Omega)} dt, \forall v \in L^2((0,T); V). \tag{1.9}$$

More detail can be referred to [45, Section 1.3].

Remark 1.1. *Although we assume that the Dirichlet boundary condition $y = 0$ holds, it should be noticed that the assumption is not a restriction and our considerations can also carry over to the more general boundary conditions of Robin type*

$$\frac{\partial y}{\partial v} + \gamma y = g \ \textit{on}\ \partial\Omega \times (0, T), \tag{1.10}$$

where $g \in L^2(\partial\Omega \times [0, T])$ is given and $\gamma \in L^\infty(\partial\Omega \times [0, T])$ is nonnegative coefficient.

Remark 1.2. *For opitimization problems with nonlinear PDE constraint, we may apply SQP approach on the continuous level to make each subproblem be an minimization problem with linear constraint. Then it becomes the type of problem in our setting, and we can apply our methods mentioned in the thesis to solve the SQP subproblems.*

The objective function of problem (P) contains both the L^1-norm and the box constraint. Hence it is nonsmooth. Here we put L^1-norm in the objective to promote the sparsity of the control variable, and it has important applications, such as actuator placement problems [25, 35]. In optimal control of distributed parameter systems, it may be impossible or undesirable to put the controllers at every point of the domain. Instead, we can decide to control the system by localizing the controls in small regions.

By introducing two artificial variables v, z, we rewrite the primal problem as

$$\begin{cases} \displaystyle\min_{y\in Y, u, v, z\in U} & J(y, u, v, z) = \dfrac{1}{2}\|y - y_d\|^2_{L^2(\Omega_T)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Omega_T)} + \beta\|v\|_{L^1(\Omega_T)} + \delta_{U_{ad}}(z) \\ \quad\text{s.t.} & \mathcal{A}y - \mathcal{B}(u + y_c) = 0, \\ & \mathcal{B}(u - v) = 0, \\ & \mathcal{B}(u - z) = 0. \end{cases} \tag{1.11}$$

Let p, λ, μ be the Lagrangian multipliers for the three equalities respectively and let us define the Lagrangian functional as below

$$\begin{aligned}\mathcal{L}(y,u,v,z;p,\lambda,\mu) =& J(y,u,v,z) + \langle p, \mathcal{A}y - \mathcal{B}(u+y_c)\rangle + \langle \lambda, \mathcal{B}(u-v)\rangle \\ &+ \langle \mu, \mathcal{B}(u-z)\rangle\end{aligned} \tag{1.12}$$

We can then obtain the Lagrangian dual problem of (1.11) by firstly minimize with respective to the the primal variables $(y,u,v,z)$and then maximize the Lagrangian multipliers (or called dual variables), and it is provided as below.

$$\max_{p,\lambda,\mu} \inf_{y,u,v,z} \mathcal{L}(y,u,v,z;p,\lambda,\mu) \tag{1.13}$$

We can simply it and obtain the equivalent minimization formulation as

$$\begin{aligned}\min_{\mu,\lambda,p} \Phi(\mu,\lambda,p) :=& \frac{1}{2}\|\mathcal{A}^*p - y_d\|^2_{L^2(\Omega_T)} + \frac{1}{2\alpha}\|\lambda+\mu-p\|^2_{L^2(\Omega_T)} + \langle \mathcal{B}y_c, p\rangle_{L^2(\Omega_T)} \\ &+ \delta_{[-\beta,\beta]}(\lambda) + \delta^*_{[a,b]}(\mu) - \frac{1}{2}\|y_d\|^2_{L^2(\Omega_T)},\end{aligned} \tag{D}$$

where

1. $\mu, \lambda \in L^2(\Omega_T), p \in L^2((0,T);H_0^1(\Omega))$.
2. $\mathcal{A}^* \in \mathcal{L}(L^2((0,T);V), Y^*)$ is the adjoint of $\mathcal{A}$.
3. $\delta^*_{U_{ad}}(\mu) := \sup\limits_{x\in U_{ad}} \langle x, \mu\rangle_{L^2(\Omega_T)}$ is the conjugate function of the indicator function $\delta_{U_{ad}}(\cdot)$ with respective to the inner product induced by L^2-norm.
4. $\delta_{[-\beta,\beta]}(\cdot)$ is in fact the conjugate function of $\beta\|\cdot\|_{L^1(\Omega_T)}$ with respective to the inner product induced by L^2-norm.

In other way, we would regard the two nonsmooth functions, $\beta\|\cdot\|_{L^2(\Omega_T)}$ and $\delta_{[a,b]}(\cdot)$, as one function q. Then similarly, we introduce one additional variable w and let λ be the Lagrangian multiplier for the equality $\mathcal{B}(u-w) = 0$, and obtain another dual problem as

$$\begin{aligned}\min_{\lambda,p} \widetilde{\Phi}(\lambda,p) :=& \frac{1}{2}\|\mathcal{A}^*p - y_d\|^2_{L^2(\Omega_T)} + \frac{1}{2\alpha}\|\lambda-p\|^2_{L^2(\Omega_T)} + \langle \mathcal{B}y_c, p\rangle_{L^2(\Omega_T)} \\ &+ q^*(\lambda) - \frac{1}{2}\|y_d\|^2_{L^2(\Omega_T)},\end{aligned} \tag{$\widetilde{\text{D}}$}$$

with $\lambda \in L^2(\Omega_T), p \in L^2((0,T);H_0^1(\Omega))$.

1.2 Literature review

The optimal control problem with control constraint, namely Problem (P) with $\alpha > 0, \beta = 0$, has been widely studied for decades, see e.g. [18, 33, 36, 44, 59, 70]. Most of the papers focus on the following three aspects, the discretization of the continuous optimal control problem, the error estimate of the discretization and the optimization methods to solve the discretized problem.

To tackle the problem (P) numerically, we have two different approaches. The first one is *first optimize then discretize*. The idea is to get the first order optimality condition of the continuous problem, then to discretize the linear equations, and solve them numerically. The second approach is *first discretize then optimize*. That is to discretize the variables and spaces to obtain a finite dimensional optimization problem, and then utilize optimization algorithms to solve the discretized problem. There are differing opinions regarding which route to take (see Collis and Heinkenschloss [24] for a discussion). And for the discretization, finite element method is often preferred because that it is often easy to obtain an error estimate as well as we can approximate the solution with good enough basis functions, such as piece-wise linear functions. To get a better error estimate, Hinze proposed a new discretization method [44], the variational discretization, which requires more computation time but often enjoys a higher order of error estimation.

Numerous results of error estimate for different optimal control problems are proposed in recent papers, see e.g. [1, 10, 15–17, 33, 36, 60, 81]. For distributed control problems, Falk [33] and Geveci [36] presented the finite element analysis and obtained the L^2 error estimate, $\mathcal{O}(h)$, for piecewise constant approximations of control variables, while Meyer and Rösch [60] proved the same order of error estimate for piecewise linear approximation. Casas, Mateos and Tröltzsch [16] presented numerical analysis for Neumann boundary control of semilinear elliptic equations and proved the error estimate, $\|u - u_h\|_{L^2(\Omega)} = \mathcal{O}(h)$, for piecewise constant control approximation. Casas and Mateos [15] further exploited the error estimate for piecewise finite element approximation, $\|u - u_h\|_{L^2(\Omega)} = o(h)$, and that for variational discretization, $\|u - u_h\|_{L^2(\Omega)} = \mathcal{O}(h^{3/2})$. For parabolic optimal control problems, Meidner and Vexler [57,58] considered the discrete approximation based on $dG(0)$ in time and finite element in space, they proved the estimate $\|u - u_{\tau,h}\|_{L^2(\Omega_T)} = \mathcal{O}(\tau + h)$ for piece-wise linear approximation and $\|u - u_{\tau,h}\|_{L^2(\Omega_T)} = \mathcal{O}(\tau + h^{3/2})$ for variational discretization. For further references we refer to the papers, [29,30,39,43,56,59,69].

Optimization method is also the key for solving optimal control problems efficiently. For the control constrained optimal control problem, semismooth Newton is often preferred due to its locally superlinear convergence rate, see e.g. [42,46,79,80]. Primal dual active set, which was later proved to be a special case of semismooth Newton method [80], is also very popular as the choice for solving optimal problems, see e.g. [5, 49]. Many other state-of-the-art algorithms have been applied to solve the optimal control problem, such as SQP [37,38,40,61,86], ADMM [73], FISTA [71] and ABCD [72].

Motivated by the optimal placement of actuators on piezoelectric plates [25,35], the sparsity property of the control variables is considered. And sparse optimal control problems with L^1-cost function has been intensively studied recently. Georg Stadler first proposed it in [74] to obtain the sparse optimal solution for the optimal control problem with elliptic equation constraints. In the paper, the author studied the optimality conditions for the problem and proposed the semismooth Newton method for solving the sparse optimal control problem when the regularization parameter $\alpha > 0$.

The first error estimate for sparse optimal control problem with L^1-norm, as far as we know, was provided by Wachsmuth and Wachsmuth [82] in 2011. In their paper, they investigated two types of approximations for their problem. Firstly, they studied the convergence of solutions if the regularization parameter α tends to zero. Secondly, they studied finite element approximations for the regularized problem. They proved that $\|u_\alpha - u_{\alpha,h}\|_{L^2(\Omega)} \leq C(h/\alpha + h^2/\alpha^{3/2})$ for piecewise constant discretization of the control variables and $\|u_\alpha - u_{\alpha,h}\|_{L^2(\Omega)} \leq Ch^2/\alpha$ for variational discretization. Moreover, to have a decoupled form of objective function when using the piecewise linear discretization, they introduced the approximation of the L^1-norm. Under this approach, they were able to obtain the above error estimate.

In [11,12], Casas, Herzog, and Wachsmuth analyzed the nonconvex case governed by semilinear elliptic equations. They made use of the technique of approximation of L^1-norm, and then derived the error estimates for the nonconvex control problem under three different discretization.

Apart from using L^1-norm to induce sparsity, Clason and Kunisch in [22, 23] investigated elliptic control problems with measure-valued controls to promote the sparsity of the control. In their model, the L^2-norm regularization is no longer necessary, that is, $\alpha = 0$. And they studied the dual problem of the sparse optimal control problem instead of the primal problem, which is quite impressive. There are also many sparse optimal control using measure-valued controls, e.g. [8,9,13,14,48, 50]. Furthermore, for parabolic equation constrained control problem, directional sparsity was introduced by Herzog et al. in [41] to promote striped sparsity patterns.

To numerical solve Problem (P), many methods have been applied. Since we have a nonsmooth term in the objective function, semismooth Newton method is often the first choice. Its fast superlinear convergence rate and its capability to obtain high accuracy of numerical solution make it a very efficient algorithm, see [45, 74, 82]. However, there are two drawbacks to use semismooth Newton for Problem (P). Firstly, before using the semismooth Newton method, we have to adopt the approximate discretization of L^1-norm to discretize the objective function of the primal problem. Hence, additional discretization error is inevitable. Secondly, to obtain the fast superlinear convergence rate, one needs a good initial point and needs to solve the semismooth Newton equation up to a very high accuracy. However, when the L^2-norm regularization parameter, i.e., α is very small or equals to zero, this requirement is very strict and often hard to satisfy. Combing our error estimate result of $\mathcal{O}(h + \sqrt{\tau})$, we see that it is no need to solve for high accuracy solution. Hence, it is necessary to consider other efficient methods than semismooth Newton method for Problem (P).

Besides semismooth Newton method, there are also many efficient first-order methods for choices, such as the recently popular ADMM method [21,27,34,54,75], FISTA [4], APG method [47, 78], ABCD method [26, 76]. The ADMM method was first applied to solve sparse optimal control problem by Song, Yu, Wang and Zhang [73]. To apply the ADMM method, they also made use of the approximate L^1-norm to obtain a decouple objective function. Moreover, they use different choices of the penalty term for different variable when using the ADMM. Hence it is called inexact heterogeneous ADMM (ihADMM) in the paper. Numerical results showed that the ihADMM method outperforms semismooth Newton method to obtain a moderate accuracy. Later FISTA method was also utilized to solve sparse optimal control problem by Schindele in [71], where it was called Fast Inexact Proximal method (FIP). In theory, we see that FISTA method has a computation complexity of $\mathcal{O}(1/k^2)$, which is faster than ADMM. However, in theory, the convergence of sequences for FISTA is not guaranteed. And in numerical implementation, we need to spend extra time to compute the Lipschitz constant at each iteration, which is quite expensive.

Later in 2017, Song and Chen [72] applied ABCD method to solve sparse optimal elliptic control problem. The ABCD method has the same $\mathcal{O}(1/k^2)$ complexity of optimal value, but different with FIP method, our method does not need to compute the Lipschitz constant, and we can avoid the discretization error caused by approximation of L^1-norm.

1.3 Contributions

In this thesis, I aim at filling the gaps mentioned above.

In the first part, I focus on how to avoid the error caused by the approximate discretization of L^1-norm, and I extend the result of [72] to SPOCPs. Since Problem (P) is a convex optimization problem, I am inspired to look into its dual problem, for which the L^1-norm is changed to be a indicator function of a unit ball. Hence instead of discretizing the primal problem, I discretize the dual problem. And when exploiting the predual of the discretized dual problem, I find that actually I discretize the L^1-norm in a new way. The new approximation is proved to be a better one than the conventional approximation of L^1-norm. And for the new discretization of SPOCPs, I also provide the error estimate, which is $\mathcal{O}(h + \sqrt{\tau})$. As the discretized dual problem turns out to be an unconstrained three-block optimization problem, motivated by [26], I find that the state-of-the-art method, symmetric Gauss-Seidel based inexact majorized accelerated block coordinate descent method (sGS-imABCD) can be a suitable algorithm to solve Problem (P). I illustrate the implementation of sGS-imABCD. In the theory, I prove the convergence of the primal variable and the first order optimality conditions. If SPOCPs are discretized with the general approximation of L^1-norm, I apply the imABCD method to solve it. The difference is that I regard the two nonsmooth terms as one, and obtain an unconstrained two-block dual problem. Convergence of both the primal and dual variables are provided, and under proper assumptions, I also prove that the imABCD method has the uniformly mesh-independence property, which means the convergence rate of all the discretized problems is independent of the mesh-size when the mesh-size is fine enough.

In the second part, I consider the case when the L^2 regularization parameter α is small or is zero. Both semismooth Newton method and ABCD method are sensitive to the parameter α, hence I need to consider another method for this situation. Impressed by the linear convergence rate of Semismooth Newton Augmented Lagrangian (SSNAL) method, and motivated by recently published paper [53], we come up with the idea to transfer the unconstrained optimization problem into a linear equation constrained optimization problem and apply the SSNAL method. When augmented Lagrangian method is applied to solve the dual problem of the model problem, the dimension of SSNAL subproblem can be reduced due to the sparsity structure of problem (P). Besides, the SSNAL method enjoys a fast linear convergence rate for the solution sequence obtained from the outer iteration. And for the subproblem, I apply the semismooth Newton method to solve it efficiently to achieve the stopping criteria. I illustrate the details of implementation of the algorithm and present the convergence results. Furthermore, given the proper choice of initial penalty parameter σ, I prove the uniformly mesh-independence of the method. And the SSNAL method also shows the robustness to the regularization parameter α. For the case $\alpha = 0$, the SSNAL method also works very efficiently.

1.4 Thesis organization

The rest of the thesis is organized as follows. In Chapter 2, I provide the preliminaries that will be used in the subsequent chapters. In Chapter 3, I illustrate the finite element discretization. Existence and uniqueness of Problem (P) are shown in sequence. Later I give an error estimate for our discretization of the SPOCPs. In Chapter 4, I proposed the sGS-imABCD method for the SPOCPs and the imABCD method for the decoupled SPOCPs. Convergence results are provided in the subsections. Furthermore, I prove the uniformly mesh-independence for solving the decoupled SPOCPs. In Chapter 5, devoted to solving the case α is very small or is zero, I propose the SSNAL method, which applies semismooth Newton method for its subproblem. Convergence results are established in sequence. For the case $\alpha > 0$, I prove the uniformly mesh-independence and show the robustness to the regularization parameter α. For the case $\alpha = 0$, I also illustrate the implementation and the convergence results. In Chapter 6, I provide some comparison methods, and some examples. Numerical results reconfirm the theory provided and show that imABCD method works well for α not very small, and the SSNAL method, for any choice of α, outperforms other method provided. And we conclude in Chapter 7.

Chapter 2

Preliminaries

2.1 Notations

- Let n be a given integer. We use $\mathcal{S}^n$ to denote the space of all $n \times n$ symmetric matrices, $\mathcal{S}^n_+$ to denote the space of all $n \times n$ positive semidefinite matrices,

- Denote $\mathcal{X}$ as a finite dimensional Euclidean space endowed with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$, and $\mathcal{M} : \mathcal{X} \to \mathcal{X}$ as a self-adjoint positive semidefinite linear operator. We write $\mathcal{M}^{\frac{1}{2}}$ as a self-adjoint positive semidefinite linear operator such that $\mathcal{M}^{\frac{1}{2}}\mathcal{M}^{\frac{1}{2}} = \mathcal{M}$, which always exists. For any $x, y \in \mathcal{X}$, we define $\langle x, y \rangle_{\mathcal{M}} := \langle x, \mathcal{M}y \rangle$ and $\|x\|_{\mathcal{M}} := \sqrt{\langle x, \mathcal{M}x \rangle}$. Furthermore, we denote the induced norm on the space $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ as follow:

$$\|A\| = \sup\{\|Ax\| : x \in \mathcal{X} \text{ with } \|x\| = 1\}.$$

- Given a closed convex set $\mathcal{C} \subseteq \mathcal{X}$ and a point $x \in \mathcal{C}$, denote $\mathcal{T}_{\mathcal{C}}(x)$ as the tangent cone of C at x and $\mathcal{N}_{\mathcal{C}}(x)$ as the normal cone of C at x. We define $\text{dist}(x, \mathcal{C}) := \inf_{y \in \mathcal{C}} \|x - y\|$.

- Given a closed convex cone $\mathcal{K} \subseteq \mathcal{X}$, denote $\mathcal{K}^*$ as the dual cone of $\mathcal{K}$ and $\mathcal{K}^\circ$ as the polar cone of $\mathcal{K}$.

- Given a convex function $f : \mathcal{X} \to (-\infty, +\infty]$, we use dom$f$ to denote the effective domain of f, and epi$f$ to denote the epigraph of f. We also use the notation f^* to denote the Fenchel's conjugate function of f, and Prox_f as the proximal mapping of f. Furthermore, we say f is a $L\mathcal{C}^1$ function if it is continuously differentiable and its gradient is Lipschitz continuous, and we say f is $\mathcal{C}^2$ if it is twice continuously differentiable.

- Given a set of matrices $X := (X_1, X_2, \ldots X_s) \in \mathcal{R}^{n_1 \times m_1} \times \mathcal{R}^{n_2 \times m_2} \times \ldots \times \mathcal{R}^{n_s \times m_s}$ for some positive integers s, $n_1, n_2, \ldots, n_s$ and $m_1, m_2, \ldots, m_s$, we denote Diag(X) as a block diagonal matrix whose ith main block diagonal is given by X_i for $i \in \{1, 2, \ldots, s\}$.

2.2 An inexact block symmetric Gauss-Seidel iteration

In this section, we first introduce the symmetric Gauss-Seidel (sGS) technique proposed recently by Li, Sun and Toh [51]. It is a powerful tool to solve a convex minimization problem whose objective is the sum of a multi-block quadratic function and a non-smooth function involving only the first block, which plays a vital role in our subsequent algorithm designing.

Let $s \geq 2$ be a given integer and $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_s$ where $\mathcal{X}_i$ are real finite dimensional Euclidean spaces. The sGS technique aims to solve the following unconstrained nonsmooth convex optimization problem approximately

$$\min \phi(x_1) + \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle r, x\rangle, \tag{2.1}$$

where $x := (x_1, ..., x_s) \in \mathcal{X}$ with $x_i \in \mathcal{X}_i$, $i = 1, ..., s$, $\phi : \mathcal{X}_1 \to (-\infty, +\infty]$ is a closed proper convex function, $\mathcal{Q} : \mathcal{X} \to \mathcal{X}$ is a given self-adjoint positive semidefinite linear operator and $r := (r_1, ..., r_s) \in \mathcal{X}$ is a given vector.

For notational convenience, we denote the quadratic function in (2.1) as

$$h(x) := \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle r, x\rangle, \tag{2.2}$$

and the block decomposition of the operator $\mathcal{Q}$ as

$$\mathcal{Q}x := \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1s} \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} & \cdots & \mathcal{Q}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{Q}_{1s}^* & \mathcal{Q}_{2s}^* & \cdots & \mathcal{Q}_{ss} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_s \end{pmatrix}, \tag{2.3}$$

where $\mathcal{Q}_{ii} : \mathcal{X}_i \to \mathcal{X}_i, i = 1, ..., s$ are self-adjoint positive semidefinite linear operators, $\mathcal{Q}_{ij} : \mathcal{X}_j \to \mathcal{X}_i, i = 1, ..., s-1, j > i$ are linear maps whose adjoints are given by $\mathcal{Q}_{ij}^*$. Here, we assume that $\mathcal{Q}_{ii} \succ 0, \forall i = 1, ..., s$. Then, we consider a splitting of $\mathcal{Q}$

$$\mathcal{Q} = \mathcal{D} + \mathcal{U} + \mathcal{U}^*, \tag{2.4}$$

where

$$\mathcal{U} := \begin{pmatrix} 0 & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1s} \\ & \ddots & \cdots & \mathcal{Q}_{2s} \\ & & \ddots & \mathcal{Q}_{(s-1)s} \\ & & & 0 \end{pmatrix} \tag{2.5}$$

denotes the strict upper triangular part of $\mathcal{Q}$ and $\mathcal{D} := \mathrm{Diag}(\mathcal{Q}_{11}, ..., \mathcal{Q}_{ss}) \succ 0$ is the diagonal of $\mathcal{Q}$. For later discussions, we also define the following self-adjoint positive semidefinite linear operator

$$\mathrm{sGS}(\mathcal{Q}) := \mathcal{T} = \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^*. \tag{2.6}$$

For any $x \in \mathcal{X}$, we define

$$x_{\leq i} := (x_1, x_2, ..., x_i), \quad x_{\geq i} := (x_i, x_{i+1}, ..., x_s), \quad i = 0, 1, ..., s+1,$$

with the convention $x_{\leq 0} = x_{\geq 0} = \emptyset$. Moreover, in order to solve the problems inexactly, we introduce the following two error tolerance vectors:

$$\delta' :\equiv (\delta_1', ..., \delta_s'), \quad \delta :\equiv (\delta_1, ..., \delta_s),$$

with $\delta_1' = \delta_1$. Define

$$\Delta(\delta', \delta) = \delta + \mathcal{U}\mathcal{D}^{-1}(\delta - \delta').$$

Given $\bar{x} \in \mathcal{X}$, we consider solving the following problem

$$x^+ := \arg\min_x \left\{ \phi(x_1) + h(x) + \frac{1}{2}\|x - \bar{x}\|_{\mathcal{T}}^2 - \langle x, \Delta(\delta', \delta)\rangle \right\}. \tag{2.7}$$

Then, the following sGS decomposition theorem, which is established by Li, Sun and Toh in [52], shows that computing x^+ in (2.7) is equivalent to computing in an inexact block symmetric Gauss-Seidel type sequential updating of the variables $x_1, ..., x_s$.

Theorem 2.1. [52, Theorem 2.1]. *Assume that the self-adjoint linear operators $\mathcal{Q}_{ii}$ are positive definite for all $i = 1, ..., s$. Then, it holds that*

$$\mathcal{Q} + \mathcal{T} = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*) \succ 0. \tag{2.8}$$

Furthermore, given $\bar{x} \in \mathcal{X}$, for $i = s, ..., 2$, suppose we have computed $x_i' \in \mathcal{X}_i$ defined as follow,

$$\begin{aligned} x_i' :&= \arg\min_{x_i \in \mathcal{X}_i} \phi(\bar{x}_1) + h(\bar{x}_{\le i-1}, x_i, x'_{\ge i+1}) - \langle \delta_i', x_i\rangle, \\ &= \mathcal{Q}_{ii}^{-1}\left(r_i + \delta_i' - \sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* \bar{x}_j - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j' \right), \end{aligned} \tag{2.9}$$

then the optimal solution x^+ defined by (2.7) can be obtained exactly via

$$\begin{cases} x_1^+ = \arg\min_{x_1 \in \mathcal{X}_1} \phi(x_1) + h(x_1, x'_{\ge 2}) - \langle \delta_1, x_1\rangle, \\ x_i^+ = \arg\min_{x_i \in \mathcal{X}_i} \phi(x_1^+) + h(x^+_{\le i-1}, x_i, x'_{\ge i+1}) - \langle \delta_i, x_i\rangle, \\ \quad = \mathcal{Q}_{ii}^{-1}\left(r_i + \delta_i - \sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* x_j^+ - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j' \right), \quad i = 2, ..., s. \end{cases} \tag{2.10}$$

Remark 2.1. (a). *In (2.9)and (2.10), x_i' and x_i^+ should be regarded as inexact solutions to the corresponding minimization problems without the linear error terms $\langle \delta_i', x_i\rangle$ and $\langle \delta_i, x_i\rangle$. Once these approximate solutions have been computed, they* *would generate the error vectors* δ_i' *and* δ_i *as follows:*

$$\delta_i' = \mathcal{Q}_{ii}x_i' - \left(r_i - \sum_{j=1}^{i-1} \mathcal{H}_{ji}^* \bar{x}_j - \sum_{j=i+1}^{s} \mathcal{Q}_{ij}x_j' \right), \quad i = s, ..., 2,$$

$$\delta_1 \in \partial\phi(x_1^+) + \mathcal{Q}_{11}x_1^+ - \left(r_1 - \sum_{j=2}^{s} \mathcal{H}_{1j}x_j' \right),$$

$$\delta_i = \mathcal{Q}_{ii}x_i^+ - \left(r_i - \sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* x_j^+ - \sum_{j=i+1}^{s} \mathcal{Q}_{ij}x_j' \right), \quad i = 2, ..., s.$$

With the above known error vectors, we have that x_i' *and* x_i^+ *are the exact solutions to the minimization problems in (2.9) and (2.10).*

(b). *In actual implementations, assuming that for* $i = s, ..., 2$, *we have computed* x_i' *in the backward GS sweep for solving (2.9), then when solving the subproblems in the forward GS sweep in (2.10) for* $i = 2, ..., s$, *we may try to estimate* x_i^+ *by using* x_i', *and in this case the corresponding error vector* δ_i *would be given by*

$$\delta_i = \delta_i' + \sum_{j=1}^{i-1} \mathcal{Q}_{ji}^*(x_j' - \bar{x}_j).$$

In practice, we may accept such an approximate solution $x_i^+ = x_i'$ *for* $i = 2, ..., s$, *if the corresponding error vector satisfies an admissible condition such as* $\|\delta_i\| \leq c\|\delta_i'\|$ *for some constant* $c > 1$, *say* $c = 10$.

For the latter purpose, we present the following proposition, which can be found in [52].

Proposition 2.1. [**52, Proposition 2.1**]. *Suppose that* $\widehat{\mathcal{Q}} = \mathcal{Q} + \mathcal{T}$ *is positive definite. Let* $\xi = \|\widehat{\mathcal{Q}}^{-1/2}\Delta(\delta', \delta)\|$. *It holds that*

$$\xi \leq \|\mathcal{D}^{-1/2}(\delta - \delta')\| + \|\widehat{\mathcal{Q}}^{-1/2}\delta'\|. \tag{2.11}$$

2.3 Accelerated block coordinate descent method

Let us consider a general class of unconstrained, multi-block convex optimization problems with coupled objective function, that is

$$\min_{u,v} \theta(u, v) := p_1(u) + p_2(v) + \phi(u, v), \tag{2.12}$$

where $p_1 : \mathcal{U} \to (-\infty, +\infty]$ and $p_2 : \mathcal{V} \to (-\infty, +\infty]$ are two convex functions (possibly nonsmooth), $\phi : \mathcal{U} \times \mathcal{V} \to (-\infty, +\infty]$ is a smooth convex function with Lipschitz continuous gradient mapping, and $\mathcal{U}$, $\mathcal{V}$ are real finite dimensional Hilbert spaces.

To solve (2.12), in 2015, Chambolle and Pock [19] proposed the accelerated alternative descent (AAD) algorithm for this situation that the joint objective function is quadratic. However, their method does not take the inexactness of the solutions of associated subproblems into account. Hence, it is not suitable for the practical application. Later, Sun, Toh and Yang [76] proposed an inexact accelerated block coordinate descent (iABCD) method when the subproblem $\arg\min_u p_1(u) + \phi(u, v)$ has an explicit solution. They applied the Danskin-type theorem to reduce the two-block problem into one block and then utlized the accelerated proximal gradient method to solve the reduced one-block problem. With the help of the symmetric Gauss-Seidel technique [51], the model problem (2.12) can be extended to a multi-block unconstrained optimization problem with only two block nonsmooth terms. They proved that the iABCD method has a $\mathcal{O}(1/k^2)$ iteration complexity of the optimal value. For a more general case when the subproblem mentioned above can not be solved exactly, Cui [26, Chapter 3] in 2016 proposed the inexact majorized accelerated block coordinate descent (imABCD) method. Under suitable assumption on ϕ and some inexact criteria, the $\mathcal{O}(1/k^2)$ complexity of optimal value can also be obtained.

Next, let us give a brief sketch of the inexact majorized ABCD method. To deal with the general model (2.12), we need some more conditions and assumptions on ϕ. Let us denote $\omega := (u, v) \in \mathcal{U} \times \mathcal{V}$. Assume $\nabla\phi$ be globally Lipschitz continuous, and hence there exist two self-adjoint positive semidefinite linear operators $\mathcal{Q}$ and $\hat{Q} : \mathcal{U} \times \mathcal{V} \to \mathcal{U} \times \mathcal{V}$ such that for any $\omega, \omega' \in \mathcal{U} \times \mathcal{V}$, it holdes

$$\phi(\omega) \geq \phi(\omega') + \langle \nabla\phi(\omega'), \omega - \omega' \rangle + \frac{1}{2}\|\omega' - \omega\|^2_{\mathcal{Q}},$$

and

$$\phi(\omega) \leq \hat{\phi}(\omega; \omega') := \phi(\omega') + \langle \nabla\phi(\omega'), \omega - \omega' \rangle + \frac{1}{2}\|\omega' - \omega\|^2_{\hat{\mathcal{Q}}}.$$

We further decompose the operators $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ into the following block structures:

$$\mathcal{Q}\omega := \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}, \quad \widehat{\mathcal{Q}}\omega := \begin{pmatrix} \widehat{\mathcal{Q}}_{11} & \widehat{\mathcal{Q}}_{12} \\ \widehat{\mathcal{Q}}_{12}^* & \widehat{\mathcal{Q}}_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}, \quad \forall \omega = (u, v) \in \mathcal{U} \times \mathcal{V},$$

and assume $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ satisfy the following conditions.

Assumption 2.1. [**26**, **Assumption 3.1**] *There exist two self-adjoint positive semidefinite linear operators $\mathcal{D}_1 : \mathcal{U} \to \mathcal{U}$ and $\mathcal{D}_2 : \mathcal{V} \to \mathcal{V}$ such that*

$$\widehat{\mathcal{Q}} := \mathcal{Q} + \mathrm{Diag}(\mathcal{D}_1, \mathcal{D}_2).$$

Furthermore, $\widehat{\mathcal{Q}}$ satisfies that $\widehat{\mathcal{Q}}_{11} \succ 0$ and $\widehat{\mathcal{Q}}_{22} \succ 0$.

