# Important Notice

UNIVERSITY OF CALGARY

# Novel Optimization Schemes for Full Waveform Inversion: Optimal Transport and Inexact Gradient Projection

by

Da Li

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA
MARCH, 2021

# Abstract

Full waveform inversion (FWI) is an important seismic inversion technique that provides high-resolution estimates of underground physical parameters. However, high-accuracy inverse results are not guaranteed due to the essential non-convexity characteristics of the FWI problem. This thesis focuses on designing novel optimization schemes for the FWI problem which improve the inverse results.

Applying optimal transport (OT) based distances to the FWI problem is popular because they provide additional geometric information. The OT distances are designed for positive measures with equal mass, and the unbalanced optimal transport (UOT) distance can overcome the mass equality condition. A mixed distance is constructed which can also overcome the mass equality condition, and the convex properties for the shift, dilation, and amplitude change are proved. Both UOT distance and the proposed distance are applied to the FWI problem with normalization methods transforming the signals into positive functions. Numerical examples show that the optimal transport based distances outperform the traditional L2 distance in certain cases.

The gradient projection methods are often used to solve constrained optimization problems, and the closed-form projection function is necessary since the projection has to be evaluated exactly. A constraint set expanding strategy is designed for the gradient projection methods such that the projection can be evaluated inexactly, which extends the application scope of the gradient projection methods. The convergence analysis is provided with proper assumptions.

A priori information of the model is important to improve the inverse result, and an optimization scheme is proposed for incorporating multiple a priori information into the FWI problem. The optimization scheme is a combination of the scaled gradient projection method and a projection onto convex sets algorithm. Also, the L-BFGS Hessian approximation and the above constraint set expanding strategy are used. Numerical examples show that the proposed optimization scheme is flexible for integrating multiple types of constraint sets such as total variation constraint, sparsity constraint, box constraint, and hyperplane constraint into the FWI problem.

# Preface

This thesis is an original work by the author during his Ph.D. program at the University of Calgary. Some portions of this thesis are submitted as a journal paper, published as conference abstracts and as research reports. All of these research works were carried out under the supervision of Dr. Michael Lamoureux and Dr. Wenyuan Liao at the University of Calgary.

A portion of the research work in Chapter 2, Section 4 is in the submitted paper: Li, D., Li, K., and Liao, W., Efficient and stable finite difference modelling of acoustic wave propagation in variable-density media. This work is included as part of the introduction in the thesis, and not in the contributions of this thesis.

A portion of the research work in Chapter 4 is published in the technical report: Li, D., Lamoureux, M. P., and Liao, W., 2019, Full waveform inversion with unbalanced optimal transport distance: CREWES Research Report, 31, 40.1–40.13. This portion is also in the conference abstract: Li, D., Lamoureux, M. P., and Liao, W., 2020, Full waveform inversion with unbalanced optimal transport distance: GeoConvention, Conference Abstracts.

A portion of the research work in Chapter 5 and Chapter 6 is published in the technical report: Li, D., Lamoureux, M. P., and Liao, W., 2020, Incorporating multiple a priori information for full waveform inversion: CREWES Research Report, 32, 38.

# Acknowledgements

First and foremost I would like to thank my supervisor Dr. Michael Lamoureux. I thank Dr. Michael Lamoureux for choosing me as his graduate student, giving me the opportunity to join the Ph.D. program for mathematical research, as this has always been a dream of mine. Through the past five years, his guidance and encouragements have been invaluable to me, especially during the most difficult time of my Ph.D. program. I thank him for introducing the seismic imaging and inverse problem to me, which is the main topic of this thesis. I also thank him for encouraging me to study microlocal analysis in order to have a deeper understanding of imaging problem. I am moved by the beauty of the theory, and I understand how precious these experiences are for a graduate student. Apart from research, I am also very grateful for his help in my daily life.

At the same time, I would like to express my gratitude to my co-supervisor Dr. Wenyuan Liao for his guidance during my Ph.D. program. His courses "Scientific Computation (MATH 661.03)" and "Numerical Differential Equations (MATH 661.05)" significantly enhanced my theoretical background in forward modeling of seismic wave propagation, which plays an essential role in my research work. He often organizes discussions and creates opportunities for graduate students to communicate and work together, from which I learned a lot. I have benefited greatly from the collaborative research work he invited me into. When I was lost in the research work, he always reminded me of what is the most important thing and what should I do. I truly appreciate his advice. I can not finish this thesis without the help and support of my supervisors. The Ph.D. program is a precious experience of my life, and I appreciate both of my supervisors for the patient guidance, continuous support, and everything they taught me.

The faculty members and colleagues at the Department of Mathematics and Statistics supported me in various ways. I would like to thank Dr. Cristian Rios and Dr. Yuriy Zinchenko for helping me to build up my knowledge on the differential equations, measure theory, and optimization. Especially I want to thank my analysis teacher, Dr. Elena Braverman for her inspiring class "Analysis III (MATH 603)". The research work in Chapter 5 could not have been finished without the knowledge she taught me. In addition, I want to thank Dr. Yaoting Lin, Dr. Keran Li, Dr. Ebrahim Ghaderpour, and Dr. Hatef Dastour for many discussions

and collaborations. Also, special thanks to Yanmei Fei for helping me get used to a new life in University of Calgary. I thank Matthew Adams for sharing the office with me and for his countless kind help. Thank you to Dr. Peng Yong for countless discussions on the full waveform inversion, optimal transport, and total variation.

I am very grateful to be a member of the Consortium for Research in Elastic Wave Exploration Seismology (CREWES) group at the University of Calgary. Especially, I want to thank Dr. Kristopher Innanen for providing me many opportunities to communicate with geophysical researchers. His class "Inverse Theory & Applications II (GOPH 673)" opened the door of the full waveform inversion problem for me. The Friday talk each week is a great place to understand the frontiers of geophysics research and inspired me a lot. Also thanks to Dr. Wenyong Pan, Dr. Raul Cova, Dr. Huaizhen Chen, Dr. Junxiao Li, Dr. Jian Sun, Xin Fu, Qi Hu, Shang Huang, Liu He, Zhan Niu, Luping Qu, Ziguang Su for valuable help and discussions.

I also gratefully acknowledge the funding support from the China Scholarship Council, the Natural Sciences and Engineering Research Council of Canada (CRDPJ 461179-13, CRDPJ 532227-18, the Discovery grants RGPIN-2015-06038, RGPIN-2019-04830, RGPIN-2020-04561), and the CREWES sponsors.

I would not have been able to finish my Ph.D. program without continuous supports from my family members. As the single child of my parents, I always feel guilty about not being able to always accompany them. However, their selfless love for me has always encouraged me to pursue my dreams. Finally, I want to thank my wife Wen Chen for her love, this thesis can not be finished without her companionship. She is always by my side in my most difficult times and her smile is like the warm winter sunlight scattering in my eye that comforts me a lot. I want to tell her: I love you.

To my father Guoxiang Li, mother Jianying Wang, and wife Wen Chen.

# Contents

# List of Figures

# Chapter 1

# Introduction

Geophysics is a long-standing discipline, and the seismic wave is one of the main research objects. The seismic wave propagating through the Earth brings the physical information of the underground medium. Especially in the exploration geophysics, the seismic wave generated by the artificial sources is one of the most important approaches to reveal the geophysical target structures such as oil and gas reservoirs.

With the seismic data recorded by the receivers, the seismic inversion can be performed to recover the subsurface properties. There are different types of seismic inversion methods. The seismic tomography takes advantage of the travel time information in the seismic data to determine the locations of reflection and refraction of the Earth model. In the impedance inversion method, the physical model of seismic wave propagation is simplified as a one-dimensional convolution between the seismic wavelet and the underground impedance. Then the physical properties can be estimated by comparing the synthetic data to the well-log data. In the amplitude versus offset (AVO) method the physical model is the Zoeppritz equations which approximately describe the seismic wave reflection behavior at an interface. The reflection coefficients can be estimated with the relation of amplitude and the angle of incidence, then the physical parameters can be determined.

Compared to the above conventional seismic inversion methods, the full waveform inversion (FWI) technique takes advantage of the "full" information of the seismic data, including travel time, amplitude, phase, time-frequency information, etc, so that it is expected to provide accurate and high-resolution information of the underground structures in the target area. The FWI technique was developed by Lailly [78] and Tarantola [128] in the early 1980s. In the FWI problem, the PDE that governs the seismic propagation, the recorded seismic data, the function of the seismic sources, and the locations of the receivers and the sources are known a priori. Given an initial model, the FWI problem is to minimize the difference between

the seismic data simulated by the model and the recorded data. Then the inverse result will be updated iteratively by an optimization algorithm starting from the initial model. When the difference between the simulated data and the recorded data is decreasing, the model that generated the simulated data can be expected to become closer to the true model that generated the recorded data.

From the optimization point of view, the FWI problem is a PDE constrained optimization problem. Similar to the optimal control problem, the objective function is the difference between the simulated data and the recorded data, and the constraint PDE is the PDE that describes the seismic propagation such as a wave equation, linear elastic wave equation, Helmholtz equation, etc. The FWI problem can be written in a compact form as

$$\min_{y,u} J(y,u) = \frac{1}{2}\|Qy - y_{\mathrm{d}}\|^2, \tag{1.1}$$

$$\text{such that } e(y,u) = 0, \tag{1.2}$$

where $u$ is the control parameter, representing the physical properties of the model, such as velocity, density, Lamé parameters, etc. The state parameter $y$ is the seismic wavefield, and $Q$ is the recording operator determined by the position of the receivers. The $L^2$ norm is usually used in the above equation. The constrained PDE is written in a compact form. Since the PDE is well-posed, a parameter-to-state map can be well-defined as $y = F(u)$. The above optimization problem can be written in a compact form

$$\min_{u} f(u) = J(F(u),u), \tag{1.3}$$

where $f(u)$ is smooth, nonlinear, and nonconvex.

Tools from different mathematical branches are needed to solve the above PDE constrained optimization problem. The analysis of the PDE constrained optimization problem and the constraint PDE provides the solution properties and the connection between the optimization problem in a continuous setting and the discrete optimization problem. The numerical methods for PDE such as finite difference and finite elements are needed to simulate the seismic data with the model. The gradient and Hessian of the reduced objective function can be evaluated efficiently through the adjoint state method. The FWI problem is large-scale on two aspects. First, the control parameter $u$ and the state parameter $y$ are high-dimensional vectors after the discretization. Second, extensive computation is required since the numerical PDE solver is performed intensively. Due to the large scale of the FWI problem, numerical optimization methods in a deterministic form are needed. Different optimization methods can be applied based on the different formulation of the objective function. In this thesis work, we focus on developing novel optimization schemes for the FWI

problem that can improve the inversion results, such as improving the piece-wise constant structure.

## 1.1 Developments and issues of the FWI problem

Although the basic scheme of the FWI problem was fixed in the early 1980s, it is still evolving in different aspects.

The FWI problem was initialed and discussed with the acoustic approximation of the seismic wave propagation [128, 60], later it was extended to the elastic model [129, 97]. In the work [110, 111], Pratt came up with the FWI problem in the frequency domain. Instead of working with a linear evolution equation, the constraint PDE in the frequency domain FWI is the Helmholtz equation. Both time domain and frequency domain FWI problems are equivalent since the data and the model in time and frequency domain are connected by Fourier transform. The frequency domain approach easily leads to the multi-scale strategies.

The gradient-based methods such as steepest descent and nonlinear conjugate gradient, are the conventional optimization method used for the FWI problem [128, 60]. The update direction is the inverse of the gradient in the steepest descent method and is the linear combination of the gradient of the current iteration and the previous iteration in the nonlinear conjugate gradient method. Despite that the gradient-based method is efficient to evaluate and stable for the large-scale optimization problem, it suffers the slow convergence speed. The second-order information of the objective function is expected to improve the convergence speed [121]. In Newton's method and Gauss-Newton method, the update direction is achieved by the inverse Hessian matrix times the gradient vector, which behaves as a "deconvolution" operator compensating the geometrical spreading effects and deblurring the gradient [112]. However, the inverse Hessian matrix can never be evaluated explicitly due to the large-scale of the FWI problem.

There are two major ways to introduce the second-order information to the optimization algorithm in the FWI problem. In the truncated Newton method, the multiplication between Hessian matrix and the update direction is evaluated in an abstract form by the adjoint state method, then the update direction is evaluated approximately with the conjugate gradient method [92]. Another way is to use the quasi-Newton methods such as Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method [86], in which the inverse Hessian matrix is approximated with the model and the gradient in the previous iterations. Besides the above methods, novel optimization methods can be applied for the case when specific constraints and regularizations are introduced to the FWI problem, such as primal-dual hybrid gradient (PDHG) method [55, 140], fast iterative shrinkage-thresholding algorithm (FISTA) [2], etc. Some popular optimization methods for the FWI problem are reviewed in Chapter 1.

The computational requirements depend primarily on the numerical simulation of the constraint PDE in

the FWI problem. In practice, the objective function is constrained by several PDEs, and each PDE corresponds to a seismic source located at different positions. To reduce the computation cost, a simultaneous-source (or phase-encoding) strategy can be introduced [68, 77, 130, 101]. The basic idea is: instead of solving the PDE one by one, multiple sources are added together and simulated as one PDE [5]. In this case, the number of forward modeling simulations can be largely reduced for each iteration during the inversion. Specifically for the truncated Newton method, another way to decrease the computational cost is the preconditioning technique for solving the update direction with the conjugate gradient method [98, 92, 102].

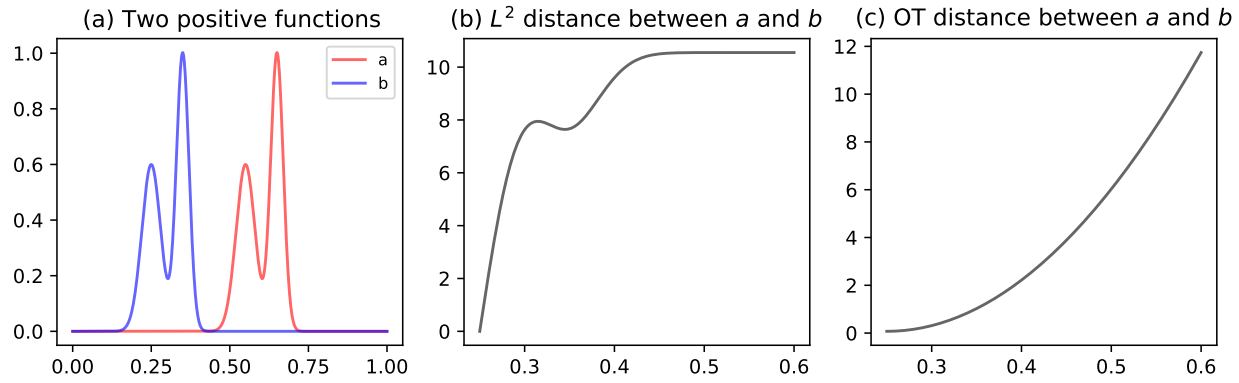


Figure 1.1: (a): Two positive functions $a$ and $b$. The position of $b$ is fixed. (b): $L^2$ distance between $a$ and $b$ as $a$ is shifted from left to right. The cycle-skipping issue exists at the local minimum near 0.35. (c): Optimal transport distance between $a$ and $b$ as $a$ is shifted from left to right. Only a global minimum exists.

Among all of the problems of the FWI technique, one of the fundamental problems is: can we get the "right" inverse result? From the optimization point of view, because of the intrinsic non-convexity of the objective function, only the local minimum can be guaranteed with the optimization methods in the deterministic approach. In this case, the reasonable goal is how can we achieve a local minimum inverse result that is close to the global minimum. There are two kinds of problems caused by the non-convexity that are widely discussed: the cycle-skipping issue and the parameter cross-talk issue for multi-parameter FWI. We focus on the single parameter FWI problem in this thesis work, and especially we developed optimization schemes based on the optimal transport distance to mitigate the cycle-skipping issue.

There are several ways to improve the inverse result. First, the multi-scale approach is developed to improve the inverse result of the FWI problem [25, 123, 23]. The multi-scale approach is based on the heuristic justifications when the difference between the seismic wavelets within half a wavelength, the nearest minimum in the objective function should be the global minimum. Then the converging area will be relatively large as the frequency of the seismic signal is relatively low. The multi-scale approach is to perform the inversion algorithm starting from the low-frequency components then move to the high-frequency components.

However, the seismic data is naturally bandlimited in practice and the low-frequency information is missing in the recorded data, which means the multi-scale approach can not fully eliminate the cycle-skipping issue.

Another way to mitigate the cycle-skipping issue is to enlarge the parameter space. In the work [131, 132], instead of introducing the constraint PDE with the parameter-to-state map, the author suggested using the constraint PDE as a penalty term in the objective function. In this case, the optimization variable from $u$ extends to $(y, u)$. With the name Wavefield Reconstruction Inversion (WRI), this technique has been successfully applied to the full waveform inversion problem [135, 1]. Other extension methods are developed, such as the extended source approach [72, 70, 71, 127].

Special objective functions can be designed for the FWI problem. For example, the normalized integration method (NIM) [87], and the integral wavefields misfit functional [65] have been used. Recently, the optimal transport (OT) distance or named Wasserstein distance has been introduced to describe the difference between seismic signals [50], and later it has been applied to the seismic imaging [52, 141] and the FWI problem [53]. Although it requires certain prerequisites such as mass conservation and normalization, the OT distance has the convexity property with respect to shift, dilation, and amplitude change in signals [53] which is one of the main motivations for introducing OT distance to the FWI problem. For example as shown in Figure 1.1.

The optimal transport distance is designed to describe the difference between two positive finite measures with equal total mass, and this is denoted as the mass equality limitation. Despite the promising properties, the seismic signal is oscillating around 0 and usually, the condition of equal mass is not satisfied. Several works have been proposed to overcome those restrictions and integrate the OT distance to the FWI problem. In the first strategy, the non-negative and equal mass restrictions are overcame by connecting the 1-Wasserstein distance to the KR norm [22] with the dual form of the Kantorovich problem. And then the distance is computed by a proximal splitting strategy called the simultaneous descent method of multipliers (SDMM) [94, 93]. In [142], the 1-Wasserstein distance is evaluated through the dynamic formulation [13] and then solved by a primal-dual hybrid gradient method (PDHG) with line search method. Another strategy is to normalize the seismic signals into positive functions with equal mass first, then compute the OT distance. In [113, 114, 138, 139], the seismic signals are normalized into positive functions with equal mass through normalization methods such as linear, quadratic, and exponential functions. Then the 2-Wasserstein distance between seismic signals can be evaluated either through a trace-by-trace technique or through the numerical computation of Monge-Ampère equation.

To overcome the mass equality limitation, the unbalanced optimal transport (UOT) problem is raised in [10] based on a dynamic approach. Later several works have been proposed in both static and dynamic approaches [107, 38, 39]. In Chapter 3 and Chapter 4, we introduce the UOT distance to the FWI problem

to overcome the mass equality limitation based on the work [38, 39]. And a new mixed $L^1$/Wasserstein distance is constructed and applied to the FWI problem.

Incorporating the a priori information of the model to the FWI problem can lead to better inversion results. Suppose for the grid point $x_0$ in the physical model, we know the average value of the ball $B(x_0, r)$ which centered at $x_0$ with radius $r$. Suppose for each of the grid points in the physical domain, we know this a priori average value information. As $r \to 0$, we actually have the values at all grid points. In other words, we already know the true solution of the inverse problem. This toy example suggests that a more accurate inverse result can be expected with more a priori information.

There are two equivalent ways to introduce the a priori information to the optimization problem: regularization and constraints. In the work [54, 104], the total variation (TV) constraint is used to reduce the cycle-skipping issues and build salt structures. Both box constraint and TV constraint are considered in [140, 55]. Also, adaptive regularization strategies are studied in [2]. In the work [105], the author developed an algorithm as a combination of the spectral projected gradients and Dykstra's algorithm, which can impose multiple constraints for the optimization problem. However, when the projection algorithm is evaluated inexactly, the update might be outside of the constraint sets and the constraint may not work in this case. This phenomena will be discussed in detail and a special constraint set expanding strategy is designed in Chapter 5 to solve this problem. In Chapter 6, a new optimization scheme is designed for incorporating multiple a priori information into the FWI problem.

## 1.2 Organization

The objective of this thesis is to develop novel optimization schemes that can provide better inverse results for the FWI problem. The thesis work can be divided into two parts:

- Introduce the optimal transport based distances to the FWI problem.
- Develop the optimization methods for incorporating multiple a priori information into the FWI problem.

The thesis is organized as follows:

Chapter 2 presents the background of the FWI problem. The acoustic approximation with constant density of the seismic propagation is used in this work, and the Born approximation is discussed which is the main reason for the non-convexity of the objective function. The formulation of the reduced form FWI problem is presented. To simulate the seismic wave propagation in an unbounded domain, the perfectly matched layer (PML) technique is used, and the acoustic wave equation with PML is derived which is a part

of the collaborative work [83]. The gradient of the objective function is derived through the adjoint state method. In the end, several popular optimization algorithms for the FWI problem are reviewed.

The background of the optimal transport problem is provided in Chapter 3. The Monge problem, Kantorovich problem, and dual Kantorovich problem are discussed. The convex properties of the 2-Wasserstein distance with respect to shift and dilation was proved with the Monge problem in the work [138, 139]. However, the Monge problem is not well-defined for comparing the discrete vectors. The above convex properties are proved with the Kantorovich problem in this chapter with the similar method. The definition of the unbalanced optimal transport (UOT) distance based on [38, 39] is reviewed. The numerical methods based on the entropy regularization of the optimal transport problem is reviewed, which is used for the evaluation of the UOT and the proposed mixed $L^1$/Wasserstein distance.

In Chapter 4, the optimal transport based distances are introduced to the FWI problem. A mixed $L^1$/Wasserstein distance is constructed, which inherits the convex properties of the 2-Wasserstein distance and overcomes the mass equality limitation. Normalization methods are discussed to introduce the UOT distance and the proposed distance to the seismic signals. The computation methods of the adjoint sources for both distances in the FWI problem are provided. Numerical examples including both the cross-well model and reflective wave model show that the UOT distance and the mixed distance outperform the conventional $L^2$ distance in certain cases. Compared with the current literature on the optimal transport distance and the FWI problem, this thesis work focuses on overcoming the mass equality limitation and computing through the entropy regularization approach. Parts of the work in this chapter is in the preprint work [82].

Chapter 5 focuses on the gradient projection methods with inexact projections. When projecting a point onto the intersection of several convex sets, a projection algorithm such as Dykstra's algorithm is generating a convergent sequence. This kind of algorithm has to be terminated after several iterations in practical usage, and this makes the projection process actually an inexact projection. When the constrained optimization problem is solved with the gradient projection methods, and the projection can only be evaluated with an inexact projection, the update points can not be guaranteed to be in the constraint set. We first review the set convergence and set-valued mapping results, then the convergence of the projection mapping sequence is analyzed. A set expanding strategy is developed for both gradient projection method and scaled gradient projection method with inexact projection. The convergence analyses are provided for both algorithms under several proper assumptions.

In Chapter 6, an optimization scheme is developed which is a combination of the scaled gradient projection method and the projection algorithm developed in [44, 45]. Multiple constraint sets including both the set that has the closed-form projection function and the set that has subgradient projection function can be incorporated in this scheme. The FWI problem is solved with the proposed optimization scheme. The

numerical examples are provided for both cross-well model and reflective wave model with TV constraint, $l_1$ constraint, box constraint, and hyperplane constraint. Compared to the work [104], the proposed method solves the inner projection problem with only one algorithm instead of two (Dykstra's algorithm and the ADMM method) when there is no closed-form projection function for the constraint set, which simplifies the structure of the optimization algorithm. The set expanding strategy developed in Chapter 5 is applied such that proper stopping criteria are available for the inner projection algorithm. The second-order information of the objective function can be involved with the L-BFGS Hessian approximation, which provides a faster convergence speed.

The contributions and innovations of this thesis work are summarized in Chapter 7. Future studies are discussed in the end.

## 1.3 Contributions

Contributions of this thesis are summarized as follows:

- Developing a methodology to introduce the unbalanced optimal transport (UOT) distance to the full waveform inversion (FWI) problem. Numerical examples show that the UOT distance provides more accurate inverse results compared with the conventional $L^2$ distance in certain cases.

- A mixed $L^1$/Wasserstein distance is constructed, which inherits the convex properties and overcomes the mass equality limitation of the optimal transport distance. The proposed distance is applied to the FWI problem, and similar results to the UOT distance can be achieved.

- A set expanding strategy is developed for the gradient projection methods when the inexact projection algorithm is used. The convergence results are proved under proper assumptions.

- An optimization scheme that can incorporate multiple a priori information as convex constraint sets is developed. This optimization scheme is applied to the FWI problem, and numerical examples show that the proposed scheme is sufficiently flexible to introduce multiple constraint sets at the same time with both closed-form projection and subgradient projection.

- All code used in this thesis work is developed by the author with the programming language Julia. The code can be found on the Github page: `https://github.com/zzar43`.

# Chapter 2

# Background of full waveform inversion problem

In this chapter, we provide the background material of the full waveform inversion (FWI) problem. First, we discuss the acoustic model of the reflective seismic wave and the Born approximation that explains the nonconvex behavior of the FWI problem. Then we formulate the FWI problem as an optimization problem. The gradient of the objective function can be evaluated through the adjoint state method, which is discussed in Section 2.3. Specific techniques and boundary conditions are required for numerically simulating the seismic wave propagating in an unbounded domain. We review the perfectly matched layer (PML) technique in Section 2.4, and the numerical scheme based on the finite difference method is discussed. In the end, several popular optimization methods are reviewed for the FWI problem.

Since this thesis focuses on designing novel numerical methods that can provide better inverse results compared with the conventional methods, we focus on the discrete optimization problem instead of building the theory in functional spaces. We start with the abstract function spaces setting for the convenience of describing the computation methods formally. FWI includes many components such as geophysics, scattering theory of wave equation, numerical methods for PDE, optimization methods, etc, and it is impossible to give a full description in one chapter. Only the fundamental contents required for this thesis work are reviewed in this chapter. We refer to review paper [134] and monograph [56] for a more detailed introduction.

## 2.1 Acoustic model of reflective seismic wave

Before the discussion of the inverse problem, we discuss the forward problem that characterizes the seismic propagation. The linear acoustic wave equation governs the wave propagation in the medium with small transient deformation, such as fluids and gases. Denote the time interval as $I = (0, T)$, $T > 0$. Consider a spatial domain $\Omega \subset \mathbb{R}^d$ where the seismic waves are propagating in, here $d = 1, 2, 3$. The acoustic wave equation in the time domain is given by,

$$\frac{1}{\rho(x)c(x)^2}\frac{\partial^2}{\partial t^2}y(x,t) - \nabla \cdot \left(\frac{1}{\rho(x)}\nabla y(x,t)\right) = s(x,t) \quad \text{in } \Omega \times I, \tag{2.1}$$

where $y(x,t)$ is the wavefield, $c(x)$ is the acoustic velocity, $\rho(x)$ represents the density, and $s(x,t)$ is the source term. When the density is homogeneous with $\rho(x) = 1$, from equation (2.1), we can have the scalar wave equation

$$\frac{1}{c(x)^2}\frac{\partial^2}{\partial t^2}y(x,t) - \Delta y(x,t) = s(x,t) \quad \text{in } \Omega \times I. \tag{2.2}$$

Since we focus on developing new methods to improve the results of the FWI problem, we only consider the simple case when the constraint PDE is the scalar wave equation.

Here we make a reasonable assumption that the spatial domain $\Omega$ is large enough such that the seismic waves never reach the boundary of the domain so that the Dirichlet boundary condition can be equipped for the convenience of analysis.

$$y(x,t) = 0 \quad \text{in } \partial\Omega \times I, \tag{2.3}$$

In the practical numerical simulation of the forward problem, special techniques are required to simulate the wave propagation in an unbounded domain, such as the perfectly matched layer (PML) which will be discussed later. Assume the medium is in an equilibrium state when $t < 0$, in this case we can assume the causal initial condition

$$y(x,0) = 0, \quad y_t(x,0) = 0 \quad \text{on } \Omega \times \{t = 0\}. \tag{2.4}$$

Also, the spatial size of the source is much smaller than the medium domain and the seismic wavelength. In

this case, a point source can be assumed as

$$s(x,t) = \tilde{s}(t)\delta(x_s), \tag{2.5}$$

where $\tilde{s}(t)$ is a function with respect to time only and $s(t) = 0$ as $t < 0$. We use $u$ instead of $c$ to represent the physical parameter that needs to be revealed through the optimization problem. When the multiple physical parameters are considered such as the acoustic wave equation, denote $u = (c, \rho)$, the equation (2.1) can be written in a compact form,

$$e(y,u) = L[u]y - s = 0, \tag{2.6}$$

where $L[u]$ is the linear differential operator depending on $u$.

The forward modeling problem can be described as: given the physical parameter $u(x)$, the source term, the initial and boundary condition of the system, compute the acoustic wavefield $y(x,t)$.

Since the above PDE system is well-posed, a parameter-to-state map can be defined as:

$$y = F(u). \tag{2.7}$$

Although the constraint PDE is linear, it is clear that the parameter-to-state map is not linear with respect to the parameter $u$, and it is natural to study the linearization of the parameter-to-state map.

Next, we discuss a formal linearization of the parameter-to-state map based on the scalar wave equation (2.2). Denote $c_1(x) = c_0(x) + \delta c(x)$, where $c_0$ is the reference velocity model and $\delta c(x)$ is the perturbation of the model. Denote $y_1(x,t) = y_0(x,t) + \delta y(x,t)$, where $\delta y(x,t)$ is the scattering wavefield. By equation (2.2),

$$\frac{1}{(c_0(x) + \delta c(x))^2}\frac{\partial^2}{\partial t^2}(y_0(x,t) + \delta y(x,t)) - \Delta(y_0(x,t) + \delta y(x,t)) = s(x,t). \tag{2.8}$$

Using the Tayler series of the term

$$\frac{1}{(c_0(x) + \delta c(x))^2} \approx \left(\frac{1}{c_0(x)^2} - \frac{2\delta c(x)}{c_0(x)^3}\right), \tag{2.9}$$

and equation (2.2), then

$$\frac{1}{c_0(x)^2}\frac{\partial^2}{\partial t^2}\delta y(x,t) - \Delta\delta y = \frac{2\delta c(x)}{c_0(x)^3}\frac{\partial^2}{\partial t^2}y_1(x,t). \tag{2.10}$$

Notice that the scattering wavefield $\delta y(x,t)$ is still implicitly in the right-hand side of the above equation. We can have the solution of $\delta y$ formally by the Green's function,

$$\delta y(x,t) = \int_I \int_\Omega G_0(x,y;t-\tau)\frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial \tau^2} y_1(y,\tau)\ \mathrm{d}y\mathrm{d}\tau, \tag{2.11}$$

where $G_0(x,y;t-\tau)$ is the Green's function with respect to the reference model, i.e.

$$\frac{1}{c_0(y)^2}\frac{\partial^2}{\partial t^2} y_0(x,t) - \Delta y_0(x,t) = s(x,t). \tag{2.12}$$

Denote the linear operator that arises by integration with $G_0(x,y;t-\tau)$ as $\mathcal{G}_0$, we obtain

$$\delta y(x,t) = \mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2} y_1(y,t). \tag{2.13}$$

Replace $\delta y(x,t) = y_1(x,t) - y_0(x,t)$, then,

$$y_0(x,t) = y_1(x,t) - \mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2} y_1(y,t), \tag{2.14}$$

which is named the Lippmann-Schwinger equation. Then we can formally have the following operator equation,

$$y_1(x,t) = \left(I - \mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2}\right)^{-1} y_0(x,t). \tag{2.15}$$

The following Born series can be achieved by the expansion of the Neumann series,

$$\begin{aligned} y_1(x,t) = y_0(x,t) &+ \left(\mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2}\right) y_0(x,t) \\ &+ \left(\mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2}\right)\left(\mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2}\right) y_0(x,t) + \cdots . \end{aligned} \tag{2.16}$$

When the perturbation $\delta c(y)$ is small enough such that $\|\mathcal{G}_0 \frac{2\delta c(y)}{c_0(y)^3}\frac{\partial^2}{\partial t^2}\| < 1$ in some norm, the Neumann series converges.

On the other hand, suppose the parameter-to-state map (2.7) is Fréchet differentiable we can have

$$y_1 = y_0 + \delta y = F(c_1) = F(c_0 + \delta c) = F(c_0) + DF[c_0]\delta c + D^2F[c_0](\delta c, \delta c) + \cdots , \tag{2.17}$$

It is easy to check that $DF[c_0]\delta c$ is formally equal to the first term of the Born series. Consider the scalar

wave equation (2.2) with the reference model, replace $y_0$ by $F(c_0)$,

$$\frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}F(c_0) - \Delta F(c_0) = s. \tag{2.18}$$

Then we compute the directional derivative with respect to $c$ in the direction $\delta c$ of the above equation,

$$\frac{-2\delta c}{c_0^3}\frac{\partial^2}{\partial t^2}F(c_0) + \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}DF[c_0]\delta c - \Delta DF[c_0]\delta c = 0. \tag{2.19}$$

With the Green's function defined above, it is easy to check that

$$DF[c_0]\delta c = \mathcal{G}_0\frac{2\delta c}{c_0^3}\frac{\partial^2}{\partial t^2}y_0. \tag{2.20}$$

When the linearization is accurate, we can have the approximation

$$\delta y \approx DF[c_0]\delta c, \tag{2.21}$$

which is known as the Born approximation. The accuracy of the Born approximation is the most important topic of the seismic imaging problem and seismic inverse problem. Here we quote the work [126]: based on the heuristic, physical reasoning, computational experience, the linearization relation (2.21) is well-approximated as long as

1. the reference model $c(x)$ is slowly-varying (smooth) relative to a typical data wavelength;
2. the perturbations $\delta c$ is oscillatory ("rough").

This scale-separation phenomenon explains the reason that the accurate initial (reference) model is crucial for the result of the seismic imaging and inversion problem.

We demonstrate this scale-separation phenomenon with the following two-dimensional numerical example. The reference model $c(x)$ is shown in Figure 2.1 (a), which is a part of the standard Marmousi model. Two approximations are considered as

$$c \approx c_1 + \delta c_1, \quad \text{and } c \approx c_2 + \delta c_2. \tag{2.22}$$

Here $c_1$ is a smoothed velocity model generated by the Gaussian filter and the reference model $c$, which is shown in Figure 2.1 (b). The perturbation $\delta c_1$ is shown in subfigure (c). The second approximation is achieved by letting $\delta c_2(x)$ be a constant, such that $\|\delta c_2\|_2 = \|\delta c_1\|_2$, which means that the perturbation $\delta c_1$

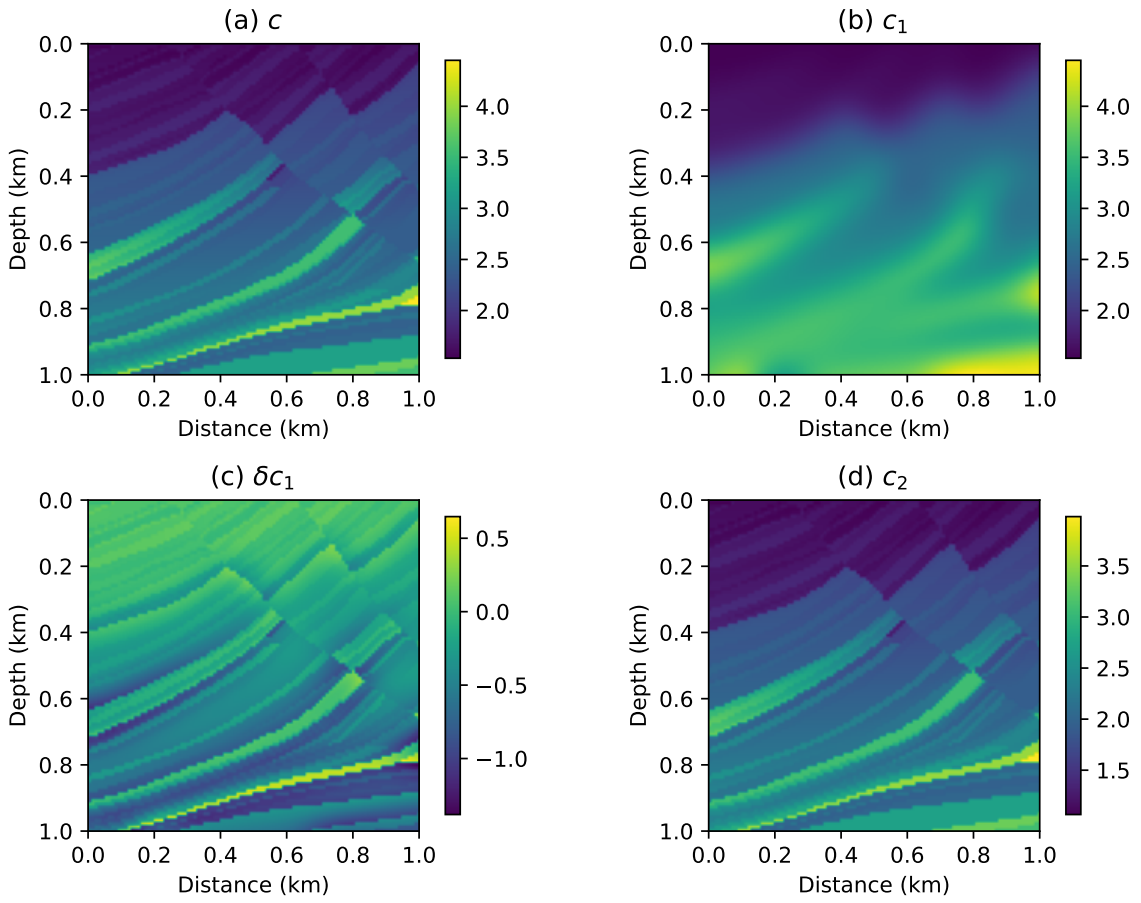


Figure 2.1: (a): The reference velocity model $c$. (b): The first approximated velocity model $c_1$. (c): The perturbation $\delta c_1$. (d): The second approximated velocity model $c_2$. Notice that the color scale is different to (a).

and $\delta c_2$ have the same $L^2$ norm. The velocity model $c_2$ is shown in Figure 2.1 (d). One 8 Hz Ricker wavelet is placed in the middle of the domain with a depth of 0.05 km. There are 101 equally spaced sources placed on the top of the domain.

Figure 2.2 shows the received data. As we can see, with the same $L^2$ norm perturbations $\delta c_1$ and $\delta c_2$, and $F(c_1) + DF[c_1]\delta c_1$ is close to $F(c)$, which means the Born approximation is accurate in this case. However, there are significant differences between $F(c)$ and $F(c_2) + DF[c_2]\delta c_2$.

The theoretical results on the accuracy of the Born approximation and the scale-separation phenomenon in the one-dimensional case can be found in [81]. To this author's best knowledge, there are no similar results for higher dimensional settings. More discussion on the linearization of the acoustic wave model can be found in [126]. We refer to the thesis work [125] for more information on the well-posedness and the smoothness of the parameter-to-state map, in which a linear hyperbolic equation is discussed with the coefficients are in

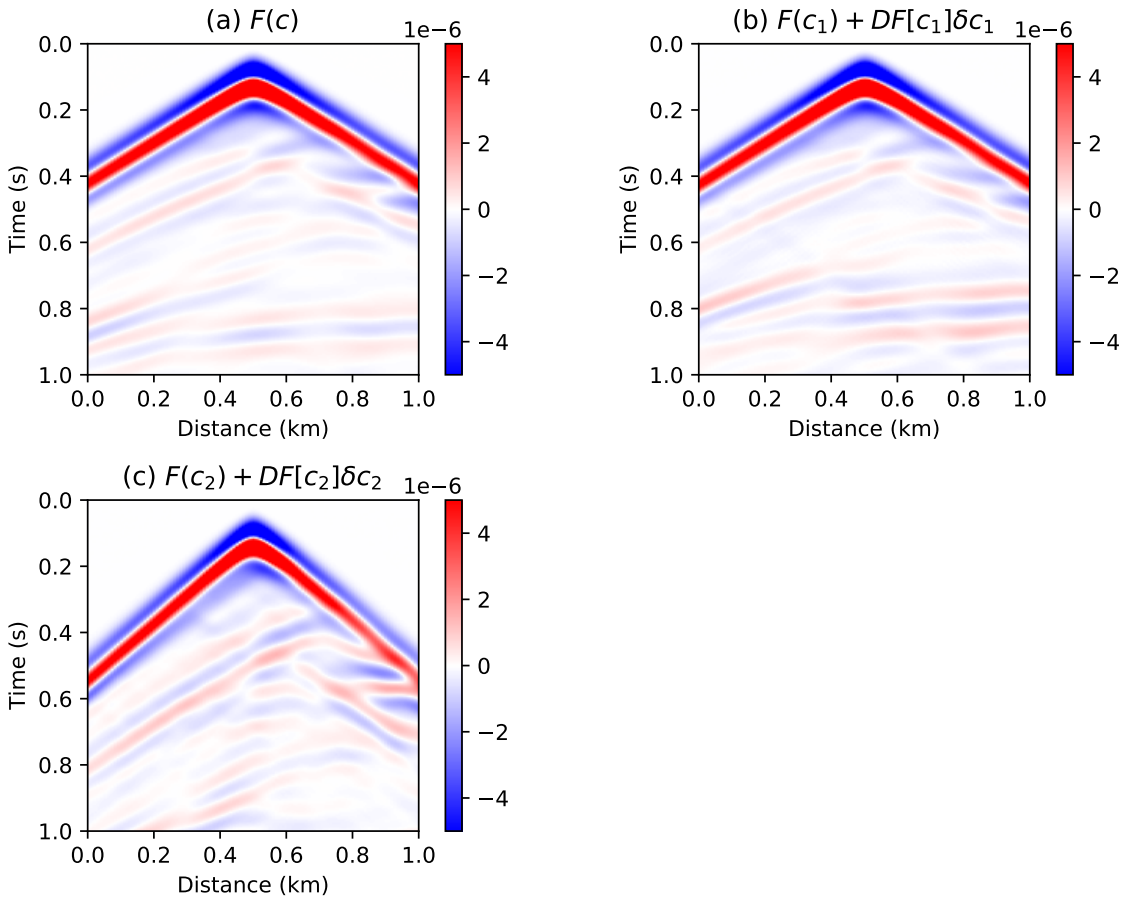


Figure 2.2: (a): Recorded wavefield $F(c)$ generated by the reference velocity model $c$. (b): The first order approximation with velocity model $c_1$ and perturbation $\delta c_1$. (c): The first order approximation with velocity model $c_2$ and perturbation $\delta c_2$.

$L^\infty$ (measurable and essentially bounded).

## 2.2 Formulation of the full waveform inversion

We discuss the inverse problem in this section by formulating the full waveform inversion (FWI) problem as a PDE constrained optimization problem. FWI can also be illustrated as a parameter estimation problem, since the objective of the FWI problem is to estimate the physical parameters of the underground medium such as acoustic velocity, density, Lamé parameters, etc. The seismic data generated by the artificial seismic sources or the natural seismic event is given at first. An initial model that is an estimation of the underground physical parameters is provided. The FWI problem is to minimize the distance between the simulated seismic data generated by the initial (and updated) model and the received data.

With the scalar wave equation as the constraint PDE and the $L^2$ distance is used in the objective function,

the FWI problem can be written as:

$$\min_{c} J(y,c) = \frac{1}{2}\int_I \int_\Omega \left(Qy(x,t) - y_d(x,t)\right)^2 \, \mathrm{d}x\mathrm{d}t + \lambda_r \mathcal{R}(c), \tag{2.23}$$

such that,

$$\begin{aligned} \frac{1}{c(x)^2}\frac{\partial^2}{\partial t^2} y(x,t) - \Delta y(x,t) &= s(x,t) \quad \text{in } \Omega \times I, \\ y(x,0) = 0, \quad y_t(x,0) &= 0 \quad \text{on } \Omega \times \{t=0\}, \\ y(x,t) &= 0 \quad \text{in } \partial\Omega \times I. \end{aligned} \tag{2.24}$$

Here $y_d(x,t)$ is the recorded data, $Q$ is the recording operator that maps the forward modeling seismic wavefield to the recorded seismic signal. The $\mathcal{R}(c)$ is a regularization term, and $\lambda_r$ is the regularization parameter. In the objective function, the $c(x)$ is implicitly contained in the forward modeling wavefield $y(x,t)$.

The constraint PDE is not limited to the scalar wave equation. When the constraint PDE is the acoustic wave equation (2.1), we minimize both the velocity $c(x)$ and the density $\rho(x)$ in the objective function. More sophisticated physical models can be considered. For example, consider the linear elastic wave equation of isotropic medium, the constraint PDE is given by

$$\rho(x)\frac{\partial^2}{\partial t^2} y(x,t) - \nabla(\lambda(x)\nabla \cdot y(x,t)) - \nabla \cdot (\mu(x)(\nabla y(x,t) + (\nabla y(x,t))^{'})) = s(x,t), \tag{2.25}$$

here the displacement $y(x,t)$ is a time-dependent vector field. The $\lambda(x)$ and $\mu(x)$ are the Lamé parameters. In this case, we optimize the triples $(\rho(x), \lambda(x), \mu(x))$ in the objective function. Also, other parameterizations are available, for example, consider the p-wave velocity and s-wave velocity as

$$c_p = \sqrt{\frac{\lambda + 2\mu}{\rho}}, \quad c_s = \sqrt{\frac{\mu}{\rho}}. \tag{2.26}$$

For convenience, we rewrite the problem (2.23) and (2.24) in a compact form with abstract function space setting. Let $Y$ and $U$ be the spaces representing seismic wavefield and physical parameters. Instead of working on the specific physical parameter, denote $u$ as the parameter we are looking for. Suppose there is no regularization term in the objective function, then the FWI problem can be formulated formally as

$$\begin{aligned} \min_{(y,u)\in Y\times U_{\text{ad}}} J(y,u) &= \frac{1}{2}\|Qy - y_d\|_Y^2, \\ \text{such that } e(y,u) &= 0. \end{aligned} \tag{2.27}$$

Here, $y_d$ is the received data, $Q : Y \to Y$ is the observation operator that maps the seismic wavefield $y$ to the seismic signals recorded by the receivers. The set $U_{\text{ad}} \subset U$ is the feasible set. The constraint PDE is written in a compact form with $e : Y \times U \to Z$. For example, consider the scalar wave equation with initial and boundary condition in (2.24), then,

$$\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{u^2} y_{tt} - \Delta y - s \\ y(\cdot, 0) \\ y_t(\cdot, 0) \\ y(x,t)|_{x \in \partial\Omega} \end{pmatrix} = 0. \tag{2.28}$$

Since the initial and boundary condition can be discussed separately, we use another compact form for the following computation,

$$e(y, u) = L(u)y - s = \frac{1}{u^2} y_{tt} - \Delta y - s = 0, \tag{2.29}$$

where $L(u)$ is the linear differential operator determined by $u$, and $s$ is the source term.

The optimization problem (2.27) has the same form as the optimal control problem, where $y$ is denoted as the state parameter, and $u$ is denoted as the control parameter. Since the constraint PDE is well-posed in our case, a parameter-to-state map can be defined as $y = F(u)$. Instead of solving problem (2.27) with respect to both $y$ and $u$, a reduced problem can be considered

$$\min_{u \in U} f(u) = J(F(u), u), \quad \text{such that } u \in U_{\text{ad}}. \tag{2.30}$$

When $U_{\text{ad}} = U$, this is an unconstrained optimization problem, otherwise it is a constrained optimization problem.

The objective function $f(u)$ is smooth, nonlinear, and nonconvex. The non-convexity of $f(u)$ follows that the parameter-to-state map $y = F(u)$ is nonlinear as discussed in the previous section. Loosely speaking, when the initial model is close to the true model and the difference between the initial model and the true model is mainly oscillatory structures, the linearization of $y = F(u)$ is nearly accurate. Then, the objective function $f(u)$ around the initial model is nearly convex, which means that the solution of the inverse problem is more likely to be "close" to the true solution. On the contrary, when the initial model is far from the true model, or the difference between the initial model and the true model is mainly large-scale structures, the linearization of $y = F(u)$ is not accurate. Then, the optimization algorithm will be trapped in a local minimum. The typical size of the FWI problem is relatively large and the evaluation of the objective function

is relatively expensive since a PDE system needs to be solved numerically, only the deterministic optimization algorithm can be implemented. In this case, the local minimum problem is unavoidable. The initial idea of this thesis research work is to find a way such that the inverse results (local minima) are close to the true solution (global minimum). In this thesis work, we focus on developing new algorithms for the FWI problem, which provides better results compared to the conventional methods. We focus on the scalar wave equation as the constraint PDE, and we work on the reduced problem (2.30).

Notice that, instead of solving the optimization in the reduced form, the penalty approach can also be considered which transform the constraint PDE as a penalty term in the objective function [131, 132]. It has been successfully implemented in solving the FWI problem with the name Wavefield Reconstruction Inversion (WRI). For the precise analysis work on the seismic inverse problem, we refer to [20]. The work [21] provides a detailed function space setting for the FWI problem with the elastic wave equation, which includes a wide range of the choice of constraint PDE for the FWI problem.

## 2.3 Adjoint state method

In this section, we review the adjoint state method for formally evaluating the first-order derivative of the objective function $f(u)$ in equation (2.30). The adjoint state method is a technique that evaluates the gradient of a function efficiently. It was developed from the control theory [85] and then was applied to the inverse problem [36]. In the early 80s, the adjoint state method was introduced to the exploration geophysics by the work [78] and [128] as an important component of the FWI problem. For a complete review of the adjoint state method in the seismic inverse problem, we refer to [109].

There are different equivalent approaches to derive the adjoint state method. The Lagrangian approach is used in this section, and for the sensitivity approach and the adjoint approach for the general inverse problem, we refer to [69]. We start the derivation based on the abstract form of the FWI problem (2.27) for convenience. Let $Y$, $U$, $Z$ be Banach spaces, define the Lagrangian function: $\mathcal{L} : Y \times U \times Z^* \to \mathbb{R}$ as

$$\mathcal{L}(y, u, v) = J(y, u) + \langle v, e(y, u) \rangle_{Z^*, Z}\,, \tag{2.31}$$

where $J(y, u)$ is the objective function in (2.27). Since the parameter-to-state map $y = F(u)$ is well defined, and $e(F(y), u) = 0$, the reduced problem is

$$f(u) = J(F(u), u) + \langle v, e(F(u), u) \rangle_{Z^*, Z} = \mathcal{L}(F(u), u, v). \tag{2.32}$$

Then the derivative of $f(u)$ can be written as

$$\langle f(u)', \delta u\rangle_{U^*,U} = \langle \mathcal{L}_y(F(u), u, v), DF[u]\delta u\rangle_{Y^*,Y} + \langle \mathcal{L}_u(F(u), u, v), \delta u\rangle_{U^*,U}\,. \tag{2.33}$$

The idea of the adjoint state method is to find an adjoint state $v(u)$ that is depending $u$ such that $\mathcal{L}_y(F(u), u, v) = 0$. Notice that,

$$\begin{aligned}\langle \mathcal{L}_y(y, u, v), \delta y\rangle_{Y^*,Y} &= \langle J_y(y, u), \delta y\rangle_{Y^*,Y} + \langle v, e_y(y, u)\delta y\rangle_{Z^*,Z} \\ &= \langle J_y(y, u) + e_y(y, u)^* v, \delta y\rangle_{Y^*,Y}\,.\end{aligned} \tag{2.34}$$

Then we can have the following adjoint equation in an abstract form,

$$\mathcal{L}_y(F(u), u, v) = J_y(F(u), u) + e_y(F(u), u)^* v = 0. \tag{2.35}$$

Given the state parameter $u$, compute the adjoint state parameter $v(u)$ with the above adjoint equation, the derivative of $f(u)$ can be written as

$$f'(u) = \mathcal{L}_u(F(u), u, v(u)) = J_u(F(u), u) + e_u(F(u), u)^* v(u). \tag{2.36}$$

Next, we discuss the special case when the constraint PDE is the scalar wave equation, which can be written as

$$e(y, u) = \frac{1}{u(x)^2}\frac{\partial^2}{\partial t^2} y(x, t) - \Delta y(x, t) - s(x, t) = 0. \tag{2.37}$$

Compute the derivative of $e(y, u)$ as

$$e_y(y, u)\delta y = \left(\frac{1}{u(x)^2}\frac{\partial^2}{\partial t^2} - \Delta\right)\delta y(x, t), \tag{2.38}$$

$$e_u(y, u)\delta u = \frac{-2\delta u(x)}{u(x)^3}\frac{\partial^2}{\partial t^2} y(x, t). \tag{2.39}$$

Recall that $y = F(u)$, consider

$$\begin{aligned}
\langle e_y(F(u),u)^* v, \delta y\rangle_{Y^*,Y} &= \langle v, e_y(F(u),u)\delta y\rangle_{Z^*,Z} \\
&= \int_I \int_\Omega v(x,t)\left(\frac{1}{u(x)^2}\frac{\partial^2}{\partial t^2} - \Delta\right)\delta y(x,t)\ \mathrm{d}x\mathrm{d}t \\
&= \int_I \int_\Omega \delta y(x,t)\left(\frac{1}{u(x)^2}\frac{\partial^2}{\partial t^2} - \Delta\right) v(x,t)\ \mathrm{d}x\mathrm{d}t \\
&\quad + \int_\Omega v(x,t)\left(\frac{\partial}{\partial t}\delta y(x,t) - \delta y(x,t)\frac{\partial}{\partial t}v(x,t)\right)\Big|_0^T\ \mathrm{d}x \\
&\quad + \int_I \int_{\partial\Omega} v(x,t)\frac{\partial \delta y(x,t)}{\partial \vec{\nu}} - \delta y(x,t)\frac{\partial v(x,t)}{\partial \vec{\nu}}\ \mathrm{d}S\mathrm{d}t.
\end{aligned} \tag{2.40}$$

The last equation follows from integrating by parts and Green's formula.

By the assumption in the previous sections, the source term is concentrating on a point and the domain $\Omega$ is large enough such that the wavefields never reach the boundary. Also, the initial condition is in an equilibrium status. We can say that,

$$\delta y(x,t) = \frac{\partial \delta y(x,t)}{\partial \vec{\nu}} = 0, \quad \forall x \in \partial\Omega. \tag{2.41}$$

With the initial condition, we have,

$$\delta y(x,t) = \frac{\partial}{\partial t}\delta y(x,t) = 0, \quad \text{as } t = 0. \tag{2.42}$$

Also, we can assume that

$$v(x,t) = \frac{\partial}{\partial t}v(x,t) = 0, \quad \text{as } t = T, \tag{2.43}$$

and

$$v(x,t) = \frac{\partial v(x,t)}{\partial \vec{\nu}} = 0, \quad \forall x \in \partial\Omega. \tag{2.44}$$

Then the equation (2.40) is

$$\langle e_y(F(u),u)^* v, \delta y\rangle_{Y^*,Y} = \int_I \int_\Omega \delta y(x,t)\left(\frac{1}{u(x)^2}\frac{\partial^2}{\partial t^2} - \Delta\right) v(x,t)\ \mathrm{d}x\mathrm{d}t. \tag{2.45}$$

By equation (2.35), the adjoint equation can be written explicitly as

$$\frac{1}{u(x)^2}v(x,t) - \Delta v(x,t) = -J_y(F(u),u), \tag{2.46}$$

where the adjoint source is

$$-J_y(F(u),u) = -Q^*\left(QF(u) - y_d\right), \tag{2.47}$$

by equation (2.27). The boundary condition is given by equation (2.44), and the initial condition is given by equation (2.43). Notice that the initial condition is given when the time is $t = T$, which is actually the final status of the forward modeling system. The adjoint equation (2.46) should be solved with time-reversed.

