UNIVERSITY OF CALIFORNIA

Los Angeles

The Structure of Inverse Problems and Unnormalized Optimal Transport

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Michael A. Puthawala

2019

ABSTRACT OF THE DISSERTATION

The Structure of Inverse Problems and Unnormalized Optimal Transport

by

Michael A. Puthawala
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2019
Professor Stanley J. Osher, Chair

In this thesis we consider the solution of inverse problems, especially the components of a numerical inversion, and detection of forward operator error by the use of an extension optimal transport that accepts unnormalized arguments. We improve the inversion in [42] in both speed and quality of reconstruction and motivated by the desire to improve reconstruction on experimental data we propose a method for fixing forward operator error. We introduce a new tool called the *s*tructure, based on the Wasserstein distance, and propose the use of this to diagnose and remedy forward operator error. Finally we extend the work of [5] and develop an Unnormalized Wasserstein distance measures the distance between two functions of possibly different integral.

The dissertation of Michael A. Puthawala is approved.

Cory D. Hauck

Wotao Yin

Luminita Aura Vese

Christopher R. Anderson

Stanley J. Osher, Committee Chair

University of California, Los Angeles

2019

to my mother, father, and the memory of Joyce McLaughlin.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my parents Mary and Anwer as well as siblings Christy, Stephanie, Daniel, Emily and Joy for their boundless support, infinite patience and invaluable friendship.

I would also like to thank all of the friends that I've made throughout graduate school. Especially I would like to thank Shea Cheng and Chichi Huang for their very close friendship, Denali Molitor for sharing so many memorable meals, Stephanie Wang for the exciting times that we share especially during the first couple of (very stressful) years of graduate school, Zach and Katie Boyd for their warm understanding and openness, Allen David Boozer and Jeremy Brightbill for their extremely stimulating discussions over innumerable Covel lunches.

Finally I would like to thank all of the faculty who have shaped me though my schooling. I would especially like to thank. Dr. Christopher Anderson for always being available to chat about both life and academics throughout my entire graduate career. Dr. Cory Hauck for the time that he spent going through countless iterations of the struc paper and taking his time to carefully and purposely review everything that I wrote. Also Dr. Olga Radko for running the Los Angeles Math Circle always having my back. Finally my advisor Dr. Stanley Osher for giving my so much mathematical guidance throughout my time at UCLA, starting from the open house right up to and including graduation.

VITA

2013, 2014 Summer Research Inter, MIT Lincoln Laboratory, Lexington, MA. I worked in group 10-05, Airborne Radar Systems and Techniques, under Dr. Jennifer Watson.

2014 B.S. Mathematics, Rensselaer Polytechnic Institute, Troy, NY.

2016 M.A. Applied Mathematics, University of California, Los Angeles, CA.

2016, 2017 Summer Research Intern, Oak Ridge National Laboratory, Oak Ridge, TN. I worked under Dr. Cory Hauck in the Computational Applied Math (CAM) group on an imaging problem in fusion energy.

2018 Technical Intern, Google LLC, Venice CA. I worked under the supervision of Dr. Nathan Grigg in a team which worked on the back-end algorithms used to forecast Google's Ads.

2019 (Expected) Ph.D. Applied Mathematics, University of California, Los Angeles, CA. Adviser, Dr. Stanley Osher.

PUBLICATIONS

Cory D. Hauck, Stanley J. Osher and Michael A. Puthawala. *Diagnosing Forward Operator Error Using Optimal Transport.* Submitted for Publication.

Wilfrid Gangbo, Wuchen Li, Stanley J. Osher, and Michael A. Puthawala. *Unnormalized Optimal Transport* Submitted for Publication.

CHAPTER 1

Introduction and Overview

Inverse problems are some of the most well studies subdisciplines of applied math. The core of every inverse problem is to recover the state of some physical quantity or system given observable measurements of that system. Such problems are encountered by physicists doing X-ray crystallography, doctors who take CT scans or oil firms that search for oil underground. The operation that takes as input the state of the system and produces a measurements is called the forward operator and it is this operator whose inversion is required. Often the forward operator does not have a direct inverse or else said inverse if not (feasibly) computable or is otherwise not useful. These problems are often overcome by finding and approximate inverse that exhibits desirable mathematical properties.

Given a discrete forward operator $L \cdot \mathbb{R}^n \to \mathbb{R}^m$, a noise contaminated measurement $b + \eta \in \mathbb{R}^m$ the task is to recover an approximate reconstruction $\tilde{u} \in \mathbb{R}^n$. One way to do this is to solve for

$$\tilde{u} = \underset{v \in \mathbb{R}^n}{\operatorname{argmin}} \|L(v) - (b + \eta)\|^2 + \Phi(v) \tag{1.1}$$

where $\Phi \colon \mathbb{R}^n \to \mathbb{R}^m$ is some regularizer chosen so that when v exhibits undesirable properties (perhaps e.g. large norm, sharp corners or large support) $\Phi(v)$ is large. This work begins with the solution of Eq. 1.1 for the specific problem of Tokamak imaging, specifically the efficient solution of the system and choice of regularizer.

We investigate consider Eq. 1.1 when the forward operator may not be given accurately. We introduce a new tool called the *s*tructure, based on the Wasserstein-1 distance, and propose the use of this to diagnose and remedy forward operator error. Computing the structure turns out to use an easy calculation for a Euclidean homogeneous degree one distance, the Earth Mover's Distance, based on recently developed algorithms. The structure is proven to distinguish between noise and signals in the residual and gives a plan to help recover the true direct operator in some interesting cases. We expect to use this technique not only to diagnose the error, but also to correct it, which we do in some simple cases presented below.

Finally we take this work and propose an extension of the computational fluid mechanics approach to the Monge-Kantorovich mass transfer problem, which was developed by Benamou-Brenier in [5]. Our extension allows optimal transfer of unnormalized and unequal masses. We obtain a one-parameter family of simple modifications of the formulation in [5]. This leads us to a new Monge-Ampére type equation and a new Kantorovich duality formula. These can be solved efficiently by, for example, the Chambolle-Pock primal-dual algorithm [8]. This solution to the extended mass transfer problem gives us a simple metric for computing the distance between two unnormalized densities.

1.1 The GSVD and Tokamak imaging

The purpose of this project was to improve upon the work on done in [42]. In broad terms, the task was to find a way to improve the quality of reconstruction of an inverse problem in Tokamak imaging. Specifically to reconstruct a three dimensional plasma bulk using two dimensional measurements from a pin-hole camera. On its surface this problem is at least difficult or perhaps intractable because one has to reconstruct the three dimensional bulk using a two dimensional measurement. This problem is saved by the physical symmetry of the plasma bulk within the Tokamak. In accordance with plasma physics, one can deduce that the bulk is symmetric along magnetic field lines. This implies that the bulk is uniquely determined from a set of unknowns living on a two dimensional manifold. Therefore, the original problem is indeed tractable.

For this problem the forward operator has two components. The first is derived from the geometry and optics of the measurement apparatus. The other is given by the assumption that plasma is symmetric about prescribed magnetic field lines. We did not work on the construction of the forward operator, my task was to treat this operator opaquely and improve reconstruction of this problem by applying different regularizations.

We found that we were able to improve both the quality of the reconstruction by a fair bit, and the speed of the reconstruction by an extremely large amount (a factor of over 100 times). Further we rediscovered a number of useful properties about the Generalized Singular Value Decomposition (GSVD) and also discovered a new role for it in non-linear regularization. We improved the quality of reconstruction by changing the regularizer. We replaced a Laplacian regularizer with Dirichlet boundary conditions with one with both Dirichlet and Neumann boundary conditions. The speed improvement came from respecting the sparsity of the forward operator and regularizer. The previous state-of-the-art method used the GSVD to take a dense factorization of a sparse linear system to solve said system. We replaced this factorization with an iterative procedure to solve said system. This alone speeds up the reconstruction by a factor of over 100 times. The reconstruction speed and quality results are several steps ahead of the previous state-of-the-art however they are as of yet unpublished. In Chapter 2, we describe all of the particulars of the problem and my improvements. Furthermore we, explain how the GSVD can be used to great effect in modern inverse problems.

During the course of solving this Tokamak problem we discovered that the main thing that was standing in the way of great experimental reconstruction was forward operator error. We discovered that the experimentalists had mismodeled their forward operator! If there is even a modest error in the forward operator, then even the most powerful inverse procedures falter. This problem occurs not only in Tokamak imaging but any kind of inverse problems that are prone to miscalibration or other kinds of mismodeling. This realization led me to shift my focus from solving specific inverse problems to trying to devise a scheme that could be used to diagnose forward operator error.

1.2 Optimal Transport and Inverse Problems

Suppose that one is solving an inverse problem where the forward operator present in the measurement is unknown, but one has a parameterization of candidate forward operators $\{L_\theta\}_{\theta \in \Theta}$ for some parameter set Θ. For a given model (i.e. choice of $\theta \in \Theta$) the reconstruction is given as the solution to the variational problem

$$\tilde{u}_\theta = \tilde{L}_\theta^{-1} b \equiv \underset{v}{\operatorname{argmin}} \, \|L_\theta v - b\|_2^2 + \Phi(v) \tag{1.2}$$

where $L_{\hat{\theta}} u = b$ where u is the ground truth, and some regularizing functional Φ. For almost all inverse problems there is some noise in the measurement, so typically $b + \eta$ is given where η is some corrupting noise. In that case, the problem becomes

$$\tilde{u}_{\theta,\eta} = \tilde{L}_\theta^{-1}(b + \eta) \equiv \underset{v}{\operatorname{argmin}} \, \|L_\theta v - b - \eta\|_2^2 + \Phi(v). \tag{1.3}$$

We denote the residual in the reconstruction as

$$r_{\theta,\eta} = (b + \eta) - L_\theta(\tilde{u}_{\theta,\eta}). \tag{1.4}$$

In Chapter 3 we show that the residual is made up of two distinct components. The first is due to operator error (i.e. a mismatch between θ and $\hat{\theta}$) and the second is due to noise and regularization. Chapter 3 details the development of a functional that we call the structure and its implications for forward operator error detection. The structure is defined as

$$\operatorname{struc}[f] = \operatorname{EMD}\left(\max\left(f - \int f dx, 0\right), \max\left(\int f dx - f, 0\right)\right) \tag{1.5}$$

where the Earth Mover's Distance (a.k.a. Wasserstein-1 distance) is defined as

$$
\begin{aligned}
\mathrm{EMD}(\rho_1, \rho_2) = \min_{m} \int_{\Omega} \|m(x)\|_2 \, dx,& \\
\text{subject to:} \quad \nabla \cdot m(x) + \rho_2(x) - \rho_1(x) = 0,& \\
m(x) \cdot n(x) = 0 \quad \forall x \in \partial\Omega&
\end{aligned}
\tag{1.6}
$$

when $\int \rho_0 dx = \int \rho_1 dx$. The structure satisfies many theoretically important properties, the most important of which are

1. The structure of i.i.d. discrete noise is asymptotically small
2. The structure of a non-constant piecewise continuous function (of note, smooth artifacts introduced through forward operator mismatch) is not small, and is asymptotically bounded away from zero.

These two facts together with the observation concerning the two components of the residual of an inverse problem suggest that the structure could be a useful tool for measuring only the part of the residual that comes from mismodeling the forward operator. Therefore, it is reasonable to expect that the structure of the residual of an inverse problem can be used as a proxy for the correctness of a forward operator. In other words that

$$\hat{\theta} \approx \operatorname*{argmin}_{\theta \in \Theta} \operatorname{struc}\left[r_{\theta,\eta}\right]. \tag{1.7}$$

Chapter 3 contains the further development of this idea as well as the proofs of all necessary facts and a battery of numerical experiments to show that this conclusion is numerically valid. Further, although it is not proven in Chapter 3, it is also true that $\operatorname{struc}[f]$ is differentiable w.r.t. f.

Through the definition of the structure we have extended the utility of the Wasserstein-1 distance to the unnormalized case when $\int_\Omega \rho_0 dx \neq \int_\Omega \rho_1 dx$ and when $\rho_0(x), \rho_1(x) \not\geq 0$ for all $x \in \Omega$. That is

$$W_1(\rho_0, \rho_1) = \operatorname{struc}\left[\rho_0 - \rho_1\right] \quad \text{when} \quad \int_\Omega \rho_0 dx = \int_\Omega \rho_1 dx \text{ and } \rho_0, \rho_1 \geq 0. \tag{1.8}$$

Note that the l.h.s. of Eq. 1.8 is defined even when the constraints are violated. Leveraging this we define an unbalanced Wasserstein-1 with no constraints in the inputs as

$$UW_1(\rho_0, \rho_1) = \mathrm{struc}\left[\rho_0 - \rho_1\right] + \left|\int_\Omega \rho_0 - \rho_1 dx\right|. \tag{1.9}$$

We realized that in this way the structure was a nice extension of the Wasserstein-1 distance, and further a similar trick one could generalize the Wasserstein-p distance in a similar way to get a similarly natural generalization. Additionally this generalization preserves many of the desirable theoretical properties of the Wasserstein-p distances including a Lagrange formulation, a Monge-Ampére equation and Kantorovich duality. Finally this distance can actually be easily computed as well.

1.3 Unbalanced Wasserstein Distance

Building off of the developments in the [36], we propose an extension of the Wasserstein distance to an unnormalized case. In the Benamou-Brenier [5] formulation of the $W_2(\rho_0, \rho_1)$ distance,

$$W_2(\rho_0, \rho_1) = \min_{u,m} \left(\int_\Omega \int_0^1 \frac{\|m\|_2^2}{\mu} dt dx\right)^{1/2} \tag{1.10}$$

where the minimum is taken w.r.t. all μ and m which satisfy

$$\mu_t(t,x) + \nabla \cdot m(t,x) = 0, \tag{1.11}$$

$$\mu(0,x) = \rho_0(x), \mu(1,x) = \rho_1(x) \text{ on } \Omega \tag{1.12}$$

$$m \cdot n = 0 \text{ on } \partial\Omega \tag{1.13}$$

where n is the unit normal vector on $\partial\Omega$. Integrating Eqn. 1.11 in space and time in time yields

$$\int_\Omega \rho_0 dx = \int_\Omega \rho_1 dx \tag{1.14}$$

and so any choice of μ or m that satisfies the constraints must necessarily conserve mass. Thus, to extend the W_2 distance to be well defined when $\int_\Omega \rho_0 dx \neq \int_\Omega \rho_1 dx$, we must change the constraint. One natural thing to change is to add another term f on the right side of 1.11 so that it becomes

$$\mu_t(t,x) + \nabla \cdot m(t,x) = f. \tag{1.15}$$

For our extension, we choose to add an f such that $f = f(t)$. Finally, we also modify Eqn. 1.10 to minimize over all f which satisfy the constraint, as well as bias the objective functional to prefer f which are close to zero. Putting it all together our Unnormalized Wasserstein-2 distance is

$$UW_2\rho_0, \rho_1) = \min_{\mu,m,f} \left(\int_\Omega \int_0^1 \frac{\|m\|_2^2}{\mu} dtdx + \frac{|\Omega|}{\alpha} \int_0^1 |f(t)|^2 \right)^{1/2} \tag{1.16}$$

where μ, m and f must satisfy

$$\mu_t(t,x) + \nabla \cdot m(t,x) = f(t), \tag{1.17}$$

$$\mu(0,x) = \rho_0(x), \mu(1,x) = \rho_1(x) \text{ on } \Omega \tag{1.18}$$

$$m \cdot n = 0 \text{ on } \partial\Omega. \tag{1.19}$$

Let's note a few things about this extension of the W_2 distance.

1. If $\int_\Omega \rho_0 dx = \int_\Omega \rho_1 dx$ and μ, m satisfy the constraints in Eqn. 1.11 - 1.13 then μ and m also satisfy Eqn. 1.17 - 1.19. In other words, when $W_2(\rho_0, \rho_1)$ are defined then

$$UW_2(\rho_0, \rho_1) \leq W_2(\rho_0, \rho_1). \tag{1.20}$$

2. The W_2 distance is sometimes motivated with use of the following analogy. A certain quantity of sand is distributed according to ρ_0. The goal is to move the sand to a

new configuration ρ_1 while doing the minimal possible work, where work is defined as the sum of sand moved times the distance that it is moved. $W_2(\rho_0, \rho_1)$ is the minimum amount of work that *must* be done to move the sand. $UW_2(\rho_0, \rho_1)$ has a similar physical intuition involving snow. Suppose instead that one wanted to move snow from one place to another, but with the additional flexibility of the weather. At any time snow can fall from the sky (i.e. $f(t) > 0$) uniformly or melt (i.e. $f(t) < 0$) on the ground. The question then becomes how could you move about snow if you could also control the snowfall for a cost.

3. As a function of α, $UW_2(\rho_0, \rho_1)$ is monotone decreasing, and (formally) one can easily see that for all $\rho_0, \rho_1 \in L_1(\Omega)$ if $\int_\Omega \rho_0 dx = \int_\Omega \rho_1 dx$ then. Further, we have found numerically that as $\alpha \to \infty$, $UW_2(\rho_0, \rho_1)$ does converge, and converges to a non-zero value provided that $\rho_0 - \rho_1$ is not identically constant.

4. The W_2 has a bevy of formulations which make it possible to analyze it from any number of possible angles. UW_2 also has many of those formulations, and thus many of the theoretical results about the formulations extend to out unbalanced formulation as well. Examples include a Lagrange formulation, a Monge-Ampére equation and Kantorovich duality.

CHAPTER 2

The Generalized Singular Value Decomposition and Split Bregman

2.1 Introduction

The purpose of this chapter is to expand upon the work described in [42]. In that work, the authors solve an inverse problem arising in the field of plasma imaging by the use of Tikhonov regularization. My work is concerned with improving the quality of reconstruction on that problem by using different regularizers. We find that one can achieve higher quality reconstructions by using Tikhonov regularization with Laplace regularization with special boundary conditions. We rediscovered an important application of the Generalized Singular Value Decomposition (GSVD) when solving both linearly and non-linearly regularized inverse problems. Specifically we show that when one is solving an inverse problems with a dense forward operator and dense regularization, the GSVD can be used to quickly solve the resulting systems. Finally we explain how we were able to increase the speed of reconstruction in the case of [42] by over a hundredfold by using iterative methods.

2.2 The problem

In [42], the authors analysed the inverse problem which arises from imaging plasma within a DIII-D Tokamak. The plasma itself occupies a three dimensional volume (called the plasma bulk) within a toroidal reactor chamber. The bulk is imaged using a pinhole camera. By using plasma physics one can exploit symmetries of the plasma within the chamber and reduce the dimension of the problem of computing the entire three dimensional bulk to the problem of a two dimensional cross section. From there, the problem becomes a standard inverse problem. Given an operator L which represents the projection of the plasma's emission onto the pinhole camera s, recover the underlying emission ϵ subject to

$$Lu = b \tag{2.1}$$

$L \in \mathbb{R}^{m \times n}$ where $m > n$, and $\text{rank}(L) < n$. In order to solve this problem we must regularize it in some way. The state-of-the-art approach as described in [42] is to use Tikhonov regularization in order to regularize the minimization problem where the choice of regularizer is a discrete differential operator C. Wingen et al. minimize the problem

$$J_w(v) = \|Lv - b\|_2^2 + \frac{1}{\lambda}\|Cv\|_2^2 \tag{2.2}$$

where $\lambda \in \mathbb{R}^+$, is a parameter that controls the strength of the regularization and $C \in \mathbb{R}^{o \times n}$ is a regularizing matrix. The authors choose C to be a discrete approximation to the Laplacian, so $o = n$. The minimizer of $J_w(v)$ is given by the normal equations

$$(\lambda^2 L^* L + C^* C)v = \lambda L^* b. \tag{2.3}$$

These are then solved to obtain the actual reconstruction, the minimizer of Eqn. 2.2. One can use the GSVD (which is defined & described in section 2.4.1) to great effect on this problem to speed up the inversion. Further, one can also use the GSVD combined with L-curve theory to choose a value of λ which is optimal. This avoids the difficult process of manual calibration.

2.3 Non-linear Regularization

Motivated by the ROF model [37] our approach is to regularize J with an L_1 type term, so our functional to minimize becomes

$$J_P(v) = \|Lv - b\|_2^2 + \frac{1}{\lambda}\|Cv\|_1 \tag{2.4}$$

The L_1 norm is sparsity inducing, and so a minimizer to this problem will produce a v such that Cv is mostly zero. This in turn may produce desirable effects on the final picture, such as preserving sharp edges (if C is gradient like) or linear changes changes in v (if C is Laplacian like).