Now we present the inexact majorized acclerated block coordinate descent algorithm as follow.

Algorithm 1: (Inexact majorized ABCD algorithm for (2.12))

Input: $(u^1, v^1) = (\tilde{u}^0, \tilde{v}^0) \in \mathrm{dom}(p) \times \mathrm{dom}(q)$. Let $\{\epsilon_k\}$ be a summable sequence of nonnegative numbers, and set $t_1 = 1$, $k = 1$.

Output: $(\tilde{u}^k, \tilde{v}^k)$

Iterate until convergence:

Step 1 Choose error tolerance $\delta_u^k \in \mathcal{U}, \delta_v^k \in \mathcal{V}$ such that

$$\max\{\delta_u^k, \delta_v^k\} \leq \epsilon_k / \max\{\|\hat{\mathcal{Q}}_{11}^{-1}\|, \|\hat{\mathcal{Q}}_{22}^{-1}\|\}.$$

Compute

$$\begin{cases} \tilde{u}^k = \arg\min\limits_{u \in \mathcal{U}} \{p_1(u) + \hat{\phi}(u, v^k; \omega^k) - \langle \delta_u^k, u \rangle\}, \\ \tilde{v}^k = \arg\min\limits_{v \in \mathcal{V}} \{p_2(v) + \hat{\phi}(\tilde{u}^k, v; \omega^k) - \langle \delta_v^k, v \rangle\}. \end{cases}$$

Step 2 Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$ $\forall$ k. Compute

$$u^{k+1} = \tilde{u}^k + \beta_k(\tilde{u}^k - \tilde{u}^{k-1}), v^{k+1} = \tilde{v}^k + \beta_k(\tilde{v}^k - \tilde{v}^{k-1}).$$

Here we state the convergence result without proving. For the detailed proof, one can see [26, Chapter 3].

Theorem 2.2. [26, **Theorem 3.2**] *Suppose that Assumption 2.1 holds and the solution set* Ω *of the Problem (2.12) is non-empty. Let* $\omega^* \in \Omega$*. Assume that* $\sum\limits_{i=1}^{\infty} i\epsilon_i < \infty$*. Then the sequence* $\{\tilde{\omega}^k\} := \{(\tilde{u}^k, \tilde{v}^k)\}$ *generated by the imABCD algorithm satisfies that*

$$\theta(\tilde{\omega}^k) - \theta(\omega^*) \leq \frac{2\|\tilde{\omega}^0 - \omega^*\|_{\mathcal{H}}^2 + c_0}{(k+1)^2}, \quad \forall k \geq 1,$$

where c_0 *is a constant number, and* $\mathcal{H} := \mathrm{Diag}(\mathcal{D}_1, \mathcal{D}_2 + \mathcal{Q}_{22})$*.*

2.4 Augmented Lagrangian method

In this section, we introduce the augmented Lagrangian method and its convergence theory.

Firstly, let us introduce the proximal point algorithm for solving a fundamental problem of determining an element z such that $0 \in T(z)$.

The proximal point algorithm is presented as

$$z^{k+1} \approx P_k(z^k), \text{ where } P_k = (I + c_k T)^{-1} \tag{PPA}$$

and for any starting point z^0, it generates a convergent sequence $\{z^k\}$ in H.

To obtain its inexact form, we need two general criteria for the approximate calculation of $P_k(z^k)$ as follow,

$$\|z^{k+1} - P_k(z^k)\| \leq \epsilon_k, \qquad \sum_{k=1}^{\infty} \epsilon_k < \infty, \tag{2.13}$$

$$\|z^{k+1} - P_k(z^k)\| \leq \delta_k \|z^{k+1} - z^k\|, \qquad \sum_{k=1}^{\infty} \delta_k < \infty. \tag{2.14}$$

For more details, one can refer to [55, 67].

Definition 2.1. *Let* $\mathbb{X}, \mathbb{Y}$ *be real Hilbert space with inner product* $\langle \cdot, \cdot \rangle$*. A multifunction* $T : \mathbb{X} \to \mathbb{Y}$ *is said to be metrically subregular at* $\bar{z} \in \mathbb{X}$ *for* $y \in \mathbb{Y}$ *with modulus* $\kappa > 0$ *if* $(\bar{x}, \bar{y}) \in gph(T)$ *and there exists neighborhoods* $\mathcal{U}$ *of* $\bar{x}$ *and* $\mathcal{V}$ *of* $\bar{y}$ *such that*

$$dist(x, T^{-1}(\bar{y})) \leq \kappa dist(\bar{y}, T(x) \cap \mathcal{V}), \forall x \in \mathcal{U}, \tag{2.15}$$

or equivalently, T is said to metrically subregular at $\bar{z}$ for $\bar{y}$ with modulus $\kappa > 0$ if there exists a nieghborhood $\mathcal{U}'$ of $\bar{z}$ such that

$$dist(x, T^{-1}(\bar{y})) \le \kappa dist(\bar{y}, T(x)), \forall x \in \mathcal{U}'. \tag{2.16}$$

Proposition 2.2. [28, Proposition 2.1] *Let $\mathcal{H}$ be a real Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle$ and $\theta : \mathcal{H} \to (-\infty, +\infty]$ be a proper lower semicontinuous convex function. Let $\bar{v}, \bar{x} \in \mathcal{H}$ satisfy $(\bar{x}, \bar{v}) \in gph(\partial\theta)$. Then $\partial\theta$ is metrically subregular at $\bar{x}$ for $\bar{v}$ if and only if there exists a neighborhood $\mathcal{U}$ of $\bar{x}$ and a constant $\kappa > 0$ such that*

$$\theta(x) \ge \theta(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle + \kappa dist^2(x, (\partial\theta)^{-1}(\bar{v})), \forall x \in \mathcal{U}. \tag{2.17}$$

Remark 2.2. *From* [2], *if the second order condition is valid with parameter $\kappa > 0$, one can also estimate the metric subregularity parameter to be at least $1/\kappa$.*

Hence we can see that if θ is strongly convex with modulus $a > 0$, then for any $(\bar{x}, 0) \in gph(\partial\theta)$, we have the second order growth condition

$$\theta(x) \ge \theta(\bar{x}) + \frac{a}{2}\|x - \bar{x}\|^2, \forall x \in \mathcal{H}. \tag{2.18}$$

This equation above is equivalent to say that equation (2.17) in Proposition 2.2 is satisfied with $\bar{x} = \bar{x}, \bar{v} = 0$ and $\kappa = a/2$. Then by the remark above or by [2], one has that $\partial\theta$ is metrically subregular at $\bar{x}$ for origin with the parameter at least $2/a$.

Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ be a lower semicontinuous convex function. For $i = 1, 2, \cdots, m$, let $f_i : \mathbb{R}^n \to \mathbb{R}$ be an affine function. We consider the convex programming problem

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} \quad f_0(x) \\ \text{s.t.} \quad f_j(x) = 0, j = 1, \cdots, m. \end{cases} \tag{2.19}$$

For Problem (2.19), the Lagrangian function and augmented Lagrangian function are defined as follow respectively,

$$\begin{aligned} l(x, \omega) &= f_0(x) + \sum_{i=1}^{m} \omega_i f_i(x), \forall x \in \mathbb{R}^n, \omega \in \mathbb{R}^m \\ l_\sigma(x, \omega) &= f_0(x) + \sum_{i=1}^{m} \omega_i f_i(x) + \frac{\sigma}{2} \sum_{i=1}^{m} |f_i(x)|^2, \forall x \in \mathbb{R}^n, \omega \in \mathbb{R}^m, \sigma > 0. \end{aligned}$$

Then we can propose the augmented Lagrangian method for the problem (2.19)

$$\begin{cases} x^{k+1} \approx \arg\min_{x\in\mathbb{R}^n} l_{\sigma_k}(x,\omega^k), \\ \omega_i^{k+1} = \omega_i^k + \sigma_k f_i(x^{k+1}), \text{ for } i = 1,2,\cdots,m. \end{cases} \tag{ALM}$$

The inexact forms are executed with the following stopping criteria

$$\phi_k(x^{k+1}) - \inf_x \phi_k(x) \leq \epsilon_k^2/2\sigma_k, \epsilon_k \geq 0, \sum_{k=0}^{\infty}\epsilon_k < \infty, \tag{2.20}$$

$$\phi_k(x^{k+1}) - \inf_x \phi_k(x) \leq (\delta_k^2/2\sigma_k)\|\omega^{k+1}-\omega^k\|^2, \delta_k \geq 0, \sum_{k=0}^{\infty}\delta_k < \infty, \tag{2.21}$$

$$dist(0, \partial\phi_k(x^{k+1})) \leq (\delta_k'/\sigma_k)\|\omega^{k+1}-\omega^k\|^2, 0 \leq \delta_k \to 0. \tag{2.22}$$

where $\phi_k(x) := l_{\sigma_k}(x,\omega^k)$.

The ordinary dual problem associated with (2.19) is

$$\text{maximize } g(\omega) \text{ over all } \omega \in \mathbb{R}^m, \tag{2.23}$$

where $g(\omega) := \inf_{x\in\mathbb{R}^n} l(x,\omega)$.

The following result shows the relation between augmented Lagrangian method and the general proximal point algorithm (PPA) in the case of $T = T_g$, or in other words,

$$P_k(\omega) := (I + \sigma_k T_g)^{-1}(\omega) = \arg\max_{\omega\in\mathbb{R}^m}\{g(\omega) - (1/2\sigma_k)|\omega - \omega^k|^2\}. \tag{2.24}$$

Proposition 2.3. [66, Proposition 6] *For P_k as defined as in (2.24), and sequence $\{(x^k,\omega^k)\}$ generated from (ALM) for (2.19) , we have*

$$|\omega^{k+1} - P_k(\omega^k)|^2/2\sigma_k \leq L_{\sigma_k}(x^k;\omega^k) - \inf_x L_{\sigma_k}(x;\omega^k). \tag{2.25}$$

Remark 2.3. *From the above proposition, we can see that $\{(x^k,\omega^k)\}$ can also be regarded as generated from the inexact PPA for Problem (2.24). Or in other word, inexact ALM method can be regarded as a special case of inexact PPA method.*

Hence we can analysis the convergence rate of inexact ALM through the convergence result for inexact PPA.

We define some useful multifunctions as follow:

$$T_f = \partial f, T_g = -\partial g, T_l(x,\omega) = \{(u,v) | (u,-v) \in \partial l(x,\omega)\},$$

and

$$\begin{cases} T_f^{-1}(v) := \arg\min\limits_{x\in\mathbb{R}^n}\{f(x) - x\cdot v\}, \\ T_g^{-1}(\mu) := \arg\max\limits_{\omega\in\mathbb{R}^n}\{g(\omega) + \omega\cdot\mu\}, \\ T_l^{-1}(v,\mu) := \arg\min\limits_{x\in\mathbb{R}^n}\max\limits_{\omega\in\mathbb{R}^m}\{l(x,\omega) - x\cdot v + \omega\cdot\mu\}, \end{cases} \tag{2.26}$$

where $l(x,\omega) = f_0(x) + \sum\limits_{i=1}^{m}\omega_i f_i(x), f(x) = \sup\limits_{\omega\in\mathbb{R}^m} l(x,\omega), g(\omega) = \inf\limits_{x\in\mathbb{R}^n} l(x,\omega)$.

Theorem 2.3. [28, Theorem 4.2] *Suppose optimal solution set $T_f^{-1}(0)$ to Problem (2.19) is nonempty. Let $\{(x^k,\omega^k)\}$ be an infinite sequence generated by the Algorithm (ALM) with stopping criteria (2.20). Then the whole sequence $\{x^k\}$ is bounded and converges to some $\bar{x} \in T_f^{-1}(0)$, and the sequence $\{\omega^k\}$ satisfies for all $k \geq 0$,*

(a) *If T_g is metrically subregular at $\bar{\omega}$ for the origin with modulus κ_g, then under criterion (2.21), there exists $k \geq 0$ such that for all $k \geq \bar{k}, \eta_k < 1$ and*

$$dist(\omega^{k+1}, T_g^{-1}(0)) \leq \theta_k dist(\omega^k, T_g^{-1}(0)), \tag{2.27}$$

where

$$1 > \theta_k = \left(\kappa_g/\sqrt{\kappa_g^2 + \sigma_k^2 + 2\delta_k}\right)(1-\delta_k)^{-1} \to \theta_\infty = \kappa_g/\sqrt{\kappa_g^2 + \sigma_\infty^2}$$
$$(\theta_\infty = 0,\ \textit{if } \sigma_\infty = \infty).$$

(b) *If in addition to stopping criteria (2.21) and the metric subregularity of T_g at $\bar{\omega}$ for the origin, one has stopping criteria (2.22), $T_f^{-1}(0)$ is non-empty and bounded and the following condition on T_l: there exist two constants $\kappa_l \geq 0$ and $\epsilon > 0$, any (x,ω) satisfying $dist((x,\omega), T_f^{-1} \times \{\bar{\omega}\}) \leq \epsilon$,*

$$dist((x,\omega), T_l^{-1}(0)) \leq \kappa_l dist(0, T_l(x,\omega)). \tag{2.28}$$

Then there exists $\tilde{k} > 0$ such that for all $k \geq \tilde{k}, \delta_k < 1$, and

$$dist(x^{k+1}, T_f^{-1}(0)) \leq \theta'_k dist(\omega^k, T_g^{-1}(0)), \tag{2.29}$$

where $\theta'_k = \kappa_l\sigma_k^{-1}(1+\delta'_k)(1-\delta'_k) \to \theta'_\infty = \kappa_l/\sigma_\infty(\theta'_\infty = 0,\ \textit{if } \sigma_\infty = \infty)$.

Chapter 3

Discretization

To solve the continuous optimization problem (P), we need to discretize it first. There are two approaches. One is *first optimize then discretize*, another is *first discretize then optimize* [24]. In the thesis, I choose the latter approach. And for the discretization, we consider the finite element discretization for the space and backward Euler for the time.

In this chapter, I provide a new discretization for both the primal problem (P) and its dual (D). For numerical comparison, I also provide the conventional discretization in which I apply an approximation of L^1-norm in the objective function of the primal problem (P). Then I study the uniqueness and existence of the optimal solution and also exploit the first order optimality conditions. Finally, I propose the error estimate of the new discretization.

3.1 Finite element discretization

Firstly, let us give some assumptions.

We consider a family of regular and quasi-uniform triangulations $\{\mathcal{T}_h\}_{h>0}$ of $\bar{\Omega}$. For each cell $R \in \mathcal{T}_h$, let us define the diameter of the set R by $\rho_R := diamR$ and define σ_R to be the diameter of the largest ball contained in R. The mesh size of the grid is defined by $h = \max_{R\in\mathcal{T}_h} \rho_R$. We suppose that the following regularity assumptions on the triangulation are satisfied which is standard in the context of error estimates:

Assumption 3.1 (Regular and quasi-uniform triangulations). *There exist two positive constants κ and δ such that*

$$\frac{\rho_T}{\sigma_T} \le \kappa, \quad \frac{h}{\rho_T} \le \delta,$$

hold for all $R \in \mathcal{T}_h$ and all $h > 0$. Moreover, let us define $\bar{\Omega}_h = \bigcup_{R\in\mathcal{T}_h} R$, and let $\Omega_h \subset \Omega$ and Γ_h be its interior and its boundary, respectively. We assumed that $\bar{\Omega}_h$ is convex and that all boundary vertices of $\bar{\Omega}_h$ are contained in Γ_h, such that

$$|\Omega \setminus \Omega_h| \le ch^2, \tag{3.1}$$

where $|\cdot|$ denotes the measure of the set and $c > 0$ is a constant.

On account of the homogeneous boundary condition of the state equation, we define

$$\mathcal{W}_h = \{w_h \in C(\bar{\Omega}) | w_h|_T \in \mathcal{P}_1, \text{ for all } T \in \mathcal{T}_h, \text{ and } w_h = 0 \text{ in } \bar{\Omega} \setminus \Omega_h\}, \tag{3.2}$$

where $\mathcal{P}_1$ denotes the space of polynomials of degree less than or equal to 1.

Let τ be the uniform time grid, $N_T := \frac{T}{\tau}$, and $0 = t_0 < t_1 < \cdots < t_{N_T} = T$. Denote $\Omega_{h,T} = \Omega_h \times [0, T]$, and set $I_j := (t_{j-1}, t_j]$, then we seek

$$Y_{h,\tau} = \{y_{h,\tau} \in C(\bar{\Omega}_T) | y_{h,\tau}(\cdot, t)|_{\bar{\Omega}} \in \mathcal{W}_h \text{ and } y_{h,\tau}(x, \cdot)|_{I_j} \in \mathcal{P}_0, j = 1, \cdots, N_T\} \tag{3.3}$$

as the discrete space, where $\mathcal{P}_0$ denotes the space of piecewise constant polynomials. As mentioned above, we also use the same discrete space to discretize the control u, and define

$$U_{h,\tau} = \{u_{h,\tau} \in C(\bar{\Omega}_T) | u_{h,\tau}(\cdot, t)|_{\bar{\Omega}} \in \mathcal{W}_h \text{ and } u_{h,\tau}(x, \cdot)|_{I_j} \in \mathcal{P}_0, j = 1, \cdots, N_T\}. \tag{3.4}$$

For a given regular and quasi-uniform triangulation $\mathcal{T}_h$ with nodes $\{x_i\}_{i=1}^{N_h}$, let $\{\phi_i\}_{i=1}^{N_h}$ be a set of nodal basis functions, which span $Y_{h,\tau}$ as well as $U_{h,\tau}$ and satisfy the following properties:

$$\phi_i \geq 0, \|\phi_i\|_\infty = 1, i = 1, 2, \cdots, N_h, \sum_{i=1}^{N_h} \phi_i = 1.$$

The elements $u_{h,\tau} \in U_{h,\tau}$ and $y_{h,\tau}, y_{d,h} \in Y_{h,\tau}$ can be represented in the following forms, respectively, $\forall t \in [0, T]$,

$$u_{h,\tau}(\cdot, t) = \sum_{i=1}^{N_h} u_i(t)\phi_i, y_{h,\tau}(\cdot, t) = \sum_{i=1}^{N_h} y_i(t)\phi_i, \tag{3.5}$$

and

$$y_{d,h}(\cdot, t) = \sum_{i=1}^{N_h} y_d^i(t)\phi_i, y_{c,h}(\cdot, t) := \sum_{i=1}^{N_h} y_c^i(t)\phi_i. \tag{3.6}$$

For $i = 1, 2, \cdots, N_T, j = 1, 2, \cdots, N_h$, let us denote that

$$\begin{aligned} u_{i,j} &= u_j(t_i), u_i = (u_{i,1}, u_{i,2}, \cdots, u_{i,N_h})^T \in \mathbb{R}^{N_h}, \\ y_{i,j} &= y_j(t_i), y_i = (y_{i,1}, y_{i,2}, \cdots, y_{i,N_h})^T \in \mathbb{R}^{N_h}. \end{aligned} \tag{3.7}$$

Let $U_{ad,h,\tau}$ denote the discrete feasible set, which is defined by

$$U_{ad,h,\tau} := U_{h,\tau} \cap U_{ad} \subset U_{ad}. \tag{3.8}$$

where $U_{ad} := \{u \in U | a \leq u(x,t) \leq b, \text{ a.e. } x \in \Omega, t \in [0,T]\}$, with $-\infty \leq a < 0 < b \leq +\infty$.

And let $y_{c,h}$, $y_{d,h}$ be the L^2-projection of y_c and y_d onto $Y_{h,\tau}$, respectively,

$$y_{c,h,\tau} = \sum_{i=1}^{N_h} y_c^i \phi_i, \quad y_{d,h,\tau} = \sum_{i=1}^{N_h} y_d^i \phi_i. \tag{3.9}$$

Then we can obtain the weak form of the problem (P),

$$\left\{ \begin{aligned} \min_{y_{h,\tau} \in Y_{h,\tau}, u_{h,\tau} \in U_{h,\tau}} \quad J(y_{h,\tau}, u_{h,\tau}) =& \frac{1}{2}\|y_{h,\tau} - y_{d,h,\tau}\|^2_{L^2(\Omega_T)} + \frac{\alpha}{2}\|u_{h,\tau}\|^2_{L^2(\Omega_T)} \\ & + \beta\|u_{h,\tau}\|_{L^1(\Omega_T)} + \delta_{U_{ad,h,\tau}}(u_{h,\tau}) \\ \text{s.t.} \qquad \langle \mathcal{A}_{h,\tau} y_{h,\tau}, \phi_i \rangle_{U_{h,\tau}} =& \langle \mathcal{B}_{h,\tau}(u_{h,\tau} + y_{c,h,\tau}), \phi_i \rangle_{U_{h,\tau}}, \forall v_{h,\tau} \in U_{h,\tau}. \end{aligned} \right. \tag{3.10}$$

From the perspective of numerical implementation, we introduce the stiffness matrix and the mass matrix as follow:

$$K_h = \left(\int_{\Omega_h} \langle \nabla\phi_i(x), \nabla\phi_j(x)\rangle dx\right)_{i,j=1}^{N_h}, M_h = \left(\int_{\Omega_h} \phi_i(x)\phi_j(x) dx\right)_{i,j=1}^{N_h}. \tag{3.11}$$

We use backward Euler for the time and rewrite the problem (3.10) in the following way,

$$\left\{\begin{aligned} \min_{y,u\in\mathbb{R}^m} \quad J(y,u) =& \frac{T}{N_T}\sum_{i=1}^{N_T}\{\frac{1}{2}\|y_i - y_{d,i}\|_{M_h}^2 + \frac{\alpha}{2}\|u_i\|_{M_h}^2 + \beta\|\sum_{j=1}^{N_h} u_{i,j}\phi_j\|_{L^1(\Omega)} \\ & + \delta_{[a,b]}(u_i)\} \\ \text{s.t.} \qquad F_1 y_{i+1} =& F_2 y_i + M(u_{i+1} + y_{c,i+1}), i = 0, 1, \cdots, N_T - 1. \end{aligned}\right. \tag{3.12}$$

where $m = N_h \times N_T, F_1 = \frac{N_T}{T}M_h + K_h, F_2 = \frac{N_T}{T}M_h$,

$$\left\{\begin{aligned} y &= (y_1^T, y_2^T, \cdots, y_{N_T}^T)^T \in \mathbb{R}^m, \\ u &= (u_1^T, u_2^T, \cdots, u_{N_T}^T)^T \in \mathbb{R}^m, \\ y_d &= (y_{d,1}^T, y_{d,2}^T, \cdots, y_{d,N_T}^T)^T \in \mathbb{R}^m, \\ y_c &= (y_{c,1}^T, y_{c,2}^T, \cdots, y_{c,N_T}^T)^T \in \mathbb{R}^m. \end{aligned}\right. \tag{3.13}$$

Define $B = \text{Diag}(M_h, M_h, \cdots, M_h) \in \mathbb{R}^{m\times m}$, and

$$A = \begin{pmatrix} F_1 & 0 & 0 & \cdots & 0 \\ -F_2 & F_1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & -F_2 & F_1 & 0 \\ 0 & \cdots & 0 & -F_2 & F_1 \end{pmatrix} \tag{3.14}$$

From the invertibility of the mass matrix M_h, and stiffness matrix K_h, we can see that F_1 is invertible and thus obtain that A is also invertible. And from the symmetric positive definiteness of M_h, we obtain that B is also symmetric positive definite.

Applying the notation A, B, we can derive the simplified form of problem (3.12),

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} \quad J(y,u) = & \frac{1}{2}\|y-y_d\|_B^2 + \frac{\alpha}{2}\|u\|_B^2 + \beta\sum\limits_{i=1}^{N_T}\|\sum\limits_{j=1}^{N_h} u_{i,j}\phi_j\|_{L^1(\Omega)} + \delta_{[a,b]}(u) \\ \text{s.t.} \qquad Ay = & B(u+y_c). \end{cases} \tag{P$_{\mathrm{h},\tau}$}$$

It is clear that the discrete L^1-norm $\|\sum\limits_{j=1}^{N_h} u_{i,j}\phi_i\|_{L^1(\Omega)}$ is a coupled form with respect to $(u_{i,j})$, which can not be written as a matrix-vector form. And the subgradient with respect to the function $u_{h,\tau}$, i.e. $\nu_{h,\tau} \in \partial_{u_{h,\tau}}\|u_{h,\tau}\|_{L^1(\Omega_T)}$, will not belong to a finite-dimensional subspace. Hence, if directly solving (P$_{\mathrm{h},\tau}$), it is inevitable to bring some difficulties for the numerical calculation. To overcome these difficulties, in [82], the authors introduced a lumped mass matrix W_h which is a diagonal matrix defined as

$$W_h := \mathrm{Diag}\left(\int_{\Omega_h}\phi_i(x)\mathrm{d}x\right)_{i=1}^{N_h}. \tag{3.15}$$

They defined an alternative discretization of the L^1-norm:

$$\|x_{h,\tau}\|_{L_h^1(\Omega_{h,T})} := \sum_{j=1}^{N_T}\sum_{i=1}^{N_h}|x_{j,i}|\int_{\Omega_h}\phi_i(x)\mathrm{d}x = \sum_{j=1}^{N_T}\|W_h x_j\|_1, \text{ for all } x_{h,\tau}\in U_{h,\tau}, \tag{3.16}$$

which is a weighted l^1-norm for the coefficients of $x_{h,\tau}$. More importantly, the following results about the mass matrix M_h and the lumped mass matrix W_h hold.

Proposition 3.1. [**83, Table 1**] $\forall\ z \in \mathbb{R}^{N_h}$, *the following inequalities hold:*

$$\|z\|_{M_h}^2 \le \|z\|_{W_h}^2 \le \gamma\|z\|_{M_h}^2, \ where\ \gamma = \begin{cases} 4 & if \quad n=2, \\ 5 & if \quad n=3. \end{cases} \tag{3.17}$$

$$\int_{\Omega_h}|\sum_{i=1}^{n} z_i\phi_i(x)|\ \mathrm{d}x \le \|W_h z\|_1. \tag{3.18}$$

Thus we obtain another discretization of Problem (3.10) as follow

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u)=\frac{T}{N_T}\sum\limits_{j=1}^{N_T}\{\frac{1}{2}\|y_j-y_{d,j}\|_M^2+\frac{\alpha}{2}\|u_j\|_M^2+\beta\|W_hu_j\|_1+\delta_{[a,b]}(u_j)\} \\ \text{s.t.} & F_1y_{j+1}=F_2y_j+M(u_{j+1}+y_{c,j+1}), j=0,1,\cdots,N_T-1. \\ & \quad y_0=0. \end{cases} \tag{3.19}$$

Furthermore, we define $q(u):=\beta\|Cu\|_1+\delta_{[a,b]}(u)$ for convenience, and then simplify Problem (3.19) as follow,

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u)=\frac{1}{2}\|y-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+q(u) \\ \text{s.t.} & Ay=B(u+y_c). \end{cases} \tag{$\tilde{\mathrm{P}}_{\mathrm{h},\tau}$}$$

where $C=\mathrm{Diag}(W_h,W_h,\cdots,W_h)\in\mathbb{R}^{m\times m}$.

This is the conventional way of discretization for the primal problem (P), where the l^1-norm is approximated by a decoupled function.

Noticed that the conjugate function of the l^1 norm is a unit ball indicator function, we consider to discretize the dual problem instead of the primal problem. We present the discretized dual problem (D) directly as follow,

$$\begin{aligned} \min_{p,\lambda,\mu\in\mathbb{R}^m} \quad \Phi(p,\lambda,\mu)=&\frac{1}{2}\|A^*p-By_d\|_{B^{-1}}^2+\frac{1}{2\alpha}\|p-\lambda-\mu\|_B^2+\langle By_c,p\rangle \\ &+\delta_{[-\beta,\beta]}(\lambda)+\delta_{[a,b]}^*(B\mu)-\frac{1}{2}\|y_d\|_B^2. \end{aligned} \tag{$\widehat{\mathrm{D}}_{\mathrm{h},\tau}$}$$

To mention a bit more, its predual is

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u)=\frac{1}{2}\|y-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+\beta\|Bu\|_1+\delta_{[a,b]}(u) \\ \text{s.t.} & Ay=B(u+y_c). \end{cases} \tag{$\widehat{\mathrm{P}}_{\mathrm{h},\tau}$}$$

or

$$\begin{cases} \min\limits_{y_{h,\tau}\in Y_{h,\tau},u_{h,\tau}\in U_{h,\tau}} & J(y_{h,\tau},u_{h,\tau})=\frac{1}{2}\|y_{h,\tau}-y_{d,h,\tau}\|_{L^2(\Omega_T)}^2+\frac{\alpha}{2}\|u_{h,\tau}\|_{L^2(\Omega_T)}^2 \\ & \qquad +\beta\|u_{h,\tau}\|_{\tilde{L}_h^1(\Omega_T)}+\delta_{U_{ad,h,\tau}}(u_{h,\tau}) \\ \text{s.t.} & \mathcal{A}_{h,\tau}y_{h,\tau}=\mathcal{B}_{h,\tau}(u_{h,\tau}+y_{c,h,\tau}). \end{cases} \tag{$\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$}$$

where

$$\|u\|_{\tilde{L}_h^1(\Omega_T)} := \tau \sum_{j=1}^{N_T} \sum_{i=1}^{N_h} | \int_\Omega u_{h,\tau}(x,t_j)\phi_i(x)dx|, \forall u_{h,\tau} \in U_{h,\tau}. \tag{3.20}$$

It is important to know that Problem ($\widehat{\mathrm{P}}_{\mathrm{h},\tau}$) is different from the Problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$) defined above in the l_1-norm term of the objective functional, the former is $\|Bu\|_1$ is while the latter is $\|Cu\|_1$.

In the conventional approximate discretization, the l^1-norm term is decoupled with respective to u, which is more convenient to study and choose suitable algorithm, like ADMM method. In the new discretiztaion, though the l^1-norm term is not decoupled, we will prove it in Section 3.3 that this new approximation for l^1-norm turns out to be a better one, and later in Chapter 6, the numerical experiments show the new discretization has slightly small error than the new discretization.

In this section, we will only present details of discretization for Problem ($\widehat{\mathrm{P}}_{\mathrm{h},\tau}$) and Problem ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$). Similar results can be obtained for Problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$) and its dual problem.

3.2 Existence and uniqueness as well as optimality conditions

Before the discretization, we would like to exploit the existence and uniqueness of optimal solutions to Problem (P) and the corresponding discretized problems ($\widehat{\mathrm{P}}_{\mathrm{h},\tau}$).

Firstly, we provide the existence and uniqueness of the weak solution defined in the equation (1.3).

Theorem 3.1. [32, Theorem 5]. *For every $u \in L^2(\Omega_T)$ and $y_c \in L^2(\Omega_T)$, the parabolic equation has a unique weak solution $y \in H_0^1(\Omega)$. Furthermore,*

$$\|y\|_{L^2(H_0^1(\Omega),[0,T])} \leq C_p(\|u\|_{L^2(\Omega_T)} + \|y_c\|_{L^2(\Omega_T)}) \tag{3.21}$$

for a constant C_p depending only on Ω and T.

We present the existence of a minimizer for Problem (P) in the following proposition.

Proposition 3.2. *Problem* (P) *has a unique solution* $u^* \in L^2(\Omega_T)$.

Proof. We consider a feasible and minimizing sequence $\{(y^n, u^n)\}_{n\in\mathbb{N}} \subset L^2(\Omega_T)$ of (P), with $u^0 = 0$, and y^0 be the weak solution for the state equation (1.3) with u^0.