Since there is no regularization term in equation (2.27), we have $J_u(F(u),u) = 0$. Also,

$$\begin{aligned}\langle e_u(F(u),u)^* v, \delta u\rangle_{U^*,U} &= \langle v, e_u(F(u),u)\delta u\rangle_{Z^*,Z} \\ &= \int_I\int_\Omega v(x,t)\frac{-2\delta u(x)}{u(x)^3}\frac{\partial^2}{\partial t^2}y(x,t)\ \mathrm{d}x\mathrm{d}t \\ &= \int_\Omega \delta u(x)\int_I v(x,t)\frac{-2}{u(x)^3}\frac{\partial^2}{\partial t^2}y(x,t)\mathrm{d}t\mathrm{d}x.\end{aligned} \tag{2.48}$$

When there is no regularization term, by equation (2.36), the derivative of the objective function can be given as

$$f'(u) = \int_I v(x,t)\frac{-2}{u(x)^3}\frac{\partial^2}{\partial t^2}y(x,t)\mathrm{d}t. \tag{2.49}$$

For the summary, the adjoint state method for computing the derivative $f'(u)$ can be written as:

1. Given model $u(x)$, evaluate $F(u)$ by computing the scalar wave equation (2.24).
2. Compute the adjoint wavefield $v(x,t)$ by solving the adjoint equation (2.46).
3. Compute the derivative $f'(u)$ with the equation (2.49).

## 2.4 Forward modeling with perfectly matched layer

To solve the FWI problem numerically, the "first discretize, then optimize" approach is used in this work. All the quantities in equation (2.23) and (2.24) are discretized at first. Then the FWI problem given by (2.27) and (2.30) turns into a finite-dimensional optimization problem. Meanwhile, the constraint PDE can be evaluated numerically with the same discretization scheme.

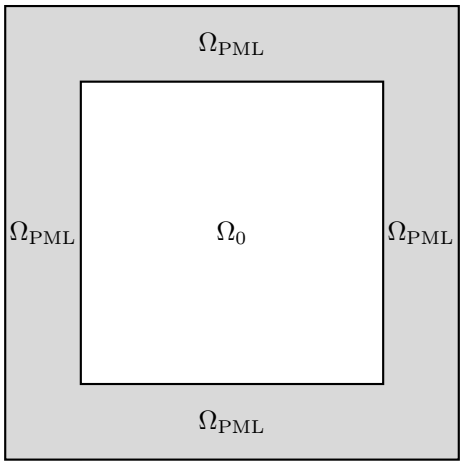


Figure 2.3: The computation domain with PML: $\Omega = \Omega_{\text{PML}} \cup \Omega_0$, citing from [83].

We assume the spatial domain $\Omega$ be large enough such that the wavefield never reaches the boundary $\partial\Omega$. However, this is infeasible for the practical algorithm since the enlargement of the domain will largely increase the computation time of the numerical simulation for the PDE system. As discussed in the previous sections, the numerical solutions of the scalar wave equations are needed for both the evaluation of the objective function $f(u)$ and the evaluation of the derivative $f'(u)$, and this is the main bottleneck of the numerical methods for the FWI problem. Two special techniques can be applied to simulate the wave propagating with a bounded domain effectively and efficiently: absorbing boundary condition (ABC) [41, 51] and perfectly matched layer (PML) [16]. The PML technique is used in this work.

The PML technique is to add an artificial absorbing layer for the PDE in the spatial domain, such that the energy of the seismic wave can be approximately reduced to 0 as propagating in the absorbing layer. Denote the absorbing layer as $\Omega_{\text{PML}}$ and the interior domain as $\Omega_0$, then the computation domain $\Omega = \Omega_{\text{PML}} \cup \Omega_0$. Figure 2.3 provides a demonstration of the computation domain with PML in a two-dimensional spatial space.

It can be illustrated with a one-dimensional sinusoidal wave propagation: $e^{\mathrm{i}(kx-\omega t)}$. By the analytic continuation, we extend the domain to the complex plane with

$$x \to x + \frac{\mathrm{i}}{\omega} \int_0^x \sigma(x')\ \mathrm{d}x', \tag{2.50}$$

where the damping function $\sigma(\cdot)$ controls the attenuation of the sinusoidal wave. The damping function $\sigma(x) > 0$ when $x$ in the absorbing layer, and $\sigma(x) = 0$ when $x$ in the interior domain. Then we can have,

$$e^{\mathrm{i}(kx-\omega t)} \to e^{\mathrm{i}(kx-\omega t)-\frac{k}{\omega}\int_0^x \sigma(x')\ \mathrm{d}x'}. \tag{2.51}$$

It can be seen that the amplitude of the sinusoidal wave is attenuated in the area when $\sigma(x) \neq 0$. Popular choices of $\sigma(x)$ including the linear damping function

$$\sigma(x) = \begin{cases} \sigma_0 x, & \text{if } x \in \Omega_{\text{PML}}, \\ 0, & \text{if } x \in \Omega_0, \end{cases} \tag{2.52}$$

and the inverse distance damping function

$$\sigma(x) = \begin{cases} \frac{\sigma_0}{x}, & \text{if } x \in \Omega_{\text{PML}}, \\ 0, & \text{if } x \in \Omega_0. \end{cases} \tag{2.53}$$

here the $\sigma_0$ is the coefficient to control the attenuation of the absorbing layer.

Based on the analytic continuation (2.51), the following transformation of the differential operator can be used to apply the PML technique to the constraint PDE:

$$\frac{\partial}{\partial x} \to \frac{1}{1 + \frac{\mathrm{i}\sigma(x)}{\omega}} \frac{\partial}{\partial x}. \tag{2.54}$$

Next, we demonstrate the PML technique with the acoustic wave equation (2.1) in a three-dimensional spatial domain. This is based on the work [83] by the author and collaborators. For convenience, rewrite the acoustic wave equation with spatial parameter $(x, y, z)$ as

$$\frac{1}{\rho c^2} \frac{\partial^2}{\partial t^2} u - \nabla \cdot \left( \frac{1}{\rho} \nabla u \right) = s, \tag{2.55}$$

where $\rho(x, y, z)$ and $c(x, y, z)$ are the density function and acoustic velocity function defined, $u(x, y, z, t)$ represents the wavefield, $s(x, y, z, t)$ represents the source term. Replace the spatial differential operator by

$$\frac{\partial}{\partial x} \to \frac{1}{1 + \frac{\mathrm{i}\sigma(x)}{\omega}} \frac{\partial}{\partial x} = \frac{1}{\eta_x} \frac{\partial}{\partial x}, \tag{2.56}$$

$$\frac{\partial}{\partial y} \to \frac{1}{1 + \frac{\mathrm{i}\sigma(y)}{\omega}} \frac{\partial}{\partial y} = \frac{1}{\eta_y} \frac{\partial}{\partial y}, \tag{2.57}$$

$$\frac{\partial}{\partial z} \to \frac{1}{1 + \frac{\mathrm{i}\sigma(z)}{\omega}} \frac{\partial}{\partial z} = \frac{1}{\eta_z} \frac{\partial}{\partial z}. \tag{2.58}$$

Then the equation (2.55) turns into

$$\begin{aligned}
&\eta_x\eta_y\eta_z\frac{1}{\rho c^2}\frac{\partial^2}{\partial t^2}u \\
&-\left[\left(\frac{\partial}{\partial x}\frac{1}{\rho}\right)\left(\frac{\eta_y\eta_z}{\eta_x}\frac{\partial}{\partial x}u\right)+\left(\frac{\partial}{\partial y}\frac{1}{\rho}\right)\left(\frac{\eta_x\eta_z}{\eta_y}\frac{\partial}{\partial y}u\right)+\left(\frac{\partial}{\partial z}\frac{1}{\rho}\right)\left(\frac{\eta_x\eta_y}{\eta_z}\frac{\partial}{\partial z}u\right)\right] \\
&-\left[\frac{1}{\rho}\frac{\partial}{\partial x}\left(\frac{\eta_y\eta_z}{\eta_x}\frac{\partial}{\partial x}u\right)+\frac{1}{\rho}\frac{\partial}{\partial y}\left(\frac{\eta_x\eta_z}{\eta_y}\frac{\partial}{\partial y}u\right)+\frac{1}{\rho}\frac{\partial}{\partial z}\left(\frac{\eta_x\eta_y}{\eta_z}\frac{\partial}{\partial z}u\right)\right]=\eta_x\eta_y\eta_z s.
\end{aligned} \tag{2.59}$$

Consider the temporal derivative term in the above equation,

$$\begin{aligned}
\eta_x\eta_y\eta_z\frac{1}{\rho c^2}\frac{\partial^2}{\partial t^2}u &= \frac{1}{\rho c^2}\left(1+\frac{\sigma_x}{\mathrm{i}\omega}\right)\left(1+\frac{\sigma_y}{\mathrm{i}\omega}\right)\left(1+\frac{\sigma_z}{\mathrm{i}\omega}\right)\frac{\partial^2}{\partial t^2}u \\
&= \frac{1}{\rho c^2}\left(\frac{\partial^2}{\partial t^2}u+(\sigma_x+\sigma_y+\sigma_z)\frac{\partial}{\partial t}u+(\sigma_x\sigma_y+\sigma_x\sigma_z+\sigma_y\sigma_z)u+\sigma_x\sigma_y\sigma_z\frac{u}{\mathrm{i}\omega}\right).
\end{aligned} \tag{2.60}$$

For the spatial derivative term along $x$-direction in equation (2.59),

$$\begin{aligned}
\frac{\eta_y\eta_z}{\eta_x}\frac{\partial}{\partial x}u &= \frac{\left(1+\frac{\sigma_y}{\mathrm{i}\omega}\right)\left(1+\frac{\sigma_z}{\mathrm{i}\omega}\right)}{\left(1+\frac{\sigma_x}{\mathrm{i}\omega}\right)}\frac{\partial}{\partial x}u \\
&= \frac{-\sigma_x+\sigma_y+\sigma_z+\frac{\sigma_y\sigma_z}{\mathrm{i}\omega}}{\mathrm{i}\omega+\sigma_x}\frac{\partial}{\partial x}u+\frac{\partial}{\partial x}u := v_x+\frac{\partial}{\partial x}u.
\end{aligned} \tag{2.61}$$

For the variable $v_x$, one has

$$\mathrm{i}\omega v_x+\sigma_x v_x(\sigma_x-\sigma_y-\sigma_z)\frac{\partial}{\partial x}u-\sigma_y\sigma_z\frac{\partial}{\partial x}\left(\frac{u}{\mathrm{i}\omega}\right)=0. \tag{2.62}$$

By the Fourier transform, we have $\mathrm{i}\omega v_x=\frac{\partial}{\partial t}v_x$. The same computation can be carried for the rest two directions. Let $w=\frac{u}{\mathrm{i}\omega}$, we have $\frac{\partial}{\partial t}w=u$ by Fourier transform. Together with the above equations, the acoustic wave equation (2.55) can be transformed as

$$\begin{aligned}
\frac{1}{\rho c^2}\left(\frac{\partial^2}{\partial t^2}u+\alpha\frac{\partial}{\partial t}u+\beta u+\gamma w\right)-\nabla\cdot\left(\frac{1}{\rho}(\vec{v}+\nabla u)\right)&=s, \\
\frac{\partial}{\partial t}\vec{v}+A\vec{v}+B\nabla u-C\nabla w&=0, \\
\frac{\partial}{\partial t}w-u&=0,
\end{aligned} \tag{2.63}$$

where $\vec{v} = (v_x, v_y, v_z)$, $\alpha = \sigma_x + \sigma_y + \sigma_z$, $\beta = \sigma_x\sigma_y + \sigma_x\sigma_z + \sigma_y\sigma_z$, $\gamma = \sigma_x\sigma_y\sigma_z$, and

$$\begin{aligned} A &= \begin{pmatrix} \sigma_x & & \\ & \sigma_y & \\ & & \sigma_z \end{pmatrix}, \quad B = \begin{pmatrix} \sigma_y\sigma_z & & \\ & \sigma_x\sigma_z & \\ & & \sigma_x\sigma_z \end{pmatrix}, \\ C &= \begin{pmatrix} \sigma_x - \sigma_y - \sigma_z & & \\ & \sigma_y - \sigma_x - \sigma_z & \\ & & \sigma_z - \sigma_x - \sigma_y \end{pmatrix}. \end{aligned} \tag{2.64}$$

Replacing the partial differential operator with the finite difference operator, the standard finite difference scheme can be applied for the above equations. Higher spatial and temporal accuracy schemes can also be designed [83].

## 2.5 Optimization algorithms for full waveform inversion problem

As discussed in the previous sections, the FWI problem can be discretized and solved in a reduced form:

$$\min_{u \in \mathbb{R}^n} f(x), \quad \text{such that } x \in C. \tag{2.65}$$

For the two-dimensional or three-dimensional case, the physical parameter can be reshaped into a $n$-dimensional vector. The objective function $f(x)$ is nonlinear and nonconvex, $C$ is a convex set, and the gradient of $f(x)$ can be achieved through the adjoint state method as discussed before. Problem (2.65) is a constrained optimization problem when $C$ is a subset of $\mathbb{R}^n$, otherwise, it is an unconstrained optimization problem.

In this section, we review some popular optimization methods for the FWI problem, especially for the reduced problem (2.65). A convergent sequence $\{x^k\}$ is generated by the optimization method for $k \in \mathbb{N}$ that converges to a local minimum of the objective function $f(x)$. The solution $x_0$ satisfies the first-order optimality condition, i.e.

$$\nabla f(x_0) = 0, \tag{2.66}$$

when $C = \mathbb{R}^n$, and

$$\langle \nabla f(x_0), x - x_0 \rangle \geq 0, \quad \forall x \in C, \tag{2.67}$$

when $C$ is a subset of $\mathbb{R}^n$. Instead of the convergence results of the optimization methods, we only focus on the implementation in this section.

The steepest descent method is one of the most commonly used methods for the unconstrained optimization problem. Suppose at the $k$-th iteration, the update direction is represented as $\delta x^k$, the negative of the gradient is used for the update direction:

$$\delta x^k = -\nabla f(x^k). \tag{2.68}$$

Then the update step can be written as

$$x^{k+1} = x^k + \alpha^k \delta x^k, \tag{2.69}$$

where the step size $\alpha^k$ is achieved through the line search algorithms, such as Armijo condition

$$f(x^k + \alpha^k \delta x^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)' \delta x^k, \tag{2.70}$$

or the Wolfe conditions

$$f(x^k + \alpha^k \delta x^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)' \delta x^k, \tag{2.71}$$

$$\nabla f(x^k + \alpha^k \delta x^k)' \delta x^k \geq c_2 \nabla f(x^k)' \delta x^k, \tag{2.72}$$

here $c_1$ and $c_2$ are the line search coefficients.

Another most commonly used gradient-based method is the nonlinear conjugate gradient (NCG) method. The update direction is defined as a linear combination of the gradient in the current iteration and the update direction in the previous iteration:

$$\delta x^k = -\nabla f(x^k) + \beta_k \delta x^{k-1}, \tag{2.73}$$

where $\beta_k$ is defined such that the $\delta x^k$ and $\delta x^{k-1}$ are conjugate. Popular choices of the scalar $\beta_k$ including the Fletcher-Reeves method [57]:

$$\beta_k^{\mathrm{FR}} = \frac{(\nabla f(x^k))' \nabla f(x^k)}{(\nabla f(x^{k-1}))' \nabla f(x^{k-1})}, \tag{2.74}$$

the Polak-Ribière method

$$\beta_k^{\mathrm{PR}} = \frac{(\nabla f(x^k))^{'}(\nabla f(x^k) - \nabla f(x^{k-1}))}{\|\nabla f(x^k)\|^2}, \tag{2.75}$$

and the Hestenes–Stiefel formula

$$\beta_k^{\mathrm{HS}} = \frac{(\nabla f(x^k))^{'}(\nabla f(x^k) - \nabla f(x^{k-1}))}{(\nabla f(x^k) - \nabla f(x^{k-1}))^{'}\delta x^k}. \tag{2.76}$$

For more discussion about the $\beta_k$ we refer to the textbook [99]. The Fletcher-Reeves method is used in this work for numerical experiments.

With the line search algorithm, the gradient-based methods are known to converge globally. However, the convergence speed is possibly very slow. Compared to the gradient-based methods, Newton's method provides faster convergence speed. The update direction of Newton's method is given by

$$\delta x^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k), \tag{2.77}$$

where $\nabla^2 f(x^k)$ is the Hessian matrix. A second-order adjoint state method is needed for the full Newton step update [92].

Recall the reduced objective function of the FWI problem:

$$f(x) = \frac{1}{2}\|QF(x) - y_d\|^2, \tag{2.78}$$

then the gradient can be given as

$$\nabla f(x) = DF[x]^* Q^*(QF(x) - y_d), \tag{2.79}$$

The Gauss-Newton method is to approximate full Hessian $\nabla^2 f(x)$ with

$$B = DF[x]^* Q^* Q DF[x]. \tag{2.80}$$

Here the $DF[x]$ and $DF[x]^*$ can be derived through the adjoint method, by differentiating $e(F(x), x) = 0$,

$$e_y(F(x), x)DF[x] + e_x(F(x), x) = 0, \tag{2.81}$$

where $e_y(F(x), x)$ and $e_x(F(x), x)$ are defined by equation (2.38) and (2.39). The $DF[x]^*$ follows the same

method in equation (2.48). Usually, the matrix-vector production between inverse Hessian and the gradient is evaluated in a Hessian free form. The update direction $\delta x^k$ is achieved approximately by solving the following linear system with a conjugate gradient method:

$$B\delta x^k = -\nabla f(x^k). \tag{2.82}$$

Also, preconditioning methods can be introduced for solving the above Newton equation [102]. Although Newton's method and the Gauss-Newton method provide fast convergence speed, huge numbers of the forward modeling are proceed in each of the iteration, which is expensive for the large-scale problem.

Compared to Newton's method and the Gauss-Newton method, quasi-Newton methods provide both the Hessian information and the efficient computation at the same time. At the $k$-th iteration, the objective function $f(x)$ is approximated by a quadratic form

$$f(x^k + \delta x^k) \approx f(x^k) + \nabla f(x^k)^{'} \delta x^k + \frac{1}{2}(\delta x^k)^{'} B_k \delta x^k, \tag{2.83}$$

where $B_k$ is a symmetric positive definite matrix approximating the Hessian $\nabla^2 f(x^k)$.

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) method is one of the most popular quasi-Newton methods. Denote the inverse Hessian matrix as $H_k = B_k^{-1}$, the $H_k$ is approximated by $x_k$, $\nabla f(x^k)$, and the Hessian approximation $H_{k-1}$ at the previous iteration [99]. The BFGS inverse Hessian approximation is given by

$$H_{k+1} = (I - \rho_k s_k y_k^{'}) H_k (I - \rho_k y_k s_k^{'}) + \rho_k s_k s_k^{'}, \tag{2.84}$$

where $I$ is the identity matrix, $s_k = x^{k+1} - x^k$, $y_k = \nabla f(x^{k+1}) - \nabla f(x^k)$, $\rho_k = 1/(y_k^{'} s_k)$. Then the update direction is given by

$$\delta x^k = -H_k \nabla f(x^k). \tag{2.85}$$

The Wolfe line search condition can be used for the update which can keep the BFGS Hessian approximation be symmetric positive definite during the iteration.

When the model $x$ is a $n$-dimensional vector, the size of the inverse Hessian matrix is $n \times n$. In most cases, it is unrealistic to save the inverse Hessian matrix $H_k$ explicitly due to the large-scale size of the FWI problem. A limited memory quasi-Newton method for the large-scale optimization problem has been designed in the work [86] with the name L-BFGS method. Instead of saving the full inverse Hessian matrix $H_k$ at each iteration, only limited numbers of vectors are required. Denote $V = I - \rho_k y_k s_k^{'}$, the L-BFGS

inverse Hessian approximation is given by

$$\begin{aligned} H_k = & \left(V_{k-1}^{'} \cdots V_{k-m}^{'}\right) H_k^0 \left(V_{k-m} \cdots V_{k-1}\right) \\ & + \rho_{k-m} \left(V_{k-1}^{'} \cdots V_{k-m+1}^{'}\right) s_{k-m} s_{k-m}^{'} \left(V_{k-m+1} \cdots V_{k-1}\right) \\ & + \rho_{k-m+1} \left(V_{k-1}^{'} \cdots V_{k-m+2}^{'}\right) s_{k-m+1} s_{k-m+1}^{'} \left(V_{k-m+2} \cdots V_{k-1}\right) \\ & + \cdots \\ & + \rho_{k-1} s_{k-1} s_{k-1}^{'}. \end{aligned} \tag{2.86}$$

Only the vectors $s_k$ and $y_k$ are needed for $m$ most recent iterations, which is much more efficient than the storage of the full inverse Hessian matrix. Modest values of $m$ such as $3 \leq m \leq 20$ can provide satisfactory results in practice [99].

When the feasible set $C \neq \mathbb{R}^n$, algorithms for the constrained optimization problem are needed. The gradient projection method is one of the most popular numerical schemes for the constrained problem [17], which can be written as

$$\bar{x}^k = P_C(x^k - \beta_k \nabla f(x^k)), \tag{2.87}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k), \tag{2.88}$$

where $\alpha^k$ is the line search parameter, $\beta_k$ is a positive scalar. The projection operator $P_C(x_0)$ is defined as

$$P_C(x_0) = \arg\min_{x \in C} \|x - x_0\|. \tag{2.89}$$

Similar to the quasi-Newton method, the second-order information can be introduced and this is denoted as the scaled gradient projection method, which is

$$\bar{x}^k = \arg\min_{x \in C} \nabla f(x^k)^{'}(x - x^k) + \frac{1}{2\beta_k}(x - x^k)^{'} B_k (x - x^k), \tag{2.90}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k), \tag{2.91}$$

where $B_k$ is a symmetric positive definite matrix approximating the Hessian matrix $\nabla^2 f(x^k)$. We discuss the gradient projection method and scaled gradient projection method in detail in Chapter 5, and novel numerical schemes are developed for the case when the projection operator can not be evaluated exactly.

Denote the regularization term as $g(x)$, consider the regularized problem as

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + g(x). \tag{2.92}$$

Define a proximal operator as

$$\begin{aligned} x^k &= \text{prox}_{\beta_k}[g](x^{k-1} - \beta_k \nabla f(x^{k-1})), \\ &= \arg\min_x g(x) + \frac{1}{2\beta_k}\|x - (x^{k-1} - \beta_k \nabla f(x^{k-1}))\|^2. \end{aligned} \tag{2.93}$$

It can be derive that, when the $g(x)$ is the indicator function of set $C$, i.e.,

$$g(x) = \delta_C(x) = \begin{cases} x, & \text{as } x \in C, \\ 0, & \text{as } x \notin C, \end{cases} \tag{2.94}$$

Equation (2.93) is equivalent to

$$x^k = P_C(x^{k-1}\beta_k \nabla f(x^{k-1})), \tag{2.95}$$

which is the gradient projection method discussed above [8].

Consider the $l_1$ regularization with $g(x) = \lambda\|x\|_1$, then the equation (2.93) is equivalent to

$$x^k = \mathcal{T}_{\lambda\beta_k}(x^{k-1} - \beta_k \nabla f(x^{k-1})), \tag{2.96}$$

where $\mathcal{T}_{\lambda\beta_k}(\cdot)$ is a shrink operator defined by

$$\mathcal{T}_\alpha(x)_i = (|x_i| - \alpha)_+ \text{sgn}(x_i). \tag{2.97}$$

This update is denoted as the iterative shrinkage-thresholding algorithm (ISTA). An improved version fast iterative shrinkage-thresholding Algorithm (FISTA) is provided in [9],

$$x^k = \text{prox}_{1/L}[g]\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \tag{2.98}$$

$$t^{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right), \tag{2.99}$$

$$y^{k+1} = x^k + \frac{t_k - 1}{t_{k+1}}\left(x^k - x^{k-1}\right). \tag{2.100}$$

The application of the FISTA algorithm for the FWI problem can be found in [144, 2].

The total variation (TV) technique is known for generating the results with piece-wise constant structures, which is important for the seismic inverse problem. Consider the constrained optimization problem with

$$\min_x f(x), \quad \text{such that } \|x\|_{\text{TV}} \leq \theta, \tag{2.101}$$

here $\theta$ is the radius of the TV ball that can be used to control the inverse results. The primal-dual hybrid gradient (PDHG) method is developed for solving the above problem [145]. A PDHG algorithm with adaptive step size is designed in [62]. To apply the PDHG algorithm, define a Lagrangian function [140] as

$$\min_{\delta x} \max_y \mathcal{L}(\delta x, y, \lambda) = \nabla f(x^k)^{'} \delta x + \frac{1}{2} \delta x^{'} + B_k \delta x^{'} + \lambda \left( p^{'} D(x^k + \delta x) - \theta \right), \tag{2.102}$$

where $D$ is the finite difference operator. Then we have the primal problem

$$\arg\min_{\delta x} f(x^k + \delta x) + \delta x^{'} D^{'} y, \tag{2.103}$$

and the dual problem

$$\arg\sup_y y^{'} D(x^k + \delta x) - \theta \|y\|_\infty. \tag{2.104}$$

Then the PDHG algorithm is to iteratively solve the above primal problem and dual problem. For more information about the adaptive PDHG algorithm applied in the FWI problem, we refer to [140]. And for the case when both box constraint and TV constraint are considered, please refer to [55].

# Chapter 3

# Background of optimal transport problem

The classical optimal transport (OT) problem is defined for comparing the difference between two probability measures. Then based on the solution of the optimal transport problem, the optimal transport distance or so-called Wasserstein distance is well defined for probability measures. Compared to the conventional $L^2$ and $L^1$ distance, the OT distance holds better geometric properties, such as the convexity with respect to shift and dilation operations. However, to generalize the OT distance to the general variational problem, two problems need to be considered:

1. Generalize the OT distance to positive measures or positive functions which can describe the physical properties with positive values.

2. Generalize the OT distance to signed measures or the functions with positive, negative, and 0-valued parts which can describe the wavefield and signals.

We focus on the first problem in this chapter and discuss how to partially solve the second problem in the next chapter.

To overcome the mass equality constraint, the unbalanced optimal transport (UOT) problem is studied in recent works [10, 107, 38, 39]. With different approaches to define the UOT problem, the unbalanced optimal transport (UOT) distance can be used to measure the difference between two positive measures. Also, the entropy regularization methods of the original optimal transport problem can be introduced to compute the UOT distance, which provides an efficient way to approximately evaluate the UOT distance and the gradient.

We review the optimal transport problem first. Since the convexity of OT distance with respect to shift and dilation was proved in the work [139] through the Monge problem, which is not an ideal choice as we will discuss later. We reprove the convex properties through the Kantorovich problem. Then the entropy regularization of the OT problem and the Sinkhorn algorithm are reviewed. The UOT distance is introduced in the last section, and the numerical algorithm to evaluate the UOT distance and the gradient is provided.

## 3.1 Review of optimal transport problem

First, we review the optimal transport problem and provide the preliminary results which are useful to describe the metric properties of the proposed mixed $L^1$/Wasserstein distance. For more analysis results on the topic of optimal transport, we refer to the monographs [133, 119]. We refer to the book [106] for more discussion of the computation methods and applications of the optimal transport problem.

Start with the general setting of the optimal transport problem, set $X$ and $Y$ be Polish spaces, we study how to transport a measure $\mu$ on a space $X$ to another measure $\nu$ on a space $Y$.

**Definition 3.1** (Polish space)**.** *A Polish space is a topological space that is homeomorphic to some complete separable metric space. Equivalently, a topological space is a Polish space if it is separable and completely metrizable.*

One of the important examples of Polish space is the Euclidean space $\mathbb{R}^d$ with the usual topology. Since we are interested in the case comparing two functions that representing some physical properties in the real world, we can focus on the special case when there is some nonempty closed subset $\Omega \subset \mathbb{R}^d$ which is also compact. Also, discretization is needed for the numerical algorithms and results. In this case, we are working on the discrete measure defined on $\Omega$ or $\mathbb{R}^d$ in this work.

Denote the set of Radon measures on $X$ as $\mathcal{M}(X)$, and the set of all positive measures on $X$ as $\mathcal{M}_+(X)$. The set of probability measure on $X$ is denoted as $\mathcal{P}(X)$. We will work on the discrete measure in $\mathbb{R}^d$, first, define the probability simplex as

$$\Sigma_n = \left\{ a \in \mathbb{R}_+^n \;\middle|\; \sum_{i=1}^n a_i = 1 \right\}. \tag{3.1}$$

A discrete probability measure with weights $a \in \Sigma_n$ and locations $x_1, \cdots, x_n \in X$ is

$$\alpha = \sum_{i=1}^n a_i \delta_{x_i}, \tag{3.2}$$

where $\delta_{x_i}$ is the Dirac measure concentrated on the point $x_i$. When the coefficient $a$ has not to be restricted

in the probability simplex $\Sigma_n$, the general discrete measure can be defined as:

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \tag{3.3}$$

here $a \in \mathbb{R}^n$. Since our goal is to compare the difference between two discrete signals which have the values as a vector in $\mathbb{R}^n$ and sampling at the spatial (or temporal) points $(x_1, \cdots, x_n) \in (\mathbb{R}^d)^n$, the above discrete measure fulfills our demand. We use $\alpha$, $\beta$ to represent the discrete measures, and use $\mu$, $\nu$ as the general probability measures and Radon measures in this work.

For a continuous map $T : X \to Y$, the push forward operator $T_\sharp : \mathcal{M}(X) \to \mathcal{M}(Y)$ is defined as

$$T_\sharp \mu(E) = \mu(T^{-1}(E)), \quad \forall \text{ Borel set } E \subset Y. \tag{3.4}$$

The push forward operator can be characterized by

$$\int_Y f(y) \, \mathrm{d}T_\sharp \mu(y) = \int_X f \circ T(x) \, \mathrm{d}\mu(x), \quad \forall f \in \mathcal{C}(Y). \tag{3.5}$$

Denote the continuous map $T$ as a transport map. For the discrete probability measure in equation (3.2), the push forward operator is to move the position of all the points in the support of the measure with the transport map, i.e.,

$$T_\sharp \alpha = \sum_{i=1}^{n} a_i \delta_{T(x_i)}. \tag{3.6}$$

To represent the distance between $x_i$ and $T(x_i)$, a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ can be equipped for the space $X$ and $Y$. We focus on the case when $c$ is the square of the Euclidean distance on $\mathbb{R}^d$, i.e., $c(x, y) = |x - y|^2$ for $x \in X$ and $y \in Y$. The optimal transport problem is a classical problem proposed by Monge [96] in 1781, and we write it in the modern language as:

**Problem 3.2** (Monge problem)**.** *Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, minimize*

$$T \to \int_X c(x, T(x)) \, \mathrm{d}\mu(x), \tag{3.7}$$

*for all transport maps $T$ such that $T_\sharp \mu = \nu$.*

One of the issues of the Monge problem (3.7) is the optimal transport map $T$ may not exists. For example, if $\mu$ is a Dirac measure and $\nu$ is not. And also, when we are working on the discrete measures, the Monge

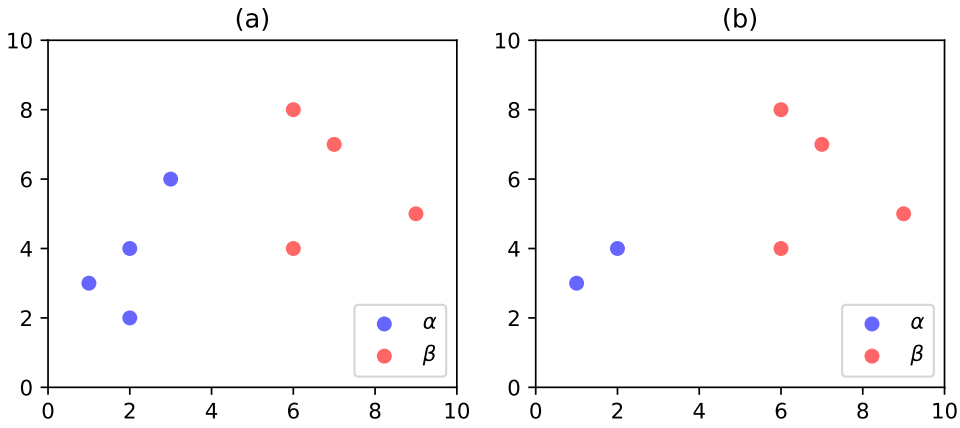


Figure 3.1: The empirical measure $\alpha$ and $\beta$ are shown in blue and red. In the case of subfigure (a), the Monge problem can be well defined. In the case of subfigure (b), the Monge problem can not be well defined.

problem can only be used to compare the uniform histograms with the same size, for example as Figure 3.1. Another issue is the constraint in the Monge problem (3.7) is nonconvex. The reason of the Monge problem is not working well on discrete measures is one source point $x_i$ can only be assigned to another point $T(x_i)$ and no mass can be split [106]. In 1942, Kantorovich proposed a relaxed transportation problem [74]. Instead of working on the transport map $T$, a transport plan $\gamma$ which is a measure on the product space $X \times Y$ is considered, and this allows the mass can be split from a source $x_i$ toward several target points. Define the set of the transport plans as

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) \mid P_{X\sharp}\pi = \mu \text{ and } P_{Y\sharp}\pi = \nu\}, \tag{3.8}$$

where $P_{X\sharp}$ and $P_{Y\sharp}$ are the projections of $X \times Y$ onto $X$ and $Y$ respectively. The set of transport plan is nonempty since there exists a transport plan $\pi$ such that

$$\pi(A \times B) = \mu(A)\nu(B), \quad \forall \text{ Borel sets } A \subset X, B \subset Y. \tag{3.9}$$

Then the Kantorovich problem is given by

**Problem 3.3** (Kantorovich problem)**.** *Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, minimize*

$$\pi \to \int_{X\times Y} c(x, y)\, \mathrm{d}\pi(x, y), \tag{3.10}$$

*such that $\pi \in \Pi(\mu, \nu)$, where the set of transport plans is defined by equation* (3.8)*.*

The following theorems give the existence of the Kantorovich problem.

**Theorem 3.4** ([119], Theorem 1.4). *Let $X$ and $Y$ be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X \times Y \to \mathbb{R}$ is a continuous function. Then the Kantorovich problem admits a solution.*

**Theorem 3.5** ([119], Theorem 1.7). *Let $X$ and $Y$ be Polish spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X \times Y \to \mathbb{R} \cap \{+\infty\}$ is lower semicontinuous. Then the Kantorovich problem admits a solution.*

Notice that the uniqueness is not guaranteed for the Kantorovich problem, one example is shown in Figure 3.2.

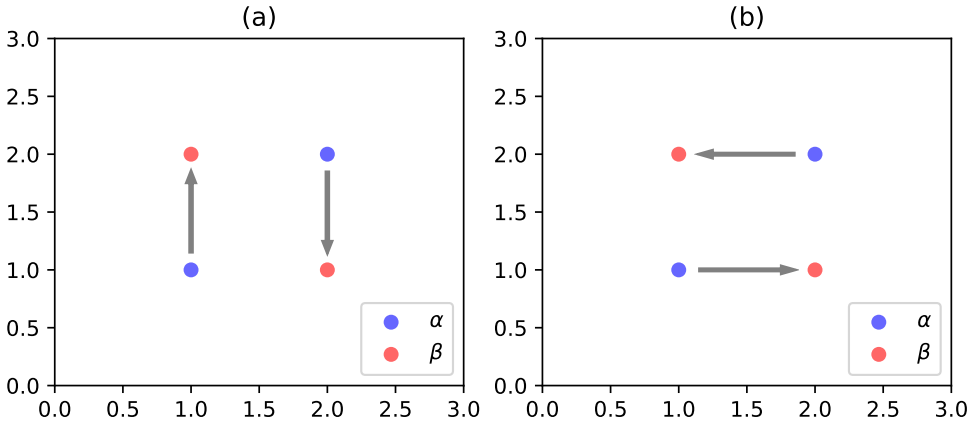


Figure 3.2: The empirical measure $\alpha$ and $\beta$ are shown in blue and red. The transport plan shown in subfigures (a) and (b) are different, but the costs of two transport plan are equal due to the symmetry.

The transport plans "include" transport maps. For two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, suppose we have a transport map $T : X \to Y$ between $\mu$ and $\nu$. A transport plan $\gamma$ can have the form of $(\mathrm{id}, T_\sharp)\mu$. In this case, the Kantorovich problem is a relaxation of the Monge problem. And since we focus on the discrete measure, the Kantorovich problem is a better choice since it is well defined for the discrete measures. The following definitions and corollary provide the connection between the Monge problem and the Kantorovich problem.

**Definition 3.6** (Atomless measure). *When $\Omega \subset \mathbb{R}^d$ be a compact subset, the measure $\mu$ is called atomless if for every singleton in $\Omega$, the measure is $0$, i.e.,*

$$\mu(\{x\}) = 0, \quad \forall x \in \Omega. \tag{3.11}$$

**Corollary 3.7** ([119], Corollary 1.29). *If $\mu$, $\nu$ are two probability measures on $\mathbb{R}^d$ and $\mu$ is atomless, then there exists at least a transport map $T$ such that $T_\sharp \mu = \nu$.*

Obviously, the discrete measure (3.2) is not an atomless measure. The above corollary provides a sufficient condition that the Monge problem is equivalent to the Kantorovich problem. However, we do not have this equivalence for the case of discrete measure.

**Definition 3.8** ([133], Definition 5.1, $c$-cyclically monotonicity)**.** *Let $X$, $Y$ be arbitrary sets, and $c : X \times Y \to (-\infty, \infty]$ be a function. A subset $\Gamma \subset X \times Y$ is said to be c-cyclically monotone if, for any $N \in \mathbb{N}$, and any family $(x_1, y_1), \cdots, (x_N, y_N)$ of points in $\Gamma$, holds the inequality*

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}), \tag{3.12}$$

*holds, with the convention $y_{N+1} = y_1$. A transport plan is said to be c-cyclically monotone if it is concentrated on a c-cyclically monotone set.*

**Theorem 3.9** ([3], Theorem 1.13)**.** *Assume that $c : X \times Y \to \mathbb{R}$ is continuous and bounded from below, and let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be such that*

$$c(x, y) \leq a(x) + b(y), \quad \forall (x, y) \in X \times Y, \tag{3.13}$$

*for some $a \in L^1(\mu)$, $b \in L^1(\nu)$. Also, let $\pi \in \Pi(\mu, \nu)$. Then the following statements are equivalent:*
*(i) The transport plan $\pi$ is optimal.*
*(ii) The set transport plan $\pi$ is c-cyclically monotone.*

Notice that when $X$ and $Y$ are compact subsets in $\mathbb{R}^d$, the condition (3.13) is natural. We will restrict the measure $\mu$ and $\nu$ be the probability measure with a finite second-order moment when $X$ and $Y$ are the Euclidean space $\mathbb{R}^d$. The condition (3.13) is for the cases when the cost function admits $\infty$ values, see Remark 1.14 in [3]. The Theorem 3.9 is an important tool to determine whether a transport plan is the optimal transport plan. And it will be used in the following sections to show that the convex properties of 2-Wasserstein distance.

Since the equation (3.10) is a linear functional for $\pi$, and the constraints of the Kantorovich problem is affine, a dual problem can be achieved.

**Problem 3.10** ([3], Problem 1.16, Dual of Kantorovich problem)**.** *Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, maximize*

$$\int \phi(x) \, \mathrm{d}\mu(x) + \int \psi(y) \, \mathrm{d}\nu(y), \tag{3.14}$$

*among all functions* $\phi \in L^1(\mu)$, $\psi \in L^1(\nu)$ *such that*

$$\phi(x) + \psi(y) \leq c(x, y), \quad \forall x \in X, y \in Y. \tag{3.15}$$

The following theorem provides the existence of a solution to the dual problem and provides the connection between the Kantorovich problem and the dual problem.

**Theorem 3.11** ([119], Theorem 1.39)**.** *Suppose that* $X$ *and* $Y$ *are Polish spaces and that* $c : X \times Y \to \mathbb{R}$ *is uniformly continuous and bounded. Then the dual of Kantorovich problem 3.10 admits a solution. And we have the maximum value of equation* (3.14) *is equal to the minimum of equation* (3.10)*, i.e. strong duality holds.*

**Corollary 3.12.** *Suppose* $X$ *and* $Y$ *are compact subspaces in* $\mathbb{R}^d$*, and the cost function* $c(x, y) = |x - y|^2$ *is the squared Euclidean distance. Then the Kantorovich problem 3.3 and the dual of Kantorovich problem 3.10 admit a solution. The strong duality holds.*

*Proof.* Since the cost function $c(x, y) = |x-y|^2$ is defined on the compact set $X \times Y$, $c$ is uniformly continuous and bounded on $X \times Y$. The existence follows by the Theorem 3.4 and Theorem 3.11. The strong duality follows by the Theorem 3.11. □

Now we discuss the choice of the set $X$ and $Y$. Suppose $X$ and $Y$ are the Euclidean space $\mathbb{R}^d$, and the cost function $c(x, y) = |x - y|^2$ is the square of Euclidean distance. In this case, we have the existence of Kantorovich problem 3.3 by the Theorem 3.5. But the existence of dual Kantorovich problem is not achieved since a uniformly continuous and bounded cost function is needed based on Theorem 3.11. Consider the case when $X$, $Y$ are closed and bounded subset of $\mathbb{R}^d$, the compactness follows by the Heine–Borel theorem. Also, the uniformly continuous of the square of Euclidean distance cost function follows the Heine–Cantor theorem. Then by Theorem 3.4 and Corollary 3.12, we have the existence of both the primal problem and the dual problem.

From the above discussion, the Kantorovich problem is a better choice for our work since it is well defined for the discrete measures. The concept of $c$-transform and $c$-concave functions play an important role in the discussion of the dual Kantorovich problem, the proof of the existence of the dual problem, and the connection between of Kantorovich problem and the dual problem. We refer the monographs [133] and [3] for more detailed discussion.

## 3.2 Metric properties of discrete 2-Wasserstein distance

In this section, we review the Kantorovich problem in the discrete form first, then the $p$-Wasserstein distance for the discrete probability measures is provided. We focus on the 2-Wasserstein distance in this work. Then the metric property and subdifferentiability of the 2-Wasserstein distance are provided. The convexity of 2-Wasserstein distance with respect to shift and dilation was provided in the work [138, 139] through the Monge problem. The convex properties are the main reason for us to study the optimal transport distance. However, the Monge problem is not an ideal choice for the discrete probability measure as we discussed in the previous subsection. The proof of the convexity results based on the Kantorovich problem is provided at the end of this subsection.

Let $\Omega \subset \mathbb{R}^d$ be nonempty, closed, and bounded. Let $p \in [1, \infty)$, we restrict our work on the set of probability measures defined on $\Omega$ with finite $p$-th order moment:

$$\mathcal{P}_p(\Omega) = \left\{ \mu \in \mathcal{P}(\Omega) \;\middle|\; \int_\Omega |x|^p \, \mathrm{d}\mu < +\infty \right\}. \tag{3.16}$$

Next, we define the $p$-Wasserstein distance on $\mathcal{P}_p$ as:

**Definition 3.13.** *Given two measure $\mu, \nu \in \mathcal{P}_p$ and the cost function $c(x, y) = |x - y|^p$, the $p$-Wasserstein distance between $\mu$ and $\nu$ is defined as*

$$W_p(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \left( \int_{\Omega \times \Omega} |x - y|^p \, \mathrm{d}\pi \right)^{1/p}, \tag{3.17}$$

*where the set of transport plans $\Pi(\mu, \nu)$ is given by equation* (3.8).

**Proposition 3.14** ([119], Proposition 5.1)**.** *The quantity $W_p$ defined above is a distance over $\mathcal{P}_p(\Omega)$.*

In this work, we focus on the discrete measure defined on $\Omega$. To define the discrete measures, first denote the set of sampling points as $X = \{x_1, \cdots, x_n\} \subset \Omega$ and $Y = \{y_1, \cdots, y_m\} \subset \Omega$. Define two probability measures as:

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \beta = \sum_{i=1}^{m} b_i \delta_{y_i}, \tag{3.18}$$

where $a \in \Sigma_n$, $b \in \Sigma_m$. Denote $\pi$ as a transport plan between $\alpha$ and $\beta$, i.e. $\pi \in \Pi(\alpha, \beta)$. It is straight forward to see that the transport plan has the discrete form as

$$\pi = \sum_{i,j} P_{i,j} \delta_{(x_i, y_j)}, \quad x_i \in X \text{ and } y_j \in Y. \tag{3.19}$$

We use the matrix $P \in \mathbb{R}_+^{n\times m}$ to represent the transport plan in the discrete form. The set of the transport plans in the discrete form is given by:

$$\Pi_d(\alpha,\beta) = \left\{ P \in \mathbb{R}_+^{n\times m} \;\middle|\; P\mathbf{1}_m = a \text{ and } P^{'}\mathbf{1}_n = b \right\}, \tag{3.20}$$

where $\mathbf{1}_n$ is the row vector with $n$ entries and every entry is 1. The feasible set $\Pi_d(\alpha,\beta)$ is bounded and defined by $n+m$ equality constraints, then it is a convex polytope [106].

Next, let $p \in [1,\infty)$, define the cost matrix $C \in \mathbb{R}^{n\times m}$ as

$$C_{i,j} = c(x_i, y_j) = |x_i - y_j|^p, \tag{3.21}$$

where the distance function $c : X \times Y \to \mathbb{R} \cap \{+\infty\}$ is the $p$-th order of Euclidean distance.

When the sampling points of $\alpha$ and $\beta$ are fixed, the Kantorovich problem 3.15 has a discrete form as:

**Problem 3.15** (Discrete Kantorovich problem)**.** *Given two discrete probability measure $\alpha$ and $\beta$ in $\mathbb{R}^d$ as* (3.18)*, and the cost matrix $C$ is defined by equation* (3.21) *with $p \in [1,\infty)$. The discrete Kantorovich problem is*

$$\min_{P\in\Pi_d(\alpha,\beta)} \langle P, C\rangle = \sum_{i,j} C_{i,j}P_{i,j}, \tag{3.22}$$

*where the set of transport plan $\Pi_d(\alpha,\beta)$ is given by equation* (3.20)*.*

The definition of $p$-Wasserstein distance between discrete probability measures is given by:

**Definition 3.16.** *Given two discrete measure $\alpha$ and $\beta$ as* (3.18)*, and the cost matrix $C$ is defined by equation* (3.21) *with $p \in [1,\infty)$. The $p$-Wasserstein distance between $\alpha$ and $\beta$ is defined as*

$$W_p(\alpha,\beta) = \left( \min_{P\in\Pi_d(\alpha,\beta)} \langle P, C\rangle \right)^{1/p}, \tag{3.23}$$

*where the set of transport plans $\Pi_d(\alpha,\beta)$ is given by equation* (3.20)*.*

**Proposition 3.17.** *The $p$-Wasserstein distance defined in equation* (3.23) *is a distance over the set of discrete probability measures in $\mathcal{P}_p(\Omega)$.*

**Problem 3.18** (Discrete dual problem)**.** *Given two discrete probability measure $\alpha$ and $\beta$ as equation* (3.18)*, and the cost matrix $C$ is defined by equation* (3.21) *with $p \in [1,\infty)$. The dual Kantorovich problem is given*

*by*

$$\max_{(\phi,\psi)\in R_C} \phi^{'} a + \psi^{'} b, \tag{3.24}$$

*where the polyhedron $R_C$ of dual variables is*

$$R_C = \{(\phi,\psi) \in \mathbb{R}^n \times \mathbb{R}^m \mid \phi_i + \psi_j \leq C_{i,j},\ 1 \leq i \leq n,\ 1 \leq j \leq m\}. \tag{3.25}$$

In this work, we focus on the case when $p = 2$. By Corollary 3.12, both the solution of the primal problem 3.15 and the dual problem 3.18 exist, and the strong duality holds. The value of 2-Wasserstein distance can be evaluated with the dual problem 3.18.

Let $(\phi^*, \psi^*)$ solves the dual problem 3.18, it is straightforward to see that $(\phi^* + k\mathbf{1}_n, \psi^* - k\mathbf{1}_m)$ is also a solution of the dual problem. To remove this freedom, we set $\sum_{i=1}^n \phi_i = 0$ [49].

Notice that, for the case of discrete probability measures, the $W_2^2(\alpha,\beta)$ is actually controlled by the vectors $a$, $b$, and sets $X$, $Y$. Then the square of 2-Wasserstein distance for $\alpha$ and $\beta$ in equation (3.18) is

$$W_2^2(\alpha,\beta) = W_2^2(a,b,X,Y) = \max_{(\phi,\psi)\in R_C} \phi^{'} a + \psi^{'} b, \tag{3.26}$$

where the feasible set $R_C$ is given by equation (3.25). The subdifferentiability of $W_2^2$ as a function of $a$ is given by the following proposition.

**Proposition 3.19** ([49], Proposition 1)**.** *Given two discrete probability measure $\alpha$ and $\beta$ as equation* (3.18)*, and the cost matrix $C$ is defined by equation* (3.21) *with $p = 2$. Any optimal dual variable $\phi^*$ of the dual problem* (3.24) *is a subgradient of $W_2^2$ with respect to $a$.*

One of the reasons we introduce the 2-Wasserstein distance to the inverse problem is that it maintains the convexity with respect to shift and dilation compared to the usual $L^2$ distance which is popular in the inverse problem. The convexity results has been shown in the work of [138, 139]. However, in the above work, the Wasserstein distance and the convex properties were built with the Monge problem which is not the ideal choice for the discrete measures used in this work. We show the convex properties of the 2-Wasserstein distance with the discrete Kantorovich problem 3.15 next.

The following theorem provides the convexity with respect to shift between two discrete probability measures. The shift process is described in Figure 3.3 (a).

**Theorem 3.20.** *Suppose $\alpha$ and $\beta$ are two discrete probability measures defined by equation* (3.18)*, let $\pi = \sum_{i,j} P_{i,j}\delta_{(x_i,y_j)} \in \Pi(\alpha,\beta)$ be the optimal transport plan that rearranges $\alpha$ to $\beta$, where the matrix*

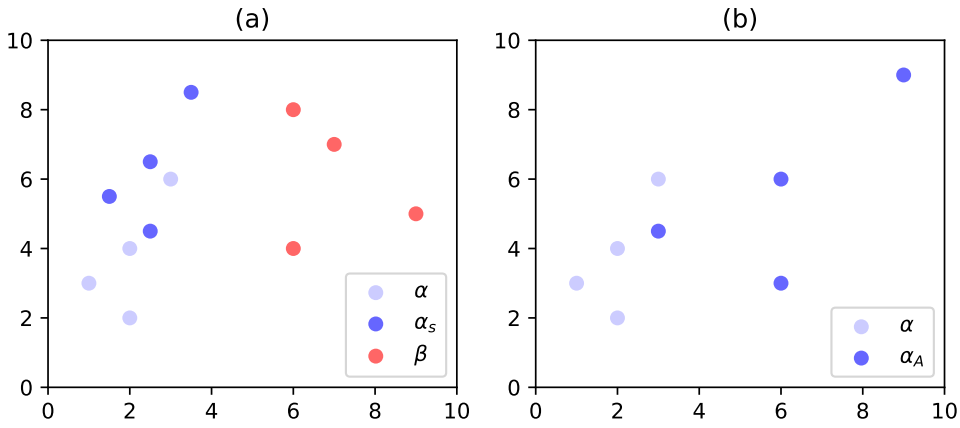


Figure 3.3: (a): The empirical measure $\alpha_s$ is the shift of measure $\alpha$ with the direction $\eta = (1, 5)$ and the length $s = 0.5$. (b): The empirical measure $\alpha_A$ is the dilation of measure $\alpha$ with the transform $A = \begin{bmatrix} 3 & 0 \\ 0 & 1.5 \end{bmatrix}$.

$P \in \Pi_d(\alpha, \beta)$. *The shift of discrete measure* $\alpha$ *with the direction* $\eta \in \mathbb{R}^d$ *and shift size* $s > 0$ *is defined as*

$$\alpha_s = \sum_{i=1}^{n} a_i \delta_{x_i + s\eta}. \tag{3.27}$$

*Here the shift size* $s$ *is small enough such that* $x_i + s\eta \in \Omega$ *for* $i = 1, \cdots, n$. *Then* $W_2^2(\alpha_s, \beta)$ *is convex with respect to the shift size* $s$.

*Proof.* Denote $\pi_s = \sum_{i,j} (P_s)_{i,j} \delta_{(x_i + s\eta, y_j)} \in \Pi(\alpha_s, \beta)$ as a transport map between $\alpha_s$ and $\beta$, where $P_s \in \Pi_d(\alpha_s, \beta)$. Since measure $\alpha_s$ and $\alpha$ have the same value $a$, then $\Pi_d(\alpha_s, \beta) = \Pi_d(\alpha, \beta)$. Then $P_s = P$, i.e., the discrete transport plan is not changing with the shift operator. And $\pi_s = \sum_{i,j} P_{i,j} \delta_{(x_i + s\eta, y_j)}$ is a transport plan between $\alpha_s$ and $\beta$.

Next, we show that $\pi_s = \sum_{i,j} P_{i,j} \delta_{(x_i + s\eta, y_j)}$ is the optimal transport plan. By Theorem 3.9, it is equivalent to show that the transport plan $\pi_s$ is $c$-cyclically monotone. Denote $X_s = \{x_1 + s\eta, \cdots, x_N + s\eta\}$ be the set where $\alpha_s$ is concentrated. Suppose for any $N \in \mathbb{N}$, and any family of points $(x_1 + s\eta, y_1), \cdots (x_N + s\eta, y_N) \in X_s \times Y$, by Definition 3.8 and $c(x, y) = |x - y|^2$,

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}), \tag{3.28}$$

$$\sum_{i=1}^{N} \langle x_i - y_i, x_i - y_i \rangle \leq \sum_{i=1}^{N} \langle x_i - y_{i+1}, x_i - y_{i+1} \rangle, \tag{3.29}$$

$$\sum_{i=1}^{N} \langle x_i, y_i \rangle \geq \sum_{i=1}^{N} \langle x_i, y_{i+1} \rangle. \tag{3.30}$$

Then,

$$\begin{aligned}\sum_{i=1}^{N}\langle x_i+s\eta,y_i\rangle &= \sum_{i=1}^{N}\langle x_i,y_i\rangle+\sum_{i=1}^{N}\langle s\eta,y_i\rangle\\ &\geq \sum_{i=1}^{N}\langle x_i,y_{i+1}\rangle+\sum_{i=1}^{N}\langle s\eta,y_i\rangle\\ &= \sum_{i=1}^{N}\langle x_i,y_{i+1}\rangle+\sum_{i=1}^{N}\langle s\eta,y_{i+1}\rangle\\ &= \sum_{i=1}^{N}\langle x_i+s\eta,y_{i+1}\rangle\,.\end{aligned} \tag{3.31}$$

Then,

$$W_2^2(\alpha_s,\beta)=\langle P,C\rangle=\sum_{i,j}P_{i,j}|x_i+s\eta-y_j|^2, \tag{3.32}$$

which is convex with respect to $s$. □

The next theorem provides the convex property of the square of 2-Wasserstein distance with respect to the dilation. The dilation process is described in Figure 3.3 (b).

**Theorem 3.21.** *Given $a\in\Sigma_n$ and a discrete set $X=\{x_1,\cdots,x_n\}\subset\Omega$, two discrete probability measures are defined by*

$$\alpha=\sum_{i=1}^{n}a_i\delta_{x_i},\quad \alpha_A=\sum_{i=1}^{n}a_i\delta_{Ax_i}.$$

*Here $A$ is a dilation transform matrix which is symmetric positive definite, and $\alpha_A$ is the dilation of $\alpha$. Also, suppose $x_i\in\Omega$ and $Ax_i\in\Omega$ for $i=1,\cdots,n$. Let $\pi=\sum_{i,j}P_{i,j}\delta_{(Ax_i,x_j)}\in\Pi(\alpha_A,\alpha)$ be the transport plan that rearranges $\alpha_A$ to $\alpha$, where $P\in\Pi_d(\alpha_A,\alpha)$. Then, $\pi$ is the optimal transport plan with $P=diag(a)$, and $W_2^2(\alpha_A,\alpha)$ is convex with respect to the eigenvalues of $A$.*

*Proof.* From the definition of $\alpha$ and $\alpha_A$, it is easy to see that when $P=\mathrm{diag}(a)$, $\pi\in\Pi(\alpha_A,\alpha)$.