The cost of this approach is that this minimization problem eq. (2.4) is harder to solve than eq. (2.2). We also don't have L-curve theory to guide the choice of λ, the analogous split Bregman 'step size' parameter. Determining a good choice of λ necessitates either good heuristics or computationally expensive calibration. Later in this report we show how one can reuse work from run to run via the GSVD, which mitigates the cost of doing many runs with different values of λ. In order to compute a minimizer of eq. (2.4), we used the split Bregman method introduced in [20].

2.4 The Generalized Singular Value Decomposition (GSVD)

2.4.1 Definition

Introduced in [33], the GSVD is defined as follows. Given two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{o\times n}$ such that $\ker(A) \cap \ker(B) = \emptyset$, then there exists matrices $U \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{o\times n}$ $S, C, X \in \mathbb{R}^{n\times n}$ such that X is invertable and the following properties hold:

- $A = UCX^T$
- $B = VSX^T$
- $U^TU = V^TV = I$
- $C =$diag$(c_1, c_2, \ldots, c_n)$ where $1 \geq c_1 \geq c_2 \geq \cdots \geq c_n \geq 0$
- $S =$diag$(s_1, s_2, \ldots, s_n)$ where $0 \leq s_1 \leq s_2 \leq \cdots \leq s_n \leq 1$

- $S^T S + C^T C = I$.

Even if the kernals do not have trivial intersection, one can still define the GSVD differently so that it exists [3], but for simplicity we use this definition. Requiring that $\ker(A) \cap \ker(B) = \emptyset$ is the same as assuming that $J(x) = \|Ax - b\|_p + \|Bx - c\|_q$ is strictly convex where $b \in \mathbb{R}^m, c \in \mathbb{R}^o$ and $1 \leq p, q < \infty$.

The authors of [33] point out that the GSVD could be useful for computing a solution of $AB^{-1}x = b$ where B is square. The utility here is that S is a diagonal matrix, so as long as B is non-singular, one can solve the $AB^{-1}x = UCS^{-1}V^*x = b$ without having to actually invert B directly. Another use of the GSVD which is very useful in our application is that if one has precomputed the GSVD of a (A, B), then one can solve the system:

$$(\lambda^2 A^* A + B^* B)x = b \tag{2.5}$$

very quickly, as $(\lambda^2 A^* A + B^* B) = X(\lambda^2 C^2 + S^2)X^*$. This trick is also useful whenever one is solving a system of the form

$$(\lambda^2 F + G)x = b \tag{2.6}$$

provided that F and G are symmetric positive definite by computing the matrix square root of F and G.

2.4.2 Scaling property

Given that the matrix pair (A, B) has GSVD $A = UCX^T$ and $B = VSX^T$, then for any $\lambda \in \mathbb{R} > 0$ the matrix pair $\lambda A, B$ has a GSVD $A = UC_\lambda X_\lambda^T$, $B = VS_\lambda X_\lambda^T$. where

$$D_\lambda = diag(\sqrt{s_1^2 + \lambda^2 c_1^2}, \sqrt{s_2^2 + \lambda^2 c_2^2}, \ldots, \sqrt{s_n^2 + \lambda^2 c_n^2}) \tag{2.7}$$

$$C_\lambda = \lambda C D^{-1} \tag{2.8}$$

$$S_\lambda = S D^{-1} \tag{2.9}$$

$$X_\lambda = DX \tag{2.10}$$

2.5 GSVD and Split Bregman

One can also leverage the GSVD in an interesting way in order to speed up one of the intermediate steps of split Bregman as the cost of some precomputation. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{o\times n}$ have a GSVD given by $A = UCX^*, B = VSX^*$ as described in section 2.4.1, then the solution to the equation:

$$(\lambda^2 A^* A + B^* B)u = b \tag{2.11}$$

is

$$u = X^{-*}(\lambda^2 C^* C + S^* S)^{-1} X^{-1} b. \tag{2.12}$$

In order to implement split Bregman, one has to repeatedly solve

$$(\lambda^2 L^* L + C^* C)u^{k+1} = \lambda L^* s + C^*(d^k - b^k). \tag{2.13}$$

. If we apply the observation in eq. (2.12), and expand the left hand side in terms of the GSVD of (L, C) then we obtain

$$u^{k+1} = \gamma + B(d^k - b^k) \tag{2.14}$$

where

$$\gamma = \lambda X^{-*}(\lambda^2 C^* C + S^* S)^{-1} C U^* s \tag{2.15}$$

$$B = X^{-*}(\lambda^2 C^* C + S^* S)^{-1} S V^*. \tag{2.16}$$

Note that for any choice of λ, once can reuse the same GSVD of (L, C) by exploiting the scaling property of the GSVD in Section 2.4.2

For a given Bregman run γ can always be computed ahead of time at the cost of 2 dense matrix by vector multiplies (N.B. C and S are diagonal matrices). If one intends to do large number of Bregman iterations, then one can also compute B ahead of time at the cost of 2 dense matrix by matrix multiplications. If one is doing a small number of Bregman iterations, then it is probably favorable to avoid computing B ahead of time.

This will also work in solving systems of the form $(\lambda F + G)u = b$ quickly whenever one can find a way to factor $F = A^*A, G = B^*B$. This includes cases where F, C are positive semi-definite using the matrix square root, and also certain differential operators such as the Laplacian and Biharmonic operators which can be factorized analytically.

Finally, in section section 2.4.2 we state a scaling property of the GSVD which can be used to derive the GSVD of the matrix pair $(\lambda A, B)$ from the pair (A, B). This fact was first shown in [23] this fact could potentially be very useful in applications where one must compute the GSVD of families matrices.

2.6 Results

To generate out results, we used both the observation operator (L) and simulated emission (u) given to us by the authors of [42]. A plot of the emission is shown in section 2.6. Notice that the left boundary of the emission is highly irregular. These irregularities preclude one from using techniques like the FFT that rely on the regularity of the domain.

For all of these numerical experiments, we will use the purely additive noise model. Let u be the true emission, and $b = Lu$ be the exact measurement by the pinhole camera. The problem we will be solving is computing $\tilde{u}$ given $b + c\eta$ where n is white noise and $c = \frac{\|b\|_2}{\|\eta\|_2}$. For the first few plots, we will use a SNR of 10. Below each reconstruction are the weights used to generate the reconstruction and the errors. For each reconstruction, a rudimentary search for optimal parameters was done. The L_1 error is $\|\tilde{u} - u\|_1 = \frac{1}{dx \cdot dy} \sum_{i,j} |\tilde{u}_{i,j} - u_{i,j}|$, and the L_2 error is $\|\tilde{u} - u\|_2 = \frac{1}{dx \cdot dy} \sqrt{\sum_{i,j} (\tilde{u}_{i,j} - u_{i,j})^2}$ where dx and dy are the distance between the pixels in the x and y dimension respectively. For all of the following plots, I did exactly 100 Bregman iterations; regardless of convergence.

The below table compares the quality of the reconstruction between the four regularizers.

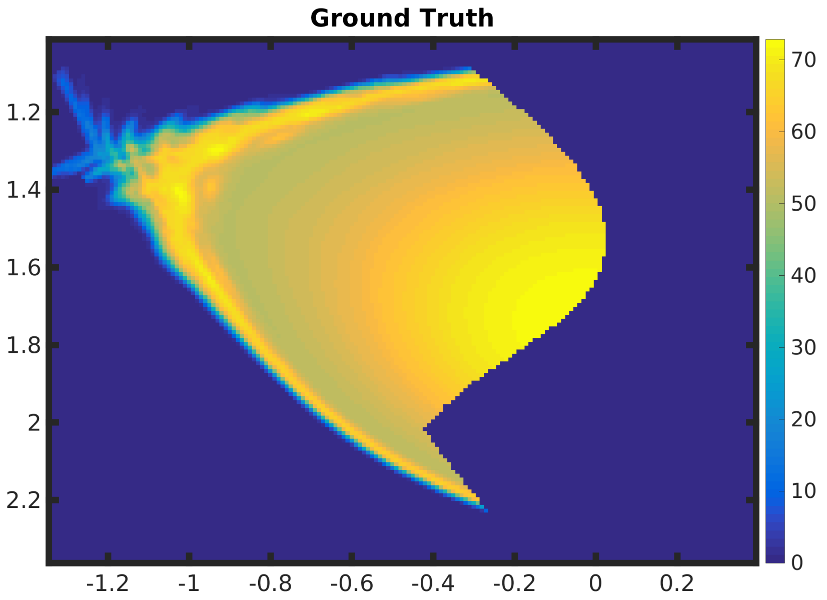


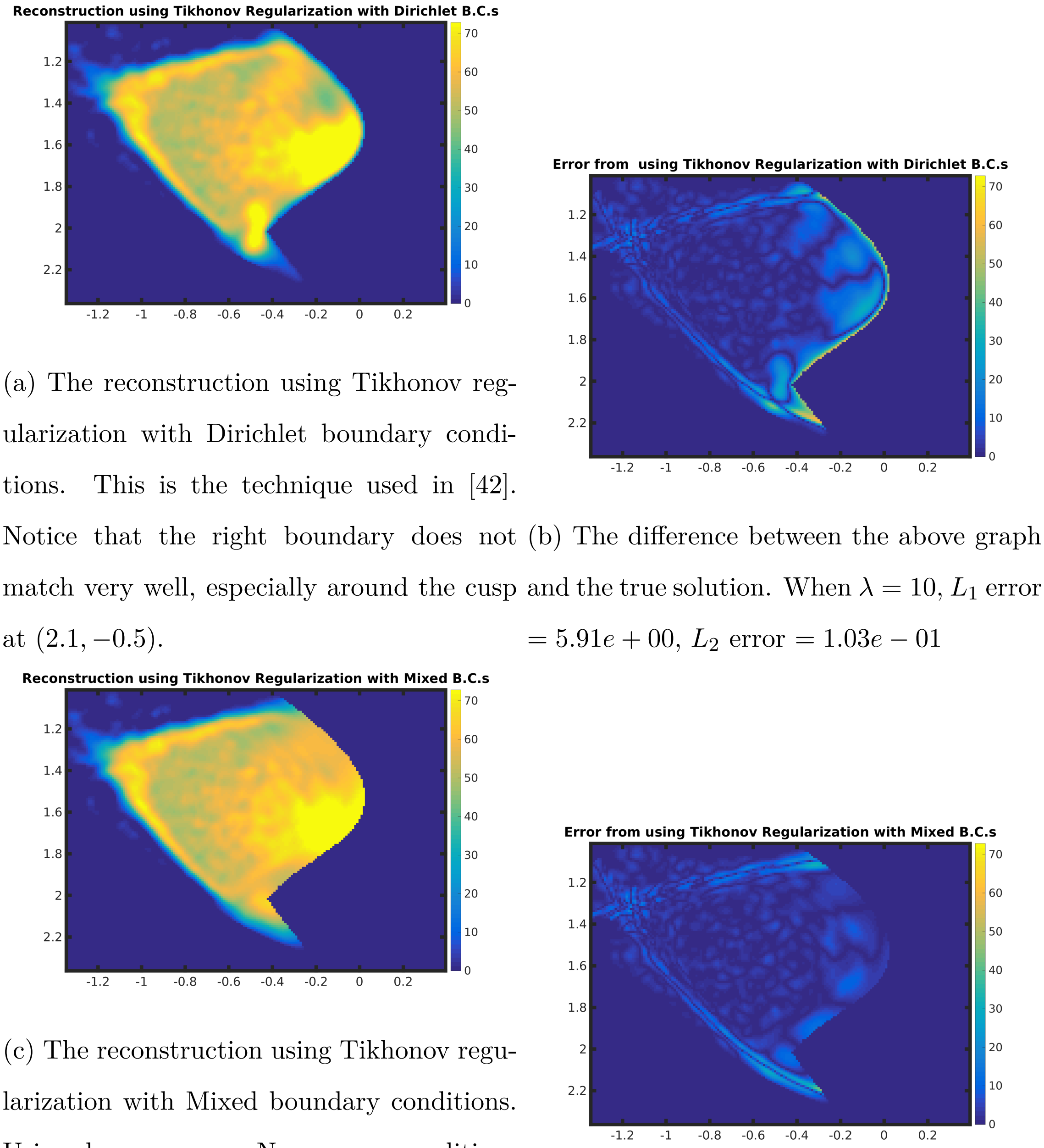
Figure 2.1: The simulated data which was use throughout the numerical experiments.

Notice that for large SNRs, using the L_1 of the Laplacian or total variation both outperform the regularizer used in [42]. As the SNR decreases, the difference between the regularizers decreases. For the true solution, $\|u\|_1 = 4.63e + 01$, $\|u\|_2 = 5.23e - 01$

Regularizer	SNR 100		SNR 10		SNR 5.6	
	L_1 error	L_2 error	L_1 error	L_2 error	L_1 error	L_2 error
L_2 Dirichlet	4.61e+00	9.20e-02	5.91e+00	1.03e-01	8.55e+00	1.20e-01
L_2 Mixed	2.17e+00	4.27e-02	3.41e+00	5.18e-02	7.21e+00	9.55e-02
L_1 Laplacian	2.00e+00	3.79e-02	3.41e+00	5.01e-02	9.08e+00	1.19e-01
TV	3.22e+00	6.40e-02	4.90e+00	8.47e-02	7.85e+00	1.05e-01

2.7 Failure of GSVD for Sparse Problems

In spite of the useful properties of the GSVD, the GSVD is still not a good fit for problems that are highly sparse. Like the SVD, the GSVD does not respect sparsity. If A and B is sparse the various matrices produced by the GSVD are usually not sparse. This presents two problems for both time and computer memory. Consider the case when we solve



(a) The reconstruction using Tikhonov regularization with Dirichlet boundary conditions. This is the technique used in [42]. Notice that the right boundary does not match very well, especially around the cusp at $(2.1, -0.5)$.

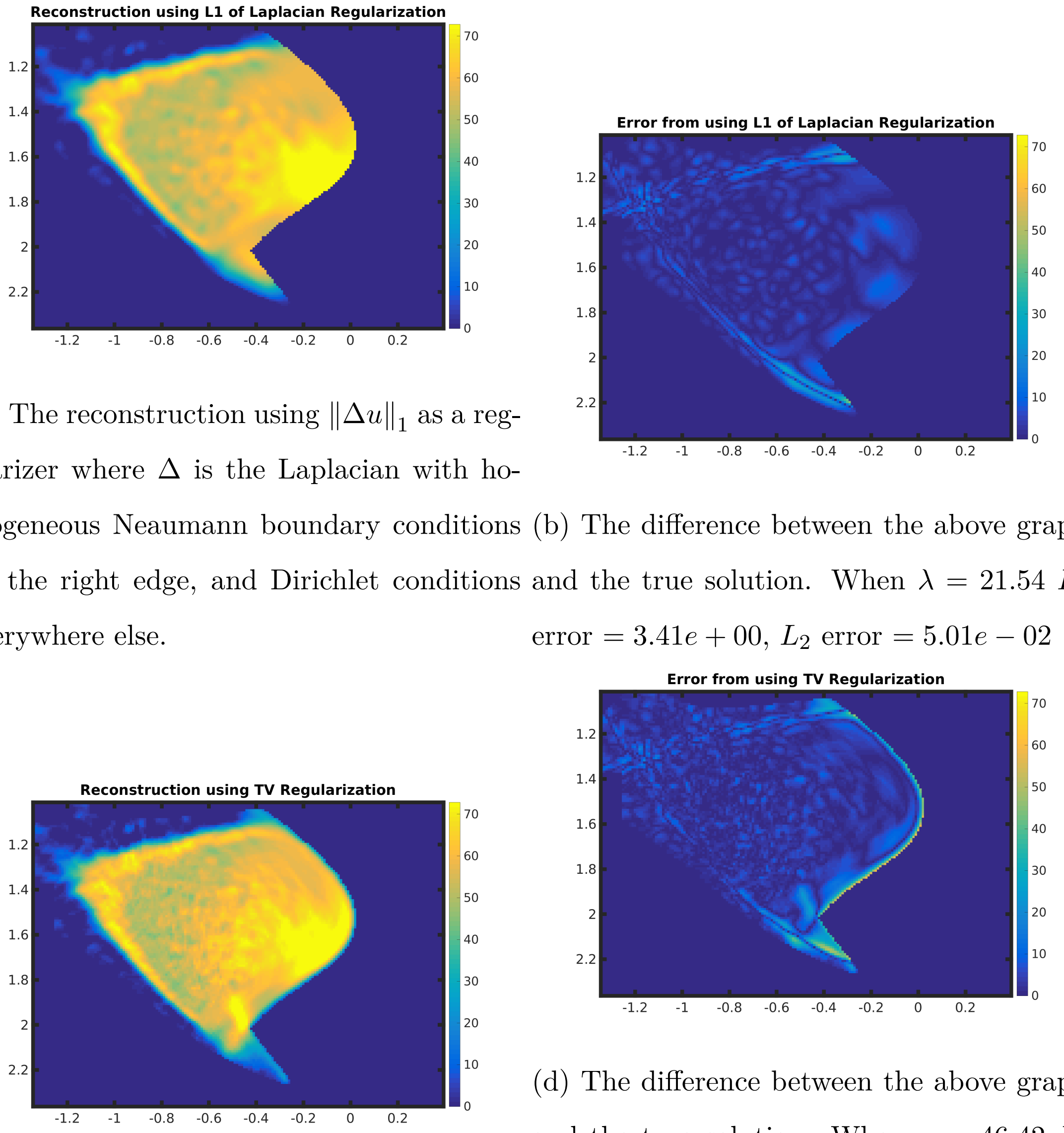
(b) The difference between the above graph and the true solution. When $\lambda = 10$, L_1 error $= 5.91e + 00$, L_2 error $= 1.03e - 01$

(c) The reconstruction using Tikhonov regularization with Mixed boundary conditions. Using homogeneous Neaumann conditions on the right side of the boundary resulted in considerably better reconstruction on that edge.

(d) The difference between the above graph and the true solution. When $\lambda = 10$, L_1 error $= 3.41e + 00$, L_2 error $= 5.18e - 02$

Figure 2.2: Reconstructions with Tikhonov regularization

$$(A^*A + B^*B)x = b. \tag{2.17}$$



(a) The reconstruction using $\|\Delta u\|_1$ as a regularizer where Δ is the Laplacian with homogeneous Neaumann boundary conditions on the right edge, and Dirichlet conditions everywhere else.

(b) The difference between the above graph and the true solution. When $\lambda = 21.54$ L_1 error $= 3.41e + 00$, L_2 error $= 5.01e - 02$

(c) The reconstruction using $\text{TV}(u)$.

(d) The difference between the above graph and the true solution. When $\mu = 46.42$ L_1 error $= 4.90e + 00$, L_2 error $= 8.47e - 02$

Figure 2.3: Reconstructions with L_1 regularization

when $A^*A + B^*B$ has 5% non-zero elements.

1. The time that it takes to compute U, V, X, C and S can be orders of magnitude larger then the cost of solving the system Eq. 2.17 using existing sparse linear solvers. For example, for a $A \in \mathbb{R}^{128\times128}$ Line Integral Operator (the construction is discussed in Section 3.7) where C is the numerical Laplacian solving the system 2.17 using Matlab's build in backslash operator takes 3 seconds, whereas computing the GSVD of A and

B takes about 20 minutes. A factor of about 300. Therefore, one has to solve a lot of linear systems before it is worth paying the amortized cost of one GSVD factorization.

2. If A and B are both square, then the computer memory needed to store U, V and X is 30 times larger. Thus if they are large enough just storing U, V and X can be a real challenge even if A and B are not too large. In the case of [42] the GSVD occupies 140 times more memory than the forward operator and regularizer alone.

For the Tokamak problem in [42] both the forward operator and the differential regularizer were sparse, and so I was able to speed up the reconstruction by a factor of over 100 by replacing the GSVD solve with an iterative solve.

Still, the GSVD is still a useful factorization in the following cases:

1. Either the forward operator or the regularizer is dense. In this case, neither of the above two considerations apply. An example of this would be when the forward operator or regularizer is given as an integral equation.

2. When one particular forward operator and regularizer pair is used many times. Eventually, the amortized cost of computing one GSVD becomes less than many iterative solves.

3. The GSVD can not only be used to do an inversion. It can also be used to regularize 'optimally' [25]. Thus, the GSVD can still be worth using when finding the optimal regularization is difficult.