Then $(y, u) = (y^0, 0)$ is feasible to problem (P). Since $\{(y^n, u^n)\}_{n\in\mathbb{N}}$ is a minimizing sequence, we have that the objective function $J(y, u)$ should have a upper bound $\frac{1}{2}\|y^0 - y_d\|^2_{L^2(\Omega_T)}$ for $\{(y^n, u^n)\}_{n>k}$ with k large enough. Hence we can obtain that $\{u^n\}_{n>k}$ should be bounded by a constant $\frac{1}{\sqrt{\alpha}}\|y^0 - y_d\|_{L^2(\Omega_T)}$.

Since $L^2(\Omega_T)$ is a Hilbert space, we can extract a subsequence, $\{u^{n_i}\}_{i\in\mathbb{N}}$ of the bounded sequence $\{u^n\}_{n\in\mathbb{N}}$, such that $\{u^{n_i}\}_{n\in\mathbb{N}}$ converges to a $u^* \in L^2(\Omega_T)$ in weak topology.

Let us define $y^*, \{y^n\}_{n\in\mathbb{N}}$ to be the corresponding weak solution for state equation (1.3) with respect to $u^*, \{u^n\}_{n\in\mathbb{N}}$. Then by the weak lower semicontinuity of the objective function J, we deduce

$$\liminf_{i\to\infty} J(y^{n_i}, u^{n_i}) \geq J(y^*, u^*).$$

Hence (y^*, u^*) is a minimizer of (P). And the uniqueness of the optimal solution is due to the strong convexity of the objective function. □

In the next proposition, we provide the optimality condition. It can be derived easily from results in [45, Lemma 1.12, Theorem 1.51].

Proposition 3.3 (First-order optimality condition)**.** *The pair function* (y^*, u^*) *is the optimal solution of* (P)*, if and only if there exists adjoint variables* (p^*, λ^*)*, such that the following conditions hold*

$$\begin{cases} \mathcal{A}y^* = \mathcal{B}(u^* + y_c), \\ \mathcal{A}^*p^* = y_d - y^*, \\ \left.\begin{aligned} \alpha u^* &= \mathcal{B}^*(p^* - \lambda^*), \\ u^* &= \Pi_{U_{ad}}(soft((u^* + \mathcal{B}^*\lambda^*), \beta)). \end{aligned}\right\} \Leftrightarrow (*) \end{cases} \tag{3.22}$$

where

$$\Pi_{U_{ad}}(v(x)) = \max\{a, \min\{v(x), b\}\},$$
$$soft(v(x), \beta) = sgn(v(x)) \cdot \max(|v(x)| - \beta, 0).$$
$$(*) : \langle -p^* + \alpha u^*, u - u^* \rangle + \beta(\|u\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)}) \geq 0, \forall u \in U_{ad}.$$

Remark 3.1. *For the continuous problem, we can also obtain the relationship between optimal control and optimal adjoint function as*

$$u^* = \Pi_{U_{ad}}(\frac{1}{\alpha} soft(p^*, \beta)). \tag{3.23}$$

And this property is useful for the estimation of upper bound of $\|u^*\|_{L^2(H^1(\Omega),[0,T])}$ and for the construction of numerical examples.

Hence, from the first order optimality (3.22) and the uniqueness of optimal solution (y^*, u^*), we can conclude the existence and uniqueness of the solution for our problem ($\widetilde{\mathrm{D}}$).

Proposition 3.4. *Problem ($\widetilde{\mathrm{D}}$) has a unique optimal solution (p^*, λ^*).*

Similarly, we can prove the existence and uniqueness of the discretized problem ($\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$) and the first order optimality conditions. Here we provide the propositions without proof as below.

Proposition 3.5. *Problem ($\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$) has a unique solution $u^*_{h,\tau} \in L^2(\Omega_T)$.*

Proposition 3.6 (Discrete first-order optimality condition)**.** *Let the pair function $(y^*_{h,\tau}, u^*_{h,\tau})$ be the optimal solution of ($\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$), if and only if there exists adjoint variables (p^*, λ^*), such that the following conditions hold in the weak sense*

$$\begin{cases} \mathcal{A}_{h,\tau} y^*_{h,\tau} = \mathcal{B}_{h,\tau}(u^*_{h,\tau} + y_c), \\ \mathcal{A}^*_{h,\tau} p^*_{h,\tau} = y_d - y^*_{h,\tau}, \\ \left.\begin{aligned} \alpha u^*_{h,\tau} &= \mathcal{B}^*_{h,\tau}(p^*_{h,\tau} - \lambda^*_{h,\tau}), \\ u^*_{h,\tau} &= \Pi_{U_{ad,h,\tau}}(soft((u^*_{h,\tau} + \mathcal{B}^*_{h,\tau}\lambda^*_{h,\tau}), \beta)). \end{aligned}\right\} \Leftrightarrow (**) \end{cases} \tag{3.24}$$

where

$$(**) : \langle -p^*_{h,\tau} + \alpha u^*_{h,\tau}, u - u^*_{h,\tau} \rangle + \beta(\|u\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)}) \geq 0, \forall u \in U_{ad,h,\tau}.$$

3.3 Error estimate

In this section, we give an error estimate for our new discretization problem $(\widehat{\mathrm{P}}_{\mathrm{h},\tau})$.

For the convenience, we introduce several interpolation operators as follow:

$$\begin{cases} (\Pi_h u)(x,t) := \sum_{i=1}^{N_h} \pi_i(u)(t)\phi_i(x), \pi_i(u)(t) := \dfrac{\int_\Omega u(x,t)\phi_i(x)dx}{\int_\Omega \phi_i(x)dx}, \\ (I_h u)(x,t) := \sum_{i=1}^{N_h} u(x_i,t)\phi_i(x), \\ (I_\tau u)(x,t) = (\Pi_\tau u)(x,t) := u(x,t_j), \forall t \in (t_{j-1},t_j], \\ \Pi_{h,\tau} = \Pi_\tau \circ \Pi_h, I_{h,\tau} = I_\tau \circ I_h. \end{cases} \tag{3.25}$$

where $\circ$ is the function composition operator, $\{x_i\}_{i=1}^{N_h} \subset \Omega$ are the nodal points corresponding to the nodal basis $\{\phi_i\}_i^{N_h}$, such that

$$\phi_i(x_j) = \delta_{i,j}, \forall i,j = 1,\cdots,N_h. \tag{3.26}$$

We called $\Pi_{h,\tau}, \Pi_h$ the quasi-interpolation operator, $I_{h,\tau}, I_h$ the nodal interpolation operator.

We have $I_h \circ I_h = I_h$, but both of Π_h, I_h are not orthogonal projections.

To prove the error estimate, we provide some propositions as follow.

Proposition 3.7. *For all $v \in L^2((0,T);H^1(\Omega))$, such that v is Lipschitz continuous with respect to t, there exist constant C_I^1, C_I^2, independent of h, τ, such that*

$$\begin{aligned} \|v - I_h v\|_{L^2(\Omega_T)} &\le C_I^1 h |v|_{L^2((0,T);H^1(\Omega))}, \\ \|I_\tau v - v\|_{L^2(\Omega_T)} &\le C_I^2 \tau. \end{aligned} \tag{3.27}$$

where $\Omega_T := \Omega \times [0,T]$,

$$L^2((0,T);H^1(\Omega)) := \left\{ y \in L^2(\Omega_T) : \int_0^T \int_\Omega |y_x(t,x)|^2 dxdt < \infty \right\},$$

and $|v|_{L^2((0,T);H^1(\Omega))} := \left(\int_0^T \int_\Omega |v_x(t,x)|^2 dxdt \right)^{\frac{1}{2}}$.

Therefore, there exists C_I independent of h, τ, such that

$$\|v - I_{h,\tau}v\|_{L^2(\Omega_T)} \leq C_I(h|v|_{L^2((0,T);H^1(\Omega))} + \tau), \tag{3.28}$$

Furthermore, there exists C_I^3 independent of h, τ, such that

$$|\|v\|_{L^1(\Omega_T)} - \|I_{h,\tau}v\|_{L^1(\Omega_T)}| \leq C_I^3(h|v|_{L^2((0,T);H^1(\Omega))} + \tau). \tag{3.29}$$

Proof. By [6, Corollary 4.4.24], we can derive the first inequality.

And

$$\begin{aligned}\|I_\tau v - v\|^2_{L^2(\Omega_T)} &= \int_\Omega \sum_{j=1}^{N_T} \int_{t_{j-1}}^{t_j} |v(x,t_j) - v(x,t)|^2 dtdx \\ &\leq \int_\Omega \sum_{j=1}^{N_T} \int_{t_{j-1}}^{t_j} L^2(t_j - t)^2 dtdx \\ &= (\mu(\Omega_T)L^2/3)\tau^2\end{aligned} \tag{3.30}$$

where L is the Lipschitz constant for z with respect to t.

Here $C_I^2 := \sqrt{\mu(\Omega_T)L^2/3}$ is the constant number we want.

Therefore

$$\begin{aligned}|\|v\|_{L^1(\Omega_T)} - \|I_{h,\tau}v\|_{L^1(\Omega_T)}| \leq & \|v - I_{h,\tau}v\|_{L^1(\Omega_T)} \\ \leq & \sqrt{\mu(\Omega_T)}\|v - I_{h,\tau}v\|_{L^2(\Omega_T)} \\ \leq & \sqrt{\mu(\Omega_T)}(\|v - I_h v\|_{L^2(\Omega_T)} + \|I_h v - I_{h,\tau}v\|_{L^2(\Omega_T)}) \\ \leq & \sqrt{\mu(\Omega_T)}(C_\pi^1 h|v|_{L^2((0,T);H^1(\Omega))} + C_\pi^2\tau) \\ \leq & C_I^3(h|v|_{L^2((0,T);H^1(\Omega))} + \tau).\end{aligned} \tag{3.31}$$

where $C_\pi^3 := \max\{C_\pi^1, C_\pi^2\}\sqrt{\mu(\Omega_T)}$. □

Proposition 3.8. *For all $z \in L^2((0,T);H^1(\Omega))$, such that z is Lipschitz continuous with respect to t, we have*

$$\begin{aligned}&h\|z - \Pi_h z\|_{L^2(\Omega_T)} + \|z - \Pi_h z\|_{L^2(H^{-1}(\Omega),[0,T])} \leq C_\pi^1 h^2 \|z\|_{L^2((0,T);H^1(\Omega))}, \\ &\|z - \Pi_\tau z\|_{L^2(\Omega_T)} \leq C_\pi^2 \tau, \|z - \Pi_\tau z\|_{L^2(H^{-1}(\Omega),[0,T])} \leq C_\pi^3 \tau.\end{aligned} \tag{3.32}$$

Hence,

$$\begin{aligned} &\|z-\Pi_{h,\tau}z\|_{L^2(\Omega_T)} \leq C_\pi(h\|z\|_{L^2((0,T);H^1(\Omega))}+\tau), \\ &\|z-\Pi_{h,\tau}z\|_{L^2(H^{-1}(\Omega),[0,T])} \leq C'_\pi(h^2\|z\|_{L^2((0,T);H^1(\Omega))}+\tau). \end{aligned} \tag{3.33}$$

The constant numbers $c^1_\pi, c^2_\pi, c^3_\pi, c_\pi, c'_\pi$ are all independent of h, τ.

Proof. The first inequality can be referred to [7]. And the second inequality is similar to that in Proposition 3.7. □

Let us denote S be the solution mapping of the parabolic equation (1.3), and $S_{h,\tau}$ be the solution mapping of its discretization. From [77, Theorem 1.5], we have

Proposition 3.9. *For any $z \in L^2(\Omega_T)$, there exist C^1_s, C^2_s, independent of h, τ, such that*

$$\begin{aligned} &\|(S-S_{h,\tau})z(t_n)\|_{L^2(\Omega)} \leq C^1_s(h^2+\tau), n=1,\cdots,N_T, \\ &\|(S^*-S^*_{h,\tau})z(t_n)\|_{L^2(\Omega)} \leq C^2_s(h^2+\tau), n=1,\cdots,N_T. \end{aligned} \tag{3.34}$$

Given the projections, we can define the approximate L^1-norm as follow

$$\begin{aligned} &\|u\|_{L^1_h(\Omega_T)} := \|I_{h,\tau}|u|\|_{L^1(\Omega_T)}, \\ &\|u\|_{\tilde{L}^1_h(\Omega_T)} := \tau\sum_{j=1}^{N_T}\sum_{i=1}^{N_h}|\int_\Omega (I_{h,\tau}u)(\cdot,t_j)\phi_i(x)dx|. \end{aligned} \tag{3.35}$$

Remark 3.2. *Here we extend the $\tilde{L}^1_h$-norm defined in (3.20) to the whole space $L^2(\Omega_T)$.*

In the following proposition, we prove the relationship between different approximations of L^1-norm.

Proposition 3.10. *For any $u \in L^2(\Omega_T)$, we have*

$$\begin{aligned} &\|\Pi_{h,\tau}u\|_{\tilde{L}^1_h(\Omega_T)} \leq \|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} \leq \|u\|_{\tilde{L}^1_h(\Omega_T)} \leq \|I_{h,\tau}u\|_{L^1(\Omega_T)} \leq \|u\|_{L^1_h(\Omega_T)}, \\ &\|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} \leq \|u\|_{L^1(\Omega_T)} + C^3_I(h^2\|u\|_{L^2((0,T);H^1(\Omega))}+\tau)+ch^2, \end{aligned} \tag{3.36}$$

where C^3_I is defined in (3.29), c is defined in (3.1).

Furthermore, for any $u \in L^2((0,T);H^1(\Omega))$, we have

$$||\|u\|_{\tilde{L}^1_h(\Omega_T)} - \|u\|_{L^1(\Omega_T)}| \leq \sqrt{\mu(\Omega_T)}C'_\pi(h^2\|u\|_{L^2((0,T);H^1(\Omega))} + \tau) + ch^2, \tag{3.37}$$

where C'_π is defined in (3.33).

Proof. By definition, we have

$$\begin{aligned}
\|\Pi_{h,\tau}u\|_{\tilde{L}^1_h(\Omega_T)} &= \tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}|\int_{\Omega_h}\sum_{i=1}^{N_h}\frac{\langle u(\cdot,t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)\phi_k(x)dx|,\\
\|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} &= \tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}\frac{\langle u(\cdot,t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)|dx,\\
\|u\|_{\tilde{L}^1_h(\Omega_T)} &= \tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}|\int_{\Omega_h}\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)\phi_k(x)dx| = \tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}|\langle u(\cdot,t_j),\phi_k\rangle|,\\
\|I_{h,\tau}u\|_{L^1(\Omega_T)} &= \int_0^T\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t)\phi_i(x)|dxdt = \tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)|dx,\\
\|u\|_{L^1_h(\Omega_T)} &= \tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}|u(x_i,t_j)|\phi_i(x)|dx = \tau\sum_{j=1}^{N_T}\sum_{i=1}^{N_h}\int_{\Omega_h}|u(x_i,t_j)|\phi_i(x)dx.
\end{aligned} \tag{3.38}$$

Hence we have

$$\begin{aligned}
\|\Pi_{h,\tau}u\|_{\tilde{L}^1_h(\Omega_T)} =&\tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}|\int_{\Omega_h}\sum_{i=1}^{N_h}\frac{\langle u(t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)\phi_k(x)dx|\\
\leq&\tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}\int_{\Omega_h}|\sum_{i=1}^{N_h}\frac{\langle u(t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)|\phi_k(x)dx\\
=&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}\frac{\langle u(t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)|dx\\
=&\|\Pi_{h,\tau}u\|_{L^1(\Omega_T)}.
\end{aligned} \tag{3.39}$$

Then

$$\begin{aligned}
\|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} =&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}\frac{\langle u(t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)|dx\\
\leq&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}\sum_{i=1}^{N_h}\frac{|\langle u(t_j),\phi_i\rangle|}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)dx\\
=&\|u\|_{\tilde{L}^1_h(\Omega_T)}.
\end{aligned} \tag{3.40}$$

Later, we deduce

$$\begin{aligned}
\|u\|_{\tilde{L}_h^1(\Omega_T)} =&\tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}|\int_{\Omega_h}\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)\phi_k(x)dx|\\
\leq&\tau\sum_{j=1}^{N_T}\sum_{k=1}^{N_h}\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)|\phi_k(x)dx\\
=&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)|\sum_{k=1}^{N_h}\phi_k(x)dx\\
=&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)|dx\\
=&\|u\|_{L^1(\Omega_T)}.
\end{aligned} \tag{3.41}$$

And we have

$$\begin{aligned}
\|u\|_{L^1(\Omega_T)} =&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}u(x_i,t_j)\phi_i(x)|dx\\
\leq&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}\sum_{i=1}^{N_h}|u(x_i,t_j)|\phi_i(x)dx\\
=&\|u\|_{L_h^1(\Omega_T)}.
\end{aligned} \tag{3.42}$$

For the second inequality, we obtain

$$\begin{aligned}
\|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} =&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|\sum_{i=1}^{N_h}\frac{\langle u(\cdot,t_j),\phi_i\rangle}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)|dx\\
\leq&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}\sum_{i=1}^{N_h}\frac{|\langle u(\cdot,t_j),\phi_i\rangle|}{\int_{\Omega_h}\phi_i(x)dx}\phi_i(x)dx\\
=&\tau\sum_{j=1}^{N_T}\sum_{i=1}^{N_h}|\langle u(\cdot,t_j),\phi_i\rangle|\\
=&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}\sum_{i=1}^{N_h}|u(x,t_j)|\phi_i(x)dx\\
=&\tau\sum_{j=1}^{N_T}\int_{\Omega_h}|u(x,t_j)|dx\\
\leq&C_I^3(h^2\|u\|_{L^2((0,T);H^1(\Omega))}+\tau)+ch^2+\|u\|_{L^1(\Omega_T)}.
\end{aligned} \tag{3.43}$$

where the last inequality is by Proposition 3.7 and equation (3.1).

Finally,

$$\begin{aligned}
|\|u\|_{\tilde{L}_h^1(\Omega_T)} - \|u\|_{L^1(\Omega_T)}| \le & \|u\|_{L^1(\Omega_T)} - \|\Pi_{h,\tau}u\|_{L^1(\Omega_T)} + ch^2 \\
\le & \|u - \Pi_{h,\tau}u\|_{L^1(\Omega_T)} + ch^2 \\
\le & \sqrt{\mu(\Omega_T)}\|u - \Pi_{h,\tau}u\|_{L^2(H^{-1}(\Omega),[0,T])} + ch^2 \\
\le & \sqrt{\mu(\Omega_T)}C'_\pi(h^2\|u\|_{L^2((0,T);H^1(\Omega))} + \tau) + ch^2
\end{aligned} \tag{3.44}$$

where the last inequality is by Proposition 3.8. □

Remark 3.3. *From the inequality (3.29), we can derive that*

$$\begin{aligned}
|\|u\|_{L_h^1(\Omega_T)} - \|u\|_{L^1(\Omega_T)}| = & |\|I_{h,\tau}|u|\|_{L^1(\Omega_T)} - \|u\|_{L^1(\Omega_T)}| \\
\le & C_I^3(h\|u\|_{L^2((0,T);H^1(\Omega))} + \tau).
\end{aligned} \tag{3.45}$$

Thus, comparing the above inequality and inequality (3.37), we can see that $\|u\|_{\tilde{L}_h^1(\Omega_T)}$ is a better approximate of $\|u\|_{L^1(\Omega_T)}$ than $\|u\|_{L_h^1(\Omega_T)}$, especially for the situation $\tau = 0$, when the parabolic equation degenerates into an elliptic equation.

Before we prove the error estimate, we give the bounds on the condition number of the mass matrix and the stiffness matrix.

Proposition 3.11. [31, Theorem 1.29, 1.32]. *For $\mathcal{P}1$ approximation on a regular and quasi-uniform subdivision of $\mathbb{R}^n$ which satisfies Assumption 3.1, and for any $x \in \mathbb{R}^{N_h}$, the mass matrix M_h approximates the scaled identity matrix in the sense that*

$$c_1h^2 \le \frac{x^TM_hx}{x^Tx} \le c_2h^2,\ \textit{if}\ n=2,\ \text{and}\ c_1h^3 \le \frac{x^TM_hx}{x^Tx} \le c_2h^3,\ \textit{if}\ n=3.$$

The stiffness matrix K_h satisfies

$$d_1h^2 \le \frac{x^TK_hx}{x^Tx} \le d_2,\ \textit{if}\ n=2,\ \text{and}\ d_1h^3 \le \frac{x^TK_hx}{x^Tx} \le d_2h,\ \textit{if}\ n=3.$$

where the constants c_1, c_2, d_1 and d_2 are independent of the mesh size h.

We further exploit the relationship between the $L^2((0,T);H^1(\Omega))$ norm and the $L^2(\Omega_T)$ norm.

Proposition 3.12. *Given $u_{h,\tau} \in U_{h,\tau}$, $h_0 > 0$, for any $0 < h < h_0$, we have*

$$\|u_{h,\tau}\|_{L^2((0,T);H^1(\Omega))} \le C_H \|u_{h,\tau}\|_{L^2(\Omega_T)} h^{-1}. \tag{3.46}$$

Proof.

$$\begin{aligned}\|u\|^2_{L^2((0,T);H^1(\Omega))} =&\tau \sum_{j=1}^{N_T} \int_\Omega (\sum_{i=1}^{N_h} u(x_i,t_j)\phi_i(x))^2 dx + \tau \sum_{j=1}^{N_T} \int_\Omega (\sum_{i=1}^{N_h} u(x_i,t_j)\nabla\phi_i(x))^2 dx \\ =&\tau(u^T B u + u^T \tilde{A} u) \le c_2 h_0^2 u^T u + d_2 u^T u,\end{aligned} \tag{3.47}$$

where $u = (u_1^T, u_2^T, \cdots, u_{N_T}^T)$, $u_j^T = \{u(x_i,t_j)\}_{i=1}^{N_h}$, $\tilde{A} = Diag(K_h, \cdots, K_h)$.

Similarly, we have

$$\|u\|^2_{L^2(\Omega_T)} = \tau \sum_{j=1}^{N_T} \int_\Omega (\sum_{i=1}^{N_h} u(x_i,t_j)\phi_i(x))^2 dx = \tau u^T B u \ge c_1 h^2 u^T u. \tag{3.48}$$

Hence

$$\begin{aligned}\|u\|^2_{L^2((0,T);H^1(\Omega))} \le& \frac{c_2 h_0^2 u^T u + d_2 u^T u}{c_1 h^2 u^T u} \|u\|^2_{L^2(\Omega_T)} \\ =& \frac{c_2 h_0^2 + d_2}{c_1} \|u\|^2_{L^2(\Omega_T)} h^{-2}\end{aligned} \tag{3.49}$$

□

Under the propositions above, we can then prove the error estimate as below.

Theorem 3.2. *Let $(y^*, u^*), (y^*_{h\tau}, u^*_{h,\tau})$ be the optimal solution for the continuous problem* (P) and ($\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$) respectively. Given $h_0 > 0, \tau_0 > 0$, for any $0 < h < h_0, 0 < \tau < \tau_0$, there exists a constant C_0 such that,

$$\alpha\|u^* - u^*_{h,\tau}\|^2_{L^2(\Omega_T)} + \|y^* - y^*_{h,\tau}\|^2_{L^2(\Omega_T)} \le C_0(\frac{\beta}{\alpha}h^2 + \frac{1}{\alpha}(h^2+\tau) + \frac{1}{\alpha^2}(h^2+\tau)^2). \tag{3.50}$$

Proof. By the optimality conditions for the continuous and discretized problems, we have

$$\begin{aligned}\langle -p^* + \alpha u^*, u^*_{h,\tau} - u^* \rangle_{L^2(\Omega_T)} + \beta(\|u^*_{h,\tau}\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)}) \ge 0, \\ \langle -p^*_{h,\tau} + \alpha u^*_{h,\tau}, \Pi_{h,\tau} u^* - u^*_{h,\tau} \rangle_{L^2(\Omega_T)} + \beta(\|\Pi_{h,\tau} u^*\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)}) \ge 0.\end{aligned} \tag{3.51}$$

Add them together we have

$$\begin{aligned}\langle p^*_{h,\tau}-p^*, u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)} - \alpha\|u^*_{h,\tau}-u^*\|^2_{L^2(\Omega_T)} + \langle p^*_{h,\tau}-\alpha u^*_{h,\tau}, u^*-\Pi_{h,\tau}u^*\rangle_{L^2(\Omega_T)}\\ +\beta(\|u^*_{h,\tau}\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)} + \|\Pi_{h,\tau}u^*\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)}) \geq 0.\end{aligned} \tag{3.52}$$

That is

$$\begin{aligned}\alpha\|u^*_{h,\tau}-u^*\|^2_{L^2(\Omega_T)} \leq & \langle \alpha u^*_{h,\tau}-p^*_{h,\tau}, \Pi_{h,\tau}u^*-u^*\rangle_{L^2(\Omega_T)} + \langle p^*_{h,\tau}-p^*, u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\ & + \beta(\|u^*_{h,\tau}\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)} + \|\Pi_{h,\tau}u^*\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)})\\ = & \langle \alpha u^*-p^*, \Pi_{h,\tau}u^*-u^*\rangle_{L^2(\Omega_T)} + \alpha\langle u^*_{h,\tau}-u^*, \Pi_{h,\tau}u^*-u^*\rangle_{L^2(\Omega_T)}\\ & + \langle p^*_{h,\tau}-p^*, u^*-\Pi_{h,\tau}u^*\rangle_{L^2(\Omega_T)} + \langle p^*_{h,\tau}-p^*, u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\ & + \beta(\|u^*_{h,\tau}\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)} + \|\Pi_{h,\tau}u^*\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)})\\ := & I_1 + I_2 + I_3 + I_4 + I_5.\end{aligned} \tag{3.53}$$

From the box constraint, we know $\|u^*\|_{L^2(\Omega_T)}$ and $\|u^*_{h,\tau}\|_{L^2(\Omega_T)}$ are bounded. Then H^1-norm of the states y^*, $y^*_{h,\tau}$, and the adjoint states $p^*, p^{h,\tau}$ are bounded. Hence we can choose a large enough constant $C_1 > 0$, independent of α, h, τ and a constant α_0, such that: for all $0 < \alpha \leq \alpha_0$ and $0 < h < h_0$, $0 < \tau < \tau_0$,

$$\begin{aligned}& 2\|p^*\|_{L^2((0,T);H^1(\Omega))} + \|y^*\|_{L^2((0,T);H^1(\Omega))} + \|y^*_{h,\tau}\|_{L^2((0,T);H^1(\Omega))}\\ & + T(\beta+\alpha(b-a))\mu(\Omega)^{\frac{1}{2}} + \|y^*-y_d\|_{L^2(\Omega_T)} + \|u^*_{h,\tau}\|_{L^2(\Omega_T)} \leq C_1.\end{aligned} \tag{3.54}$$

By (3.23), we have

$$\|u^*\|_{L^2((0,T);H^1(\Omega))} \leq \frac{1}{\alpha}\|p^*\|_{L^2((0,T);H^1(\Omega))} + T(\frac{\beta}{\alpha}+b-a)\mu(\Omega)^{\frac{1}{2}} \leq \alpha^{-1}C_1. \tag{3.55}$$

And we can then obtain

$$\|\alpha u^*-p^*\|_{L^2((0,T);H^1(\Omega))} \leq 2\|p^*\|_{L^2((0,T);H^1(\Omega))} + T(\beta+\alpha(b-a))\mu(\Omega)^{\frac{1}{2}} \leq C_1. \tag{3.56}$$

By Proposition 3.8, we see that

$$\begin{aligned} I_1 =& \langle \alpha u^* - p^*, \Pi_{h,\tau} u^* - u^* \rangle_{L^2(\Omega_T)} \\ \leq & \|\alpha u^* - p^*\|_{L^2((0,T);H^1(\Omega))} \|\Pi_{h,\tau} u^* - u^*\|_{L^2(H^{-1}(\Omega),[0,T])} \\ \leq & C_1 C'_\pi (C_1 \alpha^{-1} h^2 + \tau) \\ \leq & C_2(\alpha^{-1} h^2 + \tau), \end{aligned} \tag{3.57}$$

where $C_2 := \max\{C_1^2 C'_\pi, C_1 C'_\pi, C'_\pi\}$.

And

$$\begin{aligned} I_2 =& \alpha \langle u^*_{h,\tau} - u^*, \Pi_{h,\tau} u^* - u^* \rangle_{L^2(\Omega_T)} \\ \leq & \frac{\alpha}{4} \|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega_T)} + \alpha \|\Pi_{h,\tau} u^* - u^*\|^2_{L^2(\Omega_T)} \\ \leq & \frac{\alpha}{4} \|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega_T)} + \alpha C'_\pi (h \|u^*\|_{L^2((0,T);H^1(\Omega))} + \tau)^2 \\ \leq & \frac{\alpha}{4} \|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega_T)} + \alpha C'_\pi (\alpha^{-1} C_1 h + \tau)^2 \\ \leq & \frac{\alpha}{4} \|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega_T)} + \alpha C_2 (\alpha^{-1} h + \tau)^2, \end{aligned} \tag{3.58}$$

where the second inequality is by Proposition 3.8.

For I_3, we have

$$\begin{aligned} I_3 =& \langle p^*_{h,\tau} - p^*, u^* - \Pi_{h,\tau} u^* \rangle_{L^2(\Omega_T)} \\ =& \langle S^*_{h,\tau}(y_d - S_{h,\tau}(u^*_{h,\tau} + y_c)) - S^*(y_d - S(u^* + y_c)), u^* - \Pi_{h,\tau} u^* \rangle_{L^2(\Omega_T)} \\ =& \langle (S - S_{h,\tau})(u^*_{h,\tau} + y_c), S(u^* - \Pi_{h,\tau} u^*) \rangle_{L^2(\Omega_T)} \\ &+ \langle (S^* - S^*_{h,\tau})(y^*_{h,\tau} - y_d), u^* - \Pi_{h,\tau} u^* \rangle_{L^2(\Omega_T)} \\ &- \langle S(u^*_{h,\tau} - u^*), S(u^* - \Pi_{h,\tau} u^*) \rangle_{L^2(\Omega_T)} \\ &+ \langle (S^*_{h,\tau} - S^*) y_d, u^* - \Pi_{h,\tau} u^* \rangle_{L^2(\Omega_T)} \\ \leq & \frac{1}{2} \|(S - S_{h,\tau})(u^*_{h,\tau} + y_c)\|^2_{L^2(\Omega_T)} + \|S(u^* - \Pi_{h,\tau} u^*)\|^2_{L^2(\Omega_T)} \\ &+ \frac{1}{\alpha} \|(S^* - S^*_{h,\tau})(y^*_{h,\tau} - y_d)\|^2_{L^2(\Omega_T)} + \frac{\alpha}{2} \|u^* - \Pi_{h,\tau} u^*\|^2_{L^2(\Omega_T)} \\ &+ \frac{1}{\alpha} \|(S^*_{h,\tau} - S^*) y_d\|^2_{L^2(\Omega_T)} + \frac{1}{2} \|S(u^*_{h,\tau} - u^*)\|^2_{L^2(\Omega_T)}. \end{aligned} \tag{3.59}$$

By Proposition 3.8 and Proposition 3.9, we know

$$\begin{aligned}
&\|(S_{h,\tau}-S)(u^*_{h,\tau}+y_c)\|_{L^2(\Omega_T)} \le \max_{t_n}\|(S_{h,\tau}-S)(u^*_{h,\tau}+y_c)(t_n)\|_{L^2(\Omega)} \le C_s^1(h^2+\tau),\\
&\|(S^*_{h,\tau}-S^*)z_{h,\tau}\|_{L^2(\Omega_T)} \le \max_{t_n}\|(S^*_{h,\tau}-S^*)z_{h,\tau}(t_n)\|_{L^2(\Omega)} \le C_s^2(h^2+\tau),\\
&\|\Pi_{h,\tau}z-z\|_{L^2(\Omega_T)} \le C_\pi(h\|z\|_{L^2(H^1(\Omega),[0,T)}+\tau),\\
&\|\Pi_{h,\tau}z-z\|_{L^2(H^{-1}(\Omega),[0,T])} \le C'_\pi(h^2\|z\|_{L^2(H^1(\Omega),[0,T])}+\tau).
\end{aligned} \tag{3.60}$$

Hence

$$\begin{aligned}
I_3 \le&\|S(u^*-\Pi_{h,\tau}u^*)\|^2_{L^2(\Omega_T)}+\frac{1}{2}\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}\\
&+(C_3)^2((h^2+\tau)^2+\frac{1}{\alpha}(h^2+\tau)^2+\alpha(\alpha^{-1}h+\tau)^2),\\
\le&\|S\|^2_{\mathcal{L}(H^{-1},L^2)}\|u^*-\Pi_{h,\tau}u^*\|^2_{L^2(H^{-1}(\Omega),[0,T])}+\frac{1}{2}\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}\\
&+(C_3)^2((h^2+\tau)^2+\frac{1}{\alpha}(h^2+\tau)^2+\alpha(\alpha^{-1}h+\tau)^2)\\
\le&(C_4)^2((\alpha^{-1}h^2+\tau)^2+(h^2+\tau)^2+\frac{1}{\alpha}(h^2+\tau)^2+\alpha(\alpha^{-1}h+\tau)^2)+\frac{1}{2}\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}.
\end{aligned} \tag{3.61}$$

where $C_3 := \max\{C_s^1, 2C_s^2, C_\pi, C'_\pi, C_\pi C_1, C'_\pi C_1\}$, $C_4 := \max\{C_3, \|S\|_{\mathcal{L}(H^{-1},L^2)}\}$.