Next we show that $\pi=\sum_{i,j}P_{i,j}\delta_{(Ax_i,x_j)}$ is the optimal transport plan. For any $N\in\mathbb{N}$ and $N\leq n$ and any family of points $\{x_1,\cdots,x_N\}\subset X$, by the construction above $\{Ax_1,\cdots,Ax_N\}\subset\Omega$, denote $y_i=Ax_i$,

for $i = 1, \cdots, N$.

$$\begin{aligned}\sum_{i=1}^{N} \langle y_i, x_i \rangle - \sum_{i=1}^{N} \langle y_i, x_{i+1} \rangle &= \sum_{i=1}^{N} \langle Ax_i, x_i \rangle - \sum_{i=1}^{N} \langle Ax_i, x_{i+1} \rangle \\ &= \frac{1}{2} \left( \sum_{i=1}^{N} \langle Ax_i, x_i \rangle + \sum_{i=1}^{N} \langle Ax_{i+1}, x_{i+1} \rangle - 2 \sum_{i=1}^{N} \langle Ax_i, x_{i+1} \rangle \right) \\ &= \frac{1}{2} \sum_{i=1}^{N} \langle A(x_i - x_{i+1}), x_i - x_{i+1} \rangle \geq 0.\end{aligned} \tag{3.33}$$

The last inequality holds since $A$ is symmetric positive definite. Then, by inequality (3.30), the transport plan $\pi = \sum_{i,j} P_{i,j} \delta_{(Ax_i, x_j)}$ is concentrated on a $c$-cyclically monotone set. By Theorem 3.9, the transport plan $\pi$ is optimal.

Since $A$ is a real symmetric matrix, by eigendecomposition,

$$A = QDQ^{'}, \tag{3.34}$$

where $Q$ is a real orthogonal matrix, and $D$ is a diagonal matrix whose entries are the eigenvalues of $A$, denoted as $D = \mathrm{diag}\{\lambda_1, \cdots, \lambda_d\}$. Also,

$$\begin{aligned}W_2^2(\alpha_A, \alpha) &= \langle \mathrm{diag}(a), C \rangle \\ &= \sum_{i=1}^{n} a_i \left| x_i - Ax_i \right|^2 \\ &= \sum_{i=1}^{n} a_i \left\langle (I - A)x_i, (I - A)x_i \right\rangle \\ &= \sum_{i=1}^{n} a_i (Q^{'} x_i)^{'} \mathrm{diag}\left( (1 - \lambda_1)^2, \cdots, (1 - \lambda_d)^2 \right) Q^{'} x_i,\end{aligned} \tag{3.35}$$

where $\lambda_1, \cdots, \lambda_d$ are the eigenvalues of $A$. The convex properties follows the above equation. □

## 3.3 Entropy regularization of the optimal transport problem

In this subsection, we review the entropy regularization for the optimal transport problem, which leads to a fast approximate evaluation of the Wasserstein distance. The numerical methods in this section are not innovative, but it is important for the numerical evaluation of the following unbalanced optimal transport distance and the proposed mixed $L^1$/Wasserstein distance.

As we discussed before, the Wasserstein distance holds several desirable geometry properties compared to the conventional $L^2$ and $L^1$ distance for discrete vectors. However, the evaluation of the Wasserstein

distance is not straight forward. The computational cost of solving the discrete Kantorovich problem 3.15 through network simplex or interior point methods for two $n$-dimensional vectors is at least $O(n^3 \log(n))$ [48, 103], that prevents the widespread use of the Wasserstein distance for large-scale problem. A dynamic formulation of optimal transport problem is proposed in the work [11] which provides a connection between the optimal transport problem and the computational fluid mechanics. By transforming the transport plan as a geodesic, the optimal transport problem can be solved as a convex optimization problem. However, an extra time dimension is needed for this approach and this expands the dimension of the problem and is expensive to compute. When the optimal transport problem is considered with the squared Euclidean distance as the cost function, it can be solved through the connection between the optimal transport plan and the Monge-Ampère equation [14]. The Monge-Ampère equation can be solved with Newton's method [88, 59] approximately, and additional regularity assumptions on the density and domain are needed [15]. And this approach has been successfully applied in the geophysics domain and the seismic inversion problem [139, 138].

Besides the above methods, the entropy regularization method is the most popular numerical method for the optimal transport problem. Introducing the regularization to the optimal transport problem is a natural choice and can date back to the 1960s [136]. The work [48, 49] provide a smoothed Wasserstein distance by solving the Kantorovich problem with an entropy regularization. Those works make the evaluation of large-scale optimal transport problem become possible and popularized the Wasserstein distance in the machine learning society. By introducing the regularization term to the Kantorovich problem 3.15, a strict convex problem can be solved with matrix scaling algorithms. In [48], the author suggests to use the Sinkhorn's fixed point iteration algorithm which has a linear convergence [58, 75]. In this case, the smoothed Wasserstein distance is also denoted as Sinkhorn distance. We provide a short review of using the Sinkhorn algorithm to solve the entropy regularized optimal transport problem.

Let $F : \Omega \to \mathbb{R}$ be a continuously differentiable, strictly convex function, and the set $\Omega$ is closed and convex. The Bregman divergence associated with $F$ for points $a, b \in \Omega$ is given by

$$D_F(a, b) = F(a) - F(b) - \langle \nabla F(b), a - b \rangle . \tag{3.36}$$

Given a vector $a \in \mathbb{R}^n$, with $a_i \geq 0$ for $i = 1, \cdots, n$, the entropy function is defined as:

$$E(a) = -\sum_{i=1}^{n} a_i(\log(a_i) - 1), \tag{3.37}$$

here the convention $0 \log(0) = 0$ is used. The Kullback–Leibler (KL) divergence (also called relative entropy)

is defined as the negative Bregman divergence associated with the entropy function $E$. Given vector $a, b \in \mathbb{R}^n$, with $a_i \geq 0, b_i \geq 0$, for $i = 1, \cdots, n$, the Kullback–Leibler (KL) divergence is defined as:

$$KL(a|b) = \sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) - a_i + b_i, \tag{3.38}$$

here the convention $0 \log(0/0) = 0$ is used. For the special case when $a, b \in \Sigma_n$,

$$KL(a|b) = \sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right). \tag{3.39}$$

Since the matrix can be rearranged into a vector, the entropy function and the KL divergence can be defined for the matrix with non-negative entries with the same method.

Given discrete probability measures $\alpha$ and $\beta$ defined by equation (3.18), consider the entropy regularized Kantorovich problem:

$$P_\varepsilon = \arg\min_{P \in \Pi_d(\alpha,\beta)} \langle P, C\rangle - \varepsilon E(P), \tag{3.40}$$

where $\varepsilon > 0$ is the regularization parameter, $\Pi_d(\alpha, \beta)$ is the set of optimal transport plan defined by (3.20). Since we focus on the case when the cost matrix $C$ is defined by the squared Euclidean distance, we denote the square of the regularized 2-Wasserstein distance as:

$$W_{2,\varepsilon}^2(\alpha, \beta) = \langle P_\varepsilon, C\rangle, \tag{3.41}$$

$$\text{where } P_\varepsilon = \arg\min_{P \in \Pi_d(\alpha,\beta)} \langle P, C\rangle - \varepsilon E(P). \tag{3.42}$$

This is also denoted as the Sinkhorn distance in the work [48].

Figure 3.4 shows that the convergence behavior of the transport plan in the regularized Kantorovich problem as $\varepsilon \to 0$. In Figure 3.4 (a), the Gaussian density $\alpha$ is centered at 0.4, and $\beta$ is centered at 0.6. The transport plan between $\alpha$ and $\beta$ should be a shift as discussed in the previous section, which should be sparse as shown in the transport plan matrix. As shown in Figure 3.4 (b), (c), (d), the transport plan converges to a sparse matrix as the regularization parameter $\varepsilon$ goes to 0. For more analysis results of the convergence behavior of the entropy regularized OT problem, please refer to the monograph [106], Chapter 4.

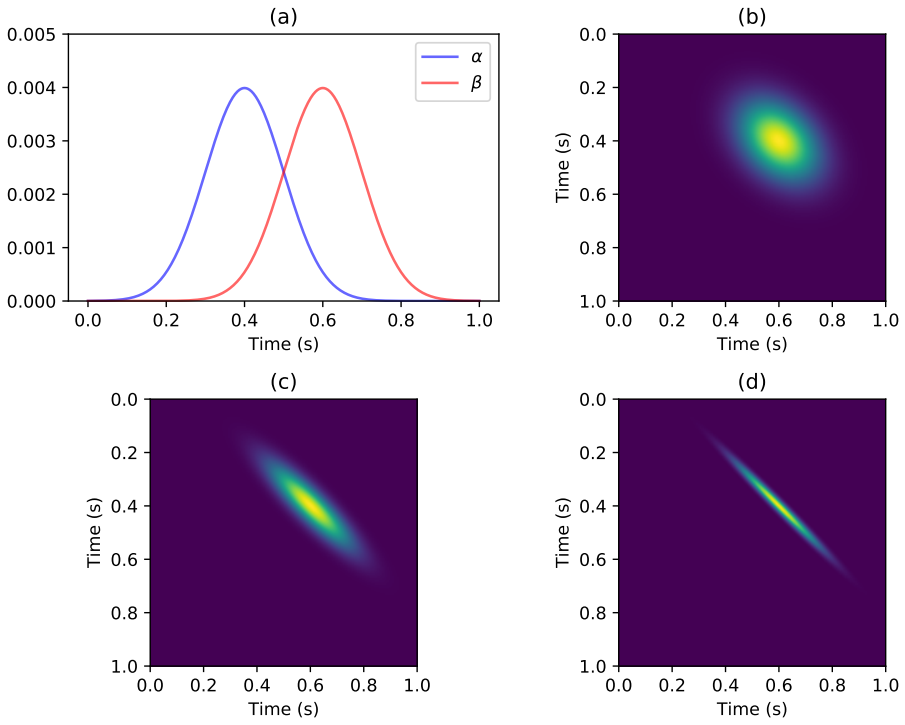


Figure 3.4: (a): Two Gaussian densities $\alpha$ and $\beta$ centered at 0.4 s and 0.6 s. (b): The regularized transport plan with $\varepsilon = 5 \times 10^{-2}$. (c): The regularized transport plan with $\varepsilon = 5 \times 10^{-3}$. (d): The regularized transport plan with $\varepsilon = 5 \times 10^{-4}$.

It is straightforward to show that

$$\begin{aligned} P_\varepsilon &= \arg \min_{P \in \Pi_d(\alpha,\beta)} KL(P|K) \\ &= \arg \min_{P \in \Pi_d(\alpha,\beta)} \sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j} \log \left( \frac{P_{i,j}}{K_{i,j}} \right) - P_{i,j} + K_{i,j}, \end{aligned} \tag{3.43}$$

where $K_{i,j} = e^{-C_{i,j}/\varepsilon}$. Denote,

$$C_1 = \left\{ P \in \mathbb{R}_+^{n \times m} \;\middle|\; P\mathbf{1}_m = a \right\}, \quad C_2 = \left\{ P \in \mathbb{R}_+^{n \times m} \;\middle|\; P^{'}\mathbf{1}_n = b \right\}, \tag{3.44}$$

then $\Pi_d(\alpha, \beta) = C_1 \cap C_2$. The regularized Kantorovich problem (3.40) can be explained as finding a projection of $K$ onto the intersection of $C_1$ and $C_2$, which can be solved by the Bregman iterative projections [15]. This is equivalent to the following matrix scaling algorithm approach, and we refer to the work [15] for the details of this approach.

Consider the Lagrangian of the problem (3.40),

$$\mathcal{L}(P_\varepsilon, \phi, \psi) = \sum_{i,j}(P_\varepsilon)_{i,j}C_{i,j} + \varepsilon(P_\varepsilon)_{i,j}\left(\log(P_\varepsilon)_{i,j} - 1\right) + \phi'\left(a - P_\varepsilon \mathbf{1}_m\right) + \psi'\left(b - P_\varepsilon^{'}\mathbf{1}_n\right). \tag{3.45}$$

The first order optimality condition provides that

$$\frac{\partial \mathcal{L}(P_\varepsilon, \phi, \psi)}{\partial (P_\varepsilon)_{i,j}} = C_{i,j} + \varepsilon \log(P_\varepsilon)_{i,j} - \phi_i - \psi_j = 0, \tag{3.46}$$

which equivalents to

$$P_\varepsilon = \text{diag}(u)K\text{diag}(v), \tag{3.47}$$

where $u_i = e^{\phi_i/\varepsilon}$, $v_j = e^{\psi_i/\varepsilon}$, and $K_{i,j} = e^{-C_{i,j}/\varepsilon}$.

The dual problem can be derived with the Lagrangian:

$$\max_{\phi,\psi \in \mathbb{R}^n \times \mathbb{R}^m} \phi^{'} a + \psi^{'} b - \varepsilon \sum_{i,j} e^{-(C_{i,j} - \phi_i - \psi_j)/\varepsilon}. \tag{3.48}$$

The following proposition provides the connection between the regularized Kantorovich problem (3.40) and the dual problem.

**Proposition 3.22** ([49], Proposition 2)**.** *Given discrete probability measures $\alpha, \beta$, the matrix $K$ is defined by $K_{i,j} = e^{-C_{i,j}/\varepsilon}$. Then there exists a pair of vectors $(u, v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ such that the optimal solutions of the primal problem* (3.40) *and the dual problem* (3.48) *is given by*

$$P_\varepsilon^* = \textit{diag}(u)K\,\textit{diag}(v). \tag{3.49}$$

*The dual parameter $\phi^*$ is given by*

$$\phi_i^* = \varepsilon \log(u_i). \tag{3.50}$$

Then, solving the regularized Kantorovich problem (3.40) is equivalent to find vector $u \in \mathbb{R}_+^n$, $v \in \mathbb{R}_+^m$ such that $P = \text{diag}(u)K\text{diag}(v)$ satisfies $P\mathbf{1}_m = a$ and $P^{'}\mathbf{1}_n = b$. This matrix scaling problem can be solved with the Sinkhorn fixed point iteration. The following lemma provides the matrix scaling computation of the regularized Kantorovich problem (3.40). We refer to [48, 49, 106] for more details.

**Lemma 3.23** ([49], Lemma 1; [122])**.** *For any positvie matrix $K \in \mathbb{R}_+^{n \times m}$ and positive probability vectors*

*$a \in \Sigma_n$ and $b \in \Sigma_m$, there exist positive vectors $u \in \mathbb{R}_+^n$ and $v \in \mathbb{R}_+^m$, unique up to scalar multiplication, such that $diag(u)K\,diag(v) \in \Pi_d(\alpha, \beta)$. Such a pair $(u, v)$ can be recovered as a fixed point of the Sinkhorn map*

$$u_i = \phi_i/(Kv)_i, \quad v_j = \psi_j/(K'u)_j. \tag{3.51}$$

The Sinkhorn algorithm of regularized Kantorovich problem (3.40) is given by Algorithm 1.

The stopping criteria can be designed in a variety of ways, for example, the iteration process can be terminated after a certain number of iterations. If the distance is evaluated for each iterations as $d_1$, by introducing a temporary variable $d_0$ to save the distance in the previous iteration. The iteration can be terminated when $|d_1/d_0 - 1| \leq \eta$, here $\eta$ is a threshold variable to control the accuracy of the distance.

**Algorithm 1:** Sinkhorn algorithm.

Input: $\alpha$, $\beta$, $C$, $\varepsilon$.
Initialization: matrix $K$ with $K_{i,j} = e^{-C_{i,j}/\varepsilon}$, $u = \mathbf{1}_n$, $v = \mathbf{1}_m$.
**while** *not converged* **do**
  Update vector $u$ with $u_i = a_i/(Kv)_i$.
  Update vector $v$ with $v_j = b_j/(K^{'}u)_j$.
**end**
Compute the transport plan matrix $P_\varepsilon = \mathrm{diag}(u)K\mathrm{diag}(v)$.
**return** *The Sinkhorn distance* $W_{2,\varepsilon}^2(\alpha, \beta) = \langle P_\varepsilon, C\rangle$.
*The gradient of* $W_{2,\varepsilon}^2(\alpha, \beta)$ *with respect to* $a$ *is* $(\nabla_a W_{2,\varepsilon}^2(\alpha, \beta))_i = \varepsilon \log(u_i) - \varepsilon/n \sum_j \log(u_j)$.

Two matrix-vector productions $Kv$ and $K^{'}u$ are evaluated for each of the iterations in Algorithm 1. Suppose $u$ and $v$ are the row vectors rearranged by the matrices with $n = N \times M$ size, then the computational complexity of the matrix-vector production be $O(N^2M^2)$. Also, the $MN \times MN$ matrix $K$ is stored for the matrix-vector production which is impossible for large-scale problems. When the discrete probability measure $\alpha$ and $\beta$ are defined on $\mathbb{R}^d$, a special technique can be applied. This technique is denoted as convolutional Wasserstein distance and extends to the case when the optimal transport problem is studied on manifolds [124]. The central idea is to replace the matrix-vector production with a kernel convolution. Since the feasibility of large-scale problem is common for the variational problem, we review this technique by using the tools of the Kronecker product.

Suppose the discrete probability measures $\alpha$ and $\beta$ are defined on $\mathbb{R}^2$,

$$\alpha = \sum_i a_i \delta_{x_i}, \quad \beta = \sum_j b_j \delta_{y_j}, \tag{3.52}$$

here $a, b \in \Sigma_n$. And the index $i$ and $j$ are multi-index here as $x_i = (x_{i_1}^1, x_{i_2}^2) \in \mathbb{R}^2$, $y_j = (y_{j_1}^1, y_{j_2}^2) \in \mathbb{R}^2$.

Define the matrix $C_1$ and $C_2$ as

$$(C_1)_{i_1,j_1} = (x^1_{i_1} - y^1_{j_1})^2, \quad (C_2)_{i_2,j_2} = (x^2_{i_2} - y^2_{j_2})^2. \tag{3.53}$$

Then the cost matrix $C$ is defined as

$$C_{i,j} = |x_i - y_j|^2 = (x^1_{i_1} - y^1_{j_1})^2 + (x^2_{i_2} - y^2_{j_2})^2 = (C_1)_{i_1,j_1} + (C_2)_{i_2,j_2}. \tag{3.54}$$

Define the matrix $K_1$ and $K_2$ as

$$(K_1)_{i_1,j_1} = e^{-(C_1)_{i_1,j_1}/\varepsilon}, \quad (K_2)_{i_2,j_2} = e^{-(C_2)_{i_2,j_2}/\varepsilon}, \tag{3.55}$$

by the definition of matrix $K$,

$$K_{i,j} = e^{-C_{i,j}/\varepsilon} = (K_1)_{i_1,j_1}(K_2)_{i_2,j_2}. \tag{3.56}$$

Then matrix $K$ can be represented as the Kronecker product between $K_2$ and $K_1$,

$$K = K_2 \otimes K_1. \tag{3.57}$$

**Lemma 3.24** ([73], Section 2.8)**.** *For the $N \times M$ dimension matrix $U$ and $V$, let $u$ and $v$ be the row-ordered vector of matrix $U$ and $V$. Given $N \times N$ matrix $A$ and $M \times M$ matrix $B$, if $V = AUB^{'}$, then $v = (A \otimes B)u$.*

Taking advantage of the above lemma, we can have the following algorithm designed for computing the Sinkhorn distance for two-dimensional images.

## 3.4 Unbalanced optimal transport distance

To overcome the mass equality limitation, the unbalanced optimal transport (UOT) problem is raised in [10] based on a dynamic approach. Later several works have been proposed in both static and dynamic approaches [107, 38, 39]. In this subsection, we introduce the UOT distance mainly based on the work in [39], and then apply it to the FWI problem.

Let $\Omega$ be a nonempty, closed and bounded subset in $\mathbb{R}^n$, given two sampling set $X = \{x_1, \cdots, x_n\} \subset \Omega$

**Algorithm 2:** Sinkhorn algorithm in $\mathbb{R}^2$.

Input: discrete probability measures $\alpha, \beta$ on $\mathbb{R}^2$ with densities $a, b \in \mathbb{R}^{N\times M}$, $\varepsilon > 0$. The cost matrix $C_1$ and $C_2$ defined by equation (3.53).
Initialization: matrix $K_1$, $K_2$ with $(K_1)_{i,j} = e^{-(C_1)_{i,j}/\varepsilon}$, $(K_2)_{i,j} = e^{-(C_2)_{i,j}/\varepsilon}$. Matrix $u, v$ are initialized as $N \times M$ matrices with all entries be 1. Let $d = 0$.
**while** *not converged* **do**
  Update vector $u$ with $u_{i,j} = a_{i,j}/(K_2 v K_1^{'})_{i,j}$.
  Update vector $v$ with $v_{i,j} = b_{i,j}/(K_2^{'} u K_1)_{i,j}$.
**end**
**for** $i_1 = 1 : N$ **do**
  **for** $i_2 = 1 : M$ **do**
    **for** $j_1 = 1 : N$ **do**
      **for** $j_2 = 1 : M$ **do**
        $d = d + u_{i_1,i_2}(K_2)_{i_2,j_2}(K_1)_{i_1,j_1} v_{j_1,j_2}((C_1)_{i_1,j_1} + (C_2)_{i_2,j_2})$.
      **end**
    **end**
  **end**
**end**
**return** *The Sinkhorn distance* $W^2_{2,\varepsilon}(\alpha, \beta) = d$.
*The gradient of* $W^2_{2,\varepsilon}(\alpha, \beta)$ *with respect to* $a$ *is* $(\nabla_a W^2_{2,\varepsilon}(\alpha, \beta))_{i,j} = \varepsilon \log(u_{i,j}) - \varepsilon/n \sum_{k,l} \log(u_{k,l})$.

and $Y = \{y_1, \cdots, y_m\} \subset \Omega$. Define two positive discrete measure on $\Omega$ as

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \beta = \sum_{i=1}^{m} b_i \delta_{y_i}, \tag{3.58}$$

where $a = (a_1, \cdots, a_n) \in \mathbb{R}^n_+$ and $b = (b_1, \cdots, b_m) \in \mathbb{R}^m_+$. Denote the set of discrete measure defined in equation (3.58) on $\Omega$ as $\mathcal{M}_d(\Omega)$.

When $a$ and $b$ are the density functions of probability measures $\alpha$ and $\beta$, the equal mass condition $\sum_i a_i = \sum_i b_i$ is satisfied intrinsically. The unbalanced optimal transport problem is a generalization of the optimal transport problem to overcome the mass equality limitation between $\alpha$ and $\beta$. The unbalanced optimal transport problem used in this work is based on the work in [39]. To relax the marginal constraints in the discrete Kantorovich problem 3.15, the unbalanced optimal transport problem is defined as:

$$\min_{P \in \mathbb{R}^{n\times m}} \langle P, C\rangle + F_a(P\mathbf{1}_m) + F_b(P^{'}\mathbf{1}_n), \tag{3.59}$$

here both $F_a$ and $F_b$ are proper convex functions.

For example, when $F_a$ and $F_b$ are the indicator function:

$$F_a(P\mathbf{1}_m) = \iota_{\{=\}}(P\mathbf{1}_m|a), \quad F_b(P^{'}\mathbf{1}_n) = \iota_{\{=\}}(P^{'}\mathbf{1}_n|b), \tag{3.60}$$

where the indecator function between vectors $a, b \in \mathbb{R}^n$ is defined as

$$\iota_{=}(a|b) = \begin{cases} 0, & a = b, \\ \infty, & a \neq b. \end{cases} \tag{3.61}$$

It can be easily checked that the unbalanced optimal transport problem (3.59) coincides with the discrete Kantorovich problem (3.15) when $\sum_i a_i = \sum_i b_i$. In this work we consider the case when

$$F_a(P\mathbf{1}_m) = \varepsilon_u KL(P\mathbf{1}_m|a), \quad F_b(P^{'}\mathbf{1}_n) = \varepsilon_u KL(P^{'}\mathbf{1}_n|b). \tag{3.62}$$

Here $F_a$ and $F_b$ are the Kullback-Leibler divergence between vectors given in equation (3.38) which measures the differences between $P\mathbf{1}_m$ and $a$, $P^{'}\mathbf{1}_n$ and $b$. And the parameter $\varepsilon_u$ controls the weight of the mass balancing term in (3.59).

Similar to the Wasserstein distance (3.23), the unbalanced optimal transport distance based on the square Euclidean ground cost between vector is:

$$W^2_{2,\varepsilon_u}(\alpha, \beta) = \min_{P \in \mathbb{R}^{n \times m}} \langle P, C \rangle + \varepsilon_u KL(P\mathbf{1}_m|a) + \varepsilon_u KL(P^{'}\mathbf{1}_n|b), \tag{3.63}$$

where the cost matrix $C$ is defined as $C_{i,j} = |x_i - y_j|^2$.

As we discussed in the previous subsection, the entropy regularization method is a proper choice to evaluate the Wasserstein distance and the gradient. Given the regularization parameter $\varepsilon > 0$, consider the entropy regularized UOT problem:

$$\min_{P \in \mathbb{R}^{n \times m}} \langle P, C \rangle - \varepsilon E(P) + \varepsilon_u KL(P\mathbf{1}_m|a) + \varepsilon_u KL(P^{'}\mathbf{1}_n|b). \tag{3.64}$$

Same as we discussed in the previous subsection, the above equation can be rewritten as

$$\min_{P \in \mathbb{R}^{n \times m}} \varepsilon KL(P|K) + \varepsilon_u KL(P\mathbf{1}_m|a) + \varepsilon_u KL(P^{'}\mathbf{1}_n|b), \tag{3.65}$$

where $K$ is defined as $K_{i,j} = e^{-C_{i,j}/\varepsilon}$.

Then we have the definition of regularized unbalanced optimal transport distance used in this work:

**Definition 3.25.** *Given positive discrete measures $\alpha, \beta$ with equation* (3.58)*. Define the ground cost matrix $C$ by $C_{i,j} = |x_i - y_j|^2$. With the regularization parameter $\varepsilon$ and the mass balancing parameter $\varepsilon_u$, the*

*regularized unbalanced optimal transport distance between $\alpha$ and $\beta$ is*

$$\begin{aligned} W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta) =& \langle P_\varepsilon, C\rangle + \varepsilon_u KL(P_\varepsilon \mathbf{1}_m|a) + \varepsilon_u KL(P_\varepsilon^{'}\mathbf{1}_n|b), \\ \text{where } P_\varepsilon =& \arg\min_{P\in\mathbb{R}^{n\times m}} \varepsilon KL(P|K) + \varepsilon_u KL(P\mathbf{1}_m|a) + \varepsilon_u KL(P^{'}\mathbf{1}_n|b). \end{aligned} \tag{3.66}$$

*Here $KL(\cdot|\cdot)$ is the Kullback-Leibler divergence between two matrices or vectors. And $K_{i,j} = e^{-C_{i,j}/\varepsilon}$.*

Notice that, when the support $X$ and $Y$ of $\alpha$ and $\beta$ are fixed, the UOT distance $W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta)$ is defined with respect to the vector $a$ and $b$, that can also be denoted as $W^2_{2,\varepsilon_u,\varepsilon}(a,b)$.

Equation (3.66) is denoted as the primal problem. The dual problem is needed to compute the unbalanced optimal transport distance.

**Theorem 3.26** ([39], Theorem 3.2)**.** *The dual problem of* (3.66) *is*

$$\max_{\phi,\psi\in\mathbb{R}^n_+} \sum_{i,j} -\varepsilon_u a_i\left(e^{-\phi_i/\varepsilon_u}-1\right) - \varepsilon_u b_j\left(e^{-\psi_j/\varepsilon_u}-1\right) - \varepsilon K_{i,j}\left(e^{\phi_i/\varepsilon}e^{\psi_j/\varepsilon}-1\right), \tag{3.67}$$

*where the matrix $K$ is defined by $K_{i,j} = e^{-C_{i,j}/\varepsilon}$. Strong duality holds for the primal and the dual problem. The minimization is attained for a unique $P^*_\varepsilon$ for the primal problem* (3.66)*. And $\phi^*$, $\psi^*$ maximize the dual problem* (3.67) *if and only if:*

$$(P^*_\varepsilon)_{i,j} = e^{\phi_i^*/\varepsilon} K_{i,j} e^{\psi_j^*/\varepsilon}. \tag{3.68}$$

**Proposition 3.27.** *Given matrix $K$, coefficient $\varepsilon$ and $\varepsilon_u$ in consistent with Theorem 3.26. Suppose $\phi^*$, $\psi^*$ solve the dual problem* (3.67)*, let $(u^*, v^*) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+$ with $u_i^* = e^{\phi_i^*/\varepsilon}$ and $v_j^* = e^{\psi_j^*/\varepsilon}$. Then,*

$$u_i^* = \left(\frac{a_i}{\sum_j K_{i,j} v_j^*}\right)^{\frac{\varepsilon_u}{\varepsilon_u+\varepsilon}}, \quad v_j^* = \left(\frac{b_j}{\sum_i K_{i,j} u_i^*}\right)^{\frac{\varepsilon_u}{\varepsilon_u+\varepsilon}}. \tag{3.69}$$

The above proposition can be easily checked by computing the first order optimality condition of the dual problem (3.67). The following remark provides an algorithm to compute the unbalanced optimal transport distance with the entropy regularization as Definition 3.25.

**Remark 3.28.** *Starting with an initial value $v^{(0)} = \mathbf{1}_m$, the dual problem can be computed through a coordinate ascent algorithm: for the $k$-th iteration,*

$$u_i^{(k+1)} = \left(\frac{a_i}{\sum_j K_{i,j} v_j^{(k)}}\right)^{\frac{\varepsilon_u}{\varepsilon_u+\varepsilon}}, \quad v_j^{(k+1)} = \left(\frac{b_j}{\sum_i K_{i,j} u_i^{(k+1)}}\right)^{\frac{\varepsilon_u}{\varepsilon_u+\varepsilon}}. \tag{3.70}$$

*Suppose the coordinate ascent algorithm converges with $u^*, v^*$, the transport plan matrix $P_\varepsilon^*$ in* (3.66) *can be computed as*

$$(P_\varepsilon^*)_{i,j} = u_i^* K_{i,j} v_j^*. \tag{3.71}$$

Also, the gradient of regularized unbalanced optimal transport distance can be achieved through the following remark.

**Remark 3.29.** *Suppose $P^*$, $\phi^*$ and $\psi^*$ solve the primal problem* (3.66) *and dual problem* (3.67)*, the gradient of regularized unbalanced optimal transport distance with respect to a is:*

$$\nabla_a W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta) = -\varepsilon_u \left(e^{-\phi^*/\varepsilon_u} - 1\right). \tag{3.72}$$

The algorithm to compute the regularized unbalanced optimal transport distance and the gradient is given by Algorithm 3. For more information about the above remarks and Algorithm 3, please refer to the work [39].

**Algorithm 3:** Scaling algorithm for regularized UOT distance and gradient

Input: $C$, $\varepsilon_u$, $\varepsilon$
Initialization: $K_{i,j} = e^{-C_{i,j}/\varepsilon}$, $v = \mathbf{1}_m$,
**while** *not converged* **do**
  Update vector $u$ with $u_i = a_i/(Kv)_i^{\varepsilon_u/(\varepsilon_u+\varepsilon)}$.
  Update vector $v$ with $v_j = b_j/(K^{'}u)_j^{\varepsilon_u/(\varepsilon_u+\varepsilon)}$.
**end**
Compute transport matrix $P_\varepsilon$ with $(P_\varepsilon)_{i,j} = u_i K_{i,j} v_j$
**return** *The regularized UOT distance:*

$$\begin{aligned} W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta) = &\sum_{i,j}(P_\varepsilon)_{i,j}C_{i,j} + \varepsilon_u\left((P_\varepsilon\mathbf{1}_m)_i \log\left(\frac{(P_\varepsilon\mathbf{1}_m)_i}{a_i}\right) - (P_\varepsilon\mathbf{1}_m)_i + a_i\right) \\ &+ \varepsilon_u\left((P_\varepsilon^{'}\mathbf{1}_n)_j \log\left(\frac{(P_\varepsilon^{'}\mathbf{1}_n)_j}{b_j}\right) - (P_\varepsilon^{'}\mathbf{1}_n)_j + b_j\right). \end{aligned}$$

*The gradient of $W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta)$ with respect to a is $(\nabla_a W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta))_i = -\varepsilon_u\left(u_i^{-\varepsilon/\varepsilon_u} - 1\right)$.*

Similar to the Sinkhorn algorithm, a special algorithm can be derived to compare the discrete measures defined on $\mathbb{R}^2$.

**Algorithm 4:** Scaling algorithm for regularized UOT distance and gradient in $\mathbb{R}^2$

Input: discrete probability measures $\alpha, \beta$ on $\mathbb{R}^2$ with densities $a, b \in \mathbb{R}^{N\times M}$, $\varepsilon > 0$. The cost matrix $C_1$ and $C_2$ defined by equation (3.53).

Initialization: matrix $K_1$, $K_2$ with $(K_1)_{i,j} = e^{-(C_1)_{i,j}/\varepsilon}$, $(K_2)_{i,j} = e^{-(C_2)_{i,j}/\varepsilon}$. Matrix $u, v$ are initialized as $N \times M$ matrices with all entries be 1. Let $d = 0$, $A, B \in \mathbb{R}^{N\times M}$ with entries be 1.

**while** *not converged* **do**

  Update vector $u$ with $u_{i,j} = a_{i,j}/(K_2 v K_1^{'})_{i,j}^{\varepsilon_u/(\varepsilon_u+\varepsilon)}$.

  Update vector $v$ with $v_{i,j} = b_{i,j}/(K_2^{'} u K_1)_{i,j}^{\varepsilon_u/(\varepsilon_u+\varepsilon)}$.

**end**

**for** $i_1 = 1 : N$ **do**

  **for** $i_2 = 1 : M$ **do**

    **for** $j_1 = 1 : N$ **do**

      **for** $j_2 = 1 : M$ **do**

        $A_{i_1,i_2}$ += $u_{i_1,i_2}(K_2)_{i_2,j_2}(K_1)_{i_1,j_1}v_{j_1,j_2}$.

        $B_{j_1,j_2}$ += $u_{i_1,i_2}(K_2)_{i_2,j_2}(K_1)_{i_1,j_1}v_{j_1,j_2}$.

        $d$ += $u_{i_1,i_2}(K_2)_{i_2,j_2}(K_1)_{i_1,j_1}v_{j_1,j_2}((C_1)_{i_1,j_1} + (C_2)_{i_2,j_2})$.

      **end**

    **end**

  **end**

**end**

**return** *The regularized UOT distance:*

$$W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta) = d + \sum_{i_1,i_2} \varepsilon_u \left( A_{i_1,i_2} \log\left(\frac{A_{i_1,i_2}}{a_{i_1,i_2}}\right) - A_{i_1,i_2} + a_{i_1,i_2} \right) + \sum_{j_1,j_2} \varepsilon_u \left( B_{j_1,j_2} \log\left(\frac{B_{j_1,j_2}}{b_{j_1,j_2}}\right) - B_{j_1,j_2} + b_{j_1,j_2} \right).$$

*The gradient of* $W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta)$ *with respect to* $a$ *is* $(\nabla_a W^2_{2,\varepsilon_u,\varepsilon}(\alpha,\beta))_{i_1,i_2} = -\varepsilon_u \left( u_{i_1,i_2}^{-\varepsilon/\varepsilon_u} - 1 \right)$.

# Chapter 4

# Full waveform inversion with optimal transport based distance

In this chapter, we introduce the method of solving the full waveform inversion problem with optimal transport (OT) based distance and the entropy regularization approach.

Although the unbalanced optimal transport (UOT) distance generalizes the OT distance, the geometric properties are only approximately inherited. A new mixed $L^1$/Wasserstein distance which is a combination of Wasserstein distance and a mass balancing term is proposed in this chapter, and it is well defined for the discrete measure in $\mathbb{R}^d$ without the mass equality constraint. Furthermore, we show that the proposed mixed distance is indeed a metric for the discrete measures. When the mixed distance is applied in the variational problem, we show that the objective function is convex with respect to shift, dilation, and amplitude change. Following the entropy regularization of the optimal transport problem, the approximated gradient of the mixed distance can be evaluated efficiently. Due to the historical issue, the name of optimal transport distance and the Wasserstein distance are interchangeably used. We call the UOT distance with the name "optimal transport" in this work based on the conventions in the main references [38, 39]. And we call the proposed mixed distance with the name "Wasserstein" based on the convention in the work [12], which proposed a mixed $L^2$/Wasserstein distance.

Since we focus on developing numerical optimization methods, the discrete measures are considered in this chapter. The optimal transport problem was designed to compare the difference between probability measures, i.e. positive measures with equal total mass. The UOT distance and the proposed mixed $L^1$/Wasserstein distance can overcome the mass equality limitation. However, normalization methods that transform the signals to positive functions are still needed to compare the difference between seismic signals.

Similar procedures are used in several works [142, 50, 53, 138, 139, 137, 93, 95].

The mixed $L^1$/Wasserstein distance is proposed in Section 4.1. In Section 4.2, we review some reliable normalization methods and discuss the normalization parameter selection with numerical examples. Then, we introduce the method to compute the adjoint source which is an important part of the adjoint state method for the gradient computation. The trace-by-trace technique is used due to the large size of the seismic signal in the real case. Numerical examples of the full waveform inversion problem are provided, including a two-parameter two-layer toy model, a cross-well model, and a more realistic Marmousi model.

## 4.1 Mixed $L^1$/Wasserstein distance

A mixed $L^1$/Wasserstein distance is constructed in this section. The concept of the mixed distance with Wasserstein distance is not new. In the work of [12], a mixed $L^2$/Wasserstein is provided through the dynamic form of optimal transport problem. Our initial idea is to generalize the 2-Wasserstein distance and maintain two properties:

1. The mixed distance can handle the discrete measures defined on a nonempty, closed, and bounded subset in $\mathbb{R}^d$. In this case, the distance can be used to represents some physical properties and can leads to a discrete form for computation.

2. The objective function of the mixed distance should keep the convex properties with respect to the shift and dilation as the square of 2-Wasserstein distance.

Given a closed and bounded set $\Omega \subset \mathbb{R}^d$, we start with the positive discrete measures defined as

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \beta = \sum_{i=1}^{m} b_i \delta_{y_i}, \tag{4.1}$$

where $a = (a_1, \cdots, a_n) \in \mathbb{R}_+^n$, $b = (b_1, \cdots, b_m) \in \mathbb{R}_+^m$, and the sampling points $X = \{x_1, \cdots, x_n\} \subset \Omega$, $Y = \{y_1, \cdots, y_m\} \subset \Omega$. Since $a$ and $b$ are vectors with positive entries, we can use the $l_1$ norm to denote the mass of the measure as $\sum_{i=1}^n a_i = \|a\|_1$, and $\sum_{i=1}^m b_i = \|b\|_1$. Denote the normalized $\alpha$ and $\beta$ as

$$\hat{\alpha} = \frac{1}{\|a\|_1} \sum_{i=1}^{n} a_i \delta_{x_i} = \sum_{i=1}^{n} \hat{a}_i \delta_{x_i}, \quad \hat{\beta} = \frac{1}{\|b\|_1} \sum_{i=1}^{m} b_i \delta_{y_i} = \sum_{i=1}^{m} \hat{b}_i \delta_{y_i}. \tag{4.2}$$

It is straight forward to see that $\hat{\alpha}, \hat{\beta} \in \mathcal{P}_2(\Omega)$, and $\hat{a} = (\hat{a}_1, \cdots, \hat{a}_n) \in \Sigma_n$, $\hat{b} = (\hat{b}_1, \cdots, \hat{b}_m) \in \Sigma_m$. The 2-Wasserstein distance between $\hat{\alpha}$ and $\hat{\beta}$ is well defined.

The following definition provides the mixed $L^1$/Wasserstein distance.

**Definition 4.1** (Mixed $L^1$/Wasserstein distance)**.** *Given two discrete measures $\alpha$ and $\beta$ as equation* (4.1)*, the cost matrix $C$ is defined by $C_{i,j} = |x_i - y_j|^2$. The mixed $L^1$/Wasserstein distance between $\alpha$ and $\beta$ is*

$$\bar{W}_2(\alpha, \beta) = W_2(\hat{\alpha}, \hat{\beta}) + |\|a\|_1 - \|b\|_1|, \tag{4.3}$$

*where $\hat{\alpha}$ and $\hat{\beta}$ are the normalized measures given by equation* (4.2)*.*

In the above definition, the first term is the 2-Wasserstein distance between normalized measures $\hat{\alpha}$ and $\hat{\beta}$, which describes the shape difference between $\alpha$ and $\beta$. The second term is the $L^1$ term, which is the absolute value of the mass difference between $\alpha$ and $\beta$. The following proposition describes the metric property of the above mixed distance.

**Proposition 4.2.** *The mixed $L^1$/Wasserstein distance defined in equation* (4.3) *is a distance over the set of positive discrete measures in $\mathcal{M}_d(\Omega)$.*

*Proof.* It is clear that $\bar{W}_2(\alpha, \beta)$ is nonnegative. Suppose $\alpha = \beta$, then $\|a\|_1 = \|b\|_1$, and $|\|a\|_1 - \|b\|_1| = 0$. Also $W_2(\hat{\alpha}, \hat{\beta}) = 0$ by the metric property of $W_2$, we have $\bar{W}_2(\alpha, \beta) = 0$. On the other hand, if $\bar{W}_2(\alpha, \beta) = 0$ then $W_2(\alpha, \beta) = 0$ and $|\|a\|_1 - \|b\|_1| = 0$, which leads to $\alpha = \beta$.

The symmetric property of $\bar{W}_2(\alpha, \beta)$ comes directly from the metric property of $W_2(\alpha, \beta)$.

Denote $\gamma \in \mathcal{M}_d(\Omega)$ with $\gamma = \sum_{i=1}^k c_i \delta_{z_i}$, where $c \in \mathbb{R}_+^k$ and the set of sampling point $Z = \{z_1, \cdots, z_k\} \subset \Omega$. We then observe:

$$\begin{aligned}
\bar{W}_2(\alpha, \gamma) &= W_2(\hat{\alpha}, \hat{\gamma}) + |\|a\|_1 - \|c\|_1| \\
&= W_2(\hat{\alpha}, \hat{\gamma}) + |\|a\|_1 - \|b\|_1 + \|b\|_1 - \|c\|_1| \\
&\leq W_2(\hat{\alpha}, \hat{\beta}) + W_2(\hat{\beta}, \hat{\gamma}) + |\|a\|_1 - \|b\|_1| + |\|b\|_1 - \|c\|_1| \\
&= \bar{W}_2(\alpha, \beta) + \bar{W}_2(\beta, \gamma).
\end{aligned} \tag{4.4}$$

The inequality follows by the metric property of 2-Wasserstein distance and the triangle inequality. □

The Wasserstein distances have been widely applied in different areas, for instance, imaging restoration, tomographic inversion, density regularization, sparse recovery, and seismic inversion [106]. To extend the application scene from the discrete probability measures to the discrete measures, the above mixed $L^1$/Wasserstein distance can be used. For the variational problem when a discrete measure $\beta$ is given, we are going to apply the optimization algorithm to find a measure $\alpha$ which is close to $\beta$. Since

$$\bar{W}_2^2(\alpha, \beta) = \left(W_2(\hat{\alpha}, \hat{\beta}) + |\|a\|_1 - \|b\|_1|\right)^2 \leq 2\left(W_2^2(\hat{\alpha}, \hat{\beta}) + (\|a\|_1 - \|b\|_1)^2\right), \tag{4.5}$$

define

$$J(\alpha;\beta) = W_2^2(\hat{\alpha},\hat{\beta}) + \left(\|a\|_1 - \|b\|_1\right)^2. \tag{4.6}$$

In this case, to minimize the mixed $L^1$/Wasserstein distance between $\alpha$ and $\beta$, it is sufficient to minimize the objective function $J(\alpha;\beta)$ with respect to $\alpha$.

By the construction of the discrete measure $\alpha$, the objective function is controlled by vector $a$ and sampling set $X$, i.e.

$$J(\alpha;\beta) = J(a,X). \tag{4.7}$$

The following proposition provides the subdifferentiability of the objective function with respect to $a$.

**Proposition 4.3.** *The objective function $J(a,X)$ defined by equation (4.6) and (4.7) is subdifferentiable with respect to $a$.*

*Proof.* The subdifferentiability of $J(a,X) = J(\alpha;\beta)$ with respect to vector $a$ follows directly from Proposition 3.19. □

The following propositions show that the objective function (4.7) with the proposed mixed distance retains the convex properties of the square of 2-Wasserstein distance.

**Proposition 4.4.** *Given two discrete measures $\alpha$ and $\beta$. Given the shift direction $\eta \in \mathbb{R}^d$ and the shift size $s > 0$, the shift of discrete measure $\alpha$ is denoted as*

$$\alpha_s = \sum_{i=1}^{n} a_i \delta_{x_i + s\eta}, \tag{4.8}$$

*here the shift size $s$ is small enough such that $\alpha_s$ is defined on $\Omega$. The objective function $J(s) = J(\alpha_s;\beta)$ is defined by equation (4.6). Then $J(s)$ is convex with respect to $s$.*

*Proof.* By the construction of the objective function

$$J(s) = J(\alpha_s,\beta) = W_2^2(\hat{\alpha}_s,\hat{\beta}) + \left(\|a\|_1 - \|b\|_1\right)^2. \tag{4.9}$$

The convexity of $J(s)$ follows since $W_2^2(\hat{\alpha}_s,\hat{\beta})$ is convex with respect to $s$ by Theorem 3.20. □

**Proposition 4.5.** *Given a discrete measure $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$, where $a \in \mathbb{R}_+^n$ and $x_i \in \Omega$ for $i = 1, \cdots, n$.*

*The dilation of measure $\alpha$ is given by*

$$\alpha_A = \sum_{i=1}^{n} a_i \delta_{Ax_i}, \tag{4.10}$$

*here $A$ is a dilation transform matrix which is symmetric positive definite, and $Ax_i \in \Omega$ for $i = 1, \cdots, n$. Let $\lambda_1, \cdots, \lambda_d$ be the eigenvalues of matrix $A$. The objective function $J(\lambda_1, \cdots, \lambda_d) = J(\alpha_A; \alpha)$ is defined by equation (4.6). Then $J(\lambda_1, \cdots, \lambda_d)$ is convex with respect to $\lambda_1, \cdots, \lambda_d$.*

*Proof.* By the construction of the objective function

$$J(\lambda_1, \cdots, \lambda_d) = J(\alpha_A; \alpha) = W_2^2(\hat{\alpha}_A, \hat{\alpha}) + \left(\|a\|_1 - \|a\|_1\right)^2. \tag{4.11}$$

The convexity of $J(\lambda_1, \cdots, \lambda_d)$ follows that $W_2^2(\hat{\alpha}_A, \hat{\alpha})$ is convex with respect to $\lambda_1, \cdots, \lambda_d$ by Theorem 3.21. □

**Proposition 4.6.** *Given two discrete measures $\alpha$ and $\beta$. Denote the mass change of measure $\alpha$ by*

$$\alpha_k = \sum_{i=1}^{n} (ka_i)\delta_{x_i}, \tag{4.12}$$

*where $k \in (0, \infty)$. The objective function $J(k) = J(\alpha_k; \beta)$ is defined by equation (4.6). Then $J(k)$ is convex with respect to $k$.*

*Proof.* By the construction of the objective function

$$J(k) = J(\alpha_k; \beta) = W_2^2(\hat{\alpha}_k, \hat{\beta}) + \left(k\|a\|_1 - \|b\|_1\right)^2. \tag{4.13}$$

Since $\hat{\alpha}_k = \hat{\alpha}$, then the convexity of $J(k)$ follows by the second term of the above equation. □

The numerical evaluation of $J(\alpha; \beta)$ is straightforward. Suppose the support of $\alpha$ and $\beta$ are fixed, the entropy regularization method can be used to define a regularized $J_\varepsilon$ with

$$J_\varepsilon(\alpha; \beta) = W_{2,\varepsilon}^2(\hat{a}, \hat{b}) + (\|a\|_1 - \|b\|_1)^2. \tag{4.14}$$

The first part of the $J_\varepsilon(\alpha; \beta)$ can be evaluated with the Sinkhorn algorithm 1 and 2. When the supports of $\alpha$ and $\beta$ are fixed, the regularized objective function $J_\varepsilon(\alpha; \beta)$ can be written as $J_\varepsilon(a, b)$, and the gradient of

$J_\varepsilon(a;b)$ with respect to $a$ is given by

$$\nabla J_\varepsilon(a;b) = (D\hat{a})'\nabla_{\hat{a}} W_{2,\varepsilon}^2(\hat{a},\hat{b}) + 2(\|a\|_1 - \|b\|_1)\mathbf{1}_n, \tag{4.15}$$

where $\nabla_{\hat{a}} W_{2,\varepsilon}^2(\hat{a},\hat{b})$ is the gradient of the regularized Wasserstein distance with respect to the first entry. The Jacobian matrix $D\hat{a}$ is given by

$$(D\hat{a})_{i,j} = \frac{\partial \hat{a}_i}{\partial a_j} = \begin{cases} -\frac{a_i}{(\sum_k a_k)^2}, & \text{if } i \neq j, \\ \frac{1}{\sum_k a_k} - \frac{a_i}{(\sum_k a_k)^2}, & \text{if } i = j. \end{cases} \tag{4.16}$$

In practice, a mass balancing parameter $\lambda_m > 0$ can be introduced to the $J_\varepsilon$ as

$$J_{\varepsilon,\lambda_m}(\alpha;\beta) = W_{2,\varepsilon}^2(\hat{a},\hat{b}) + \lambda_m(\|a\|_1 - \|b\|_1)^2. \tag{4.17}$$

## 4.2 Normalization methods for signals

In this section, we discuss several normalization methods that transform the seismic signals into positive functions. Instead of focusing on the theoretical properties of the normalization methods, we are working with numerical examples to show how the normalizations behaves. The normalization methods here can only partially solve the problem that how to generalize the optimal transport distance to compare the difference between signals. However, the numerical examples provided in this section and in the following sections show that by introducing the optimal transport based distance, the inverse results of the full waveform inversion problem are indeed improved in certain cases.

Since the trace-by-trace strategy is going to be used, we focus on comparing the difference between one-dimensional signals with UOT distance and the mixed $L^1$/Wasserstein distance. The signal $a(t)$ and $b(t)$ are defined on the time domain. When $t = (\delta_{t_1}, \cdots, \delta_{t_n})$ is fixed, the signals $a(t)$ and $b(t)$ can be represented as $n$-dimensional vectors $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$, $b = (b_1, \cdots, b_n) \in \mathbb{R}^n$. And the discrete signals are defined by $\alpha = \sum_i a_i \delta_{t_i}$, $\beta = \sum_j b_j \delta_{t_j}$.

The $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance are going to be compared in several numerical experiments, and we use $d(\cdot,\cdot)$ to represent the distance used in the objective functions. Instead of working on the distance between $\alpha$ and $\beta$, we are working on the distance between $a$ and $b$ since the vector

$t$ is fixed as the sampling settings of the signals. Let $d_u(\cdot,\cdot)$ represent the UOT distance

$$d_u(a,b) = W^2_{2,\varepsilon_u,\varepsilon}(a,b), \tag{4.18}$$

and $d_m(\cdot,\cdot)$ represent the mixed $L^1$/Wasserstein distance,

$$d_m(a,b) = J(a;b) = W^2_{2,\varepsilon}(\hat{a},\hat{b}) + \lambda_m(\|a\|_1 - \|b\|_1)^2, \tag{4.19}$$

where $J(a;b)$ is defined by equation (4.6). As discussed in the previous chapter, both UOT distance and mixed $L^1$/Wasserstein distance can be evaluated through the entropy regularization approach. And the smaller the regularization coefficient $\varepsilon$ is used, the more accurate result we can achieve. On the other hand, the regularization coefficient $\varepsilon$ can not be too small due to the machine precision. In this work, the regularization coefficients are chosen as small as possible in the numerical experiments.

### 4.2.1 Review of some normalization methods

The first normalization method in our discussion is the Mainini strategy [90], which separates the signals into positive and negative parts. We set

$$a = a^+ - a^-, \quad b = b^+ - b^-, \tag{4.20}$$

where $\cdot^+$ is the positive part of the vector, and $\cdot^-$ is the absolute value of the negative part of the vector. Then the Mainini strategy is given by

$$d(a,b) = W_p^p(a^+ + b^-, b^+ + a^-). \tag{4.21}$$

A special form can be derived when $p = 1$. Consider the continuous case for the density function $a(t)$ and $b(t)$, the dual of the Kantorovich problem is given by:

$$W_p^p(a,b) = \max_{\phi,\psi} \int \phi(t)a(t) + \psi(t)b(t)\ \mathrm{d}t, \tag{4.22}$$

such that $\phi(t_1) + \psi(t_2) \le |t_1 - t_2|$. As $p = 1$, we can claim that $\psi = -\phi$, and $\phi \in \mathrm{Lip}_1$ which is the space of all 1-Lipschitz functions. In this case, we have

$$W_1(a,b) = \max_{\phi\in\mathrm{Lip}_1} \int \phi(t)(a(t) - b(t))\ \mathrm{d}t. \tag{4.23}$$

We refer to the monograph [106] (Chapter 6) for more information about the Kantorovich problem as $p = 1$. Based on the above equation, we can have the following equation

$$\begin{aligned} d(a,b) &= W_1(a^+ + b^-, b^+ + a^-) \\ &= \max_{\phi \in \mathrm{Lip}_1} \int \phi(t)(a^+(t) + b^-(t) - b^+(t) - a^-(t))\, \mathrm{d}t \\ &= W_1(a^+ - a^-, b^+ - b^-) = W_1(c,d), \end{aligned} \tag{4.24}$$

where $c = a^+ + b^-$ and $d = b^+ + a^-$. The above equations extend the definition of the Kantorovich problem from the positive measures to the signed measures. Based on the above equation, the Mainini strategy is actually comparing the vector $c$ and $d$. The connection between Mainini strategy and Kantorovich–Rubinstein (KR) norm, and the application can be found in the work [79, 93]. This strategy has been also introduced in the full waveform inversion problem in the work [142].

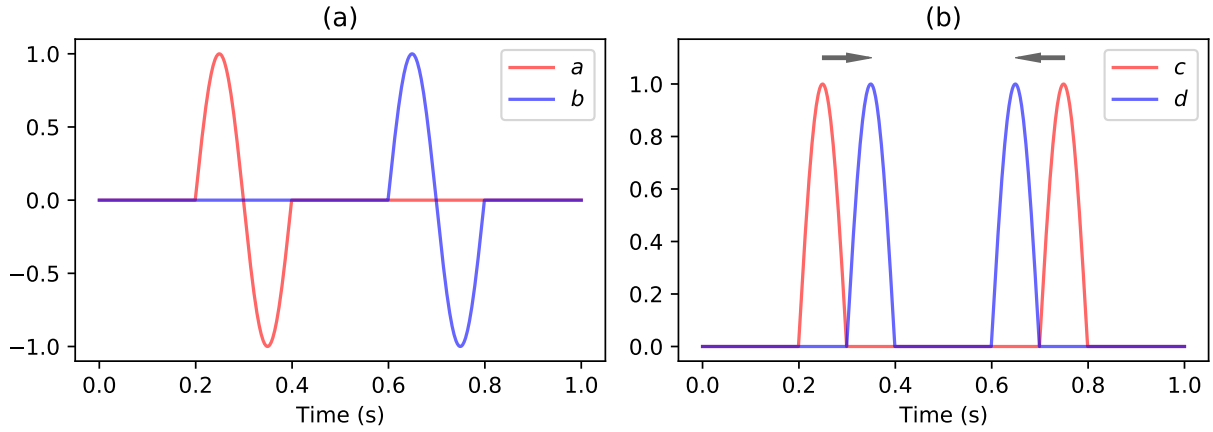


Figure 4.1: (a): Signal $a$ and $b$. (b): Comparing signal $a$ and $b$ with Mainini strategy.