CHAPTER 3

Diagnosing Forward Operator Error Using Optimal Transport

3.1 Introduction

3.1.1 Motivation

From medical imaging [2] to petroleum engineering [32] to meteorology [7], inverse problems are ubiquitous in science, engineering and mathematics. The goal of such problems is to recover an unknown quantity u given a known forward operator L and measurement b such that $L(u) = b$. In this work we consider the case where L is a linear operator and write $L(u) \equiv Lu$. While this choice facilitates a simple analysis in some places, the computational techniques developed here can be extended to consider non-linear operators.

A considerable amount of work has been dedicated to solving inverse problems for a variety of forward operators, especially when L is linear. Powerful techniques have been developed that perform well in the presence of noise in b, singularities in L and various constraints on the solution u [28].

Despite some great successes in the field of inverse problems, there are still mathematical challenges that are difficult to address. One of these, which is important in a bevy of applications, is the calibration of forward operators. For example, computed tomography (CT) machines are calibrated using known phantoms for which the desired reconstruction is known exactly [39]; in synthetic aperture radar, reflectors provide a known ground truth on which devices and reconstruction algorithms are tuned [17]; and in some plasma imaging problems, the forward model has unknown parameters, and the model itself is possibly incomplete [42].

Often the calibration problem can be formulated mathematically by considering a family of forward operators L_θ, parameterized by $\theta \in \Theta \subset \mathbb{R}^p$, with a unique $\hat{\theta}$ such that $L_{\hat{\theta}}$ best represents the underlying physical system. In other words, there exists a $\hat{\theta}$ such that $L = L_{\hat{\theta}}$ [15, 42]. If $\hat{\theta}$ is estimated poorly, then an accurate approximation of u is often impossible, even with very sophisticated inverse procedures.

The problem of detecting forward operator error is similar to that of blind deconvolution in image processing [9], where the task is to identify a blurring kernel and recover an image from a given blurry signal. The application of the blurring operator with the image can also be represented in the form $Lu = b$ where the action of L gives the convolution with the blurring kernel. One important difference between the calibration problem considered here and the problem of blind deconvolution is that we will be considering overdetermined problems. By overdetermined we mean that in the reconstruction process there are more knowns than unknowns, even if the forward operator has a non-zero nullspace.

3.1.2 Prior Work

Methods for detecting and correcting for errors within the forward operator exist. One approach is total least squares [22], which generalizes the standard least squares method by allowing for error in L. This is expressed by the minimization problem

$$\min_{\mathbf{v},\mathbf{J}} \|\mathbf{L} - \mathbf{J}\|_F^2 + \|\mathbf{b} - \mathbf{J}\mathbf{v}\|_2^2, \tag{3.1}$$

where $\mathbf{L}$ is the matrix representations of L, $\mathbf{b}$ is the vector representation of b, and $\|\cdot\|_F$ is the Frobenius norm.

This approach has the advantage of being relatively easy to analyze, robust under noise in the entries of $\mathbf{L}$ and solvable using standard linear algebra software. However, for calibration problems, the goal is not to remove entry-wise error in $\mathbf{L}_\theta$. Instead we seek a value of $\theta \approx \hat{\theta}$. Total least squares provides good reconstructions when $\mathbf{L}$ is a matrix whose entries are corrupted by noise. However it requires modification in order to be applied to the parametric calibration problem. In particular, adding the requirement $J = L_\theta$ for $\theta \in \Theta$ to Eq. 3.1 make the resulting minimization problem more difficult to solve, and so may require code beyond standard linear algebra software.

Another common approach for calibration is based on Bayesian techniques [27]. In this setting measured data (possibly noisy) is assumed to be the sum of model output and a discrepancy function, both of which are modeled as Gaussian processes. We do not go into details of the Bayesian approach in this paper but intend to make comparisons with the EMD approach in future work. However, it is worth noting that the results in this paper do not rely on a Gaussian noise model.

Our work is motivated in part by [11, 12, 43], where the authors use the quadratic Wasserstein metric to solve Full-Waveform Inversion (FWI) problems. In particular, it is demonstrated that the quadratic Wasserstein metric, as opposed to the L^2 norm, provides an effective measure of the misfit between given data and computed solution.

3.1.3 Our contribution

In this paper we introduce a new tool, called the structure, that is based on the Earth Mover's Distance (EMD) from optimal transport. We show that the structure is sensitive to modeling errors in L, but insensitive to noise in b. For simple functional forms of L_θ, we demonstrate that the structure can successfully recover the correct parameter $\hat{\theta}$. The method can be implemented as a wrapper around existing inverse problem solvers and thus can be easily integrated into preexisting work flows for solving inverse problems with minimal modifications to existing code bases. Moreover, due to recent advancements in the calculation of the EMD [29, 30], the additional cost is reasonable.

Our work extends that of [11, 12, 43] by considering different inverse problems, a more general noise model, and we use a different Wasserstein metric. See section 3.3.4 for more detail. We also show that new algorithms for computing the EMD can be combined with inverse problem solvers to diagnose forward operator error in general inverse problems.

3.2 Background

3.2.1 Inverse Problems

Let $\mathcal{U} \subset L^\infty(X)$ and $\mathcal{B} \subset L^\infty(Y)$ be function spaces defined over bounded rectangular domains $X \subset \mathbb{R}^{d_x}$ and $Y \subset \mathbb{R}^{d_y}$, respectively. We consider problems which come from the discretization of the linear equation

$$\mathcal{L} f = g \tag{3.2}$$

where $f \in \mathcal{U}$, $g \in \mathcal{B}$, and $\mathcal{L} : \mathcal{U} \to \mathcal{B}$ is a bounded linear operator.

To discretize Eq. 3.2, we assume that for some $\Delta x > 0$ and $\Delta y > 0$, X and Y can be partitioned into hypercubes K^x and K^y, respectively, of size $= \Delta x^{d_y}$ and Δy^{d_y}, respectively, such that $X = \cup_i \overline{K_i^x}$ and $Y = \cup_j \overline{K_j^y}$. We then let

$$\mathcal{U}_{\Delta x} = \{ f_{\Delta x} \in \mathcal{U} : f_{\Delta x}|_{K_x} \text{ is constant for all } K_x \subset X \} \tag{3.3}$$

$$\mathcal{B}_{\Delta y} = \{ g_{\Delta y} \in \mathcal{B} : g_{\Delta y}|_{K_y} \text{ is constant for all } K_y \subset Y \}. \tag{3.4}$$

The discrete version of Eq. 3.2 takes the form

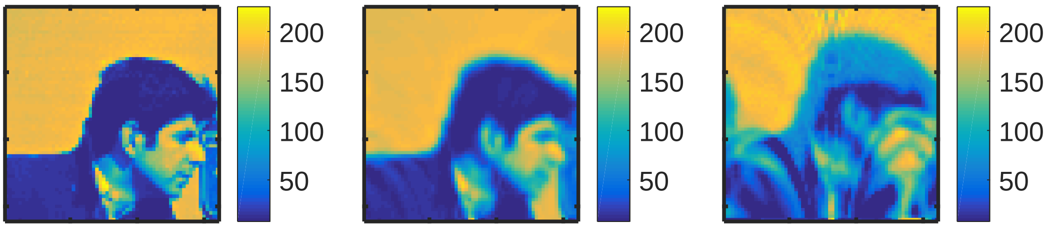
$$Lu = b, \tag{3.5}$$

where $u \in \mathcal{U}_{\Delta x}$, $b \in \mathcal{B}_{\Delta y}$, and $L \colon \mathcal{U}_{\Delta x} \to \mathcal{B}_{\Delta y}$ is a bounded linear operator that approximates $\mathcal{L}$. The exact forms of L, u, and b depend on the discretization. In the appendix, we present a discretization based on the assumption that $\mathcal{L}$ is generated by line integrals over paths $\mathcal{P}_y \subset X$ that are parameterized by elements $y \in Y$.

Solving Eq. 3.5 directly may not be practical if the condition number of L is large, as noise in b can be strongly amplified in the inversion process. A variational approach to address this difficulty is instead to solve

$$\tilde{u} = \tilde{L}^{-1} b \equiv \operatorname*{argmin}_{v \in \mathcal{U}_{\Delta x}} \|Lv - b\|_2^2 + \Phi(v; \lambda), \tag{3.6}$$

where $\Phi \colon \mathcal{U}_{\Delta x} \to \mathbb{R}^+$ is a regularizing functional with parameter $\lambda \in \mathbb{R}^+$. If $\Phi = 0$, then Eq. 3.6 gives the least squares solution of Eq. 3.5. Nontrivial examples of Φ (which may require more regularity than $L^\infty(X)$) include



(a) Ground truth, u. (b) u_θ when $\theta = 2.3 = \hat{\theta}$ (c) u_θ when $\theta = 2.4 \not\approx \hat{\theta}$

Figure 3.1: Demonstration of the sensitivity in the reconstruction in Eq. 3.6 to errors in the forward operator. In this example $L = L_{\hat{\theta}}$ is the 'academic operator' from [15], θ is the parameter R in [15, Table 1], and $\hat{\theta} = 2.3$. In this problem Tikhonov regularization was used to define the approximate inverse in Eq. 3.6.

1. $\Phi(v; \lambda) = \lambda \left\|Cv\right\|_2^2$, where the linear operator C approximates a differential operator (Generalized Tikhonov regularization);
2. $\Phi(v; \lambda) = \lambda \, \mathrm{TV}(()v)$ (Total Variation regularization [37]);
3. $\Phi(v; \lambda) = \lambda \left\|Cv\right\|_1$, where C is a transformation to a space in which u is known to be sparse (Basis Pursuit in Compressed Sensing [20]);
4. a weighted sum of the coefficients in some basis of U (such as a wavelet basis [31, 10] or singular vectors [25]).

These regularization methods are able to stably invert the operator L, at least approximately in the sense that $L\tilde{u} = L\tilde{L}^{-1}b \approx b$. Moreover, solutions of Eq. 3.6 are able to mitigate the effect of error within b; that is, even if b is corrupted (e.g. by noise), $\tilde{u}$ will be a reasonable reconstruction. In contrast, a modest error in L will likely result in a terrible reconstruction, regardless of the choice of Φ. An example of this behavior is given in Fig. 3.1.

For the purposes of this paper, we assume that there exists a family $\{L_\theta\}_{\theta \in \Theta}$ of forward operators parameterized by $\theta \in \Theta$, and a unique $\hat{\theta} \in \Theta$ such that $L_{\hat{\theta}} = L$. Given a noisy measurement $b + \eta$, where η is the noise, and a model parameter θ, the approximate reconstruction of u, based on the regularization in Eq. 3.6 with operator L_θ, is given by

$$\tilde{u}_{\theta,\eta} = \tilde{L}_\theta^{-1}(b+\eta). \tag{3.7}$$

where the tilde denotes the solution to a regularized problem of the form in Eq. 3.6 (where the choice of Φ is understood). This notation will be used throughout the remainder of the paper.

We define the residual as

$$r_{\theta,\eta} = (b+\eta) - L\tilde{u}_{\theta,\eta} = (I - L_\theta \tilde{L}_\theta^{-1})(b+\eta) \tag{3.8}$$

where I is the identity operator. The residual is the main object that we study to determine when the parameter θ is poorly chosen.

3.2.2 Earth Mover's Distance

A key tool in the analysis of forward operator error is the Earth Mover's Distance. Below we summarize the presentation in [30].

Definition 1 (Wasserstein Distance)**.** *Let $\Omega \subset \mathbb{R}^d$ be convex and compact, and let $c\colon \Omega\times\Omega \to [0,+\infty)$ be a distance. Given two non-negative distributions $\rho_1\colon \Omega \to \mathbb{R}^+, \rho_2\colon \Omega \to \mathbb{R}^+$ such that $\int_\Omega \rho_1 = \int_\Omega \rho_2$. For a given $p \in \mathbb{N}$ the p'th Wasserstein distance is*

$$\begin{aligned} W_p(\rho_1,\rho_2) = \left(\min_{\pi\geq 0} \int_{\Omega\times\Omega} c(x^{(1)},x^{(2)})^p \pi(x^{(1)},x^{(2)}) dx^{(1)} dx^{(2)}\right)^{1/p}&, \\ \textit{subject to:} \quad \int_\Omega \pi(x^{(1)},x^{(2)}) dx^{(2)} = \rho_1(x^{(1)})&, \\ \int_\Omega \pi(x^{(1)},x^{(2)}) dx^{(1)} = \rho_2(x^{(2)})&. \end{aligned} \tag{3.9}$$

The function c is called the ground metric and each feasible function π is referred to as a transport plan. In this work we set $c(x^{(1)},x^{(2)}) = \left\| x^{(1)} - x^{(2)} \right\|_2$. The Earth Mover's Distance we define here is a special case of the Wasserstein distance where $p = 1$.

Definition 2 (Earth Mover's Distance)**.** *Let $\Omega \subset \mathbb{R}^d$ be convex and compact, and let $c\colon \Omega \times \Omega \to [0,+\infty)$ be a distance. Given two non-negative distributions $\rho_1\colon \Omega \to \mathbb{R}^+, \rho_2\colon \Omega \to \mathbb{R}^+$ such that $\int_\Omega \rho_1 = \int_\Omega \rho_2$. The Earth Mover's Distance (EMD) between ρ_1 and ρ_2 is*

$$\mathrm{EMD}(\rho_1,\rho_2) = W_1(\rho_1,\rho_2). \tag{3.10}$$

The EMD can also be written in the equivalent form [14]

$$\begin{aligned}\mathrm{EMD}(\rho_1,\rho_2) = \min_m \int_\Omega \|m(x)\|_2 \, dx, \\ \text{subject to:} \quad \nabla\cdot m(x) + \rho_2(x) - \rho_1(x) = 0, \\ m(x)\cdot n(x) = 0 \quad \forall x \in \partial\Omega,\end{aligned} \tag{3.11}$$

where $n(x)$ is the normal vector at $x \in \partial\Omega$. This formulation is the basis for recently developed algorithms in [29, 30].

3.3 Applying EMD to inverse problems

3.3.1 Residual and operator correctness

In a variational reconstruction procedure, the quality of the fit can be investigated by an analysis of $r_{\theta,\eta}$ and $\Phi(\tilde{u}_{\theta,\eta})$. Generally, the larger λ the larger the first term and the smaller the second and vice-versa. Typically the value of λ is chosen in an attempt to balance these contributions [24, 25]. However if an incorrect forward operator is used, $r_{\theta,\eta}$ will have an additional contribution that does not depend on λ.

The characterization above can be made precise in the case of Tikhonov regularization by introducing a matrix notation and using Generalized Singular Value Decomposition [21, Chapter 8.7.3]. To this end, let $n = \dim(\mathcal{U}_{\Delta x})$ and $m = \dim(\mathcal{B}_{\Delta y})$, and expand u and b in terms of characteristic basis functions:

$$u(x) = \sum_{j=1}^{n} u_j \chi_{K_j^x}(x) \quad \text{and} \quad b(y) = \sum_{i=1}^{m} b_i \chi_{K_i^y}(y). \tag{3.12}$$

Then Eq. 3.5 becomes

$$\mathbf{L}\mathbf{u} = \mathbf{b}. \tag{3.13}$$

where $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{b} = (b_1, \ldots, b_m)$, and $\mathbf{L}$ has components

$$L_{i,j} = \frac{1}{\Delta y^{d_y}} \int_Y \chi_{K_i^y} L \chi_{K_j^x} dy. \tag{3.14}$$

Definition 3 (GSVD). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{o\times n}$ *be two matrices such that* $\mathrm{null}(\mathbf{A}) \cap \mathrm{null}(\mathbf{B}) = \emptyset$. *The Generalized Singular Value Decomposition (GSVD) of the matrix pair*

$(\mathbf{A}, \mathbf{B})$ *is given by*

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{Z}^T \quad \textit{and} \quad \mathbf{B} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{Z}^T, \tag{3.15}$$

where $\mathbf{U} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{V} \in \mathbb{R}^{o\times n}$ *are semi-orthogonal;* $\mathbf{Z} \in \mathbb{R}^{n\times n}$ *is invertible; and*

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n} \quad \textit{and} \quad \boldsymbol{\Gamma} = \operatorname{diag}(\gamma_1, \ldots, \gamma_n) \in \mathbb{R}^{n\times n} \tag{3.16}$$

are diagonal matrices such that

$$1 \geq \sigma_1 \geq \cdots \geq \sigma_n \geq 0 \quad \textit{and} \quad 0 \leq \gamma_1 \leq \cdots \leq \gamma_n \leq 1, \tag{3.17}$$

with $\boldsymbol{\Sigma}^2 + \boldsymbol{\Gamma}^2 = \mathbf{I}$.

Using the GSVD, we obtain the following proposition that is proven in the appendix:

Proposition 1 (Residual with Tikhonov regularization)**.** *Suppose* $\mathbf{L}\mathbf{u} = \mathbf{b}$, *where* $\mathbf{L} \in \mathbb{R}^{m\times n}$ *and* $m > n$. *Let* $\tilde{\mathbf{u}}_{\theta,\eta}$ *be defined by Eq. 3.7 with* $\Phi(\mathbf{v}; \lambda) = \lambda \|\mathbf{C}\mathbf{v}\|_2^2$, *where* $\mathbf{C} \in \mathbb{R}^{o\times n}$, *and a noise vector* $\boldsymbol{\eta} \in \mathbb{R}^m$ *whose elements are independent and spherically symmetric—that is,* $\boldsymbol{\eta}$ *and* $\mathbf{Q}\boldsymbol{\eta}$ *have the same probability distribution function for any orthogonal matrix* $\mathbf{Q} \in \mathbb{R}^{m\times m}$. *Assume that* $\operatorname{null}(\mathbf{L}_\theta) \cap \operatorname{null}(\mathbf{C}) = \emptyset$ *so that the GSVD*

$$\mathbf{L}_\theta = \mathbf{U}_\theta\boldsymbol{\Sigma}_\theta\mathbf{Z}_\theta^T \qquad \mathbf{C} = \mathbf{V}_\theta\boldsymbol{\Gamma}_\theta\mathbf{Z}_\theta^T \tag{3.18}$$

for the matrix pair $(\mathbf{L}_\theta, \mathbf{C})$ *is well-defined. Then the residual* $\mathbf{r}_{\theta,\eta}$ *associated to* $\tilde{\mathbf{u}}_{\theta,\eta}$ *satisfies the bound*

$$\begin{aligned} \|\mathbf{r}_{\theta,\eta}\|_2^2 \leq & \left\|(\mathbf{I} - \mathbf{U}_\theta\mathbf{U}_\theta^T)\mathbf{b}\right\|_2^2 + \|(\mathbf{b} - \mathbf{L}_\theta\mathbf{u})\|_2^2 \\ & + \frac{1}{4}\lambda \left\|\mathbf{Z}_\theta^T\mathbf{u}\right\|_2^2 + \frac{m - n + \operatorname{Tr}(\hat{\mathbf{D}}_{\theta,\lambda}^2))}{m}\mathbb{E}\left[\|\boldsymbol{\eta}\|_2^2\right]. \end{aligned} \tag{3.19}$$

where

$$\hat{\mathbf{D}}_{\theta,\lambda} := \frac{\lambda\boldsymbol{\Gamma}_\theta^2}{\boldsymbol{\Sigma}_\theta^2 + \lambda\boldsymbol{\Gamma}_\theta^2}. \tag{3.20}$$

This result shows how calibration error can induce $O(1)$ terms (with respect to the regularization parameter λ) into the residual, the first two terms in Eq. 3.19. The noise that is orthogonal to the image of $\mathbf{L}_\theta$ also induces $O(1)$ terms, even if $\theta = \hat{\theta}$. Thus it is important to develop tools that can differentiate between these two contributions. For completeness, one should also consider regularization with more general forms of Φ. Unfortunately in many situations, the operator $\tilde{\mathbf{L}}_\theta^{-1}$ is nonlinear, and a rigorous analysis in this vein is much more difficult.

3.3.2 Introduction to the structure

We introduce a mathematical tool to detect contributions to $r_{\theta,\eta}$ that are due to errors in the operator L, i.e., when $\theta \neq \hat{\theta}$, and is insensitive to noise in the residual. This tool, which we call the structure, is a functional built using the Earth Mover's Distance (EMD).

Definition 4 (Structure)**.** *For any $f \in L^1(\Omega)$, the structure of f is*

$$\operatorname{struc}[f] = \operatorname{EMD}(f^+, f^-), \tag{3.21}$$

where

$$f^+(x) = \max(f(x) - \mu, 0) \quad \textit{and} \quad f^-(x) = \max(\mu - f(x), 0) \tag{3.22}$$

and $\mu = \frac{1}{\|\Omega\|} \int_\Omega f(x)dx$.