Similarly, we obtain

$$\begin{aligned}
I_4 =&\langle p^*_{h,\tau}-p^*, u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\
=&\langle S^*_{h,\tau}(y_d-S_{h,\tau}(u^*_{h,\tau}+y_c))-S^*(y_d-S(u^*+y_c)), u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\
=&\langle (S-S_{h,\tau})(u^*_{h,\tau}+y_c), S(u^*_{h,\tau}-u^*)\rangle_{L^2(\Omega_T)}+\langle (S^*-S^*_{h,\tau})(y^*_{h,\tau}-y_d), u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\
&-\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}+\langle (S^*_{h,\tau}-S^*)y_d, u^*_{h,\tau}-u^*\rangle_{L^2(\Omega_T)}\\
\le&\|(S-S_{h,\tau})(u^*_{h,\tau}+y_c)\|^2_{L^2(\Omega_T)}+\frac{2}{\alpha}\|(S^*-S^*_{h,\tau})(y^*_{h,\tau}-y_d)\|^2_{L^2(\Omega_T)}+\frac{\alpha}{4}\|u^*_{h,\tau}-u^*\|^2_{L^2(\Omega_T)}\\
&+\frac{2}{\alpha}\|(S^*_{h,\tau}-S^*)y_d\|^2_{L^2(\Omega_T)}-\frac{3}{4}\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}\\
\le&4(C_3)^2((h^2+\tau)^2+\frac{1}{\alpha}(h^2+\tau)^2)+\frac{\alpha}{4}\|u^*_{h,\tau}-u^*\|^2_{L^2(\Omega_T)}-\frac{3}{4}\|S(u^*_{h,\tau}-u^*)\|^2_{L^2(\Omega_T)}.
\end{aligned} \tag{3.62}$$

Finally, for I_5, by Proposition 3.10, we have

$$\begin{aligned} I_5 =&\beta(\|u^*_{h,\tau}\|_{L^1(\Omega_T)} - \|u^*\|_{L^1(\Omega_T)} + \|\Pi_{h,\tau}u^*\|_{\tilde{L}^1_h(\Omega_T)} - \|u^*_{h,\tau}\|_{\tilde{L}^1_h(\Omega_T)}) \\ \leq&\beta(\sqrt{\mu(\Omega_T)}C_I^3h^2\|u^*_{h,\tau}\|_{L^2((0,T);H^1(\Omega))} + C'_\pi h^2\|u^*\|_{L^2((0,T);H^1(\Omega))} + 2\tau + 2ch^2) \\ \leq&C_4\beta(h^2\|u^*_{h,\tau}\|_{L^2((0,T);H^1(\Omega))} + \alpha^{-1}h^2 + h^2 + \tau) \\ \leq&C_4\beta(h^2(\|u^*\|_{L^2((0,T);H^1(\Omega))} + \|u^*_{h,\tau} - u^*\|_{L^2((0,T);H^1(\Omega))}) + \alpha^{-1}h^2 + h^2 + \tau) \end{aligned} \tag{3.63}$$

where $C_4 := \max\{\sqrt{\mu(\Omega_T)}C_I^3, C'_\pi C_1, 2, 2c\}$.

Hence, it remains to estimate $\|u^*_{h,\tau} - u^*\|_{L^2((0,T);H^1(\Omega))}$.

$$\begin{aligned} \|u^*_{h,\tau} - u^*\|_{L^2((0,T);H^1(\Omega))} \leq&\|u^*_{h,\tau} - I_{h,\tau}u^*\|_{L^2((0,T);H^1(\Omega))} + \|I_{h,\tau}u^* - u^*\|_{L^2((0,T);H^1(\Omega))} \\ \leq&C_H\|u^*_{h,\tau} - I_{h,\tau}u^*\|_{L^2(\Omega)}h^{-1} + C_I^4\|u^*\|_{L^2(H^1(\Omega,[0,T]))} + C_I^2\tau \\ \leq&C_Hh^{-1}(\|u^*_{h,\tau} - u^*\|_{L^2(\Omega)} + \|u^* - I_{h,\tau}u^*\|_{L^2(\Omega)}) \\ &+ C_I^4\|u^*\|_{L^2(H^1(\Omega,[0,T])} + C_I^2\tau \\ \leq&C_Hh^{-1}\|u^*_{h,\tau} - u^*\|_{L^2(\Omega)} + C_HC_I(\|u^*\|_{L^2((0,T);H^1(\Omega))} + h^{-1}\tau) \\ &+ C_I^4\|u^*\|_{L^2(H^1(\Omega,[0,T])} + C_I^2\tau \\ \leq&C_5h^{-1}(\|u^*_{h,\tau} - u^*\|_{L^2(\Omega)} + \tau) + 2\alpha^{-1}. \end{aligned} \tag{3.64}$$

where $C_5 := \max\{C_H, C_HC_I, C_I^4C_1, C_I^2, \tau_0\}$.

Hence we have

$$\begin{aligned} I_5 \leq&C_4\beta(h^2(\|u^*\|_{L^2((0,T);H^1(\Omega))} + \|u^*_{h,\tau} - u^*\|_{L^2((0,T);H^1(\Omega))}) + \alpha^{-1}h^2 + h^2 + \tau) \\ \leq&C_4\beta(h^2\alpha^{-1}C_1 + C_5h(\|u^*_{h,\tau} - u^*\|_{L^2(\Omega)} + \tau) + 3\alpha^{-1}h^2 + h^2 + \tau) \\ \leq&C_6\beta h\|u^*_{h,\tau} - u^*\|_{L^2(\Omega)} + C_6\beta(2\alpha^{-1}h^2 + h\tau + h^2 + \tau) \\ \leq&2(C_6)^2\beta^2h^2 + \frac{1}{8}\|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega)} + C_6\beta(2\alpha^{-1}h^2 + h\tau + h^2 + \tau). \end{aligned} \tag{3.65}$$

From (3.57), (3.58), (3.61), (3.62) and (3.65), we can conclude that

$$\begin{aligned}
&\frac{\alpha}{8}\|u^*_{h,\tau} - u^*\|^2_{L^2(\Omega_T)} + \frac{1}{4}\|S(u^*_{h,\tau} - u^*)\|^2_{L^2(\Omega_T)} \\
\leq &C_2((\alpha^{-1}h^2 + \tau) + \alpha(\alpha^{-1}h + \tau)^2) + 5(C_4)^2((\alpha^{-1}h^2 + \tau)^2 + (h^2 + \tau) + \frac{1}{\alpha}(h^2 + \tau)^2) \\
&+ (2(C_6)^2\beta^2h^2 + C_6\beta(2\alpha^{-1}h^2 + h\tau + h^2 + \tau)) \\
\leq &C_0((\alpha^{-1} + 1 + \alpha^{-1}\beta + \beta^2)h^2 + (\alpha^{-2} + \alpha^{-1})h^4 + (1 + \beta)\tau + (\alpha + 1 + \alpha^{-1})\tau^2 + \beta h\tau),
\end{aligned} \tag{3.66}$$

where the last inequality is by the fact that when h, τ are chosen small enough, many terms can be omitted, and $C_0 := \max\{C_2, 5(C_4)^2, 2(C_6)^2, 2C_6\}$. □

Corollary 3.1. *Let (y^*, u^*) be the optimal solution of problem (P), $(y^*_{h,\tau}, u^*_{h,\tau})$ be the optimal solution of problem $(\widehat{\mathrm{P}}_{\mathrm{h},\tau})$, then for each $h_0 > 0, \tau_0 > 0, \alpha_0 > 0, \beta_0 > 0$ small enough, there exists a constant C_0, such that for all $0 < \alpha \leq \alpha_0, 0 < h \leq h_0, 0 < \tau \leq \tau_0, 0 < \beta \leq \beta_0$, it holds*

$$\begin{aligned}
\|u^*_{h,\tau} - u^*\|_{L^2(\Omega_T)} &\leq C_0(\frac{h+\tau}{\alpha} + \frac{h^2}{\alpha^{\frac{3}{2}}} + \sqrt{\frac{\tau}{\alpha}} + \sqrt{\frac{\beta h\tau}{\alpha}}), \\
\|S(u^*_{h,\tau} - u^*)\|_{L^2(\Omega_T)} &\leq C_0(\frac{h+\tau}{\sqrt{\alpha}} + \frac{h^2}{\alpha} + \sqrt{\tau} + \sqrt{\beta h\tau}), \\
\|y^*_{h,\tau} - y^*\|_{L^2(\Omega_T)} &\leq C_0(\frac{h+\tau}{\sqrt{\alpha}} + \frac{h^2}{\alpha} + \sqrt{\tau} + \sqrt{\beta h\tau}).
\end{aligned} \tag{3.67}$$

Remark 3.4. *Later in the numerical experiment, we will check the error order with respect to the mesh-size h to be at least 1, given fixed α, β, τ.*

Chapter 4

Accelerated block coordinate descent method

In this chapter, we consider solving the dual problems of the discretized primal problems ($\mathrm{P}_{\mathrm{h},\tau}$) and ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$).

Conventionally, we can discretize the primal problem with the approximate L^1-norm, and then derive the dual problem. In this way, we will apply the inexact majorized ABCD method introduced in Chapter 2 to solve the discretized problem. As the approximate discretized L^1-norm is decoupled, we call problem ($\mathrm{P}_{\mathrm{h},\tau}$) the decoupled SPOCP.

Besides, to avoid the approximation of L^1-norm, we study the dual problem (D). Through solving the dual problem, we avoid the error caused by approximation. For implementation, we utilize the symmetric Gauss-Seidel technique, which is introduced in Section 2.2, with the inexact majorized ABCD method.

4.1 The sGS-imABCD method for solving SPOCPs

We discretize the dual problem and obtain the following discretized problem,

$$\begin{aligned} \min_{\mu,\lambda,p\in\mathbb{R}^m} \Phi(\mu,\lambda,p) :=& \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|\lambda+\mu-p\|^2_B + \langle By_c, p\rangle \\ &+ \delta_{[-\beta,\beta]}(\lambda) + \delta^*_{[a,b]}(B\mu) - \frac{1}{2}\|y_d\|^2_B. \end{aligned} \tag{$\widehat{\mathrm{D}}_{\mathrm{h},\tau}$}$$

It is an unconstrained optimization problem with two separable nonsmooth terms. Therefore we apply the previous mentioned imABCD method to solve it.

For later use, we also provide its pre-dual problem as follow,

$$\begin{cases} \min\limits_{y,u,v,w\in\mathbb{R}^m} & J(y,u,v,w) = \frac{1}{2}\|y-y_d\|_B^2 + \frac{\alpha}{2}\|u\|_B^2 + \beta\|Bu\|_1 + \delta_{[a,b]}(u) \\ \text{s.t.} & Ay = B(u+y_c), \\ & B(u-v) = 0, \\ & B(u-w) = 0. \end{cases} \tag{4.1}$$

where p, λ, μ are the Lagrangian multipliers for the equalities constraints, respectively.

4.1.1 Numerical implementation

Now, we can apply Algorithm 1 to ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$), with μ taken as one block, and (λ, p) as the other one. For convenience, we denote $z = (\mu, \lambda, p)$ and

$$\phi(z) = \phi(\mu,\lambda,p) = \frac{1}{2\alpha}\|\lambda+\mu-p\|_B^2 - \frac{1}{2}\|y_d\|_B^2. \tag{4.2}$$

Obviously ϕ is quadratic with the Hessian matrix

$$\mathcal{Q} = \frac{1}{\alpha}\begin{pmatrix} B & B & -B \\ B & B & -B \\ -B & -B & B \end{pmatrix}. \tag{4.3}$$

Additionally, we assume that there exist two self-adjoint positive semidefinite operators $\mathcal{D}_1$ and $\mathcal{D}_2$, such that Assumation 2.1 holds. Then we can majorize $\phi(\mu,\lambda,p)$ at $z' = (\mu', \lambda', p')$ as

$$\phi(z) \le \hat{\phi}(z;z') = \phi(z) + \frac{1}{2}\|\mu-\mu'\|_{\mathcal{D}_1}^2 + \frac{1}{2}\left\|\begin{pmatrix}\lambda\\ p\end{pmatrix} - \begin{pmatrix}\lambda'\\ p'\end{pmatrix}\right\|_{\mathcal{D}_2}^2. \tag{4.4}$$

The framework of imABCD for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$) is given below:

Algorithm 2: (imABCD algorithm for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$))

Input: $(\mu^1, \lambda^1, p^1) = (\tilde{\mu}^0, \tilde{\lambda}^0, \tilde{p}^0) \in \mathrm{dom}(\delta^*_{[a,b]}) \times [-\beta, \beta] \times \mathbb{R}^{N_h}$. Set $k = 1, t_1 = 1$.

Output: $(\tilde{\mu}^k, \tilde{\lambda}^k, \tilde{p}^k)$.

Iterate until convergence

Step 1 Compute

$$\begin{aligned}
\tilde{\mu}^k &= \arg\min \delta^*_{[a,b]}(B\mu) + \phi(\mu, \lambda^k, p^k) + \frac{1}{2}\|\mu - \mu^k\|^2_{\mathcal{D}_1} - \langle \delta^k_\mu, \mu \rangle, \\
(\tilde{\lambda}^k, \tilde{p}^k) &= \arg\min \delta_{[-\beta,\beta]}(\lambda) + \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \langle By_c, p \rangle \\
&\quad + \phi(\tilde{\mu}^k, \lambda, p) + \frac{1}{2}\left\| \begin{pmatrix} \lambda \\ p \end{pmatrix} - \begin{pmatrix} \lambda^k \\ p^k \end{pmatrix} \right\|^2_{\mathcal{D}_2} - \langle \delta^k_\lambda, \lambda \rangle - \langle \delta^k_p, p \rangle.
\end{aligned}$$

Step 2 Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$, Compute

$$\begin{aligned}
&\mu^{k+1} = \tilde{\mu}^k + \beta_k(\tilde{\mu}^k - \tilde{\mu}^{k-1}),\ p^{k+1} = \tilde{p}^k + \beta_k(\tilde{p}^k - \tilde{p}^{k-1}), \\
&\lambda^{k+1} = \tilde{\lambda}^k + \beta_k(\tilde{\lambda}^k - \tilde{\lambda}^{k-1}).
\end{aligned}$$

Now we discuss the issue on how to choose the two operators $\mathcal{D}_1$ and $\mathcal{D}_2$.

From the perspective of both theoretical analysis and numerical implementation, it is very important to choose two appropriate and effective operators $\mathcal{D}_1$ and $\mathcal{D}_2$. For the numerical efficiency, the general principle is to achieve the goal that both $\mathcal{D}_1$ and $\mathcal{D}_2$ can be as small as possible while the corresponding subproblems can be solved relatively easily.

Firstly, for the proximal term $\frac{1}{2}\|\mu - \mu^k\|^2_{\mathcal{D}_1}$, we choose

$$\mathcal{D}_1 := \frac{1}{\alpha}\gamma BC^{-1}B - \frac{1}{\alpha}B, \quad \text{where } \gamma = \begin{cases} 4 & if\ n = 2, \\ 5 & if\ n = 3. \end{cases} \tag{4.5}$$

Here C is defined in section 3.1 as $C = Diag(W_h, W_h, \cdot, W_h)$, with W_h being the stiffness matrix.

The μ-subproblem, though, does not have an analytical solution, we can solve it in a rather cheap way. The details will be explained later.

Next, we focus on the choice of the operator $\mathcal{D}_2$. The objective function of (λ, p)-subproblem is a sum of a two-block quadratic function and a non-smooth function involving only the first block λ. Inspired by the inexact sGS technique, which is introduced in Section 2.2, we hope to find a proper $\mathcal{D}_2$ such that solving the above subproblem is equivalent to solve the following subproblems by order:

$$\begin{cases} \hat{p}^k = \arg\min \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|p - \lambda^k - \tilde{\mu}^k\|^2_B + \langle By_c, p\rangle - \langle \hat{\delta}^k_p, p\rangle, \\ \tilde{\lambda}^k = \arg\min \frac{1}{2\alpha}\|\lambda - (\hat{p}^k - \tilde{\mu}^k)\|^2_B + \delta_{[-\beta,\beta]}(\lambda) - \langle \delta^k_\lambda, \lambda\rangle + \frac{1}{2}\|\lambda - \lambda^k\|^2_{\mathcal{D}_\lambda}, \\ \tilde{p}^k = \arg\min \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|p - \tilde{\lambda}^k - \tilde{\mu}^k\|^2_B + \langle By_c, p\rangle - \langle \delta^k_p, p\rangle. \end{cases} \tag{4.6}$$

Let

$$\mathcal{Q}' := \frac{1}{\alpha}\begin{pmatrix} B & -B \\ -B & B \end{pmatrix} \tag{4.7}$$

It is obvious that $Q'_{11} = Q'_{22} = B$, and thus Q'_11, Q'_22 are both positive definite. Then by Theorem 2.1, we only need to choose the $\mathcal{D}_2 = sGS(\mathcal{Q}') + \mathcal{D}_\lambda$.

Here we add a proximal term $\mathcal{D}_\lambda$, in order to make the λ-subproblem in (4.6) have a decouple form. To address this problem and make the subproblem having a closed form solution, we take advantage of the relationship between the matrix B and the matrix C and set $\mathcal{D}_\lambda = \frac{1}{\alpha}(C - B)$.

In summary, we choose $\mathcal{D}_2$ to be

$$\mathcal{D}_2 = sGS(\mathcal{Q}') + \frac{1}{\alpha}\begin{pmatrix} C-B & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{\alpha}\begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.8}$$

where $sGS(\cdot)$ can refer to Section 2.2.

Based on the choice of $\mathcal{D}_1$ and $\mathcal{D}_2$, we get the majorized Hessian matrix $\widehat{\mathcal{Q}}$ as

$$\widehat{\mathcal{Q}} = \mathcal{Q} + \frac{1}{\alpha}\begin{pmatrix} \gamma BC^{-1}B - B & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4.9}$$

Then, according to the choice of $\mathcal{D}_1$ and $\mathcal{D}_2$, we give the detailed framework of our inexact sGS based majorized ABCD method (called sGS-imABCD) for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$) as follows.

Algorithm 3: (sGS-imABCD algorithm for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$))

```
Input: $(\mu^1, \lambda^1, p^1) = (\tilde{\mu}^0, \tilde{\lambda}^0, \tilde{p}^0) \in \mathrm{dom}(\delta^*_{[a,b]}) \times [-\beta, \beta] \times \mathbb{R}^{N_h}$. Let $\{\epsilon_k\}$ be a
    nonincreasing sequence of nonnegative numbers such that $\sum_{k=1}^{\infty} k\epsilon_k < \infty$. Set
    $k = 1, t_1 = 1$.
Output: $(\tilde{\mu}^k, \tilde{\lambda}^k, \tilde{p}^k)$
Iterate until convergence,
Step 1 Choose error tolerance $\delta^k_\mu, \hat{\delta}^k_p, \delta^k_p$ such that
        $\max\{\|\delta^k_\mu\||, \|\hat{\delta}^k_p\||, \|\delta^k_p\||\} \le \epsilon_k.$
    Compute
        $\tilde{\mu}^k = \arg\min_{\mu} \frac{1}{2\alpha}\|\mu - (p^k - \lambda^k)\|^2_B + \delta^*_{[a,b]}(B\mu) + \frac{1}{2}\|\mu - \mu^k\|^2_{\mathcal{D}_1} - \langle \delta^k_\mu, \mu \rangle,$
        $\hat{p}^k = \arg\min_{p} \frac{1}{2}\|A^* p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|p - \lambda^k - \tilde{\mu}^k\|^2_B + \langle By_c, p \rangle - \langle \hat{\delta}^k_p, p \rangle,$
        $\tilde{\lambda}^k = \arg\min_{\lambda} \frac{1}{2\alpha}\|\lambda - (\hat{p}^k - \tilde{\mu}^k)\|^2_B + \delta_{[-\beta,\beta]}(\lambda) + \frac{1}{2\alpha}\|\lambda - \lambda^k\|^2_{C-B},$
        $\tilde{p}^k = \arg\min_{p} \frac{1}{2}\|A^* p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|p - \tilde{\lambda}^k - \tilde{\mu}^k\|^2_B + \langle By_c, p \rangle - \langle \delta^k_p, p \rangle.$
Step 2 Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$, Compute
        $\mu^{k+1} = \tilde{\mu}^k + \beta_k(\tilde{\mu}^k - \tilde{\mu}^{k-1}), \quad p^{k+1} = \tilde{p}^k + \beta_k(\tilde{p}^k - \tilde{p}^{k-1}),$
        $\lambda^{k+1} = \tilde{\lambda}^k + \beta_k(\tilde{\lambda}^k - \tilde{\lambda}^{k-1}).$
```

Numerical computation of the block μ and λ subproblems

For the first subproblem of Algorithm 3 in kth iteration, at first glance, there is no closed form solution for the variable μ. However, if we introduce an artificial variable $\xi = B\mu$, we can obtain a closed form solution about ξ. Thus, solving the subproblem about the variable μ can be tranformed to solving the following subproblem first,

$$\begin{aligned}\tilde{\xi}^k &= \arg\min \frac{1}{2\alpha}\|\xi - B(p^k - \lambda^k)\|^2_{B^{-1}} + \delta^*_{[a,b]}(\xi) + \frac{1}{2\alpha}\|\xi - \xi^k\|^2_{\gamma C^{-1} - B^{-1}} \\ &= \arg\min \frac{1}{2\alpha}\|\xi - (\xi^k + \frac{1}{\gamma}C(p^k - \lambda^k - \mu^k))\|^2_{\gamma C^{-1}} + \delta^*_{[a,b]}(\xi).\end{aligned} \tag{4.10}$$

then solve for $\tilde{\mu}^k = B^{-1}\tilde{\xi}^k$.

To solve (4.10), we first introduce the proximal mapping $\text{prox}^f_{\mathcal{M}}(\cdot)$ with respect to a self-adjoint positive definite linear operator $\mathcal{M}$, which is defined as,

$$\text{prox}^f_{\mathcal{M}}(x) = \arg\min_z\{f(z) + \frac{1}{2}\|z - x\|^2_{\mathcal{M}}\}, \quad \forall x \in \mathcal{X}, \tag{4.11}$$

where f is a closed proper convex function f and $\mathcal{X}$ is a finite-dimensional real Euclidean space.

For the proximal mapping, we have the following Moreau identity which is shown in [51, Proposition 2.4]:

$$x = \text{prox}^f_{\mathcal{M}}(x) + \mathcal{M}^{-1}\text{prox}^{f^*}_{\mathcal{M}^{-1}}(\mathcal{M}x), \tag{4.12}$$

where f^* is the conjugate function of f. Thus, making use of the Moreau identity (4.12), we derive

$$\tilde{\xi}^k = v^k - \frac{\alpha}{\gamma}C\Pi_{[a,b]}(\frac{\gamma}{\alpha}C^{-1}v^k), \tag{4.13}$$

where

$$v^k \quad = \quad B\mu^k + \frac{1}{\gamma}C(p^k - \lambda^k - \mu^k). \tag{4.14}$$

Hence we obtain the analytic form of solution for the z-subproblem (4.10). And this is the important reason why we choose the proximal term $\frac{1}{2\alpha}\|\mu - \mu^k\|^2_{\gamma BC^{-1}B - B}$ for μ.

Now we discuss the numerical method to solve for $\tilde{\mu}^k = B^{-1}\tilde{\xi}^k$. Based on the eigenvalues bounds for the mass matrix given in [83], we suggest that it is an appropriate choice to use a fix number steps of Chebyshev semi-iteration to represent approximation of B^{-1}. For more details on the Chebyshev semi-iteration method we refer to [65, 84]. In actual numerical implementation, we use 20 steps of Chebyshev semi-iteration and set the error tolerance to be 10^{-12}, which can guarantee the error vector $\|\delta_{\mu}^{k}\|_2 \leq \epsilon_k$.

For the block λ, since the matrix C is a diagonal positive definite matrix, we easily derive that

$$\tilde{\lambda}^k = \Pi_{[-\beta,\beta]}(s^k),$$

where $s^k = \lambda^k + C^{-1}B(\hat{p}^k - \tilde{\mu}^k - \lambda^k)$.

For the approximate discretized problem, we compute the λ-subproblem in a similar way,

$$\begin{cases} w^k = \dfrac{1}{\gamma}(p^k - \mu^k + \gamma C^{-1}B\lambda^k - \lambda^k), \\ z = \Pi_{[-\beta,\beta]}(w^k), \\ \lambda = B^{-1}C(z). \end{cases} \tag{4.15}$$

An efficient iteration method and preconditioner for the block $\hat{p}$ subproblem

The main computation time of our sGS-imABCD algorithm is on solving p-subproblem. Thus, it is crucial to improve the efficiency of our algorithm by employing a fast strategy to solve the p-subproblems. For the subproblem about $\hat{p}^k$, if we ignore the error vector $\hat{\delta}_p^k$, it is obvious to see that solving the subproblem is equivalent to solving the following system:

$$AB^{-1}(A^*\hat{p}^k - By_d) + \frac{1}{\alpha}B(\hat{p}^k - \lambda^k - \tilde{\mu}^k) + By_c = 0. \tag{4.16}$$

Since $A^*p = B(y_d - y)$, then (4.16) can be rewritten as:

$$\mathcal{A}w^{k+1} \equiv \begin{pmatrix} B & -\alpha A \\ A^* & B \end{pmatrix} \begin{pmatrix} \hat{p}^k \\ \hat{y}^k \end{pmatrix} = \begin{pmatrix} B(\lambda^k + \tilde{\mu}^k - \alpha y_c) \\ By_d \end{pmatrix}. \tag{4.17}$$

Clearly, the linear system (4.17) is a special case of the generalized saddle-point system, thus some Krylov-based methods can be used to inexactly solve the linear system by constructing a good preconditioner. From the structure of the equation, the preconditioner is given as

$$\mathcal{P} = \frac{1}{\alpha}\begin{pmatrix} B & -\alpha A \\ A^* & \sqrt{\alpha}(A + A^*) \end{pmatrix}. \tag{4.18}$$

which is introduced in [3], is employed to precondition the generalized minimal residual (GMRES) method for solving (4.17). About the spectral properties of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$, we introduce the following theorem, see [3] for more details.

Theorem 4.1. [3, Proposition 4]. *When $\mathcal{P}$ is used to precondition the matrix $\mathcal{A}$, the eigenvalues of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ are contained within the interval $[\frac{1}{2}, 1]$ if $\alpha > 0$ and $ker(A) \cap ker(B) = \{0\}$. Therefore, the bound $\kappa(\mathcal{P}^{-1}\mathcal{A}) \leq 2$.*

Since B is positive definite, we can see that $ker(A) \cap ker(B) \subset ker(B) = \{0\}$.

In actual implementations, the action of the preconditioning matrix, when used to precondition the *Krylov* subspace methods, is realized through solving a sequence of generalized residual equations of the form

$$\mathcal{P}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{4.19}$$

By making use of the structure of the matrix $\mathcal{P}$, we obtain the following procedure for computing the vector $\begin{pmatrix} x \\ y \end{pmatrix}$,

Algorithm 4: Numerical implementation for (4.19)

```
1 . Solve $H_1 g = f_1 + \sqrt{\alpha} f_2$.
2 . Compute $Bg$ and $f_1 - Bg$.
3 . Solve $H_2 h = f_1 - Bg$.
4 . Compute $x = g + h$ and $y = -h/\sqrt{\alpha}$.
```

Here $H_1 = B + \sqrt{\alpha}A^*, H_2 = B + \sqrt{\alpha}A$.

Remark 4.1. *Noticed that B is block diagonal matrix, A^* is block upper triangular matrix, hence H_1 is a block upper triangular matrix. Thus we can solve the linear equation $H_1 g = f_1 + \sqrt{\alpha} f_2$ from the last block to the first block. And each small linear equations can be inexactly handled with some alternative efficient methods, e.g., preconditioned conjugate gradient (PCG) method, Chebyshev semi-iteration or some multigrid scheme. Similarly, we can see that H_2 is a block lower triangular matrix, and we can solve the equation from the first block to the last block.*

It is well known that the convergence behavior of iterative solution methods will be precisely characterized in terms of $\kappa(M_h)$ and $\kappa(K_h)$, which represents the condition number of M_h and K_h, respectively. Then about the bounds on the condition number, we can have the following results, one can see Proposition 1.29 and Theorem 1.32 in [31] for more details.

In addition, let $(\hat{r}_1^k, \hat{r}_2^k)$ be the residual error vector, which means:

$$\begin{pmatrix} B & -\alpha A \\ A^* & B \end{pmatrix} \begin{pmatrix} \hat{p}^k \\ \hat{y}^k \end{pmatrix} = \begin{pmatrix} B(\lambda^k + \tilde{\mu}^k - \alpha y_c) + \hat{r}_1^k \\ B y_d + \hat{r}_2^k \end{pmatrix}, \tag{4.20}$$

and $\hat{\delta}_p^k = \hat{r}_1^k / \alpha + A B^{-1} \hat{r}_2^k$, thus in the numerical implementation we can require

$$\|\hat{r}_1^k\|_2 + \|\hat{r}_2^k\|_2 < \frac{\epsilon_k}{\max\{1/\alpha, \|A\|_2 \|B^{-1}\|_2\}}, \tag{4.21}$$

to guarantee the error vector $\|\hat{\delta}_p^k\|_2 \leq \epsilon_k$.

An efficient predictor for the block $\tilde{p}$ subproblem

In Step 1 of Algorithm 3, actually we do not need to solve the p block subproblem twice in the majority situation. In practice, to improve the efficiency of our sGS-imABCD algorithm, we design an efficient predictor for the block $\tilde{p}$ subproblem to check whether to solve it.