However, problems occur when the Mainini strategy is applied for the signals. For example, two signals $a$ and $b$ with one period of the sine function are given in Figure 4.1 (a). The case when comparing $a$ and $b$ with the Mainini strategy is given by figure (b). When the support of $a$ and $b$ are far away enough from each other, the transport plan is given by the black arrows in figure (b), which moves the positive parts of $a$ to the negative parts of $a$, and move the negative parts of $b$ to the positive parts of $b$. Suppose the signal $b$ is shifted towards the right direction, the transport plan stays the same, so does the transport cost. That means, when the signal $b$ moves in the right direction, the distance between $a$ and $b$ is not changing under the Mainini strategy. This is not ideal for describing the waveform propagation.

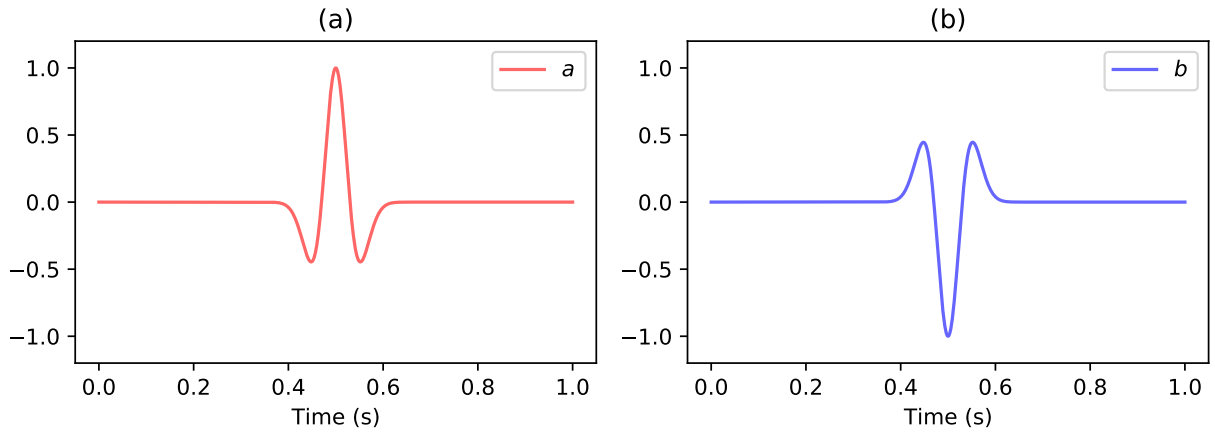


Figure 4.2: (a): Ricker wavelet $a$. (b): Ricker wavelet $b$.

The normalization method with square scaling function has been discussed in [138], defined as

$$\hat{a}(t) = a(t)^2. \tag{4.25}$$

The square normalization can not distinguish the case when the compared signals have the phase difference of $\pi$, as shown in Figure 4.2. This case is important for the seismic signals since the reflectivity of the earth medium will change the phase of reflection signals. The same reason exists for the absolute value normalization, i.e.,

$$h(a)(t) = |a(t)|. \tag{4.26}$$

The next normalization method is to compare the Wasserstein distance for the signals with positive and negative parts separately:

$$d(a, b) = W_2^2(a^+, b^+) + W_2^2(a^-, b^-), \tag{4.27}$$

where the $\cdot^+$ and $\cdot^-$ are the positive and negative parts of the entries as equation (4.20). Suppose for $a \in \mathbb{R}^n$, there exists a linear operator such that $a^- = P^-(a)a = \langle p^-(a), a\rangle$, where $p^-(a) \in \mathbb{R}^n$. Notice that the linear operator $P^-(a)$ is depending on $a$, and it is not differentiable with respect to $a$.

One counter example can be given as: let $a = (0, 0) \in \mathbb{R}^2$, direction $v_1 = (1, 0)$ and $v_2 = (-1, 0)$. Let

$A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, then we have

$$\lim_{t\to 0} \frac{\|p^-(a+tv_1) - p^-(a) - tAv_1\|}{\|tv_1\|} = \lim_{t\to 0} \frac{\|(0,0)' - (0,0)' - tA(1,0)'\|}{\|t(1,0)\|} = 0. \tag{4.28}$$

On the other hand,

$$\lim_{t\to 0} \frac{\|p^-(a+tv_2) - p^-(a) - tAv_1\|}{\|tv_2\|} = \lim_{t\to 0} \frac{\|(1,0)' - (0,0)' - tA(1,0)'\|}{\|t(-1,0)\|} \neq 0. \tag{4.29}$$

The objective function will not be differentiable when this normalization is applied in the variational problem. The same problem exists for another sign-sensitive normalization proposed by the work [137] (Section 5.2.3) and the absolute value normalization (4.26). Given coefficients $k$ and $l$, the sign-sensitive normalization in [137] is defined as

$$h(a)(t) = \begin{cases} \frac{a(t)+\frac{1}{k}}{l}, & \text{if } a(t) > 0, k > 0, \\ \frac{1}{k}\frac{e^{kf(t)}}{l}, & \text{if } a(t) < 0. \end{cases} \tag{4.30}$$

We focus on the linear and exponential normalizations in this work. Given a normalization parameter $k$, the linear normalization is defined as

$$h_{\mathrm{l}}(a,k)(t) = a(t) + k, \tag{4.31}$$

and the exponential normalization is defined as

$$h_{\mathrm{e}}(a,k)(t) = e^{ka(t)}. \tag{4.32}$$

We demonstrate the behavior of the above two normalizations with numerical examples in the following subsection.

### 4.2.2 Numerical examples for the normalization methods

As discussed in the previous chapter, the optimal transport distance is convex for shift and dilation, and that is our initial idea to introduce the optimal transport distance to the seismic inverse problem. The normalization methods are needed to extend the optimal transport distance to signals. However, the normalization methods will destroy the convex properties. In this section, we discuss the convex behavior with both linear

and exponential normalization for both UOT distance and mixed $L^1$/Wasserstein distance. Usually, the seismic event can be approximated with a linear combination of the Ricker wavelet, i.e.,

$$s(t) = \left(1 - \frac{(t-t_0)^2}{\sigma^2}\right) e^{\frac{-(t-t_0)^2}{2\sigma^2}}. \tag{4.33}$$

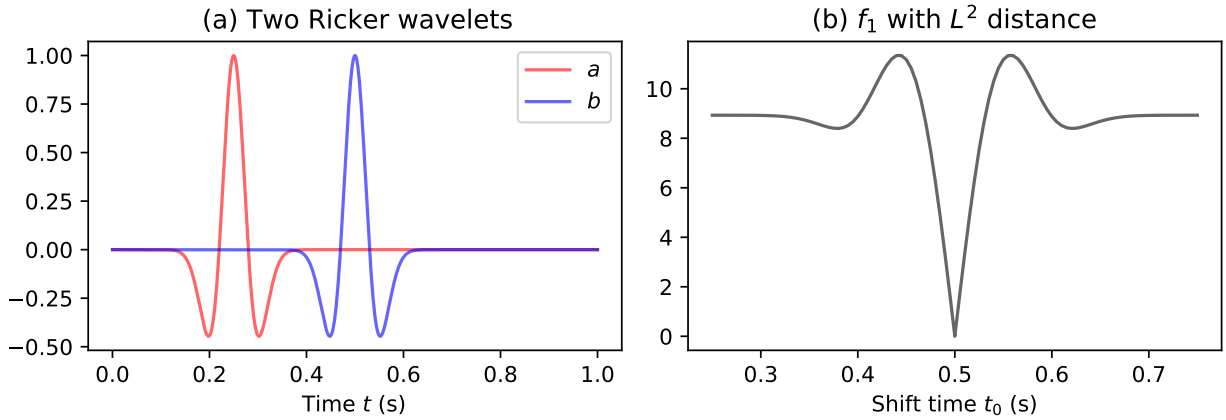


Figure 4.3: (a): Ricker wavelets $a$ and $b$. (b): The objective function $f_1(t_0)$ with $L^2$ distance.

First, we investigate the behavior for the time-shift of Ricker wavelets with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance. Let $\sigma = 0.03$, $t_0 \in [0.25, 0.75]$, the sampling frequency is 1000 Hz. Let $b$ be fixed with the center at 0.5 s, and $a$ is shifting from left to right, denote $a$ and $b$ as

$$a(t_0, t) = \left(1 - \frac{(t-t_0)^2}{0.03^2}\right) e^{\frac{-(t-t_0)^2}{2\times 0.03^2}}, \tag{4.34}$$

$$b(t) = \left(1 - \frac{(t-0.5)^2}{0.03^2}\right) e^{\frac{-(t-0.5)^2}{2\times 0.03^2}}, \tag{4.35}$$

as in Figure 4.3 (a). We fix $b$ as the reference signal and shift the center of $a$ from 0.25 s to 0.75 s. Define the objective function as

$$f_1(t_0, k) = d(h(a(t_0, t), k), h(b(t), k)), \tag{4.36}$$

where $d$ can be UOT distance and mixed $L^1$/Wasserstein distance as equation (4.18) and (4.19). No normalization method is needed for the $L^2$ distance. The normalization function $h(\cdot, \cdot)$ can be linear and the exponential normalization is defined by equation (4.31) and (4.32).

To evaluate the UOT distance and mixed $L^1$/Wasserstein distance we set the entropy regularization parameter $\varepsilon = 1 \times 10^{-3}$ to guarantee that the optimal transport distance is evaluated accurately. We set

$\varepsilon_u = 1$ in the UOT distance and set $\lambda_m = 1 \times 10^{-10}$ such that both UOT distance and mixed $L^1$/Wasserstein distance have notable results for the time-shift.

The normalized objective function $f_1(t_0)$ with $L^2$ distance is shown in Figure 4.3 (b). One global minimum and two local minima are observed, which is a sign of the cycle-skipping artifact.

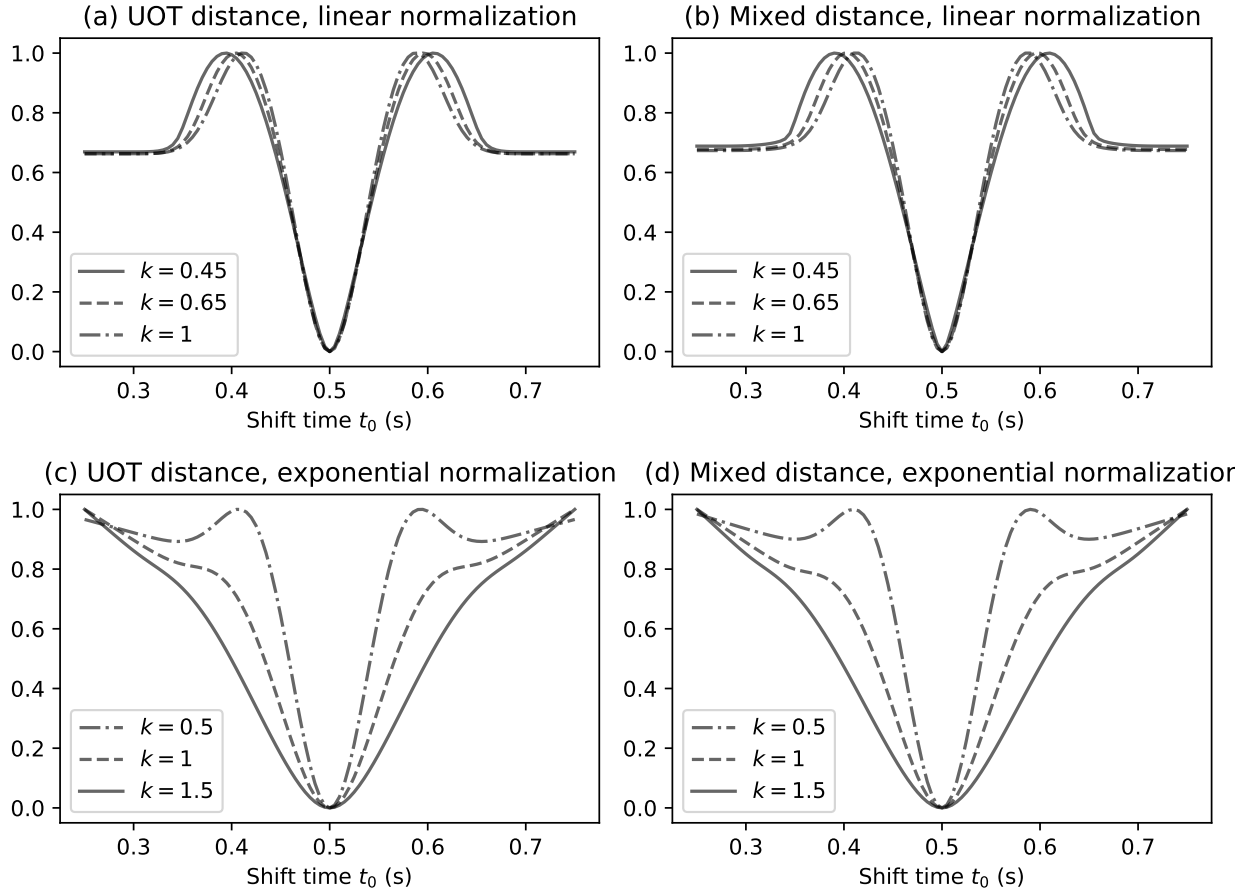


Figure 4.4: The normalized objective function $f_1(t_0, k)$ with UOT distance, mixed Wasserstein distance and linear normalization, exponential normalization.

The numerical results of normalized objective function $f_1(t_0, k)$ of both optimal transport based distances with both linear and exponential normalization are shown in Figure 4.4. Comparing subfigures (a), (b) with subfigures (c) (d), the shape of normalized objective functions are similar for both normalization methods. Compared to $L^2$ distance, the cycle-skipping artifact is slightly reduced by both distances with linear normalization as shown in subfigures (a) and (b). The smaller the normalization coefficient $k$ is used, the better performance can be achieved. However, $k$ can not be less than the absolute value of the minimal value of $a$ and $b$, which is approximately 0.446259 in this example. In subfigures (c) and (d), as $k = 0.5$, the normalized objective function is similar to the case of (a) and (b), i.e., with one global minimum and two

local minima. Only one global minimum is obtained with the case $k = 1$ and $k = 1.5$, which means no cycle-skipping issue occurs in this case. Compared to $L^2$ distance, both UOT distance and mixed $L^1$/Wasserstein distance can mitigate the cycle-skipping artifact with proper normalization coefficient $k$. Also, compared to the previous work in [93, 142], the UOT distance and mixed distance provide more convex behavior than the 1-Wasserstein distance with respect to the time-shift.

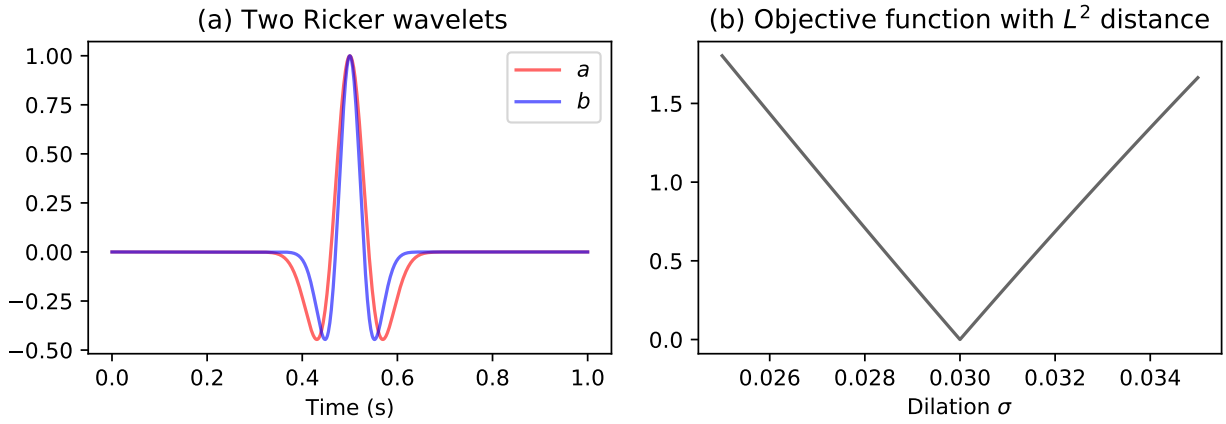


Figure 4.5: (a): Two Ricker wavelets $a$ and $b$. (b): The objective function $f_2(\sigma_0)$ with $L^2$ distance.

In the following example, we investigate the behavior with respect to the dilation of the Ricker wavelet. Fix $t_0 = 0.5$, let $\sigma_0 \in [0.02, 0.04]$, the sampling frequency is still 1000 Hz. Let $b$ be fixed with $\sigma = 0.03$, and a is dilating with the change of the $\sigma_0$, denote $a$ and $b$ as

$$a(\sigma_0, t) = \left(1 - \frac{(t-0.5)^2}{\sigma_0^2}\right) e^{\frac{-(t-0.5)^2}{2\sigma_0^2}}, \tag{4.37}$$

$$b(t) = \left(1 - \frac{(t-0.5)^2}{0.03^2}\right) e^{\frac{-(t-0.5)^2}{2\times 0.03^2}}. \tag{4.38}$$

One example is shown in Figure 4.5 (a). Define the objective function as

$$f_2(\sigma_0, k) = d(h(a(\sigma_0, t), k), h(b(t), k)), \tag{4.39}$$

where $d$ can be UOT distance and mixed $L^1$/Wasserstein distance as equation (4.18) and (4.19). No normalization method is needed for the $L^2$ distance. The normalization function $h(\cdot, \cdot)$ can be linear and exponential normalization is defined by equation (4.31) and (4.32). The computation coefficients $\varepsilon_u$, $\lambda_m$ are the same as in the previous shift Ricker example.

The normalized objective function $f_2(\sigma_0)$ with $L^2$ distance is shown in Figure 4.5 (b). Only one global minimum is observed and it is located at point $\sigma_0 = 0.03$, and in this case $a(\sigma_0, t) = b(t)$.

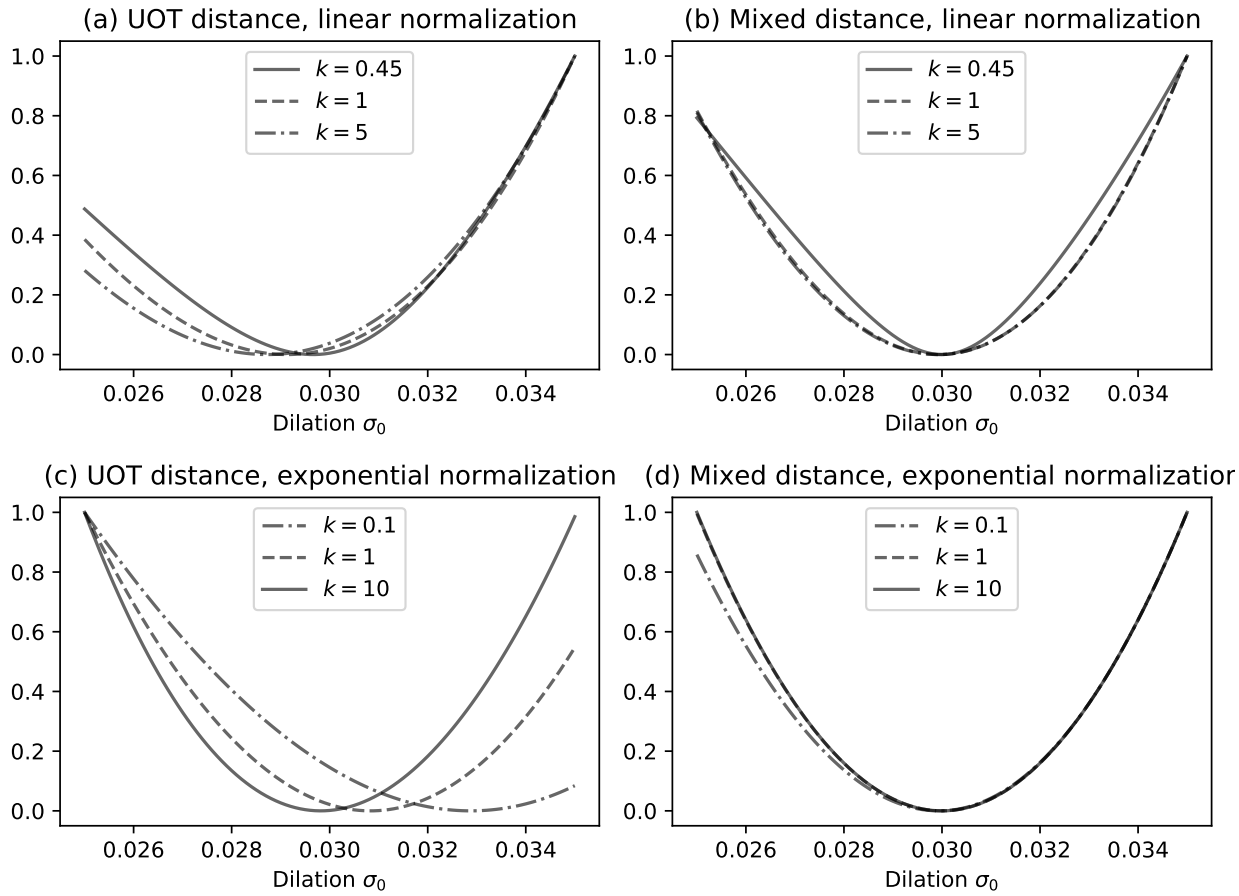


Figure 4.6: The normalized objective function $f_2(\sigma_0, k)$ with UOT distance, mixed $L^1$/Wasserstein distance and linear normalization, exponential normalization.

The numerical results of the normalized objective function $f_2(\sigma_0)$ of both optimal transport based distances with both linear and exponential normalization are shown in Figure 4.6. Subfigure (a) shows the result of UOT distance with linear normalization, only one global minimum can be observed for each $k$. However, when $k$ is larger, the position of the global minima tends to be less than 0.03 which is the global minima we expect. Subfigure (c) is the case of UOT distance with exponential normalization. The position of the global minima is larger than 0.03 when $k$ is small. The position of the global minima is gradually decreasing when $k$ is increasing, and it will be less than 0.03 when $k$ is large enough. For the case of mixed $L^1$/Wasserstein distance with both normalizations are shown in subfigures (b) and (d). There is only one global minimum in each of subfigures (b) and (d), and the global minima is close to (but may not equal to) the point $\sigma_0 = 0.03$ for different normalization coefficients $k$. Compared to the $L^2$ distance, both UOT distance and mixed $L^1$/Wasserstein distance can retain the convex property with respect to $\sigma_0$ when proper normalization and coefficient are chosen.

In conclusion, when the linear normalization method is used, the smaller normalization coefficient $k$ will lead to better convex behavior with respect to the time-shift for the UOT distance. However, the $k$ can not be arbitrarily small since it has to be larger than the absolute value of the minimum value of the signals. Therefore the linear normalization is not encouraged for use with the optimal transport based distances. On the other hand, both distances with a larger exponential normalization coefficient $k$ will retain the convex properties with respect to the time-shift and dilation. To avoid the significant distortion of the waveform, the normalization parameter of exponential normalization should not be too small or large. In practice, the normalization parameter should be chosen such that the normalized signal is approximately in the interval from 1 and 10. In this case, the wavefront of the seismic signal can be slightly amplified and the waveform is not significantly distorted.

Notice that the above experiments are designed to analyze the behavior of normalization methods based on the fact that the seismic signals can be approximated with a linear combination of Ricker wavelets. Although both UOT distance and mixed $L^1$/Wasserstein distance with proper normalization fulfill the requirement, these experiment results should not be considered as the theoretical results. And more detailed mathematical analysis is still expected to show how the normalization works for generalizing the optimal transport distance to signed measures and signals.

## 4.3 Applying the optimal transport based distances in full waveform inversion

In this section, we formulate the FWI problem with UOT distance and mixed $L^1$/Wasserstein distance by introducing the normalization methods discussed in the previous section. The wave equation in a two-dimensional domain is used as the constraint PDE, and the computation of adjoint sources is provided.

Consider there are $N_s$ sources and $N_r$ receivers in the domain, and let $s = 1, \cdots, N_s$, $r = 1, \cdots, N_r$ be the indexes of the sources and the receivers. Denote the objective function as

$$J(c, y_1, \cdots, y_{N_s}) = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} d(h(Q_r y_s), h(y_{\mathrm{d},s,r})), \tag{4.40}$$

where here $d$ is chosen as one of $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance. The function $h$ can be linear normalization or exponential normalization, and the normalization coefficient $k$ is omitted. When the $L^2$ distance is used in the objective function, no normalization method is needed. The operator $Q_r$ is the recording operator that maps the wavefield generated by the $s$-th source $y_s$ to the signals

received by the $r$-th receiver. The $y_{\mathrm{d},s,r}$ represents the received data by the $s$-th source and the $r$-th receiver.

The constraint PDE is given by

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}y_s - \Delta y_s = f_s, \quad s = 1, \cdots, N_s, \tag{4.41}$$

where $f_s$ is the function of the $s$-th source. In practice, a special technique such as absorbing boundary condition (ABC) or perfectly matched layer (PML) is needed to simulate the seismic wave propagating in an unbounded domain. A numerical PDE method such as finite difference or finite element method is needed to discretize the system and numerically simulate the wave propagation. We focus on the discrete form in this work.

Since the PDE is well-posed, it can be written in a compact form as $F_s(c) = y_s$. Then the reduced objective function is given by

$$f(c) = J(c, F_1(c), \cdots, F_{N_s}(c)). \tag{4.42}$$

The gradient of $f(c)$ can be achieved through the adjoint state method:

$$\nabla f(c) = \sum_{s=1}^{N_s} \int \frac{-2}{c^3}\left(\frac{\partial^2}{\partial t^2}u_s\right) v_s \, \mathrm{d}t. \tag{4.43}$$

Here $v_s$ is the solution of the adjoint equation with $s$-th source

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}v_s - \Delta v_s = \tilde{f}_s, \tag{4.44}$$

where $\tilde{f}_s$ is the adjoint source with respect to the $s$-th constraint equation. When $L^2$ distance is applied in the objective function, the adjoint source is given by

$$\tilde{f}_s = -\sum_{r=1}^{N_r} Q_r^{'}(Q_r y_s - y_{\mathrm{d},s,r}). \tag{4.45}$$

When the UOT distance and mixed $L^1$/Wasserstein distance with linear normalization is used in the objective function, the adjoint source is given by

$$\tilde{f}_s = -\sum_{r=1}^{N_r} Q_r^{'}\nabla_1 d(h(Q_r y_s), h(y_{\mathrm{d},s,r})), \tag{4.46}$$

where the $\nabla_1$ is the gradient of $d(\cdot,\cdot)$ with respect to the first term. When the UOT distance and mixed

$L^1$/Wasserstein distance with exponential normalization is used in the objective function, the adjoint source is given by

$$\tilde{f}_s = -\sum_{r=1}^{N_r} Q_r^{'} \left( k e^{k Q_r y_s} \right) \nabla_1 d(h(Q_r y_s), h(y_{\mathrm{d},s,r})). \tag{4.47}$$

Once the gradient $\nabla f(c)$ can be computed, the PDE constrained optimization problem can be solved by the gradient based optimization methods such as conjugate gradient method or L-BFGS method.

## 4.4 Numerical examples and discussion

Three full waveform inversion examples are provided in this section based on the formulation in the previous section. We compare the numerical results generated by $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance. The exponential normalization method is used for UOT distance and mixed $L^1$/Wasserstein distance. A two-parameter two-layer model is designed to compare the objective function generated by different distances in the first example. Next, a cross-well example is provided to show the update direction generated by different distances. The third numerical example is based on the standard Marmousi model. Both UOT distance and mixed $L^1$/Wasserstein distance outperforms the conventional $L^2$ distance in this example. In the end, we discuss the practical strategy for the general seismic inverse problem.

### 4.4.1 Example 1: Two-parameter two-layer model

This example shows the difference of objective functions between the $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance of a toy model. Due to the large size and nonlinear behavior of the FWI problem, we build a simplified two-parameter two-layer velocity model in two-dimension:

$$c(\delta c, z) = c_0(x, z) + \delta c H(z), \tag{4.48}$$

where $H(z)$ is the Heaviside step function along the $z$ direction. The factor $\delta c$ is the velocity perturbation of the bottom part for the background velocity $c_0(x, z)$. The background velocity is chosen to be homogeneous with $c_0(x, z) = 1$ km/s. The model is in a region with 1 km wide and 1 km deep, discretized into $101 \times 101$ grid points. Only one source is used in this example, located at the center of the model and 0.05 km depth with a 6 Hz Ricker wavelet. The sampling frequency is 300 Hz and the sampling time is 2 seconds. There are 11 equally spaced receivers at the top of the region.

Define the objective function as:

$$f_3(\delta c, z) = f(c(\delta c, z)), \tag{4.49}$$

where $f(\cdot)$ is defined by equation (4.42). The true model of this example is $\delta c = 0.05$, $z = 0.51$ which is shown in Figure 4.7. We set $\delta c \in [-0.1, 0.2]$ with step size 0.005, and $z \in [0.4, 0.6]$ with step size 0.01. Since there is a velocity perturbation between the two layers at the depth $z$, a reflective seismic wave is generated as the seismic wave propagating through the interface, and it will be recorded by the receivers at the top of the model. For different velocity models $c(\delta c, z)$, the position of the reflector $z$ controls the arriving time of the reflective wave, and the velocity difference $\delta c$ controls the amplitude of the reflective wave. We generate the recorded data with the true model $c(0.05, 0.51)$. As the $\delta c$ and $z$ are changing, the reflective waves will interact with the above recorded data, which will cause the cycle-skipping artifact. We evaluate $f_3$ for each $(\delta c, z)$ by using $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance respectively, the results are shown in Figure 4.8. Similar numerical examples with other kinds of optimal transport based distance are provided in the work [50, 95, 93].

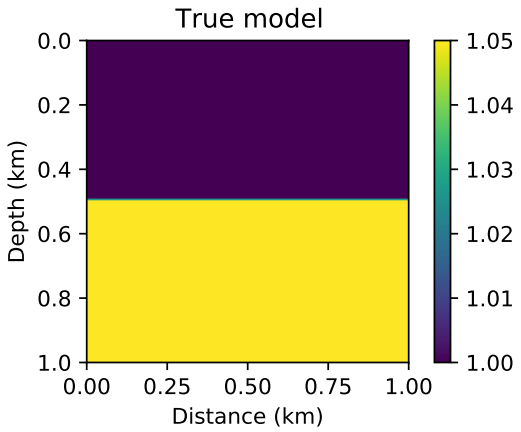


Figure 4.7: The true velocity model $c(0.05, 0.51)$.

In Figure 4.8, the $z$ axis is the normalized objective function $f_3(\delta c, z)$, and the other two axes are the perturbation $\delta c$ and the position $z$. The objective function with $L^2$ distance is shown in subfigure (a). Notice the global minimum is located at the point $(0.05, 0.51)$, and there are several wrinkles in the surface of the objective function around the global minimum. This suggests that when an initial model that is not close to the global minimum is provided, the optimization algorithm will be trapped in a local minimum due to the wrinkles.

The exponential normalization method is used in this example to compare the difference between signals

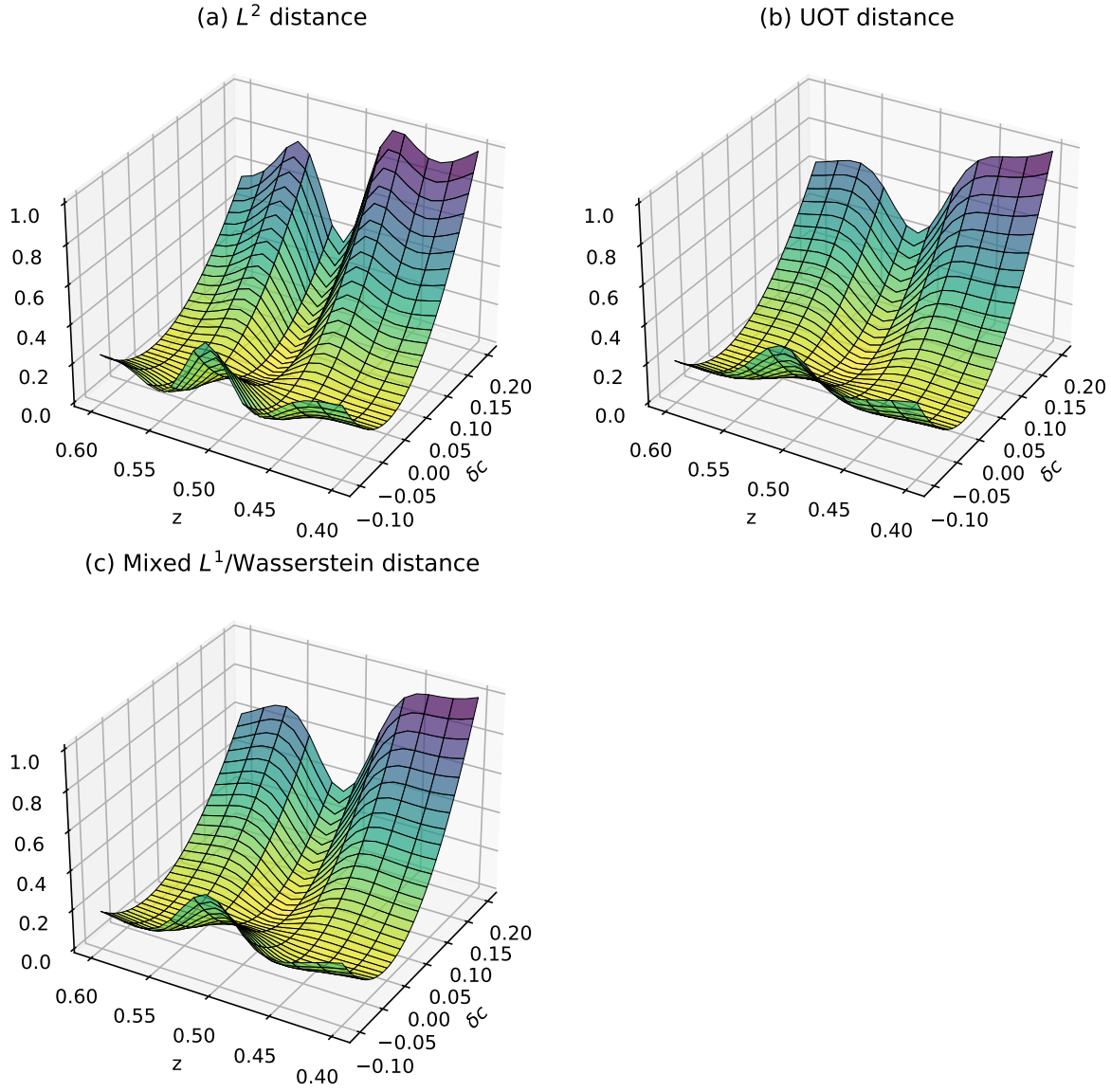


Figure 4.8: (a), (b), (c): the normalized objective function $f_3(\delta c, z)$ with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance.

with UOT distance and mixed $L^1$/Wasserstein distance. We set the normalization parameter $k = 5 \times 10^4$ such that the maximal value of the normalized signal is approximately in the interval between 1 and 10, and the entropy regularization parameter is $\varepsilon = 1 \times 10^{-4}$. For UOT distance, we set the coefficient of mass balancing term to be $\varepsilon_u = 1$. And we set the $\lambda_m = 1 \times 10^{-8}$ in the mixed $L^1$/Wasserstein distance. There are 500 iterations performed for the computation of UOT and mixed distance. This numerical example is performed on a server with the cpu model Intel Xeon CPU E7-8891 v4 @ 2.80GHz, and the code is written in the programming language Julia. There are 12 workers used in this numerical example. For each $(\delta c, z)$, $61 * 21 = 1281$ experiments are performed for UOT distance and the mixed distance, and 11 signals

are compared for each experiment. The computation time of the UOT example is 273 seconds, and the computation time of the mixed $L^1$/Wasserstein distance is 228 seconds.

The objective functions with UOT distance and mixed $L^1$/Wasserstein distance are shown in subfigures (b) and (c) respectively. Compared with subfigure (a), the surface in subfigures (b) and (c) have fewer wrinkle structures. For an initial model with $\delta c \in [-0.1, 0.2]$ and $z \in [0.4, 0.6]$, the optimization algorithm is less likely to be trapped in the local minima with both UOT distance and mixed $L^1$/Wasserstein distance. As the position of the perturbation is controlled by $z$, so too is the travel time of the reflection seismic events. The results in Figure 4.8 are consistent with the shift Ricker wavelet examples in the previous section.

### 4.4.2 Example 2: Cross-well model

In this subsection, we perform the full waveform inversion in a two-dimensional cross-well model to investigate the behavior of the update step in the optimization algorithm with direct wave. When the initial model is close to the true model, the difference between the simulated data and the received data is small. In this case, the Born approximation is relatively accurate and the inverse result is less likely to be trapped into a local minimum which is far away from the global minimum. However, the inverse result may be very different from the global minimum when the initial model is inaccurate. This phenomenon is demonstrated by the Camembert model [60]. The previous research shows that the 2-Wasserstein distance provides more accurate update steps compared to the $L^2$ distance [139]. We repeat the Camembert model experiment here to show the optimal transport based distances have the same advantage.

The model size is 2 km by 2 km, discretized into $101 \times 101$ grids with spatial grid size 0.02 km. The true velocity model is given by Figure 4.9 (a). In the true model, the background velocity is 3 km/s, and a single circle velocity anomaly is located at the center of the model with radius 0.5 km and velocity 3.6 km/s. There are 11 are equally spaced sources located on the left boundary of the domain, and 101 equally spaced receivers located on the right boundary of the domain. The synthetic data is generated with 10 Hz Ricker wavelets and a homogeneous initial velocity model is used with velocity 3 km/s.

The inverse results with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance is compared, the exponential normalization method is used for the optimal transport based distances. The L-BFGS method with a memory parameter of 5 is used as the optimization algorithm, and we perform 5 iterations to show the directions of the velocity model updates. Figure 4.10 shows the 6-th adjoint source at the first iteration with different distances. The adjoint sources generated by UOT distance and mixed $L^1$/Wasserstein distance provide slow transitions on the positions of the seismic wavefront. The frequency component of the seismic data is lower compared to the $L^2$ case. This leads to gradients with fewer large-scale components due to the

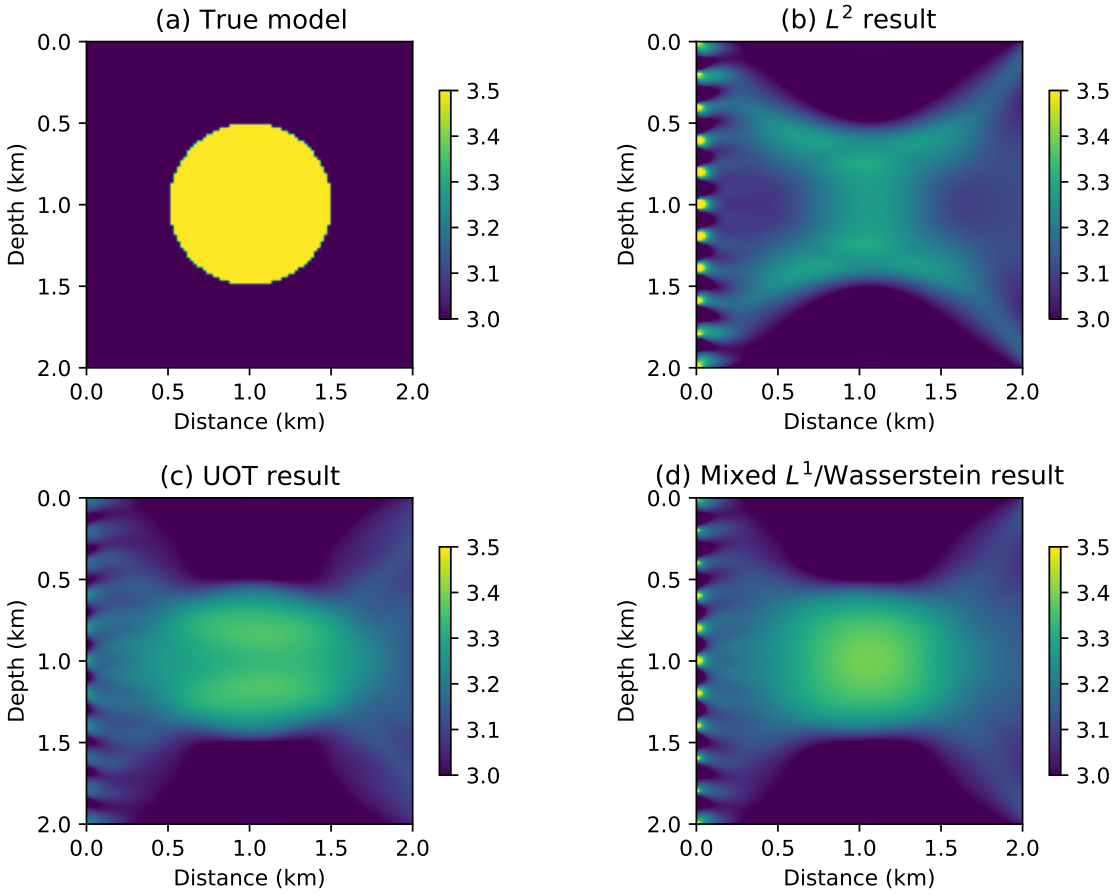


Figure 4.9: (a): True velocity model. (b): Inverse result with $L^2$ distance. (c): Inverse result with UOT distance and exponential normalization. (d): Inverse result with mixed $L^1$/Wasserstein distance and exponential normalization.

adjoint state method (4.43). Also, compared to the trace-by-trace strategy used in [139], the adjoint sources in subfigures (b) and (c) are more regular by the proposed method.

Figure 4.9 (b), (c), (d) display the inverse results with $L^2$ distance, UOT distance, mixed $L^1$/Wasserstein distance respectively. All three results describe the presence of the velocity anomaly. However, the $L^2$ result contains abnormal disturbances at the left and right parts of the center, which will provide a wrong velocity update in future iterations. Compared to the $L^2$ result, both UOT distance and mixed $L^1$/Wasserstein distance provide more regular updates with the shape similar to the velocity anomaly. This experiment shows that both UOT distance and mixed $L^1$/Wasserstein distance with exponential normalization can reduce the risk of wrong velocity updates, which may cause the optimization algorithm to be trapped in the local minima.

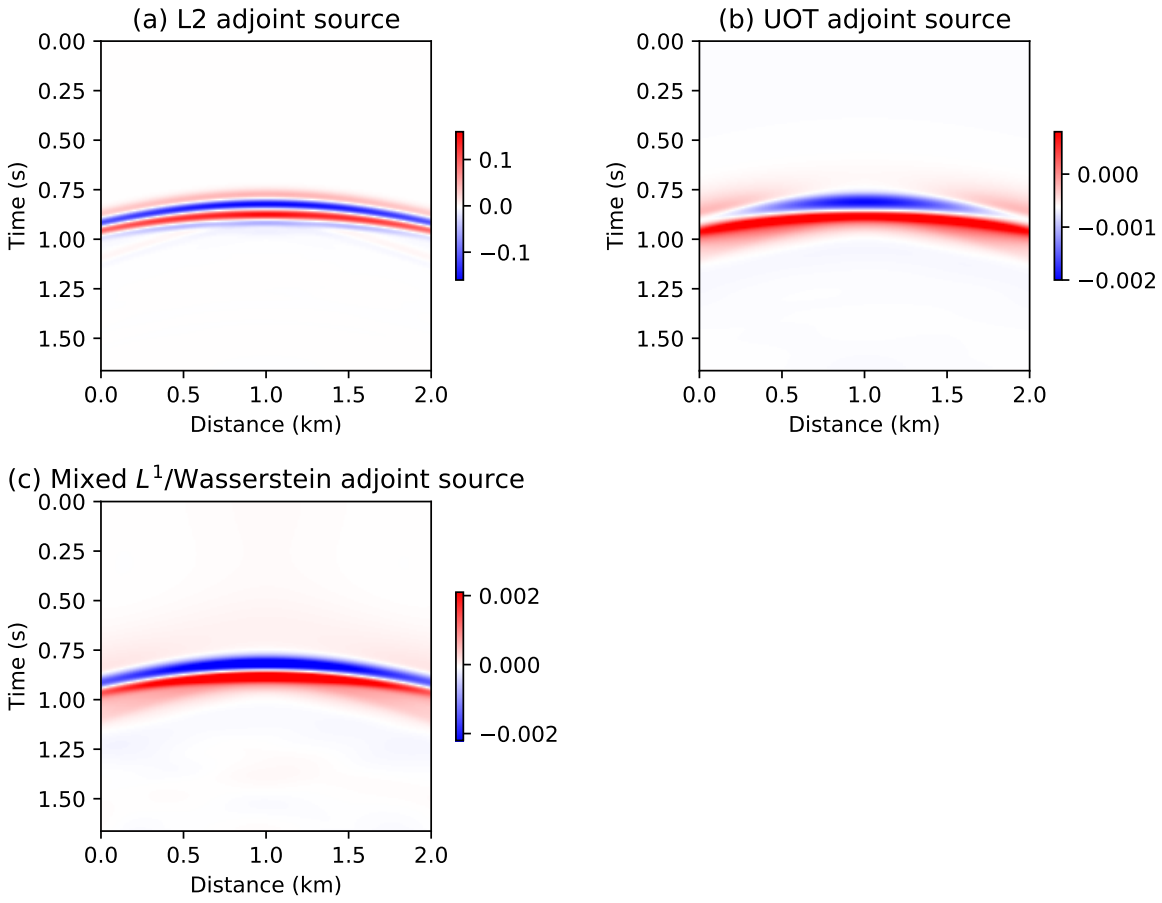


Figure 4.10: (a), (b), (c): The 6-th adjoint sources at the first iteration with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance, the exponential normalization method is used.

### 4.4.3 Example 3: Marmousi model

In this subsection, we compare the inverse results with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance through a two-dimensional reflection model. The exponential normalization method is used for the optimal transport based distances.

As shown in Figure 4.11 (a), the true velocity model is a part of the Marmousi 2 model [91] that provides strong velocity differences in both vertical and horizontal directions. The velocity model is discretized into $84 \times 301$ grids with the spatial size 0.03 km. There are 11 equally spaced sources and 101 equally spaced receivers located on the surface of the model. The initial model is achieved through a two-dimensional Gaussian filter applied to the true model which is strongly smoothened, as shown in Figure 4.11 (b). The sampling frequency is 400 Hz, and the recording time is 3 s. The synthetic data is generated by the Ricker wavelet with central frequency 5 Hz as the source function. The perfectly matched layer technique is performed to simulate the seismic wave propagating in an unbounded domain.

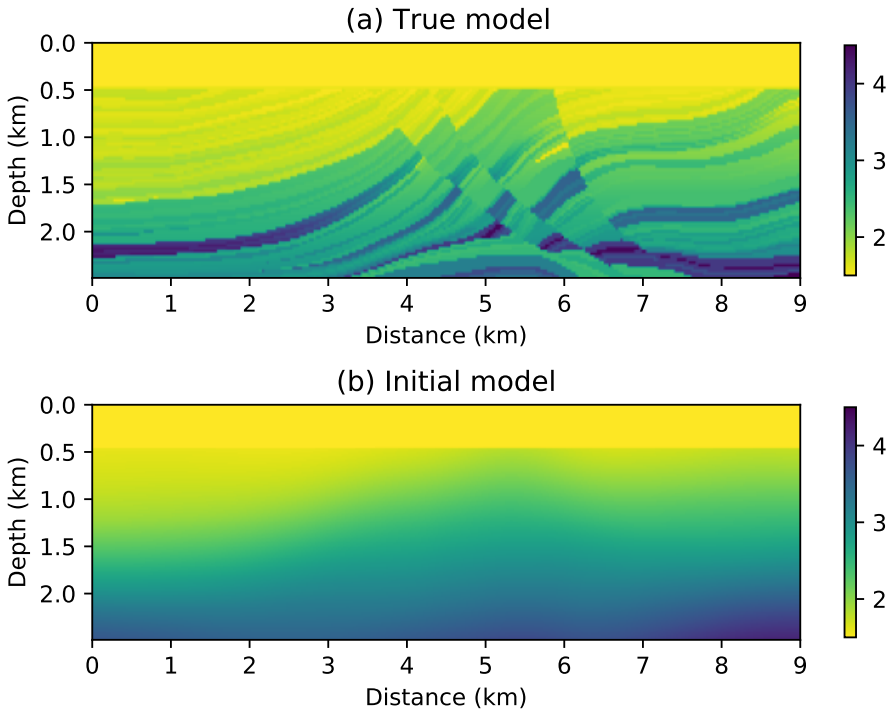


Figure 4.11: (a): True velocity model. (b): Initial velocity model.

Figure 4.12 shows snapshots of the synthetic wavefield, demonstrating the seismic wave propagating in the domain. Figure 4.13 shows the adjoint sources of $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance at the first iteration. Similar to the previous example, the energy of the adjoint sources generated by the optimal transport based distance concentrates on the location of the seismic wavelet, and provides a smoothed waveform of the seismic events. The first iteration gradients of $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance are shown in Figure 4.14. Compared to the $L^2$ gradient, the UOT gradient and mixed Wasserstein gradient provide more large-scale structures, which is more sensible on the bottom of the domain. These large-scale structures will increase the stability of the optimization algorithm.

The nonlinear conjugate gradient (CG) method is performed to minimize the objective function with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance. The inverse results with $L^2$ distance are shown in Figure 4.15. The inverse result after 20 iterations and 40 iterations are shown in subfigures (a) and (b) separately. Compared to the true velocity model, there is a velocity anomaly that exists at near depth 0.75 km, distance 6.25 km. This can be explained as the cycle-skipping issue since the velocity distribution in the initial model at this area is inaccurate compared to the true velocity model. In this case, the $L^2$ distance inversion failed to recover the velocity structure of the domain.

The inverse results of UOT distance and mixed $L^1$/Wasserstein distance are provided in Figure 4.16

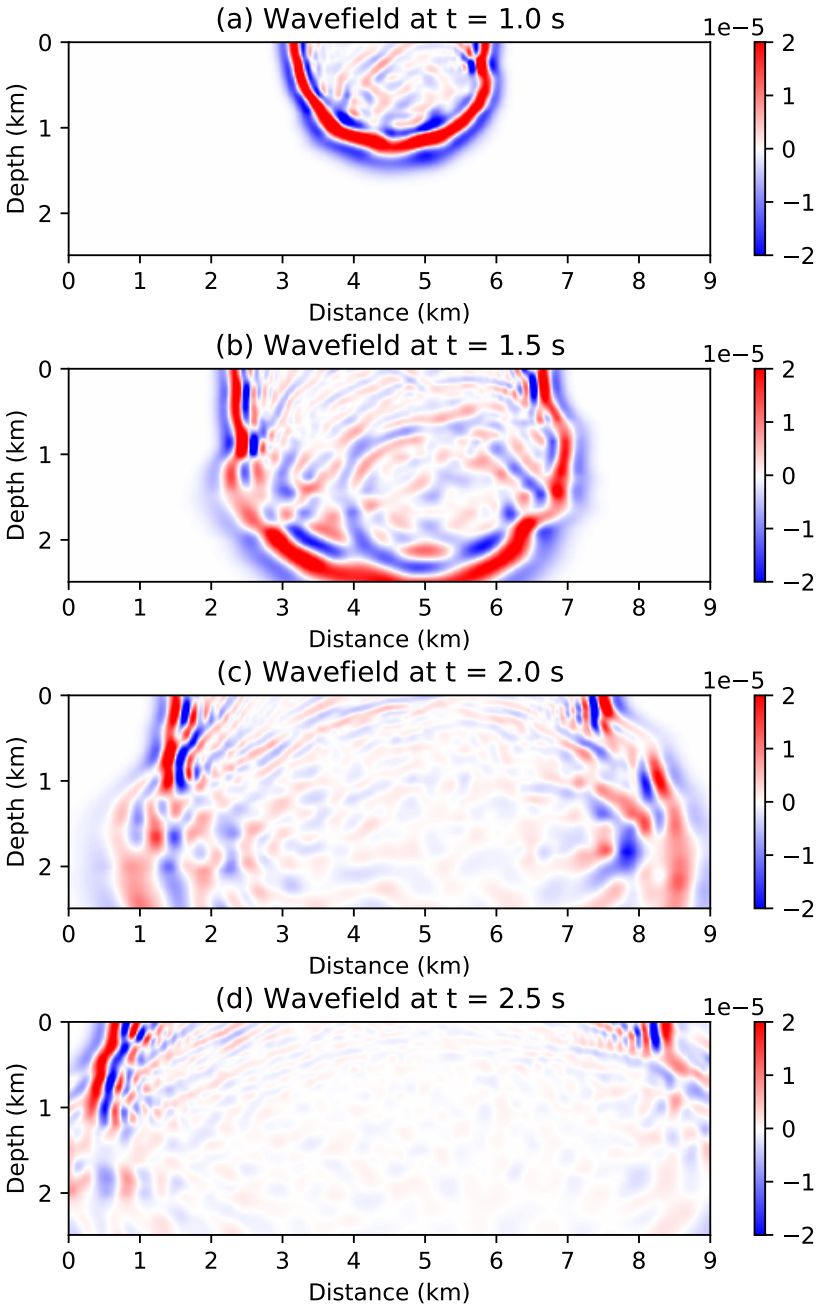


Figure 4.12: Snapshots of seismic wave generated by the 6-th source propagating in the domain.

and 4.17 (a) and (b). Both optimal transport based distances recovered the structure of the true velocity model after 40 iterations. Since the evaluation of UOT distance and mixed $L^1$/Wasserstein distance is much more expensive than the conventional $L^2$ distance. After the large-scale structure is accurately revealed, the optimal transport based distances can be replaced by the $L^2$ distance to achieve the inverse result more efficiently. With the inverse results Figure 4.16 (b) and Figure 4.17 (b) as the initial model, 80 nonlinear

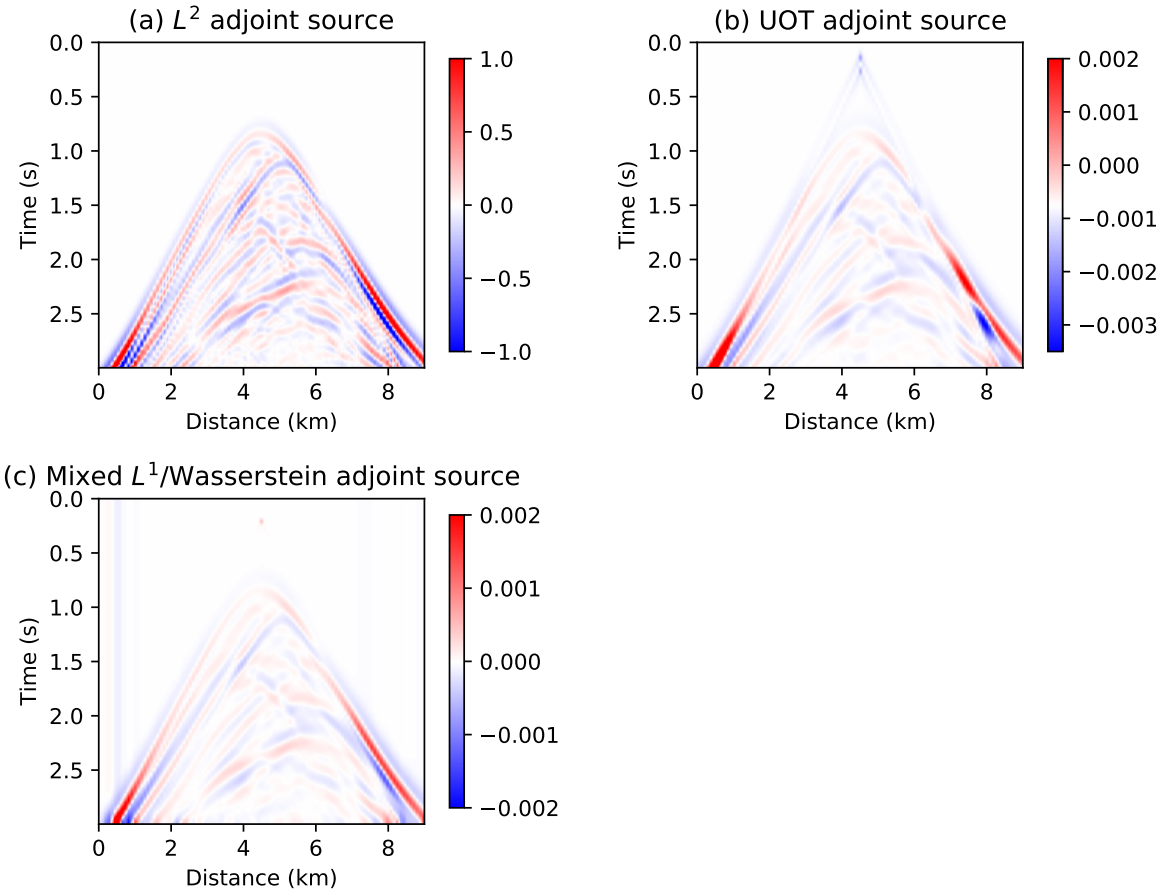


Figure 4.13: (a), (b), (c): The 6-th adjoint source at the first iteration with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance.

conjugate gradient iterations are performed with the $L^2$ distance. The final inverse results are shown in Figure 4.16 (c) and Figure 4.17 (c) with more detailed velocity structures be revealed. The difference before and after the additional $L^2$ nonlinear CG iterations are shown in Figure 4.16 (d) and Figure 4.17 (d).