The following proposition is proven in the appendix.

Proposition 2 (Basic Properties of Structure)**.** *The operator* struc $[\cdot]$ *satisfies the following properties:*

1. *it is a semi-norm on $L^1(\Omega)$;*

2. *for all $g \in L^1(\Omega)$ and $c \in \mathbb{R}$,*

$$\operatorname{struc}[g] = \operatorname{struc}[g + c]; \tag{3.23}$$

3. struc $[c] = 0$ *for any constant $c \in \mathbb{R}$;*

4. if $\rho_1 \colon \Omega \to \mathbb{R}^+$, $\rho_2 \colon \Omega \to \mathbb{R}^+$ *and* $\int_\Omega \rho_1 = \int_\Omega \rho_2$,

$$\operatorname{struc}[\rho_2 - \rho_1] = \operatorname{EMD}(\rho_1, \rho_2). \tag{3.24}$$

Using struc $[\cdot]$ is a good strategy for detecting operator error for several reasons:

- The struc $[\cdot]$ is small when applied to piecewise noise and large when applied to a (non-constant) smooth function. (Rigorous statements this effect are made in Section 3.3.3 below). Thus struc $[r_{\theta,\eta}]$ will be small when the forward operator is correct and large when it is not. Although the struc $[\cdot]$ of a constant is zero, any such contribution to the residual can be discerned by applying a standard norm to its spatial average.
- With recent algorithmic advances [26, 29, 30], the underlying EMD calculation for computing struc $[\cdot]$ can be performed quickly. If $\mathbf{b} \in \mathbb{R}^{256} \times \mathbb{R}^{256}$, then struc $[\mathbf{b}]$ can be computed in less than a second using an intel i7-4770 processor. In general, the limiting factor in computing the struc $[\cdot]$ is the fast Fourier transform. Hence if $\mathbf{b} \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_j}$ then struc $[\mathbf{b}]$ is computed in $O\left(\prod_{i=1}^{j} d_i \log(d_i)\right)$ time [26].
- Because its evaluation does not affect the actual inverse procedure, the structure calculation can be incorporated into existing work flows without altering old code. Thus it can be quickly integrated into an existing toolbox for solving inverse problems.
- The struc $[r_{\theta,\eta}]$ calculation produces not only a number, but also outputs a transport plan (see Figs. 3.4b, 3.4d). For certain classes of forward operators this additional information can be leveraged to correct forward operators with minimal tuning. This idea will be explored in future work.

3.3.3 Theoretical Results

In this section we establish some theoretical results which support the use of the structure as a tool for diagnosing structural errors in the forward operator of an inverse problem. The proofs of Theorems 1–2 are given in Appendix. 3.6.

Theorem 1 (Characterization of noise by structure)**.** *Given non-negative integers d and ℓ, let $\Omega = [0,1)^d$ and let $\mathcal{O}_\ell = \left\{\omega_{\ell,1}, \ldots, \omega_{\ell,2^{\ell d}}\right\}$ partition Ω into $2^{\ell d}$ hypercubes of volume $2^{-\ell d}$. Define $h_\ell : \Omega \to \mathbb{R}$ by*

$$h_\ell(y) = \eta_{\ell,1}\chi_{\ell,1}(y) + \cdots + \eta_{\ell,2^{\ell d}}\chi_{\ell,2^{\ell d}}(y) \tag{3.25}$$

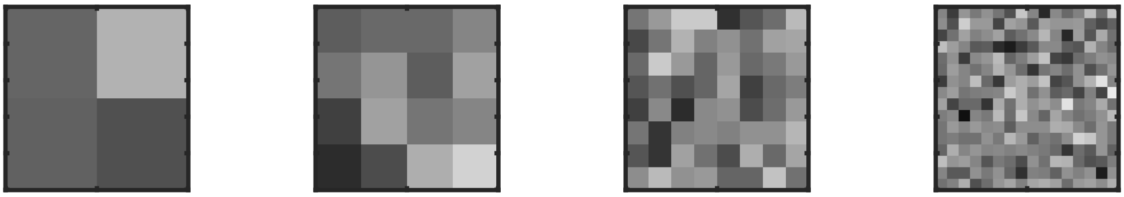
where

$$\chi_{\ell,i}(y) = \begin{cases} 1, & x \in \omega_{\ell,i}, \\ 0, & x \notin \omega_{\ell,i}, \end{cases} \tag{3.26}$$

and $\{\eta_{\ell,i}\}_{i=1}^{2^{\ell d}}$ is a set i.i.d. random variables with mean μ and variance σ^2 (See Fig. 3.2 for a visualization of h_ℓ). If $\epsilon_\ell = 2^{-\ell}$,

$$\mathbb{E}\left[\mathrm{struc}\left[h_\ell\right]\right] \leq \sigma \begin{cases} -\epsilon_\ell \log \epsilon_\ell, & d = 2, \\ 2\sqrt{d}\epsilon_\ell, & d > 2, \end{cases} \tag{3.27}$$

where the expectation is with respect to the weights $\eta_{\ell,i}$.



(a) h_1 (b) h_2 (c) h_3 (d) h_4

Figure 3.2: Example of h_ℓ when $d = 2$, $\mu = 0$, and $\sigma = 1$.

Lemma 1 (L^2 norm of Noise)**.** *Given the assumptions of Thm. 1, suppose further that $\mu = 0$. Then*

$$\mathbb{E}\left[\|h_\ell\|_2^2\right] = \sigma^2, \tag{3.28}$$

where the expectation is with respect to the weights $\eta_{\ell,i}$.

Theorem 2 (Characterization of a smooth function by structure)**.** *Given the assumptions of Thm. 1, let $R_\ell \colon \mathcal{B} \to \mathcal{B}_{\epsilon_\ell}$. If*

$$R_\ell\phi(y) = \frac{1}{\omega_{\ell,i}} \int_{\omega_{\ell,i}} \phi(z)dz, \quad \forall y \in \omega_{\ell,i}. \tag{3.29}$$

where $\phi \in C^1\left(\overline{Y}\right)$ *then*

$$|\text{struc}\,[R_\ell\phi] - \text{struc}\,[\phi]| \leq C(|\nabla\phi|)\, d\epsilon_\ell^2, \tag{3.30}$$

where the constant C *depends on the maximum of* $\nabla\phi$ *on* $\overline{Y}$*. In particular,*

$$\text{struc}\,[R_\ell\phi] \to \text{struc}\,[\phi] \;\; \textit{as } \ell \to +\infty. \tag{3.31}$$

3.3.4 Comparison with prior work

The work here is inspired, in part, by the study of seismic imaging inverse problems in [11, 12, 43]. There the authors measure the misfit between simulated and measured data using the Wasserstein distance squared $W_2^2(\rho_1, \rho_2) = (W_2(\rho_1, \rho_2))^2$. To handle the possibly negative distributions, the authors in [11, 12, 43] introduce the *m*isfit function

$$\begin{aligned} d(f,g) = W_2^2 &\left(\frac{\max(f,0)}{\int \max(f,0)dx}, \frac{\max(g,0)}{\int \max(g,0)dx} \right) \\ + W_2^2 &\left(\frac{\max(-f,0)}{\int \max(-f,0)dx}, \frac{\max(-g,0)}{\int \max(-g,0)dx} \right) \end{aligned} \tag{3.32}$$

which plays a similar role to struc $[f - g]$ in this work. In [11, Section 2.6] the authors show that d is insensitive to noise, with a scaling result that is similar to Thm. 1 up to a logarithmic factor. Specifically, if f and g are two non-negative functions such that $f - g$ has the form of h_ℓ, defined in Eq. 3.25), with uniformly distributed noise, then

$$d(f,g) = O(\epsilon_\ell). \tag{3.33}$$

The approach taken in [11, 12, 43] differs from the approach in this paper in at least two key ways. First is the choice of W_2^2 rather than W_1. This has the following consequences:

- W_2 and W_2^2 have the property of *cyclic monotonicity* (see [13, Sec. 2.1] for a definition and proof), which can be used to show convexity of d with respect to shifts, dilation and partial amplitude loss. In this work we make no such claims about the convexity of struc $[\cdot]$.

- As a semi-norm, the EMD (like all W_p for $p \in [1, \infty)$) is a degree-one homogeneous functional and satisfies a triangle inequality (see [41, p. 94]. The functional W_2^2 has neither property. For example of the latter, let $f = 2\chi_{0,1/2}$, $h = 2\chi_{1/2,1}$ and $g = 2\chi_{1,3/2}$. Then $W_2^2(f, h) = \frac{1}{4}$, $W_2^2(h, g) = \frac{1}{4}$ but $W_2^2(f, g) = 1$, then

$$W_2^2(f, g) > W_2^2(f, h) + W_2^2(h, g). \tag{3.34}$$

- Redefining d with W_2 instead of W_2^2 would recover a triangle inequality and degree-one homogeneity. However, the cost of such a modification would be to increase the sensitivity of d to noise. Indeed, the scaling in Eq. 3.33 would change from $O(\epsilon_\ell)$ to $O(\epsilon_\ell^{1/2})$, which is significantly slower than the scaling in Thm. 1.

- Finally, W_1 is more directly analogous to the definition of work used throughout physics, distance times effort. Consider the case when

$$f(x) = \frac{1}{2}\chi_{[0,2]}(x) \quad g(x) = \frac{1}{2}\chi_{[1,3]}(x) \tag{3.35}$$

  and the two transport plans

$$\pi_1(x_1, x_2) = \begin{cases} 1/2 \text{ if } x_2 = 1 + x_1 \text{ and } x_1 \in [0, 2] \\ 0 \text{ otherwise} \end{cases} \tag{3.36}$$

$$\pi_2(x_1, x_2) = \begin{cases} 1/2 \text{ if } x_2 = 2 + x_1 \text{ and } x_1 \in [0, 1] \\ 0 \text{ otherwise} \end{cases} \tag{3.37}$$

  The cost of π_1 as measured by W_2 is twice that of π_2. Both plans cost the same as measured by W_1. In words W_2 'prefers' to make many smaller movements as opposed to fewer larger movements, while W_1 is agnostic to such differences.

The second key difference between the approach in [11, 12, 43] and the approach taken here lies in the definition of d and struc $[\cdot]$, both of which are used to address the fact that the Wasserstein metric is only defined for non-negative distributions with the same mass. It is worth noting that $d(f, g)$ and struc $[\cdot]$ could be defined using any Wassterstein metric. However, d introduces several undesirable artifacts.

- The normalization in the definition means that

  $$d(\lambda f, \lambda g) = d(f, g), \quad \forall \lambda \in \mathbb{R}^+. \tag{3.38}$$

  In particular, unlike struc $[\cdot]$, it is not degree-one homogeneous.

- Special care is required in the case that $\max(f, 0) \equiv 0$ but $\max(g, 0) \not\equiv 0$. Indeed one of the reasons that the results in Eq. 3.33 require f and g to be positive and differ only by uniform noise is that small changes is the noise can alter the support of $\max(f, 0)$ and $\max(g, 0)$. The struc $[\cdot]$ has no such restrictions on the noise model.

- The struc $[\cdot]$ is continuous w.r.t. the $L^1(\Omega)$ norm provided that Ω is bounded (see Lemma 5). $d(f, g)$, however, is not. For example consider, the functions

  $$f_\epsilon = \chi_{[\epsilon,1-\epsilon]} - \epsilon\chi_{(1-\epsilon,1]}, \quad g_\epsilon = -\epsilon\chi_{[0,\epsilon)} + \chi_{[\epsilon,1-\epsilon]} - \epsilon\chi_{(1-\epsilon,1]}. \tag{3.39}$$

  Clearly $f_\epsilon - g_\epsilon \to 0$ in $L^1(\Omega)$ as $\epsilon \to 0$; however,

  $$\lim_{\epsilon\to 0} d(f_\epsilon, g_\epsilon) \geq \lim_{\epsilon\to 0} \frac{1}{2}\left(1 + \frac{\epsilon}{4}\right)^2 = \frac{1}{2}. \tag{3.40}$$

  This lack of continuity due to sign changes is one of the reasons for having restrictions on the noise model for $d(f, g)$.

- The kernel of struc $[\cdot]$ consists of constant functions, and so struc $[f - g] = 0 \iff f = g + c$ for some constant c. This c is easily recovered by computing the difference between the averages if f and g.On the other hand, the kernel of d is

  $$\mathrm{Ker}(d) = \left\{ \begin{array}{ll} (f, g) \in L^1 \times L^1 : & \max(f, 0) = \lambda_+ \max(g, 0) \text{ and} \\ & \max(-f, 0) = \lambda_- \max(-g, 0) \quad \text{for } \lambda_+, \lambda_- \in \mathbb{R}^+ \end{array} \right\} \tag{3.41}$$

  It is more difficult to account for such a kernel.

3.4 Numerical Results

In this section we present the results of several numerical experiments. We make two simplying assumptions. First, we let X and Y be two dimensional domains. This choice is motivated by ease of visualization as well as the availability of code to quickly compute the EMD in two dimensions. We, however, believe that the results generalize well to high dimensional problems. Second, we assume that L_θ is linear in θ. This choice is for simplicity, but it also is a reasonable approximation for finding a local optimum. Indeed, if L_θ smoothly depends on θ, then L is locally linear:

$$L_{\hat{\theta}+\delta\theta} = L_{\hat{\theta}} + \nabla_\theta L(\hat{\theta}) \cdot \delta\theta + O(\delta\theta^2). \tag{3.42}$$

For each experiment, we provide with a known signal u and a family of operators $\{L_\theta\}_{\theta\in\Theta}$. We then set $L = L_{\hat{\theta}}$ for some $\hat{\theta} \in \Theta$, generate a measurement $b = L_{\hat{\theta}} u$, and examine the behavior of $\text{struc}\left[r_{\theta,\eta}\right]$ as a function of θ. The expectation is that

$$\hat{\theta} \approx \theta^* := \operatorname*{argmin}_{\theta\in\Theta} \text{struc}\left[r_{\theta,\eta}\right]. \tag{3.43}$$

The first two experiments show that indeed $\theta^* \approx \hat{\theta}$ even with relatively high noise. The final experiment illustrates that the method performs better as the problem becomes more overdetermined or more regularized (i.e. λ in Eqn. 3.6 increases). We report a figure of merit, the contrast, defined as:

$$\text{cont}(F) = \frac{\max(F) - \min(F)}{\max(F) + \min(F)} \tag{3.44}$$

for any $F\colon \Theta \to \mathbb{R}^+$ that is not identically zero. The contrast measures the depth of a minimum, and the greater the contrast, the less the location of the minimum changes in the presence of additive noise in F. In all three experiments we compare the contrast of $\text{struc}\left[\cdot\right]$ with the discrete norms $\|\cdot\|_1$ and $\|\cdot\|_2$. For any $z \in \mathcal{B}_{\Delta y}$ these norms are given by,

$$\|z\|_1 = \Delta y^2 \sum_{i_1,i_2} |z_{i_1,i_2}| \quad \text{and} \quad \|z\|_2 = \Delta y \left(\sum_{i_1,i_2} z_{i_1,i_2}^2\right)^{1/2} \tag{3.45}$$

We also generate plots of all three (semi-) norms as a function of the parameter θ.

3.4.1 Implementation Details

The implementation of each of these experiments involves four basic steps: (i) the generation of the random forward operators L_θ; (ii) generation of the signal u, measurement b and noise

Parameter	Value	Parameter	Value	Ref.	Parameter	Value	Ref.
Discretization[1]		**Inversion**			struc $[\cdot]$		
Δx	1/64	$\Phi(\cdot, \lambda)$	$\lambda\,\mathrm{TV}(()u)$	[37]	Max Iter	8000	[30]
Δy	1/100	λ	10	[37]	EMD_μ	7e-6	[30]
		μ	100	[20]	EMD_τ	3	[30]
		Bregman Iterations	10	[20]			

Table 3.1: Numerical parameters for Experiments 1 - 3.

η; (iii) calculation of $\tilde{u}_{\theta,\eta}$; and (iv) computation of the struc $[\cdot]$. The specific values of parameters needed to recreate the results are given in Table 3.1.

1. **G**eneration of the random forward operators. Recall the definitions in Section 3.2.1. A forward operator L_θ, even an academic one, is a discretization of an operator $\mathcal{L} \colon \mathcal{U} \to \mathcal{B}$. In applications, L_θ models the action of some physical process which produces a measurement. For example in seismic imaging the forward operator is the propagation of a seismic wave [11], and in plasma imaging in tokamaks the forward operator couples the optics of the camera with the symmetries of the plasma [42].

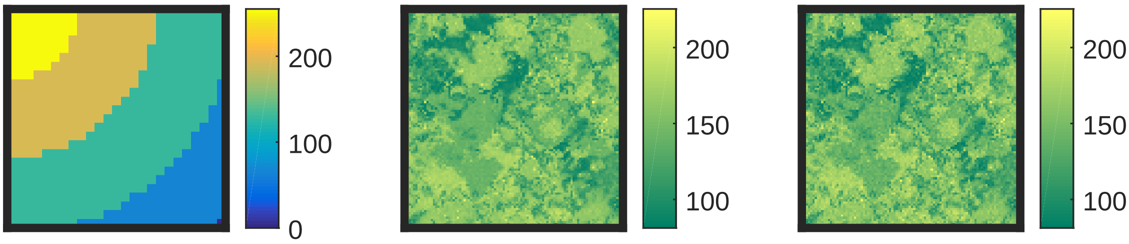
   For the experiments, we presume that $\mathcal{L}$ is a Line Integral Operator (LIO). (See Appendix 3.7 for details.) If $f \colon X \to \mathbb{R}$ and $g \colon Y \to \mathbb{R}$, then for each $y \in Y$, $g(y)$ represents the integral of f over some path $p(y)$. Some examples of common LIO are the Radon, Abel and Helical Abel transforms [15].

2. **G**eneration of the signal, measurement and noise. The underlying signal $u \in \mathcal{U}_{\Delta x}$ is a series of concentric rings (see Fig. 3.3a). Then we apply $L_{\hat{\theta}}$ to u to obtain a noiseless measurement $b \in \mathcal{B}_{\Delta y}$ (see Fig. 3.3b). The noisy signal (see Fig. 3.3c) is generated by adding independent white noise η with mean zero and variance σ to each element of b so that

$$\mathrm{SNR} = \frac{\|b\|_2}{\|\eta\|_2} \tag{3.46}$$

[1] Δx and Δy both change for Experiment 3, however the other parameters are fixed.

is at a specified level.



(a) u.

(b) $b = L_{\hat{\theta}} u$.

(c) $b + \eta$.

Figure 3.3: The signal u, measurement b, and noisy measure $b + \eta$ for Experiment 1.

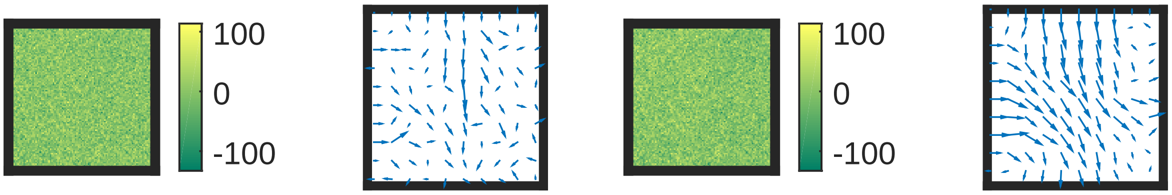
3. **C**omputation of $\tilde{u}_{\theta,\eta}$. Throughout these experiments, we use the inversion procedure of the form of Eq. 3.6 with $\Phi(v; \lambda) = \lambda \left\| \mathbf{C} v \right\|_1$ where $\mathbf{C}$ is a one-sided discrete approximation of the gradient operator:

$$\begin{aligned} (Cv)_{2i,j} &= \frac{1}{dx} \left(v_{i,j} - v_{\ell-1,j} \right) \\ (Cv)_{2i+1,j} &= \frac{1}{dy} \left(v_{i,j} - v_{i,j-1} \right) \end{aligned} \tag{3.47}$$

where $v_{i,j}$ is the i'th x and j'th y component of the vector $\mathbf{v}$, and likewise for $(\mathbf{Cv})_{i,j}$.This is TV regularization and has found wide success within image processing, especially when the underlying signal to be recovered is piecewise constant [20, 37]. Finally for experiments 1 and 2 we fix $\lambda = 1$. For experiment 3 we study how the results change as λ does.