Obviously, to solve the block $\tilde{p}$ subproblem, we only need to replace λ^k by $\tilde{\lambda}^k$ in the right-hand term of (4.17). Then we have

$$\begin{pmatrix} B & -\alpha A \\ A^* & B \end{pmatrix} \begin{pmatrix} \tilde{p}^k \\ \tilde{y}^k \end{pmatrix} = \begin{pmatrix} B(\tilde{\lambda}^k + \tilde{\mu}^k - \alpha y_c) \\ B y_d \end{pmatrix}. \tag{4.22}$$

Hence, all the numerical techniques for the block $\hat{p}$ is also applicable for the block $\tilde{p}$.

We can often avoid solving the linear system if $\hat{p}^k$ is already sufficiently close to $\tilde{p}^k$. More specifically, if we employ $\hat{p}^k$ to approximate $\tilde{p}^k$, then the residual vector for (4.22) is given by

$$\begin{pmatrix} \tilde{r}_1^k \\ \tilde{r}_2^k \end{pmatrix} = \begin{pmatrix} B(\tilde{\lambda}^k - \lambda^k) - \hat{r}_1^k \\ -\hat{r}_2^k \end{pmatrix}, \tag{4.23}$$

which means $\tilde{\delta}_p^k = \frac{1}{\alpha}(B(\tilde{\lambda}^k - \lambda^k) - \hat{r}_1^k) - A^* B^{-1} \hat{r}_2^k$.

If the condition

$$\|\tilde{r}_1^k\|_2 + \|\tilde{r}_2^k\|_2 < \frac{\epsilon_k}{\max\{1/\alpha, \|A\|_2 \|B^{-1}\|_2\}}, \tag{4.24}$$

is satisfied, which also guarantees that $\|\tilde{\delta}_p^k\|_2 \leq \epsilon_k$, then we need not solve the linear system (4.22) and take $\tilde{p}^k = \hat{p}^k$.

In order to measure the accuracy of an approximate optimal solution (μ, λ, p) for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$), we can introduce the KKT condition for ($\mathrm{P}_{\mathrm{h},\tau}$) and ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$) as below

$$\begin{cases} 0 = B(y - y_d) + A^* p, \\ 0 = \alpha u - p + \lambda + \mu, \\ 0 = Ay - Bu - By_c, \\ 0 = u - \Pi_{[a,b]}(u + B\mu), \\ 0 = \lambda - \Pi_{[-\beta,\beta]}(\lambda + Bu). \end{cases} \tag{4.25}$$

Thus, let ϵ be a given accuracy tolerance, we terminate our sGS-imABCD method when $\eta < \epsilon$.

The relative residual η is given by

$$\eta = \max\{\eta_1, \eta_2, \eta_3, \eta_4\}, \tag{4.26}$$

where

$$\eta_1 = \frac{\|B(y - y_d) + A^* p\|}{1 + \|By_d\|}, \quad \eta_2 = \frac{\|Ay - Bu - By_c\|}{1 + \|By_c\|},$$
$$\eta_3 = \frac{\|u - \Pi_{[a,b]}(u + B\mu)\|}{1 + \|u\|}, \quad \eta_4 = \frac{\|\lambda - \Pi_{[-\beta,\beta]}(\lambda + Bu)\|}{1 + \|\lambda\|},$$

and $u = (p - \lambda - \mu)/\alpha$.

4.1.2 Convergence results

From Chapter 2, we know that we only get the convergence of optimal value for the method to solve the general problems. However, the convergence of KKT or solution sequence can be obtained for our SPOCPs and decoupled SPOCPs.

The result is based on the convergence of optimal value. Firstly we prove the second order growth condition, then use it to show the convergence of KKT condition and primal solution sequence.

Second order growth condition for bounded sets

Proposition 4.1. *Given any $\bar{x} \in \mathbb{R}^n$, assume that U_1 is a bounded neighborhood of $\bar{x}$, $\bar{v}$ is a bounded vector in $\mathbb{R}^n$, such that $U_1 \subset \eta B_1(0), \bar{v} \in \partial\delta^*_{[a,b]}(\bar{x})$ for some $\eta > 0$, then there exists $\kappa > 0$, the following inequality holds,*

$$\delta^*_{[a,b]}(x) - \delta^*_{[a,b]}(\bar{x}) \geq \langle \bar{v}, x - \bar{x} \rangle + \kappa dist^2(x, \partial\delta^*_{[a,b]}(\bar{v})), \forall x \in U_1, \tag{4.27}$$

where κ depends on η and $\bar{v}$ only.

Proof.

$$\bar{v} \in \partial\delta^*_{[a,b]}(\bar{x}) \Leftrightarrow \bar{x} \in \partial\delta_{[a,b]}(\bar{x}). \tag{4.28}$$

It is easy to compute that

$$\partial\delta_{[a,b]}(\bar{v}_i) = \begin{cases} \{0\}, & \bar{v}_i \in (a,b), \\ \mathbb{R}^-, & \bar{v}_i = a, \\ \mathbb{R}^+, & \bar{v}_i = b. \end{cases} \tag{4.29}$$

(1) If $\bar{v}_i \in (a,b) \Rightarrow \bar{x}_i = 0$.

Case 1: $x_i < 0 \Rightarrow \delta^*_{[a,b]}(x_i) = ax_i$,

The inequality (4.27) is then

$$\langle a, x_i \rangle - 0 \geq \langle \bar{v}_i, x_i - 0 \rangle + \kappa |x_i|^2. \tag{4.30}$$

Since $x \in U_1 \subset CB_1(0)$, we can choose $\kappa = (\bar{v}_i - a)/\eta$.

Case 2: $x_i \geq 0 \Rightarrow \delta^*_{[a,b]}(x_i) = bx_i$.

The inequality (4.27) is then

$$\langle b, x_i\rangle - 0 \geq \langle \bar{v}_i, x_i - 0\rangle + \kappa|x_i|^2. \tag{4.31}$$

Similarly, we choose $\kappa = (b - \bar{v}_i)/\eta$.

(2) If $\bar{v}_i = b \Rightarrow (\partial\delta^*_{[a,b]})^{-1}(\bar{v}_i) = \mathbb{R}^+$.

Case 1: $x_i \geq 0$, then $dist(x_i, \mathbb{R}^+) = 0$, the inequality (4.27) is obvious.

Case 2: $x_i < 0 \Rightarrow \delta^*_{[a,b]}(x_i) = ax_i$,

The inequality (4.27) is then

$$\langle a, x_i\rangle - \langle b, \bar{x}_i\rangle \geq \langle b, x_i - \bar{x}_i\rangle + \kappa|x_i|^2. \tag{4.32}$$

We just need to choose $\kappa = (b - a)/\eta$.

(3) If $\bar{v}_i = a$, similarly, we pick $\kappa = (b - a)/\eta$, then the inequality (4.27) holds.

Hence, if we seek $\kappa = \min\limits_{a<\bar{v}_i<b}\{b - \bar{v}_i, \bar{v}_i - a\}$, we show that the inequality (4.27) is valid for all $x \in U_1 \subset \eta B_1(0)$. □

Proposition 4.2. *Given any $\bar{x} \in \mathbb{R}^n$, assume that $\bar{v}$ is a bounded vector in $\mathbb{R}^n$, such that $\bar{v} \in \partial\delta_{[-\beta,\beta]}(\bar{x})$ for some $\eta > 0$, then there exists $\kappa > 0$, the following inequality holds,*

$$\delta_{[-\beta,\beta]}(x) - \delta_{[-\beta,\beta]}(\bar{x}) \geq \langle \bar{v}, x - \bar{x}\rangle + \kappa dist^2(x, (\partial\delta_{[-\beta,\beta]})^{-1}(\bar{v})), \forall x \in U_2, \tag{4.33}$$

where κ depends on $\beta, \bar{v}$ only.

Proof. It is easy to compute that

$$\partial\delta_{[-\beta,\beta]}(\bar{x}_i) = \begin{cases} \{0\}, & \bar{x}_i \in (-\beta, \beta), \\ \mathbb{R}^-, & \bar{x}_i = -\beta, \\ \mathbb{R}^+, & \bar{x}_i = \beta. \end{cases} \tag{4.34}$$

(1) If $\bar{x}_i \in (-\beta, \beta) \Rightarrow \bar{v}_i = 0 \Rightarrow (\partial\delta_{[-\beta,\beta]})^{-1}(\bar{v}_i) = (-\beta, \beta)$.

So $dist(x, (\partial\delta_{[-\beta,\beta]})^{-1}(\bar{v}_i)) = 0$, the inequality (4.33) holds automatically.

(2) If $\bar{x}_i = \beta \Rightarrow \bar{v}_i \in \mathbb{R}^+ \Rightarrow (\partial\delta_{[-\beta,\beta]})^{-1}(\bar{v}_i) = \{\beta\}$.

The inequality (4.33) is then

$$\langle \bar{v}_i, x_i - \beta\rangle + \kappa|x_i - \beta|^2 \leq 0. \tag{4.35}$$

Since $x_i \in [-\beta, \beta]$, we take $\kappa = \bar{v}_i/(2\beta)$, then

$$\kappa|x_i - \beta|^2 \leq \langle \bar{v}_i, \beta - x_i\rangle. \tag{4.36}$$

(3) If $\bar{x}_i = -\beta$, similarly, we obtain that $\kappa = -\bar{v}_i/(2\beta)$.

Hence we choose $\kappa = \min\limits_{\bar{v}_i \neq 0}\{|\bar{v}_i|/(2\beta)\}$, the inequality (4.33) holds for all $x \in [-\beta, \beta]$. □

Proposition 4.3. *Let $f_1(\mu) = \delta^*_{[a,b]}(B\mu)$, given any $\bar{x} \in \mathbb{R}^n$, assume that U_2 is a bounded neighborhood of $\bar{x}$, $\bar{v}$ is a bounded vector in $\mathbb{R}^n$, such that $U_2 \subset \eta B_1(0), \bar{v}_1 \in \partial f_1(\bar{\mu})$ for some $\eta > 0$, then there exists $\kappa > 0$, the following inequality holds,*

$$\delta^*_{[a,b]}(B\mu) - \delta^*_{[a,b]}(B\bar{\mu}) \geq \langle \bar{v}_1, \mu - \bar{\mu}\rangle + \kappa dist^2(\mu, (\partial f_1)^{-1}(\bar{v}_1)), \forall \mu \in U_2, \tag{4.37}$$

where κ depends on $\eta, \bar{v}_1, \lambda_{\min}(B)$.

Proof. By Proposition 4.1, we obtain

$$\begin{aligned}\delta^*_{[a,b]}(B\mu) - \delta^*_{[a,b]}(B\bar{\mu}) &\geq \langle \bar{v}, B\mu - B\bar{\mu}\rangle + \kappa_1 dist^2(B\mu, (\partial\delta^*_{[a,b]})^{-1}(\bar{v}))\\ &\geq \langle \bar{v}_1, \mu - \bar{\mu}\rangle + \kappa_1\lambda^2_{\min}(B) dist^2(\mu, (\partial f_1)^{-1}(\bar{v}_1)),\end{aligned} \tag{4.38}$$

where $\bar{v}_1 = B\bar{v}, \kappa = \kappa_1\lambda^2_{\min}(B)$, and the last inequality is because that $(\partial\delta^*_{[a,b]})^{-1}(\bar{v}) \subset B(\partial f_1)^{-1}(\bar{v}_1)$ and $\|B\mu - B\bar{\mu}\| \geq \lambda_{\min}(B)\|\mu - \bar{\mu}\|$. □

Theorem 4.2. *Let us denote*

$$\begin{cases} f_1(\mu) = \delta^*_{[a,b]}(B\mu), \\ f_2(\lambda) = \delta_{[-\beta,\beta]}(\lambda), \\ f_3(p) = \dfrac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \langle p, By_c\rangle, \\ g(\mu,\lambda,p) = \dfrac{1}{2\alpha}\|p-\lambda-\mu\|^2_B, \\ \Phi(\mu,\lambda,p) = g(\mu,\lambda,p) + f_1(\mu) + f_2(\lambda) + f_3(p), \end{cases} \tag{4.39}$$

then we have the second order growth condition for Φ*,*

$$\Phi(\mu,\lambda,p) - \Phi(\bar{\mu},\bar{\lambda},\bar{p}) \geq \kappa dist^2((\mu,\lambda,p),(\partial\Phi)^{-1}(0)), \tag{4.40}$$

for all $p \in U_1, \mu \in U_2, \lambda \in [-\beta,\beta]$*, where* $U_1, U_2 \subset \eta B_1(0)$ *are bounded sets,* κ *depends on* $\eta, \beta, \bar{u} = (\bar{p}-\bar{\lambda}-\bar{\mu})/\alpha$ *and the bounded linear subregularity subdifferential sets.*

Proof. By Proposition 4.2, 4.3 and the strong convexity of f_3, we have

$$\begin{aligned} \Phi(\mu,\lambda,p) - \Phi(\bar{\mu},\bar{\lambda},\bar{p}) =& g(\mu,\lambda,p) - g(\bar{\mu},\bar{\lambda},\bar{p}) + f_1(\mu) - f_1(\bar{\mu}) + f_2(\lambda) - f_2(\bar{\lambda}) + f_3(p) \\ & - f_3(\bar{p}) \\ \geq& \langle \nabla g(\bar{\mu},\bar{\lambda},\bar{p}), (\mu-\bar{\mu}, \lambda-\bar{\lambda}, p-\bar{p})\rangle + \kappa_1 dist^2((\mu,\lambda,p), D_1) \\ & + \langle \bar{v}_1, \mu-\bar{\mu}\rangle + \kappa_2 dist^2(\mu, D_2) + \langle \bar{v}_2, \lambda-\bar{\lambda}\rangle + \kappa_3 dist^2(\lambda, D_3) \\ & + \langle \nabla f_3(\bar{p})\rangle + \frac{1}{2}\lambda_{\min}(AB^{-1}A^*)\|p-\bar{p}\|^2 \\ =& \langle \nabla g(\bar{\mu},\bar{\lambda},\bar{p}) + (\bar{v}_1,\bar{v}_2,\bar{v}_3), (\mu-\bar{\mu},\lambda-\bar{\lambda},p-\bar{p})\rangle + \kappa_1 dist^2(\mu, D_1) \\ & + \kappa_2 dist^2(\mu, D_2) + \frac{1}{2}\lambda_{\min}(AB^{-1}A^*)\|p-\bar{p}\|^2, \end{aligned} \tag{4.41}$$

where $\kappa_1 = \frac{1}{2}\lambda^+_{\min}(B)$ is half of the smallest positive eigenvalue of B.

$$\begin{cases} D_1 = \{(\mu,\lambda,p) | B(p-\lambda-\mu)/\alpha = -\bar{v}_1 = \bar{v}_2 = \bar{v}_3\}, \\ D_2 = (\partial f_1)^{-1}(\bar{v}_1), \\ D_3 = (\partial \delta_{[-\beta,\beta]})^{-1}(\bar{v}_2). \end{cases} \tag{4.42}$$

Then we have $-\bar{v}_1 = \bar{v}_2 = \bar{v}_3 = B\bar{u}$. Thus κ_2, κ_3 depend on $\eta, \beta, B\bar{u}, a, b$ only. And we denote

$$D_4 = \{(\mu, \lambda, p) | \mu \in D_2, \lambda \in D_3, AB^{-1}(A^*p - By_d) + By_c = B\bar{u}\}. \tag{4.43}$$

Then from the bounded linear subregularity of the two polyhedral sets $\{D_1, D_4\}$, we have

$$\max\{dist((\mu, \lambda, p), D_1), dist((\mu, \lambda, p), D_4), \} \geq \kappa_4 dist((\mu, \lambda, p), D_1 \cap D_4). \tag{4.44}$$

Therefore, we get

$$\Phi(\mu, \lambda, p) - \Phi(\bar{\mu}, \bar{\lambda}, \bar{p}) \geq \kappa dist^2((\mu, \lambda, p), (\partial\Phi)^{-1}(0)), \forall p \in U_1, \mu \in U_2, \lambda \in [-\beta, \beta], \tag{4.45}$$

where κ is chosen as $\min\{\kappa_1, \kappa_2, \kappa_3, \frac{1}{2}\lambda_{\min}(AB^{-1}A^*)\} * \kappa_4$. □

Convergence results for discretized problem

From Cui's thesis [26, Theorem 3.2], we obtain the iteration complexity of optimal value as below.

Theorem 4.3. *Assume that* $\sum\limits_{i=k}^{\infty} k\epsilon_k < \infty$. *Let* $\{\xi^k\} := \{(p^k, \lambda^k, \mu^k)\}$ *be the sequence generated by the Algorithm 3 to solve Problem* ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$), $\xi^* \in (\partial\Phi)^{-1}(0)$. *Then we have*

$$\Phi(\xi^k) - \inf_{\xi} \Phi(\xi) \leq \frac{2\|\xi^0 - \xi^*\|^2_{\mathcal{H}} + c_0}{(k+1)^2} \tag{4.46}$$

where c_0 *is a constant number,* $\mathcal{H} := \mathrm{Diag}(\mathcal{D}_1, \mathcal{D}_2 + \mathcal{Q}_{22})$, *and* $\Phi(\cdot)$ *is the objective function of the dual problem* ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$).

From the previous section, we have the second order growth condition as follow,

$$\kappa dist^2(\xi^k, (\partial\Phi)^{-1}(0)) \leq \Phi(\xi^k) - \inf_{\xi} \Phi(\xi) = \mathcal{O}(1/k^2). \tag{4.47}$$

Hence we arrive at that

$$dist(\xi^k, (\partial\Phi)^{-1}(0)) = \mathcal{O}(1/k). \tag{4.48}$$

By the definition of $(\partial\Phi)^{-1}(0)$, we know it is the set of points $\xi = (p, \lambda, \mu)$ satisfying the following equations

$$\begin{cases} AB^{-1}A^*p - Ay_d + B(p - \lambda - \mu)/\alpha = 0, \\ B(p - \lambda - \mu)/\alpha \in \partial f_2(\lambda), \\ B(p - \lambda - \mu)/\alpha \in \partial f_1(\mu). \end{cases} \tag{4.49}$$

From equation (4.48) and property of the distance function, we know there exists $\hat{\xi}^k = (\hat{p}^k, \hat{\lambda}^k, \hat{\mu}^k) \in (\partial\Phi)^{-1}(0)$, such that $\|\xi^k - \hat{\xi}^k\| = \mathcal{O}(1/k)$.

Let $r^k = (r_p^k, r_\lambda^k, r_\mu^k) = \hat{\xi}^k - \xi^k$. For the sack of simplicity, we also denote $r_0^k = (r_p^k - r_\lambda^k - r_\mu^k)/\alpha, x^k = (p^k - \lambda^k - \mu^k)/\alpha$, then we have

$$\begin{cases} AB^{-1}A^*(p^k + r_p^k) - Ay_d + B(x^k + r_0^k) = 0, \\ B(x^k + r_0^k) \in \partial f_2(\lambda^k + r_\lambda^k), \\ B(x^k + r_0^k) \in \partial f_1(\mu^k + r_\mu^k). \end{cases} \tag{4.50}$$

From Algorithm 3, we see that $\lambda^k \in \mathrm{Dom} f_2 = [-\beta, \beta], \mu^k \in \mathrm{Dom} f_1 = \mathbb{R}^m$. And from above equation (4.50), we get $\lambda^k + r_\lambda^k \in \mathrm{Dom} f_2, \mu^k + r_\mu^k \in \mathrm{Dom} f_1$

Since f_1, g_1 are closed convex polyhedral functions, it is not hard to obtain the properties as follow.

Proposition 4.4. *If f is closed convex polyhedral function, then f are Lipschitz continuous on its effective domain.*

Proof. From [68, page 172], we know that any polyhedral function f can be expressed in the form

$$f(x) = h(x) + \delta_C(x), \tag{4.51}$$

where

$$\begin{aligned} h(x) &= \max\{\langle x, b_1\rangle - \beta_1, \cdots, \langle x, b_k\rangle - \beta_k\}, && (4.52) \\ C &= \{x | \langle x, b_{k+1}\rangle \le \beta_{k+1}, \cdots, \langle x, b_l\rangle \le \beta_l\}. && (4.53) \end{aligned}$$

It is clear that its domain C is polyhedral convex set, which, therefore, is a locally simplicial set.

By [68, Theorem 10.2], we have f is continuous on its domain C.

And by the definition of h, we see that f is Lipschitz continuous in C with $L = \max\{b_1, \cdots, b_k\}$. □

Now we want to prove our main result, the convergence results as below.

Proposition 4.5. *Let $\{\xi^k\} := \{(p^k, \lambda^k, \mu^k)\}$ be the sequence generated by the Algorithm 3 to solve Problem $(\widehat{\mathrm{D}}_{\mathrm{h},\tau})$, $u^k = (p^k - \lambda^k - \mu^k)/\alpha$, and denote $\tilde{u}^k = \Pi_D(u^k)$, where $D = \mathrm{Dom} f_1 \cap \mathrm{Dom} f_2$, then*

$$F(\tilde{x}^k) - F(x^*) = \mathcal{O}(1/k), \tag{4.54}$$

where u^* is the unique optimal solution of Problem $(\widehat{\mathrm{P}}_{\mathrm{h},\tau})$, and

$$F(u) := \frac{1}{2}\|A^{-1}B(u + y_c) - y_d\|_B^2 + \frac{\alpha}{2}\|u\|_B^2 + \delta_{[a,b]}(u) + \beta\|Bu\|_1.$$

Therefore,

$$\|\tilde{u}^k - u^*\| = \mathcal{O}(1/\sqrt{k}). \tag{4.55}$$

Moreover, the KKT condition of Problem $(\widehat{\mathrm{P}}_{\mathrm{h},\tau})$ has an $\mathcal{O}(1/k)$ iteration complexity.

Proof. From the coercivity of Φ, we obtain that $\{\xi^k\}, \{u^k\}$ are both bounded sequences.

For simplicity, we denote

$$\begin{aligned} \tilde{F}(y, u, v, w) &:= \frac{1}{2}\|A^{-1}B(u + y_c) - y_d\|_B^2 + \frac{\alpha}{2}\|u\|_B^2 + \delta_{[a,b]}(w) + \beta\|Bv\|_1, && (4.56) \\ h(v) &:= \beta\|Bv\|_1. && (4.57) \end{aligned}$$

It is easy to obtain that

$$f_2(\lambda) \;=\; \delta_{[-\beta,\beta]}(\lambda) = h^*(B\lambda). \tag{4.58}$$

As we know

$$-\Phi(\xi^k) = \inf_{y,u,v,w} L(y, u, v, w; \xi^k) = L(\bar{y}^k, \bar{u}^k, \bar{v}^k, \bar{w}^k; \xi^k). \tag{4.59}$$

where $L(y,u,v,w;\xi) := \tilde{F}(y,u,v,w) + \langle p, Ay - Bu\rangle + \langle \lambda, B(u-v)\rangle + \langle \mu, B(u-w)\rangle$ is the Lagrangian function of primal problem ($\widehat{\mathrm{P}}_{\mathrm{h},\tau}$), $(\bar{y}^k, \bar{u}^k, \bar{v}^k, \bar{w}^k)$ satisfying the following conditions,

$$\begin{cases} B(\bar{y}^k - y_d) + A^* p^k = 0, \\ \alpha \bar{u}^k - B(p^k - \lambda^k - \mu^k) = 0, \\ B\lambda^k \in \partial h(\bar{v}^k), \\ B\mu^k \in \partial \delta_{[a,b]}(\bar{w}^k). \end{cases} \tag{4.60}$$

Then

$$\begin{aligned} F(\tilde{x}^k) + \Phi(\xi^k) =& F(\tilde{x}^k) - L(\bar{y}^k, \bar{x}^k, \bar{v}^k, \bar{w}^k; \xi^k) \\ =& (\frac{1}{2}\|A^{-1}B(\tilde{u}^k + y_c) - y_d\|_B^2 - \frac{1}{2}\|\bar{y}^k - y_d\|_B^2 \\ & - \langle p^k, A\bar{y}^k - B(\bar{u}^k + y_c)\rangle) + (\frac{\alpha}{2}\|\tilde{u}^k\|_B^2 - \frac{\alpha}{2}\|\bar{u}^k\|_B^2) \\ & + (h(\tilde{u}^k) - h(\bar{v}^k) - \langle \lambda^k, B(\bar{u}^k - \bar{v}^k)\rangle) \\ & + (\delta_{[a,b]}(\tilde{u}^k) - \delta_{[a,b]}(\bar{w}^k) - \langle \mu^k, B(\bar{u}^k - \bar{w}^k)\rangle) \\ :=& I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{4.61}$$

From equation (4.60), we obtain that

$$\begin{cases} \bar{u}^k = (p^k - \lambda^k - \mu^k)/\alpha = u^k, \\ \bar{y}^k = y_d - B^{-1}A^* p^k), \\ h(\bar{v}^k) - \langle \bar{v}^k, B\lambda^k\rangle = -h^*(B\lambda^k), \\ \delta_{[a,b]}(\bar{w}^k) - \langle \bar{w}^k, B\mu^k\rangle = -\delta^*_{[a,b]}(B\mu^k). \end{cases} \tag{4.62}$$

From the choice of $\xi^k, \tilde{x}^k$ and equation (4.50), we get the following conditions

$$\begin{cases} AB^{-1}A^*(p^k + r_p^k) - Ay_d + B(u^k + r_0^k) = 0, \\ h(u^k + r_0^k) - \langle u^k + r_0^k, B(\lambda^k + r_\lambda^k)\rangle = -h^*(B(\lambda^k + r_\lambda^k)) = f_2(\lambda^k + r_\lambda^k), \\ \delta_{[a,b]}(u^k + r_0^k) - \langle u^k + r_0^k, B(\mu^k + r_\mu^k)\rangle = -\delta^*_{[a,b]}(B(\mu^k + r_\mu^k)). \end{cases} \tag{4.63}$$

We have that $u^k + r_0^k \in D = \mathrm{Dom} f_1 \cap \mathrm{Dom} f_2$, so

$$\|\tilde{u}^k - \bar{u}^k\| = \|\Pi_D(u^k) - u^k\| = dist(u^k, D) \le \|r_0^k\| = \mathcal{O}(1/k). \tag{4.64}$$

Hence

$$\begin{aligned}
I_1 =&\frac{1}{2}\|A^{-1}B(\tilde{u}^k + y_c) - y_d\|_B^2 - \frac{1}{2}\|\bar{y}^k - y_d\|_B^2 - \langle p^k, A\bar{y}^k - B(\bar{u}^k + y_c)\rangle \\
=&\langle BA^{-1}(B(\tilde{u}^k + y_c) - A\bar{y}^k), A^{-1}B(\tilde{u}^k + y_c) + \bar{y}^k - 2y_d\rangle \\
&- \langle p^k, A\bar{y}^k - B(\bar{u}^k + y_c)\rangle \\
\leq& C(\|A\bar{y}^k - B(\tilde{u}^k + y_c)\| + \|A\bar{y}^k - B(\bar{u}^k + y_c)\|) \\
\leq& C(2\|A\bar{y}^k - B(u^k + y_c)\| + \|B(\tilde{u}^k - u^k)\|) \\
=&C(2\|Ay_d - AB^{-1}A^*p^k - B(u^k + y_c)\| + \|B(\tilde{u}^k - u^k)\|) \\
\leq& C(2\|AB^{-1}A^*r_p^k + Br_0^k\| + \|Br_0^k\|) \\
\leq& C(2\|AB^{-1}A^*\|\|r_p^k\| + 3\|B\|\|r_0^k\|) = \mathcal{O}(1/k).
\end{aligned} \tag{4.65}$$

where the constant C is the upper bound of the sequences $\{\xi^k\}_{k\geq 1}, \{x^k\}_{k\geq 1}$.

And

$$I_2 = \frac{\alpha}{2}\|\tilde{u}^k\|_B^2 - \frac{\alpha}{2}\|u^k\|_B^2 \leq \alpha\|B(\tilde{u}^k + u^k)\|\|\tilde{u}^k - u^k\| = \mathcal{O}(1/k). \tag{4.66}$$

Later, we compute

$$\begin{aligned}
I_3 =&h(\tilde{u}^k) - h(\bar{v}^k) - \langle \bar{\lambda}^k, B(\bar{u}^k - \bar{v}^k)\rangle \\
=&(h(\tilde{u}^k) - h(u^k + r_0^k)) + (h(u^k + r_0^k) - \langle B(\lambda^k + r_\lambda^k), u^k + r_0^k\rangle) \\
&- (h(\bar{v}^k) - \langle \bar{v}^k, B\lambda^k\rangle) + (\langle B(\lambda^k + r_\lambda^k), r_0^k\rangle + \langle Br_\lambda^k, u^k\rangle) \\
=&(h(\tilde{u}^k) - h(u^k)) - (h^*(B(\lambda^k + r_\lambda^k)) - h^*(B(\lambda^k))) \\
&+ (\langle B(\lambda^k + r_\lambda^k), r_0^k\rangle + \langle Br_\lambda^k, x^k\rangle) \\
\leq&(L + C)(\|r_0^k\| + \|r_\lambda^k\|) = \mathcal{O}(1/k).
\end{aligned} \tag{4.67}$$

where the last inequality is due to the Lipschitz continuity of h and h^* in their effective domains, with L being the maximum of their Lipschitz constants.

Similarly, we prove $I_4 = \mathcal{O}(1/k)$.

Hence

$$\begin{aligned}
|F(\tilde{u}^k) - \inf_u F(u)| &= |F(\tilde{u}^k) + \inf_\xi \Phi(\xi)| \\
&\leq |F(\tilde{x}^k) + \Phi(\xi^k)| + |\Phi(\xi^k) - \inf_\xi \Phi(\xi)| \\
&= \mathcal{O}(1/k) + \mathcal{O}(1/k^2) = \mathcal{O}(1/k).
\end{aligned} \tag{4.68}$$

Thus, from the strongly convexity of F, we can easily deduce that

$$\|\tilde{x}^k - x^*\| = \mathcal{O}(1/\sqrt{k}). \tag{4.69}$$

Furthermore, we also obtain the convergence of KKT equations of the dual problem.

$$\|AB^{-1}A^*p^k - Ay_d + B(p^k - \lambda^k - \mu^k)/\alpha\| = \|AB^{-1}A^*r_p^k + Br_0^k\| = \mathcal{O}(1/k) \tag{4.70}$$

Therefore, we arrive at the estimate

$$\begin{aligned} \|u^k - Prox_h(u^k + B\lambda^k)\| &= \|u^k + r_0^k - r_0^k - Prox_h(u^k + B\lambda^k)\| \\ &\le \|r_0^k\| + \|Prox_h(u^k + r_0^k + B(\lambda^k + r_\lambda^k)) - Prox_h(u^k + B\lambda^k)\| \\ &\le 2\|r_0^k\| + \|B\|\|r_\lambda^k\| = \mathcal{O}(1/k). \end{aligned} \tag{4.71}$$

Similarly we obtain the inequality

$$\|u^k - \Pi_{[a,b]}(u^k + B\mu^k)\| \le 2\|r_0^k\| + \|B\|\|r_\mu^k\| = \mathcal{O}(1/k). \tag{4.72}$$

□

4.2 Inexact majorized ABCD method for solving decoupled SPOCPs

Moreover, for the sake of comparison of numerical experiments, we also apply the inexact majorized ABCD method to solve the dual problem of ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$) in this section.