### 4.4.4 Discussion

In this work, the methodology of integrating UOT distance and mixed $L^1$/Wasserstein distance to the full waveform inversion problem is provided. The normalization methods for transforming the signals into positive functions are needed, and several normalization methods are discussed with numerical examples. We formulate the full waveform inversion problem with UOT distance and mixed $L^1$/Wasserstein distance, and the computation methods of adjoint sources are provided. The numerical examples show that UOT distance and mixed $L^1$/Wasserstein distance with exponential normalization can mitigate the cycle-skipping issue efficiently compared to the conventional $L^2$ distance. With a poor initial model, the optimal transport

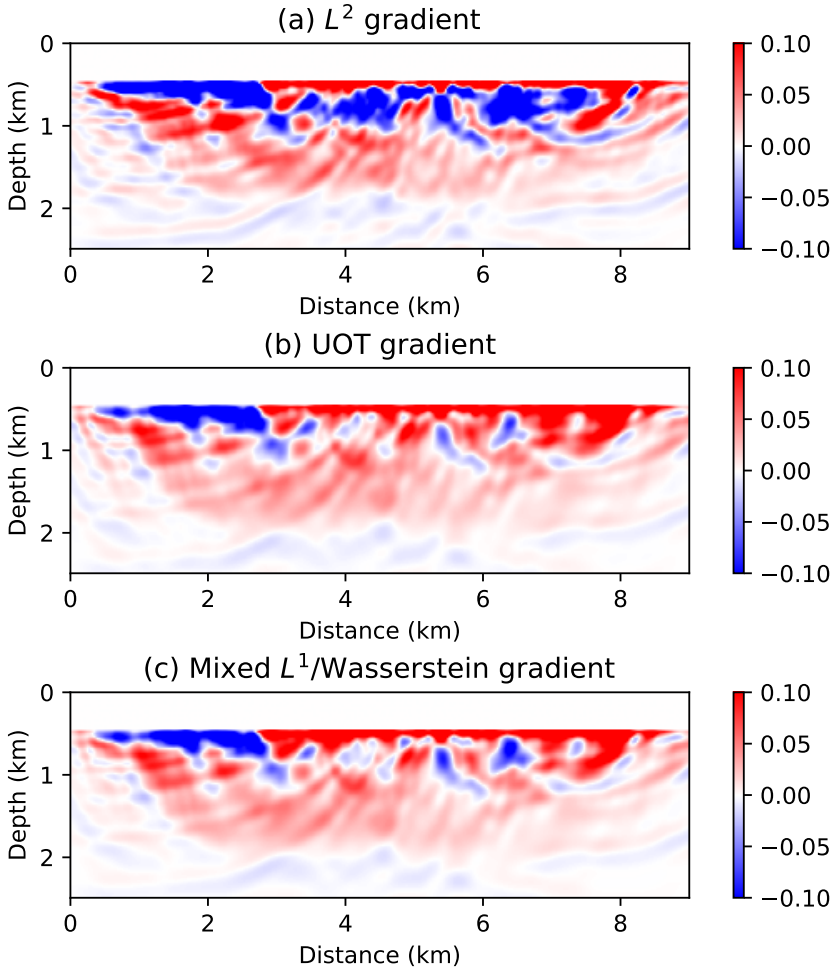


Figure 4.14: (a), (b), (c): The gradient at the first iteration with $L^2$ distance, UOT distance, and mixed $L^1$/Wasserstein distance.

based distances can provide a more accurate update step of the objective function and increase the stability of the optimization algorithm. Compared to the optimal transport based distances, $L^2$ distance objective function is sensitive to the initial model but can be evaluated efficiently. In practice, the inverse problem can be solved in two parts. First, use UOT distance or mixed $L^1$/Wasserstein distance objective function in the first few iterations to improve the initial model. Second, use $L^2$ distance objective function in the following iterations to increase the resolution of the inverse result.

There are two reasons that might explain the better performance of the optimal transport based distance. First, the velocity anomaly between the true velocity model and the initial velocity model dilates the shape of the seismic wavelet and changes the arriving time of the seismic event. Based on the previous discussion, the optimal transport based distances have more convex behavior compared to the $L^2$ distance. Another reason is the adjoint sources generated by the optimal transport based distances have more low-frequency

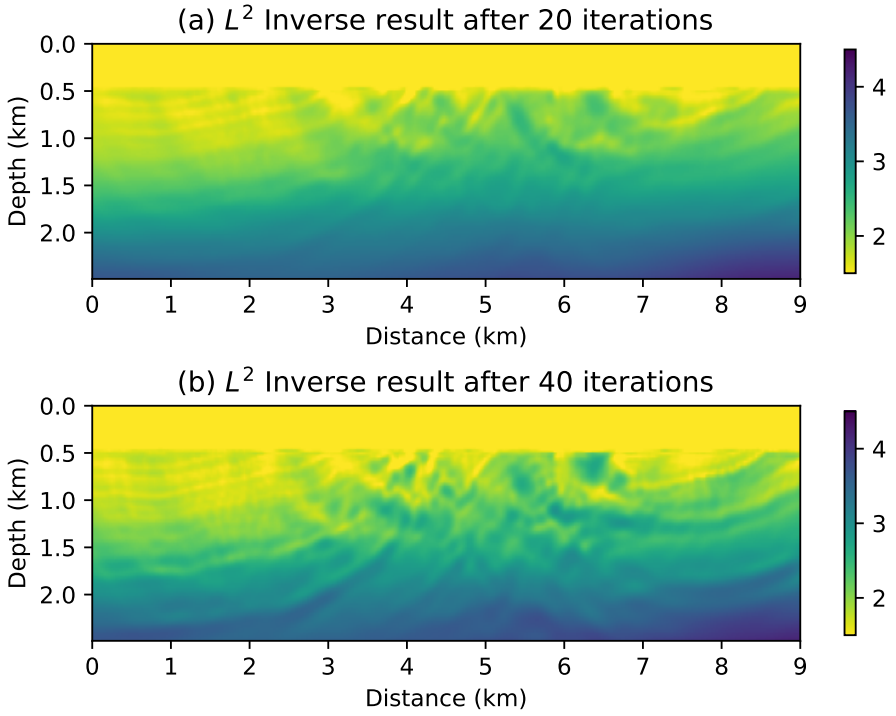


Figure 4.15: Nonlinear conjugate gradient inverse results with $L^2$ distance after 20 and 40 iterations.

components, so the update steps have more large-scale components based on the adjoint state method. This will decrease the nonlinearity of the optimization problem.

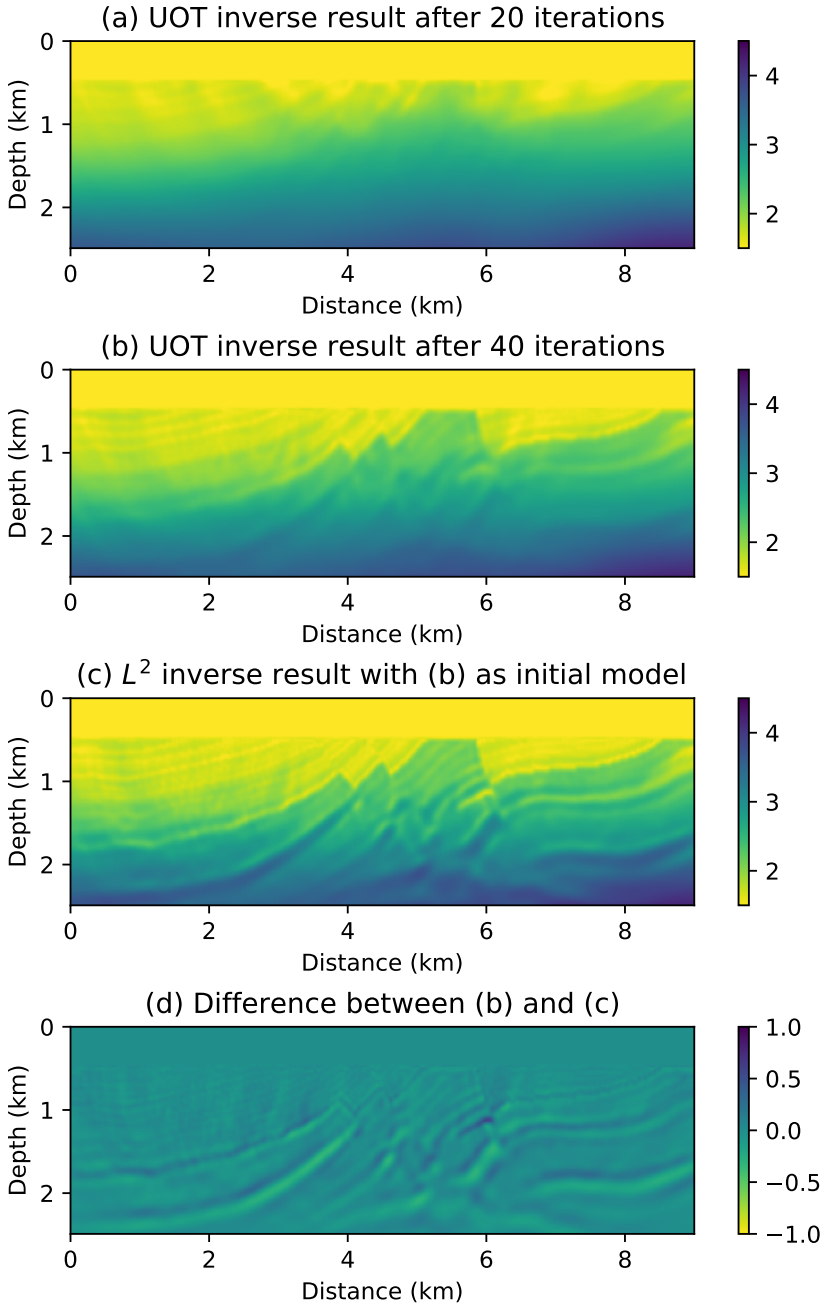


Figure 4.16: (a), (b): Nonlinear conjugate gradient inverse results with UOT distance after 20 and 40 iterations. (c): Nonlinear conjugate gradient inverse result with $L^2$ distance and (b) as the initial model after 80 iterations. (d): The difference between (b) and (c).

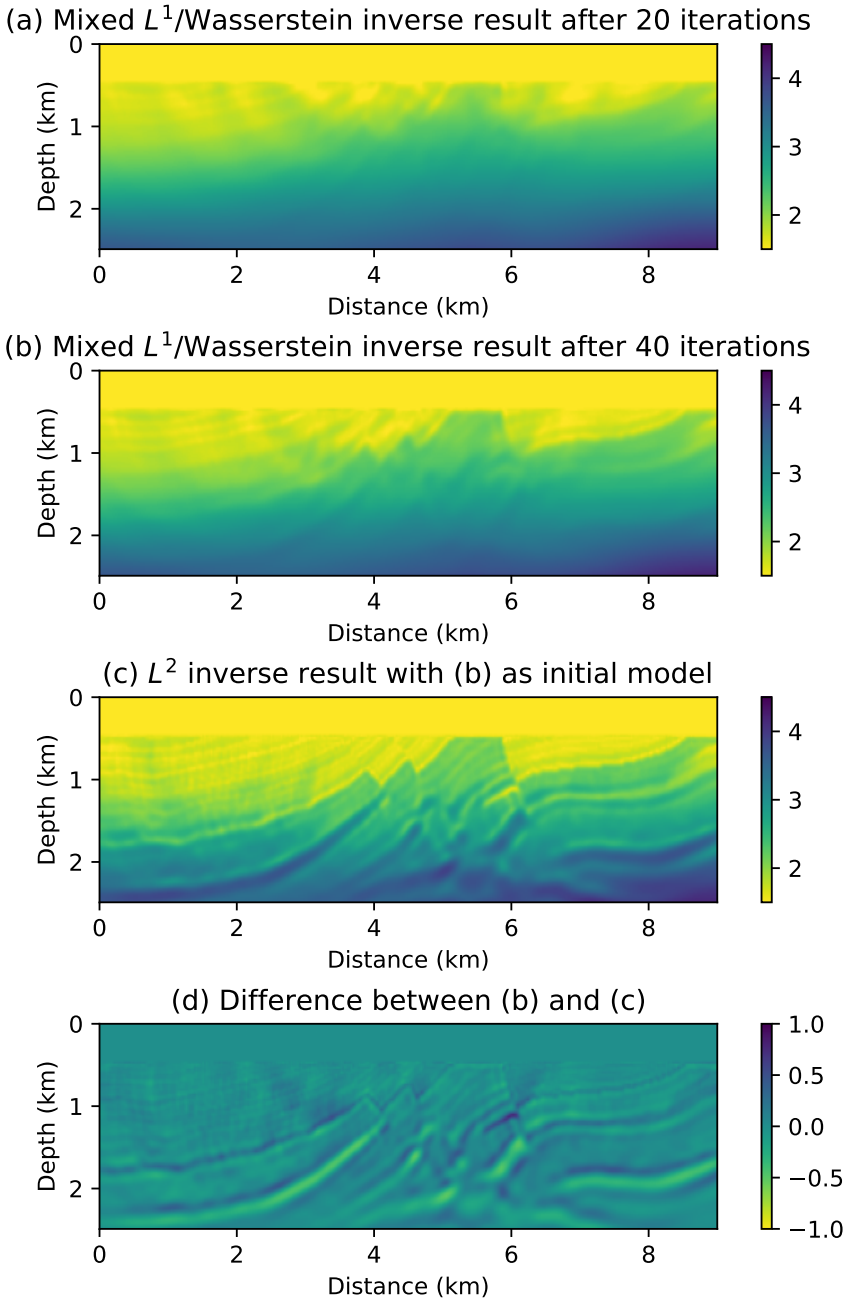


Figure 4.17: (a), (b): Nonlinear conjugate gradient inverse results with mixed $L^1$/Wasserstein distance after 20 and 40 iterations. (c): Nonlinear conjugate gradient inverse result with $L^2$ distance and (b) as the initial model after 80 iterations. (d): The difference between (b) and (c).

# Chapter 5

# Gradient projection methods with inexact projection

## 5.1 Introduction

In this chapter, we focus on solving the constrained optimization problem:

$$\min_x f(x), \quad \text{such that } x \in X, \tag{5.1}$$

where $f$ is a smooth nonlinear objective function which might be nonconvex. The feasible set $X$ is a nonempty, convex, and closed subset in $\mathbb{R}^n$.

To solve the above constrained optimization problem, one of the most straightforward methods is the gradient projection method. At the $k$-th iteration, compute

$$\bar{x}^k = P_X(x^k - \beta^k \nabla f(x^k)), \tag{5.2}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k), \tag{5.3}$$

here $\beta^k$ is a scalar parameter and $\alpha^k \in [0, 1]$ is a stepsize achieved through a line search method. This algorithm is demonstrated in Figure 5.1.

One of the most important parts of the gradient projection method is the evaluation of the projection function $P_X(\cdot)$. When the constraint is simple, like the box constraint, the closed-form projection function is available and the evaluation of the projection function is efficient. However, only iterative projection

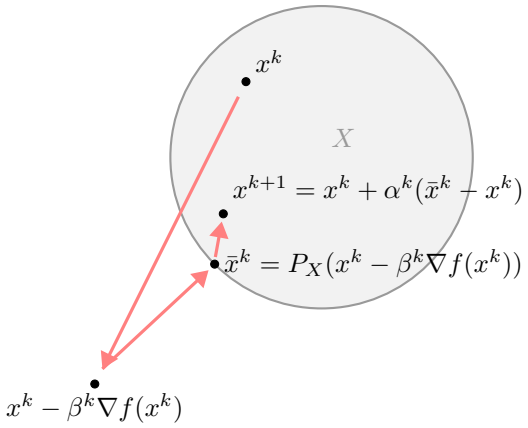


Figure 5.1: Gradient projection method at the $k$-th iteration.

algorithms like Dykstra's algorithm are available when the feasible set $X$ is an intersection of several convex constraint sets, and the iteration process has to be ended after a stopping criterion is met. In this case, the projection function $P_X(\cdot)$ is an inexact projection. For example in the gradient projection method defined in equation (5.2), the projection $\bar{x}^k$ may only be close to the accurate projection $P_X(x^k - \beta^k \nabla f(x^k))$ and may not be in $X$. We use the notation $\bar{P}_X(\cdot)$ to represent the inexact projection function onto set $X$. Figure 5.2 is an illustration of the inexact projection.

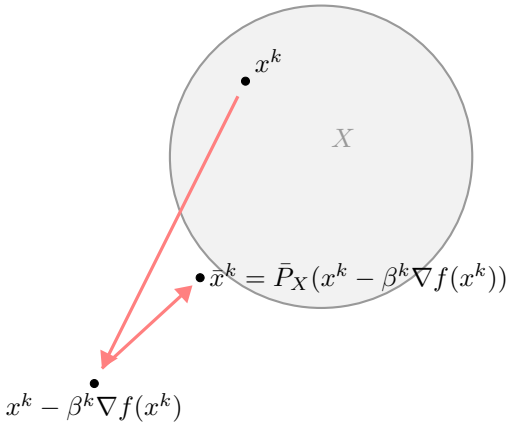


Figure 5.2: Gradient projection method at the $k$-th iteration with inexact projection.

In this chapter, we develop a set expanding strategy for the gradient projection methods with inexact projection. This set expanding strategy determines the stopping criterion of the iterative projection process when the constraint set $X$ is an intersection of several constraint sets and the iterative projection process is used. First, we construct an increasing constraint set sequence $\{X^k\}$:

$$X = \lim_{k \to \infty} X^k, \quad X^k \subset X^{k+1}, \quad X^{k+1} \neq X^k, \tag{5.4}$$

where $X^k$ is nonempty, closed, and convex for each $k \in \mathbb{N}$. At the $k$-th iteration, we evaluate the projection process towards the set $X^k$ such that the inexact projection $\bar{P}_X(\cdot) \in X^{k+1}$. With this construction, each set $X^k$ acts as a constraint set, and $X^k$ is expanding along with the iterations. In this case, the expanding set sequence $\{X^k\}$ adds a "soft" constraint information to the overall algorithm and might increase the result of the optimization problem when the objective function is nonconvex.

In Section 5.2, the basic set convergence and set-valued mapping results are reviewed, and preliminary results for the projection function are provided. We discuss the gradient projection method with inexact projection in Section 5.3. To increase the convergence speed, the Hessian matrix information can be introduced to the projection gradient method, and this leads to the scaled gradient projection method. We introduce the scaled gradient projection method with inexact projection in Section 5.4.

## 5.2 Preliminary results

In this section, we review some of the set convergence and set-valued mapping results based on the monograph [116], Chapter 3 to 5. We consider a sequence of sets $C^\nu \subset \mathbb{R}^n$ and focus on the convergence behavior to $C$. Later, the projection operator on sets can be considered as set-valued mappings. We review some results of the projection function and then prove the basic results which are used to show the convergence of the proposed algorithm.

Since we are working with both $n$-dimensional Euclidean space and the scaled Euclidean space, denote $\mathcal{H} = \mathbb{R}^n$ be the $n$-dimensional Euclidean space with the inner product $\langle x, y\rangle = x'y$ and the norm $\|x\| = \sqrt{x'x}$. Given a symmetric positive definite matrix $B$, let the space $\mathcal{H}_B$ be the scaled Euclidean space with the elements in $\mathbb{R}^n$, the inner product $\langle x, y\rangle_B = \langle Bx, y\rangle = x'By$ and the norm $\|x\|_B = \sqrt{x'Bx}$.

### 5.2.1 Set convergence and set-valued mapping convergence

The following subsets of $\mathbb{N}$ are useful in this section:

$$\mathcal{N}_\infty = \{N \subset \mathbb{N} \mid \mathbb{N}\backslash N \text{ finite}\}, \tag{5.5}$$

$$\mathcal{N}_\infty^\sharp = \{N \subset \mathbb{N} \mid N \text{ infinite}\}, \tag{5.6}$$

where $\mathcal{N}_\infty$ represents the set of subsequences of $\mathbb{N}$ containing all $\nu$ beyond some $\bar{\nu}$, and $\mathcal{N}_\infty^\sharp$ represents all subsequences of $\mathbb{N}$. Then we can have the definition of the limit of set sequence:

**Definition 5.1** ([116] 4.1, inner and outer limits)**.** *For a sequence $\{C^\nu\}_{\nu\in\mathbb{N}}$ of subsets of $\mathbb{R}^n$, the outer limit*

*is the set*

$$\limsup_{\nu\to\infty} C^\nu = \left\{ x \;\middle|\; \exists N \in \mathcal{N}_\infty^\sharp,\ \exists x^\nu \in C^\nu (\nu \in N) \text{ with } x^\nu \underset{N}{\to} x \right\}, \tag{5.7}$$

*while the inner limit is the set*

$$\liminf_{\nu\to\infty} C^\nu = \left\{ x \;\middle|\; \exists N \in \mathcal{N}_\infty,\ \exists x^\nu \in C^\nu (\nu \in N) \text{ with } x^\nu \underset{N}{\to} x \right\}, \tag{5.8}$$

*The limit of sequence exists if the outer and inner limit sets are equal:*

$$\lim_{\nu\to\infty} C^\nu := \limsup_{\nu\to\infty} C^\nu = \liminf_{\nu\to\infty} C^\nu \tag{5.9}$$

The set convergence defined by the above definition is denoted as Painlevé-Kuratowski convergence [116].

For any nonempty, closed set $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$, the distance between a point $x$ to $C$ is denoted as a distance function:

$$d_C(x) = \inf_{z\in C} \|x - z\|. \tag{5.10}$$

We denote the distance function $d_C(x) = d(x, C)$ sometimes.

Next, we discuss the distance between sets and the set convergence in a metric space. For more detailed discussion, we refer to [116]. Given a parameter $\rho \in \mathbb{R}_+ = [0, \infty)$ and a pair of nonempty sets $C$ and $D$, define

$$d_\rho(C, D) = \max_{|x|\le\rho} |d_C(x) - d_D(x)|. \tag{5.11}$$

Then, the (integrated) set distance between $C$ and $D$ is defined as

$$d(C, D) = \int_0^\infty d_\rho(C, D) e^{-\rho}\, \mathrm{d}\rho. \tag{5.12}$$

Fix the notation of the sets of nonempty, closed subsets of $\mathbb{R}^n$ as

$$\text{cl-sets}_{\neq\emptyset}(\mathbb{R}^n) = \text{the space of all nonempty, closed subsets of } \mathbb{R}^n. \tag{5.13}$$

The following definition characterizes the set sequence eventually departs from any bounded region of $\mathbb{R}^n$.

**Definition 5.2** ([116] 4.11, escape to the horizon)**.** *The condition $C^\nu \to \emptyset$ (or equivalently, $\limsup_\nu C^\nu = \emptyset$)*

*holds for a sequence* $\{C^\nu\}_{\nu\in\mathbb{N}}$ *in* $\mathbb{R}^n$ *if and only if for every* $\rho > 0$ *there is an index set* $N \in \mathcal{N}_\infty$ *such that* $C^\nu \cap \rho\mathbb{B} = \emptyset$ *for all* $\nu \in N$.

The following theorem gives the metric description of set convergence.

**Theorem 5.3** ([116] 4.42, metric description of set convergence)**.** *The expression* $d$ *gives a metric on* cl-sets$_{\neq\emptyset}(\mathbb{R}^n)$ *which characterizes ordinary set convergence:*

$$C^\nu \to C \Longleftrightarrow d(C^\nu, C) \to 0. \tag{5.14}$$

*Furthermore,* cl-sets$_{\neq\emptyset}(\mathbb{R}^n)$ *is a complete metric space in which a sequence* $\{C^\nu\}_{\nu\in\mathbb{N}}$ *escapes to the horizon if and only if for some set* $C$ *in this space one has* $d(C^\nu, C) \to \infty$.

The following corollary provides the boundedness of the set limit.

**Corollary 5.4** ([116] 4.12, limits of connected sets)**.** *Let* $C^\nu \subset \mathbb{R}^n$ *be connected with* $\limsup_\nu C^\nu$ *bounded and no subsequence escaping to the horizon. Then there is a bounded set* $B \subset \mathbb{R}^n$ *such that* $C^\nu \subset B$ *for all* $\nu$ *in some* $N \in \mathcal{N}_\infty$.

The above corollary is useful because convex sets in $\mathbb{R}^n$ are connected.

The projection function is a set-valued mapping, denoted as $P_C(\cdot)$ for an nonempty, closed set $C$. Next, we review the results of the set-valued mapping to analyse the convergence behavior of $P_{C^\nu}(x)$ to $P_C(x)$ as $C^\nu \to C$. Consider a set-valued mapping $S$ which maps the element in space $X$ to the elements in space $U$, $S(x)$ is a subset in $U$ and point $x \in X$. The graph of $S$ is a subset of space $X \times U$ as

$$\text{gph } S := \{(x, u) \mid u \in S(x)\}. \tag{5.15}$$

Denote the set-valued mapping $S : X \rightrightarrows U$:

$$S(x) = \{u \mid (x, u) \in \text{gph } S\}. \tag{5.16}$$

The double arrow notation is used in the textbook [116] in order to distinguish the set-valued mapping from regular function.

**Definition 5.5** ([116] 5.31, pointwise limits of mappings)**.** *For a sequence of mappings* $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$*, the pointwise outer limit and the pointwise inner limit are the mappings* $p\text{-}\limsup_\nu S^\nu$ *and* $p\text{-}\liminf_\nu S^\nu$ *defined*

*at each point $x$ by*

$$\left(p\text{-}\limsup_{\nu} S^{\nu}\right)(x) := \limsup_{\nu} S^{\nu}(x), \tag{5.17}$$

$$\left(p\text{-}\liminf_{\nu} S^{\nu}\right)(x) := \liminf_{\nu} S^{\nu}(x). \tag{5.18}$$

*When the pointwise outer and inner limits agree, the pointwise limit $p\text{-}\lim_{\nu} S^{\nu}$ is said to exist; thus, $S = p\text{-}\lim_{\nu} S^{\nu}$ if and only if $S \supset p\text{-}\limsup_{\nu} S^{\nu}$ and $S \subset p\text{-}\liminf_{\nu} S^{\nu}$. We summarize this definition with the notation*

$$S^{\nu} \overset{\mathrm{p}}{\to} S \Longleftrightarrow S^{\nu}(x) \to S(x) \text{ for all } x. \tag{5.19}$$

The graphical convergence is obtained by applying the theory of set convergence to the graph of sets.

**Definition 5.6** ([116] 5.32, graphical limits of mappings)**.** *For a sequence of mappings $S^{\nu} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, the graphical outer limit denoted by $\mathrm{g}\text{-}\limsup_{\nu} S^{\nu}$ is the mapping*

$$\operatorname{gph}\left(\mathrm{g}\text{-}\limsup_{\nu} S^{\nu}\right) = \limsup_{\nu} \left(\operatorname{gph} S^{\nu}\right), \tag{5.20}$$

$$\left(\mathrm{g}\text{-}\limsup_{\nu} S^{\nu}\right)(x) = \left\{u \;\middle|\; \exists N \in \mathcal{N}_{\infty}^{\sharp}, x^{\nu} \underset{N}{\to} x, u^{\nu} \underset{N}{\to} u, u^{\nu} \in S^{\nu}(x^{\nu})\right\}. \tag{5.21}$$

*The graphical inner limit, denoted by $\mathrm{g}\text{-}\liminf_{\nu} S^{\nu}$ is the mapping*

$$\operatorname{gph}\left(\mathrm{g}\text{-}\liminf_{\nu} S^{\nu}\right) = \liminf_{\nu} \left(\operatorname{gph} S^{\nu}\right), \tag{5.22}$$

$$\left(\mathrm{g}\text{-}\liminf_{\nu} S^{\nu}\right)(x) = \left\{u \;\middle|\; \exists N \in \mathcal{N}_{\infty}, x^{\nu} \underset{N}{\to} x, u^{\nu} \underset{N}{\to} u, u^{\nu} \in S^{\nu}(x^{\nu})\right\}. \tag{5.23}$$

*If these outer and inner limits agree, the graphical limit $\mathrm{g}\text{-}\lim_{\nu} S^{\nu}$ exists. Thus $S = \mathrm{g}\text{-}\lim_{\nu} S^{\nu}$ if and only if $S \supset \mathrm{g}\text{-}\limsup_{\nu} S^{\nu}$ and $S \subset \mathrm{g}\text{-}\liminf_{\nu} S^{\nu}$. We summarize this definition with the notation*

$$S^{\nu} \overset{\mathrm{g}}{\to} S \Longleftrightarrow \operatorname{gph} S^{\nu} \to \operatorname{gph} S. \tag{5.24}$$

**Proposition 5.7** ([116] 5.33, graphical limit formulas at a point)**.** *For any sequence of mappings $S^{\nu} : \mathbb{R}^n \rightrightarrows$*

$\mathbb{R}^m$, *one has*

$$\left(\text{g-}\liminf_{\nu} S^\nu\right)(x) = \bigcup_{\{x^\nu \to x\}} \liminf_{\nu\to\infty} S^\nu(x^\nu) = \lim_{\delta\downarrow 0}\left[\liminf_{\nu\to\infty} S^\nu\left(x+\delta\mathbb{B}\right)\right], \tag{5.25}$$

$$\left(\text{g-}\limsup_{\nu} S^\nu\right)(x) = \bigcup_{\{x^\nu \to x\}} \limsup_{\nu\to\infty} S^\nu(x^\nu) = \lim_{\delta\downarrow 0}\left[\limsup_{\nu\to\infty} S^\nu\left(x+\delta\mathbb{B}\right)\right], \tag{5.26}$$

*where the unions are taken over all sequences* $x^\nu \to x$. *Thus,* $S^\nu$ *converges graphically to* $S$ *if and only if, at each point* $\bar{x} \in \mathbb{R}^n$, *one has*

$$\bigcup_{\{x^\nu \to x\}} \limsup_{\nu\to\infty} S^\nu(x^\nu) \subset S(\bar{x}) \subset \bigcup_{\{x^\nu \to x\}} \liminf_{\nu\to\infty} S^\nu(x^\nu). \tag{5.27}$$

The above proposition is characterized by the graphical convergence of $S^\nu$ to $S$ at a single point $\bar{x}$ by equation (5.27). We say that $S^\nu$ converges graphically to $S$ relative to a set $X$ if equation (5.27) holds for every $\bar{x} \in X$ with corresponding sequence $x^\nu$ in $X$.

The following definitions describe the convergence behavior of $S^\nu$ converges to $S$ as $x^\nu$ converges to $\bar{x}$. Let $\mathcal{N}(x)$ be the set of neighborhood of point $x$.

**Definition 5.8** ([116] 5.41, continuous limits of mappings)**.** *A sequence of mappings* $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *is said to converge continuously to a mapping* $S$ *at* $\bar{x}$ *if* $S^\nu(x^\nu) \to S(\bar{x})$ *for all sequences* $x^\nu \to \bar{x}$. *It can be identified with the condition that for every* $\varepsilon > 0$ *and* $\rho > 0$ *there exists* $N \in \mathcal{N}_\infty$ *along with a neighborhood* $V \in \mathcal{N}(\bar{x})$ *such that*

$$S^\nu(x) \cap \rho\mathbb{B} \subset S(\bar{x}) + \varepsilon\mathbb{B}, \tag{5.28}$$

$$S(\bar{x}) \cap \rho\mathbb{B} \subset S^\nu(x) + \varepsilon\mathbb{B}, \tag{5.29}$$

*for all* $x \in V$ *when* $\nu \in N$. *If this holds at all* $\bar{x} \in \mathbb{R}^n$, *the sequence* $S^\nu$ *converges continuously to* $S$. *It does so relative to a set* $X \subset \mathbb{R}^n$ *if this holds at all* $\bar{x} \in X$ *when* $x^\nu \in X$.

**Definition 5.9** ([116] 5.41, uniform limits of mappings)**.** *The mappings* $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *converge uniformly to* $S$ *on a subset* $X$ *if for every* $\varepsilon > 0$ *and* $\rho > 0$ *there exists* $N \in \mathcal{N}_\infty$ *such that*

$$S^\nu(x) \cap \rho\mathbb{B} \subset S(x) + \varepsilon\mathbb{B}, \tag{5.30}$$

$$S(x) \cap \rho\mathbb{B} \subset S^\nu(x) + \varepsilon\mathbb{B}, \tag{5.31}$$

*for all* $x \in X$ *when* $\nu \in N$.

The following theorem provides a connection between continuous convergence and uniform convergence of the set-valued mapping sequence.

**Theorem 5.10** ([116] 5.43, continuous versus uniform convergence)**.** *For mappings $S$, $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and a set $X \subset \mathbb{R}^n$, the following conditions are equivalent:*

1. *$S^\nu$ converges continuously to $S$ relative to $X$;*
2. *$S^\nu$ converges uniformly to $S$ on all compact subsets of $X$, and $S$ is continuous relative to $X$.*

The next theorem provides a connection between graphical convergence and continuous convergence at a point.

**Theorem 5.11** ([116] 5.44, graphical versus continuous convergence)**.** *For mappings $S$, $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and a set $X \subset \mathbb{R}^n$, the following properties at $\bar{x} \in X$ are equivalent:*

1. *$S^\nu$ converges continuously to $S$ at $\bar{x}$ relative to $X$;*
2. *$S^\nu$ converges graphically to $S$ at $\bar{x}$ relative to $X$, and the sequence is asymptotically equicontinuous at $\bar{x}$ relative to $X$.*

The gradient projection method is one of the most popular methods to solve the constrained optimization problem (5.1), and a projection function is evaluated at each of the iterations. Later we will show the projection function is a single-valued mapping when projecting onto the nonempty, closed, and convex sets. The following corollary plays an important role in our work.

**Corollary 5.12** ([116] 5.45, graphical convergence of single-valued mappings)**.** *For single-valued mappings $F, F^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, the following conditions are equivalent:*

1. *$F^\nu$ converges continuously to $F$ at $\bar{x}$;*
2. *$F^\nu$ converges graphically to $F$ at $\bar{x}$, and the sequence is eventually locally bounded at $\bar{x}$, i.e., there exist $V \in \mathcal{N}(\bar{x})$, $N \in \mathcal{N}_\infty$ and a bounded set $B$ such that $F^\nu(x) \in B$ for all $x \in V$ when $\nu \in N$.*

### 5.2.2 Convergence of projection mapping sequence

Given a nonempty, closed set $C \subset \mathbb{R}^n$, define the projection function as

$$P_C(x) = \arg\min_{z \in C} \|x - z\|. \tag{5.32}$$

It is a set-valued function consisting of the points in $C$ nearest to $x$.

When $C$ is convex, the projection function is a single-valued mapping, $P_C(x)$ is the closest point to point $x$ on $C$. The following proposition gives the basic properties of the projection functions onto convex sets.

**Proposition 5.13** ([17] Proposition 2.1.3, projection theorem)**.** *Let $C$ be a nonempty, closed, and convex subset of $\mathbb{R}^n$,*

1. *For every $x \in \mathbb{R}^n$, there exists a unique $z = P_C(x)$ that minimize $\|x - z\|$ over all $z \in C$.*
2. *Given some $x \in \mathbb{R}^n$, a point $\bar{x}$ is equal to the projection $P_C(x)$ if and only if*

$$\langle x - \bar{x}, z - \bar{x} \rangle \leq 0, \quad \forall z \in C. \tag{5.33}$$

3. *The mapping $P_C : \mathbb{R}^n \to C$ is continuous and nonexpansive, that is*

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \tag{5.34}$$

4. *In the case when $C$ is a subspace, a point $\bar{x}$ is equal to the projection $P_C(x)$ if and only if $x - \bar{x}$ is orthogonal to $C$, that is*

$$\langle x - \bar{x}, z \rangle = 0, \quad \forall z \in C. \tag{5.35}$$

**Proposition 5.14** ([116] 5.35, graphical convergence of projection mappings)**.** *For closed sets $C^\nu, C \subset \mathbb{R}^n$, one has $P_{C^\nu} \xrightarrow{\text{g}} P_C$ if and only if $C^\nu \to C$.*

**Proposition 5.15** ([116] 5.23, local boundedness of projection mappings)**.** *For any nonempty set $C \subset \mathbb{R}^n$, the projection mapping $P_C$ is locally bounded.*

**Lemma 5.16** (Eventually locally boundedness)**.** *Suppose $C^\nu, C \subset \mathbb{R}^n$ are nonempty, closed, and convex, $C^\nu \to C$. Then the sequence of projection function $\{P_{C^\nu}\}$ is eventually locally bounded.*

*Proof.* Fix arbitrary $\bar{x} \in \mathbb{R}^n$, by Proposition 5.15, since $C$ is nonempty, $P_C$ is locally bounded at $\bar{x}$, i.e., $\exists V \in \mathcal{N}(\bar{x})$ and a bounded set $B_1$, such that $P_C(x) \in B$ for all $x \in V$.

By Proposition 5.13, $P_{C^\nu}, P_C$ are single-valued continuous functions. The map $x \to (x, P_{C^\nu}(x))$ is continuous, and $x \in V$ is connected, then the graph gph $P_{C^\nu} \subset V \times \mathbb{R}^n$ is connected.

By Proposition 5.14, $P_{C^\nu}$ converge graphically to $P_C$, that is equivalent to gph $P_{C^\nu} \to$ gph $P_C$ by equation (5.24). Since gph $P_C$ is nonempty, there is no subsequence escaping to the horizon in $\{\text{gph } P_{C^\nu}\}$. Also, gph $P_C \subset V \times B$ is bounded.

Then, by Corollary 5.4, there exists a bounded set $\bar{B} \subset \mathbb{R}^n$ and $N \in \mathcal{N}_\infty$ such that gph $P_{C^\nu} \subset \bar{B}$ for all $\nu \in N$. That means, there exists a bounded set $B_2$ such that gph $P_{C^\nu} \subset V \times B_2 \subset \bar{B}$ for all $\nu \in N$.

In summary, for arbitrary $\bar{x}$, there exists the above $V \in \mathcal{N}(\bar{x})$, the above $N \in \mathcal{N}_\infty$, and the above bounded set $B_2$, such that $P_{C^\nu}(x) \in B_2$ for all $x \in V$ when $\nu \in N$. □

The following proposition is used for the convergence analysis.

**Proposition 5.17.** *Suppose $C^\nu, C \subset \mathbb{R}^n$ are nonempty, closed, and convex, $C^\nu \to C$. Then the projection mapping sequence $P_{C^\nu}$ converges continuously to $P_C$.*

*Furthermore, suppose $X$ is a subset in $\mathbb{R}^n$, the projection mapping sequence $P_{C^\nu}$ converges uniformly to $P_C$ on all compact subsets of $X$.*

*Proof.* Since $C^\nu, C$ are nonempty and convex, the projection functions $P_{C^\nu}, P_C$ are single-valued mapping by Proposition 5.13.

For any $\bar{x} \in \mathbb{R}^n$, since $C^\nu, C \subset \mathbb{R}^n$ are closed and $C^\nu \to C$, we have $P_{C^\nu}$ converges graphically to $P_C$ at point $\bar{x}$ by Proposition 5.14. By Lemma 5.16, the projection function sequence $\{P_{C^\nu}\}$ is eventually locally bounded at $\bar{x}$. By Corollary 5.12, $P_{C^\nu}$ converges continuously to $P_C$ at $\bar{x}$. The proof is finished since $\bar{x}$ is an arbitrary point in $\mathbb{R}^n$.

By Theorem 5.10, $P_{C^\nu}$ converges uniformly to $P_C$ on all compact subsets of $X$.

□

### 5.2.3 Convergence of scaled projection mapping sequence

Next, we discuss the projection mappings defined on the scaled Euclidean space $\mathcal{H}_B$. For any nonempty closed convex set $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$, the scaled distance function is denoted as:

$$d_{B,C}(x) = \inf_{z \in C} \|x - z\|_B. \tag{5.36}$$

The scaled projection mapping is denoted as

$$P_{B,C}(x) = \arg\min_{z \in C} \|x - z\|_B. \tag{5.37}$$

The following proposition gives the basic properties of scaled projection mapping with respect to nonempty closed convex sets.

**Proposition 5.18** (Scaled projection theorem)**.** *Suppose $C \subset \mathbb{R}^n$ is nonempty, closed, and convex. Given a symmetric positive definite matrix $B$, suppose for all $x \in \mathbb{R}^n$, there exists positive constants $\beta_1, \beta_2, \beta_3 > 0$,*

*such that*

$$\|Bx\| \leq \beta_1 \|x\|, \quad \textit{and} \quad \beta_2 \|x\|^2 \leq \langle Bx, x\rangle \leq \beta_3 \|x\|^2, \quad \forall x \in \mathbb{R}^n. \tag{5.38}$$

*Then following four statements hold:*

1. *Given $x \in \mathbb{R}^n$, there exists a unique vector $z \in C$ that minimize $\|z - x\|_B$ over all $z \in C$. Denote $z = P_{B,C}(x)$ as the projection of $x$ on $C$ in $\mathcal{H}_B$, i.e.*

$$P_{B,C}(x) = \arg\min_{z \in C} \|z - x\|_B. \tag{5.39}$$

2. *Given $x \in \mathbb{R}^n$, a vector $\bar{x} \in C$ is equal to $P_{B,C}(x)$ if and only if,*

$$\langle x - \bar{x}, z - \bar{x}\rangle_B \leq 0, \quad \forall z \in C. \tag{5.40}$$

3. *The function $P_{B,C}(x) : \mathbb{R}^n \to C$ is continuous and,*

$$\|P_{B,C}(x) - P_{B,C}(y)\| \leq \frac{\beta_1}{\beta_2} \|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \tag{5.41}$$

4. *When $C$ is a subspace in $\mathbb{R}^n$, for all $x \in \mathbb{R}^n$, there exists $\bar{x} \in C$ equal to $P_{B,C}(x)$ if and only if $x - \bar{x}$ is orthogonal to $C$ in $\mathcal{H}_B$, i.e.*

$$\langle x - \bar{x}, z\rangle_B = 0, \quad \forall z \in C. \tag{5.42}$$

*Proof.* Note that $\beta_1 = \beta_3$ can be taken as the maximum eigenvalue of $B$, and $\beta_2$ as the minimum eigenvalue. Let's examine each statement:

1. Fix $x \in \mathbb{R}^n$, then the statement is equivalent to

$$\text{find } z \in C, \quad \text{s.t. } \|x - z\|_B \leq \|x - y\|_B, \ \forall y \in C. \tag{5.43}$$

   Define function $g(z) = \|z - x\|_B$ and $g$ is continuous on a closed set $C$. By the Weierstrass' extreme value theorem, there exists a minimizing vector for function $g$ on $C$. The uniqueness follows since $g$ is strictly convex.

2. For all $x, y \in C$,

$$\|z - x\|_B^2 = \|z - \bar{x}\|_B^2 + \|\bar{x} - x\|_B^2 - 2\langle x - \bar{x}, z - \bar{x}\rangle_B \geq \|\bar{x} - x\|_B^2 - 2\langle x - \bar{x}, z - \bar{x}\rangle_B . \tag{5.44}$$

If $\langle x - \bar{x}, z - \bar{x}\rangle_B \leq 0$ for all $z \in C$, then we have,

$$\|z - x\|_B^2 \geq \|\bar{x} - x\|_B^2, \quad \forall z \in C, \tag{5.45}$$

which means $\bar{x} = P_{B,C}(x)$. On the other hand, suppose $\bar{x} = P_{B,C}(x)$, denote $h(\alpha) = \alpha z + (1-\alpha)\bar{x}$, $h$ is continuous and $h(0) = \bar{x}$.

$$\|x - h(\alpha)\|_B^2 = \alpha^2\|x - z\|_B^2 + (1-\alpha)^2\|x - \bar{x}\|_B^2 + 2\alpha(1-\alpha)\langle x - z, x - \bar{x}\rangle_B . \tag{5.46}$$

Consider the derivative

$$\begin{aligned} \frac{\partial}{\partial \alpha}\left(\|x - h(\alpha)\|_B^2\right)_{\alpha=0} &= -2\|x - \bar{x}\|_B^2 + 2\langle x - z, x - \bar{x}\rangle_B \\ &= -2\langle x - \bar{x}, z - \bar{x}\rangle_B . \end{aligned} \tag{5.47}$$

Suppose $\langle x - \bar{x}, z - \bar{x}\rangle_B > 0$ then $\frac{\partial}{\partial \alpha}\left(\|x - h(\alpha)\|_B^2\right)_{\alpha=0} < 0$. By continuity of $\|x - h(\alpha)\|_B^2$, there exists an $\alpha_0$ small enough such that

$$\|x - h(\alpha)\|_B^2 \leq \|x - h(0)\|_B^2 = \|x - \bar{x}\|_B^2, \tag{5.48}$$

contradict to $\bar{x} = P_{B,C}(x)$.

3. For all $x, y \in \mathbb{R}^n$, by 1 and 2 we have,

$$\langle x - P_{B,C}(x), P_{B,C}(y) - P_{B,C}(x)\rangle_B \leq 0, \tag{5.49}$$

$$\langle y - P_{B,C}(y), P_{B,C}(x) - P_{B,C}(y)\rangle_B \leq 0. \tag{5.50}$$

Then we have,

$$\langle y - P_{B,C}(y) - x + P_{B,C}(x), P_{B,C}(x) - P_{B,C}(y)\rangle_B \leq 0, \tag{5.51}$$

$$\|P_{B,C}(x) - P_{B,C}(y)\|_B^2 \leq \langle B(x - y), P_{B,C}(x) - P_{B,C}(y)\rangle . \tag{5.52}$$

By the assumption of matrix $B$ and Cauchy–Schwarz inequality, we have

$$\|P_{B,C}(x) - P_{B,C}(y)\| \leq \frac{\beta_1}{\beta_2}\|x - y\|. \tag{5.53}$$

The continuity follows by the above inequality.

4. Since $C$ is a subspace of $\mathbb{R}^n$, $\bar{x} + z$ and $\bar{x} - z$ in $C$ for all $z \in C$. Then by the equation in 2,

$$\langle x - \bar{x}, \bar{x} + z - \bar{x}\rangle_B \leq 0, \quad \text{and } \langle x - \bar{x}, \bar{x} - z - \bar{x}\rangle_B \leq 0, \tag{5.54}$$

then, $\langle x - \bar{x}, z\rangle_B = 0$.

□

Suppose the matrix $B_\nu$ is symmetric positive definite for each $\nu$ and converges to $B$ as $\nu \to \infty$. Next, we describe the behavior of scaled projection mappings $P_{B_\nu,C}$, $P_{B,C^\nu}$, and $P_{B,C}$ as $\nu \to \infty$. Because every symmetric positive definite matrix has a unique Cholesky decomposition, matrix $B_\nu, B$ can be written as

$$B_\nu = L_\nu L_\nu^{'}, \quad B = LL^{'}, \tag{5.55}$$

here $L_\nu$ and $L$ are lower triangular, invertible, and with real and positive diagonal entries, $L_\nu \to L$ as $\nu \to \infty$. Suppose matrix $B_\nu, B$ satisfy the assumption in (5.38), then

$$\beta_2\|x\| \leq \|L_\nu x\| \leq \beta_3\|x\|, \quad \beta_2\|x\| \leq \|Lx\| \leq \beta_3\|x\|. \tag{5.56}$$

Also, $L_\nu$ and $L$ can be written as bounded linear set-valued mappings map $\mathbb{R}^n \to \mathbb{R}^n$ as

$$L(C) = \left\{x \mid L^{-1}x \in C\right\}. \tag{5.57}$$

**Proposition 5.19.** *Given symmetric positive definite matrix $B$, $L$ is an invertible matrix given by the Cholesky decomposition with $B = LL^{'}$. Suppose $C \subset \mathbb{R}^n$ is nonempty, closed, and convex, then for any $x \in \mathbb{R}^n$, $P_{B,C}(x) = P_{L^{'}(C)}(L^{'}x)$.*

*Proof.*

$$P_{B,C}(x) = \arg\min_{z\in C}\|x - z\|_B = \arg\min_{z\in C}\|L^{'}(x - z)\| = \arg\min_{z\in L^{'}(C)}\|L^{'}x - z\| = P_{L^{'}(C)}(L^{'}x). \tag{5.58}$$

□

**Lemma 5.20.** *Suppose $C^\nu, C \subset \mathbb{R}^n$ are nonempty, closed, and convex, $C^\nu \to C$. And suppose $L_\nu, L$ are bounded linear operators derived by Cholesky decomposition* (5.55)*, and satisfies* (5.56)*, $L_\nu \to L$ as $\nu \to \infty$.*

1. *Set $L_\nu(C)$, $L(C^\nu)$, $L(C)$ are nonempty, closed, and convex for all $\nu$.*
2. *$L(C^\nu) \to L(C)$.*
3. *$L_\nu(C) \to L(C)$.*

*Proof.* 1. We only need to show the case of $L(C)$. $L(C)$ is nonempty follows that $C$ is nonempty. Since $L$ is a bijection and is continuous, then it is a closed map, $L(C)$ is closed. The convexity of $L(C)$ follows $L$ is a linear operator and $C$ is convex.

2. For all $x \in L(C)$, we have $L^{-1}x \in C$. By Definition 5.1, there exists $N_1 \in \mathcal{N}_\infty$ and sequence $x_1^\nu \in C^\nu$ with $x_1^\nu \underset{N_1}{\to} L^{-1}x$. Then we have $Lx_1^\nu \in L(C^\nu)$ with $Lx_1^\nu \underset{N_1}{\to} x$, that is $x \in \liminf_\nu L(C^\nu)$. So $L(C) \subset \liminf_\nu L(C^\nu)$.

   For all $x \in \limsup_\nu L(C^\nu)$, by Definition 5.1, there exists $N_2 \in \mathcal{N}_\infty^\sharp$ and sequence $x_2^\nu \in L(C^\nu)$ with $x_2^\nu \underset{N_2}{\to} x$. Then we have $L^{-1}x_2^\nu \in C^\nu$ with $L^{-1}x_2^\nu \underset{N_2}{\to} L^{-1}x$, that is $L^{-1}x \in \limsup_\nu C = C$. Then $x \in L(C)$ and $\limsup_\nu L(C^\nu) \subset L(C)$.

   Then

$$L(C) = \liminf_\nu L(C^\nu) = \limsup_\nu L(C^\nu). \tag{5.59}$$

3. First, we show $L(C) \subset \liminf_\nu L_\nu(C)$. For all $x \in L(C)$, let $N_1 = \{1, 2, 3, \cdots\} \in \mathcal{N}_\infty$, there exists $x^\nu = L_\nu L^{-1}x \in L_\nu(C)$ and $x^\nu \underset{N_1}{\to} x$. By Definition 5.1, $x \in \liminf_\nu L_\nu(C)$, then $L(C) \subset \liminf_\nu L_\nu(C)$. Also by Definition 5.1 we have $\liminf_\nu L_\nu(C) \subset \limsup_\nu L_\nu(C)$. Suppose there exists a point $x_0 \in \limsup_\nu L_\nu(C)$ but $x_0 \notin L(C)$. Then there exists $N_2 \in \mathcal{N}_\infty^\sharp$, and $x_0^\nu \in L_\nu(C)$ with $x_0^\nu \underset{N_2}{\to} x_0$ as $\nu \to \infty$. By

$$\begin{aligned} \|L_\nu^{-1}x_0^\nu - L^{-1}x_0\| &= \|L_\nu^{-1}x_0^\nu - L_\nu^{-1}x_0 + L_\nu^{-1}x_0 - L^{-1}x_0\| \\ &\le \|L_\nu^{-1}\|\|x_0^\nu - x_0\| + \|L_\nu^{-1} - L^{-1}\|\|x_0\|, \end{aligned} \tag{5.60}$$

   we have $L_\nu^{-1}x_0^\nu \underset{N_2}{\to} L^{-1}x_0 \in C$ since $C$ is closed, contradicting that $x_0 \notin L(C)$. Then we have $\limsup_\nu L_\nu(C) \subset L(C)$. In summary,

$$L(C) = \liminf_{\nu\to\infty} L_\nu(C) = \limsup_{\nu\to\infty} L_\nu(C). \tag{5.61}$$

□

The following propositions are useful for convergence analysis.

**Proposition 5.21.** *Given a symmetric positive definite matrix $B$. Suppose $C^\nu, C \subset \mathbb{R}^n$ are nonempty, closed, and convex, $C^\nu \to C$. Then the scaled projection mapping sequence $P_{B,C^\nu}$ converges continuously to $P_{B,C}$.*

*Furthermore, suppose $X$ is a subset in $\mathbb{R}^n$, the scaled projection mapping sequence $P_{B,C^\nu}$ converges uniformly to $P_{B,C}$ on all compact subsets of $X$.*

*Proof.* By the Cholesky decomposition, there exists an invertible positive definite matrix $L$ with $B = LL^{'}$. For all $\bar{x} \in \mathbb{R}^n$, $L^{'}\bar{x} \in \mathbb{R}^n$.

By Lemma 5.20, $L^{'}(C^\nu)$, $L^{'}$ are nonempty, closed, and convex sets, $L^{'}(C^\nu) \to L(C)$ and no subsequence escapes to the horizon. By Proposition 5.14, $P_{L^{'}(C^\nu)}$ converges graphically to $P_{L^{'}(C)}$ at $L^{'}\bar{x}$. By Lemma 5.16, the projection function sequence $\{P_{L^{'}(C^\nu)}\}$ is eventually locally bounded at $\bar{x}$.