To solve the resulting non-linear variational problem, we use the Split-Bregman algorithm, specifically the Generalized Split-Bregman Algorithm (GSBA) of [20], which requires specification of a step size parameter μ (called λ in [20]). GSBA requires the repeated solution of the linear system $(\mathbf{L}^T\mathbf{L} + \lambda^2 \mathbf{C}^T\mathbf{C})x = y$. The matrix $(\mathbf{L}^T\mathbf{L} + \lambda^2 \mathbf{C}^T\mathbf{C})$ is sparse and so we solve it using the L-BFGS [4, 44] method (limited memory Broyden-Fletcher-Goldfarb-Shanno[6, 16, 19, 40]).

4. **C**omputation of the struc $[\cdot]$. Computing struc $[\cdot]$ requires computing EMD. The algorithm that we use is given in [29, 30, 38].



(a) $r_{0.04,\eta}$ (b) $\hat{m}_{0.04}$ (c) $r_{0.20,\eta}$ (d) $\hat{m}_{0.20}$

Figure 3.4: Results from Experiment 1. The residual and flow $\hat{m}_\theta$ that minimizes Eq. 3.11 for a given θ. In Figs. 3.4b and 3.4d, the orientation of the arrows represents the direction $\hat{m}_\theta$, and the length of the arrows is proportional to the magnitude.

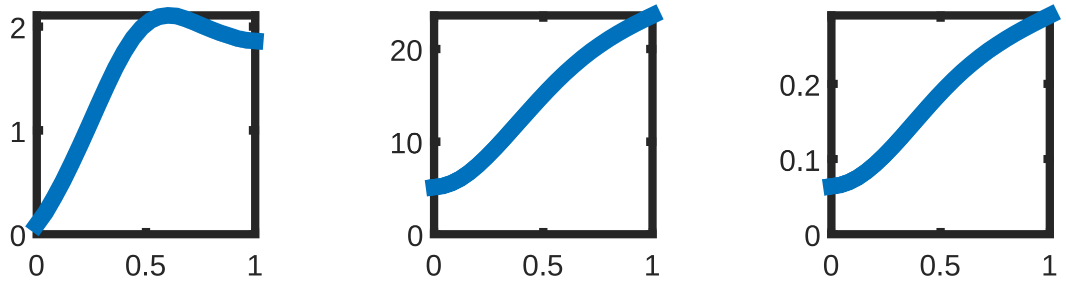
3.4.2 Experiment 1

This experiment is based on a normalized Eq. 3.42 where $p = 1$. Let L_0 and L_1 be two operators generated as described in Appendix 3.7. We define $\theta \in [0, 1]$ and

$$L_\theta = (1 - \theta)L_0 + \theta L_1. \tag{3.48}$$

Fig. 3.4 is a plot of the residual for different values of θ. In Fig. 3.4a, $\theta = 0.04$, and in Fig. 3.4c $\theta = 0.20$. Upon close inspection, one can see that from Fig. 3.4a that when θ is small the residual visually looks like white noise, whereas from Fig. 3.4c when θ is large the residual has underlying structure in addition to the noise. It is, however, difficult to see. Despite these two plots appearing similar they have very different structures, $\text{struc}\left[r_{0.04,\eta}\right] \approx 0.06$ and $\text{struc}\left[r_{0.20,\eta}\right] \approx 0.54$. The structure is also evident by looking at Figs. 3.4b, 3.4d, which are m from Eq. 3.11. Note that when $\theta = 0.04$, m is higgledy-piggledy, whereas when $\theta = 0.20$, m appears much more orderly.

A plot of $\text{struc}\left[r_{\theta,\eta}\right]$ vs θ is given in Fig. 3.5. Clearly, $\text{struc}\left[r_{\theta,\eta}\right]$ is minimized when $\theta \approx 0$. Further, we note that $\text{struc}\left[r_{\theta,\eta}\right]$ is increasing as a function of θ when $\theta \in [0, 0.5]$, however then decreases. This is expected behavior around the minimum, however the problem is evidently not convex away from $\hat{\theta}$. This is important to keep in mind for future work.

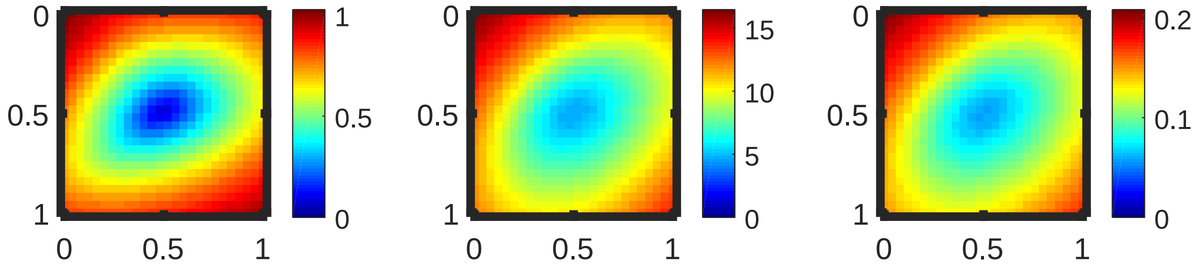


(a) struc $[r_{\theta,\eta}]$ vs θ. (b) $\|r_{\theta,\eta}\|_1$ vs θ. (c) $\|r_{\theta,\eta}\|_2$ vs θ.

Figure 3.5: Results from Experiment 1. The value of $r_{\theta,\eta}$ as measured by struc $[\cdot]$, $\|\cdot\|_1$ and $\|\cdot\|_2$. In all examples the minimum occurs when $\theta = 0$ however the contrast is greatest for struc $[\cdot]$.

3.4.3 Experiment 2

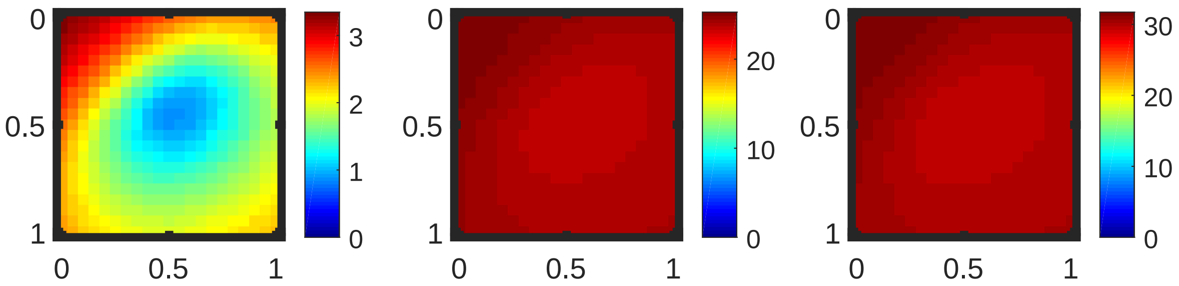
Experiment 2 is also based on a normalized Eq. 3.42, however in this case $p = 2$ and $\hat{\theta} = \left(\frac{1}{2}, \frac{1}{2}\right)$. The true signal used in Experiment 2 is the same as in Experiment 1 (see Fig. 3.3a). This experiment studies the change in the contrast for struc $[\cdot]$, $\|\cdot\|_1$ and $\|\cdot\|_2$ as the SNR decreases. The results are summarized in Table 3.2.



(a) struc $[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

Figure 3.6: Results from Experiment 2. In these plots SNR $= 25$.

In all cases, the contrast of struc $[\cdot]$ is greatest, and the contrast of struc $[\cdot]$ relative to $\|\cdot\|_1$ of $\|\cdot\|_2$ increases as the problem becomes more noisy. This suggests that struc $[\cdot]$ is a more robust choice of semi-norm for measuring the level of miscalibration of L_θ, especially when noise levels are high.



(a) $\mathrm{struc}\,[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

Figure 3.7: Results from Experiment 2. In these plots SNR = 5.

Contrast	$\mathrm{struc}\,[r_{\theta,\eta}]$	$\|r_{\theta,\eta}\|_1$	$\|r_{\theta,\eta}\|_2$
SNR = 25	0.7547	0.3493	0.3544
SNR = 5	0.5917	0.0398	0.0404

Table 3.2: Results from Experiment 2. The contrast for different choices of (semi)norms. Larger is better.

3.4.4 Experiment 3

The final experiment examines the necessity of the overdetermined assumption of L_θ as well as the role of λ. We repeat the setup of Experiment 2; however we fix the SNR = 25 and adjust Δy so that $L_\theta : \mathcal{U}_{\Delta x} \to \mathcal{B}_{\Delta y}$ becomes a square operator and independently let $\lambda = 0.1$, 1 and 10. We start with a fixed reference Δy_0, and consider

$$\mathcal{B}_{\Delta y_0} \cong \mathbb{R}^{100\times 100} \quad \mathcal{B}_{4/3\Delta y_0} \cong \mathbb{R}^{75\times 75} \quad \mathcal{B}_{2\Delta y_0} \cong \mathbb{R}^{50\times 50} \quad \mathcal{B}_{4\Delta y_0} \cong \mathbb{R}^{25\times 25}. \tag{3.49}$$

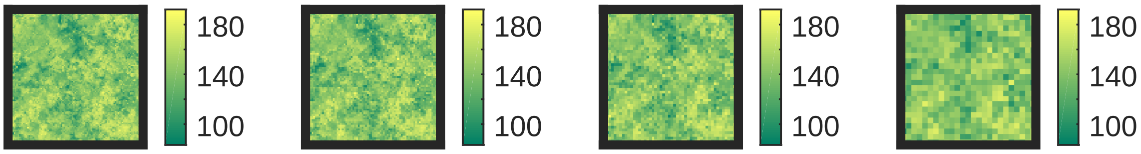
In all cases, $\mathcal{U}_\Delta x \cong \mathbb{R}^{25\times 25}$ is fixed. Each of the $\mathcal{B}$ in Eq. 3.49 are plotted in Fig. 3.8.

$$\theta^s = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{struc}\,[r_{\theta,\eta}] \quad \theta^1 = \operatorname*{argmin}_{\theta\in\Theta} \|r_{\theta,\eta}\|_1 \quad \theta^2 = \operatorname*{argmin}_{\theta\in\Theta} \|r_{\theta,\eta}\|_2 \tag{3.50}$$

The contrast is recorded in Tables 3.3 - 3.5. Finally for $\lambda = 1$, plots of $\mathrm{struc}\,[r_{\theta,\eta}]$, $\|r_{\theta,\eta}\|_1$, and $\|r_{\theta,\eta}\|_2$ vs θ as Δy changes are given in Figs. 3.9 - 3.12.

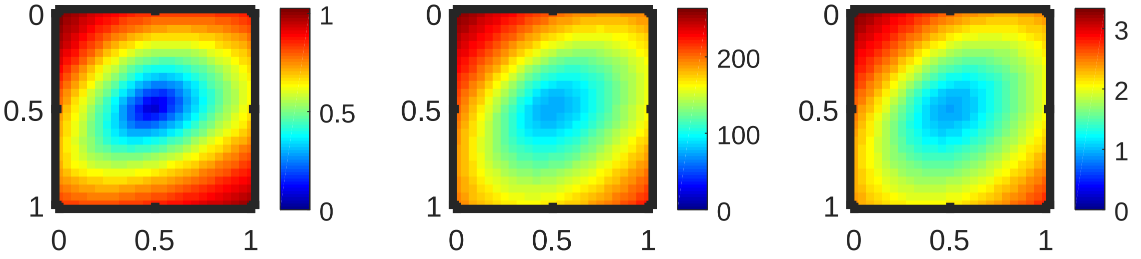
Below we give some more numerical results, when $\lambda = 0.1$ and $\lambda = 10$ for completeness.

From Tables 3.3 - 3.5 we observe two things. First as the problem becomes more overdetermined, the contrast improves for all three metrics, but especially for the $\mathrm{struc}\,[\cdot]$. Indeed



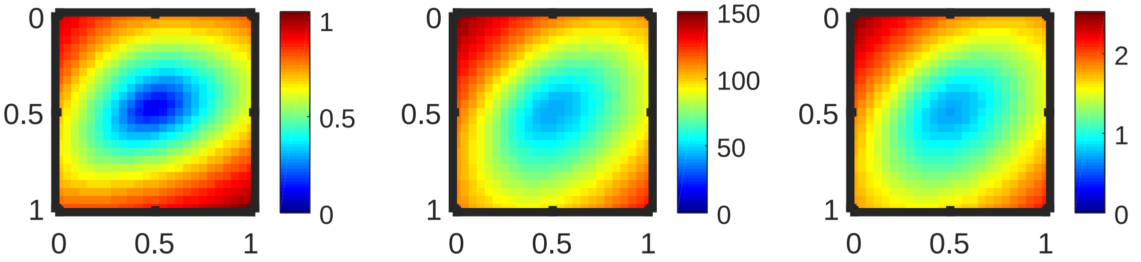
(a) $b \in \mathbb{R}^{100\times 100}$ (b) $b \in \mathbb{R}^{75\times 75}$ (c) $b \in \mathbb{R}^{50\times 50}$ (d) $b \in \mathbb{R}^{25\times 25}$

Figure 3.8: Plot of b for various choices of Δy (see Eq. 3.49).



(a) $\mathrm{struc}\,[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

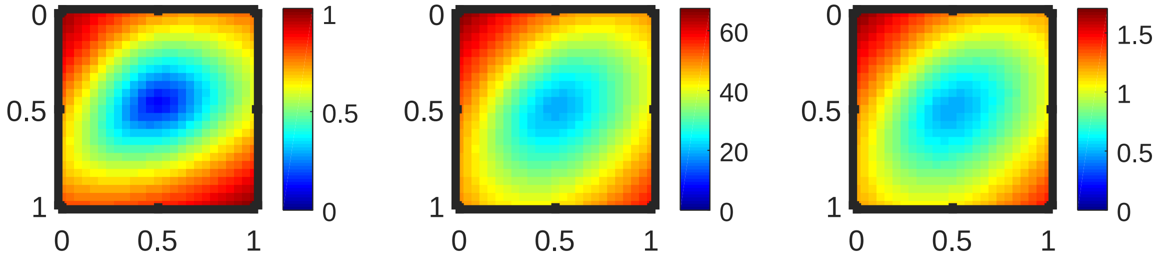
Figure 3.9: Results from Experiment 3 when $\lambda = 1$. In these plots $L\colon \mathbb{R}^{25\times 25} \to \mathbb{R}^{100\times 100}$. See Table 3.4 for the contrast.



(a) $\mathrm{struc}\,[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

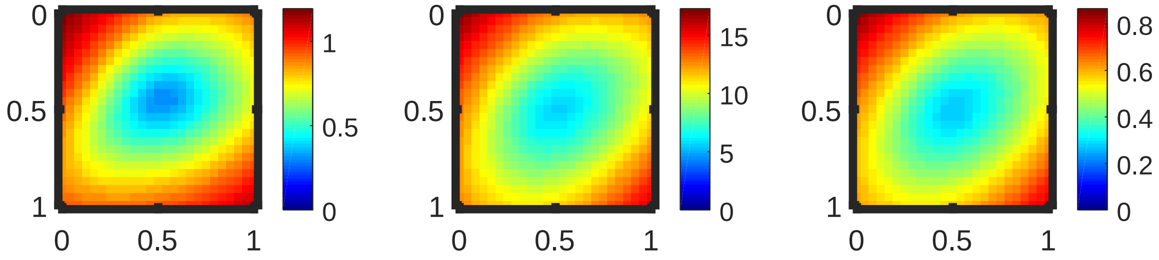
Figure 3.10: Results from Experiment 3 when $\lambda = 1$. In these plots $L\colon \mathbb{R}^{25\times 25} \to \mathbb{R}^{75\times 75}$. See Table 3.4 for the contrast.

the more overdetemined the problem the more $\mathrm{struc}\,[\cdot]$ outperforms the L^1 or L^2 discrete norms. These results are consistent with Thms. 1 - 2, which together suggest that as Δy decreases, the ability of $\mathrm{struc}\,[\cdot]$ to distinguish between noise and structure increases. Second the contrast is higher for all three metrics when λ increases. This observation suggests that



(a) $\operatorname{struc}[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

Figure 3.11: Results from Experiment 3 when $\lambda = 1$. In these plots $L\colon \mathbb{R}^{25\times 25} \to \mathbb{R}^{50\times 50}$. See Table 3.4 for the contrast.



(a) $\operatorname{struc}[r_{\theta,\eta}]$ vs θ (b) $\|r_{\theta,\eta}\|_1$ vs θ (c) $\|r_{\theta,\eta}\|_2$ vs θ

Figure 3.12: Results from Experiment 3 when $\lambda = 1$. In these plots $L\colon \mathbb{R}^{25\times 25} \to \mathbb{R}^{25\times 25}$. See Table 3.4 for the contrast.

it is easier to identify the correct operator when an inverse problem is heavily regularized. However this topic is beyond the scope of this manuscript and will be the subject of future work.

3.5 Conclusion

In this work we have developed a new functional called the structure, which is suitable for detecting forward operator error as it arises in inverse problems. The structure is defined by use of the Earth Mover's Distance (EMD), using a very rapid algorithm and a homogeneous degree one distance. The structure takes as input the residual from an existing inverse procedure, and can be computed quickly. We prove some apparently new results concerning

Contrast, $\lambda = 0.1$	struc $[r_{\theta,\eta}]$	$\|r_{\theta,\eta}\|_1$	$\|r_{\theta,\eta}\|_2$
$\mathbf{b} \in \mathbb{R}^{100\times 100}$	0.6715	0.5414	0.5450
$\mathbf{b} \in \mathbb{R}^{75\times 75}$	0.6099	0.5184	0.5217
$\mathbf{b} \in \mathbb{R}^{50\times 50}$	0.4777	0.4738	0.4734
$\mathbf{b} \in \mathbb{R}^{25\times 25}$	0.3261	0.3713	0.3741

Table 3.3: Results from Experiment 3. when $\lambda = 0.1$, which indicate the contrast. The larger the contrast the better.

Contrast, $\lambda = 1$	struc $[r_{\theta,\eta}]$	$\|r_{\theta,\eta}\|_1$	$\|r_{\theta,\eta}\|_2$
$\mathbf{b} \in \mathbb{R}^{100\times 100}$	0.8598	0.5607	0.5645
$\mathbf{b} \in \mathbb{R}^{75\times 75}$	0.8259	0.5604	0.5630
$\mathbf{b} \in \mathbb{R}^{50\times 50}$	0.7809	0.5520	0.5567
$\mathbf{b} \in \mathbb{R}^{25\times 25}$	0.5930	0.5068	0.5180

Table 3.4: Results from Experiment 3. when $\lambda = 1$, which indicate the contrast. The larger the contrast the better.

Contrast, $\lambda = 10$	struc $[r_{\theta,\eta}]$	$\|r_{\theta,\eta}\|_1$	$\|r_{\theta,\eta}\|_2$
$\mathbf{b} \in \mathbb{R}^{100\times 100}$	0.8908	0.5543	0.5567
$\mathbf{b} \in \mathbb{R}^{75\times 75}$	0.8357	0.5559	0.5568
$\mathbf{b} \in \mathbb{R}^{50\times 50}$	0.8018	0.5558	0.5599
$\mathbf{b} \in \mathbb{R}^{25\times 25}$	0.7717	0.5498	0.5568

Table 3.5: Results from Experiment 3. when $\lambda = 10$, which indicate the contrast. The larger the contrast the better.

the treatment of noise by EMD. Further, we consistent with these theoretical results we perform numerical experiments and show that the structure is able to distinguish between error in the modeling of a forward operator, and noise in the signal of an inverse problem.

The numerical results concern a model linear forward operator. On these problems the structure of the residual is indeed minimized when the correct forward operator is used. The

L^1 or L^2 norms of the residual are also minimized around the correct forward operator, the structure, however, is more localized and has better contrast around the minimum. Further, we observe that the degree to which the inverse problem is overdetermined and degree of regularization is critical to the success of the procedure. The more over determined the problem, the more useful the structure. This is borne out by the analysis in the case of linear regularization, as well as the numerical results on more sophisticated problem.