Since the proximal mapping of q has a explicit form, we consider only introducing one variable v, such that $B(u - v) = 0$. Thus we rewrite the primal problem as,

$$\left\{ \begin{array}{ll} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u) = \frac{1}{2}\|y - y_d\|_B^2 + \frac{\alpha}{2}\|u\|_B^2 + q(v) \\ \text{s.t.} & Ay = B(u + y_c), \\ & B(u - v) = 0. \end{array} \right. \tag{4.73}$$

Let us denote p, λ to be the Lagrangian multipliers for two equalities constraints, respectively, then we obtain the dual problem, in its equivalent minimization form, as below,

$$\begin{aligned}\min_{\lambda,p\in\mathbb{R}^m} \widetilde{\Phi}(\lambda,p) :=& \frac{1}{2}\|A^*p - By_d\|_{B^{-1}}^2 + \frac{1}{2\alpha}\|\lambda - p\|_B^2 + \langle By_c, p\rangle \\ &+ q^*(B\lambda) - \frac{1}{2}\|y_d\|_B^2.\end{aligned} \tag{$\widetilde{\mathrm{D}}_{h,\tau}$}$$

We present the details of the implementation and then prove the convergence result of the primal and dual variables and the optimality condition in the following subsections. Finally we exploit the uniformly mesh-independence property, which means the number of iteration is independent of the mesh-size when the mesh is fine enough.

4.2.1 Numerical implementation

Now, we can apply Algorithm 1 to ($\widetilde{\mathrm{D}}_{h,\tau}$), with $u = p, v = \lambda$, and

$$\begin{aligned} p_1(u) &= \frac{1}{2}\|A^*p - By_d\|_{B^{-1}}^2 + \langle By_c, p\rangle - \frac{1}{2}\|y_d\|_B^2, \\ p_2(v) &= q^*(B\lambda), \phi(u,v) = \frac{1}{2\alpha}\|\lambda - p\|_B^2. \end{aligned} \tag{4.74}$$

Then we present the algorithm framework as follows.

Algorithm 5: (inexact majorized ABCD algorithm for ($\widetilde{\mathrm{D}}_{h,\tau}$))

Input: $(\lambda^1, p^1) = (\tilde{\lambda}^0, \tilde{p}^0) \in \mathrm{dom} q^* \times \mathbb{R}^{N_h}$. Let $\{\epsilon_k\}$ be a nonincreasing sequence of nonnegative numbers such that $\sum\limits_{k=1}^{\infty} k\epsilon_k < \infty$. Set $k = 1, t_1 = 1$.

Output: $(\tilde{\lambda}^k, \tilde{p}^k)$

Iterate until convergence

Step 1 Choose error tolerance $\hat{\delta}_p^k, \delta_p^k$ such that $\max\{\|\delta_p^k\|, \|\delta_\lambda^k\|\} \leq \epsilon_k$. Compute

$$\begin{aligned} \tilde{p}^k &= \arg\min_p \frac{1}{2}\|A^*p - By_d\|_{B^{-1}}^2 + \frac{1}{2\alpha}\|p - \lambda^k\|_B^2 + \langle By_c, p\rangle - \langle \tilde{\delta}_p^k, p\rangle, \\ &= (\alpha A B^{-1} A^* + B)^{-1}(\alpha A y_d + B(\lambda^k - \alpha y_c) + \alpha \hat{\delta}_p^k), \\ \tilde{\lambda}^k &= \arg\min_\lambda q^*(B\lambda) + \frac{1}{2\alpha}\|\lambda - \tilde{p}^k\|_B^2 + \frac{1}{2}\|\lambda - \lambda^k\|_{\mathcal{D}_\lambda}^2 + \langle \lambda, \delta_\lambda^k\rangle. \end{aligned}$$

Step 2 Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$, Compute

$$p^{k+1} = \tilde{p}^k + \beta_k(\tilde{p}^k - \tilde{p}^{k-1}), \quad \lambda^{k+1} = \tilde{\lambda}^k + \beta_k(\tilde{\lambda}^k - \tilde{\lambda}^{k-1}).$$

Remark 4.2. *In general, we only need* $\sum_{k=1}^\infty k\epsilon_k < \infty$*. In practice, we tend to compute the p-subproblem exact enough, such as let* $\epsilon_k = \min\{\frac{1}{k^3}, 10^{-8}\}$*. When we let* $\epsilon_k = 0$*, we call this method exact ABCD method.*

We now discuss the issue of the choice the proximal term $\mathcal{D}_\lambda$.

It is an important thing to choose a proper proximal term. We hope to add a proximal term as small as possible while expecting the subproblem can be solved efficiently.

Since we have a nonsmooth functional q^* in the subproblem, it is reasonable to choose the proximal term such that the subproblem has an analytic solution. Firstly, we introduce $z = B\lambda$, then solving the λ-subproblem is equivalent to solving the follow systems

$$\begin{cases} \tilde{z}^k = \arg\min_z q^*(z) + \frac{1}{2\alpha}\|z - B\tilde{p}^k\|^2_{B^{-1}} + \frac{1}{2}\|z - B\lambda^k\|^2_{B^{-1}\mathcal{D}_\lambda B^{-1}}, \\ \tilde{\lambda}^k = B^{-1}\tilde{z}^k + \delta_\lambda \end{cases} \tag{4.75}$$

To make sure that the z-subproblem has a explicit solution, we choose

$$\mathcal{D}_\lambda = \frac{1}{\alpha}(\gamma B C^{-1} B - B), \gamma = \begin{cases} 4, \text{ if } n = 2, \\ 5, \text{ if } n = 3. \end{cases} \tag{4.76}$$

Then we have

$$\begin{aligned} \tilde{z}^k =& \arg\min_z q^*(z) + \frac{1}{2\alpha}\|z - B\tilde{p}^k\|^2_{B^{-1}} + \frac{1}{2}\|z - B\lambda^k\|^2_{B^{-1}\mathcal{D}_\lambda B^{-1}} \\ =& \frac{1}{\gamma}C(\tilde{p}^k - \lambda^k) + z^k - \frac{\alpha}{\gamma}C Prox_q^{\frac{\alpha}{\gamma}C}(\frac{1}{\alpha}(\hat{p}^k - \lambda^k) + \frac{\gamma}{\alpha}C^{-1}z^k) \end{aligned} \tag{4.77}$$

where

$$\begin{aligned} (Prox_q^{\frac{\alpha}{\gamma}C}(x)) :=& \Pi_{[a,b]}(soft(x, \frac{\gamma}{\alpha}\beta)), \\ soft(x, c) :=& \max(0, |x| - c) \cdot sign(x). \end{aligned} \tag{4.78}$$

Detail of the implementations of p-subproblem is the same as that in the previous section. And so is the preconditioner issue. We omit the details here.

For the approximate discretized problem, the KKT is given as follow,

$$\begin{cases} 0 = B(y - y_d) + A^* p, \\ 0 = \alpha u - p + \lambda + \mu, \\ 0 = Ay - Bu - By_c, \\ 0 = u - \Pi_{[a,b]}(\alpha^{-1} soft(\alpha u + C^{-1}B(p - \alpha u), \beta)). \end{cases} \tag{4.79}$$

4.2.2 Convergence results

In previous section, we see that ABCD method has the convergence results of both the primal variable and KKT equations when is applied to solve the SPOCPs. For the decoupled SPOCPs, we can also obtain the convergence results of the dual variables, and the uniformly mesh-independence of the method.

Our result is based on the convergence of the optimal value. Firstly we prove convergence of the primal and dual variables and the KKT equations, and then we show that the imABCD method has the uniformly mesh-independence property.

To prove convergence and uniformly mesh-independence, we also need the following property. You are referred to [85, Theorem 1.12] for the detail.

Proposition 4.6. *Let M be a symmetric matrix given by*

$$M = \begin{pmatrix} A & B \\ B^* & C \end{pmatrix}, \tag{4.80}$$

Assume C positive definite, let us denote Schur complement of C by $M/C := A - BC^{-1}B^*$, then

$$\begin{aligned} &(1)\ M \text{ is positive semidefinite } \Leftrightarrow M/C \text{ is positive semidefinite}, \\ &(2)\ M \text{ is positive definite } \Leftrightarrow M/C \text{ is positive definite.} \end{aligned} \tag{4.81}$$

Let us define

$$\psi(p, \lambda) := \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + \frac{1}{2\alpha}\|\lambda - p\|^2_B + \langle By_c, p\rangle. \tag{4.82}$$

We obtain its Hessian

$$\begin{pmatrix} AB^{-1}A^* + \frac{1}{\alpha}B & -\frac{1}{\alpha}B \\ -\frac{1}{\alpha}B & \frac{1}{\alpha}B \end{pmatrix}. \tag{4.83}$$

From the above proposition, we know $\frac{1}{\alpha}B$, which is in the bottom right corner of the matrix, is positive definite. And its Schur complement, $AB^{-1}A^*$, is also positive definite. Thus the Hessian is positive definite.

Since both q^* is a convex function with respect to (p, λ), and $\tilde{\Phi}(p, \lambda) = \psi(p, \lambda) + q^*(\lambda)$, the objective function $\tilde{\Phi}$ in the dual problem is also strongly convex with modulus at least the smallest eigenvalue of $\nabla^2\psi$. Combining it with the convergence of objective function value, we can obtain the convergence of dual variables, the KKT conditions and the primal variable $u := (p - \lambda)/\alpha$. The details are provided in the following theorem.

Theorem 4.4. *Assume that $u^*_{h,\tau}$ is the optimal solution of Problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$), $\xi^*_{h,\tau} := (p^*_{h,\tau}, \lambda^*_{h,\tau})$ is the optimal solution of its dual problem. Let $\{\xi^k_{h,\tau}\} := \{(p^k_{h,\tau}, \lambda^k_{h,\tau})\}$ be the sequence generated by the exact majorized ABCD method, that is $\epsilon_k = 0$ for all k, then there exists a constant c_2, independent of the mesh-size h, τ and the regularization parameter α, such that*

$$\begin{aligned} \|\xi^*_h - \xi^k_h\| &\leq c_2/\alpha k, \\ \|u^*_h - u^k_h\| &\leq c_2/\alpha^2 k. \end{aligned} \tag{4.84}$$

Proof. From the convergence of the optimal value, we have

$$\tilde{\Phi}(\xi^k_h) - \tilde{\Phi}(\xi^*_h) \leq \frac{2\|\xi^0_h - \xi^*_h\|^2_{\mathcal{H}}}{(k+1)^2} \leq \frac{Ch^2}{\alpha(k+1)^2}, \tag{4.85}$$

where

$$\mathcal{H} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\alpha}B + \frac{\gamma}{\alpha}(BC^{-1}B - B) \end{pmatrix}. \tag{4.86}$$

As we analysis before the theorem, $\tilde{\Phi}$ is strongly convex, then we can obtain

$$\|\xi^k_h - \xi^*_h\| \leq \frac{\sqrt{C/c}}{\alpha(k+1)}, \tag{4.87}$$

where ch^2 is the strongly convexity modulus for $\tilde{\Phi}$.

Hence it sufficient to prove the ψ strongly convexity modulus ch^2, with C independent of the meshsize h, or equivalently, the hessian $\nabla^2\psi$ is positive definite with the modulus $c\lambda_{min}(B)$, where c is independent of the mesh-size.

Therefore, it is sufficient to prove that $M - cDiag(B, B)$ is positive semidefinite. By the previous proposition, we need to prove that there exists $c > 0$ such that both $(\frac{1}{\alpha} - c)B \succ 0$ and $AB^{-1}A^* + (\frac{1}{\alpha} - c - \frac{1}{\alpha - \alpha^2 c})B \succeq 0$.

Since we have $\lambda_{\min}(AB^{-1}A^*) = \mathcal{O}(h^2), \lambda_{\max}(B) = \mathcal{O}(h^2)$. There exists $c_1 > 0$, such that $AB^{-1}A^* \succ c_1 B$.

Thus, we require that c satisfies the following inequality

$$\begin{cases} 0 < c < \dfrac{1}{\alpha}, \\ c - \dfrac{1}{\alpha} + \dfrac{1}{\alpha - \alpha^2 c} \leq c_1. \end{cases} \tag{4.88}$$

Let $x = 1 - \alpha c \in (0,1)$, then the inequality system is equivalent to prove the existence of $x \in (0,1)$, such that $f(x) := x^2 + \alpha c_1 x - 1 \geq 0$. It is not hard to conclude that, when $x \in [\frac{-\alpha c_1 + \sqrt{\alpha^2 c_1^2 + 1}}{2\alpha c_1}, 1)$, $f(x) \geq 0$. We hope to find as large c as possible, so we set

$$c = (1 - x_{\min})/\alpha = \frac{2c_1}{2 + \alpha c_1 + \sqrt{\alpha^2 c_1^2 + 4}}. \tag{4.89}$$

And because $u^* = (p^* - \lambda^* - \mu^*)/\alpha$ and $u^*_{h,\tau} = (p^*_{h,\tau} - \lambda^*_{h,\tau} - \mu^*_{h,\tau})/\alpha$, we can obtain the second inequality easily. □

From [26], we know c_0 is just the sum of $i\epsilon_i$. In theory, we require $\sum\limits_{i=1}^{N_h} i\epsilon_i < \infty$. However, in practice, we often choose ϵ_i very small, like 10^{-10}. If we specifically choose $\epsilon_k = h^2/k^3$, we can obtain the result below.

Corollary 4.1. *Assume $u^*_{h,\tau}$ be the optimal solution of the problem ($\widehat{\mathrm{P}}'_{\mathrm{h},\tau}$), $\xi^*_{h,\tau} := (p^*_{h,\tau}, \lambda^*_{h,\tau})$ be the optimal solution of its dual problem. Let $\{\xi^k_{h,\tau}\} := \{(p^k_{h,\tau}, \lambda^k_{h,\tau})\}$ be the sequence generated by the imABCD method, with $\epsilon_k = h^2/k^3$ for all k, then there exists a constant c_3, independent of the mesh-size h, τ and regularization parameter α, such that*

$$\begin{aligned} \|\xi^*_h - \xi^k_h\| &\leq c_3/\alpha k, \\ \|u^*_h - u^k_h\| &\leq c_3/\alpha^2 k. \end{aligned} \tag{4.90}$$

Remark 4.3. *From the corollary, we see that the convergence rate of the majorized ABCD method can be affected by the parameter α. And we reconfirm it in the numerical experiment part.*

Chapter 5

Semismooth Newton augmented Lagrangian method

From previous chapter, we know convergence rate of imABCD method would be affected by the L^2 regularization parameter α. Hence when α is very small or is zero, it is necessary to consider other methods.

Impressed by the linear convergence rate of augmented Lagrangian method(ALM), and motivated by the recently published paper [53], we see that the SSNAL method would be a suitable method for choice. For the outer iteration, the SSNAL method uses the augmented Lagrangian method, which has a better and better linear convergence rate. For the inner iteration, we solve the ALM subproblem with semismooth Newton method, which is very efficient when the initial point is good enough. Furthermore, the SSNAL method makes good use of the sparsity structure of our problem and thus it can reduce dimension of the linear equation for the ALM subproblem.

In this chapter, we present the details of the implementation of the SSNAL method. And we proposed the convergence theory of the primal and dual variables. Then we give the uniform mesh-independence property of the SSNAL method. Moreover, we also obtain the robustness of our method to the parameter α, given a proper initial penalty parameter σ.

For the case α equals to zero, the SSNAL method still solve the problem. And we also present the convergence theory.

5.1 The SSNAL method for decoupled SPOCPs

In this section, we apply the SSNAL method to solve the approximate discretized problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$).

We introduce one additional variable v, such that $B(u-v)=0$ and we rewrite problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$) as

$$\left\{\begin{array}{ll} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u)=\frac{1}{2}\|y-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+q(v) \\ \text{s.t.} & Ay=B(u+y_c), \\ & B(u-v)=0. \end{array}\right. \tag{5.1}$$

Denote p,λ to be the Lagrangian multipliers for two equalities constraints, respectively, we obtain the dual problem, in its equivalent minimization form, as below,

$$\min_{\lambda,p,z\in\mathbb{R}^m} \theta(p,\lambda,z) := \frac{1}{2}\|A^*p-By_d\|_{B^{-1}}^2+\frac{1}{2\alpha}\|\lambda-p\|_B^2+\langle By_c,p\rangle+q^*(z)-\frac{1}{2}\|y_d\|_B^2$$
$$\text{such that} \quad C^{-\frac{1}{2}}(B\lambda-z)=0. \tag{$\widetilde{\mathrm{D}}'_{h,\tau}$}$$

Remark 5.1. *Here actually we use the scaling technique to make the equality having the same scale with the objective function, and moreover, when we look at its dual problem, we obtain a better strong convexity modulus bounded below by a constant independent of mesh-size.*

5.1.1 Numerical implementation

The Lagrangian function of Problem ($\widetilde{\mathrm{D}}'_{h,\tau}$) is given as

$$L(p,\lambda,z;\omega)=\frac{1}{2\alpha}\|p-\lambda\|_B^2+\frac{1}{2}\|A^*p-By_d\|_{B^{-1}}^2+q^*(z)+\langle p,By_c\rangle+\langle\omega,C^{-\frac{1}{2}}(B\lambda-z)\rangle,$$

where ω is the Lagrangian multiplier for the equality $C^{-\frac{1}{2}}(B\lambda - z) = 0$. Actually you will find that the optimal control u satisfies $u = C^{-\frac{1}{2}}\omega$.

Applying the SSNAL method, we obtain that,

Algorithm 6: SSNAL method for problem ($\widetilde{\mathrm{D}}'_{h,\tau}$)

Iterate the following steps for $k = 1, 2, \cdots$

Step 1

$$\begin{aligned}
(p^k, \lambda^k) &= \arg\min_{p,\lambda} L_{\sigma_k}(p, \lambda, z(p, \lambda); u^k) + \langle \delta_p, p\rangle + \langle \delta_\lambda, \lambda\rangle && (5.2)\\
&= \arg\min_{p,\lambda} \phi(p, \lambda) + \langle \delta_p^k, p\rangle + \langle \delta_\lambda^k, \lambda\rangle, && (5.3)
\end{aligned}$$

or equivalently to solve the equations (inexactly) for (p^k, λ^k),

$$\begin{cases}
h_1 := \dfrac{1}{\alpha}Bp^k + AB^{-1}(A^*p^k - By_d) - \dfrac{1}{\alpha}B\lambda^k + By_c = 0,\\
h_2 := -\dfrac{1}{\alpha}Bp^k + \dfrac{1}{\alpha}B\lambda^k + BProx_q^{C/\sigma_k}(\sigma_k C^{-1}B\lambda^k + u^k) = 0.
\end{cases} \tag{5.4}$$

Step 2

$$z^k = Prox_{q^*}^{\sigma_k C^{-1}}(B\lambda^k + Cu/\sigma_k) = B\lambda + \frac{1}{\sigma_k}Cu - \frac{1}{\sigma_k}CProx_q^{C/\sigma_k}(\sigma_k C^{-1}B\lambda^k + u).$$

Step 3 Update the multiplier

$$u^{k+1} = u^k + \sigma_k C^{-1}(B\lambda^k - z^k) = Prox_q^{C/\sigma_k}(\sigma_k C^{-1}B\lambda^k + u^k).$$

and the penalty parameter $\sigma_{k+1} = \tau\sigma_k$.

Then check if residual of the KKT condition is less than given tolerance.

Here we define

$$\begin{aligned}
(Prox_q^{C/\sigma}(x)) &:= \Pi_{[a,b]}(soft(x, \sigma\beta)),\\
soft(x, \sigma\beta) &:= sgn(x) \cdot \max\{|x| - \sigma\beta, 0\},\\
\phi(p, \lambda) &:= \min_z L_\sigma(p, \lambda, z; u^k).
\end{aligned}$$

And the KKT equation is given as

$$\left\{\begin{aligned} A^*p + B(y - y_d) = 0,\\ Ay - B(u + y_c) = 0,\\ u - \Pi_{[a,b]}(\alpha^{-1}soft(\alpha u + C^{-1}B(p - \alpha u), \beta)) = 0. \end{aligned}\right. \tag{5.5}$$

5.1.2 Efficient computation of subproblems

For the outer iteration, we know that augmented Lagrangian method has a linear convergence rate. To make the whole algorithm efficient, we solve the inner subproblem (5.4) with the famous semismooth Newton method.

Algorithm 7: Semismooth Newton method

Iterate the following steps for $j = 0, 1, \cdots$

Step 1 Select an element $V_j \in \partial_B F(x^j)$ and find an approximate solution d^j to

$$F(x^j) + V_j d = 0.$$

Step 2 Let m_j be the smallest nonnegative integer m, such that

$$f(x^j + \rho^m d^j) - f(x^j) \leq \tau \rho^m \langle \nabla f(x^j), d^j \rangle.$$

Step 3 Set $x^{j+1} = x^j + \rho^m d^j$.

For our problem, $x = (p, \lambda), F(x) = (h_1(x)^T, h_2(x)^T)^T$ and $f(x) = \phi(p, \lambda)$. We then obtain the Newton equation as follow

$$\begin{pmatrix} \frac{1}{\alpha}B + AB^{-1}A^* & -\frac{1}{\alpha}B \\ -\frac{1}{\alpha}B & \frac{1}{\alpha}B + \sigma BQB \end{pmatrix} \begin{pmatrix} d_p \\ d_\lambda \end{pmatrix} = \begin{pmatrix} -h_1 \\ -h_2 \end{pmatrix} \tag{5.6}$$

where $Q = C^{-\frac{1}{2}}\hat{Q}, \hat{Q} \in \partial Prox_q^{C/\sigma}(\sigma C^{-\frac{1}{2}}B\lambda + u_1) = \partial\Pi_{[a,b]}(soft(\sigma C^{-\frac{1}{2}}B\lambda + u_1, \sigma\beta))$.

As C and $\hat{Q}$ are both positive semi-definite diagonal matrices, Q is still diagonal, and also positive semi-definite. Hence there exists a low rank matrix $Q_1 \in \mathbb{R}^{m\times r}$, such that $Q = Q_1Q_1^*$. Define $r = rank(Q)$.

There are two cases to consider, $r = 0$ and $r \neq 0$. To make use of the sparsity structure, we apply the Schur complement technique and the Sherman-Morrison formulas.

If $r = 0$, the Newton equation then degenerates into

$$\begin{pmatrix} \frac{1}{\alpha}B + AB^{-1}A^* & -\frac{1}{\alpha}B \\ -\frac{1}{\alpha}B & \frac{1}{\alpha}B \end{pmatrix} \begin{pmatrix} d_p \\ d_\lambda \end{pmatrix} = \begin{pmatrix} -h_1 \\ -h_2 \end{pmatrix} \tag{5.7}$$

Obviously, it is equivalent to solve these equations in order

$$\begin{cases} Ady = h_1 + h_2, \\ A^* d_p = -Bd_y, \\ d_\lambda = d_p - \alpha H_2. \end{cases} \tag{5.8}$$

If $r \neq 0$, we make use of the Schur complement to simplify the computation.

Firstly we obtain

$$(\frac{1}{\alpha}B + \sigma BQB)^{-1} = \alpha(B^{-1} - \alpha\sigma Q_1 D^{-1} Q_1^*) \tag{5.9}$$

with Q_1 satisfies $Q = Q_1 Q_1^*$ and $D := I_r + \alpha\sigma Q_1^* B Q_1$.

Hence by the Schur complement, one obtains

$$(AB^{-1}A^* + \sigma BQ_1 D^{-1} Q_1^* B)d_p = -h_1 - h_2 + \alpha\sigma BQ_1 D^{-1} Q_1^* h_2 := h_p. \tag{5.10}$$

Then we introduce an additional variable $dy := -B^{-1}A^*dp$, and get

$$\begin{pmatrix} \sigma BQ_1 D^{-1} Q_1^* B/\zeta & -A/\zeta \\ A^* & B \end{pmatrix} \begin{pmatrix} d_p \\ d_y \end{pmatrix} = \begin{pmatrix} -h_p/\zeta \\ 0 \end{pmatrix} \tag{5.11}$$

where $h_p := -h_1 - h_2 + \alpha\sigma BQ_1 D^{-1} Q_1^* h_2, \zeta = (1 + \alpha\sigma)/\sigma$.

To solve this non-symmetric two blocks system efficiently, we need to find a proper preconditioner. We approximate the term $\sigma B^* Q_1 D^{-1} Q_1^* B/\zeta$ by B, then use the preconditioner proposed in [3].

If we introduce one more variable $d_x = D^{-1} Q_1^* B d_p$, and then obtain a three block linear equation

$$\begin{pmatrix} B & 0 & A^* \\ 0 & \sigma D & -\sigma(BQ_1)^* \\ A & -\sigma BQ_1 & 0 \end{pmatrix} \begin{pmatrix} d_y \\ d_x \\ d_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ h_p \end{pmatrix}. \tag{5.12}$$

For this system, we choose the preconditioner to be $Diag(B, \sigma D, AB^{-1}A^*)$.

Later we calculate d_λ by $d_\lambda = d_p - \alpha H_2 - \alpha\sigma Q_1 D^{-1} Q_1^*(Bd_p - \alpha h_2)$.

Finally we use the line search to get proper step length.

We summarize the whole process as the following algorithm.

Algorithm 8: SSNCG for subproblem (5.4)

Iterate the following steps for $j = 0, 1, \cdots$.

Step 1. Solve the equation for d_p^j,

$$\begin{pmatrix} B & 0 & A^* \\ 0 & \sigma D & -\sigma(BQ_1)^* \\ A & -\sigma BQ_1 & 0 \end{pmatrix} \begin{pmatrix} d_y^j \\ d_x^j \\ d_p^j \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ h_p \end{pmatrix}.$$

and compute d_λ^j by

$$d_\lambda^j = d_p^j - \alpha H_2 - \alpha\sigma Q_1 D^{-1} Q_1^*(Bd_p^j - \alpha h_2).$$

Step 2. Line search: Let m_k is the first nonnegative integer m, such that

$$\phi(p^j + \rho^m d_p^j, \lambda^j + \rho^m d_\lambda^j) \le \phi(p^j, \lambda^j) + \tau\rho^m \langle \nabla\phi(p^j, \lambda^j), (d_p^j, d_\lambda^j)\rangle$$

Then set $(p^{j+1}, \lambda^{j+1}) = (p^j, \lambda^j) + \rho^{m_k}(d_p^j, d_\lambda^j)$;

Check stopping criteria (B):

$$dist^2(0, \partial\phi(p^{j+1}, \lambda^{j+1})) \le \min\{\frac{b}{\sigma_k}\delta_k^2, \frac{\delta_k'^2}{\sigma_k^2}\}\|u^{k+1} - u^k\|_{C^{-1}}^2.$$

If satisfied, stop and let $(p^k, \lambda^k) = (p^{j+1}, \lambda^{j+1})$. Otherwise go to Step 1.

where $\nabla\phi(p, \lambda) = (h_1, h_2)$, with h_1, h_2 given in (5.4).

5.2 Convergence and uniformly mesh-independence

In this section, we will prove some convergence results for SSNAL method, including the convergence of primal vairables u, the uniformly mesh-independence and the robustness to the parameter α.

Similar to Section 2.4, we introduce some notations for Problem ($\widetilde{\mathrm{D}}'_{h,\tau}$).

Let $x := (p, \lambda, z)$, and define

$$l(x, \omega) := L(p, \lambda, z; \omega), f(x) := \sup_{u} l(x; \omega) g(\omega) := \inf_{x} l(x, \omega)$$

$$T_f = \partial f, T_g = -\partial g, T_l(x_1, x_2) = \{(u_1, u_2) | (u_1, -u_2) \in \partial l(x_1, x_2)\}$$

and

$$\begin{aligned} T_f^{-1}(v) &:= \arg\min_{x \in \mathbb{R}^n} \{f(x) - x \cdot v\}, \\ T_g^{-1}(\mu) &:= \arg\max_{\omega \in \mathbb{R}^n} \{g(\omega) + \omega \cdot \mu\}, \\ T_l^{-1}(v, \mu) &:= \arg\min_{x \in \mathbb{R}^n} \max_{\omega \in \mathbb{R}^m} \{l(x, \omega) - x \cdot v + \omega \cdot \mu\}, \end{aligned}$$

According to Theorem 2.3, we only need to prove the metric subregularity of the multi-functions T_g and T_l.

The following property gives us an approach to prove the subregularity of T_l.

Proposition 5.1. [62, Theorem 7.5]. *Let $\bar{\omega}$ be a feasible solution to the dual problem ($\widetilde{\mathrm{D}}'_{h,\tau}$), and q be a convex piece-wise linear function. Then T_l is metric subregular at $((0, 0), (\bar{x}, \bar{v}))$ and $\bar{x}$ is a local optimal solution to problem ($\tilde{\mathrm{P}}_{\mathrm{h},\tau}$) if and only if the following conditions are valid:*

(i) *the collection of Lagrange multipliers for problem ($\widetilde{\mathrm{D}}'_{h,\tau}$) at $\bar{x}$ given by*

$$\Lambda_{com}(\bar{\omega}) := \{v \in \mathbb{R}^n | \nabla_x l(\bar{x}, v) = 0, v \in \partial q^*(\bar{z})\} \text{ is a singleton }, \{\bar{v}\}. \tag{5.13}$$

(ii) *second order sufficient optimality condition* (SOSC) *holds, i.e.*

$$\langle \nabla_{xx} l(\bar{x}, \bar{\omega}) y, y \rangle > 0, \text{ for all } 0 \neq y \in \mathbb{R}^m \text{ with } By \in \mathcal{K}(\bar{z}, \bar{v}). \tag{5.14}$$

where $\bar{z} = B\bar{x}, \mathcal{K}(\bar{z}, \bar{v}) := \{w \in T(\bar{z}, \mathrm{Dom} q^*) | \langle \bar{v}, w \rangle = dq^*(\bar{z})(w)\}$ is the critical cone for q^* at $(\bar{z}, \bar{v})$ with $dq^*(\bar{z})(w) := \liminf_{u \to w, t \downarrow 0} (q^*(\bar{z} + tu) - q^*(\bar{z}))/t$.

Under this proposition, we can prove the metric subregularity of T_g, T_l for our problem.

Proposition 5.2. *For Problem* ($\widetilde{\mathrm{D}}'_{h,\tau}$)*,* T_g *is metrically subregular at* $\bar{\omega}$ *for origin, and* T_l *is is metrically subregular at* $(\bar{x}, \bar{\omega})$ *for origin.*

Proof. We see that

$$\begin{aligned} T_g(\omega) &= -\partial g(\omega) \\ &= BA^{*-1}B(A^{-1}B(C^{-\frac{1}{2}}\omega + y_c) - y_d) + \alpha C^{-\frac{1}{2}}BC^{-\frac{1}{2}}\omega + C^{-\frac{1}{2}}\partial q(C^{-\frac{1}{2}}\omega) \end{aligned} \tag{5.15}$$

It is easy to verify that T_g is strongly monotone with modulus $\alpha/4$. Hence T_g is metric subregular at x for y, for all $(x, y) \in gph(T_g)$.

To prove T_l is metric subregular, we only need to check the two conditions in Proposition 5.1.

By Proposition 4.6, we can obtain that

$$\nabla^2_{xx} l(\bar{x}, \bar{\omega}) = \begin{pmatrix} \frac{1}{\alpha}B + AB^{-1}A^* & -\frac{1}{\alpha}B \\ -\frac{1}{\alpha}B & \frac{1}{\alpha}B \end{pmatrix} \succ 0$$

Thus (5.14) is then easy to verify.