Then, by Corollary 5.12, the projection function sequence $P_{L^{'}(C^\nu)}$ converges continuously to $P_{L^{'}(C)}$ at $L^{'}\bar{x}$, i.e., for all $\varepsilon > 0$, $\rho > 0$, there exists $N \in \mathcal{N}_\infty$ along with a neighborhood $V \in \mathcal{N}(L^{'}\bar{x})$, such that

$$P_{L^{'}(C^\nu)}(x) \cap \rho\mathbb{B} \subset P_{L^{'}(C)}(L^{'}\bar{x}) + \varepsilon\mathbb{B}, \tag{5.62}$$

$$P_{L^{'}(C)}(L^{'}\bar{x}) \cap \rho\mathbb{B} \subset P_{L^{'}(C^\nu)}(x) + \varepsilon\mathbb{B}, \tag{5.63}$$

for all $x \in V$ when $\nu \in N$. By Proposition 5.19, $P_{L^{'}(C^\nu)}(x) = P_{B,C^\nu}((L^{'})^{-1}x)$, $P_{L^{'}(C)}(L^{'}\bar{x}) = P_{B,C}(\bar{x})$. Also, $(L^{'})^{-1}x \in (L^{'})^{-1}(V)$, and $(L^{'})^{-1}(V) \in \mathcal{N}((L^{'})^{-1}L^{'}\bar{x}) = \mathcal{N}(\bar{x})$. Then we have the following equivalence of the above equations,

$$P_{B,C^\nu}(x) \cap \rho\mathbb{B} \subset P_{B,C}(\bar{x}) + \varepsilon\mathbb{B}, \tag{5.64}$$

$$P_{B,C}(\bar{x}) \cap \rho\mathbb{B} \subset P_{B,C^\nu}(x) + \varepsilon\mathbb{B}, \tag{5.65}$$

for all $x \in (L^{'})^{-1}(V)$ when $\nu \in N$. That is the scaled projection function sequence $P_{B,C^\nu}$ converges continuously to $P_{B,C}$ at $\bar{x}$. Since $\bar{x}$ is an arbitrary point in $\mathbb{R}^n$, $P_{B,C^\nu}$ converges continuously to $P_{B,C}$.

By Theorem 5.10, $P_{B,C^\nu}$ converges uniformly to $P_{B,C}$ on all compact subsets of $X$.

□

**Proposition 5.22.** *Given symmetric positive definite matrices $B_\nu, B$ with sequence $B_\nu$ converges to $B$ as $\nu \to \infty$. Suppose $C \subset \mathbb{R}^n$ is nonempty, closed, and convex. Then the scaled projection mapping sequence $P_{B_\nu,C}$ converges continuously to $P_{B,C}$.*

*Furthermore, suppose $X$ is a subset in $\mathbb{R}^n$, the scaled projection mapping sequence $P_{B_\nu,C}$ converges uniformly to $P_{B,C}$ on all compact subsets of $X$.*

*Proof.* The proof is similar to the proof of Proposition 5.21.

□

## 5.3 Gradient projection method with inexact projection

In this section, we discuss the gradient projection method with inexact projection. Before introducing the proposed algorithm, we give a review of the traditional gradient projection method. Consider the constrained optimization problem:

$$\min_x f(x), \quad \text{such that } x \in X, \tag{5.66}$$

where $f$ is a continuous differentiable nonlinear function, possibly nonconvex, bounded from below. The constraint set $X \subset \mathbb{R}^n$ is nonempty, closed, and convex. First, we review some basic properties.

**Proposition 5.23** ([17] proposition 2.1.2, optimality condition)**.** *For problem* (5.66)*,*

1. *If $x^*$ is a local minimum of $f$ over $X$, then*

$$\nabla f(x^*)'(x - x^*) \geq 0, \quad \forall x \in X. \tag{5.67}$$

2. *If $f$ is convex over $X$, then the above condition is also sufficient for $x^*$ to minimize $f$ over $X$.*

The above proposition gives the optimality condition for the constrained optimization problem (5.66).

**Definition 5.24** (Stationary point)**.** *The point $x^*$ is a stationary point of the constrained optimization problem* (5.66) *if the optimality condition Proposition 5.23 is satisfied.*

**Definition 5.25** (Feasible direction method)**.** *For the constrained optimization problem* (5.66)*, a feasible direction method starts with a feasible point $x^0 \in X$, then generating a series of points with equation*

$$x^{k+1} = x^k + \alpha^k d^k. \tag{5.68}$$

*If $x^k$ is a nonstationary point, then $d^k$ is chosen as a feasible direction at $x^k$ which is a descent direction*

*with*

$$\left\langle \nabla f(x^k), d^k \right\rangle < 0. \tag{5.69}$$

*Moreover, there exists a stepsize* $\alpha^k > 0$ *such that*

$$x^k + \alpha^k d^k \in X. \tag{5.70}$$

*If* $x^k$ *is stationary, the method stops with* $x^{k+1} = x^k$.

The gradient projection method can be given by the following equations. At the $k$-th iteration, compute

$$\bar{x}^k = P_X(x^k - \beta^k \nabla f(x^k)), \tag{5.71}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k), \tag{5.72}$$

where $P_X$ is a single-valued projection mapping by Proposition 5.13, $\beta^k$ is a positive scalar and $\alpha^k \in (0, 1]$ is the stepsize parameter generally achieved through the line search algorithm. The positive scalar $\beta^k$ can be chosen as a constant, and more sophisticated strategy such as Armijo rule along the projection arc can be used, for more information please refer to [17]. We discuss two line search methods in this work.

1. Armijo rule [99]:

$$f(x^k + \alpha^k d^k) \leq f(x^k) + c_1 \alpha^k \nabla f(x^k)' d^k, \tag{5.73}$$

   with $c_1 \in (0, 1)$, and $\alpha^k = \eta^j \alpha$, here $\alpha$ is the initial step size, $\eta \in (0, 1)$ and $j$ is the corresponding line search time.

2. Wolfe conditions [99]:

$$f(x^k + \alpha^k d^k) \leq f(x^k) + c_1 \alpha^k \nabla f(x^k)' d^k, \tag{5.74}$$

$$\nabla f(x^k + \alpha^k d^k)' d^k \geq c_2 \nabla f(x^k)' d^k, \tag{5.75}$$

   where $0 < c_1 < c_2 < 1$. The line search stepsize $\alpha^k$ is chosen as the above Armijo rule.

**Proposition 5.26** (Existence of the feasible direction)**.** *The gradient projection method* (5.71) *is a feasible direction method.*

*Proof.* Notice that if $x^k$ is a nonstationary point, $x^{k+1} \neq x^k$, then $\bar{x}^k \neq x^k$ and $\nabla f(x^k) \neq 0$. Denote the feasible direction as $d^k = \bar{x}^k - x^k$.

If $x^k - \beta^k \nabla f(x^k) \in X$, $d^k = -\beta^k \nabla f(x^k)$ and $\left\langle \nabla f(x^k), d^k \right\rangle < 0$, then $d^k$ is a feasible direction. Also we can have $x^k + \alpha^k d^k = x^k - \alpha^k \beta^k \nabla f(x^k) \in X$ since $\alpha^k \in (0, 1]$ and $X$ is convex, then there exists an update $x^{k+1} \in X$.

On the other hand, if $x^k - \beta^k \nabla f(x^k) \notin X$, $d^k = \bar{x}^k - x^k$, by Proposition 5.13

$$\left\langle x^k - \beta^k \nabla f(x^k) - \bar{x}^k, x - \bar{x}^k \right\rangle \leq 0, \quad \forall x \in X, \tag{5.76}$$

let $x = x^k$,

$$\left\langle x^k - \beta^k \nabla f(x^k) - \bar{x}^k, x^k - \bar{x}^k \right\rangle \leq 0, \tag{5.77}$$

$$\left\langle \nabla f(x^k), d^k \right\rangle \leq 0. \tag{5.78}$$

Since $x^k$ is nonstationary then $\nabla f(x^k) \neq 0$, also $x^{k+1} \neq x^k$ then $x^k \neq \bar{x}^k$, $\left\langle \nabla f(x^k), d^k \right\rangle < 0$. Then $d^k$ is a feasible direction. Also $x^k + \alpha^k d^k = x^k + \alpha^k (\bar{x}^k - x^k) \in X$ since $\alpha^k \in (0, 1]$ and $X$ is convex. □

**Definition 5.27** (Gradient related direction)**.** *For a feasible direction method, the direction sequence $\{d^k\}$ is gradient related to $\{x^k\}$ if for any subsequence $\{x^k\}_{k \in \mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k \in \mathcal{K}}$ is bounded and satisfies*

$$\limsup_{k \to \infty, k \in \mathbb{K}} \left\langle \nabla f(x^k), d^k \right\rangle < 0. \tag{5.79}$$

An important proposition showing the convergence behavior of the feasible direction method is provided below.

**Proposition 5.28** ([17], proposition 2.2.1, stationarity of limit points for feasible direction methods)**.** *Let $\{x_k\}$ be a sequence generated by the feasible direction method $x^{k+1} = x^k + \alpha^k d^k$. Assume $\{d^k\}$ is gradient related and $\alpha^k$ is chosen by the Armijo rule. Then every limit point of $\{x_k\}$ is a stationary point.*

Proposition 5.28 is given in [17] and the proof is omitted. For the completeness we give the proof below. The proof is similar to Proposition 1.2.1 in [17].

*Proof.* Assume $\bar{x}$ is a limit point of $\{x^k\}$ and $\bar{x}$ is on the boundary of $X$. Suppose $\bar{x}$ is a nonstationary point,

i.e. there exists $x_0 \in X$ such that

$$\langle \nabla f(\bar{x}), x_0 - \bar{x} \rangle < 0. \tag{5.80}$$

In this case we have $\nabla f(\bar{x}) \neq 0$. At the $k$-th iteration

$$f(x^k) = f(x^0) + \sum_{j=1}^{k} f(x^j) - f(x^{j-1}). \tag{5.81}$$

Since $f$ is bounded from below, then $f(x^k)$ converges to a finite value then we can have $(f(x^k) - f(x^{k+1})) \to 0$, as $k \to \infty$. By the definition of Armijo rule (5.73),

$$f(x^k) - f(x^{k+1}) \geq -c_1 \alpha^k \left\langle \nabla f(x^k), d^k \right\rangle, \tag{5.82}$$

then we have $\alpha^k \left\langle \nabla f(x^k), d^k \right\rangle \to 0$ as $k \to \infty$. Let $\{x^k\}_{k \in \mathcal{K}}$ be a subsequence converges to $\bar{x}$, and since the corresponding direction $\{d^k\}_{k \in \mathcal{K}}$ is gradient related, we have

$$\limsup_{k \to \infty, k \in \mathcal{K}} \left\langle \nabla f(x^k), d^k \right\rangle < 0, \tag{5.83}$$

then $\alpha^k \to 0$ as $k \to \infty$ and $k \in \mathcal{K}$. Since $\{\alpha_k\}_{k \in \mathcal{K}}$ is decreasing to 0, from the definition of Armijo rule (5.73), there exists an index set $N \in \mathcal{N}_\infty$ such that for all $k \in N$, the corresponding Armijo stepsize $\alpha^k$ has to decrease once. That means if the stepsize is enlarged once with $\alpha^k/\eta$, the Armijo rule (5.82) will not be satisfied for $k \in N$, i.e.,

$$f(x^k) - f(x^k + \alpha^k d^k/\eta) < -c_1 \alpha^k/\eta \left\langle \nabla f(x^k), d^k \right\rangle, \quad \forall k \in N \cap \mathcal{K}. \tag{5.84}$$

Denote

$$p^k = \frac{d^k}{\|d^k\|}, \quad \bar{\alpha}^k = \frac{\alpha^k \|d^k\|}{\eta}. \tag{5.85}$$

Since $\{d^k\}$ is gradient related, $\{d^k\}_{k \in \mathcal{K}}$ is bounded, we have $\bar{\alpha}^k \to 0$ as $k \in \mathcal{K}$ and $k \to \infty$. Since $\|p^k\| = 1$ for all $k \in \mathcal{K}$, there exists a bounded subsequence $\{p^k\}_{k \in \bar{\mathcal{K}}}$ and $\bar{\mathcal{K}} \subset \mathcal{K}$ such that $p^k \to \bar{p}$ as $k \in \bar{\mathcal{K}}$ and $k \to \infty$. Also we have $\|\bar{p}\| = 1$. From equation (5.84) we have

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k p^k)}{\bar{\alpha}^k} < -c_1 \left\langle \nabla f(x^k), p^k \right\rangle. \tag{5.86}$$

Taking the limits in the above equation,

$$-\langle \nabla f(\bar{x}), \bar{p} \rangle \leq -c_1 \langle \nabla f(\bar{x}), \bar{p} \rangle , \tag{5.87}$$

then $\langle \nabla f(\bar{x}), \bar{p} \rangle \geq 0$ since $c_1 \in (0, 1)$.

On the other hand,

$$\langle \nabla f(\bar{x}), \bar{p} \rangle \leq \limsup_{k \in \bar{\mathcal{K}}, k \to \infty} \left\langle \nabla f(x^k), p^k \right\rangle = \limsup_{k \in \bar{\mathcal{K}}, k \to \infty} \frac{\left\langle \nabla f(x^k), d^k \right\rangle}{\|d^k\|} < 0, \tag{5.88}$$

the last inequality holds since $\{d^k\}_{k \in \bar{\mathcal{K}}}$ is gradient related. Then we have the contradiction.

For the case when $\bar{x}$ is in the interior of $X$, we have $\nabla f(\bar{x}) = 0$ if and only if $\bar{x}$ is stationary. By assuming $\nabla f(\bar{x}) \neq 0$, with the same discussion we can have the contradiction. □

### 5.3.1 Proposed algorithm

Next, we consider the gradient projection method with inexact projection. Rewrite the optimization problem (5.66),

$$\min_x f(x), \quad \text{such that } x \in X. \tag{5.89}$$

In the gradient projection method, at the $k$-th iteration, we compute $\tilde{x}^k = x^k - \beta^k \nabla f(x^k)$ first, then project $\tilde{x}^k$ to feasible set $X$ by $\bar{x}^k = P_X(\tilde{x}^k)$. When the projection operator can not be evaluated in a closed-form, we can only project $\tilde{x}^k$ to a point close to $P_X(x^k)$. In order to solve this problem, we construct a set expanding strategy for the feasible set $X$.

**Definition 5.29** (Set expanding strategy)**.** *Given a feasible set $X$ is nonempty, closed, and convex, construct an expanding set sequence $\{X^k\}$ such that*

$$X = \lim_{k \to \infty} X^k, \quad X^k \subset X^{k+1}, \quad X^k \neq X^{k+1}, \tag{5.90}$$

*where $X^k$ is nonempty, closed, and convex for each $k \in \mathbb{N}$.*

One example for the set expanding strategy is:

$$X^k = \left\{ x \in \mathbb{R}^2 \;\middle|\; x_1 \leq \theta(k) \right\}, \tag{5.91}$$

where $\theta(k)$ is a threshold function defined by

$$\theta(k) = \begin{cases} 0, & \text{if } k = 0, \\ \sum_{i=1}^{k} \eta^i \varepsilon, & \text{if } k \geq 1, \end{cases} \tag{5.92}$$

and

$$\lim_{k \to \infty} \theta(k) = \frac{\eta}{1-\eta} \varepsilon, \tag{5.93}$$

here $\varepsilon > 0$, $\eta \in (0, 1)$. Then the constraint set $X$ is

$$X = \left\{ x \in \mathbb{R}^2 \;\middle|\; x_1 \leq \frac{\eta}{1-\eta} \varepsilon \right\}. \tag{5.94}$$

This example is illustrated in Figure 5.3.

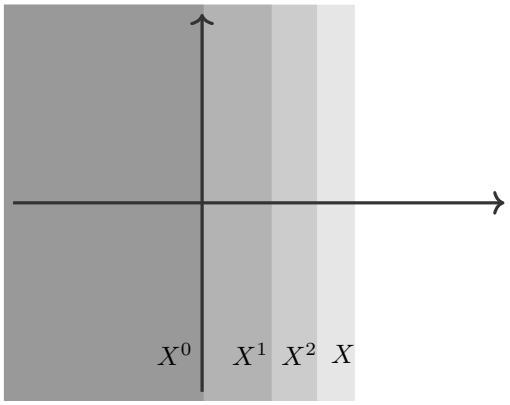


Figure 5.3: Example of a set expanding strategy in $\mathbb{R}^2$.

There are two strategies to construct the expanding set sequence for the constrained optimization problem (5.66):

1. Given the constraint set $X$, construct the initial constraint set $X^0 = X$, then the element in the expanding set sequence $\{X^k\}$ is larger than the initial constraint set $X$. A special expanding strategy like the above example can be designed such that the sequence $\{X^k\}$ does not expand too much compared to $X$. In this case, the expanding set sequence strategy adds a "soft" constraint to the optimization problem.

2. Start with a smaller initial constraint set $X^0 \subset X$, then design an expanding set sequence $\{X^k\}$ with $X^k \to X$. In this case, the constraint set $X$ is not changed.

**Proposition 5.30.** *The set sequence $\{X^k\}$ generated by the set expanding strategy 5.29 has no subsequence escaping to the horizon.*

*Proof.* Since $\lim_{k\to\infty} X^k = X$, and $X$ is nonempty, then it is obvious that no subsequence escaping to the horizon. □

Since the projection algorithm discussed in this work is an iterative process, we make an assumption for the projection algorithm.

**Assumption 5.31.** *Starting with $z^0 = \tilde{x}^k$, the projection function $P_{X^k}(\tilde{x}^k)$ generates a series of $\{z^j\}$ which converges to $P_{X^k}(\tilde{x}^k)$ strongly.*

Under the above assumption, we define an inexact projection function which projects $\tilde{x}^k$ towards $X^k$, denoted as $\bar{P}_{X^k}(\tilde{x}^k) = z^{j_0}$. Here $j_0$ is the index of sequence $\{z^j\}$ in Assumption 5.31 such that

$$z^{j_0} \in X^{k+1}, \tag{5.95}$$

$$\left\langle \tilde{x}^k - z^{j_0}, x^k - z^{j_0} \right\rangle \leq 0. \tag{5.96}$$

The above equations provide the stopping criterion of the iterative projection process.

**Proposition 5.32.** *When $x^k$ is a nonstationary point, under Assumption 5.31, there exists an index $j_0$ such that $z^{j_0}$ in the converging sequence $\{z^j\}$ such that equation* (5.95) *and* (5.96) *are satisfied.*

*Proof.* At the $k$-th iteration, $\tilde{x}^k = x^k - s^k \nabla f(x^k)$. Since $X^k$ is nonempty, closed, and convex, $P_{X^k}(\tilde{x}^k)$ exists by Proposition 5.13. Since $X^k \subset \text{int}X^{k+1}$, there exists an open ball $\mathbb{B}(P_{X^k}(\tilde{x}^k), \varepsilon) \subset X^{k+1}$. By Assumption 5.31, the sequence $\{z^j\}$ converges strongly to $P_{X^k}(\tilde{x}^k)$, then there exists an index set $N_1 \in \mathcal{N}_\infty$ such that $z^j \in X^{k+1}$ for all $j \in N_1$.

When $x^k$ is a nonstationary point, then $x^k \neq P_{X^k}(\tilde{x}^k)$, by Proposition 5.13

$$\left\langle \tilde{x}^k - P_{X^k}(\tilde{x}^k), x^k - P_{X^k}(\tilde{x}^k) \right\rangle < 0. \tag{5.97}$$

Then,

$$\begin{aligned} &\left\langle \tilde{x}^k - P_{X^k}(\tilde{x}^k), x^k - P_{X^k}(\tilde{x}^k) \right\rangle \\ =&\left\langle \tilde{x}^k - z^j, x^k - z^j \right\rangle + \left\langle \tilde{x}^k - z^j, z^j - P_{X^k}(\tilde{x}^k) \right\rangle \\ &+ \left\langle z^j - P_{X^k}(\tilde{x}^k), x^k - z^j \right\rangle + \left\langle z^j - P_{X^k}(\tilde{x}^k), z^j - P_{X^k}(\tilde{x}^k) \right\rangle < 0. \end{aligned} \tag{5.98}$$

Suppose $\left\langle \tilde{x}^k - z^j, x^k - z^j \right\rangle > 0$, then as $z^j$ converges to $P_{X^k}(\tilde{x}^k)$,

$$\left\langle \tilde{x}^k - P_{X^k}(\tilde{x}^k), x^k - P_{X^k}(\tilde{x}^k) \right\rangle \geq 0, \tag{5.99}$$

yielding a contradiction.

By Assumption 5.31, there exist an index set $N_2 \in \mathcal{N}_\infty$ such that $\left\langle \tilde{x}^k - z^j, x^k - z^j \right\rangle \leq 0$. Let $j_0$ be an index in $N_1 \cap N_2$ then proof is finished. □

The set expanding strategy can be applied for the case when the gradient projection method has an inexact projection: consider the case when $\beta^k$ is constant, at the $k$-th iteration, set the feasible set as $X^k$ then compute

$$\bar{x}^k = \bar{P}_{X^k}(x^k - \beta \nabla f(x^k)), \tag{5.100}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k). \tag{5.101}$$

This process is illustrated in Figure 5.4.

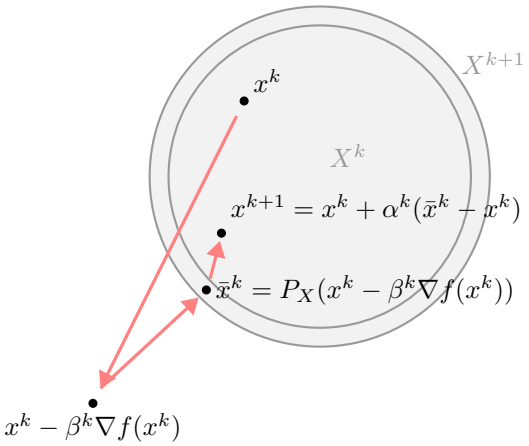


Figure 5.4: Gradient projection method with inexact projection at the $k$-th iteration.

The gradient projection method with inexact projection is described as Algorithm 5.

### 5.3.2 Convergence analysis

**Proposition 5.33.** *Algorithm 5 is a feasible direction method.*

*Proof.* With the condition (5.96), the proof is exactly the same as Proposition 5.26. □

**Algorithm 5:** Gradient projection method with inexact projection

Initialization: Given feasible set $X$, construct set sequence $\{X^k\}$ satisfies (5.90). Given the initial point $x^0 \in X^0$.

**while** *not convergent* **do**

Step 1: Compute $\tilde{x}^k = x^k - \beta \nabla f(x^k)$ ;

Step 2: Projecting $\tilde{x}^k$ to $X^k$, until equation (5.95) and (5.96) are satisfied, denote $\bar{x}^k = \bar{P}_{X^k}(\tilde{x}^k)$;

Step 3: Evaluate the line search stepsize with Armijo rule or Wolfe conditions, update with $x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k)$;

Step 4: Enlarge the feasible set $X^k = X^{k+1}$ with set expanding strategy, let $k = k + 1$.

**end**

**Assumption 5.34.** *The inexact projection function project $\tilde{x}^k$ is close enough to its exact projection $P_{X^k}(\tilde{x}^k)$ with*

$$\|\bar{P}_{X^k}(\tilde{x}^k) - P_{X^k}(\tilde{x}^k)\| \leq d(X^k, X^{k+1}), \tag{5.102}$$

*here the $d(X^k, X^{k+1})$ is the (integrated) set distance between $X^k$ and $X^{k+1}$.*

The above assumption make sense because $d(X^k, X^{k+1}) > 0$ as $X^k$ and $X^{k+1}$ are closed sets, $X^k \subset X^{k+1}$, $X^k \neq X^{k+1}$, and the projection function generating a sequence $\{z^j\}$ converges to $P_{X^k}(\tilde{x}^k)$ strongly by Assumption 5.31.

**Theorem 5.35.** *Under Assumption 5.31, 5.34, let $\{x^k\}$ be the sequence generated by Algorithm 5. Then every limit point of $\{x^k\}$ is stationary.*

*Proof.* Suppose there is a subsequence $\{x^k\}_{k\in\mathcal{K}}$ converges to a nonstationary point $x_0$. By Proposition 5.28, it is sufficient to show the update direction sequence $d^k = \bar{x}^k - x^k$ as $k \in \mathcal{K}$ is gradient related, i.e.

$$\limsup_{k\to\infty, k\in\mathcal{K}} \|\bar{x}^k - x^k\| < \infty, \tag{5.103}$$

$$\limsup_{k\to\infty, k\in\mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle < 0. \tag{5.104}$$

Denote $\tilde{x}^k = x^k - \beta\nabla f(x^k)$, $\tilde{x} = x_0 - \beta\nabla f(x_0)$, then $\tilde{x}^k \to \tilde{x}$ as $k \in \mathcal{K}$ and $k \to \infty$. We can assume that there exists a compact subset $\Omega \subset \mathbb{R}^n$ large enough such that all points in the sequence $\{\tilde{x}^k\}$ are in $\Omega$.

Recall that the inexact projection of $\tilde{x}^k$ is $\bar{x}^k = \bar{P}_{X^k}(\tilde{x}^k)$, consider the inequality,

$$\begin{aligned} &\|\bar{P}_{X^k}(\tilde{x}^k) - P_X(\tilde{x})\| \\ &\leq \|\bar{P}_{X^k}(\tilde{x}^k) - P_{X^k}(\tilde{x}^k)\| + \|P_{X^k}(\tilde{x}^k) - P_X(\tilde{x}^k)\| + \|P_X(\tilde{x}^k) - P_X(\tilde{x})\|. \end{aligned} \tag{5.105}$$

By Proposition 5.13, $P_X$ is continuous, for all $\varepsilon/3 > 0$, there exists $N_1 \in \mathcal{N}_\infty$, such that $\|P_X(\tilde{x}^k) -$

$P_X(\tilde{x})\| < \varepsilon/3$.

By Definition 5.1, Proposition 5.30, and Proposition 5.17, the projection function sequence $P_{X^k}$ converges uniformly to $P_X$ on all compact subsets of $\mathbb{R}^n$. Then the uniform convergence holds on the set $\Omega$. For $\varepsilon/3 > 0$, for all $\rho > 0$, there exists $N_2 \in \mathcal{N}_\infty$ such that

$$P_{X^k}(x) \cap \rho\mathbb{B} \subset P_X(x) + \varepsilon/3\mathbb{B}, \tag{5.106}$$

$$P_X(x) \cap \rho\mathbb{B} \subset P_{X^k}(x) + \varepsilon/3\mathbb{B}, \tag{5.107}$$

for all $x \in \Omega$ when $k \in N_2$. Then we have $\|P_{X^k}(\tilde{x}^k) - P_X(\tilde{x}^k)\| \leq \varepsilon/3$.

Since $X^k \to X$ and $X^k, X$ are nonempty closed bounded, by Theorem 5.3, $d(X^k, X) \to 0$. By Assumption 5.34, for $\varepsilon/6$, there exists $N_3 \in \mathcal{N}_\infty$ such that when $k, k+1 \in N_3$,

$$\|\bar{P}_{X^k}(\tilde{x}^k) - P_{X^k}(\tilde{x}^k)\| \leq d(X^k, X^{k+1}) \leq d(X^k, X) + d(X, X^{k+1}) < \varepsilon/6 + \varepsilon/6 = \varepsilon/3. \tag{5.108}$$

Then for the above arbitrary $\varepsilon > 0$, there exists $N = N_1 \cap N_2 \cap N_3$ such that $\|\bar{P}_{X^k}(\tilde{x}^k) - P_X(\tilde{x})\| < \varepsilon$ as $k \in N$. Which means $\bar{P}_{X^k}(\tilde{x}^k) \to P_X(\tilde{x})$ as $k \in \mathcal{K}$ and $k \to \infty$.

Then,

$$\|\bar{x}^k - x^k\| = \|\bar{P}_{X^k}(\tilde{x}^k) - x^k\| \to \|P_X(\tilde{x}) - x_0\|, \quad k \to \infty, k \in \mathcal{K}. \tag{5.109}$$

Since $P_X(\tilde{x}) \in X$ and $x_0 \in X$, then $\limsup_{k\to\infty, k\in\mathcal{K}} \|\bar{x}^k - x^k\| < \infty$.

By inequality (5.96),

$$\left\langle \tilde{x}^k - \bar{x}^k, x^k - \bar{x}^k \right\rangle \leq 0, \tag{5.110}$$

$$\left\langle x^k - \beta\nabla f(x^k) - \bar{x}^k, x^k - \bar{x}^k \right\rangle \leq 0, \tag{5.111}$$

$$\left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle \leq -\frac{1}{\beta}\|\bar{x}^k - x^k\|^2. \tag{5.112}$$

Taking the limit of above equation, since $\bar{x}^k = \bar{P}_{X^k}(\tilde{x}^k)$ and $\bar{P}_{X^k}(\tilde{x}^k) \to P_X(\tilde{x})$,

$$\limsup_{k\to\infty, k\in\mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle \leq -\frac{1}{\beta}\|P_X(\tilde{x}) - x_0\|^2. \tag{5.113}$$

Assume $P_X(\tilde{x}) = x_0$, then by Proposition 5.13,

$$\langle \tilde{x} - P_X(\tilde{x}), x - P_X(\tilde{x}) \rangle \leq 0, \quad \forall x \in X, \tag{5.114}$$

$$\langle -\beta \nabla f(x_0), x - x_0 \rangle \leq 0, \quad \forall x \in X, \tag{5.115}$$

$$\langle \nabla f(x_0), x - x_0 \rangle \geq 0, \quad \forall x \in X, \tag{5.116}$$

that means $x_0$ is stationary, yielding a contradiction. Then,

$$\limsup_{k \to \infty, k \in \mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle < 0. \tag{5.117}$$

□

## 5.4 Scaled gradient projection with inexact projection

Next, we discuss the scaled gradient projection method. For optimization problem (5.66), at the $k$-th iteration, we assume we have a symmetric positive definite matrix $B_k$ with a unique Cholesky decomposition $B_k = L_k L_k^{'}$. Here $L_k$ is an invertible positive definite lower triangular matrix. In the $\mathcal{H}_{B_k}$ space, rewrite the constrained optimization problem (5.66),

$$\min_y g_k(y) = f((L_k^{'})^{-1} y), \quad \text{such that } y \in L_k^{'}(X). \tag{5.118}$$

Perform the gradient projection method to the above problem,

$$y^{k+1} = y^k + \alpha^k (\bar{y}^k - y^k), \tag{5.119}$$

where

$$\bar{y}^k = P_{L_k^{'}(X)}(y^k - \beta^k \nabla g_k(y^k)). \tag{5.120}$$

By Proposition 5.13, the above equation is equivalent to

$$\begin{aligned}\bar{y}^k &= \arg\min_{y\in L_k'(X)} \|y-(y^k-\beta^k\nabla g_k(y^k))\|^2 \\ &= \arg\min_{y\in L_k'(X)} \|y-y^k\|^2+(\beta^k)^2\|\nabla g_k(y^k)\|^2+2\beta^k\left\langle \nabla g_k(y^k), y-y^k\right\rangle \\ &= \arg\min_{y\in L_k'(X)} \left\langle \nabla g_k(y^k), y-y^k\right\rangle+\frac{1}{2\beta^k}\|y-y^k\|^2.\end{aligned} \tag{5.121}$$

Let $y=L_k'x$, $y^k=L_k'x^k$, and $\nabla g_k(y^k)=L_k^{-1}\nabla f((L_k')^{-1}y^k)=L_k^{-1}\nabla f(x^k)$. Then,

$$\begin{aligned}\bar{x}^k &= \arg\min_{L_k'x\in L_k'(X)} \left\langle L_k^{-1}\nabla f(x^k), L_k'(x-x^k)\right\rangle+\frac{1}{2\beta^k}\left\langle L_k'(x-x^k), L_k'(x-x^k)\right\rangle \\ &= \arg\min_{x\in X} \left\langle \nabla f(x^k), x-x^k\right\rangle+\frac{1}{2\beta^k}\|x-x^k\|_{B_k}.\end{aligned} \tag{5.122}$$

The scaled gradient projection method can be written as: at the $k$-th iteration, given a symmetric positive definite matrix $B_k$, compute

$$\bar{x}^k = \arg\min_{x\in X} \left\langle \nabla f(x^k), x-x^k\right\rangle+\frac{1}{2\beta^k}\left\langle B_k(x-x^k), x-x^k\right\rangle, \tag{5.123}$$

$$x^{k+1}=x^k+\alpha^k(\bar{x}^k-x^k). \tag{5.124}$$

Here the matrix $B_k$ is chosen as an approximation of the Hessian matrix $\nabla^2 f(x^k)$. When the Hessian matrix is symmetric positive definite and $B_k=\nabla^2 f(x^k)$, the scaled gradient projection method is equivalent to the constrained Newton's method. The scale parameter $\beta^k$ can be set to 1 if the $B_k$ is an accurate approximation of the Hessian matrix. The line search parameter $\alpha^k$ can be achieved through the line search methods like Armijo rule and Wolfe conditions. The relation between line search method and the Hessian approximation will be discussed in the following subsections.

In this work, we consider the case that the accurate approximation is used for the Hessian matrix, so suppose the scale parameter $\beta^k=1$ for each iteration. Denote

$$f_k(x)=\left\langle \nabla f(x^k), x-x^k\right\rangle+\frac{1}{2}\left\langle B_k(x-x^k), x-x^k\right\rangle, \tag{5.125}$$

here the scaling matrix $B_k$ satisfies the following assumption.

**Assumption 5.36.** *The scaling matrix $B_k$ is symmetric positive definite, and there exists positive constants*

$\beta_1, \beta_2, \beta_3 \geq 0$, *such that*

$$\|B_k x\| \leq \beta_1 \|x\|, \quad \textit{and} \quad \beta_2 \|x\|^2 \leq \langle B_k x, x \rangle \leq \beta_3 \|x\|^2, \quad \forall x \in \mathbb{R}^n, \tag{5.126}$$

*for all iterations. Also, $B_k$ converges to a symmetric positive definite matrix $B$ as $k \to \infty$.*

The subproblem (5.123) of the scaled gradient projection method can be denoted as

$$\bar{x}^k = \arg\min_{x \in X} f_k(x). \tag{5.127}$$

**Proposition 5.37.** *Let $\tilde{x}^k = x^k - B_k^{-1} \nabla f(x^k)$, the equation* (5.125) *is equivalent to*

$$f_k(x) = \frac{1}{2} \|x - \tilde{x}^k\|_{B_k}^2 - \frac{1}{2} \left\langle B_k^{-1} \nabla f(x^k), \nabla f(x^k) \right\rangle. \tag{5.128}$$

*The subproblem* (5.123) *is equivalent to compute $\tilde{x}^k$ first, then project $\tilde{x}^k$ to feasible set $X$ in $\mathcal{H}_{B_k}$, i.e.*

$$\textit{find } \bar{x}^k \in X, \quad \textit{such that } \|\tilde{x}^k - \bar{x}^k\|_{B_k} \leq \|x - \bar{x}^k\|_{B_k}, \forall x \in X. \tag{5.129}$$

*There exists a unique $\bar{x}^k$ for each $\tilde{x}^k$, denote the projection function as $P_{B_k, X}(\tilde{x}^k) = \bar{x}^k$.*

*Proof.* Let $x^k = \tilde{x}^k + B_k^{-1} \nabla f(x^k)$, by equation (5.125),

$$\begin{aligned} f_k(x) = & \left\langle \nabla f(x^k), x - \tilde{x}^k \right\rangle - \left\langle \nabla f(x^k), B_k^{-1} \nabla f(x^k) \right\rangle \\ & + \frac{1}{2} \left\langle B_k (x - \tilde{x}^k), x - \tilde{x}^k \right\rangle - \frac{1}{2} \left\langle \nabla f(x^k), x - \tilde{x}^k \right\rangle \\ & - \frac{1}{2} \left\langle B_k (x - \tilde{x}^k), B_k^{-1} \nabla f(x^k) \right\rangle + \frac{1}{2} \left\langle \nabla f(x^k), B_k^{-1} \nabla f(x^k) \right\rangle \\ = & \frac{1}{2} \left\langle B_k (x - \tilde{x}^k), x - \tilde{x}^k \right\rangle - \frac{1}{2} \left\langle B_k^{-1} \nabla f(x^k), \nabla f(x^k) \right\rangle. \end{aligned} \tag{5.130}$$

Since $X$ is nonempty, closed, and convex, the existence and uniqueness of $\bar{x}^k$ is followed by Proposition 5.18. □

The scaled gradient projection method needs to compute the subproblem (5.123) at each iteration. By the above proposition, the subproblem can be solved by first computing $\tilde{x}^k = x^k - B_k^{-1} \nabla f(x^k)$, then projecting $\tilde{x}^k$ to $X$ with the metric of $\mathcal{H}_{B_k}$. Rewrite the scaled gradient projection method as: at the $k$-th iteration,

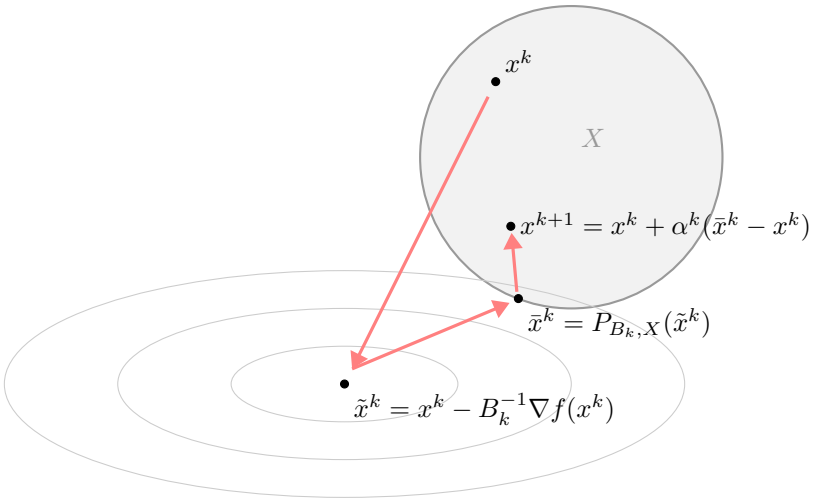


Figure 5.5: The scaled gradient projection method at the $k$-th iteration.

given a symmetric positive definite matrix $B_k$, compute

$$\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k), \tag{5.131}$$

$$\bar{x}^k = P_{B_k,X^k}(\tilde{x}^k), \tag{5.132}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k). \tag{5.133}$$

Figure 5.5 provides an illustration of the above process.

**Proposition 5.38.** *Under Assumption 5.36, the scaled gradient projection method is a feasible direction method.*

*Proof.* Consider the case when $x^k$ is nonstationary, $x^{k+1} \neq x^k$, $d^k = \bar{x}^k - x^k \neq 0$, $\nabla f(x^k) \neq 0$.

When $\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k) \in X$, $\tilde{x}^k = \bar{x}^k$,

$$\left\langle \nabla f(x^k), d^k \right\rangle = \left\langle \nabla f(x^k), \tilde{x}^k - x^k \right\rangle = \left\langle \nabla f(x^k), -B_k^{-1}\nabla f(x^k) \right\rangle = -\|\nabla f(x^k)\|^2_{B_k^{-1}} < 0, \tag{5.134}$$

then $d^k$ is a feasible direction. Also $x^k + \alpha^k d^k = x^k + \alpha^k(\tilde{x}^k - x^k) \in X$, since $X$ is convex. Then, there exists an update $x^{k+1} \in X$.

When $\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k) \notin X$, by Proposition 5.18

$$\left\langle \tilde{x}^k - \bar{x}^k, x^k - \bar{x}^k \right\rangle_{B_k} \leq 0, \tag{5.135}$$

$$\left\langle x^k - B_k^{-1}\nabla f(x^k) - \bar{x}^k, x^k - \bar{x}^k \right\rangle_{B_k} \leq 0, \tag{5.136}$$

$$\left\langle \nabla f(x^k), d^k \right\rangle = \left\langle B_k^{-1}\nabla f(x^k), d^k \right\rangle_{B_k} \leq -\|\bar{x}^k - x^k\|^2_{B_k} < 0. \tag{5.137}$$

The last inequality holds since $d^k \neq 0$, that means $d^k$ is a feasible direction. Also $x^k + \alpha^k d^k = x^k + \alpha^k(\bar{x}^k - x^k) \in X$ since $\bar{x}^k \in X$ and $X$ is convex. Then, there exists an update $x^{k+1} \in X$. □

### 5.4.1 Proposed algorithm

Next, we discuss the scaled gradient projection method with inexact projection. Rewrite the constrained optimization problem (5.66),

$$\min_x f(x), \quad \text{such that } x \in X. \tag{5.138}$$

By the same method in the gradient projection method, we construct an expanding set sequence $\{X^k\}$ by Definition 5.29 satisfying

$$X = \lim_{k\to\infty} X^k, \quad X^k \subset X^{k+1}, \quad X^k \neq X^{k+1}, \tag{5.139}$$

where $X^k$ is nonempty, closed, and convex for each $k \in \mathbb{N}$.

As with the gradient projection method with inexact projection, we make an assumption for the projection algorithm.

**Assumption 5.39.** *Starting with $z^0 = \tilde{x}^k$, the projection function $P_{B_k,X^k}(\tilde{x}^k)$ generates a series of $\{z^j\}$ which converges to $P_{B_k,X^k}(\tilde{x}^k)$ strongly.*

Under above assumption, we define an inexact projection function which projects $\tilde{x}^k$ towards $X^k$ in the metric of $\mathcal{H}_{B_k}$, denoted as $\bar{P}_{B_k,X^k}(\tilde{x}^k) = z^{j_0}$. Here $j_0$ is the index of sequence $\{z^j\}$ in Assumption 5.39 such that

$$z^{j_0} \in X^{k+1}, \tag{5.140}$$

$$\left\langle \tilde{x}^k - z^{j_0}, x^k - z^{j_0} \right\rangle_{B_k} \leq 0. \tag{5.141}$$

**Proposition 5.40.** *When $x^k$ is a nonstationary point, under Assumption 5.39, there exists an index $j_0$ such that $z^{j_0}$ in the converging sequence $\{z^j\}$ such that equation* (5.140) *and* (5.141) *are satisfied.*

*Proof.* Similar to Proposition 5.32. □

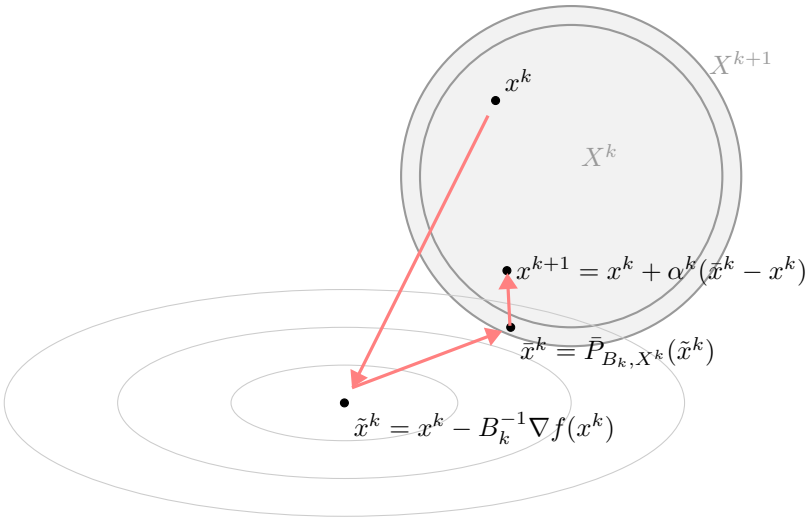


Figure 5.6: The scaled gradient projection method with inexact projection.

Similar to the scaled gradient projection method, at the $k$-th iteration, compute

$$\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k), \tag{5.142}$$

$$\bar{x}^k = \bar{P}_{B_k,X^k}(\tilde{x}^k), \tag{5.143}$$

$$x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k). \tag{5.144}$$

Figure 5.6 illustrates the above process. The scaled gradient projection method with inexact projection can be described as Algorithm 6.

**Algorithm 6:** Scaled gradient projection method with inexact projection

Initialization: Given feasible set $X$, construct set sequence $\{X^k\}$ satisfies (5.139). Given the initial point $x^0 \in X^0$.

**while** *not convergent* **do**

Step 1: Find a scaling matrix $B_k$, compute $\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k)$ ;

Step 2: Projecting $\tilde{x}^k$ towards $X^k$, until equation (5.140) and (5.141) are satisfied, denote $\bar{x}^k = \bar{P}_{B_k,X^k}(\tilde{x}^k)$;

Step 3: Evaluate the line search stepsize with Armijo rule or Wolfe conditions, update with $x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k)$;

Step 4: Enlarge the feasible set $X^k = X^{k+1}$ with set expanding strategy, let $k = k + 1$.

**end**

### 5.4.2 Convergence analysis

**Proposition 5.41.** *Algorithm 6 is a feasible direction method.*

*Proof.* With the condition (5.141), the proof is exactly the same as Proposition 5.38. □

**Assumption 5.42.** *The inexact projection function project $\tilde{x}^k$ close enough to its exact projection $P_{B_k,X^k}(\tilde{x}^k)$ with*

$$\|\bar{P}_{B_k,X^k}(\tilde{x}^k) - P_{B_k,X^k}(\tilde{x}^k)\| \leq d(X^k, X^{k+1}), \tag{5.145}$$

*here the $d(X^k, X^{k+1})$ is the (integrated) set distance between $X^k$ and $X^{k+1}$.*

The above assumption make sense since $d(X^k, X^{k+1}) > 0$ as $X^k$ and $X^{k+1}$ are closed sets, $X^k \subset X^{k+1}$, $X^k \neq X^{k+1}$, and by Assumption 5.39, the projection generating a sequence $\{z^j\}$ converges to $P_{B_k,X^k}(\tilde{x}^k)$ strongly.

**Theorem 5.43.** *Under Assumption 5.36, 5.39, and 5.42, let $\{x^k\}$ be the sequence generated by Algorithm 6. Then every limit point of $\{x^k\}$ is stationary.*

*Proof.* Suppose there is a subsequence $\{x^k\}_{k\in\mathcal{K}}$ converges to a nonstationary point $x_0$. By Proposition 5.28, it is sufficient to show the update direction sequence $d^k = \bar{x}^k - x^k$ as $k \in \mathcal{K}$ is gradient related, i.e.

$$\limsup_{k\to\infty, k\in\mathcal{K}} \|\bar{x}^k - x^k\| < \infty, \tag{5.146}$$

$$\limsup_{k\to\infty, k\in\mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle < 0. \tag{5.147}$$

Denote $\tilde{x}^k = x^k - B_k^{-1}\nabla f(x^k)$, $\tilde{x} = x_0 - B^{-1}\nabla f(x_0)$. By Assumption 5.36, $B_k \to B$ as $k \to \infty$ The objection function $f$ is smooth, $\nabla f(x^k) \to \nabla f(x_0)$ and $\tilde{x}^k \to \tilde{x}$ as $k \to \infty$. We can assume that there is a compact subset $\Omega \subset \mathbb{R}^n$ such that $\{\tilde{x}^k\} \subset \Omega$.

Recall that the inexact projection of $\tilde{x}^k$ is $\bar{x}^k = \bar{P}_{B_k,X^k}(\tilde{x}^k)$, consider the inequality,

$$\begin{aligned}\|\bar{P}_{B_k,X^k}(\tilde{x}^k) - P_{B,X}(\tilde{x})\| \leq{}& \|\bar{P}_{B_k,X^k}(\tilde{x}^k) - P_{B_k,X^k}(\tilde{x}^k)\| + \|P_{B_k,X^k}(\tilde{x}^k) - P_{B,X^k}(\tilde{x}^k)\| \\ &+ \|P_{B,X^k}(\tilde{x}^k) - P_{B,X}(\tilde{x}^k)\| + \|P_{B,X}(\tilde{x}^k) - P_{B,X}(x^k)\|.\end{aligned} \tag{5.148}$$

By Proposition 5.18, $P_{B,X}$ is continuous, for all $\varepsilon/4 > 0$, there exists $N_1 \in \mathcal{N}_\infty$, such that $\|P_{B,X}(\tilde{x}^k) - P_{B,X}(\tilde{x})\| < \varepsilon/4$.

Since $X^k \to X$ as $k \to \infty$, by Proposition 5.21, the scaled projection function sequence $P_{B,X^k}$ converges uniformly to $P_{B,X}$ on all compact subsets of $\mathbb{R}^n$. The uniform convergence holds on the set $\Omega$. For the above

$\varepsilon/4 > 0$, for all $\rho_1 > 0$, there exists $N_2 \in \mathcal{N}_\infty$ such that

$$P_{B,X^k}(x) \cap \rho_1 \mathbb{B} \subset P_{B,X}(x) + \varepsilon/4\mathbb{B}, \tag{5.149}$$

$$P_{B,X}(x) \cap \rho_1 \mathbb{B} \subset P_{B,X^k}(x) + \varepsilon/4\mathbb{B}, \tag{5.150}$$

for all $x \in \Omega$, when $k \in N_2$. Then we have $\|P_{B,X^k}(\tilde{x}^k) - P_{B,X}(\tilde{x}^k)\| \leq \varepsilon/4$.

By Assumption 5.36, $B_k \to B$ as $k \to \infty$. Then the projection function sequence $\{P_{B_k,X^k}\}$ converges uniformly to $P_{B,X^k}$ on all compact subsets of $\mathbb{R}^n$ by Proposition 5.22. The uniform convergence holds on the set $\Omega$. For the above $\varepsilon/4 > 0$, for all $\rho_2 > 0$, there exists $N_3 \in \mathcal{N}_\infty$ such that

$$P_{B_k,X^k}(x) \cap \rho_2 \mathbb{B} \subset P_{B,X^k}(x) + \varepsilon/4\mathbb{B}, \tag{5.151}$$

$$P_{B,X^k}(x) \cap \rho_2 \mathbb{B} \subset P_{B_k,X^k}(x) + \varepsilon/4\mathbb{B}, \tag{5.152}$$

for all $x \in \Omega$, when $k \in N_3$. Then we have $\|P_{B_k,X^k}(\tilde{x}^k) - P_{B,X^k}(\tilde{x}^k)\| \leq \varepsilon/4$.

Since $X^k \to X$ and $X^k$, $X$ are nonempty closed bounded, by Theorem 5.3, $d(X^k, X) \to 0$. By Assumption 5.42, for $\varepsilon/8$, there exists $N_4 \in \mathcal{N}_\infty$ such that when $k, k+1 \in N_4$,

$$\|\bar{P}_{B_k,X^k}(\tilde{x}^k) - P_{B_k,X^k}(\tilde{x}^k)\| \leq d(X^k, X^{k+1}) \leq d(X^k, X) + d(X, X^{k+1}) < \varepsilon/4. \tag{5.153}$$

Then for the above arbitrary $\varepsilon > 0$, there exists $N = N_1 \cap N_2 \cap N_3 \cap N_4$ such that $\|\bar{P}_{B_k,X^k}(\tilde{x}^k) - P_{B,X}(\tilde{x})\| < \varepsilon$ as $k \in N$. Which means $\bar{P}_{B_k,X^k}(\tilde{x}^k) \to P_{B,X}(\tilde{x})$ as $k \to \infty$ and $k \in \mathcal{K}$. Then

$$\|\bar{x}^k - x^k\| = \|\bar{P}_{B_k,X^k}(\tilde{x}^k) - x^k\| \to \|P_{B,X}(\tilde{x}) - x_0\|, \quad k \to \infty, k \in \mathcal{K}. \tag{5.154}$$

Since $P_{B,X}(\tilde{x}) \in X$ and $x_0 \in X$, then $\limsup_{k\to\infty, k\in\mathcal{K}} \|\bar{x}^k - x^k\| < \infty$.

By inequality (5.141) and Assumption 5.36,

$$\left\langle \tilde{x}^k - \bar{x}^k, x^k - \bar{x}^k \right\rangle_{B_k} \leq 0, \tag{5.155}$$

$$\left\langle x^k - B_k^{-1}\nabla f(x^k) - \bar{x}^k, x^k - \bar{x}^k \right\rangle_{B_k} \leq 0, \tag{5.156}$$

$$\left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle = \left\langle B_k^{-1}\nabla f(x^k), \bar{x}^k - x^k \right\rangle_{B_k} \leq \|\bar{x}^k - x^k\|^2. \tag{5.157}$$

Taking the limit of the above equation, since $\bar{x}^k = \bar{P}_{B_k,X^k}(\tilde{x}^k)$ and $\bar{P}_{B_k,X^k}(\tilde{x}^k) \to P_{B,X}(\tilde{x})$,

$$\limsup_{k\to\infty,k\in\mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle \leq \|P_{B,X}(\tilde{x}) - x_0\|^2. \tag{5.158}$$

Assume $P_{B,X}(\tilde{x}) = x_0$, by Proposition 5.18,

$$\langle \tilde{x} - P_{B,X}(\tilde{x}), x - P_{B,X}(\tilde{x}) \rangle_B \leq 0, \quad \forall x \in X, \tag{5.159}$$

$$\left\langle x_0 - B^{-1}\nabla f(x_0) - P_{B,X}(\tilde{x}), x - P_{B,X}(\tilde{x}) \right\rangle_B \leq 0, \quad \forall x \in X, \tag{5.160}$$

$$\left\langle B^{-1}\nabla f(x_0), x - x_0 \right\rangle_B \geq 0, \quad \forall x \in X, \tag{5.161}$$

$$\langle \nabla f(x_0), x - x_0 \rangle \geq 0, \quad \forall x \in X, \tag{5.162}$$

that means $x_0$ is a stationary point, contradiction. Then,

$$\limsup_{k\to\infty,k\in\mathcal{K}} \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle < 0. \tag{5.163}$$

□

### 5.4.3 Discussion on scaling matrix

One of the most popular Hessian approximations is the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method. At the $k$-th iteration, denote the L-BFGS Hessian matrix by $B_k$ and let $H_k = B_k^{-1}$. The information of $x^k$ and $\nabla f(x^k)$ are stored for a small number $m$ with,

$$s_k = x^{k+1} - x^k, \quad y_k = \nabla f(x^{k+1}) - \nabla f(x^k), \tag{5.164}$$

$$\rho_k = 1/y_k^{'} s_k, \quad V_k = I - \rho_k y_k s_k^{'}. \tag{5.165}$$

Then the L-BFGS Hessian inverse matrix can be written as:

$$\begin{aligned} H_k = &\left(V_{k-1}^{'} \cdots V_{k-m}^{'}\right) H_k^{(0)} \left(V_{k-m} \cdots V_{k-1}\right) \\ &+ \rho_{k-m} \left(V_{k-1}^{'} \cdots V_{k-m+1}^{'}\right) s_{k-m} s_{k-m}^{'} \left(V_{k-m+1} \cdots V_{k-1}\right) \\ &+ \rho_{k-m+1} \left(V_{k-1}^{'} \cdots V_{k-m+2}^{'}\right) s_{k-m+1} s_{k-m+1}^{'} \left(V_{k-m+2} \cdots V_{k-1}\right) \\ &\vdots \\ &+ \rho_{k-1} s_{k-1} s_{k-1}^{'}. \end{aligned} \tag{5.166}$$

Both $B_k$ and $H_k$ are needed in our work. To efficiently store and evaluate matrix multiplication of $B_k$ and $H_k$, a compact form is proposed in [27],

$$S_k = [s_{k-m}, \cdots, s_{k-1}], \quad Y_k = [y_{k-m}, \cdots, y_{k-1}], \tag{5.167}$$

$$(R_k)_{i,j} = \begin{cases} (s_{k-m-1+i})^{'}(y_{k-m-1+j}), & \text{if } i \leq j, \\ 0, & \text{otherwise,} \end{cases} \tag{5.168}$$

$$D_k = \text{diag}(s_{k-m}^{'} y_{k-m}, \cdots, s_{k-1}^{'} y_{k-1}), \tag{5.169}$$

$$(U_k)_{i,j} = \begin{cases} (s_{k-m-1+i})^{'}(y_{k-m-1+j}), & \text{if } i > j, \\ 0, & \text{otherwise.} \end{cases} \tag{5.170}$$

We write the $H_k$ and $B_k$ in a compact form.

$$B_k = \sigma_k I - \begin{bmatrix} \sigma_k S_k & Y_k \end{bmatrix} \begin{bmatrix} \sigma_k S_k^{'} S_k & U_k \\ U_k^{'} & -D_k \end{bmatrix}^{-1} \begin{bmatrix} \sigma_k S_k^{'} \\ Y_k^{'} \end{bmatrix}, \tag{5.171}$$

$$H_k = \gamma_k I + \begin{bmatrix} S_k & \gamma_k Y_k \end{bmatrix} \begin{bmatrix} (R_k^{'})^{-1} \left( D_k + \gamma_k Y_k^{'} Y_k \right) R_k^{-1} & -(R_k^{'})^{-1} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^{'} \\ \gamma_k Y_k^{'} \end{bmatrix}, \tag{5.172}$$

where $\gamma_k = y_{k-1}^{'} s_{k-1} / y_{k-1}^{'} y_{k-1}$, $\sigma_k = y_{k-1}^{'} s_{k-1} / s_{k-1}^{'} s_{k-1}$.