In the future, we will extend our work to more sophisticated non-linear operators and also use the $\operatorname{struc}[\cdot]$ to not only diagnose, but also to correct forward operator error. Assuming that the minimizer of the $\operatorname{struc}[\cdot]$ gives the correct model parameter $\hat{\theta}$ (as in Eq. 3.43), the next step is to actually solve the minimization to correct for operator error. By computing both $\operatorname{struc}[r_{\theta,\eta}]$ and $\nabla_\theta \operatorname{struc}[r_{\theta,\eta}]$ given $\nabla_\theta L_\theta$, one can do so using optimization methods such as gradient descent and BFGS [6]

3.6 Proofs

Proof of Proposition 1. Given $\Phi(\mathbf{v};\lambda) = \lambda \|\mathbf{C}\mathbf{v}\|_2^2$, the normal equations for Eq. 3.6 are

$$(\mathbf{L}_\theta^T \mathbf{L}_\theta + \lambda \mathbf{C}^T \mathbf{C})\tilde{\mathbf{u}}_{\theta,\eta} = \mathbf{L}_\theta^T(\mathbf{b} + \boldsymbol{\eta}). \tag{3.51}$$

Therefore $\tilde{\mathbf{L}}_\theta^{-1} = (\mathbf{L}_\theta^T \mathbf{L}_\theta + \lambda \mathbf{C}^T \mathbf{C}^T)^{-1} \mathbf{L}_\theta^T$. Using the GSVD in Eq. 3.18, a direct calculation gives

$$\mathbf{L}_\theta \tilde{\mathbf{L}}_\theta^{-1} = \mathbf{U}_\theta \mathbf{D}_{\theta,\lambda} \mathbf{U}_\theta^T, \quad \text{where } \mathbf{D}_{\theta,\lambda} := \frac{\boldsymbol{\Sigma}_\theta^2}{\boldsymbol{\Sigma}_\theta^2 + \lambda \boldsymbol{\Gamma}_\theta^2} \in \mathbb{R}^{n\times n}. \tag{3.52}$$

Thus according to the definition of the residual in Eq. 3.8,

$$\mathbf{r}_{\theta,\eta} = (\mathbf{I} - \mathbf{L}\tilde{\mathbf{L}}^{-1})(\mathbf{b} + \boldsymbol{\eta}) = \mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \mathbf{U}_\theta^T(\mathbf{b} + \boldsymbol{\eta}) + (\mathbf{I} - \mathbf{U}_\theta \mathbf{U}_\theta^T)(\mathbf{b} + \boldsymbol{\eta}) \tag{3.53}$$

where

$$\hat{\mathbf{D}}_{\theta,\lambda} := (\mathbf{I} - \mathbf{D}_{\theta,\lambda}) = \frac{\lambda \boldsymbol{\Gamma}_\theta^2}{\boldsymbol{\Sigma}_\theta^2 + \lambda \boldsymbol{\Gamma}_\theta^2} > 0. \tag{3.54}$$

We first bound two of the deterministic components of the residual. Using the GSVD,

$$\begin{aligned}\mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \mathbf{U}_\theta^T \mathbf{b} &= \mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \mathbf{U}_\theta^T \mathbf{L}_\theta \mathbf{u} + \mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \mathbf{U}_\theta^T(\mathbf{b} - \mathbf{L}_\theta \mathbf{u}) \\ &= \mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \boldsymbol{\Sigma}_\theta \mathbf{Z}_\theta^T \mathbf{u} + \mathbf{U}_\theta \hat{\mathbf{D}}_{\theta,\lambda} \mathbf{U}_\theta^T(\mathbf{b} - \mathbf{L}_\theta \mathbf{u}). \end{aligned} \tag{3.55}$$

Since $\left\|\hat{\mathbf{D}}_{\theta,\lambda}\right\|_2 \leq 1$ and $\mathbf{U}_\theta$ is orthogonal, it follows that

$$\left\|\mathbf{U}_\theta\hat{\mathbf{D}}_{\theta,\lambda}\mathbf{U}_\theta^T(\mathbf{b}-\mathbf{L}_\theta\mathbf{u})\right\|_2^2 \leq \|(\mathbf{b}-\mathbf{L}_\theta\mathbf{u})\|_2^2 \tag{3.56}$$

Furthermore, since

$$\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\Sigma}_\theta = \frac{\lambda\boldsymbol{\Gamma}_\theta^2\boldsymbol{\Sigma}_\theta}{\boldsymbol{\Sigma}_\theta^2+\lambda\boldsymbol{\Gamma}_\theta^2} \leq \frac{1}{2}\sqrt{\lambda}\boldsymbol{\Gamma}_\theta \leq \frac{1}{2}\sqrt{\lambda}\mathbf{I} \tag{3.57}$$

(where the inequalities between the diagonal matrices above are interpreted element-wise), it follows that

$$\left\|\mathbf{U}_\theta\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\Sigma}_\theta\mathbf{Z}_\theta^T\mathbf{u}\right\|_2^2 \leq \left\|\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\Sigma}_\theta\right\|_2^2\left\|\mathbf{Z}_\theta^T\mathbf{u}\right\|_2^2 \leq \frac{1}{4}\lambda\left\|\mathbf{Z}_\theta^T\mathbf{u}\right\|_2^2. \tag{3.58}$$

We next bound the noise component of the residual. Let $\mathbf{W}_\theta \in \mathbb{R}^{m\times(m-n)}$ be a matrix such that $\mathbf{Q} := (\mathbf{U}_\theta|\mathbf{W}_\theta) \in \mathbb{R}^{m\times m}$ is orthogonal and set

$$\boldsymbol{\alpha} = \begin{pmatrix}\boldsymbol{\alpha}_\parallel\\ \boldsymbol{\alpha}_\perp\end{pmatrix} := \mathbf{Q}^T\boldsymbol{\eta} = \begin{pmatrix}\mathbf{U}_\theta^T\boldsymbol{\eta}\\ \mathbf{W}_\theta^T\boldsymbol{\eta}\end{pmatrix}. \tag{3.59}$$

Then

$$\left\|(\mathbf{I}-\mathbf{L}\tilde{\mathbf{L}}^{-1})\boldsymbol{\eta}\right\|_2^2 = \left\|\mathbf{U}_\theta\hat{\mathbf{D}}_{\theta,\lambda}\mathbf{U}_\theta^T\boldsymbol{\eta}+(\mathbf{I}-\mathbf{U}_\theta\mathbf{U}_\theta^T)\boldsymbol{\eta}\right\|_2^2 = \left\|\mathbf{U}_\theta\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\alpha}_\parallel\right\|_2^2 + \|\mathbf{W}_\theta\boldsymbol{\alpha}_\perp\|_2^2, \tag{3.60}$$

where the last equality uses the fact that the columns of $\mathbf{U}_\theta$ and $\mathbf{W}_\theta$ are orthogonal and $\mathbf{I}-\mathbf{U}_\theta\mathbf{U}_\theta^T = \mathbf{W}_\theta\mathbf{W}_\theta^T$. Due to the spherical symmetry assumption on $\boldsymbol{\eta}$, $\boldsymbol{\alpha}_\parallel$ and $\boldsymbol{\alpha}_\perp$ are spherically symmetric random variables of dimension n and $m-n$, respectively, with components that are independent. Therefore

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{U}_\theta\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\alpha}_\parallel\right\|_2^2\right] &= \mathbb{E}\left[\left\|\hat{\mathbf{D}}_{\theta,\lambda}\boldsymbol{\alpha}_\parallel\right\|_2^2\right]\\ &= \sum_{i=1}^n\left(\frac{\lambda\gamma_i^2}{\sigma_i^2+\lambda\gamma_i^2}\right)^2\mathbb{E}\left[\boldsymbol{\eta}_i^2\right] = \frac{1}{m}\operatorname{Tr}(\hat{\mathbf{D}}_{\theta,\lambda}^2)\mathbb{E}\left[\|\boldsymbol{\eta}\|_2^2\right]\end{aligned} \tag{3.61}$$

and

$$\mathbb{E}\left[\mathbf{W}_\theta\|\boldsymbol{\alpha}_\perp\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{\alpha}_\perp\|_2^2\right] = \frac{m-n}{m}\mathbb{E}\left[\|\boldsymbol{\eta}\|_2^2\right]. \tag{3.62}$$

This completes the proof. □

Proof of Proposition 2. It is convenient to write Eq. 3.11 in the abstract form

$$\operatorname{EMD}(\rho_1,\rho_2) = \min_{m\in C(\rho_1,\rho_2)}\mathcal{T}(m), \tag{3.63}$$

where

$$\mathcal{T}(m) = \int_{\Omega} \|m\|_2 \, dx \tag{3.64}$$

$$C(\rho_1, \rho_2) = \left\{ m : \begin{array}{c} \nabla \cdot m(x) + \rho_2(x) - \rho_1(x) = 0 \quad \forall x \in \Omega, \\ m(x) \cdot n(x) = 0 \quad \forall x \in \partial\Omega \end{array} \right\}. \tag{3.65}$$

In addition, for any $f \in L^1(\Omega)$, let m_f be a minimizer of $\mathcal{T}(m)$ over $C(f^+, f^-)$ so that $\operatorname{struc}[f] = \mathcal{T}(m_f)$.

1. We check absolute homogeneity, positivity, and the triangle inequality.

   (a) To check absolute homogeneity, let $\lambda \in \mathbb{R}$ be a nonzero scalar. By linearity, $|\lambda| m \in C(|\lambda| f, |\lambda| g)$ if and only if $m \in C(f, g)$. Therefore

   $$\begin{aligned} \mathrm{EMD}(|\lambda| f, |\lambda| g) &= \min_{m \in C(|\lambda| f, |\lambda| g)} \mathcal{T}(m) \\ &= \min_{m \in C(f,g)} \mathcal{T}(|\lambda| m) = |\lambda| \min_{m \in C(f,g)} \mathcal{T}(m) = |\lambda| \, \mathrm{EMD}(f, g), \end{aligned} \tag{3.66}$$

   If $\lambda > 0$, Eq. 3.66 implies that

   $$\operatorname{struc}[\lambda f] = \mathrm{EMD}(\lambda f^+, \lambda f^-) = |\lambda| \, \mathrm{EMD}(f^+, f^-) = |\lambda| \operatorname{struc}[f] \tag{3.67}$$

   If $\lambda < 0$, then $(\lambda f)^{\pm} = |\lambda| f^{\mp}$. Again Eq. 3.66 implies that

   $$\begin{aligned} \operatorname{struc}[\lambda f] &= \mathrm{EMD}((\lambda f)^+, (\lambda f)^-) = \mathrm{EMD}(|\lambda| f^-, |\lambda| f^+) \\ &= |\lambda| \, \mathrm{EMD}(f^-, f^+) = |\lambda| \, \mathrm{EMD}(f^+, f^-) = |\lambda| \operatorname{struc}[f]. \end{aligned} \tag{3.68}$$

   Finally, if $\lambda = 0$, then the fact that $\operatorname{struc}[\lambda f] = \lambda \operatorname{struc}[f] = 0$ is trivial.

   (b) Positivity follows immediately from the positivity of EMD.

   (c) The triangle inequality follows from the fact that

   $$(f+g)^+ - (f+g)^- = (f^+ - f^-) + (g^+ - g^-) \tag{3.69}$$

   for all $f, g \in L^1(\Omega)$. Thus if $m_f \in C(f^+, f^-)$ and $m_g \in C(g^+, g^-)$, then $m_f + m_g \in C\left((f+g)^+, (f+g)^-\right)$. Along with the triangle inequality for $\mathcal{T}$, this implies that

   $$\operatorname{struc}[f+g] \equiv \mathcal{T}(m_{f+g}) \leq \mathcal{T}(m_f + m_g) \leq \mathcal{T}(m_f) + \mathcal{T}(m_g) \equiv \operatorname{struc}[f] + \operatorname{struc}[g]. \tag{3.70}$$

2. Because $\frac{1}{\|\Omega\|}\int_\Omega (g+c)dx = \frac{1}{\|\Omega\|}\int_\Omega g dx + c$, we have that $g^+ = (g+c)^+$, and $g^- = (g+c)^-$. Therefore

$$\text{struc}\,[g+c] = \text{EMD}\left((g+c)^+, (g+c)^-\right) = \text{EMD}(g^+, g^-) = \text{struc}\,[g]\,. \tag{3.71}$$

3. Let $g = 0$ in Eq. 3.71 above. Then

$$\text{struc}\,[c] = \text{struc}\,[0] = 0, \quad \forall c \in \mathbb{R}. \tag{3.72}$$

4. Because the constraint in Eq. 3.11 involves only the difference of ρ_1 and ρ_2, it follows that $\text{EMD}(\rho_1, \rho_2) = \text{EMD}(\rho_1 + f, \rho_2 + f)$ for any non-negative $f \in L^1(\Omega)$. Moreover, because ρ_2 and ρ_1 have the same mass, the average of $\rho_2 - \rho_1$ is zero. Hence,

$$\begin{aligned}\text{struc}\,[\rho_2 - \rho_1] &= \text{EMD}(\max(\rho_2 - \rho_1, 0), \max(\rho_1 - \rho_2, 0)) \\ &= \text{EMD}(\max(\rho_2 - \rho_1, 0) + \min(\rho_1, \rho_2), \max(\rho_1 - \rho_2, 0) + \min(\rho_1, \rho_2))\end{aligned} \tag{3.73}$$

Since $\forall x, y \in \mathbb{R}, \max(x - y, 0) + \min(x, y) = x$, it follows from Eq. 3.73 that

$$\text{struc}\,[\rho_2 - \rho_1] = \text{EMD}(\rho_2, \rho_1) = \text{EMD}(\rho_1, \rho_2) \tag{3.74}$$

□

Before proving Thm. 1-2, we will first prove two useful lemmas, which will be used extensively.

Lemma 2 (EMD triangle inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded set and f, g, $h \in L^\infty(\Omega)$ and $\int_\Omega f dx = \int_\Omega h dx = \int_\Omega g dx$. Then*

$$\text{EMD}(f, g) \leq \text{EMD}(f, h) + \text{EMD}(h, g). \tag{3.75}$$

Proof. Recall from Prop. 2 that $\text{struc}\,[f - g] = \text{EMD}(f, g)$, then by the triangle inequality of $\text{struc}\,[\cdot]$,

$$\text{EMD}(f, g) = \text{struc}\,[f - g] \leq \text{struc}\,[f - h] + \text{struc}\,[h - g] = \text{EMD}(f, h) + \text{EMD}(h, g) \tag{3.76}$$

□

Lemma 3 (struc $[\cdot]$ and EMD of the mean)**.** *$\Omega \subset \mathbb{R}^n$ be a bounded set and $f \in L^\infty(\Omega)$ and $\mu = \frac{1}{|\Omega|}\int_\Omega f dx$. Then*

$$\text{struc}\,[f] = \text{EMD}(f, \mu). \tag{3.77}$$

Proof. Recall from Prop. 2 that $\text{EMD}(f, g) = \text{EMD}(f + h, g + h)$, therefore

$$\text{struc}\,[f] = \text{EMD}(f^+, f^-) = \text{EMD}(f^+ + (\mu - f^-), f^- + (\mu - f^-)) = \text{EMD}(f, \mu). \tag{3.78}$$

□

Lemma 4 (EMD Subadditivity)**.** *If $\text{EMD}(f_1, g_1)$ and $\text{EMD}(f_2, g_2)$ are well defined, then so too is $\text{EMD}(f_1 + f_2, g_1 + g_2)$, and*

$$\text{EMD}(f_1 + f_2, g_1 + g_2) \leq \text{EMD}(f_1, g_1) + \text{EMD}(f_2, g_2). \tag{3.79}$$

Proof. We use the Eq. 3.10 of the EMD. Let π_1 and π_2 satisfy the constraint of Eq. 3.9 for $\text{EMD}(f_1, g_1)$ and $\text{EMD}(f_2, g_2)$ resp. Then clearly

$$\begin{aligned} \int_\Omega (\pi_1 + \pi_2) dx^{(2)} &= f_1 + f_2 \\ \int_\Omega (\pi_1 + \pi_2) dx^{(1)} &= g_1 + g_2 \\ \pi_1 + \pi_2 &\geq 0, \end{aligned} \tag{3.80}$$

and so by the minimality of the EMD,

$$\begin{aligned} \text{EMD}(f_1, g_1) + \text{EMD}(f_2, g_2) &= \int_{\Omega\times\Omega} c\pi_1 dx^{(1)} dx^{(2)} + \int_{\Omega\times\Omega} c\pi_2 dx^{(1)} dx^{(2)} \\ &= \int_{\Omega\times\Omega} c(\pi_1 + \pi_2) dx^{(1)} dx^{(2)} \\ &\geq \min_{\pi \geq 0} \int_{\Omega\times\Omega} c\pi dx^{(1)} dx^{(2)} \\ &= \text{EMD}(f_1 + f_2, g_1 + g_2) \end{aligned} \tag{3.81}$$

where π is subject to the constraints of Eq. 3.9 where $\rho_1 = f_1 + f_2$ and $\rho_2 = g_1 + g_2$. □

Lemma 5 (EMD is bounded by the L^1 norm)**.** *Let Ω be a bounded set, and $l \geq \left\| x^{(1)} - x^{(2)} \right\|_2$ for all $x^{(1)}, x^{(2)} \in \Omega$. If $f, g : \Omega \to \mathbb{R}^+$ then*

$$\text{EMD}(f, g) \leq \frac{l}{2} \left\| f - g \right\|_{L^1(\Omega)}. \tag{3.82}$$

Proof. Let $\gamma = \int_\Omega (f-g)^+ dx$ and x^{c} be such that $\|x^{\mathrm{c}} - x\|_2 \leq l/2 \ \forall x \in \Omega$ then

$$\begin{aligned}\mathrm{EMD}(f,g) = \mathrm{struc}\,[f-g] &\leq \mathrm{EMD}((f-g)^+, \gamma\delta_{x^{\mathrm{c}}}) + \mathrm{EMD}(\gamma\delta_{x^{\mathrm{c}}}, (f-g)^-) \\ &\leq \frac{l}{2}\left\|(f-g)^+\right\|_{L^1(\Omega)} + \frac{l}{2}\left\|(f-g)^-\right\|_{L^1(\Omega)} = \frac{l}{2}\|f-g\|_{L^1(\Omega)} \end{aligned} \quad (3.83)$$

The last two lines could use a few details between them. □

Lemma 6 (Expectation bound by the standard deviation). *Let* η *be a scalar random variable with zero mean such that* $\mathrm{Var}[\eta]$ *is finite. Then* $\mathbb{E}\left[|\eta|\right] \leq \sqrt{\mathrm{Var}[\eta]}$.

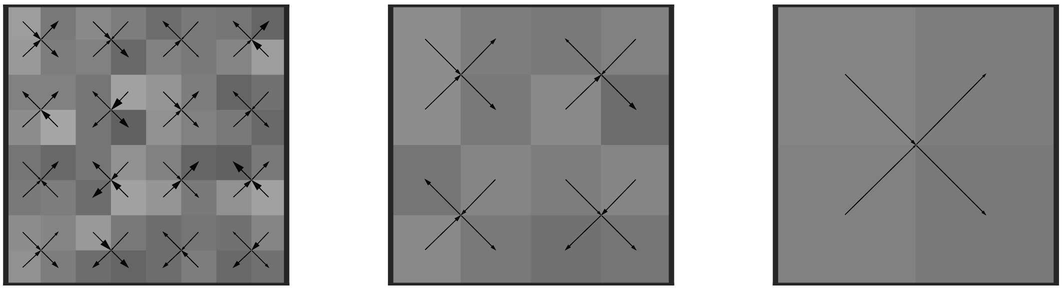
Proof. Let ψ be the probability distribution for η. By the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[|\eta|\right] \equiv \int_{-\infty}^{\infty} |x|\psi(x)dx \leq \left(\int_{-\infty}^{\infty} x^2\psi(x)dx\right)^{\frac{1}{2}} \left(\int_{-\infty}^{\infty} \psi(x)dx\right)^{\frac{1}{2}} = \Big(\mathrm{Var}[\eta]\Big)^{1/2}. \quad (3.84)$$

□

We now proceed to the proof of Theorem 1, but first it is helpful to give a brief summary. To bound the EMD from above, we give a candidate transport plan that is based on the multigrid strategy depicted in Fig. 3.13 for the case $d = 2$. In this case, the strategy is to divide the domain into square windows with two square panels per side, as shown in Figure 3.13. The mass in each window is then redistributed in such a way that the new distribution is constant on each window. Each window then becomes a panel in a window that is a factor a factor of two larger in each dimension, and the process is repeated until the distribution on the entire square is constant. For $d > 2$, the plan is the same, except that each window is a hypercube 2^d panels. The cost of the complete transport plan can be bounded by the sum of the costs of the transport plan for each step. These costs are computed in the proof below and their sum leads to the bound in Theorem 1.