For any feasible $x = (p, \lambda, z)$, we can write down the $\Lambda_{com}(x)$ explicitly as a set of point ω satisfying the following conditions

$$\begin{cases} \dfrac{1}{\alpha}B(p - \lambda) + AB^{-1}(A^*p - By_d) + By_c = 0, \\ \dfrac{1}{\alpha}B(\lambda - p) + BC^{-\frac{1}{2}}\omega = 0, \\ 0 \in \partial q^*(z) - C^{-\frac{1}{2}}\omega, \\ C^{-\frac{1}{2}}(B\lambda - z) = 0. \end{cases}$$

Since the condition is also the KKT condition for problem ($\widetilde{\mathrm{D}}'_{h,\tau}$), any point ω satisfying the condition above should be the optimal solution to the dual problem of ($\widetilde{\mathrm{D}}'_{h,\tau}$), that is the optimal solution for the problem

$$\min_{\omega \in \mathbb{R}^m} \quad \tilde{J}_h(\omega) = \frac{1}{2}\|AB^{-1}C^{-\frac{1}{2}}\omega - y_d\|^2_B + \frac{\alpha}{2}\|C^{-\frac{1}{2}}\omega\|^2_B + q(C^{-\frac{1}{2}}\omega). \tag{5.16}$$

Due to the strongly convexity of the objective function, the optimal solution should be unique. Thus (5.13) is also valid. Therefore, by Proposition 5.1, we conclude that T_l is is metrically subregular at $(\bar{z}, \bar{x})$ for origin. □

Then we can obtain the convergence result similar to Theorem 2.3.

Theorem 5.1. [28, Theorem 4.1] *Suppose optimal solution set $T_g^{-1}(0)$ to Problem (5.16) is nonempty. Let $\{(p^k, \lambda^k, z^k, u^k)\}$ be a sequence generated by the Algorithm 9,* $x^k := (p^k, \lambda^k, z^k), \omega^k := C^{1/2}u^k$. Then

(a) If T_g is metrically subregular at $\bar{\omega}$ for the origin with modulus κ_g, then there exists $k \geq 0$ such that for all $k \geq \bar{k}$,

$$dist(\omega^{k+1}, T_g^{-1}(0)) \leq \theta_k dist(\omega^k, T_g^{-1}(0)), \tag{5.17}$$

where

$$1 > \theta_k = \left(\kappa_g/\sqrt{\kappa_g^2 + \sigma_k^2 + 2\delta_k}\right)(1-\delta_k)^{-1} \to \theta_\infty = \kappa_g/\sqrt{\kappa_g^2 + \sigma_k^2}$$

(b) If in addition to the metric subregularity of T_g at $\bar{\omega}$ for the origin, one has $T_f^{-1}(0)$ is non-empty and bounded and the following condition on T_l: there exist two constants $\kappa_l \geq 0$ and $\epsilon > 0$, any (x, ω) satisfying $dist((x, \omega), T_f^{-1} \times \{\bar{\omega}\}) \leq \epsilon$,

$$dist((x, \omega), T_l^{-1}(0)) \leq \kappa_l dist(0, T_l(x, \omega)). \tag{5.18}$$

Then there exists $\tilde{k} > 0$ such that for all $k \geq \tilde{k}, \delta_k < 1$, and

$$dist(x^{k+1}, T_f^{-1}(0)) \leq \theta'_k dist(\omega^k, T_g^{-1}(0)), \tag{5.19}$$

where $\theta'_k = \kappa_l \sigma_k^{-1}(1 + \delta'_k)(1 - \delta'_k) \to \theta'_\infty = \kappa_l/\sigma_\infty(\theta'_\infty = 0,\ \text{if}\ \sigma_\infty = \infty)$.

Here, we provide a more accurate convergence rate for our specific model problem.

Proposition 5.3 (New convergence rate)**.** *Suppose $\{u^k\}$ is the sequence generated by the Algorithm 9,* $\omega^k := C^{1/2}u^k$, then

$$\|\omega^{k+1} - \bar{\omega}\| \leq \frac{(1 + \alpha\sigma_k/4)^{-1} + \delta_k}{1 - \delta_k}\|\omega^k - \bar{\omega}\|. \tag{5.20}$$

Proof. By Remark 2.3, we know the sequence $\{\omega^k\}$ can be regarded as generated from the inexact PPA for the problem below

$$P_k(\omega) := (I + \sigma_k T_g)^{-1}(\omega) = \arg\max_{\omega\in\mathbb{R}^m}\{g(\omega) - (1/2\sigma_k)|\omega - \omega^k|^2\}. \tag{5.21}$$

We have $\|\omega^{k+1} - P_k(\omega^k)\| \leq \delta_k\|\omega^{k+1} - \omega^k\|$.

For simplicity, we define $y^k = \omega^{k+1} - P_k(\omega^k)$, then we have $\|y^k\| \leq \delta_k\|\omega^{k+1} - \omega^k\|$ and $\omega^{k+1} - y^k = P_k(\omega^k)$.

And since $P_k = (I + \sigma_k T_g)^{-1}$, we obtain $(\omega^k - \omega^{k+1} + y^k)/\sigma_k \in T_g(\omega^{k+1} - y^k)$.

From the proof of Proposition 5.2, we see T_g is strongly monotone with modulus $\alpha/4$. By the definition of strong monotone and that $0 \in T_g(\bar{\omega})$ we have

$$\langle(\omega^k - \omega^{k+1} + y^k)/\sigma_k, \omega^{k+1} - y^k - \bar{\omega}\rangle \geq \frac{\alpha}{4}\|\omega^{k+1} - y^k - \bar{\omega}\|^2,$$

then

$$\langle(\omega^k - \bar{\omega}) - (\omega^{k+1} - y^k - \bar{\omega}), \omega^{k+1} - y^k - \bar{\omega}\rangle \geq \frac{\alpha\sigma_k}{4}\|\omega^{k+1} - y^k - \bar{\omega}\|^2.$$

So we get

$$\begin{aligned}(1 + \alpha\sigma_k/4)\|\omega^{k+1} - y^k - \bar{\omega}\|^2 &\leq \langle(\omega^k - \bar{\omega}), \omega^{k+1} - y^k - \bar{\omega}\rangle \\ &\leq \|\omega^k - \bar{\omega}\|\|\omega^{k+1} - y^k - \bar{\omega}\|.\end{aligned}$$

Let both sides divided by $\|\omega^{k+1} - y^k - \bar{\omega}\|$, and we obtain

$$\|\omega^{k+1} - y^k - \bar{\omega}\| \leq (1 + \alpha\sigma_k/4)^{-1}\|\omega^k - \bar{\omega}\|,$$

Hence we arrive at

$$\begin{aligned}\|\omega^{k+1} - \bar{\omega}\| &\leq (1 + \alpha\sigma_k/4)^{-1}\|\omega^k - \bar{\omega}\| + \|y^k\| \\ &\leq (1 + \alpha\sigma_k/4)^{-1}\|\omega^k - \bar{\omega}\| + \delta_k\|\omega^{k+1} - \omega^k\| \\ &\leq (1 + \alpha\sigma_k/4)^{-1}\|\omega^k - \bar{\omega}\| + \delta_k(\|\omega^{k+1} - \bar{\omega}\| + \|\omega^k - \bar{\omega}\|)\end{aligned}$$

As a result, we obtain

$$\|\omega^{k+1} - \bar{\omega}\| \leq \frac{1 + (1 + \alpha\sigma_k/4)\delta_k}{(1 + \alpha\sigma_k/4)(1 - \delta_k)}\|\omega^k - \bar{\omega}\|.$$

□

Remark 5.2. *It is easy to check that* $\frac{1}{1+\alpha\sigma/4} < \frac{1}{\sqrt{1+(\alpha\sigma/4)^2}}$. *When* $\alpha\sigma$ *is large, like* 4 *or* 40, *the difference is small. Otherwise, when* $\alpha\sigma$ *is small, like* 0.4 *or* 0.04, *the convergence rate can be a big difference. And in practice,* α *is often chosen very small. Hence it is necessary for the new estimate of convergence rate.*

It should be noticed that, although α *is often chosen very small, we will pick initial* σ *be very large, like* $0.1/\alpha$, *hence* $\alpha\sigma \to 0$ *will not happen. Thus SSNAL method will enjoy a fast linear convergence rate.*

Now we provide the convergence for the optimal control u and uniformly mesh-independence of the SSNAL method.

Theorem 5.2. *Let* $\bar{u}_{h,\tau}$ *be the optimal solution to* $(\widetilde{\mathrm{D}}'_{h,\tau})$, $\{u^k_{h,\tau}\}_{k\geq 0}$ *be sequences generated from continuous and discretized augmented Lagrangian method, then for all mesh size* $h, \tau > 0$, *we have*

$$\|C^{\frac{1}{2}}(u^{k+1}_{h,\tau} - \bar{u}_{h,\tau})\| \leq \frac{(1+\alpha\sigma_k/4)^{-1} + \delta_k}{1-\delta_k} \|C^{\frac{1}{2}}(u^k_{h,\tau} - \bar{u}_{h,\tau})\|. \tag{5.22}$$

Therefore, we have

$$\|u^{k+1}_{h,\tau} - \bar{u}_{h,\tau}\| \leq C_0 \rho^{k+1} \|u^0_{h,\tau} - \bar{u}_{h,\tau}\|,$$

where $\rho = ((1+\alpha\sigma_0/4)^{-1} + \delta_0)/(1-\delta_0), C_0 = c_2/c_1$, c_1, c_2 *are the same as that in Proposition 3.11.*

Proof. Let $w^k_{h,\tau} = C^{1/2} u^k_{h,\tau}, \bar{w}_{h,\tau} = C^{1/2}\bar{u}_{h,\tau}$. Then from the previous proposition, we can derive that

$$\|\omega^{k+1}_{h,\tau} - \bar{\omega}_{h,\tau}\| \leq \frac{(1+\alpha\sigma_k/4)^{-1} + \delta_k}{1-\delta_k} \|\omega^k_{h,\tau} - \bar{\omega}_{h,\tau}\|. \tag{5.23}$$

Therefore

$$\|C^{\frac{1}{2}}(u^{k+1}_{h,\tau} - \bar{u}_{h,\tau})\| \leq \frac{(1+\alpha\sigma_k/4)^{-1} + \delta_k}{1-\delta_k} \|C^{\frac{1}{2}}(u^k_{h,\tau} - \bar{u}_{h,\tau})\|. \tag{5.24}$$

Remember that $C = Diag(W_h, \cdots, W_h) = 2Diag(M_h, \cdots, M_h)$. By Proposition 3.11, which states bounds of the eigenvalues for the stiffness matrix and the mass matrix, we have $2c_1 \leq \lambda_{\min}(C) \leq \lambda_{max}(C) \leq 2c_2$.

Hence

$$\|u_{h,\tau}^{k+1} - \bar{u}_{h,\tau}\| \leq C_0\rho^{k+1}\|u_{h,\tau}^0 - \bar{u}_{h,\tau}\|, \tag{5.25}$$

where $\rho = ((1+\alpha\sigma_0/4)^{-1} + \delta_0)/(1-\delta_0), C_0 = c_2/c_1$. □

Remark 5.3 (**Robustness**). *If we choose $\sigma_0 = 1/\alpha$, we obtain the convergence rate $\rho = ((1.25)^{-1} + \delta_0)/(1-\delta_0)$, independent of the parameter α. Hence, we can conclude that the SSNAL method is robust to the parameter α.*

In fact, $\tilde{J}_h$ is a strongly convex function with the modulus $b = \alpha/4 + b_0 > \alpha/4$, and if again we choose the $\sigma_0 = 1/\alpha$, we have the convergence rate $\rho = ((1+\alpha\sigma_0/4)^{-1} + \delta_0)/(1-\delta_0) = ((1.25+b_0/(4\alpha))^{-1}+\delta_0)/(1-\delta_0)$. Hence when α decreases, we obtain a better convergence rate, which means the SSNAL method can deal with problems with a small α better than that with a large one.

5.3 The SSNAL method for decoupled SPOCPs with α being zero

Now we consider the model problem when $\alpha = 0$, that is

$$\begin{cases} \min\limits_{y\in Y, u\in U_{ad}} & J(y,u) = \frac{1}{2}\|y-y_d\|_{L^2(\Omega_T)}^2 + \beta\|u\|_{L^1(\Omega_T)} \\ \text{s.t.} & \mathcal{A}y = \mathcal{B}(u+y_c), \\ & u \in U_{ad} = [a,b] \end{cases} \tag{5.26}$$

We discretize the problem using the approximation discretization of L^1-norm, and obtain the discretize problem as

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} & J(y,u) = \frac{1}{2}\|y-y_d\|_B^2 + q(u) \\ \text{s.t.} & Ay = B(u+y_c). \end{cases} \tag{5.27}$$

We denote p as the Lagrangian multiplier for equality $Ay - B(u+y_c) = 0$, then derive the dual problem as

$$\begin{cases} \min\limits_{p,z} & \theta(p,z) = \frac{1}{2}\|A^*p - By_d\|_{B^{-1}}^2 + q^*(z) + \langle p, By_c\rangle \\ \text{s.t.} & C^{-\frac{1}{2}}(Bp - z) = 0. \end{cases} \tag{5.28}$$

5.3.1 Numerical implementation

In this subsection, we present the numerical implementation of SSNAL method.

The Lagrangian function of our problem is given as follow

$$L(p, z; \omega) = \frac{1}{2}\|A^*p - By_d\|^2_{B^{-1}} + q^*(z) + \langle p, By_c\rangle + \langle \omega, C^{-\frac{1}{2}}(Bp - z)\rangle.$$

where ω is the Lagrangian multiplier for the equality $C^{-\frac{1}{2}}(Bp - z) = 0$, and similar with the previous section, we know $u = C^{-\frac{1}{2}}\omega$.

Applying the SSNAL method, we obtain that

Algorithm 9: SSNAL method for Dual problem (5.28)

Iterate the following steps for $k = 0, 1, \cdots$

Step 1

$$\begin{aligned} p^k &= \arg\min_p L_{\sigma_k}(p, z(p); u^k) + \langle \delta_p, p\rangle \\ &= \arg\min_p \tilde{\phi}(p) + \langle \delta_p, p\rangle, \end{aligned} \tag{5.29}$$

or equivalently to solve the equation (inexactly) for p,

$$\tilde{h}_1 := AB^{-1}(A^*p - By_d) + By_c + BProx_q^{\frac{1}{\sigma_k}C}(\sigma_k C^{-1}Bp + u^k) = 0.$$

Step 2

$$z^k = Prox_{q^*}^{\sigma_k C^{-1}}(Bp^k + Cu^k/\sigma_k) = Bp^k + \frac{1}{\sigma_k}Cu^k - \frac{1}{\sigma_k}CProx_q^{C/\sigma_k}(\sigma_k C^{-1}Bp^k + u^k).$$

Step 3 Update the multiplier

$$u^{k+1} = u^k + \sigma_k C^{-1}(B^k - z^k) = Prox_q^{C/\sigma_k}(\sigma_k C^{-1}Bp^k + u^k),$$

and the penalty parameter $\sigma_{k+1} = \sigma_k * \tau$.

Then check the KKT conditions.

where

$$\begin{aligned} (Prox_q^{C/\sigma}(x)) &:= \Pi_{[a,b]}(soft(x, \sigma\beta)), \\ soft(x, \sigma\beta) &:= sgn(x) \cdot \max\{|x| - \sigma\beta, 0\}, \\ \tilde{\phi}(p) &:= \min_z L_\sigma(p, z(p); u^k). \end{aligned} \tag{5.30}$$

Afterward, to solve for the equations in Step 1, we use the semismooth Newton method to solve the equation below,

$$(AB^{-1}A^{*} + \sigma BQB)d_p = -\tilde{h}_1. \tag{5.31}$$

where

$$\begin{cases} Q = C^{-\frac{1}{2}}\hat{Q}C^{-\frac{1}{2}}\tilde{Q}, \\ \hat{Q} \in \partial\Pi_{[a,b]}(C^{-\frac{1}{2}}soft(\sigma C^{-\frac{1}{2}}Bp + u, \sigma\beta)), \\ \tilde{Q} \in \partial soft(\sigma C^{-\frac{1}{2}}Bp + u, \sigma\beta). \end{cases} \tag{5.32}$$

Since $C, \hat{Q}, \tilde{Q}$ are all positive semi-definite diagonal matrices, Q is also diagonal and positive semi-definite, and there still exists a diagonal $Q_1 \in \mathbb{R}^{m\times r}$, such that $Q = Q_1 Q_1^T$.

Let $r = rank(Q)$. We consider for these two cases, $r = 0$ and $r \neq 0$.

If $r = 0$, the Newton equation is then

$$AB^{-1}A^{*}d_p = -\tilde{h}_1. \tag{5.33}$$

Obviously, it is equivalent to solve these equations in order

$$\begin{cases} Ady = \tilde{h}_1, \\ A^{*}d_p = -Bd_y. \end{cases} \tag{5.34}$$

If $r \neq 0$, one obtains that

$$(AB^{-1}A^{*} + \sigma BQB)d_p = -\tilde{h}_1. \tag{5.35}$$

And we introduce an additional variable $dy := -B^{-1}A^{*}dp$, then have

$$\begin{pmatrix} B & A^{*} \\ -\frac{1}{\sigma}A & BQB \end{pmatrix} \begin{pmatrix} d_y \\ d_p \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{\sigma}\tilde{h}_1 \end{pmatrix} \tag{5.36}$$

To solve this non-symmetric two blocks system efficiently, we choose a specific preconditioner similar to that in [3].

Then we use the line search to get proper step length.

We write it as the following algorithm.

Algorithm 10: Semismooth Newton method for subproblem (5.29)

```
Iterate the following steps for $j = 0, 1, \cdots$ .

Step 1. Solve the equation for $d_p^j$,

            $(AB^{-1}A^* + \sigma BQB)d_p^j = -\tilde{h}_1.$

Step 2. Line search: let $m_j$ be the first nonnegative integer such that

            $\tilde{\phi}(p^j + \rho^m d_p^j) - \tilde{\phi}(p^j) \le \tau\rho^m \langle \nabla\tilde{\phi}(p^j), d_p^j \rangle = \tau\rho^m \langle \tilde{h}_1, d_p^j \rangle,$

        Then set $p^{j+1} = p^j + \rho^{m_j} d_p^j$.

        Check stopping criteria (B$'$):

            $dist^2(0, \partial\tilde{\phi}(p^{j+1})) \le \min\left\{ \frac{b}{\sigma_k}\delta_k^2, \left(\frac{\delta_k'}{\sigma_k}\right)^2 \right\} \|u^{k+1} - u^k\|_{C^{-1}}^2.$

        If it is satisfied, stop and let $p^k = p^{j+1}$. Otherwise go to Step 1.
```

Chapter 6

Numerical experiments

In this chapter, we demonstrate the numerical experiments and results for solving some SPOCPs. From the numerical experiments, we reconfirm the theoretical results given in previous chapters. And we also provide some comparison methods to solve the decoupled SPOCPs. Numerical experiments show that our methods, especially semismooth Newton augmented Lagrangian method(SSNAL), outperform other methods given.

6.1 Comparison methods

In this section, we provide some efficient existing methods for solving the decoupled SPOCPs of the type,

$$\begin{cases} \min\limits_{y,u\in\mathbb{R}^m} \quad J(y,u) = \dfrac{1}{2}\|y-y_d\|_B^2 + \dfrac{\alpha}{2}\|u\|_B^2 + q(u) \\ \qquad\qquad \text{s.t.} \quad Ay = B(u+y_c). \end{cases} \tag{6.1}$$

6.1.1 Inexact semi-proximal ADMM

This is an efficient algorithm proposed by Fazel, Maryam, Pong, Sun and Tseng [34] in 2013.

Here we rewrite our problem as

$$\begin{cases} \min\limits_{u,v\in\mathbb{R}^m} & J(u,v)=\frac{1}{2}\|A^{-1}B(u+y_c)-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+q(v) \\ \quad\text{s.t.} & B^{\frac{1}{2}}(u-v)=0 \end{cases} \tag{6.2}$$

We denote the $\tilde{\lambda}=B^{\frac{1}{2}}\lambda$ be the multiplier for the equality constraint. The Lagrangian function is defined as

$$L_\sigma(u,v;\lambda)=\frac{1}{2}\|A^{-1}B(u+y_c)-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+q(v)+\langle B\lambda,u-v\rangle+\frac{\sigma}{2}\|u-v\|_B^2. \tag{6.3}$$

Algorithm 11: (isPADMM algorithm for (6.2))

Input: (u^0,v^0,λ^0). Set $k=1$.

Output: $(u^{k+1},v^{k+1},\lambda^{k+1})$

Iterate until convergence

Step 1

$$\begin{aligned} u^{k+1}=&\arg\min_u\frac{1}{2}\|A^{-1}B(u+y_c)-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+\langle B\lambda^k,u\rangle+\frac{\sigma}{2}\|u-v^k\|_B^2, \\ =&(BA^{-1}B+(\alpha+\sigma)A^*)^{-1}(A^*(\sigma v^k-\lambda^k)+By_d-BA^{-1}By_c). \end{aligned} \tag{6.4}$$

Step 2

$$\begin{aligned} v^{k+1}=&\arg\min_v\frac{\sigma}{2}\|v-u^{k+1}\|_B^2-\langle B\lambda^k,v\rangle+q(v)+\frac{\sigma}{2}\|v-v^k\|_{C-B}^2, \\ =&\Pi_{[a,b]}(soft(v^k+\frac{1}{\sigma}C^{-1}B\lambda^k+C^{-1}B(u^{k+1}-v^k),\frac{\beta}{\sigma})). \end{aligned} \tag{6.5}$$

Step 3 Update the multiplier $\lambda^{k+1}=\lambda^k+\tau\sigma(u^{k+1}-v^{k+1})$.

To solve the subproblem in Step 1, we introduce one additional variable $y(=A^{-1}B(u+y_c))$, and solve the linear system as follow

$$\begin{aligned} &\begin{cases} Bu^{k+1}+By_c-Ay^{k+1}=0, \\ A^*((\alpha+\sigma)u^{k+1}-\sigma v^k+\lambda^k)+B(y^{k+1}-y_d)=0, \end{cases} \\ \Leftrightarrow&\begin{pmatrix} B & -A \\ (\alpha+\sigma)A^* & B \end{pmatrix}\begin{pmatrix} u^{k+1} \\ y^{k+1} \end{pmatrix}=\begin{pmatrix} -By_c \\ By_d+A^*(\sigma v^k-\lambda^k) \end{pmatrix} \end{aligned} \tag{6.6}$$

The KKT condition for our problem is given as

$$\begin{cases} B(y-y_d)+A^*(\alpha u+\lambda)=0, \\ Ay-B(u+y_c)=0, \\ u-\Pi_{[a,b]}(\alpha^{-1}soft(\alpha u+C^{-1}B(p-\alpha u),\beta))=0. \end{cases} \tag{6.7}$$

Obviously, this method can also deal with the case $\alpha=0$.

In this case, the third equation of the KKT system is changed to be

$$u-\Pi_{[a,b]}(Csoft(C^{-1}(u+B\lambda),\beta))=0. \tag{6.8}$$

6.1.2 Semismooth Newton method

Here we also provide the globalized semismooth Newton method for numerical comparison.

We reduce our problem (6.1) as the form given as follows

$$\min_u \quad \tilde{J}(u)=\frac{1}{2}\|A^{-1}B(u+y_c)-y_d\|_B^2+\frac{\alpha}{2}\|u\|_B^2+\beta\|Cu\|_1+\delta_{[a,b]}(u). \tag{6.9}$$

Firstly, we see the KKT equations are

$$\begin{cases} A^*p+B(y-y_d)=0, \\ Ay-B(u+y_c)=0, \\ u-\Pi_{[a,b]}(\alpha^{-1}soft(\alpha u+C^{-1}B(p-\alpha u),\beta))=0. \end{cases} \tag{6.10}$$

Or we can reduced it to be

$$F(u)=u-\Pi_{[a,b]}(\alpha^{-1}soft(\alpha u+C^{-1}B(Tu-\alpha u),\beta))=0. \tag{6.11}$$

where $Tu=Su+B(A^{-*}B(A^{-1}By_c-y_d))$, $S=BA^{-*}BA^{-1}B+\alpha B\succ 0$.

Due to the semismooth of the functions $\Pi_{[a,b]}(\cdot)$, $soft(\cdot,\beta)$, F is also semismooth. We can then use semismooth Newton method to solve our problem.

To obtain a good enough solution to equation (6.11), we need to solve the linear equations (6.10) up to a very high accuracy. For example, to obtain the accuracy of 10^{-5} for equation (6.11), we need to solve the equations (6.10) up to the accuracy

$10^{-5}(h\tau)^2\alpha$. If we choose $h = 2^{-5}, \tau = 2^{-6}$ and $\alpha = 5 \times 10^{-5}$, we need to solve equations (6.10) up to the accuracy 10^{-16}.

Hence it is very costly to apply semismooth Newton method when the α is very small or when the mesh-size is very fine.

To provide a globalized method, we will need the line search technique. Here we use the line search technique mentioned in [63, Algorithm 1]. We present the framework of globalized semismooth Newton method as follow.

Algorithm 12: (Globalized semismooth Newton method for (6.11))

Let $\sigma \in (0,1)$, iterate until convergence

Step 1. Select an $V_k \in \partial_B F(u^k)$ and find an approximate solution d^k to

$$F(u^k) + V_k d = 0, \tag{6.12}$$

such that

$$\|F(u^k) + V_k d\| \leq \eta_k \|F(u^k)\|, \tag{6.13}$$

where $\eta_k := \min\{\eta, \|F(u^k)\|\}$.

Step 2. Let m_k be the smallest nonnegative integer m such that

$$f(u^k + \rho^m d^k) - f(u^k) \leq -\sigma \rho^m f(u^k). \tag{6.14}$$

Set $u^{k+1} = u^k + \rho^{m_k} d^k$.

where $f(u) = \frac{1}{2}\|F(u)\|^2$.

Another approach to apply the line search is to use the smoothing Newton method, which require us to find a smoothing function for F. For the detail, you can referred to [64].

6.2 Numerical examples

We begin by describing the algorithmic details which are common to all examples unless otherwise mentioned.

Discretization. As shown in Section 3.1, the discretization is carried out using piecewise linear and continuous finite elements for the space and backward Euler for the time.

The assembly of the mass and the stiffness matrices, as well as the lump mass matrix, is left to the iFEM software package, see [20].

To present the finite element error estimates results, it is convenient to introduce the experimental order of convergence (EOC), which is defined as follows: given two grid sizes $h_1 \neq h_2$, let

$$\text{EOC} := \frac{\log E(h_1) - \log E(h_2)}{\log h_1 - \log h_2}, \tag{6.15}$$

where $E(h)$ with $h > 0$ is a positive error function.

It follows from this definition that if $E(h) = \mathcal{O}(h^\gamma)$, then EOC $\approx \gamma$. The error function $E(\cdot)$ investigated in the present section is given by

$$E(h) := \|u - u_h\|_{L^2(\Omega)}. \tag{6.16}$$

Initialization. For all numerical examples, we choose the initial values as zero for all algorithms.

Parameter setting. For the isPADMM method, the step-length τ for Lagrangian multipliers is chosen as $\tau = 1.68$, and the penalty parameter σ was initially chosen as $\sigma = \alpha$, and we update the σ every 10 iterations via the comparison of the primal and dual feasibility.

For the SSNAL method, initially we choose $\sigma = 1/\alpha$ when $\alpha > 0$, and increase σ by multiplying a constant ρ. When $\alpha = 0$, initially we choose $\sigma = h^2$.

Stopping criterion. In our numerical experiments, we terminate all the algorithms when the corresponding relative residual $\eta < 10^{-5}$. We choose this tolerance since the error estimate is at most $\mathcal{O}(h) \geq 10^{-4}$ when $h \geq 2^{-10}$.

Computational environment. All our computational results are obtained by MATLAB Version 9.0(R2016a) running on a computer with 64-bit MacOS 10.12.6 operation system, Intel(R) 2.7 GHz Intel Core i5 and 8GB of memory.

6.2.1 Testing examples and setting

Before giving the specific examples, we first introduce the following procedure, which can help us formulate sparse optimal control problems.

Algorithm 13: Construct the optimal control problem

Step 1 . Choose $y^* \in L^2(H_0^1(\Omega), [0,T])$ and $p^* \in L^2(H_0^1(\Omega), [0,T])$ arbitrarily.

Step 2 . Set

$$u^* := \begin{cases} \min\{\frac{p^*-\beta}{\alpha}, b\}, & \textit{on} \quad x \in \Omega_T : p^*(x) > \beta, \\ \max\{\frac{p^*+\beta}{\alpha}, a\}, & \textit{on} \quad x \in \Omega_T : p^*(x) < -\beta, \\ 0, & \textit{elsewhere}. \end{cases}$$

Step 3 . Set $y_c = \partial_t y^* - \Delta y^* - u^*$ and $y_d = -\partial_t p^* - \Delta p^* + y^*$.

According to Proposition 3.6, we find that Algorithm 13 provides an optimal solution (y^*, u^*) of the sparse optimal control problem (P). Thus we can construct examples of which we know the exact solution through the above procedure.

We rewrite the problem as follow,

$$\begin{cases} \min \; J(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_T)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega_T)} + \beta\|u\|_{L^1(\Omega_T)} \\ \qquad \text{s.t.} \qquad \partial_t y - \Delta y = u + y_c, \quad \text{in } \Omega_T := \Omega \times [0,T], \\ \qquad\qquad y(\cdot,t) = 0 \quad \text{on } \partial\Omega \times (0,T), \\ \qquad\qquad y(x,0) = 0, \quad x \in \Omega \\ \qquad\qquad u \in U_{ad} = \{v | a \le v(x,t) \le b, \text{a.e } x \in \Omega, t \in (0,T)\}. \end{cases}$$

Example 6.1. *Here, we consider the problem with control $u \in L^2(\Omega_T)$ on the unit square $\Omega = (0,1)^2, T = 1$ with choices of α, β and box constraint $[a,b]$ as below*

(i) $\alpha = 0.5$, $\beta = 0.5$, $a = -1$ *and* $b = 1$*;*

(ii) $\alpha = 5 \times 10^{-5}$, $\beta = 5 \times 10^{-3}$, $a = -100$ *and* $b = 100$.

It is a constructed problem, thus we set $y^* = 2xy\sin(2\pi x)\sin(2\pi y)t$ *and* $p^* = 2xy\sin(2\pi x)\sin(2\pi y)(1-t)$. *Then through* Algorithm 13, *we can easily get the optimal control solution* u^*, *the source term* y_c *and the desired state* y_d.

Example 6.2. *In this example, we consider the problem with control* $u \in L^2(\Omega_T)$ *with* $\Omega = B_1(0), T = 1$, *and the choices of* α, β *and box constraint* $[a, b]$ *are as below*

(i) $\alpha = 0.5$, $\beta = 0.5$, $a = -100$ *and* $b = 100$*;*

(ii) $\alpha = 5 \times 10^{-5}$, $\beta = 5 \times 10^{-3}$, $a = -100$ *and* $b = 100$.