**Remark 5.44** ([99])**.** *Under Wolfe conditions, both $B_k$ and $H_k$ remain positive definite.*

The above remark guarantees that the L-BFGS Hessian approximation can be used in practice. However, to the author's best knowledge, there are no results to show that the L-BFGS Hessian approximation satisfies Assumption 5.36. Relevant results can be found in [26, 84], which show that the Assumption 5.36 is not satisfied for the BFGS Hessian approximation.

# Chapter 6

# Scaled gradient projection method with multiple constraints

In this chapter, we introduce the scaled gradient projection method with multiple constraints, then apply the proposed method to the seismic inverse problem. First, we fix the notations. Let space $\mathcal{H}$ be the regular Euclidean space with element in $\mathbb{R}^n$, with the inner product $\langle x, y\rangle = x^{'}y$ and the norm $\|x\| = \sqrt{x^{'}x}$. Given a symmetric positive definite matrix $B$, let the space $\mathcal{H}_B$ be the scaled Euclidean space with the elements in $\mathbb{R}^n$, with inner product $\langle x, y\rangle_B = \langle Bx, y\rangle$ and norm $\|x\|_B = \sqrt{\langle Bx, y\rangle}$. When $u \in \mathbb{R}^n$ represents a digital image, it can be considered as a two dimensional discrete matrix with $N_x$ rows and $N_y$ columns with $n = N_x \times N_y$ pixels in total. In this case, $u$ can be also dealt with as a vector in $\mathbb{R}^n$. The above two representations are equivalent with $u_{(i-1)N_x+j} = u_{i,j}$, for $i = 1, \cdots, N_x$ and $j = 1, \cdots, N_y$.

Consider the constrained optimization problem:

$$\min_x f(x), \quad \text{such that } x \in \cap_i X_i, \tag{6.1}$$

where $X_i$ is nonempty, closed, and convex for each $i$. The feasible set for the above optimization problem is the intersection of $X_i$, which can describe the a priori information of the optimization problem. When $f(x)$ is nonconvex, the uniqueness properties of the optimization problem is not available. An equivalent problem can be written as:

$$\min_x f(x) + \sum_i \lambda_i \mathcal{R}_i(x), \tag{6.2}$$

where $\mathcal{R}_i(x)$ is a regularization term and $\lambda_i$ is the regularization parameter. Compared to problem (6.2):

1. Problem (6.1) is flexible when different types of constraints are considered. The gradient and Hessian evaluation of the objective is unchanged when the constraints are different.

2. The constraint set $X_i$ provides a direct description of the solution, and the radius or shape of $X_i$ can be easily estimated. On the other hand, the regularization parameter $\lambda_i$ provides an indirect description of the problem, and the choice of $\lambda_i$ might depends on the experience or multiple experiments.

We focus on the constrained optimization problem (6.1) to incorporate multiple a priori information by constraint sets. The optimization scheme proposed in this chapter is a combination of the scaled gradient projection method, the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) Hessian approximation, and the projection onto the intersection of convex sets algorithm developed in [44, 45]. The set expanding strategy proposed in the previous chapter is used because the projection algorithm can only be evaluated inexactly.

First, we discuss how to describe the a priori information as convex constraint sets. The constraint sets with closed-form projection function are discussed in Section 6.1, and the constraint sets with subgradient projection function are discussed in Section 6.2. In Section 6.3, we discuss the algorithms of projecting onto the intersection of the above convex sets based on the work [44, 45]. The proposed algorithm is provided in Section 6.4. The formulation of the proposed algorithm for the full waveform inversion (FWI) problem is provided in Section 6.5. Furthermore, numerical examples of both cross-well and reflective waves are provided at the end of the chapter.

## 6.1 Convex constraint sets with closed-form projection function

We discuss a kind of convex constraint sets named "simple set" which has the closed-form projection function. In this case, the projection function can be evaluated efficiently and accurately. First, we discuss several convex constraint sets, and each of the sets can describe the inverse problem with some kinds of a priori information. Next, the projection functions of the above sets are given. In the end, we show how to build the expanding sequence of convex constraint sets.

One of the most commonly used constraint sets in the constrained optimization problem the is box constraint which provides the lower and upper bounds of the parameter. Given $a, b \in \mathbb{R}$ with $a \leq b$, the box constraint set is given as:

$$X_{\text{box}} = \{u \in \mathbb{R}^n \mid a \leq u_i \leq b,\ i = 1, \cdots, n\}. \tag{6.3}$$

For vector $a, b \in \mathbb{R}^n$ with $a_i \leq b_i$ for index $i = 1, \cdots, n$, a more sophisticated box constraint set is given as

$$\bar{X}_{\text{box}} = \{u \in \mathbb{R}^n \mid a_i \leq u_i \leq b_i,\ i = 1, \cdots, n\}. \tag{6.4}$$

In this work we will focus on the box constraint in the simple form (6.3) to demonstrate the algorithm. Given $p \in \mathbb{R}^n$ and $\kappa \in \mathbb{R}$, the affine hyperplane is

$$X_{\text{plane}} = \{u \in \mathbb{R}^n \mid \langle u, p \rangle = \kappa\}. \tag{6.5}$$

Since we are working on the two-dimensional images with all pixels are non-negative in this work, we can set $p_i \geq 0$ for each index $i = 1, \cdots, n$, and set $\kappa > 0$. In this case, it is reasonable to use the affine hyperplane to represent the (weighted) average value of a certain area. In the same way, given $\eta, \kappa \in \mathbb{R}$ with $\eta < \kappa$, $p \in \mathbb{R}^n$ with $p_i \geq 0$, for each index $i = 1, \cdots, n$, we can have the affine hyperslab

$$X_{\text{slab}} = \{u \in \mathbb{R}^n \mid \eta \leq \langle u, p \rangle \leq \kappa\}. \tag{6.6}$$

Notice, when $\eta = -\infty$ or $\kappa = \infty$, the hyperslab is a half-space, which is a subset of $\mathbb{R}^n$ and divided by a hyperplane. Sometimes we know the exact value of certain pixels in the image, and denote the index of these values as an index set $I_{\text{subspace}}$. That is, for $i \in I_{\text{subspace}}$, we a priori know $u_i = a_i$. In this case, denote the constraint set as a subspace

$$X_{\text{subspace}} = \{u \in \mathbb{R}^n \mid u_i = a_i,\ i \in I_{\text{subspace}}\}. \tag{6.7}$$

Given $u_0 \in \mathbb{R}^n$ and $r > 0$, the $l_2$ ball with center $u_0$ and radius $r$ is

$$X_{l_2} = \{u \in \mathbb{R}^n \mid \|u - u_0\| \leq r\}. \tag{6.8}$$

**Proposition 6.1.** *All the above subsets are convex.*

Next, we give the closed-form projection function for the above convex constraint sets. The projection

function of the box constraint set (6.3) is given by

$$P_{\text{box}}(u)_i = \begin{cases} a, & \text{if } u_i < a, \\ u_i, & \text{if } a \le u_i \le b, \\ b, & \text{if } b < u_i, \end{cases} \quad \text{or } P_{\text{box}}(u)_i = \max(a, \min(u_i, b)). \tag{6.9}$$

The projection function of the affine hyperplane (6.5) is given by

$$P_{\text{plane}}(u) = u + \frac{\kappa - \langle u, p \rangle}{\|p\|^2} p. \tag{6.10}$$

The projection function of the affine hyperslab (6.6) is given by

$$P_{\text{slab}}(u) = \begin{cases} u + \frac{\kappa - \langle u,p \rangle}{\|p\|^2} p, & \text{if } \langle u, p \rangle > \kappa, \\ u, & \text{if } \eta \le \langle u, p \rangle \le \kappa, \\ u + \frac{\eta - \langle u,p \rangle}{\|p\|^2} p, & \text{if } \langle u, p \rangle < \eta. \end{cases} \tag{6.11}$$

The projection function of subspace (6.7) is given by

$$P_{\text{subspace}}(u)_i = \begin{cases} a_i, & \text{if } i \in I_{\text{subspace}}, \\ u_i, & \text{otherwise.} \end{cases} \tag{6.12}$$

The projection function of $l_2$ ball (6.8) is given by

$$P_{l_2}(u) = \begin{cases} u_0 + r \frac{u - u_0}{\|u - u_0\|}, & \text{if } \|u - u_0\| > r, \\ u, & \text{otherwise.} \end{cases} \tag{6.13}$$

A different notation $U$ is used in the optimization problem to represent the feasible set for the PDE constrained optimization problem. An expanding set sequence is needed for the proposed method, which satisfies

$$U_{\text{ad}} = \lim_{k \to \infty} U^k, \quad U^k \subset U^{k+1}, \quad U^k \neq U^{k+1}, \tag{6.14}$$

where $U^k$ is nonempty, closed, and convex. An adaptive strategy can be considered: if $u^k \in U^k$, we set $U^{k+1} = U^k$. In this case, the constraint set $U^k$ does not expand at the $k+1$ iteration. We expand the

constraint set $U^{k+1}$ such that $U^k \subset U^{k+1}$ otherwise. The adaptive strategy will be discussed in the following sections. To represent this process in a more convenient way, another index $h$ is used to describe the set expanding process.

Next, we show how to build the expanding set sequence with the convex constraint sets we showed above. Let $\varepsilon > 0$ be a threshold and $\eta \in (0, 1)$ is a parameter to control the sequence of sets expanding speed. We need the sequence is expanding as iteration goes on, but not expand to infinity large. In this case, define a threshold function as

$$\theta(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} \eta^i \varepsilon, & \text{if } h \geq 1, \end{cases} \tag{6.15}$$

and

$$\lim_{h \to \infty} \theta(h) = \frac{\eta}{1-\eta}\varepsilon, \tag{6.16}$$

here $h \in \mathbb{N}$ is the index controlling the set expanding.

First, the box constraint set (6.3), given $a, b \in \mathbb{R}$ and $a \leq b$, define

$$U_{\text{box}}^h = \{u \in \mathbb{R}^n \mid a - \theta(h) \leq u_i \leq b + \theta(h),\ i = 1, \cdots, n\}, \quad h \in \mathbb{N}. \tag{6.17}$$

In this case, we have

$$U_{\text{box}}^0 = X_{\text{box}}, \tag{6.18}$$

$$U_{\text{box}} = \lim_{h \to \infty} U_{\text{box}}^h = \left\{ u \in \mathbb{R}^n \;\middle|\; a - \frac{\eta}{1-\eta}\varepsilon \leq u_i \leq b + \frac{\eta}{1-\eta}\varepsilon,\ i = 1, \cdots, n \right\}. \tag{6.19}$$

For the affine hyperplane subset (6.5), define the sequence of hyperplane sets as

$$U_{\text{plane}}^h = \{u \in \mathbb{R}^n \mid \|u - P_{\text{plane}}(u)\| \leq \theta(h)\}, \quad h \in \mathbb{N}, \tag{6.20}$$

then,

$$U_{\text{plane}}^0 = X_{\text{plane}}, \tag{6.21}$$

$$U_{\text{plane}} = \lim_{h \to \infty} U_{\text{plane}}^h = \left\{ u \in \mathbb{R}^n \;\middle|\; \|u - P_{\text{plane}}(u)\| \leq \frac{\eta}{1-\eta}\varepsilon \right\}. \tag{6.22}$$

For the affine hyperslab subset (6.6), define the sequence of hyperslab sets as

$$U_{\text{slab}}^{h} = \{u \in \mathbb{R}^n \mid \eta - \theta(h) \leq \langle u, p \rangle \leq \kappa + \theta(h)\}, \quad h \in \mathbb{N}, \tag{6.23}$$

then,

$$U_{\text{slab}}^{0} = X_{\text{slab}}, \tag{6.24}$$

$$U_{\text{slab}} = \lim_{h \to \infty} U_{\text{slab}}^{h} = \left\{ u \in \mathbb{R}^n \;\middle|\; \eta - \frac{\eta}{1-\eta}\varepsilon \leq \langle u, p \rangle \leq \kappa + \frac{\eta}{1-\eta}\varepsilon \right\}. \tag{6.25}$$

For the subspace (6.7), define the sequence of subspaces as

$$U_{\text{subspace}}^{h} = \{u \in \mathbb{R}^n \mid \|u - P_{\text{subspace}}(u)\| \leq \theta(h)\}, \quad h \in \mathbb{N}, \tag{6.26}$$

then,

$$U_{\text{subspace}}^{0} = X_{\text{subspace}}, \tag{6.27}$$

$$U_{\text{subspace}} = \lim_{h \to \infty} U_{\text{subspace}}^{h} = \left\{ u \in \mathbb{R}^n \;\middle|\; \|u - P_{\text{subspace}}(u)\| \leq \frac{\eta}{1-\eta}\varepsilon \right\}. \tag{6.28}$$

For the $l_2$ ball, define the sequence of $l_2$ balls as

$$U_{l_2}^{h} = \{u \in \mathbb{R}^n | \|u - u_0\| \leq r + \theta(h)\}, \quad h \in \mathbb{N}, \tag{6.29}$$

then,

$$U_{l_2}^{0} = X_{l_2}, \tag{6.30}$$

$$U_{l_2} = \lim_{h \to \infty} U_{l_2}^{h} = \left\{ u \in \mathbb{R}^n \;\middle|\; \|u - u_0\| \leq r + \frac{\eta}{1-\eta}\varepsilon \right\}. \tag{6.31}$$

**Proposition 6.2.** *Every element in the above set sequences is closed and convex.*

## 6.2 Convex constraint sets with subgradient projection

### 6.2.1 Subgradient projection

In this section, we focus on the total variation (TV) constraint and $l_1$ ball constraint. Through the concept of lower level set, the TV function and $l_1$ function can be used to describe the convex constraint sets. The

closed-form projection function is not available for the TV constraint. To evaluate the projection efficiently, the subgradient projection method is introduced.

**Definition 6.3** (Lower level set)**.** *Given a continuous convex function $f : \mathbb{R}^n \to \mathbb{R}$, the lower level set function of $f$ with a height $\tau \in \mathbb{R}$ is given by*

$$\mathrm{lev}_{\leq \tau} f = \{x \in \mathbb{R}^n \mid f(x) \leq \tau\}. \tag{6.32}$$

**Proposition 6.4.** *The lower level set* (6.32) *is convex and closed.*

Before the discussion of the subgradient projection, we review the definition of subgradient and subdifferential first. For more information, we refer to [115].

**Definition 6.5** (Subgradient)**.** *Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $x^* \in \mathbb{R}^n$ is said to be a subgradient of $f$ at the point $x$ if*

$$f(z) \geq f(x) + \langle x^*, z - x \rangle, \quad \forall z \in \mathbb{R}^n. \tag{6.33}$$

*The equation* (6.33) *is also refer to as the subgradient inequality.*

*The set of all subgradients of $f$ at $x$ is called the subdifferential of $f$ at $x$, and denoted as $\partial f(x)$. The subdifferential as a set-valued mapping and $\partial f(x)$ is a closed convex set. In general, $\partial f(x)$ may be empty or it may consist of just one vector. If $\partial f(x)$ is not empty, $f$ is said to be subdifferentiable at $x$.*

The next theorem provides the relation between subdifferentiability and differentiability.

**Theorem 6.6** ([115] Theorem 25.1)**.** *Let $f$ be a convex function, and let $x$ be a point where $f$ is finite. If $f$ is differentiable at $x$, then $\nabla f(x)$ is the unique subgradient of $f$ at $x$, so that in particular*

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle, \quad \forall z \in \mathbb{R}^n. \tag{6.34}$$

*Conversely, if $f$ has a unique subgradient at $x$, then $f$ is differentiable at $x$.*

Denote $C$ as a nonempty, closed, and convex subset in $\mathbb{R}^n$ and $x$ is a vector in $\mathbb{R}^n$. The projection mapping of $x$ onto set $C$ is denoted as $P_C(x)$ in the previous chapter. When $C$ is nonempty, closed, and convex, the $P_C(x)$ is a single-valued mapping. Also, $P_C(x)$ can be described through the following inequality

$$\langle x - P_C(x), z - P_C(x) \rangle \leq 0, \quad \forall z \in C. \tag{6.35}$$

The projection of $x$ onto $C$ is equivalent to solving the following problem

$$\min_z \|z - x\|^2, \quad \text{such that } z \in C. \tag{6.36}$$

This minimization problem usually leads to some iterative structure. For the convex subsets discussed in the previous section, the projection of $x$ onto $C$ can be evaluated through a projection function in closed form which is very efficient in most cases. But for some of the convex constraint sets, the closed-form projection function is not always available. For the case when $C$ is a lower level set of a continuous convex function $f$ and height $\tau$, i.e.

$$C = \text{lev}_{\leq \tau} f = \{x \in \mathbb{R}^n \mid f(x) \leq \tau\}, \tag{6.37}$$

the subgradient projection is an economical way to approximate the projection.

**Proposition 6.7.** *Given a continuous convex function $f : \mathbb{R}^n \to \mathbb{R}$, vector $x \in \mathbb{R}^n$ and $x \notin C$, and the subgradient $x^* \in \partial f(x)$, the lower level set $C$ is defined by equation* (6.37)*. The half-space set*

$$H_x = \{z \in \mathbb{R}^n \mid f(x) + \langle x^*, z - x \rangle \leq \tau\}. \tag{6.38}$$

*is an outer approximation of set $C$. Also, we have $x \notin H_x$.*

*Proof.* Suppose $x_0 \in C$, since $C$ is a lower level set with level set function $f$ and height $r$, $f(x_0) \leq \tau$. Then by the subgradient inequality (6.33)

$$f(x) + \langle x^*, x_0 - x \rangle \leq f(x_0) \leq \tau, \tag{6.39}$$

which means $x_0 \in H_x$. On the other hand, suppose $x_0 \notin H_x$,

$$f(x) + \langle x^*, x_0 - x \rangle > \tau, \tag{6.40}$$

and by subgradient inequality (6.33), we have $f(x_0) > \tau$, which means $x_0 \notin C$.

Since

$$f(x) + \langle x^*, x - x \rangle = f(x) > \tau, \tag{6.41}$$

then $x \notin H_x$. □

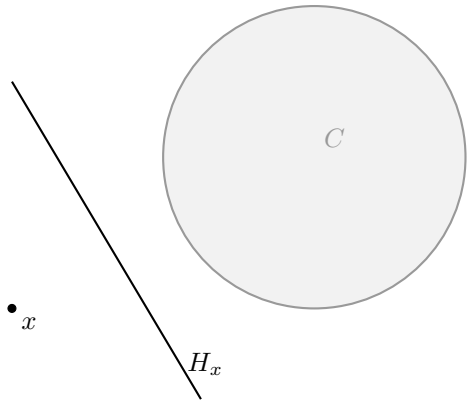


Figure 6.1: Approximation of $C$ with half-space $H_x$.

The outer approximation is demonstrated in Figure 6.1

As discussed in the previous section, the projection function onto half-space has a closed-form function

$$P_{H_x}(z) = \begin{cases} z + \frac{\tau - f(x) - \langle z - x, x^* \rangle}{\|x^*\|^2} x^*, & \text{if } z \notin H_x, \\ z, & \text{if } z \in H_x. \end{cases} \tag{6.42}$$

By Proposition 6.7, if $x \notin C$ then $x \notin H_x$. In this case, $f(x) > \tau$, we can always have

$$P_{H_x}(x) = x + \frac{\tau - f(x)}{\|x^*\|^2} x^*. \tag{6.43}$$

Next, we give the definition of subgradient projection

**Definition 6.8** (Subgradient projection)**.** *Given a continuous convex function $f : \mathbb{R}^n \to \mathbb{R}$, vector $x \in \mathbb{R}^n$, and the subgradient $x^* \in \partial f(x)$, the lower level set $C$ is defined by equation* (6.37)*. The subgradient projection function which project $x$ towards $C$ is given by*

$$\tilde{P}_C(x) = \begin{cases} x + \frac{\tau - f(x)}{\|x^*\|^2} x^*, & \textit{if } f(x) > \tau, \\ x, & \textit{if } f(x) \leq \tau. \end{cases} \tag{6.44}$$

### 6.2.2 Total variation constraint

Since first introduced in the image denoising problem by Rudin, Osher, and Fatemi [118], the total variation regularization technique has been widely applied in image processing areas such as restoration [117], segmentation [34], optical flow [24]. The total variation regularization technique is not only restricted in the image processing area but also has been successfully applied in a wide variety of inverse problems such

as computational tomography [24], magnetic resonance imaging [76], and electrical impedance tomography [40, 63]. The total variation techniques are especially suitable for image processing and imaging problem with sharp boundaries and block structures. For more application information, we refer to [35]. The total variation technique are also applied successfully in the full waveform inversion problem [66, 67, 89, 4], it provides efficient results on reducing cycle-skipping issue [55] and improve the inverse result for the salt body structure [140]. To incorporate with the total variation regularization, one of the most popular methods is to take the TV norm as a penalty term and take advantage of the dual structure of the ROF model [33]. Other methods like the primal-dual hybrid gradient method [145] are also been proposed.

Consider a two-dimensional digital image $u \in \mathbb{R}^{N_x \times N_y}$ with $N_x$ rows and $N_y$ columns, and $n = N_x \times N_y$. It is equivalent to consider $u$ as a vector in $\mathbb{R}^n$. We focus on the discrete version of the total variation norm in this work. For more analytic results of total variation and bounded variation spaces, we refer to [6]. Define the discrete gradient operator $D : \mathbb{R}^{N_x \times N_y} \to \mathbb{R}^{N_x \times N_y \times 2}$ with

$$(Du)_{i,j,1} = \begin{cases} u_{i+1,j} - u_{i,j}, & \text{if } 0 \le i < N_x, \\ 0, & \text{if } i = N_x, \end{cases} \tag{6.45}$$

$$(Du)_{i,j,2} = \begin{cases} u_{i,j+1} - u_{i,j}, & \text{if } 0 \le j < N_y, \\ 0, & \text{if } i = N_y. \end{cases} \tag{6.46}$$

Then the discrete total variation norm is given by the TV function $f_{\mathrm{TV}} : \mathbb{R}^n \to \mathbb{R}$

$$f_{\mathrm{TV}}(u) = \|u\|_{\mathrm{TV}} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left| (Du)_{i,j} \right|. \tag{6.47}$$

**Proposition 6.9.** *The above discrete total variation function $f_{TV} : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex.*

By the definition of lower level set, we can build a sequence of TV constraint sets with the TV function. Recall the threshold function in the previous section, given $\varepsilon > 0$, $\eta \in (0, 1)$, and an index $h$,

$$\theta(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} \eta^i \varepsilon, & \text{if } h \ge 1, \end{cases} \tag{6.48}$$

and

$$\lim_{h \to \infty} \theta(h) = \frac{\eta}{1 - \eta} \varepsilon, \tag{6.49}$$

Then given initial radius $\tau_{\mathrm{TV}}$, we can build a sequence of convex constraint sets as

$$U_{\mathrm{TV}}^{h} = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq \theta(h) + \tau_{\mathrm{TV}}\}. \tag{6.50}$$

In this case, we have

$$U_{\mathrm{TV}}^{0} = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq \tau_{\mathrm{TV}}\}, \tag{6.51}$$

$$U_{\mathrm{TV}} = \lim_{h\to\infty} U_{\mathrm{TV}}^{h} = \left\{u \in \mathbb{R}^n \;\middle|\; f_{\mathrm{TV}}(u) \leq \frac{\eta}{1-\eta}\varepsilon + \tau_{\mathrm{TV}}\right\}. \tag{6.52}$$

The subgradient projection function for the sequence of total variation constraint sets $U_{\mathrm{TV}}^{h}$ is

$$\tilde{P}_{U_{\mathrm{TV}}^{h}}(u) = \begin{cases} u + \frac{\theta(h)+\tau_{\mathrm{TV}}-f_{\mathrm{TV}}(u)}{\|u^*\|^2}u^*, & \text{if } f_{\mathrm{TV}}(u) > \theta(h) + \tau_{\mathrm{TV}}, \\ u, & \text{if } f_{\mathrm{TV}}(u) \leq \theta(h) + \tau_{\mathrm{TV}}. \end{cases} \tag{6.53}$$

Here $u^* \in \partial f_{\mathrm{TV}}(u)$ is a subgradient TV function $f_{\mathrm{TV}}$ at point $u$. The computation formula of subgradient $u^*$ is given by the following proposition.

**Proposition 6.10.** *Let $u \in \mathbb{R}^{N_x \times N_y}$ be a two dimensional image, the discrete total variation function $f_{TV}$ is given by* (6.47) *with the discrete gradient operator in* (6.45)*. Denote $e_{i,j}$ as the identity element of matrix in $\mathbb{R}^{N_x \times N_y}$. Given the function*

$$\begin{aligned} u^* = & \sum_{(i,j)\in I_1} \left((u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2\right)^{-1/2} \\ & \times \left((u_{i+1,j} - u_{i,j})\, e_{i+1,j} - (u_{i+1,j} - 2u_{i,j} + u_{i,j+1})\, e_{i,j} + (u_{i,j} - u_{i,j+1})\, e_{i,j+1}\right) \\ & + \sum_{(i,j)\in I_2} sgn\left(u_{N_x,j+1} - u_{N_x,j}\right)\left(e_{N_x,j+1} - e_{N_x,j}\right) \\ & + \sum_{(i,j)\in I_3} sgn\left(u_{i+1,N_y} - u_{i,N_y}\right)\left(e_{i+1,N_y} - e_{i,N_y}\right), \end{aligned} \tag{6.54}$$

*where the index sets are given by*

$$I_1 = \{(i,j) \mid u_{i,j} \neq u_{i+1,j} \text{ or } u_{i,j} \neq u_{i,j+1},\ 1 \leq i < N_x,\ 1 \leq j < N_y\}, \tag{6.55}$$

$$I_2 = \{(i,j) \mid u_{N_x,j} \neq u_{N_x,j+1},\ 1 \leq j < N_y\}, \tag{6.56}$$

$$I_3 = \left\{(i,j) \;\middle|\; u_{i,N_y} \neq u_{i+1,N_y},\ 1 \leq i < N_x\right\}. \tag{6.57}$$

*Then $u^* \in \partial f_{TV}(u)$, i.e. $u^*$ is a subgradient of $f_{TV}$ at $u$.*

*Proof.* By direct computation. □

### 6.2.3 Increase the sparsity with $l_1$ constraint

Consider $u \in \mathbb{R}^n$ is a signal or a digital image with $N_x \times N_y$ pixels, $n = N_x \times N_y$. Let $b \in \mathbb{R}^m$ be a measurement, given the measurement matrix $A \in \mathbb{R}^{m\times n}$ denote $b = Au$ as the encoding process and $u = A^{-1}b$ is the decoding matrix. Suppose $b$ is highly sparse, i.e. most of entries of $b$ are zeros. Then the decoding problem can be shown as

$$\bar{u} = \arg\min_{u\in\mathbb{R}^n} \|u\|_0, \quad \text{such that } Au = b, \tag{6.58}$$

here $\|u\|_0$ is the "$l_0$ norm" of $u$ represents the number of nonzero entries in $u$. The problem (6.58) is a combinatorial optimization problem and is impossible be solved when $n$ is large for most real applications. One way to overcome this issue is to replace the "$l_0$ norm" by $l_1$ norm as

$$\bar{u} = \arg\min_{u\in\mathbb{R}^n} \|u\|_1, \quad \text{such that } Au = b. \tag{6.59}$$

This approach has been widely applied in the geophysical area [120] and later was popularized with the name of basis pursuit by the work [37]. As $A$ is a random sampling matrix, the relation between problem (6.58) and (6.59) leads to the research area of compress sensing (CS). A rich theory has been developed for the CS area, for more analytic results we refer to the seminal work [29, 30, 31] and survey paper [28]. We only focus on the sparse structure of vector or image $u$, so we consider $A$ is the identity matrix $I$ in this work.

In most cases, a linear transformation is needed to achieve the sparsity structure of $u$. Denote the linear transformation as a matrix $\Phi \in \mathbb{R}^{m\times n}$, which can represents wavelet transform, curvelet transform, etc. Suppose we have the a priori information of the model as it has the sparsity, then we can work on the term $\|\Phi u\|_1 \leq \tau_{l_1}$. When $\Phi = I$, in this case we will focus on the $l_1$ fidelity term $\|u\|_1 \leq \tau_{l_1}$.

Next, we discuss how to describe the sparsity a priori information with $l_1$ ball as the convex constraint set. Given a matrix $\Phi \in \mathbb{R}^{m\times n}$, with $\Phi = [\phi_1, \cdots, \phi_m]'$, here each $\phi_i$, $i = 1, \cdots, m$ is a $n$-dimensional row vector represents some basis of $\mathbb{R}^n$. Here $\Phi$ represents a linear transformation that maps the signal $u$ to the coefficient space $\mathbb{R}^m$. Typical choices of $\Phi$ can be wavelet transform, or curvelet transform, etc. We define the $l_1$ function with linear operator $\Phi$ as

$$f_{l_1}(u) = \|\Phi u\|_1 = \sum_{i=1}^{m} |\langle \phi_i, u\rangle| = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} \Phi_{i,j} u_j \right|. \tag{6.60}$$

**Proposition 6.11.** *The above $l_1$ function $f_{l_1} : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex.*

By the definition of lower level set, we can build a sequence of $l_1$ constraint sets with the $l_1$ function. Given initial radius $\tau_{l_1}$, $\varepsilon > 0$, and $\eta \in (0, 1)$

$$U_{l_1}^h = \{u \in \mathbb{R}^n \mid f_{l_1}(u) \leq \theta(h) + \tau_{l_1}\}, \tag{6.61}$$

in this case, we have

$$U_{l_1}^0 = \{u \in \mathbb{R}^n \mid f_{l_1}(u) \leq \tau_{l_1}\}, \tag{6.62}$$

$$U_{l_1} = \lim_{h\to\infty} U_{l_1}^h = \left\{u \in \mathbb{R}^n \;\middle|\; f_{l_1}(u) \leq \frac{\eta}{1-\eta}\varepsilon + \tau_{l_1}\right\}. \tag{6.63}$$

The subgradient projection function for the $l_1$ constraint set $U_{l_1}^h$ is

$$\tilde{P}_{U_{l_1}^h}(u) = \begin{cases} u + \frac{\theta(h)+\tau_{l_1}-f_{l_1}(u)}{\|u^*\|^2}u^*, & \text{if } f_{l_1}(u) > \theta(h) + \tau_{l_1}, \\ u, & \text{if } f_{l_1}(u) \leq \theta(h) + \tau_{l_1}. \end{cases} \tag{6.64}$$

Here $u^* \in \partial f_{l_1}(u)$ is a subgradient of the $l_1$ function $f_{l_1}$ at point $u$, and the computation formula of subgradient $u^*$ is given by the following proposition.

**Proposition 6.12.** *Given $u \in \mathbb{R}^n$, and the linear operator $\Phi \in \mathbb{R}^{m\times n}$ with $\Phi = [\phi_1, \cdots, \phi_m]'$. The $l_1$ function is given by* (6.60)*, denote $e_i$ be the $i$-th identity element of vectors in $\mathbb{R}^n$. Given the function*

$$u^* = \sum_{i=1}^{m}\sum_{j=1}^{n} sgn\left(\langle\phi_i, u\rangle\right)\Phi_{i,j}e_j, \tag{6.65}$$

*then $u^* \in \partial f_{l_1}(u)$, i.e. $u^*$ is a subgradient of $f_{l_1}$ at $u$.*

*Proof.* By direct computation. □

When the linear operator $\Phi = I$, the $l_1$ function $f_{l_1}$ is just the $l_1$ norm of $u$

$$f_{l_1}(u) = \|u\|_1. \tag{6.66}$$

Also we can build the sequence of constraint sets as

$$U_{l_1}^0 = \{u \in \mathbb{R}^n \mid \|u\|_1 \leq \tau_{l_1}\}, \tag{6.67}$$

then,

$$U_{l_1}^h = \{u \in \mathbb{R}^n \mid \|u\|_1 \leq \theta(h) + \tau_{l_1}\}, \tag{6.68}$$

$$U_{l_1} = \lim_{h\to\infty} U_{l_1}^h = \left\{u \in \mathbb{R}^n \;\middle|\; \|u\|_1 \leq \frac{\eta}{1-\eta}\varepsilon + \tau_{l_1}\right\}. \tag{6.69}$$

The corresponding subgradient projection function to set $U_{l_1}^h$ is given by

$$\tilde{P}_{U_{l_1}^h}(u) = \begin{cases} u + \frac{\theta(h)+\tau_{l_1}-\|u\|_1}{\|u^*\|^2}u^*, & \text{if } \|u\|_1 > \theta(h) + \tau_{l_1}, \\ u, & \text{if } \|u\|_1 \leq \theta(h) + \tau_{l_1}, \end{cases} \tag{6.70}$$

where the vector $u^*$ is a subgradient of $\|u\|_1$. By Proposition 6.12, let

$$u^* = \sum_{i=1}^{n} \text{sgn}(u_i)e_i, \tag{6.71}$$

then $u^* \in \partial(\|u\|_1)$.

## 6.3 Discussion on the projection algorithm

We developed the gradient projection methods with the set expanding strategy in the previous chapter. An algorithm that projects a point towards the convex constraint set by generating a convergence sequence is needed. Since our initial idea is incorporating multiple convex constraints into the optimization problem, the projection methods onto the intersection of convex sets are needed. In this section, we discuss the projection algorithm used in our work. We begin with an initial point $x^0 \in \mathbb{R}^n$, a family of nonempty, convex, and closed sets $X_i$ with the index set $I = \{1, \cdots, N_c\}$. Suppose $X = \cap_{i\in I} X_i$ is nonempty, the projection problem is

$$\text{find } \bar{x} = \arg\min_{x\in\mathbb{R}^n} \|x - x^0\|^2, \quad \text{such that } x \in X = \cap_{i\in I} X_i. \tag{6.72}$$

### 6.3.1 Review of convex feasibility problem

Solving the convex feasibility problem is a part of the projection algorithm used in our work. Before the discussion of the projection algorithm, we give a review of numerical methods for the convex feasibility

problem based on the work [43]. The convex feasibility problem is given by

$$\text{find } \bar{x} \in X = \cap_{i \in I} X_i. \tag{6.73}$$

A class of the most popular algorithms is the projections onto convex sets (POCS) method (or named alternating projection method) [64, 80, 143]. Suppose there are $N_c$ constraint sets. Denote $P_i$ as the projection function onto set $X_i$, for $i \in I = \{1, \cdots, N_c\}$, the basic POCS generates a sequence $\{x^k\}$ with periodic projections onto the sets

$$x^{k+1} = P_{(k \mod N_c)+1}(x^k). \tag{6.74}$$

Although the POCS method has a simple form, it suffers several shortcomings as slow convergence, only processing one projection per iteration, the exact projection onto set $X_i$ is needed, etc [43]. Improvements have been made based on the basic POCS method, and new methods have been come up with such as the simultaneous iterative reconstruction technique [61],

$$x^{k+1} = \frac{1}{N_c} \sum_{i \in I} P_i(x^k), \tag{6.75}$$

and the parallel projection method [42],

$$x^{k+1} = x^k + \lambda_k \left( \sum_{i \in I} \omega_i P_i(x^k) - x^k \right). \tag{6.76}$$

Here $\lambda_k$ is a relaxation parameter, and $\omega_i$ is the weight parameter for each of the set $X_i$ with $\sum_i \omega_i = 1$.

Another way to solve the projection problem is to work in the $N_c$-fold Cartesian product space. Denote the $N_c$-fold Cartesian product space as $\mathbf{H} = \mathcal{H}^{N_c} = (\mathbb{R}^n)^{N_c}$, the inner product is given by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i \in I} \omega_i \langle x_i, y_i \rangle$, where $\omega_i$ is the weight parameter mentioned above. Denote the product set as

$$\mathbf{X} = X_1 \times \cdots \times X_{N_c} = \{ \mathbf{x} \in \mathbf{H} \mid x_i \in X_i,\ i \in I \}, \tag{6.77}$$

and

$$\mathbf{D} = \{ \mathbf{x} \in \mathbf{H} \mid x_i = x_j,\ \forall i, j \in I \}. \tag{6.78}$$

Then the convex feasibility problem (6.73) can be rewritten as

$$\text{find } \bar{\mathbf{x}} \in \mathbf{X} \cap \mathbf{D}. \tag{6.79}$$

The extrapolated parallel projection method [108, 43] is based on this construction as

$$x^{k+1} = x^k + \lambda_k \left( P_{\mathbf{D}} \circ P_{\mathbf{X}}(x^k) - x^k \right). \tag{6.80}$$

Further improvements focus on two aspects. First, since multiple projections are involved in each of the iterations, a parallel computing structure for computing all projection at the same time is expected. The problem happens when the workers (processors) in parallel computing is less than the number of sets $N_c$. The method of parallel projections [47, 32] is developed for this issue as

$$x^{k+1} = x^k + \lambda_k \left( \sum_{i \in I_k} \omega_{i,k} P_i(x^k) - x^k \right), \tag{6.81}$$

where $I_k \subset I$. More convergence analysis results can be found in [100]. Second, only an approximate projection like the subgradient projection can be evaluated for some of the sets. The generalization of extrapolated parallel projection method [43] is developed for this issue,

$$x^{k+1} = x^k + \lambda_k \left( P_{\mathbf{D}} \circ P_{\mathbf{X_k}}(x^k) - x^k \right), \tag{6.82}$$

here $X_k$ is a superset of $X$ and $x^k \notin X_k$.

As a summary to the above numerical methods, the author of [43] developed the algorithm named extrapolated method of parallel subgradient projections (EMOPSP),

$$x^{k+1} = x^k + \lambda_k \left( \sum_{i \in I_k} \omega_{i,k} P_i(x^k) - x^k \right), \tag{6.83}$$

here the weight parameter satisfies $\sum_{i \in I_k} \omega_{i,k} = 1$. For a more detailed review of the development of projection algorithms please refer to [43]. A complete study of the algorithm for the convex feasibility problem is given by [7].

### 6.3.2 Projection in scaled Euclidean space

Given a family of nonempty, convex, and closed sets $X_i$, $i \in I$, suppose $X = \cap_{i \in I} X_i$ is nonempty. Given an initial point $x^0 \in \mathbb{R}^n$, a sequence $\{x^k\}$ can be built with the basic POCS algorithm or the EMOPSP method

that towards to $X$. However, it is not guaranteed that the sequence $\{x^k\}$ converges to point $P_X(x^0)$, that is the projection onto the intersection of convex sets. More sophisticated algorithm is needed for the projection problem. In the work [105], the author suggest using the Dykstra's projection algorithm to solve similar problem.

In this work, we consider another projection algorithm based on the work [44, 45], which is compatible with both closed-form projection and subgradient projection. In [44] the author provides an outer approximation scheme in the Banach space setting to solve the projection algorithm. Then he proposed an adaptation of the outer approximation scheme to the quadratic signal recovery problem, and also the parallel computing structure is considered [45]. Later he applied the above algorithm for image restoration problem with total variation constraint [46].

Our initial idea is to represent multiple a priori information as convex constraint sets and then proceed with the optimization algorithm on the intersection of the constraint sets. Since only a finite number of constraint sets is considered, and both exact projection with respect to "simple" sets and subgradient projection are used. In this case, a simple version of the outer approximation algorithm in [44] can fulfill our demand. As discussed in the previous chapter, the gradient projection methods are performed in both Euclidean space and scaled Euclidean space. This means projection algorithms are needed for both cases. The Euclidean space can be considered as a special case of scaled Euclidean space with letting $B = I$ in $\mathcal{H}_B$. Given an initial point $x^0 \in \mathbb{R}^n$, a family of nonempty, convex, and closed sets $X_i$ with the index set $I = \{1, \cdots, N_c\}$, and $X = \cap_{i\in I} X_i$ is nonempty. We focus on the following problem in this section, let

$$f(x) = \|x - x^0\|_B^2, \tag{6.84}$$

the projection problem in $\mathcal{H}_B$ is: given $x^0$,

$$\text{find } \bar{x} = \arg\min_{x\in\mathbb{R}^n} f(x), \quad \text{such that } x \in X = \cap_{i\in I} X_i. \tag{6.85}$$

With the notations in the previous chapter, we can denote $P_{B,X}(x^0) = \bar{x}$.

Next, we introduce the projection algorithm used in this work. For each convex set $X_i$, assign a corresponding weight parameter $\omega_i \neq 0$ and $\sum_{i\in I} \omega_i = 1$. Then, the point $x^0$ is projected onto the intersection of $X = \cap_{i\in I} X_i$ with an iterative process. Suppose at the $k$-th iteration, denote $p_i = P_{X_i}(x^k)$ when set $X_i$ has a closed-form projection function. When the set $X_i$ is not simple and the subgradient projection is available, the subgradient projection is used and denote $p_i = \tilde{P}_{X_i}(x^k)$. Like the EMOPSP method, project the point

$x^k$ towards $X$ with

$$z^k = x^k + \lambda_k B^{-1}\left(\sum_{i\in I}\omega_i p_i - x^k\right), \tag{6.86}$$

where the relation parameter $\lambda_k$ is given by

$$\lambda_k = \begin{cases} \frac{\sum_{i\in I}\omega_i\|p_i - x^k\|^2}{\|x^k - \sum_{i\in I}\omega_i p_i\|_{B^{-1}}}, & \text{if } x^k \notin X, \\ 1/\|B^{-1}\|, & \text{otherwise.} \end{cases} \tag{6.87}$$

With the point $x^0$, $x^k$, and $z^k$, two half-spaces can be constructed

$$\begin{aligned} D_k &= \left\{x \in \mathbb{R}^n \;\middle|\; \left\langle x - x^k, -\nabla f(x^k)\right\rangle \le 0\right\} \\ &= \left\{x \in \mathbb{R}^n \;\middle|\; \left\langle x - x^k, x^0 - x^k\right\rangle_B \le 0\right\}, \end{aligned} \tag{6.88}$$

and

$$H_k = \left\{x \in \mathbb{R}^n \;\middle|\; \left\langle x - z^k, x^k - z^k\right\rangle_B \le 0\right\}. \tag{6.89}$$

Then an outer approximation can be constructed as the intersection of the above two closed half-spaces containing $X$ [45]. The set $H_k$ is named the surrogate half-space (or surrogate cut) [44]. For more discussion on the outer approximation we refer to [44, 45]. The following definition gives the projection function onto the intersection of the above half-spaces.

**Definition 6.13** ([45], Definition 10)**.** *Given $(u, v, w) \in (\mathbb{R}^n)^3$ such that*

$$\begin{aligned} A = &\{x \in \mathbb{R}^n \mid \langle x - v, u - v\rangle_B \le 0\} \cap \\ &\{x \in \mathbb{R}^n \mid \langle x - w, v - w\rangle_B \le 0\} \ne \emptyset, \end{aligned} \tag{6.90}$$

*Denote $Q_B(u, v, w)$ as the projection of $u$ onto $A$ in $\mathcal{H}_B$, i.e., $Q_B(u, v, w) = P_{B,A}(u)$.*

The following lemma shows how to evaluate $P_{B,A_k}(x^0)$.

**Lemma 6.14** ([45], Lemma 11, [44], eq (6.9))**.** *Set $\pi_k = \left\langle x^0 - x^k, x^k - z^k\right\rangle_B$, $\mu_k = \|x^0 - x^k\|_B^2$, $\nu_k = \|x^k - z^k\|_B^2$, and $\rho_k = \mu_k\nu_k - \pi_k^2$. Then*

$$Q_B(x^0, x^k, z^k) = \begin{cases} z^k, & \text{if } \rho_k = 0 \text{ and } \pi_k \ge 0, \\ x^0 + \left(1 + \frac{\pi_k}{\nu_k}\right)\left(z^k - x^k\right), & \text{if } \rho_k > 0 \text{ and } \pi_k\nu_k \ge \rho_k, \\ x^k + \frac{\nu_k}{\rho_k}\left(\pi_k(x^0 - x^k) + \mu_k(z^k - x^k)\right), & \text{if } \rho_k > 0 \text{ and } \pi_k\nu_k < \rho_k. \end{cases} \tag{6.91}$$

Notice that, on each of the iteration, the projection process by EMOPSP is to find a closer point $z^k$ between $x^k$ and set $X$. The projection process is actually done by the above lemma which projects the initial point $x^0$ onto the outer approximation $A_k = D_k \cap H_k$ of $X$ iteratively. Algorithm 7 is used in our proposed optimization scheme, that is a simplified version of Algorithm 14 in [45], with the name of surrogate constraint splitting algorithm.

**Algorithm 7:** Projection algorithm 1

Initialization: a family of nonempty convex closed set $X_i$, $i \in I$, initial point $x^0$, weight parameter $\omega_i \neq 0$, $i \in I$ with $\sum_{i \in I} \omega_i = 1$.
At the $k$-th iteration:
**while** *Not converge* **do**
Step 1: compute the projection of $x^k$ onto each of $X_i$ with:

$$p_i = \begin{cases} P_{X_i}(x^k), & \text{if } X_i \text{ is simple,} \\ \tilde{P}_{X_i}(x^k), & \text{if } X_i \text{ is not simple and have subgradient projection.} \end{cases} \tag{6.92}$$

Step 2: set $z^k = x^k + \lambda_k B^{-1} \left( \sum_{i \in I} \omega_i p_i - x^k \right)$, where $\lambda_k$ is given by equation (6.87).
Step 3: update $x^{k+1} = Q_B(x^0, x^k, z^k)$ with equation (6.91).
**end**

**Theorem 6.15** ([45], Theorem 16)**.** *Every sequence* $\{x^k\}$ *generated by Algorithm 7 converges strongly to the solution* $\bar{x}$ *of the projection problem* (6.85)*.*

*Proof.* Algorithm 7 is a simplified case of Algorithm 14 in [45], and it is easy to verify all assumptions in Assumption 15 which in [45] is satisfied. The above theorem is a restate of Theorem 16 in [45]. □

Observe that multiple evaluations of the matrix-vector multiplications with respect to $B$ and $B^{-1}$ are needed in equation (6.87), (6.91), and the step 2 of Algorithm 7. Algorithm 7 can be reorganized such that only one matrix-vector multiplication of $B$ and $B^{-1}$ are needed for each iteration. The reorganization has been discussed in the work [45], we restate it as Algorithm 8 for completeness. The stopping criteria are discussed in the next section for both algorithms.

## 6.4 Scaled gradient projection method with multiple constraints

In this section, we introduce the proposed algorithm which is a combination of the scaled gradient projection method in the previous chapter, the projection method in the previous section, and the L-BFGS Hessian approximation.

First, we state the problem. Denote the nonlinear objective function as $f : \mathbb{R}^n \to \mathbb{R}$ which is smooth and bounded from below, and $f$ can be a nonconvex function. Given a family of constraint sets $U_i$ and an

**Algorithm 8:** Projection algorithm 2

Initialization: a family of nonempty convex closed set $X_i$, $i \in I$, initial point $x^0$, weight parameter $\omega_i \neq 0$, $i \in I$ with $\sum_{i\in I} \omega_i = 1$.

At the $k$-th iteration:

**while** *Not converge* **do**

Step 1: compute the projection of $x^k$ onto each of $X_i$ with:

$$p_i = \begin{cases} P_{X_i}(x^k), & \text{if } X_i \text{ is simple,} \\ \tilde{P}_{X_i}(x^k), & \text{if } X_i \text{ is not simple and have subgradient projection.} \end{cases} \tag{6.93}$$

Step 2: set $v = \sum_{i\in I} \omega_i a_i$ with $a_i = p_i - x^k$.

Step 3: set $\lambda = \sum_{i\in I} \omega_i \|a_i\|^2$. If $\lambda = 0$, set $x^{k+1} = x^k$, break; otherwise set

$$b = x^0 - x^k, \quad c = Bb, \quad d = B^{-1}v, \quad \lambda = \frac{\lambda}{\langle d, v\rangle}. \tag{6.94}$$

Step 4: set $d = \lambda d$, then compute

$$\pi = -\langle c, d\rangle, \quad \mu = \langle b, c\rangle, \quad \nu = \lambda \langle d, v\rangle, \quad \rho = \mu\nu - \pi^2. \tag{6.95}$$

Step 5: update with

$$x^{k+1} = \begin{cases} x^k + d, & \text{if } \rho = 0 \text{ and } \pi \geq 0, \\ x^0 + \left(1 + \frac{\pi}{\nu}\right) d, & \text{if } \rho > 0 \text{ and } \pi\nu \geq \rho, \\ x^k + \frac{\nu}{\rho}\left(\pi b + \mu d\right), & \text{if } \rho > 0 \text{ and } \pi\nu < \rho. \end{cases} \tag{6.96}$$

**end**

index set $I = \{1, \cdots, N_c\}$, $U_i$ is nonempty, convex, and closed for each of index $i \in I$. Let the feasible set $U_{\text{ad}} = \cap_{i\in I} U_i$ be nonempty, the optimization problem is given as

$$\min_{u\in\mathbb{R}^n} f(u), \quad \text{such that } u \in U_{\text{ad}} = \cap_{i\in I} U_i. \tag{6.97}$$

We are going to implement the scaled gradient projection method to solve the above problem (6.97). The second order Taylor approximation of objection function at the $k$-th iteration is

$$f_k(u) = \left\langle \nabla f(u^k), u - u^k \right\rangle + \frac{1}{2}\left\langle B_k(u - u^k), u - u^k \right\rangle, \tag{6.98}$$

where $B_k$ is the Hessian approximation for the $k$-th iteration which is a symmetric positive definite matrix. Then the scaled gradient projection method at $k$-th iteration can be written as

$$\bar{u}^k = \arg\min_{u\in U_{\text{ad}}} f_k(u), \tag{6.99}$$

$$u^{k+1} = u^k + \alpha_k(\bar{u}^k - u^k). \tag{6.100}$$

Here $\alpha_k$ is the line search parameter achieved with Armijo rule or Wolfe conditions. As discussed in the previous chapter, the scaled gradient projection method is equivalent to

$$\tilde{u}^k = u^k - B_k \nabla f(u^k), \tag{6.101}$$

$$\bar{u}^k = P_{B_k, U_{\text{ad}}}(\tilde{u}^k), \tag{6.102}$$

$$u^{k+1} = u^k + \alpha_k(\bar{u}^k - u^k). \tag{6.103}$$

Notice that the subproblem (6.102) is equivalent to the projection problem (6.85) in the previous section. Our goal is to implement Algorithm 7 and 8 to the subproblem (6.102) in the scaled gradient projection method. Three problems arise:

1. Algorithm 7 and 8 projects $\tilde{u}^k$ to $P_{B_k, U_{\text{ad}}}(\tilde{u}^k)$ by generating a converging sequence $\{z^j\}$, with $\lim_{j\to\infty} z^j = P_{B_k, U_{\text{ad}}}(\tilde{u}^k)$. Then the projection process is actually an inexact projection and the projection result is an approximation of $P_{B_k, U_{\text{ad}}}(\tilde{u}^k)$, denoted as $\bar{P}_{B_k, U_{\text{ad}}}(\tilde{u}^k)$. In this case $\bar{P}_{B_k, U_{\text{ad}}}(\tilde{u}^k)$ might not in the feasible set $U_{\text{ad}}$.

2. We need to choose a stopping criteria for projection Algorithm 7 and 8.

3. Since the multiplications between matrix $B_k, B_k^{-1}$ and vectors are evaluated at every iteration of Algorithm 7 and 8, an efficient Hessian approximation is needed.

To solve the first problem, the set expanding strategy is introduced to the scaled gradient projection method in the previous chapter. There are two strategies to design the expanding set sequence:

1. Given the constraint set $U_i$, start with the initial constraint set $U_i^0 = U_i$, design

$$\tilde{U}_i = \lim_{k\to\infty} U_i^k, \quad U_i^k \subset U_i^{k+1}, \quad U_i^k \neq U_i^{k+1}, \tag{6.104}$$

where $U_i^k$ is nonempty, closed, and convex. In this case, $\tilde{U}_i$ is larger than the initial constraint set $U_i^0$. Denote $U_{\text{ad}}^k = \cap_{i\in I} U_i^k$, then

$$\tilde{U}_{\text{ad}} = \lim_{k\to\infty} U_{\text{ad}}^k, \quad U_{\text{ad}}^k \subset U_{\text{ad}}^{k+1}, \quad U_{\text{ad}}^k \neq U_{\text{ad}}^{k+1}. \tag{6.105}$$

In this case, the final constraint set $\tilde{U}_{\text{ad}}$ is larger than the constraint set $U_{\text{ad}}$ in problem (6.97). It is acceptable when $\tilde{U}_{\text{ad}}$ does not extend too much compared to $U_{\text{ad}}$ in practice.

2. Start with a smaller initial constraint set $U_i^0 \subset U_i$, design

$$U_i = \lim_{k\to\infty} U_i^k, \quad U_i^k \subset U_i^{k+1}, \quad U_i^k \neq U_i^{k+1}, \tag{6.106}$$

where $U_i^k$ is nonempty, closed, and convex. Denote $U_{\text{ad}}^k = \cap_{i\in I} U_i^k$, then

$$U_{\text{ad}} = \lim_{k\to\infty} U_{\text{ad}}^k, \quad U_{\text{ad}}^k \subset U_{\text{ad}}^{k+1}, \quad U_{\text{ad}}^k \neq U_{\text{ad}}^{k+1}. \tag{6.107}$$

In this case, the final constraint set $U_{\text{ad}}$ is the same as the constraint set in problem (6.97).

The construction of different set sequence is introduced in the previous sections.

Next, we describe the stopping criteria of Algorithm 7 and 8. Since the projection algorithms are generating a convergence sequence $\{z^j\}$ converges to $P_{B_k,U_{\text{ad}^h}}(\tilde{u}^k)$. Denote the inexact projection operator as $\tilde{P}_{B_k,U_{\text{ad}}^h}(\tilde{u}^k) = z^{j_0}$ for some index $j_0 \in \mathbb{N}$. With the discussion in the previous chapter, the stopping criteria of Algorithm 7 and 8 is given by

$$z^{j_0} \in U_{\text{ad}}^{k+1}, \tag{6.108}$$

$$\left\langle \tilde{u}^k - z^{j_0}, u^k - z^{j_0} \right\rangle_{B_k} \leq 0. \tag{6.109}$$

For the last problem, the L-BFGS quasi-Newton approximation of Hessian matrix is applied. Denote $H_k = B_k^{-1}$, we rewrite the L-BFGS approximation in the compact form for completeness. Denote

$$s_k = u^{k+1} - u^k, \quad y_k = \nabla f(u^{k+1}) - \nabla f(u^k). \tag{6.110}$$

The L-BFGS approximation of Hessian approximation and inverse Hessian are

$$B_k = \sigma_k I - \begin{bmatrix} \sigma_k S_k & Y_k \end{bmatrix} \begin{bmatrix} \sigma_k S_k^{'} S_k & U_k \\ U_k^{'} & -D_k \end{bmatrix}^{-1} \begin{bmatrix} \sigma_k S_k^{'} \\ Y_k^{'} \end{bmatrix}, \tag{6.111}$$

$$H_k = \gamma_k I + \begin{bmatrix} S_k & \gamma_k Y_k \end{bmatrix} \begin{bmatrix} (R_k^{'})^{-1} \left( D_k + \gamma_k Y_k^{'} Y_k \right) R_k^{-1} & -(R_k^{'})^{-1} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^{'} \\ \gamma_k Y_k^{'} \end{bmatrix}, \tag{6.112}$$

where $\gamma_k = y_{k-1}^{'} s_{k-1} / y_{k-1}^{'} y_{k-1}$, $\sigma_k = y_{k-1}^{'} s_{k-1} / s_{k-1}^{'} s_{k-1}$. The coefficients are given by

$$
\begin{aligned}
S_k &= [s_{k-m}, \cdots, s_{k-1}], \quad Y_k = [y_{k-m}, \cdots, y_{k-1}], \\
(R_k)_{i,j} &= \begin{cases} (s_{k-m-1+i})^{'} (y_{k-m-1+j}), & \text{if } i \leq j, \\ 0, & \text{otherwise}, \end{cases} \\
D_k &= \operatorname{diag}(s_{k-m}^{'} y_{k-m}, \cdots, s_{k-1}^{'} y_{k-1}), \\
(U_k)_{i,j} &= \begin{cases} (s_{k-m-1+i})^{'} (y_{k-m-1+j}), & \text{if } i > j, \\ 0, & \text{otherwise}. \end{cases}
\end{aligned}
\tag{6.113}
$$

With the above discussion, the scaled gradient projection with multiple constraints algorithm is given by the following algorithm.