Proof of Theorem 1. Since $\mathrm{struc}\,[h_\ell] = \mathrm{struc}\,[h_\ell - \mu]$ we can assume, without loss of generality, that h_ℓ has zero mean. Consider the case $\ell = 1$, which will be used for the general



(a) H_3 (b) H_2 (c) H_1

Figure 3.13: The multigrid idea of Theorem 1 when $\ell = 3$. At each step, a transport plan is computed in each 2x2 window. Then the same problem is solved at the next coarser scale. In the above figures, the arrow tip area is proportional to the mass transported at each substep. The function H_i is defined in Eq. 3.94.

setting later. We construct a two-step plan that first moves all of the mass in h_1^+ to the point $y^{\mathrm{c}} = (1/2, \ldots, 1/2)$ at the center of the domain and then moves the mass from y^{c} to h_1^-.[2]

Let $\gamma = \int_\Omega h_1^+ dy = \int_\Omega h_1^- dy$, $\mu_0 = \int_\Omega h_1 dy$, and $\gamma_{1,k} = |\eta_{1,k} - \mu_0||\omega_{1,k}|$. Then $\mathrm{EMD}(h_1^+, \gamma\delta_{y^{\mathrm{c}}})$ $= \mathrm{EMD}(\gamma\delta_{y^{\mathrm{c}}}, h_1^-)$ and

$$\begin{aligned} \operatorname{struc}[h_1] \equiv \mathrm{EMD}(h_1^+, h_1^-) &\leq \mathrm{EMD}(h_1^+, \gamma\delta_{y^{\mathrm{c}}}) + \mathrm{EMD}(\gamma\delta_{y^{\mathrm{c}}}, h_1^-) \\ &= \sum_{k=1}^{2^d} \mathrm{EMD}\left(|\eta_{1,k} - \mu_0|\chi_{1,k}, \gamma_{1,k}\delta_{y^{\mathrm{c}}}\right). \end{aligned} \tag{3.85}$$

Thus we turn our attention to computing the terms in the sum above. First,

$$\mathrm{EMD}(|\eta_{1,k} - \mu_0|\chi_{1,k}, \gamma_{1,k}\delta_{y^{\mathrm{c}}}) = |\eta_{1,k} - \mu_0|\,\mathrm{EMD}(\chi_{1,k}, |\omega_{1,k}|\,\delta_{y^{\mathrm{c}}}). \tag{3.86}$$

There is only one one admissible transport plan (see from Eq. 3.10) between $\chi_{1,k}$ and $|\omega_{1,k}|\delta_{y^{\mathrm{c}}}$; it simply moves the mass around each point of $\omega_{1,k}$ to y^{c}:

$$\pi\left(x^{(1)}, x^{(2)}\right) = \chi_{1,k}(x^{(1)}) \times \delta_{y^{\mathrm{c}}}(x^{(2)}) \tag{3.87}$$

If we consider the more general case where $\omega_{1,k}$ has side length l, then upon a change of coordinates,

[2] While the definition of the EMD in Eq. 3.10 is still well-defined for delta function, the formula in Eq. 3.11 is not. Thus while we use Eq. 3.11 for numerical calculations, we often rely on Eq. 3.10 for theoretical bounds.

$$\begin{aligned}
\mathrm{EMD}(\chi_{1,k}, |\omega_{1,k}|\delta_{y^c}) &= \int_\Omega \int_\Omega \left\|x^{(1)} - x^{(2)}\right\|_2 \chi_{1,k}(x^{(1)}) \times \delta_{y^c}(x^{(2)}) dx^{(1)} dx^{(2)} \\
&= \int_{\omega_{1,k}} \int_\Omega \left\|x^{(1)} - x^{(2)}\right\|_2 \delta_{y^c}(x^{(2)}) dx^{(1)} dx^{(2)} \\
&= \int_{\omega_{1,k}} \left\|x^{(1)} - y^c\right\|_2 dx^{(1)} = \int_{[0,l]^d} \left\|x^{(1)}\right\|_2 dx^{(1)} \\
&\leq \sqrt{d} \int_{[0,l]^d} \left\|x^{(1)}\right\|_\infty dx^{(1)} \leq \sqrt{d}\frac{l^{d+1}}{2}
\end{aligned} \tag{3.88}$$

Substituting Eq. 3.86 and Eq. 3.88 into Eq. 3.85 gives

$$\mathrm{struc}\left[h_1\right] \leq \sum_{k=1}^{2^d} |\eta_{1,k} - \mu_0| \frac{\sqrt{d} l^{d+1}}{2} = \frac{\sqrt{d}}{2^{d+2}} \sum_{k=1}^{2^d} |\eta_{1,k} - \mu_0|, \tag{3.89}$$

where we have used the fact that when $\ell = 1, l = 2^{-1}$. A standard calculation shows that

$$\mathrm{Var}\left(|\eta_{1,k} - \mu_0|\right) \leq \mathrm{Var}(|\eta_{1,k}|), \quad i = 1, \ldots, 2^d. \tag{3.90}$$

Further $\mathbb{E}\left[\eta_{1,k}\right] = 0$ and so Lemma 6 give:

$$\mathbb{E}\left[|\eta_{1,k} - \mu_0|\right] \leq \sigma \tag{3.91}$$

with Eq. 3.89 and get

$$\mathbb{E}\left[\mathrm{struc}\left[h_1\right]\right] \leq \frac{\sqrt{d}2^d}{2^{(d+2)}} \sum_{k=1}^{2^d} \mathbb{E}\left[|n_{1,k} - \mu_0|\right] \leq \frac{\sqrt{d}2^d}{2^{(d+2)}}\sigma = \frac{\sqrt{d}}{4}\sigma. \tag{3.92}$$

Now we consider the case when $\ell > 1$. Define the functions

$$H_\ell(y) = h_\ell(y) = \sum_{k=1}^{2^{\ell d}} \eta_{\ell,k} \chi_{\ell,k}(y) \tag{3.93}$$

$$H_i(y) = \sum_{k=1}^{2^{id}} \mu_{i,k} \chi_{i,k}(y), \text{ where } \mu_{i,k} = \frac{1}{|\omega_{i,k}|} \int_{\omega_{i,k}} H_{i+1}(y) dy, \quad i = 0, 1, \ldots, \ell - 1. \tag{3.94}$$

Instances of H_i are shown in Fig. 3.13. The function h_ℓ can be written as the telescoping sum

$$h_\ell = H_\ell = (H_\ell - H_{\ell-1}) + (H_{\ell-1} - H_{\ell-2}) + \cdots + (H_2 - H_1) + (H_1 - H_0) + H_0. \tag{3.95}$$

Moreover, because $H_i = \sum_{k=1}^{2^{d(i-1)}} H_i \chi_{i-1,k}$, it follows that

$$H_i - H_{i-1} = \sum_{k=1}^{2^{d(i-1)}} s_{i-1,k}, \quad \text{where } s_{i-1,k}(y) = \left(H_i(y) - \mu_{i-1,k}\right) \chi_{i-1,k}(y). \tag{3.96}$$

We apply struc $[\cdot]$ to Eq. 3.95, using Eq. 3.96, the triangle inequality, and the fact that struc $[H_0] = 0$ (because it is a constant). The result is

$$\text{struc}\left[h_\ell\right] \le \sum_{i=1}^{\ell} \text{struc}\left[H_i - H_{i-1}\right] \le \sum_{i=1}^{\ell} \sum_{k=1}^{2^{d(i-1)}} \text{struc}\left[s_{i-1,k}\right]. \tag{3.97}$$

To evaluate struc $[s_{i-1,k}]$, we repeat the argument used to generate Eq. 3.89. This gives

$$\text{struc}\left[s_{i-1,k}\right] \equiv \text{EMD}(s^+_{i-1,k}, s^-_{i-1,k}) \le \frac{\sqrt{d} l^{d+1}}{2} \sum_{k' : \omega_{i,k'} \subset \omega_{i-1,k}} |\mu_{i,k'} - \mu_{i-1,k}|. \tag{3.98}$$

By construction,

$$\mu_{i-1,k} = 2^{-d} \sum_{k' : \omega_{i,k'} \subset \omega_{i-1,k}} \mu_{i,k}. \tag{3.99}$$

It follows that the random variable $(\mu_{i+1,k'} - \mu_{i,k})$ that appears in Eq. 3.98 has zero mean. Thus Lemma 6 applies and

$$\mathbb{E}\left[|\mu_{i,k'} - \mu_{i-1,k}|\right] \le \left(\text{Var}[|\mu_{i,k'} - \mu_{i-1,k}|]\right)^{\frac{1}{2}} \le \left(\text{Var}[|\mu_{i,k'}|]\right)^{\frac{1}{2}} := \sigma_i, \tag{3.100}$$

where the last two inequalities above follows from standard probability theory. Also, because of Eq. 3.99, another standard probablity result gives

$$\sigma_i = 2^{-\frac{d}{2}} \sigma_{i+1} = \cdots = 2^{-\frac{d}{2}(\ell - i)} \sigma_\ell, \quad i = 1, \ldots, \ell. \tag{3.101}$$

We now take the expectation of Eq. 3.98, using the fact that $\omega_{i,k'}$ has side length $l = 2^{-i}$, along with the triangle inequality and Eq. 3.101. The result is

$$\mathbb{E}\left[\text{struc}\left[s_{i-1,k}\right]\right] \le \sqrt{d} 2^{-i(d+1)-1} \sum_{k' : \omega_{i,k'} \subset \omega_{i-1,k}} 2^{-\frac{d}{2}(\ell - i)} \sigma_\ell = \sqrt{d} 2^{-\frac{id}{2} - i + d - \frac{d\ell}{2} - 1} \sigma_\ell \tag{3.102}$$

Substituting this bound into Eq. 3.97 gives

$$\mathbb{E}\left[\text{struc}\left[h_\ell\right]\right] \le \sum_{i=1}^{\ell} \sum_{k=1}^{2^{d(i-1)}} \sqrt{d} 2^{-\frac{id}{2} - i + d - \frac{d\ell}{2} - 1} \sigma_\ell = \frac{\sqrt{d} \sigma_\ell}{2^{1 + \frac{\ell d}{2}}} \sum_{i=1}^{\ell} \left(2^{\frac{d}{2} - 1}\right)^i \tag{3.103}$$

If $d = 2$, then $2^{\frac{d}{2}-1} = 1$ and Eq. 3.103 becomes

$$\mathbb{E}\left[\text{struc}\left[h_\ell\right]\right] = \mathbb{E}\left[\text{struc}\left[H_\ell\right]\right] \leq \frac{2\sigma_\ell}{2^{1+i}}\ell = \frac{\sigma_\ell \ell}{2^\ell}. \tag{3.104}$$

If $d \geq 3$, then $2^{\frac{d}{2}-1}/(2^{\frac{d}{2}-1} - 1) \leq 4$, so the geometric sum in Eq. 3.103 is

$$\sum_{i=1}^{\ell} \left(2^{\frac{d}{2}-1}\right)^i = \frac{2^{\left(\frac{d}{2}-1\right)(\ell+1)} - 2^{\frac{d}{2}-1}}{2^{\frac{d}{2}-1} - 1} \leq \frac{2^{\frac{d}{2}-1} 2^{\left(\frac{d}{2}-1\right)\ell}}{2^{\frac{d}{2}-1} - 1} \leq 2^{\frac{\ell d}{2} - \ell + 2}. \tag{3.105}$$

Thus for $d \geq 3$,

$$\mathbb{E}\left[\text{struc}\left[h_\ell\right]\right] \leq \sqrt{d}\sigma_\ell \frac{2^{\frac{\ell\sqrt{d}}{2} - \ell + 2}}{2^{1+\frac{\ell\sqrt{d}}{2}}} = \sqrt{d}\sigma_\ell 2^{-\ell+1} \tag{3.106}$$

Finally, setting $\epsilon_\ell = 2^{-\ell}$ gives

$$\mathbb{E}\left[\text{struc}\left[h_\ell\right]\right] \leq \sigma \begin{cases} -\epsilon_\ell \log(\epsilon_\ell) & \text{when } d = 2 \\ 2\sqrt{d}\epsilon_\ell & \text{when } d > 2 \end{cases} \tag{3.107}$$

This completes the proof. □

Proof of Lemma 1. The proof follows directly from the definition of h_ℓ in the statement of Thm. 1:

$$\mathbb{E}\left[\|h_\ell\|_2^2\right] = \mathbb{E}\left[\int_{[0,1)^d} \left(h_\ell(y)\right)^2 dy\right] = \sum_{k=1}^{2^{\ell d}} \mathbb{E}\left[\eta_{\ell,k}^2\right] 2^{-\ell d} = 2^{-\ell d} \sum_{k=1}^{2^{\ell d}} \sigma^2 = \sigma^2. \tag{3.108}$$

□

Proof of Theorem 2. Without loss of generality, assume that ϕ is positive a.e. (If not, simply replace ϕ by $\phi - \operatorname{ess\,inf} \phi$ and use Eq. 3.71.) By construction, ϕ and $R_\ell\phi$ have the same average over Y, which we denote by μ. Thus by Lemmas 2 and 3,

$$\text{struc}\left[R_\ell\phi\right] = \text{EMD}(R_\ell\phi, \mu) \leq \text{EMD}(R_\ell\phi, \phi) + \text{EMD}(\phi, \mu) = \text{EMD}(R_\ell\phi, \phi) + \text{struc}\left[\phi\right]. \tag{3.109}$$

Hence

$$\text{struc}\left[R_\ell\phi\right] - \text{struc}\left[\phi\right] \leq \text{EMD}(R_\ell\phi, \phi). \tag{3.110}$$

One the other hand, switching the roles of $R_\ell\phi$ and ϕ Eq. 3.109 gives

$$\text{struc}\,[\phi] - \text{struc}\,[R_\ell\phi] \leq \text{EMD}(R_\ell\phi, \phi) \tag{3.111}$$

Together Eq. 3.110 and Eq. 3.110 imply the bound

$$|\,\text{struc}\,[R_\ell\phi] - \text{struc}\,[R_\ell\phi]\,| \leq \text{EMD}(R_\ell\phi, \phi). \tag{3.112}$$

We now bound $\text{EMD}(R_\ell\phi, \phi)$. For any ℓ, i $\int_{\omega_{\ell,i}} R_\ell\phi dy = \int_{\omega_{\ell,i}} \phi dy$. Thus by Lemma 4,

$$\text{EMD}(R_\ell\phi, \phi) \leq \sum_{i=1}^{2^{\ell d}} \text{EMD}(R_\ell\phi\chi_{\ell,i}, \phi\chi_{\ell,i}) \tag{3.113}$$

and further by Lemma 5, for $i = 1, \ldots, 2^{\ell d}$

$$\text{EMD}(R_\ell\phi\chi_{\ell,i}, \phi\chi_{\ell,i}) \leq \|R_\ell\phi - \phi\|_{L^1(\omega_{\ell,i})}\, d^{1/2}2^{-\ell} \tag{3.114}$$

Now we bound $\|R_\ell\phi - \phi\|_{L^1(\omega_{\ell,i})}$. Since $\phi \in C^1\left(\overline{Y}\right)$, it follows that, for $y \in \omega_{\ell,i}$

$$\begin{aligned}|R_\ell\phi(y) - \phi(y)| &= \frac{1}{|\omega_{\ell,i}|}\left|\int_{\omega_{\ell,i}} (\phi(y') - \phi(y))dy'\right| \\ &\leq \sup_{y\in\omega_{\ell,i}} |\nabla\phi(y)| \sup_{y\in\omega_{\ell,i}} |y' - y| \leq d^{1/2}2^{-\ell} \sup_{y\in\omega_{\ell,i}} |\nabla\phi(y)|\end{aligned} \tag{3.115}$$

Therefore

$$\|R_\ell\phi - \phi\|_{L^1(\omega_{\ell,i})} \leq |\omega_{\ell,i}| d^{1/2}2^{-\ell} \sup_{y\in\omega_{\ell,i}} |\nabla\phi(y)| = d^{1/2}2^{-(d+1)\ell} \sup_{y\in\omega_{\ell,i}} |\nabla\phi(y)|. \tag{3.116}$$

Combining Eq. 3.112, Eq. 3.114, and Eq. 3.116 yields

$$|\,\text{struc}\,[R_\ell\phi] - \text{struc}\,[\phi]\,| \leq \sum_{i=1}^{2^{\ell d}} d2^{-(d+2)\ell} \sup_{y\in\omega_{\ell,i}} |\nabla\phi(y)| \leq d2^{-2\ell} \sup_{y\in Y} |\nabla\phi(y)| \equiv C(|\nabla\phi|)d\epsilon_\ell^2, \tag{3.117}$$

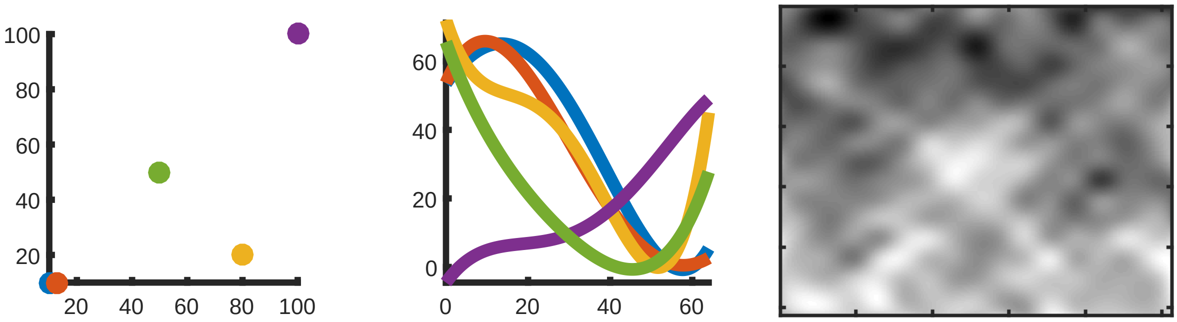
where $C(|\nabla\phi|) = \sup_{y\in Y} |\nabla\phi(y)|$ and $\epsilon_\ell = 2^{-\ell}$. This completes the proof. □

3.7 Line Integral Operators

Recall from Section 3.2 the spaces $\mathcal{U}$ and $\mathcal{B}$ of functions defined on domains X and Y, respectively. An operator $\mathcal{L} \colon \mathcal{U} \to \mathcal{B}$ is a line integral operators (LIO), if $\forall f \in \mathcal{U}$,

$$(\mathcal{L}f)(y) = \int_{P_y} f(x) d\ell_x = \int_0^1 f(\hat{x}(t;y))\hat{x}'(t;y)dt, \tag{3.118}$$

where for each $y \in Y$, $P_y = \{\hat{x}(t;y) : t \in (0,1)\} \subset X$, and $\hat{x}(t;y)$ is continuous in t and y. In particular, if f is a continuous on X, then $\mathcal{L}f$ is continuous on Y. Figs. 3.14a and 3.14b illustrate a LIO in two dimensions. The recipe we used to generate examples of $\hat{x}$ is given below.



(a) The values of y.

(b) The curves P_y.

(c) Example of Perlin noise.

Figure 3.14: An example of a LIO. Points on the right are used to generate curves on the left of the same color. Coefficients for the parameterization in Eq. 3.120 of P_y come from Perlin noise.

To discretize $\mathcal{L}$, we generate a path P_y for each hypercube $\omega \subset Y$. Line integrals along these paths are approximated via quadrature. For all LIOs, we use same the quadratures, and X, and Y.

To construct the LIO for Experiments 1 - 3, we do the following.

1. Construction of numerical grids. In all of the computational examples, the domains X and Y are unit squares in $\mathbb{R}^2$. We discretize these domains with N^x and N^y points, respectively, on each side and define grid points

$$x_{i,j} = (i\Delta x, j\Delta x), \quad 0 \leq i, j \leq N^x - 1, \tag{3.119a}$$

$$y_{k,l} = (k\Delta y, l\Delta y), \quad 0 \leq k, l \leq N^y - 1, \tag{3.119b}$$

where $\Delta x = 1/N^x$ and $\Delta y = 1/N^y$. We then generate values $u_{i,j}$ by sampling a prescribed function at the points $x_{i,j}$. An illustrative example is given in Fig. 3.3a, where piecewise smooth rings have been sampled on a 64×64 grid.