And let $r = \sqrt{x^2+y^2}$, *define*

$$p_1(r) = \begin{cases} 128\alpha r^2 + \beta - 1.5\alpha a, & \text{if} \quad 0 \le r < 1/16, \\ \beta + 16\alpha a(r - 1/8), & \text{if} \quad 1/16 \le r < 1/8, \\ a_1 r^3 + a_2 r^2 + a_3 r + a_4, & \text{if} \quad 1/8 \le r < 3/16, \\ -\beta - 16\alpha b(r - 3/16), & \text{if} \quad 3/16 \le r < 1/4, \\ 256\alpha b(r - 9/32)^2 - \beta - 5/4\alpha b, & \text{if} \quad 1/4 \le r < 5/16, \\ -\beta - 16\alpha b(3/8 - r), & \text{if} \quad 5/16 \le r < 3/8, \\ a_5(r-3/8)^3 + a_6(r-3/8)^2 + 16\alpha b(r-3/8) - \beta, & \text{if} \quad 3/8 \le r \le 1/2, \end{cases} \tag{6.17}$$

where

$$\begin{aligned} &a_1 = 16384\beta + 4096a\alpha - 4096\alpha b, a_2 = 1792\alpha b - 2048a\alpha - 7680\beta, \\ &a_3 = 1152\beta + 336a\alpha - 256\alpha b, a_4 = 12\alpha b - 18a\alpha - 55\beta, \\ &a_5 = 1024\alpha b - 1024\beta, a_6 = 192\beta - 256\alpha b. \end{aligned} \tag{6.18}$$

And the optimal adjoint function is given as

$$p(r,t) := 4p_1(r)(1-t^2). \tag{6.19}$$

Furthermore, we let

$$\begin{cases} y_6(r) = -16/9br^3 + 3/2br^2 + b_9\log(b_{10}r) + c_6; \\ y_5(r) = br^2/4 + b_7\log(b_8r) + c_5 + y_6(5/16); \\ y_4(r) = 16/9br^3 - 3/4br^2 + b_5\log(b_6r) + c_4 + y_5(1/4); \\ y_3(r) = y_4(3/16) + c_3 + b_3\log(b_4r); \\ y_2(r) = a/2r^2 - 16/9ar^3 + b_1\log(b_2r) + c_2 + y_3(1/8); \\ y_1(r) = ar^2/4 + c_1 + y_2(1/16). \end{cases} \tag{6.20}$$

where

$$\begin{aligned} &b_1 = 5/1536, b_2 = 8, b_3 = -35/1536, b_4 = 16/3, b_5 = 25/384, \\ &b_6 = 4, b_7 = -55/384, b_8 = 16/5, b_9 = -845/1536, b_{10} = 8/3, \\ &c_1 = 5/1024, c_2 = 25/1152, c_3 = 0, c_4 = 55/576, \\ &c_5 = -125/1024, c_6 = -75/128; \end{aligned} \tag{6.21}$$

And the optimal state function is given as

$$y^*(r,t) = \begin{cases} 4y_i(r)t^2, & \text{if} \quad (i-1)/16 \le r < i/16, i = 1, \cdots, 6, \\ 0, & \text{elsewhere.} \end{cases} \tag{6.22}$$

Then through Algorithm 13, we can easily get the optimal control solution u^*, the source term y_c and the desired state y_d.

Example 6.3. *Here, we consider the problem with control $u \in L^2(\Omega_T)$ on the unit square $\Omega = (0,1)^2, T = 1$, and we choose the parameters α, β and box constraint $[a,b]$ as below*

(i) $\alpha = 0.5$, $\beta = 0.5$, $a = -0.5$ *and* $b = 0.5$*;*

(ii) $\alpha = 5 \times 10^{-5}$, $\beta = 5 \times 10^{-3}$, $a = -30$ *and* $b = 30$*.*

We choose $y_c = 10^4 \sin 2\pi x \sin 2\pi y$, $y_d = \sin(2\pi x)\exp(2x)\sin(2\pi y)\sin(\pi t)$. Obviously, this is an example with explicit optimal control unknown.

Example 6.4. *Here, we consider the problem with control $u \in L^2(\Omega_T)$ on the unit square $\Omega = (0,1)^2, T = 1$ with choices of α, β and box constraint $[a,b]$ as below*

(i) *$\alpha = 0.5$, $\beta = 0.5$, $a = -1$ and $b = 1$;*

(ii) *$\alpha = 5 \times 10^{-5}$, $\beta = 5 \times 10^{-3}$, $a = -5$ and $b = 5$.*

To have a more complicate problem, we set $y^* = \sin(\pi x)\sin(\pi y)(8(t-0.5)^3+1)$ and $p^* = 2\beta\sin(2\pi x)\exp(0.5x)\sin(4\pi y)(t^2-1)$, which is related to the parameter β. Then through Algorithm 13, we can easily get the optimal control solution u^*, the source term y_c and the desired state y_d.

Remark 6.1. *Here, I will explain about the examples*

1. *Example 1, 2, 4 are examples with explicit optimal solution u constructed. We can use them to check the error estimate and uniformly mesh-independence results.*

2. *We choose large α ($= 0.5$) for two reasons. Firstly, according to Corollary 3.1, we have $\|u^*_{h,\tau} - u^*\|_{L^2(\Omega_T)} \le C_0(h/\alpha + h^2/\alpha^{3/2})$, if β, τ are fixed. If α is too small, the dominated term would be $h^2/\alpha^{3/2}$. Hence we choose $\alpha = 0.5$ to check the error order with respective to the mesh-size is at least 1. Secondly, we choose large α to check the efficiency of the imABCD method.*

3. *For different choices of α and β, we have provide the numerical results in Table 6.6 and Table 6.7, where we reconfirm the robustness of the SSNAL method.*

6.2.2 Numerical results

Error order and error estimate

Firstly, we aim to check our error estimate given in Chapter 3.

Table 6.1, Table 6.2 and 6.3 show the iteration, residual and computation time, L^2 error and the experimental order of convergence(EOC) for both imABCD method for ($\widetilde{\mathrm{D}}_{h,\tau}$) and sGS-imABCD method for ($\widehat{\mathrm{D}}_{\mathrm{h},\tau}$) with the mesh-size varying from 2^{-3} to 2^{-7}.

From those tables, we see that the convergence order with respect to the meshsize h is always greater than 1 when the mesh is fine, which confirms our error estimate of optimal control u with respective to meshsize h is at least $\mathcal{O}(h)$.

By comparing the performance of sGS-imABCD and imABCD methods for $(\widehat{\mathrm{D}}_{\mathrm{h},\tau})$ and $(\widetilde{\mathrm{D}}_{h,\tau})$, we also see the L^2 error obtained from our new discretization is often smaller than that from the approximate discretization.

Table 6.1: Example 1 (i) with $\alpha = 0.5, \beta = 0.5$: the performance of imABCD for $(\widetilde{\mathrm{D}}_{h,\tau})$ and isGSABCD for $(\widehat{\mathrm{D}}_{\mathrm{h},\tau})$. Here I choose a fixed time-step $\tau = 2^{-6}$.

Method	h	Iteration	Residual η	CPU time/s	$\|u^*_{h,\tau} - u^*\|_{L^2}$	EOC
imABCD	2^{-3}	18	5.46e-06	2.34	0.049138	-
	2^{-4}	17	8.98e-06	5.57	0.012785	1.94
	2^{-5}	18	7.12e-06	24.42	0.003910	1.54
	2^{-6}	18	9.86e-06	110.58	0.001241	1.66
	2^{-7}	19	6.68e-06	574.48	0.000411	1.59
sGS-imABCD	2^{-3}	9	6.55e-06	1.10	0.032631	-
	2^{-4}	10	2.99e-06	2.93	0.010996	1.60
	2^{-5}	12	6.67e-06	13.52	0.003548	1.63
	2^{-6}	15	4.71e-06	78.82	0.001167	1.60
	2^{-7}	15	9.54e-07	375.15	0.000398	1.55

Table 6.2: Example 2 (i) with $\alpha = 0.5, \beta = 0.5$: the performance of imABCD for $(\widetilde{\mathrm{D}}_{h,\tau})$ and isGSABCD for $(\widehat{\mathrm{D}}_{\mathrm{h},\tau})$. Here I choose a fixed time-step $\tau = 2^{-6}$.

Method	h	Iteration	Residual η	CPU time/s	$\|u^*_{h,\tau} - u^*\|_{L^2}$	EOC
imABCD	2^{-3}	17	9.67e-06	1.82	67.282657	-
	2^{-4}	16	6.25e-06	3.75	39.838774	0.76
	2^{-5}	16	6.57e-06	15.01	17.488693	1.19
	2^{-6}	18	7.51e-06	88.69	6.008151	1.54
	2^{-7}	18	9.85e-07	379.07	2.880253	1.39
sGS-imABCD	2^{-3}	17	7.69e-06	1.63	67.297358	-
	2^{-4}	17	7.11e-06	3.75	39.800728	0.76
	2^{-5}	18	6.92e-06	14.78	17.498566	1.19
	2^{-6}	20	5.35e-06	73.94	5.968034	1.55
	2^{-7}	20	7.13e-06	372.34	2.861820	1.06

Table 6.3: Example 4 (i) with $\alpha = 0.5, \beta = 0.5$: the performance of imABCD for $(\widetilde{\mathrm{D}}_{h,\tau})$ and isGSABCD for $(\widehat{\mathrm{D}}_{\mathrm{h},\tau})$. Here I choose a fixed time-step $\tau = 2^{-6}$.

Method	h	Iteration	Residual η	CPU time/s	$\|u^*_{h,\tau} - u^*\|_{L^2}$	EOC
imABCD	2^{-3}	17	8.43e-06	2.01	0.2388831	-
	2^{-4}	18	5.53e-06	6.76	0.104037	1.20
	2^{-5}	17	7.40e-06	28.08	0.026016	1.99
	2^{-6}	16	8.00e-06	117.20	0.008680	1.58
	2^{-7}	17	7.47e-06	620.83	0.002995	1.54
sGS-imABCD	2^{-3}	15	5.36e-06	1.45	0.126118	-
	2^{-4}	16	3.41e-06	3.81	0.080597	0.65
	2^{-5}	15	2.81e-06	16.89	0.023205	1.80
	2^{-6}	16	6.13e-06	80.01	0.008141	1.51
	2^{-7}	17	8.15e-06	393.83	0.002863	1.51

Now let us check the uniformly mesh-independence of the sGS-imABCD method, the imABCD method and the SSNAL method proven in Chapter 4 and Chapter 5.

From Table 6.1, Table 6.2, Table 6.3, we see that the iteration numbers of the imABCD method and the sGS-imABCD method show some consistency that they keep steady while the mesh-size refines, and we call it the uniformly mesh-independence property. Especially, we find that the iteration numbers for imABCD method stay almost the same for all the examples when mesh-size varies.

Table 6.4: The performance of SSNAL method for solving Example 1 (i), Example 2 (i) and Example 4 (i) with fixed time-step $\tau = 2^{-6}$. Here, iter is is the outer iteration, Newton iter is the number of Newton equations solved.

h	Index of performance	Example 1(i)	Example 2(i)	Example 4(i)
2^{-3}	iter(Newton iter)	8(7)	8(7)	8(7)
	residual η	1.85e-06	2.46e-06	5.45e-06
	CPU time/s	2.17	1.70	2.23
2^{-4}	iter(Newton iter)	8(7)	8(7)	8(7)
	residual η	3.07e-06	9.25e-06	2.57e-06
	CPU time/s	3.84	3.64	4.10
2^{-5}	iter(Newton iter)	8(7)	9(8)	9(8)
	residual η	3.25e-06	1.15e-06	1.16e-06
	CPU time/s	14.99	17.35	18.33
2^{-6}	iter(Newton iter)	8(7)	8(7)	9(8)
	residual η	2.52e-06	8.43e-06	1.12e-06
	CPU time/s	68.21	82.83	82.80
2^{-7}	iter(Newton iter)	8(7)	8(7)	8(7)
	residual η	1.86e-06	6.62e-06	8.20e-06
	CPU time/s	373.72	561.22	418.09

Table 6.4 provides the performance of SSNAL for the example 6.1(i), 6.2(i) and

6.4(i) with the mesh-size h varying from 2^{-3} to 2^{-7}, $\tau = 2^{-6}$. In the table, #dofs stands for the number of degrees of freedom for the control variable on each grid level.

From Table 6.4, we see that SSNAL method also shows a good mesh-independence property. Both the outer iteration number and Newton step number are almost the same for different choices of mesh-size.

Efficiency

Table 6.5: The performance of imABCD, SSNAL, isPADMM and SSN methods for the all examples with different choice of α and β. Here I choose the mesh-size of space to be $h = 2^{-5}$, and the time-step to be $\tau = 2^{-6}$.

Problem		imABCD		SSNAL		isPADMM		SSN	
		iter	time	iter	time	iter	time	iter	time
Example 1	(i)	18	24.42	8(7)	15.00	24	34.12	3	23.19
	(ii)	84	210.07	10(9)	64.40	108	206.27	6	125.34
Example 2	(i)	16	15.02	9(8)	17.35	25	23.57	6	37.11
	(ii)	97	180.75	10(9)	47.56	90	97.18	8	109.66
Example 3	(i)	17	19.94	7(6)	13.52	22	24.80	3	21.18
	(ii)	59	143.73	4(3)	10.21	137	128.81	4	21.29
Example 4	(i)	17	28.08	9(8)	18.33	24	30.24	3	21.18
	(ii)	62	122.67	8(7)	29.21	57	88.80	4	67.42

Due to the mesh-independence of our methods, we do not need to provide the numerical experiments of all mesh-sizes. Here we choose the mesh-size $h = 2^{-5}$ and the time-step $\tau = 2^{-6}$. We aim to compare our methods, imABCD method, the SSNAL method, with the state-of-art methods, the isPADMM and the globalized semismooth Newton method(SSN). And the numerical results are provided in Table 6.5.

From Table 6.5, we see that the imABCD method shows moderate efficiency for all the examples with large α, especially for Example 6.2(i). And for all examples except for Example 6.2 (i), the SSNAL method outperforms other methods. In particular, when α is small, the SSNAL method can be at least twice faster than the rest methods of all the methods provided.

Robustness

Now we also study the robustness of SSNAL method to the parameter α.

In Table 6.6 and Table 6.7, we provide the numerical performance of SSNAL method for all the examples with $\beta = 5 \times 10^{-1}$ and $\beta = 5 \times 10^{-3}$, and α being $5 \times 10^{-1}, 5 \times 10^{-2}, 5 \times 10^{-3}, 5 \times 10^{-4}$ and 5×10^{-5} respectively.

Table 6.6: The performance of SSNAL for the all examples with different choice of α and fixed $\beta = 0.5$. The mesh-size of space is $h = 2^{-5}$, and the time-step is $\tau = 2^{-6}$.

Problem	Index of performance	α				
		5e-1	5e-2	5e-3	5e-4	5e-5
Example 1	iter	9(8)	8(7)	6(5)	3(2)	2(1)
	CPU time/s	23.71	21.43	15.49	6.67	4.23
Example 2	iter	9(8)	8(7)	8(7)	7(6)	4(3)
	CPU time/s	16.42	18.40	17.30	16.15	8.24
Example 3	iter	8(7)	6(5)	3(2)	2(1)	2(0)
	CPU time/s	16.42	11.65	5.26	3.12	1.59
Example 4	iter	8(7)	8(7)	8(7)	6(5)	6(5)
	CPU time/s	17.05	20.05	25.38	19.52	20.49

Table 6.7: The performance of SSNAL for the all examples with different choice of α and fixed $\beta = 5 \times 10^{-3}$. The mesh-size of space is $h = 2^{-5}$, and the time-step is $\tau = 2^{-6}$.

Problem	Index of performance	α				
		5e-1	5e-2	5e-3	5e-4	5e-5
Example 1	iter	14(13)	14(13)	14(13)	14(13)	12(11)
	CPU time/s	41.60	50.02	74.47	85.60	76.14
Example 2	iter	15(14)	15(14)	15(14)	14(13)	12(11)
	CPU time/s	32.06	41.26	45.94	50.93	56.98
Example 3	iter	14(13)	12(11)	9(8)	2(1)	2(1)
	CPU time/s	30.89	23.18	16.98	3.18	3.55
Example 4	iter	15(14)	15(14)	15(14)	15(14)	13(12)
	CPU time/s	31.38	34.52	46.16	63.16	57.06

It can be found from Table 6.6 and Table 6.7 that, the iteration number decrease along the decrease of α. This confirms the result concluded in Remark 5.3 that SSNAL method is even better for solving decoupled SPOCPs with small L^2 regularization parameter α.

Finally, let us consider using the SSNAL method to solve problems with $\alpha = 0$. We can no longer use the Algorithm 13 to construct an example with explicit optimal solution. However, we still use the functions y_c, y_d, and parameters β, a, b the same as in the examples 6.1(i),6.2(i),6.3(i) and 6.4(i). Table 6.8 provides the comparison of the SSNAL method and the isPADMM method for solving the examples with given parameters y_c, y_d, β, a, b.

We find that the SSNAL method can solve the problems at least 6 times faster than the isPADMM method for the examples provided.

Table 6.8: The performance of SSNAL and isPADMM for the all examples with $\alpha = 0$. Here we choose the mesh-size of space to be $h = 2^{-5}$, and the time-step to be $\tau = 2^{-6}$.

Problem	β	Index of performance	SSNAL	isPADMM
Example 1	5e-1	iter	3(6)	31
		CPU time/s	7.55	69.22
Example 2	5e-1	iter	4(14)	41
		CPU time/s	11.43	65.27
Example 3	5e-1	iter	3(6)	41
		CPU time/s	7.86	57.29
Example 4	5e-1	iter	2(4)	35
		CPU time/s	8.95	69.55

Chapter 7

Conclusions and future work

7.1 Conclusions

In this thesis, we study algorithms for solving the sparse optimal parabolic control problems.

From the view of duality, we figure out a new discretization of the continuous problem, which can avoid the approximation of the L^1-norm and avoid the inevitable additional error at the same time. We give the error estimate for the new discretization, which is $\mathcal{O}(h+\sqrt{\tau})$, and we reconfirm the error order in numerical experiments.

Given the discretized dual problem is an unconstrained optimization problem, we propose the sGS-imABCD method to solve the new discretization problem. In general the convergence of the dual variables is not necessary for APG or ABCD method. However, we are able to prove the convergence of the primal variable and KKT conditions based on the convergence of the optimal value and the special structure of our problems. For the conventional discretized problem, we are able to prove the convergence of both the primal variable and dual variable. Further, we prove the uniformly mesh-independence of the imABCD, which means the iteration number is independent of the mesh when the mesh-size is at very fine level.

Later, devoted to deal with the case when α is very small or α is zero, we consider the semismooth Newton augmented Lagrangian method (SSNAL), which has a fast linear convergence rate, and we solve the subproblem efficiently by semismooth Newton method. Moreover, the SSNAL method makes good use of the sparsity structure of our problems. We illustrate the implementation of the method. And we also provide the convergence results and an improved convergence result as well. When $\alpha > 0$, we prove the uniformly mesh-independence of the SSNAL method with respect to the mesh-size h. We also obtain the robustness of the SSNAL merthod to regularization parameter α when the initial σ is properly chosen. Our SSNAL method can also solve decoupled SPOCPs for the case $\alpha = 0$, which is much more efficient than the isPADMM method.

Numerical results show that both the imABCD method and the SSNAL method present the high efficiency to solve the decoupled SPOCPs when α is not too small. And SSNAL method outperforms isPADMM, Semismooth Newton method (SSN) and imABCD method for all α, especially when α is very small. And numerical results also reconfirm the theoretical results about the error estimate of new discretization method. Both the imABCD method and the SSNAL method show the uniformly mesh-independence in numerical experiments. Moreover, the SSNAL method also shows the robustness to the parameter α.

7.2 Future work

In the future, there are still some work to be done.

1. We can consider to extend our convergence results for both ABCD method and SSNAL method to the methods in function space. In particular, we hope to extend the uniformly mesh-independence property for SSNAL method to mesh-independence.

2. We may also analysis convergence theory about the case $\alpha = 0$.

Bibliography

[1] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Computational Optimization and Applications, 23 (2002), pp. 201–229.

[2] F. J. Aragón Artacho and M. H. Geoffroy, *Characterization of metric regularity of subdifferentials*, Journal of Convex Analysis, 15 (2008), pp. 365–380.

[3] O. Axelsson, S. Farouq, and M. Neytcheva, *Comparison of preconditioned krylov subspace iteration methods for pde-constrained optimization problems*, Numerical Algorithms, 73 (2016), pp. 631–663.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[5] M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Computational Optimization and Applications, 22 (2002), pp. 193–224.

[6] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15, Springer Science and Business Media, 2007.

[7] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, ESAIM: Mathematical Modelling and Numerical Analysis, 33 (1999), pp. 1187–1202.

[8] E. Casas, C. Clason, and K. Kunisch, *Approximation of elliptic control problems in measure spaces with sparse solutions*, SIAM Journal on Control and Optimization, 50 (2012), pp. 1735–1752.

[9] ——, *Parabolic control problems in measure spaces with sparse solutions*, SIAM Journal on Control and Optimization, 51 (2013), pp. 28–63.

[10] E. Casas and L. A. Fernández, *Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state*, Applied Mathematics and Optimization, 27 (1993), pp. 35–56.

[11] E. Casas, R. Herzog, and G. Wachsmuth, *Approximation of sparse controls in semilinear equations by piecewise linear functions*, Numerische Mathematik, 122 (2012), pp. 645–669.

[12] ——, *Optimality conditions and error analysis of semilinear elliptic control problems with l^1 cost functional*, SIAM Journal on Optimization, 22 (2012), pp. 795–820.

[13] E. Casas and K. Kunisch, *Optimal control of semilinear elliptic equations in measure spaces*, SIAM Journal on Control and Optimization, 52 (2014), pp. 339–364.

[14] ——, *Parabolic control problems in space-time measure spaces*, ESAIM: Control, Optimisation and Calculus of Variations, 22 (2016), pp. 355–370.

[15] E. Casas and M. Mateos, *Error estimates for the numerical approximation of neumann control problems*, Computational Optimization and Applications, 39 (2008), pp. 265–295.

[16] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Computational Optimization and Applications, 31 (2005), pp. 193–219.

[17] E. Casas and J.-P. Raymond, *Error estimates for the numerical approximation of dirichlet boundary control for semilinear elliptic equations*, SIAM Journal on Control and Optimization, 45 (2006), pp. 1586–1611.

[18] E. Casas and F. Tröltzsch, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, Springer, 2003, pp. 89–100.

[19] A. Chambolle and T. Pock, *A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions*, SIAM Journal of Computational Mathematics, 1 (2015), pp. 29–54.

[20] L. Chen, *ifem: an integrated finite element methods package in matlab*, preprint, University of California at Irvine, CA, (2009).

[21] L. Chen, D. Sun, and K.-C. Toh, *An efficient inexact symmetric gauss–seidel based majorized admm for high-dimensional convex composite conic programming*, Mathematical Programming, 161 (2017), pp. 237–270.

[22] C. Clason and K. Kunisch, *A duality-based approach to elliptic control problems in non-reflexive banach spaces*, ESAIM: Control, Optimisation and Calculus of Variations, 17 (2011), pp. 243–266.

[23] ——, *A measure space approach to optimal source placement*, Computational Optimization and Applications, 53 (2012), pp. 155–171.

[24] S. S. Collis and M. Heinkenschloss, *Analysis of the streamline upwind/petrov galerkin method applied to the solution of optimal control problems*, CAAM TR02-01, (2002).

[25] L. Costa, I. Figueiredo, R. Leal, P. Oliveira, and G. Stadler, *Modeling and numerical study of actuator and sensor effects for a laminated piezoelectric plate*, Computers and Structures, 85 (2007), pp. 385–403.

[26] Y. Cui, *Large scale composite optimization problems with coupled objective functions: theory, algorithms and applications*, PhD thesis, 2016.

[27] Y. Cui, X. Li, D. Sun, and K.-C. Toh, *On the convergence properties of a majorized admm for linearly constrained convex optimization problems with coupled objective functions*, arXiv preprint arXiv:1502.00098, (2015).

[28] Y. Cui, D. Sun, and K.-C. Toh, *On the asymptotic superlinear convergence of the augmented lagrangian method for semidefinite programming with multiple solutions*, arXiv preprint arXiv:1610.00875, (2016).

[29] K. Deckelnick and M. Hinze, *Error estimates in space and time for tracking-type control of the instationary stokes system*, in Control and Estimation of Distributed Parameter Systems, Springer, 2003, pp. 87–103.

[30] ——, *Semidiscretization and error estimates for distributed control of the instationary navier-stokes equations*, Numerische Mathematik, 97 (2004), pp. 297–320.

[31] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, 2014.

[32] L. C. Evans, *Partial differential equations and monge-kantorovich mass transfer*, Current Developments in Mathematics, 1997 (1997), pp. 65–126.

[33] R. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, Journal of Mathematical Analysis and Applications, 44 (1973), pp. 28–47.

[34] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, *Hankel matrix rank minimization with applications to system identification and realization*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 946–977.

[35] I. M. N. Figueiredo and C. M. F. Leal, *A piezoelectric anisotropic plate model*, Asymptotic Analysis, 44 (2005), pp. 327–346.

[36] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO. Analyse Numérique, 13 (1979), pp. 313–328.

[37] R. Griesse, N. Metla, and A. Rösch, *Convergence analysis of the sqp method for nonlinear mixed-constrained elliptic optimal control problems*, ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 88 (2008), pp. 776–792.

[38] R. Griesse, N. Metla, and A. Rösch, *Local quadratic convergence of sqp for elliptic optimal control problems with mixed control-state constraints*, Control and Cybernetics, 39 (2010), pp. 717–738.

[39] M. D. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for navier–stokes flows with distributed control*, SIAM Journal on Numerical Analysis, 37 (2000), pp. 1481–1512.

[40] M. Heinkenschloss and D. Ridzal, *An inexact trust-region sqp method with applications to pde-constrained optimization*, Numerical Mathematics and Advanced Applications, (2008), pp. 613–620.

[41] R. Herzog, G. Stadler, and G. Wachsmuth, *Directional sparsity in optimal control of partial differential equations*, SIAM Journal on Control and Optimization, 50 (2012), pp. 943–963.

[42] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth newton method*, SIAM Journal on Optimization, 13 (2002), pp. 865–888.

[43] M. Hinze, *Optimal and instantaneous control of the instationary Navier-Stokes equations*, PhD thesis, Habilitation thesis, Technische Universitat Berlin, Berlin, Germany, 2000.

[44] ——, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Computational Optimization and Applications, 30 (2005), pp. 45–61.

[45] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, vol. 23, Springer Science and Business Media, 2008.

[46] M. Hinze and M. Vierling, *The semi-smooth newton method for variationally discretized control constrained elliptic optimal control problems; implementation, convergence and globalization*, Optimization Methods and Software, 27 (2012), pp. 933–950.

[47] K. Jiang, D. Sun, and K.-C. Toh, *An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp*, SIAM Journal on Optimization, 22 (2012), pp. 1042–1064.

[48] K. Kunisch, K. Pieper, and B. Vexler, *Measure valued directional sparsity for parabolic optimal control problems*, SIAM Journal on Control and Optimization, 52 (2014), pp. 3078–3108.

[49] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM Journal on Optimization, 13 (2002), pp. 321–334.

[50] K. Kunisch, P. Trautmann, and B. Vexler, *Optimal control of the undamped linear wave equation with measure valued controls*, SIAM Journal on Control and Optimization, 54 (2016), pp. 1212–1244.

[51] X. Li, *A Two-Phase Augmented Lagrangian Method For Convex Composite Quadratic Programming*, PhD thesis, 2015.

[52] X. Li, D. Sun, and K.-C. Toh, *Qsdpnal a two-phase proximal augmented lagrangian method for convex quadratic semidefinite programming*, arXiv preprint arXiv 1512 08872, (2015).

[53] ——, *A highly efficient semismooth newton augmented lagrangian method for solving lasso problems*, arXiv preprint arXiv:1607.05428, (2016).

[54] ——, *A schur complement based semi-proximal admm for convex quadratic conic programming and extensions*, Mathematical Programming, 155 (2016), pp. 333–373.

[55] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM Journal on Control and Optimization, 22 (1984), pp. 277–293.

[56] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems*, Applied Mathematics and Optimization, 8 (1982), pp. 69–95.

[57] D. Meidner and B. Vexler, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems part i: Problems without control constraints*, SIAM Journal on Control and Optimization, 47 (2008), pp. 1150–1177.

[58] ——, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems part ii: problems with control constraints*, SIAM Journal on Control and Optimization, 47 (2008), pp. 1301–1329.

[59] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM Journal on Control and Optimization, 43 (2004), pp. 970–985.

[60] ——, *L ˆ-estimates for approximated optimal control problems*, SIAM Journal on Control and Optimization, 44 (2005), pp. 1636–1649.

[61] H. D. Mittelmann and H. Maurer, *Solving elliptic control problems with interior point and sqp methods: control and state constraints*, Journal of Computational and Applied Mathematics, 120 (2000), pp. 175–195.

[62] B. S. Mordukhovich and M. E. Sarabi, *Critical multipliers in variational systems via second-order generalized differentiation*, Mathematical Programming, (2017), pp. 1–44.

[63] M. Porcelli, V. Simoncini, and M. Stoll, *Preconditioning pde-constrained optimization with l1-sparsity and control constraints*, Computers & Mathematics with Applications, 74 (2017), pp. 1059–1075.

[64] L. Qi and D. Sun, *A survey of some nonsmooth equations and smoothing newton methods*, in Progress in optimization, Springer, 1999, pp. 121–146.

[65] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for pde-constrained optimization*, SIAM Journal on Scientific Computing, 32 (2010), pp. 271–298.

[66] R. T. Rockafellar, *Augmented lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of Operations Research, 1 (1976), pp. 97–116.

[67] ——, *Monotone operators and the proximal point algorithm*, SIAM journal on control and optimization, 14 (1976), pp. 877–898.

[68] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 2015.

[69] A. Rosch, *Error estimates for parabolic optimal control problems with control constraints*, journal for analysis and its applications, 23 (2004), pp. 353–376.

[70] A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optimization Methods and Software, 21 (2006), pp. 121–134.

[71] A. Schindele and A. Borzì, *Proximal methods for elliptic optimal control problems with sparsity cost functional*, Applied Mathematics, 7 (2016), pp. 967–992.

[72] X. Song, B. Chen, and B. Yu, *An efficient duality-based approach for pde-constrained sparse optimization*, arXiv preprint arXiv:1708.09094, (2017).

[73] X. Song, B. Yu, Y. Wang, and X. Zhang, *A fe-inexact heterogeneous admm for elliptic optimal control problems with l^1-control cost*, arXiv preprint arXiv:1709.01067, (2017).

[74] G. Stadler, *Elliptic optimal control problems with l 1-control cost and applications for the placement of control devices*, Computational Optimization and Applications, 44 (2009), pp. 159–181.

[75] D. Sun, K.-C. Toh, and L. Yang, *A convergent 3-block semiproximal alternating direction method of multipliers for conic programming with 4-type constraints*, SIAM Journal on Optimization, 25 (2015), pp. 882–915.

[76] ——, *An efficient inexact abcd method for least squares semidefinite programming*, SIAM Journal on Optimization, 26 (2016), pp. 1072–1100.

[77] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, vol. 1054, Springer, 1984.

[78] K.-C. Toh and S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems*, Pacific Journal of Optimization, 6 (2010), pp. 615–640.

[79] M. ULBRICH, *Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces*, PhD thesis, Habilitation thesis, Fakultät für Mathematik, Technische Universität München, 2002.

[80] ——, *Semismooth newton methods for operator equations in function spaces*, SIAM Journal on Optimization, 13 (2002), pp. 805–841.

[81] B. VEXLER, *Finite element approximation of elliptic dirichlet optimal control problems*, Numerical Functional Analysis and Optimization, 28 (2007), pp. 957–973.

[82] G. WACHSMUTH AND D. WACHSMUTH, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM: Control, Optimisation and Calculus of Variations, 17 (2011), pp. 858–886.

[83] A. WATHEN, *Realistic eigenvalue bounds for the galerkin mass matrix*, IMA Journal of Numerical Analysis, 7 (1987), pp. 449–457.

[84] A. J. WATHEN AND T. REES, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electronic Transactions on Numerical Analysis, 34 (2009), pp. 125–135.

[85] F. ZHANG, *The Schur complement and its applications*, vol. 4, Springer Science & Business Media, 2006.

[86] J. C. ZIEMS AND S. ULBRICH, *Adaptive multilevel inexact sqp methods for pde-constrained optimization*, SIAM Journal on Optimization, 21 (2011), pp. 1–40.