**Algorithm 9:** Scaled gradient projection method with multiple constraints

Given: the objective function $f$ and initial value $u^0$; a family of nonempty, closed, convex constraint sets $U_i$, for $i \in I = \{1, \cdots, N_c\}$,
Construct: for each $U_i$ construct an increasing set sequence $\{U_i^j\}_{j \in \mathbb{N}}$; set $U_{\text{ad}}^0 = \cap_{i \in I} U_i^0$.
**while** *Not converge* **do**
  At $k$-th iteration:
  Step 1: Compute $\nabla f(u^k)$.
  Step 2: Update $s_k$ and $y_k$ with equation (6.110), $S_k$, $Y_k$, $R_k$, $U_k$ with equation (6.113).
  Step 3: Compute $\tilde{u}^k = u^k - H_k \nabla f(u^k)$ with equation (6.112).
  Step 4: Compute $\bar{u}^k = P_{B_k, U_{\text{ad}}^k}(\tilde{u}^k)$, i.e., project $\tilde{u}^k$ to $U_{\text{ad}}^k$ in $\mathcal{H}_{B_k}$ with Algorithm 8, until the stopping criteria equation (6.108) (6.109) are satisfied. The multiplication between $B_k, H_k$ and vectors are evaluated with equation (6.111) and (6.112).
  Step 5: Update $u^{k+1} = u^k + \alpha_k(\bar{u}^k - u^k)$, here $\alpha_k$ is the line search parameter achieved with the Wolfe conditions.
  Step 6: Construct $U_{\text{ad}}^{k+1} = \cap_{i \in I} U_i^{k+1}$, set $k = k + 1$.
**end**

Sometimes we do not expect the constraint set sequence to expand too fast, an adaptively expanding sequence of constraint sets can be designed. An independent index $h_i \in \mathbb{N}$ is used corresponding to each of the constraint set $U_i$ as

$$
U_i = \lim_{h_i \to \infty} U_i^{h_i}, \quad U_i^{h_i} \subset U_i^{h_i+1}, \quad U_i^{h_i} \neq U_i^{h_i+1}, \tag{6.114}
$$

where $U_i^{h_i}$ is nonempty, closed, and convex. At the $k$-th iteration, suppose $U_i^k = U_i^{h_i}$, if $u^{k+1} \in U_i^k$, set $U_i^{k+1} = U_i^{h_i}$, that means the $i$-th constraint set is not expanding at the $(k+1)$-th iteration. Otherwise, set $U_i^{k+1} = U_i^{h_i+1}$. The adaptive algorithm is given by the following algorithm.

**Algorithm 10:** Scaled gradient projection method with multiple constraints and adaptive set expanding strategy

Given: the objective function $f$ and initial value $u^0$; a family of nonempty, closed, convex constraint sets $U_i$, for $i \in I = \{1, \cdots, N_c\}$,
Construct: for each $U_i$ construct an increasing set sequence $\{U_i^{h_i}\}_{h_i \in \mathbb{N}}$; set $U_{\text{ad}}^0 = \cap_{i \in I} U_i^0$.
**while** *Not converge* **do**
At $k$-th iteration:
Step 1: Compute $\nabla f(u^k)$.
Step 2: Update $s_k$ and $y_k$ with equation (6.110), $S_k$, $Y_k$, $R_k$, $U_k$ with equation (6.113).
Step 3: Compute $\tilde{u}^k = u^k - H_k \nabla f(u^k)$ with equation (6.112).
Step 4: Compute $\bar{u}^k = P_{B_k, U_{\text{ad}}^k}(\tilde{u}^k)$, i.e., project $\tilde{u}^k$ to $U_{\text{ad}}^k$ in $\mathcal{H}_{B_k}$ with Algorithm 8, until the stopping criteria equation (6.108) (6.109) are satisfied. The multiplication between $B_k, H_k$ and vectors are evaluated with equation (6.111) and (6.112).
Step 5: Update $u^{k+1} = u^k + \alpha_k(\bar{u}^k - u^k)$, here $\alpha_k$ is the line search parameter achieved with the Wolfe conditions.
Step 6: For each $i \in I$, update the constraint sets: if $u^{k+1} \in U_i^k$, then set $U_i^{k+1} = U_i^{h_i}$; otherwise, set $U_i^{k+1} = U_i^{h_i+1}$, $h_i = h_i + 1$.
Step 7: Construct $U_{\text{ad}}^{k+1} = \cap_{i \in I} U_i^{k+1}$, set $k = k + 1$.
**end**

## 6.5 Applications with full waveform inversion problem

We are working on the full waveform inversion problem in discrete sense.

$$\min_{y_1, \cdots, y_{N_s} \in Y, u \in U_{\text{ad}}} J(y_1, \cdots, y_{N_s}, u) = \sum_{s=1}^{N_s} \frac{1}{2} \|Q y_s - y_{d,s}\|^2, \tag{6.115}$$

$$\text{such that } e_s(y_s, u) = 0, \quad s = 1, \cdots, N_s. \tag{6.116}$$

Here $Y$ is the wavefield space and $U_{\text{ad}} \subset U = \mathbb{R}^n$ is the feasible set. We have $N_s$ sources in this model, and each index $s$ is corresponding to a wave equation which is written in a compact form as $e_s(y, u) = 0$. Here $Q$ is the observation operator recording the corresponding wavefield $y_s$.

Multiple constraint sets are provided to improve the inverse result. In this case, the constrained optimization problem is processed in the intersection of several constraint sets:

$$U_{\text{ad}} = \cap_{i=1}^{N_c} U_i, \tag{6.117}$$

where $U_i \subset \mathbb{R}^n$ is nonempty, closed, and convex. From the discussion in Chapter 2, the full waveform inversion problem has a reduced form

$$\min_{u \in U_{\text{ad}}} f(u) = \min_{u \in U_{\text{ad}}} J(y_1(u), \cdots, y_{N_s}(u), u). \tag{6.118}$$

Rewrite the reduced full waveform inversion problem in the abstract form:

$$\min_u f(u), \quad \text{such that } u \in U_{\text{ad}} = \cap_{i=1}^{N_c} U_i. \tag{6.119}$$

Then, the proposed Algorithm 10 in the previous section can be applied to solve the reduced form FWI problem. The following numerical examples are the applications of the proposed algorithm. The cross-well examples are provided in Example 1 and Example 2, and the reflection seismic experiment is provided in Example 3.

### 6.5.1 Example 1: Cross-well model 1

A cross-well model is studied in this example as shown in Figure 6.2 (a), denoted as $u_{\text{true}}$. The initial velocity model is shown in Figure 6.2 (b). There are 6 equally spaced sources in the left boundary of the domain, and there are 51 equally spaced receivers in the right boundary of the domain. The model size is 1 km by 1 km and is discretized with size $101 \times 101$. A finite difference scheme is used to solve the wave equation with spatial step size 0.01 km and temporal step size 0.0005 s. The perfectly matched layer technique is used to simulate the wave propagation in a boundary-free domain. The source is a Ricker wavelet with 5 Hz peak frequency.

Three constraints are considered: box constraint, total variation constraint, and hyperplane constraint. When box constraint is used for the FWI problem, the first set expanding strategy given by equation (6.104) and (6.105) is used. The reason is: the inaccurate lower bound of the velocity model leads to an inaccurate Born approximation. The sequence of box constraint sets is given by

$$U_1^h = \{u \in \mathbb{R}^n \mid 1 - \theta_1(h) \leq u_i \leq 1.2 + \theta_1(h),\ i = 1, \cdots, n\}, \tag{6.120}$$

$$\text{where } \theta_1(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.001 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.121}$$

The sequence of total variation constraint sets is given by

$$U_2^h = \{u \in \mathbb{R}^n \mid f_{\text{TV}}(u) \leq 24 + \theta_2(h)\}, \tag{6.122}$$

$$\text{where } \theta_2(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.24 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.123}$$

Here the TV function $f_{\text{TV}}$ is given in equation (6.47). The sequence of the first hyperplane constraint sets

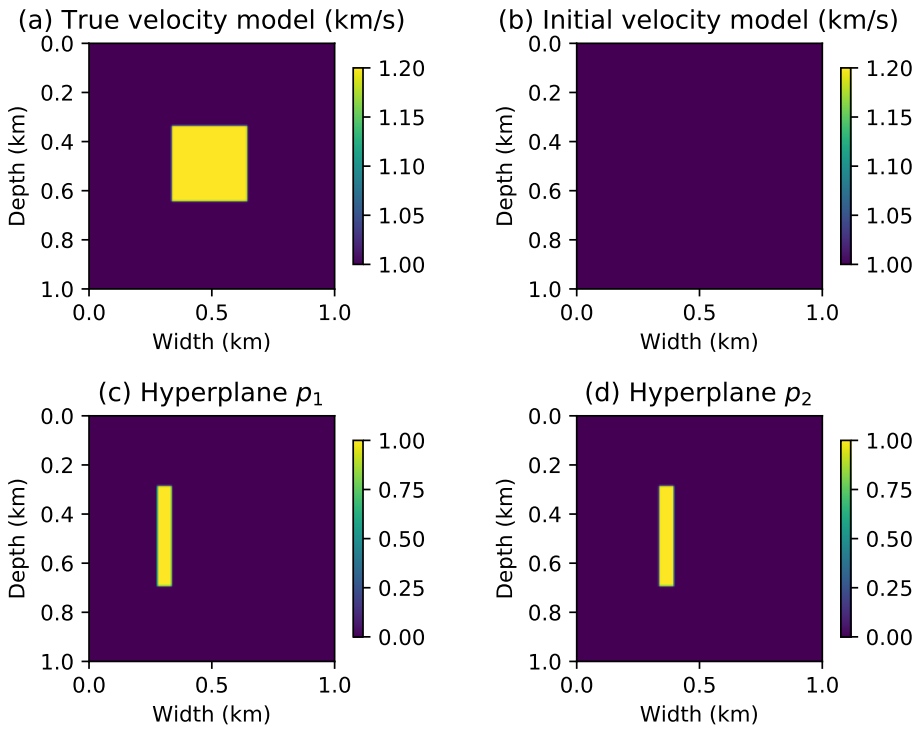


Figure 6.2: (a): True velocity model. (b): Initial velocity model used in the FWI problem. (c): The hyperplane $p_1$. (d): The hyperplane $p_2$.

is given by

$$U_3^h = \{u \in \mathbb{R}^n \mid \|u - P_1(u)\| \leq \theta_3(h) + 0.01\}, \tag{6.124}$$

$$\text{where } P_1(u) = u + \frac{\langle u_{\text{true}}, p_1\rangle - \langle u, p_1\rangle}{\|p_1\|^2} p_1, \tag{6.125}$$

$$\theta_3(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.01 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.126}$$

The sequence of the second hyperplane constraint set is given by

$$U_4^h = \{u \in \mathbb{R}^n \mid \|u - P_2(u)\| \leq \theta_4(h) + 0.01\}, \tag{6.127}$$

$$\text{where } P_2(u) = u + \frac{\langle u_{\text{true}}, p_2\rangle - \langle u, p_2\rangle}{\|p_2\|^2} p_2, \tag{6.128}$$

$$\theta_4(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.01 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.129}$$

The vector $p_1, p_2 \in \mathbb{R}^n$ are shown in Figure 6.2 (c) and (d). The hyperplane constraints used here are

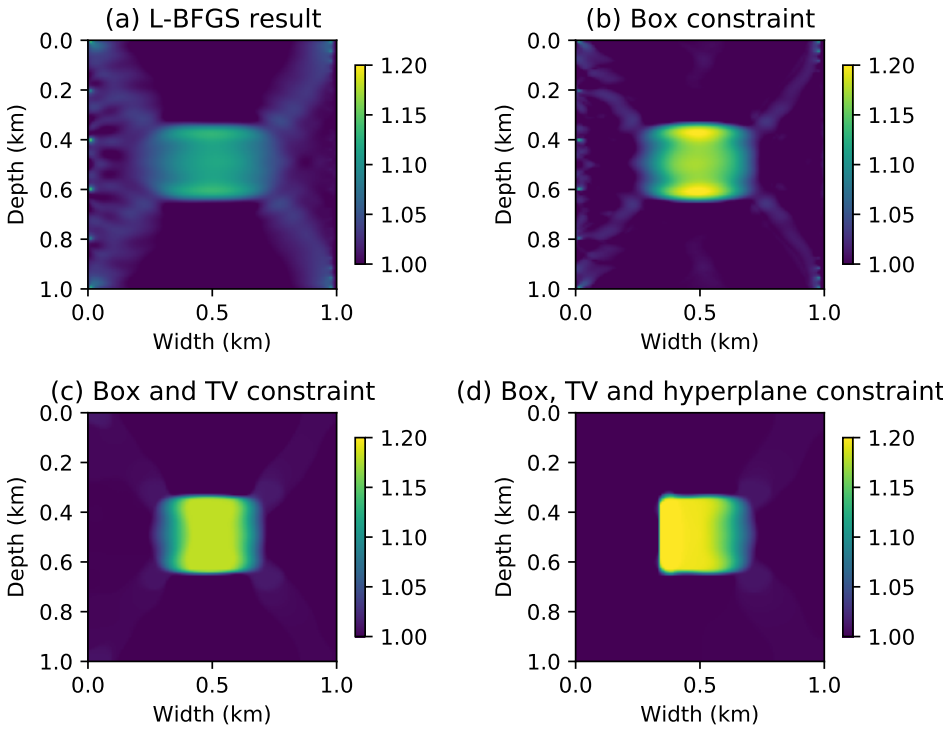


Figure 6.3: (a): Unconstrained result. (b): Inverse result with box constraint. (c): Inverse result with box and total variation constraint. (d): Inverse result with box, total variation and hyperplane constraint.

providing the average value information of the two areas near the left boundary of the velocity perturbation. In this case, an inverse result with a sharp left boundary can be expected.

The inverse results are shown in Figure 6.3. For the unconstrained result in subfigure (a), the L-BFGS algorithm is performed with 20 iterations. The proposed Algorithm 10 is performed with 20 iterations for the results in subfigures (b), (c), and (d). Subfigure (b) provides the inverse result with only box constraint sequence $\{U_1^h\}$ with the index set $I = \{1\}$ and weight parameter $\omega_1 = 1$. The inverse result in subfigure (b) provides a more accurate velocity value than the result in subfigure (a). Subfigure (c) provides the inverse result with box and total variation constraint sequences, with the index set $I = \{1, 2\}$ and the weight parameter $\omega_1 = \omega_2 = 1/2$. Compared with the result in subfigure (b), the inverse result in subfigure (c) provides a homogeneous velocity anomaly that is closed to the true model, and the faulty structure outside the anomaly is not significant. The inverse result with all four constraint sequence of sets is provided in subfigure (d), with the index set $I = \{1, 2, 3, 4\}$ and the weight parameter $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 1/4$. A sharp left boundary of the velocity anomaly is revealed because of the hyperplane constraints. For most of the inner projection process, the number of iterations ranging from hundreds to thousands. Compared with the computation cost of the PDE solver and the evaluation of gradient through the adjoint state method,

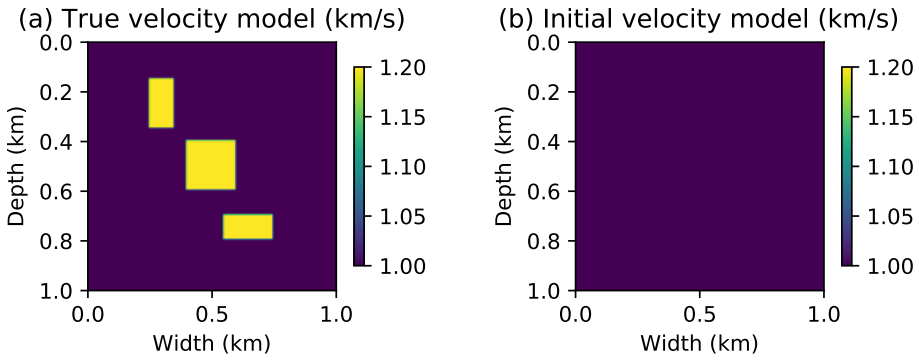


Figure 6.4: (a): True velocity model. (b): Initial velocity model.

the iterative projection process does not increase the overall computation time significantly.

All the constraints play a role in improving the inverse result compared with the unconstrained case. Although the result of hyperplane constraints is artificial, a sharp left boundary of the velocity perturbation is revealed. This provides a way to improve the local inverse results with accurate a priori information. This example shows that the proposed method can handle multiple constraints at the same time. With more information provided for the optimization algorithm, a more accurate image can be achieved.

### 6.5.2 Example 2: Cross-well model 2

In this example, we incorporate the $l_1$ constraint with the proposed method, a cross-well model similar to Example 1 is provided. The true velocity model and the initial velocity model are shown in Figure 6.4 (a) and (b). We use the initial velocity model as the reference model in the $l_1$ fidelity constraint sets, denoted as $u_{\text{ref}}$. The acquisition is the same as in Example 1.

Next, we denote the sequences of constraint sets, the first set expanding strategy given by equation (6.104) and (6.105) is used. The sequence of box constraint sets is given by

$$U_1^h = \{u \in \mathbb{R}^n \mid 1 - \theta_1(h) \leq u_i \leq 1.2 + \theta_1(h),\ i = 1, \cdots, n\}, \tag{6.130}$$

$$\text{where } \theta_1(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.001 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.131}$$

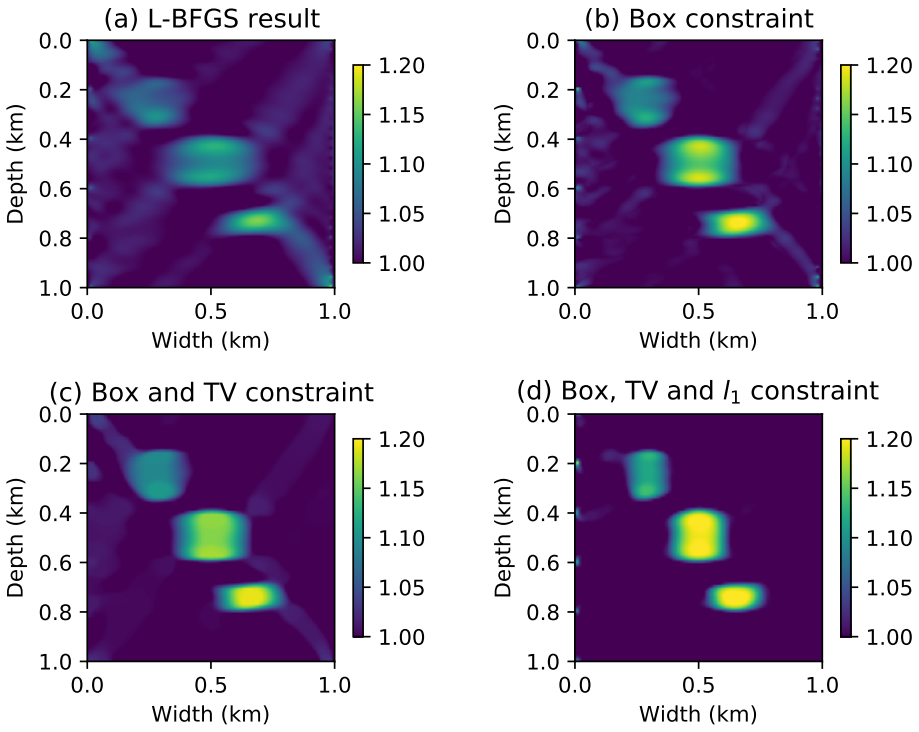


Figure 6.5: (a): Unconstrained result. (b): Inverse result with box constraint. (c): Inverse result with box and total variation constraint. (d): Inverse result with box, total variation and $l_1$ constraint.

The sequence of total variation constraint sets is given by

$$U_2^h = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq 39.5 + \theta_2(h)\}, \tag{6.132}$$

$$\text{where } \theta_2(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.395 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.133}$$

The sequence of $l_1$ constraint set is given by

$$U_3^h = \{u \in \mathbb{R}^n \mid \|u - u_{\mathrm{ref}}\|_1 \leq 128 + \theta_3(h)\}, \tag{6.134}$$

$$\text{where } \theta_3(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 1.28 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.135}$$

Numerical results are shown in Figure 6.5. For the unconstrained case, the L-BFGS algorithm is performed 20 iterations and the result is shown in subfigure (a). The proposed Algorithm 10 is performed with 20 iterations for the results in subfigures (b), (c), and (d). Subfigure (b) provides the inverse result with only

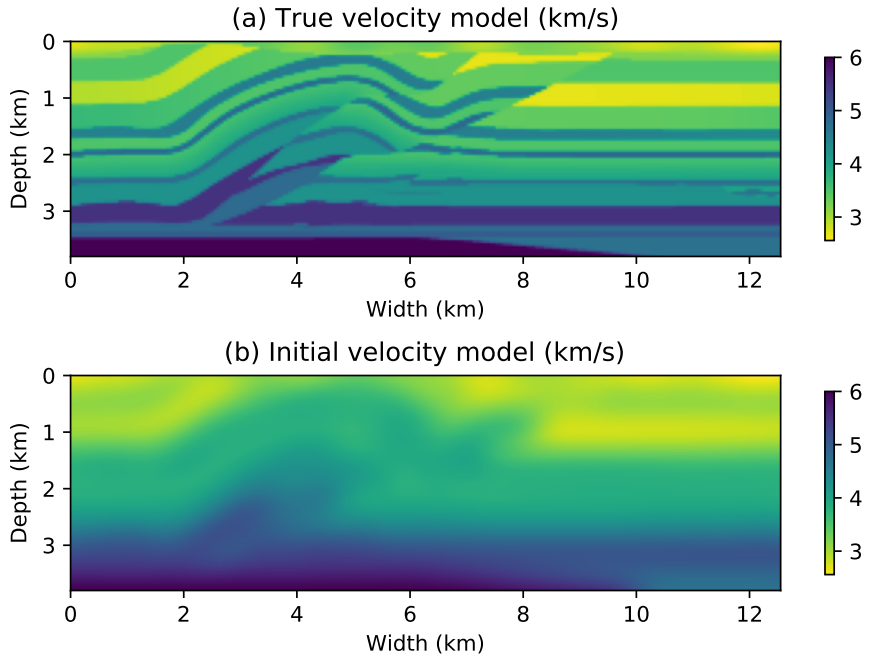


Figure 6.6: (a): True velocity model. (b): Initial velocity model.

the sequence of box constraint, with the index set $I = \{1\}$ and weight parameter $\omega_1 = 1$. Compared with the inverse result in subfigure (a), the result in subfigure (b) provides a more accurate velocity value and a clearer boundary. Subfigure (c) provides the inverse result with the sequence of box and total variation constraint, with the index set $I = \{1, 2\}$ and weight parameter $\omega_1 = \omega_2 = 1/2$. The velocity anomalies in subfigure (c) have more homogeneous structure compared with the results in subfigures (a) and (b). The inverse result with all three sequences of constraints is shown in subfigure (d), with the index set $I = \{1, 2, 3\}$ and weight parameter $\omega_1 = \omega_2 = \omega_3 = 1/3$. The faulty structure outside the velocity anomalies is not significant compared with the previous results, and the boundaries of the anomalies are clearly recovered. For the inner projection process, the number of iterations ranging from hundreds to thousands, and the iterative projection process does not increase the overall computation time significantly. Compared with different inverse results, both total variation constraint and $l_1$ constraint play an important role in providing a better inverse result.

### 6.5.3 Example 3: Overthrust model

A more realistic velocity model is provided in this example as shown in Figure 6.6 (a) that is a part of the Overthrust model. And the initial velocity model is shown in Figure 6.6 (b). Since the horizontal layer structure is prominent in the Overthrust model, it is reasonable to describe the true model as piece-wise

constant layers, and that corresponds to the total variation constraint. With 3.8 km depth and 12.55 km width, the model is discretizde into $76 \times 251$ points. There are 10 equally spaced sources and 126 equally spaced receivers on the top of the model. The finite difference method is used for the constraint equation, with spatial step size 0.05 km, and temporal step size 0.004 s. The perfectly matched layer technique is used to simulate the seismic wave propagating in a free domain. The 5 Hz Ricker wavelet is used for each of the sources.

In this example, we compare the inverse results with the different size of the constraints. The first set expanding strategy given by equation (6.104) and (6.105) is used. First, we fix the box constraint as

$$U_1^h = \{u \in \mathbb{R}^n \mid 2.5588 - \theta_1(h) \leq u_i \leq 6 + \theta_1(h),\ i = 1, \cdots, n\}, \tag{6.136}$$

$$\text{where } \theta_1(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 0.02 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.137}$$

Then we define three total variation constraint sequences with different radius

$$U_2^h = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq 1200 + \theta_2(h)\}, \tag{6.138}$$

$$\text{where } \theta_2(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 60 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.139}$$

$$U_3^h = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq 1000 + \theta_3(h)\}, \tag{6.140}$$

$$\text{where } \theta_3(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 50 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.141}$$

$$U_4^h = \{u \in \mathbb{R}^n \mid f_{\mathrm{TV}}(u) \leq 800 + \theta_4(h)\}, \tag{6.142}$$

$$\text{where } \theta_4(h) = \begin{cases} 0, & \text{if } h = 0, \\ \sum_{i=1}^{h} 40 \times 0.9^i, & \text{otherwise.} \end{cases} \tag{6.143}$$

There are 50 iterations performed for each of the following examples and the inverse results are shown in Figure 6.7. The unconstrained result with L-BFGS method is shown in subfigure (a). The proposed Algorithm 10 is performed for the case of subfigures (b), (c), and (d), and both box constraint and total

variation constraint are performed. For subfigure (b), the constraint sequence set index is $I = \{1, 2\}$, with weight parameter $\omega_1 = \omega_2 = 1/2$. For subfigure (c), the constraint sequence set index is $I = \{1, 3\}$, with weight parameter $\omega_1 = \omega_3 = 1/2$. For subfigure (d), the constraint sequence set index is $I = \{1, 4\}$, with weight parameter $\omega_1 = \omega_4 = 1/2$. For the inner projection process, the number of iterations ranged from hundreds to thousands, and the iterative projection process does not increase the overall computation time significantly. The total variation constraint is introduced to the inverse problem such that the piece-wise constant structure in the inverse result is expected. The inverse results in Figure 6.7 show that, as the total variation constraint radius is larger, the inverse result is closer to the unconstrained case. This example shows that the proposed method can control the inverse result by changing the radius of the sequence of constraint sets.

### 6.5.4 Discussion

The optimization scheme scaled gradient projection method with multiple constraints is provided in this work. The proposed scheme is a combination of scaled gradient projection method with inexact projection, L-BFGS Hessian approximation, and the iterative projection method proposed by [44, 45]. When the a priori information can be represented as convex constraint sets with closed-form projection or subgradient projection, it can be incorporated into the inverse problem with the proposed scheme. In this case, multiple a priori information of the inverse problem can be incorporated at the same time to provide a more accurate inverse result. Since the projection process in this scheme is closed-form projection or subgradient projection, and the L-BFGS Hessian approximation is used, the optimization scheme can be efficiently implemented similar to the L-BFGS method for the unconstrained optimization problem.

The full waveform inversion numerical examples provided in this work show that the a priori information of the problem indeed makes a difference compared to the unconstrained case. And the proposed optimization scheme is efficient and flexible to incorporate multiple a priori information to the problem. The box constraint is one of the most commonly used constraints for the PDE constrained optimization problem, and it provides accurate lower and upper bounds of the parameter. The hyperplane constraint can be used to provide the average value of certain areas. The total variation constraint is useful for providing a piece-wise constant structure of the inverse result, and a more homogeneous velocity anomaly can be achieved. The $l_1$ constraint can be used to enhance the sparsity of the inverse result. As shown with the numerical examples, the size of the constraint sets can control the inverse result, and it should be set based on the a priori information of the true model or the problem.

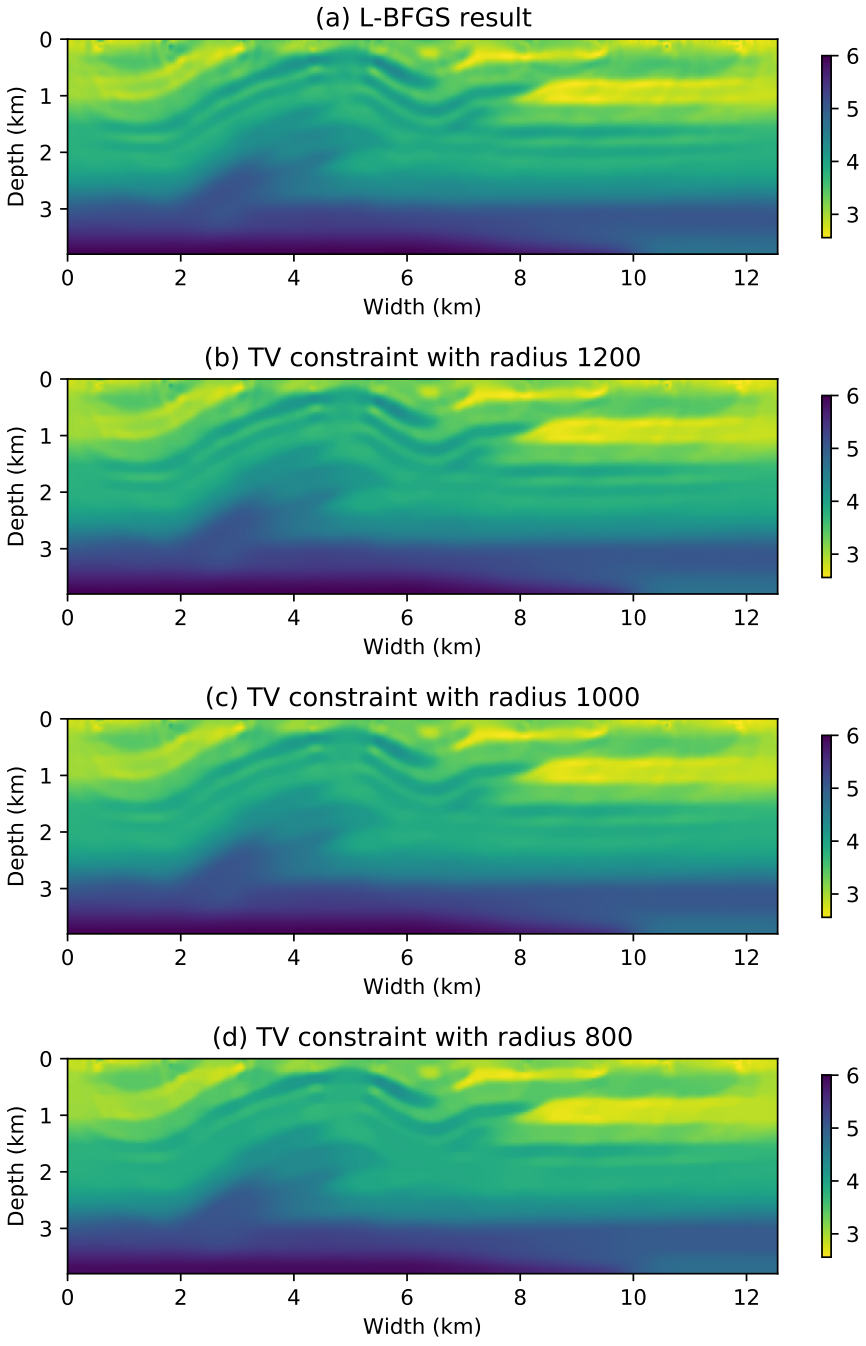


Figure 6.7: (a): Unconstrained result. (b): The inverse result with box constraint and total variation constraint with radius 1200. (c): The inverse result with box constraint and total variation constraint with radius 1000. (d): The inverse result with box constraint and total variation constraint with radius 800.

# Chapter 7

# Conclusions and future studies

## 7.1 Conclusions

As discussed in Chapter 1 and Chapter 2, the full waveform inversion (FWI) is a promising and powerful technique for the seismic inverse problem. However, how to achieve an accurate inverse result that is close to the true model is still a challenging task. In this dissertation, we focus on developing the optimization schemes for the FWI problem which can improve the inverse result. This study contains two parts, applying the optimal transport based distance to the FWI problem, and developing the optimization scheme which can incorporate multiple a priori information to the inverse problem.

In Chapter 4, the unbalanced optimal transport (UOT) distance is introduced to the objective function as the metric in the FWI problem. Also, a mixed $L^1$/Wasserstein distance is constructed to overcome the mass equality limitation for the optimal transport distance, and the convex properties with respect to shift, dilation, and amplitude change are proved for the proposed distance. Then the proposed mixed distance is also introduced to the FWI problem. Both the value and gradient of the UOT distance and the proposed mixed distance can be evaluated efficiently through the entropy regularization approach of optimal transport problem. The computation methods of the adjoint source with the UOT distance and the mixed $L^1$/Wasserstein distance are also provided. Numerical examples show that compared to the conventional $L^2$ distance, the optimal transport based distance can mitigate the cycle-skipping issue and reduce the non-convexity of the objective function.

Starting with the argument that the more a priori information are known to the model, the better inverse result can be achieved, we suggest transforming the a priori information of the model into convex constraint sets. Then the inverse problem is formulated as a constrained optimization problem, with the feasible set as

the intersection of multiple convex constraint sets. However, the projection algorithms onto the intersection of convex sets are usually worked in an iterative process by generating a convergent sequence. In practice, the iteration process has to be terminated when the stopping criteria are met, this makes the projection algorithm actually an inexact projection. In Chapter 5, a set expanding strategy is developed for the gradient projection methods, and the convergence results are proved under proper assumptions. In Chapter 6, an optimization scheme of scaled gradient projection method with multiple constraints is developed. This scheme can work with multiple convex constraint sets which can have the closed-form projection function or the subgradient projection function. Numerical examples show that the proposed optimization scheme works with several kinds of constraints and is flexible to implement, and the inverse results of the FWI problem can be improved with proper a priori information.

## 7.2 Future studies

Following the current thesis work, I have designed plans for future research works.

Despite the UOT distance and the proposed mixed $L^1$/Wasserstein distance can evaluate the difference between signals with normalization methods, theoretical results are still absent. The optimal transport problem has an intrinsic connection with the continuity equation [11] that is fundamentally different from the hyperbolic system for the signals. This might impede the ability to establish a complete analysis work for the variational problem based on the signals. New theory similar to the optimal transport problem but for the signals (or signed measures) and based on the transport equation instead of the continuity equation can be expected. On the other hand, the proposed mixed $L^1$/Wasserstein distance is well defined for the positive functions (or positive atomic measures). And the application of the mixed $L^1$/Wasserstein distance to the variational problem based on the positive quantities can be expected.

Similar to the scaled gradient projection method, the spectral projected gradient method is another generalization of the gradient projection method [18, 19]. A non-monotone inexact line search is used in the spectral projected gradient method, and the objective function value is allowed to increase temporarily, which often results in faster convergence [18]. A new optimization scheme can be designed with the combination of the spectral projected gradient method and the set expanding strategy proposed in Chapter 5 for the case when the constraint set is the intersection of several convex sets.

There are several a priori information discussed in Chapter 6 for the FWI problem. Only synthetic examples are provided in Chapter 6, and it is still unclear that what a priori information is efficient for improving the inverse results of the FWI problem with real data. When the well-log data is available, each well-log data can be represented as the average value of dozens of intervals of different depths, and this a

priori information can be represented as hyperplane constraint sets. The proposed optimization scheme is expected to largely improve the inversion results when enough well-log data are available. Another research topic is new constraints for multiple physical parameters that can be designed to decrease the cross-talk issue for the multi-parameter FWI problem.

The set expanding strategy developed in Chapter 5 provides a method to analyze the constrained optimization problem:

$$\min_x f(x), \quad \text{such that } x \in X^k, \tag{7.1}$$

where the constraint set $X^k$ is changing along the iteration process and the final constraint set $X = \lim_{k\to\infty} X^k$. This final constraint set $X$ provides the a priori information of the optimization problem. This method is not limited to the expanding sequence as long as the limit of the constraint set sequence $\{X^k\}$ exists. An adaptive changing set sequence can be constructed with the a priori information of the model for the inverse problem. An example is: the feasible set $X$ can be constructed with the well log data which is usually available for a certain target area in the FWI problem. Then, an adaptive sequence $\{X^k\}$ can be built to reduce the cycle-skipping issue and increase the stability of the inverse algorithm. Both theoretical results and practical examples can be studied with this approach.

# Bibliography


[1] Hossein S. Aghamiry, Ali Gholami, and Stéphane Operto. Improving full-waveform inversion by wavefield reconstruction with the alternating direction method of multipliers. *Geophysics*, 84(1):R139–R162, 2019.

[2] Hossein S. Aghamiry, Ali Gholami, and Stéphane Operto. Full waveform inversion with adaptive regularization. *arXiv preprint arXiv:2001.09846*, 2020.

[3] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, pages 1–155. Springer, 2013.

[4] Amsalu Y. Anagaw and Mauricio D. Sacchi. Full waveform inversion with total variation regularization. In *Recovery-CSPG CSEG CWLS Convention*, 2011.

[5] Amsalu Y. Anagaw and Mauricio D. Sacchi. Comparison of multifrequency selection strategies for simultaneous-source full-waveform inversion. *Geophysics*, 79(5):R165–R181, 2014.

[6] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014.

[7] Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[8] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[9] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[10] Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

[11] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[12] Jean-David Benamou and Yann Brenier. Mixed L2-Wasserstein optimal mapping between prescribed density functions. *Journal of Optimization Theory and Applications*, 111(2):255–271, 2001.

[13] Jean-David Benamou, Yann Brenier, and Kevin Guittet. The Monge–Kantorovitch mass transfer and its computational fluid mechanics formulation. *International Journal for Numerical Methods in Fluids*, 40(1-2):21–30, 2002.

[14] Jean-David Benamou, Brittany D. Froese, and Adam M. Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *Journal of Computational Physics*, 260: 107–126, 2014.

[15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[16] Jean-Pierre Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, 114(2):185–200, 1994.

[17] Dimitri P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

[18] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000.

[19] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Inexact spectral projected gradient methods on convex sets. *IMA Journal of Numerical Analysis*, 23(4):539–559, 2003.

[20] Kirk D. Blazek, Christiaan Stolk, and William W. Symes. A mathematical framework for inverse wave problems in heterogeneous media. *Inverse Problems*, 29(6):065001, 2013.

[21] Christian Boehm and Michael Ulbrich. A semismooth Newton-CG method for constrained parameter identification in seismic tomography. *SIAM Journal on Scientific Computing*, 37(5):S334–S364, 2015.

[22] Vladimir I. Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

[23] Chaiwoot Boonyasiriwat, Paul Valasek, Partha Routh, Weiping Cao, Gerard T. Schuster, and Brian Macy. An efficient multiscale method for time-domain waveform tomography. *Geophysics*, 74(6): WCC59–WCC68, 2009.

[24] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, pages 25–36. Springer, 2004.

[25] Carey Bunks, Fatimetou M. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, 1995.

[26] Richard H. Byrd and Jorge Nocedal. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM Journal on Numerical Analysis*, 26(3):727–739, 1989.

[27] Richard H. Byrd, Jorge Nocedal, and Robert B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, 1994.

[28] Emmanuel J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

[29] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[30] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[31] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[32] Yair Censor. Parallel application of block-iterative methods in medical imaging and radiation therapy. *Mathematical programming*, 42(1-3):307–325, 1988.

[33] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

[34] Antonin Chambolle and Jérôme Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288, 2009.

[35] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[36] Guy Chavent. Identification of functional parameters in partial differential equations. In *Joint Automatic Control Conference*, number 12, pages 155–156, 1974.

[37] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[38] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[39] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[40] Eric T. Chung, Tony F. Chan, and Xue-Cheng Tai. Electrical impedance tomography using level set representation and total variational regularization. *Journal of Computational Physics*, 205(1):357–372, 2005.

[41] Robert Clayton and Björn Engquist. Absorbing boundary conditions for acoustic and elastic wave equations. *Bulletin of the Seismological Society of America*, 67(6):1529–1540, 1977.

[42] Patrick L. Combettes. Inconsistent signal feasibility problems: Least-squares solutions in a product space. *IEEE Transactions on Signal Processing*, 42(11):2955–2966, 1994.

[43] Patrick L. Combettes. Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections. *IEEE Transactions on Image Processing*, 6(4):493–506, 1997.

[44] Patrick L. Combettes. Strong convergence of block-iterative outer approximation methods for convex optimization. *SIAM Journal on Control and Optimization*, 38(2):538–565, 2000.

[45] Patrick L. Combettes. A block-iterative surrogate constraint splitting method for quadratic signal recovery. *IEEE Transactions on Signal Processing*, 51(7):1771–1782, 2003.

[46] Patrick L. Combettes and J.-C. Pesquet. Image restoration subject to a total variation constraint. *IEEE Transactions on Image Processing*, 13(9):1213–1222, 2004.

[47] Patrick L. Combettes and Hong Puh. Iterations of parallel convex projections in Hilbert spaces. *Numerical Functional Analysis and Optimization*, 15(3-4):225–243, 1994.

[48] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[49] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. 2014.

[50] Bjorn Engquist and Brittany D. Froese. Application of the Wasserstein metric to seismic signals. *arXiv preprint arXiv:1311.4581*, 2013.

[51] Bjorn Engquist and Andrew Majda. Radiation boundary conditions for acoustic and elastic wave calculations. *Communications on Pure and Applied Mathematics*, 32:313–357, 1979.

[52] Björn Engquist and Yunan Yang. Seismic imaging and optimal transport. *arXiv preprint arXiv:1808.04801*, 2018.

[53] Bjorn Engquist, Brittany D. Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. *arXiv preprint arXiv:1602.01540*, 2016.

[54] Ernie Esser, Lluís Guasch, Felix J. Herrmann, and Mike Warner. Constrained waveform inversion for automatic salt flooding. *The Leading Edge*, 35(3):235–239, 2016.

[55] Ernie Esser, Lluis Guasch, Tristan van Leeuwen, Aleksandr Y. Aravkin, and Felix J. Herrmann. Total variation regularization strategies in full-waveform inversion. *SIAM Journal on Imaging Sciences*, 11 (1):376–406, 2018.

[56] Andreas Fichtner. *Full seismic waveform modelling and inversion*. Springer Science & Business Media, 2010.

[57] Reeves Fletcher and Colin M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1964.

[58] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[59] Brittany D. Froese and Adam M. Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic Monge–Ampère equation in dimensions two and higher. *SIAM Journal on Numerical Analysis*, 49(4):1692–1714, 2011.

[60] Odile Gauthier, Jean Virieux, and Albert Tarantola. Two-dimensional nonlinear inversion of seismic waveforms: numerical results. *Geophysics*, 51(7):1387–1403, 1986.

[61] Peter Gilbert. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1):105–117, 1972.

[62] Tom Goldstein, Min Li, Xiaoming Yuan, Ernie Esser, and Richard Baraniuk. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv preprint arXiv:1305.0546*, 2013.

[63] Gerardo González, Ville Kolehmainen, and Aku Seppänen. Isotropic and anisotropic total variation regularization in electrical impedance tomography. *Computers & Mathematics with Applications*, 74 (3):564–576, 2017.

[64] Richard Gordon, Robert Bender, and Gabor T. Herman. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, 29 (3):471–481, 1970.

[65] Huang Guanghui, Wang Huazhong, and Ren Haoran. Two new gradient precondition schemes for full waveform inversion. In *Nonrecurring Meetings 2011: International Geophysical Conference, Shenzhen, China, November 7–10, 2011*, pages 78–78. Society of Exploration Geophysicists, 2011.

[66] Antoine Guitton. Blocky regularization schemes for full-waveform inversion. *Geophysical Prospecting*, 60(5):870–884, 2012.

[67] Zhaohui Guo and Maarten V. de Hoop. Shape optimization in full waveform inversion with sparse blocky model representations. *Proceedings of the Project Review*, 1:189–208, 2012.

[68] Felix J. Herrmann, Yogi A. Erlangga, and Tim T. Lin. Compressive simultaneous full-waveform simulation. *Geophysics*, 74(4):A35–A40, 2009.

[69] Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*, volume 23. Springer Science & Business Media, 2008.

[70] Guanghui Huang and William W. Symes. Full waveform inversion via matched source extension. In *SEG Technical Program Expanded Abstracts 2015*, pages 1320–1325. Society of Exploration Geophysicists, 2015.

[71] Guanghui Huang, Rami Nammour, and William W. Symes. Full-waveform inversion via source-receiver extension. *Geophysics*, 82(3):R153–R171, 2017.

[72] Guanghui Huang, R. Nammour, William W. Symes, and M. Dollizal. Waveform inversion by source extension: 89th annual international meeting. *Expanded Abstracts, Society of Exploration Geophysicists*, 2019.

[73] Anil K. Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[74] L. Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.

[75] Philip A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[76] Florian Knoll, Kristian Bredies, Thomas Pock, and Rudolf Stollberger. Second order total generalized variation (TGV) for MRI. *Magnetic Resonance in Medicine*, 65(2):480–491, 2011.

[77] Jerome R. Krebs, John E. Anderson, David Hinkley, Ramesh Neelamani, Sunwoong Lee, Anatoly Baumstein, and Martin-Daniel Lacasse. Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, 74(6):WCC177–WCC188, 2009.

[78] P. Lailly. The seismic inverse problem as a sequence of before stack migration: Proc. Conf. on Inverse Scattering, Theory and Applications. *Expanded Abstracts, Philadelphia, SIAM*, 1983.

[79] Jan Lellmann, Dirk A. Lorenz, Carola Schonlieb, and Tuomo Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[80] Arnold Lent and Heang Tuy. An iterative method for the extrapolation of band-limited functions. *Journal of Mathematical Analysis and Applications*, 83(2):554–565, 1981.

[81] R. Michael Lewis and William W. Symes. On the relation between the velocity coefficient and boundary value for solutions of the one-dimensional wave equation. *Inverse Problems*, 7(4):597, 1991.

[82] Da Li, Michael P. Lamoureux, and Wenyuan Liao. Full waveform inversion with unbalanced optimal transport distance. *arXiv preprint arXiv:2004.05237*, 2020.

[83] Da Li, Keran Li, and Wenyuan Liao. Efficient and stable finite difference modelling of acoustic wave propagation in variable-density media. *arXiv preprint arXiv:2003.09812*, 2020.

[84] Dong-Hui Li and Masao Fukushima. On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 11(4):1054–1064, 2001.

[85] Jacques Louis Lions and Enrico Magenes. *Non-homogeneous boundary value problems and applications*, volume 1. Springer Science & Business Media, 2012.

[86] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

[87] Jiangbo Liu, Hervé Chauris, and Henri Calandra. The normalized integration method-an alternative to full waveform inversion? In *Near Surface 2011-17th EAGE European Meeting of Environmental and Engineering Geophysics*, pages cp–253. European Association of Geoscientists & Engineers, 2011.

[88] Grégoire Loeper and Francesca Rapetti. Numerical solution of the Monge–Ampère equation by a Newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, 2005.

[89] Musa Maharramov and Biondo Biondi. Robust joint full-waveform inversion of time-lapse seismic data sets with total-variation regularization. *arXiv preprint arXiv:1408.0645*, 2014.

[90] Edoardo Mainini. A description of transport cost for signed measures. *Journal of Mathematical Sciences*, 181(6):837–855, 2012.

[91] Gary S. Martin, Robert Wiley, and Kurt J. Marfurt. Marmousi2: An elastic upgrade for marmousi. *The Leading Edge*, 25(2):156–166, 2006.

[92] Ludovic Métivier, Romain Brossier, Jean Virieux, and Stéphane Operto. Full waveform inversion and the truncated newton method. *SIAM Journal on Scientific Computing*, 35(2):B401–B437, 2013.

[93] Ludovic Métivier, Romain Brossier, Quentin Merigot, E. Oudet, and Jean Virieux. An optimal transport approach for seismic tomography: Application to 3d full waveform inversion. *Inverse Problems*, 32(11):115008, 2016.

[94] Ludovic Métivier, Romain Brossier, Quentin Mérigot, Edouard Oudet, and Jean Virieux. Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, 205(1): 345–377, 2016.

[95] Ludovic Métivier, Aude Allain, Romain Brossier, Quentin Mérigot, Edouard Oudet, and Jean Virieux. Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach. *Geophysics*, 83(5):R515–R540, 2018.

[96] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[97] Peter Mora. Nonlinear two-dimensional elastic inversion of multioffset seismic data. *Geophysics*, 52 (9):1211–1228, 1987.

[98] Stephen G. Nash. A survey of truncated-Newton methods. *Journal of Computational and Applied Mathematics*, 124(1-2):45–59, 2000.

[99] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[100] N. Ottavy. Strong convergence of projection-like methods in Hilbert spaces. *Journal of Optimization Theory and Applications*, 56(3):433–461, 1988.

[101] Wenyong Pan, Kristopher A. Innanen, Gary F. Margrave, and Danping Cao. Efficient pseudo-Gauss-Newton full-waveform inversion in the $\tau$-p domain. *Geophysics*, 80(5):R225–R14, 2015.

[102] Wenyong Pan, Kristopher A. Innanen, and Wenyuan Liao. Accelerating Hessian-free Gauss-Newton full-waveform inversion via l-BFGS preconditioned conjugate-gradient algorithm. *Geophysics*, 82(2): R49–R64, 2017.

[103] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

[104] Bas Peters and Felix J. Herrmann. Constraints versus penalties for edge-preserving full-waveform inversion. *The Leading Edge*, 36(1):94–100, 2017.

[105] Bas Peters, Brendan R. Smithyman, and Felix J. Herrmann. Projection methods and applications for seismic nonlinear inverse problems with multiple constraints. *Geophysics*, 84(2):R251–R269, 2019.

[106] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[107] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[108] Guy Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28 (1):96–115, 1984.

[109] R. E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006.

[110] R. Gerhard Pratt. Seismic waveform inversion in the frequency domain; part 1, theory and verification in a physical scale model. *Geophysics*, 64(3):888–901, 1999.

[111] R. Gerhard Pratt and Richard M. Shipp. Seismic waveform inversion in the frequency domain, part 2: Fault delineation in sediments using crosshole data. *Geophysics*, 64(3):902–914, 1999.

[112] R. Gerhard Pratt, Changsoo Shin, and G. J. Hick. Gauss–Newton and full Newton methods in frequency–space seismic waveform inversion. *Geophysical Journal International*, 133(2):341–362, 1998.

[113] Lingyun Qiu, Jaime Ramos-Martínez, Alejandro Valenciano, Jan Kirkebø, and Nizar Chemingui. Mitigating the cycle-skipping of full-waveform inversion: An optimal transport approach with exponential encoding. In *SEG 2017 Workshop: Full-waveform Inversion and Beyond, Beijing, China, 20-22 November 2017*, pages 1–4. Society of Exploration Geophysicists, 2017.

[114] Lingyun Qiu, Jaime Ramos-Martínez, Alejandro Valenciano, Yunan Yang, and Björn Engquist. Full-waveform inversion with an exponentially encoded optimal-transport norm. In *SEG Technical Program Expanded Abstracts 2017*, pages 1286–1290. Society of Exploration Geophysicists, 2017.

[115] R. Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton University Press, 1970.

[116] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[117] Leonid I. Rudin and Stanley Osher. Total variation based image restoration with free local constraints. In *Proceedings of 1st International Conference on Image Processing*, volume 1, pages 31–35. IEEE, 1994.

[118] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[119] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[120] Fadil Santosa and William W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[121] Fadil Santosa and William W. Symes. Computation of the hessian for least-squares solutions of inverse problems of reflection seismology. *Inverse Problems*, 4(1):211, 1988.

[122] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[123] Laurent Sirgue and R. Gerhard Pratt. Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *Geophysics*, 69(1):231–248, 2004.

[124] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[125] Christiaan Cornelis Stolk. *On the modeling and inversion of seismic data*. PhD thesis, 2000.

[126] William W. Symes. Mathematics of reflection seismology. *Rice University*, pages 1–85, 1995.

[127] William W. Symes, Huiyi Chen, and Susan E. Minkoff. Full-waveform inversion by source extension: Why it works. In *SEG Technical Program Expanded Abstracts 2020*, pages 765–769. Society of Exploration Geophysicists, 2020.

[128] Albert Tarantola. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49 (8):1259–1266, 1984.

[129] Albert Tarantola. A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics*, 51 (10):1893–1903, 1986.

[130] Tristan van Leeuwen and Felix J. Herrmann. Fast waveform inversion without source-encoding. *Geophysical Prospecting*, 61:10–19, 2013.

[131] Tristan Van Leeuwen and Felix J. Herrmann. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(1):661–667, 2013.

[132] Tristan van Leeuwen and Felix J. Herrmann. A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, 32(1):015007, 2015.

[133] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[134] Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[135] Chao Wang, David Yingst, Paul Farmer, and Jacques Leveille. Full-waveform inversion with the reconstructed wavefield method. In *SEG Technical Program Expanded Abstracts 2016*, pages 1237–1241. Society of Exploration Geophysicists, 2016.

[136] Alan Geoffrey Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[137] Yunan Yang. *Optimal transport for seismic inverse problems*. PhD thesis, 2018.

[138] Yunan Yang and Björn Engquist. Analysis of optimal transport and related misfit functions in full-waveform inversion. *Geophysics*, 83(1):A7–A12, 2018.

[139] Yunan Yang, Björn Engquist, Junzhe Sun, and Brittany F. Hamfeldt. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62, 2018.

[140] Peng Yong, Wenyuan Liao, Jianping Huang, and Zhenchuan Li. Total variation regularization for seismic waveform inversion using an adaptive primal dual hybrid gradient method. *Inverse Problems*, 34(4):045006, 2018.

[141] Peng Yong, Jianping Huang, Zhenchun Li, Wenyuan Liao, and Luping Qu. Least-squares reverse time migration via linearized waveform inversion using a Wasserstein metric. *Geophysics*, 84(5):S411–S423, 2019.

[142] Peng Yong, Wenyuan Liao, Jianping Huang, Zhenchun Li, and Yaoting Lin. Misfit function for full waveform inversion based on the Wasserstein metric with dynamic formulation. *Journal of Computational Physics*, 399:108911, 2019.

[143] Dan C. Youla and Heywood Webb. Image restoration by the method of convex projections: Part 1 theory. *IEEE Transactions on Medical Imaging*, 1(2):81–94, 1982.

[144] Pan Zhang, Liguo Han, Zhuo Xu, Fengjiao Zhang, and Yajie Wei. Sparse blind deconvolution based low-frequency seismic data reconstruction for multiscale full waveform inversion. *Journal of Applied Geophysics*, 139:91–108, 2017.

[145] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34, 2008.