2. **G**eneration of smooth paths. To form $\hat{x}$, we first sample coefficients $\alpha_{p,r}$ for $p = 0, \dots, 4$ and $r = 1, 2$ from Perlin noise [34, 35] of order four. In Fig. 3.14c, a realization of one such coefficient as a function of y is shown on a 256×256 grid. Given these coefficients, we let $\bar{x} = (x^{(1)}, x^{(2)})$ be polynomials in t:

$$\bar{x}^{(r)}(t; y_{k,l}) = \sum_{p=0}^{4} \frac{\alpha_{p,r}(y_{k,l})}{p!} t^p, \quad r = 1, 2, \tag{3.120}$$

and then let $\hat{x}$ be the following normalization of $\bar{x}$:

$$\hat{x}^{(r)}(t; y_{k,l}) = \frac{\bar{x}^{(r)}(t; y_{k,l}) - \min_{s \in [0,1]} \bar{x}^{(r)}(s; y_{k,l})}{\max_{s \in [0,1]} \bar{x}^{(r)}(s; y_{k,l}) - \min_{s \in [0,1]} \bar{x}^{(r)}(s; y_{k,l})}, \quad r = 1, 2. \tag{3.121}$$

3. Let the paths be given as

$$\hat{x}(t; y_{k,l}) = \left(\hat{x}^{(1)}(t; y_{k,l}), \hat{x}^{(2)}(t; y_{k,l})\right) \tag{3.122a}$$

$$P_{y_{k,l}} = \{\hat{x}(t; y_{k,l}) \colon t \in [0, 1]\}. \tag{3.122b}$$

To discretize $\mathcal{L}$ we approximate the integral in Eq. 3.118, for each grid point $y_{k,l} \subset Y$, using an arc length parameterization of the curve $P_{y_{k,l}}$. The resulting quadrature takes the form

$$(\mathcal{L}f)(y_{k,l}) \approx \sum_{q} w_q f(x_q) \tag{3.123}$$

where $\{x_q\} \subset X$ and each weight $w_q > 0$. Because this quadrature involves points x_q not on the computational grid, we approximate the value $f(x_q)$ by interpolating the grid function values $f(x_{i,j})$. The result takes the form

$$(\mathcal{L}f)(y_{k,l}) \approx \sum_{i,j} L_{(k,l),(i,j)} f(x_{i,j}), \tag{3.124}$$

where the values $L_{(k,l),(i,j)}$ are now the components of the matrix operator $\mathbf{L}$.

CHAPTER 4

Unnormalized Optimal Transport

4.1 Background

This chapter is taken from the work [18] of which I am a coauthor. Specifically, in this thesis I will present a brief background of the theoretical properties and the numerics as well as their interpretation.

Recall first that the L^p Wasserstein distance can be defined as:

$$\begin{aligned} \mathrm{W}_p\left(\mu_0, \mu_1\right)^2 = \inf_{v,\mu,f} & \int_0^1 \int_\Omega \|v(t,x)\|^2 \mu(t,x) dx dt \\ \text{subject to} \quad & \partial_t \mu(t,x) + \nabla \cdot (\mu(t,x) v(t,x)) = 0, \\ & \mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x). \end{aligned} \tag{4.1}$$

Based on this definition we define the L^2 Unnormalized Wasserstein distance as a modification of Eqn. 4.1 to be:

$$\begin{aligned} \mathrm{UW}_2\left(\mu_0, \mu_1\right)^2 = \inf_{v,\mu,f} & \int_0^1 \int_\Omega \|v(t,x)\|^2 \mu(t,x) dx dt + \frac{|\Omega|}{\alpha} \int_0^1 |f(t)|^2 dt \\ \text{subject to} \quad & \partial_t \mu(t,x) + \nabla \cdot (\mu(t,x) v(t,x)) = f(t), \\ & \mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x). \end{aligned} \tag{4.2}$$

4.1.1 Unnormalized Wasserstein-1 Distance

If we define

$$m(x) = \int_0^1 v(t,x) \mu(t,x) dt \tag{4.3}$$

then by Jensen's inequality,

$$\int_0^1 \int_\Omega \|v(t,x)\| \mu(t,x) dx dt \geq \int_\Omega \| \int_0^1 v(t,x)\mu(t,x)dt\| dx = \int_\Omega \|m(x)\| dx. \tag{4.4}$$

Thus, the minimizing m is time independent, therefore Eqn. 4.2 becomes:

$$\begin{aligned} \mathrm{UW}_1\left(\mu_0,\mu_1\right) &= \int_\Omega \|m(x)\|\, dx + \frac{1}{\alpha}\left|\int_\Omega \mu_0(x) - \mu_1(x) dx\right| \\ \text{subject to}\quad & \mu_1(x) - \mu_0(x) + \nabla\cdot m(x) = \frac{1}{|\Omega|}\int_\Omega \mu_0(x) - \mu_1(x)dx. \end{aligned} \tag{4.5}$$

Of note because the problem is symmetric in time, solving for a solution to Eq. 4.5 is considerably easier numerically. The optimal m is a function of space, and constant in time. Further, we note that

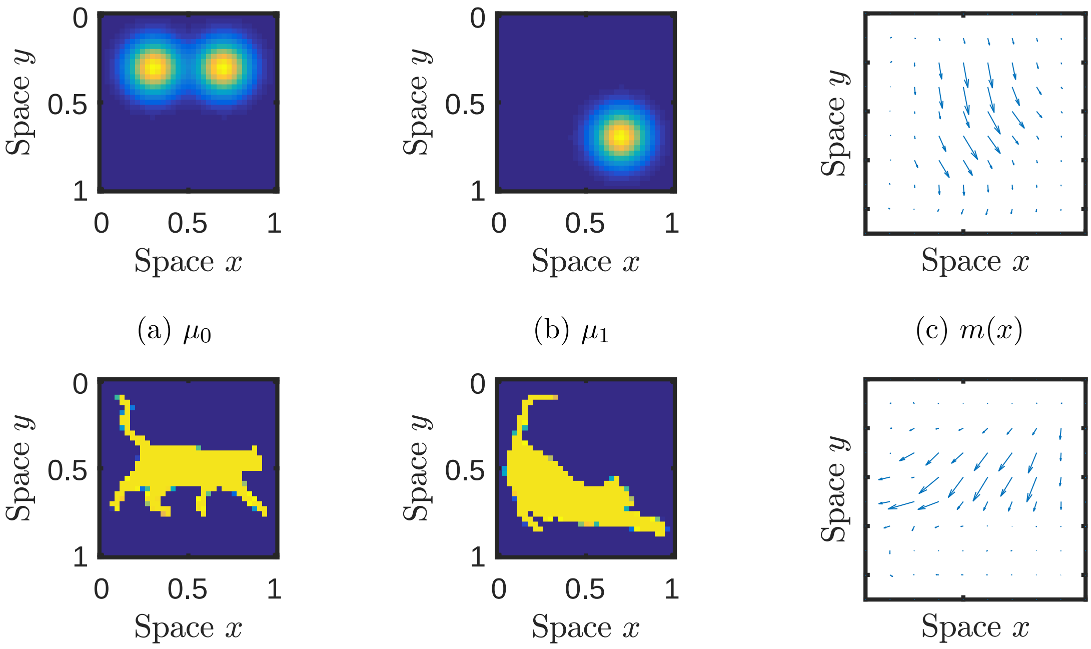
$$\mathrm{UW}_1\left(\mu_0,\mu_1\right) = \mathrm{struc}\left[\mu_0 - \mu_1\right] + \frac{1}{\alpha}\left|\int_\Omega \mu_0(x) - \mu_1(x) dx\right| \tag{4.6}$$

therefore given code that computes the struc $[\cdot]$ or EMD one can easily compute $\mathrm{UW}_1\left(\cdot,\cdot\right)$. Given the wealth of numerics for computing the EMD quickly (e.g. [38, 26]) we will not the specific algorithms for computing $\mathrm{UW}_1\left(\cdot,\cdot\right)$. We will however, present the results from some numerics in Fig. 4.1.

In Fig. 4.1 we plot two problems where we compute $\mathrm{UW}_1\left(\mu_0,\mu_1\right)$. Figs. 4.1a and 4.1a are μ_0 and μ_1, where both μ_0 and μ_1 are smooth functions. Fig. 4.1c shows the value of $m(x)$ from 4.5. Figs. 4.1d - 4.1f are the same, however they show that the numerics work just as well for non-smooth inputs as smooth ones.

4.1.2 Unnormalized Wasserstein-2 Distance

Now we discuss the $p = 2$ case. In this case the solution is not constant in time, however it is still useful to make the substitution $m = \mu v$. With this substitution Eq. 4.2 becomes



(a) μ_0 (b) μ_1 (c) $m(x)$

(d) μ_0 (e) μ_1 (f) $m(x)$

Figure 4.1: Plots of the μ_0, μ_1 and $m(x)$ for $UW_1(\mu_0,\mu_1)$ for the two Gaussian movement (A) μ_0, (B) μ_1, (C) $m(x)$ and two (D) μ_0, (E) μ_1, (F) $m(x)$.

$$\begin{aligned}\mathrm{UW}_2\left(\mu_0,\mu_1\right)^2 &= \inf_{m,\mu,f}\int_0^1\int_\Omega \frac{\|m\|^2}{\mu}dxdt + \frac{1}{\alpha}\int_0^1 |f(t)|^2 dt\\ \text{subject to} \quad & \partial_t\mu(t,x) + \nabla\cdot(m(x,t)) = 0,\\ & \mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x).\end{aligned} \tag{4.7}$$

Further we consider the zero-flux case when $m(x,t)\cdot\eta = 0$ on $\partial\Omega$ where η is the boundary normal vector. Of interest is that just as in the normalized case, there is both a Monge and Kantorovich formulation of Eqn. 4.7. We will present these alternate formulations without proof. The interested reader can see [18] for the full details. The Monge formulation of 4.7 is

$$\begin{aligned}\mathrm{UW}_2\left(\mu_0,\mu_1\right)^2 = \inf_{M,f(t)} &\int_\Omega \|M(x)-x\|^2\mu_0(x)dx + \alpha\int_0^1 f(t)^2dt\\ &+ \int_0^1\int_0^t f(s)\int_\Omega \|M(x)-x\|^2\mathrm{Det}\left(s\nabla M(x) + (1-s)\mathbb{I}\right)dsdtdx,\end{aligned} \tag{4.8}$$

where the infimum is among all one to one, invertible mapping functions $M\colon \Omega \to \Omega$ and a source function $f\colon \Omega \to \mathbb{R}$, such that the unnormalized push forward relation holds

$$\mu(1, M(x))\mathrm{Det}(\nabla M(x)) = \mu(0, x) + \int_0^1 f(t)\mathrm{Det}\Big(t\nabla M(x) + (1-t)\mathbb{I}\Big)dt. \tag{4.9}$$

The Kantorovich formulation is given implicitly in terms of $M(x) = \nabla\Psi(x)$. In this case the optimal map satisfies

$$\frac{1}{2}\mathrm{UW}_2(\mu_0, \mu_1)^2 = \sup_{\Phi} \left\{ \int_\Omega \Phi(1, x)\mu(1, x)dx - \int_\Omega \Phi(0, x)\mu(0, x)dx - \frac{\alpha}{2}\int_0^1 \Big(\int_\Omega \Phi(t, x)dx\Big)^2 dt \right\}$$

where the supremum is taken among all $\Phi\colon [0, 1] \to \Omega$ satisfying

$$\partial_t\Phi(t, x) + \frac{1}{2}\|\nabla\Phi(t, x)\|^2 \leq 0.$$

4.2 Numerics

In this section, we propose to apply a primal-dual algorithm to solve unnormalized OT numerically. We then provide several numerical examples to demonstrate the effectiveness of this procedure.

4.2.1 Algorithm

We present a primal-dual algorithm for problem Eq. 4.2. In particular, our method is based on its reformulation Eq. 4.7, named the minimal flux problem. Define the Lagrangian of Eq. 4.7:

$$\begin{aligned}\mathcal{L}(m, \mu, f, \Phi) = &\int_0^1\int_\Omega \frac{\|m(t, x)\|^2}{2\mu(t, x)}dtdx + \frac{1}{2\alpha}\int_0^1 f(t)^2 dt \\ &+ \int_0^1\int_\Omega \Phi(t, x)\Big(\partial_t\mu(t, x) + \nabla\cdot m(t, x) - f(t)\Big)dxdt,\end{aligned}$$

where $\Phi(t, x)$ is the Lagrange multiplier of the unnormalized continuity constraint in Eq. 4.7.

Convex analysis shows that $(m^*(t, x), \mu^*(t, x), f^*(t))$ is a solution to Eq. 4.7 if and only if there is a Φ^* such that (m^*, Φ^*) is a saddle point of $\mathcal{L}(m, \mu, f, \Phi)$. In other words, we can compute minimization Eq. 4.7 by solving the following minimax problem

$$\inf_{m,\mu,f} \sup_{\Phi} \ \mathcal{L}(m,\mu,f,\Phi),$$

It is clear that $\mathcal{L}$ is convex in m, μ, f and concave in Φ, and the interaction term is a linear operator. This property allows us to apply the Chambolle-Pock first order primal-dual algorithm [8], which gives the update as follows.

$$\begin{cases} m^{k+1}(t,x) = & \arg\inf_m \ \mathcal{L}(m,\mu^k,f^k,\Phi^k) + \frac{1}{2\tau_1}\int_0^1\int_\Omega \|m(t,x)-m^k(t,x)\|^2 dxdt \\ \mu^{k+1}(t,x) = & \arg\inf_\mu \ \mathcal{L}(m^k,\mu,f^k,\Phi^k) + \frac{1}{2\tau_1}\int_0^1\int_\Omega \|\mu(t,x)-\mu^k(t,x)\|^2 dxdt \\ f^{k+1}(t) = & \arg\inf_f \ \mathcal{L}(m^k,\mu^k,f,\Phi^k) + \frac{1}{2\tau_1}\int_0^1 \|f(t)-f^k(t)\|^2 dt \\ \tilde{\Phi}^{k+1}(t,x) = & \arg\sup_\Phi \ \mathcal{L}(\tilde{m},\tilde{\mu},\tilde{f},\Phi) - \frac{1}{2\tau_2}\int_0^1\int_\Omega \|\Phi(t,x)-\Phi^k(t,x)\|^2 dxdt \\ (\tilde{m},\tilde{\mu},\tilde{f}) = & 2(m^{k+1},\mu^{k+1},f^{k+1}) - (m^k,\mu^k,f^k) \end{cases} \tag{4.10}$$

where τ_1, τ_2 are given step sizes for primal, dual variables. These steps can be interpreted as a gradient descent in the primal variable (m,μ,f) and a gradient ascent in the dual variable Φ.

It turns out that the optimizations in above update Eq. 4.10 have explicit formulas. The first line becomes

$$\begin{aligned} m^{k+1}(t,x) =& \arg\inf_m \ \left\{ \frac{\|m(t,x)\|^2}{2\mu^k(t,x)} - m(t,x)\cdot\nabla\Phi(t,x) + \frac{1}{2\tau_1}\|m(t,x)-m^k(t,x)\|^2 \right\} \\ =& \frac{\mu^k(t,x)}{\mu^k(t,x)+\tau_1}\Big(\tau_1\nabla\Phi(t,x) + m^k(t,x)\Big). \end{aligned}$$

The second line of Eq. 4.10 simplifies to

$$\mu^{k+1}(t,x) = \arg\inf_\mu \ \frac{\|m^k(t,x)\|^2}{2\mu(t,x)} - \partial_t\Phi(t,x)\cdot\mu(t,x) + \frac{1}{2\tau_1}|\mu(t,x)-\mu^k(t,x)|^2.$$

The above problem has an analytical solution by solving a cubic equation. The third line of Eq. 4.10 gives

$$\begin{aligned} f^{k+1}(t) =& \arg\inf_f \ \left\{ \frac{1}{2\alpha}f(t)^2 - f(t)\int_\Omega \Phi(t,x)dx + \frac{1}{2\tau_1}\|f(t)-f^k(t)\|^2 \right\} \\ =& \frac{\alpha}{\alpha+\tau_1}\Big(\tau_1\int_\Omega \Phi(t,x)dx + f^k(t)\Big). \end{aligned}$$

The fourth line of Eq. 4.10 gives

$$\begin{aligned}\Phi^{k+1}(t,x) =& \arg\sup_{\Phi}\Big\{\Phi(t,x)\cdot(\partial_t\tilde{\mu}(t,x)+\nabla\cdot\tilde{m}(t,x)-\tilde{f}(t)) - \frac{1}{2\tau_2}\|\Phi(t,x)-\Phi^k(t,x)\|^2\Big\}\\ =&\Phi^k(t,x)+\tau_2\Big(\partial_t\tilde{\mu}^{k+1}(t,x)+\nabla\cdot\tilde{m}(t,x)-\tilde{f}(t)\Big).\end{aligned}$$

Combining all above formulas, we are now ready to state the algorithm.

Algorithm: Primal-Dual method for Unnormalized OT

Input: Unnormalized densities μ_0, μ_1;

Initial guess of m^0, μ^0, Φ^0, f^0, step size τ_1, τ_2.

Output: Minimizer $\mu(t,x)$; Dual variable $\Phi(t,x)$; Value $\mathrm{UW}_2(\mu_0,\mu_1)$.

```
For $k = 1, 2, \cdots$ Iterate until convergence
    $m^{k+1}(t,x) = \frac{\mu^k(t,x)}{\mu^k(t,x)+\tau_1}\Big(\tau_1\nabla\Phi(t,x)+m^k(t,x)\Big)$;
    Solve $\mu^{k+1}(t,x) = \arg\inf_\mu \frac{\|m^k(t,x)\|^2}{2\mu(t,x)} - \partial_t\Phi(t,x)\cdot\mu(t,x) + \frac{1}{2\tau_1}|\mu(t,x)-\mu^k(t,x)|^2$;
    $f^{k+1}(t) = \frac{\alpha}{\alpha+\tau_1}\Big(\tau_1\int_\Omega\Phi(t,x)dx + f^k(t)\Big)$;
    $\Phi^{k+1}(t,x) = \Phi^k(t,x)+\tau_2\Big(\partial_t\tilde{\mu}^{k+1}(t,x)+\nabla\cdot\tilde{m}(t,x)-\tilde{f}(t)\Big)$;
    $(\tilde{m},\tilde{\mu},\tilde{f}) = 2(m^{k+1},\mu^{k+1},f^{k+1}) - (m^k,\mu^k,f^k)$;
End
```

4.2.2 Numerical Grid

To apply the algorithm, we first define our numerical grid. For simplicity we consider the case where the space of interest is $\Omega = [0,1]^d$ and time $\mathcal{T} = [0,1]$. Further, for the following explanations we consider the problem when $d = 2$, however, our grid construction can be constructed on any dimension by extending it in the obvious way. We will use the same symbol to represent both the continuous μ, m, Φ, f and their respective discretized counterparts, as the difference between the two should be clear from context alone.

Let n_t, n_x, and n_y be given then notate $\Delta t = \frac{1}{n_t-1}$, $\Delta x = \frac{1}{n_x-1}$, and $\Delta y = \frac{1}{n_y-1}$. Using this notation we define the following sets:

$$
\begin{aligned}
\Omega_{(i,j)} &= [i\Delta x, (i+1)\Delta x] \times [j\Delta y, (j+1)\Delta y] \\
\mathcal{T}_{(k)} &= [k\Delta t, (k+1)\Delta t] \\
\Omega_{(i-1/2,j)} &= [(i-1/2)\Delta x, (i+1/2)\Delta x] \times [j\Delta y, (j+1)\Delta y] \text{ for } i = 0, \dots, n_x \\
\Omega_{(i,j-1/2)} &= [i\Delta x, (i+1)\Delta x] \times [(j-1/2)\Delta y, (j+1/2)\Delta y] \text{ for } j = 0, \dots, n_y
\end{aligned}
$$

where $i = 0, \dots, n_x - 1$, $j = 0, \dots, n_y - 1$, and $k = 0 \dots, n_t - 1$ unless otherwise specified.

For the discretized problem we consider a $f_{(k)}$ that is constant along each $\mathcal{T}_{(k)}$, and consider $\mu_{(k,i,j)}$ and $\Phi_{(k,i,j)}$ that are constant along each $\mathcal{T}_{(k)} \times \Omega_{(i,j)}$. The vector $m_{(k,i,j)}$ has two components $m_{x,(k,i-1/2,j)}$ and $m_{y,(k,i,j-1/2)}$, that are constant along $\mathcal{T}_{(k)} \times \Omega_{(i-1/2,j)}$ and $\mathcal{T}_{(k)} \times \Omega_{(i,j-1/2)}$ respectively. Numerically m quantifies the movement of density between each of the $\Omega_{(i,j)}$ and its spacial neighbors (i.e. $\Omega_{(i-1,j)}, \Omega_{(i,j-1)}, \Omega_{(i+1,j)}$, and $\Omega_{(i,j+1)}$) and so it is natural to define the components of m not on $\Omega_{(i,j)}$ but rather on $\Omega_{(i-1/2,j)}$, $\Omega_{(i+1/2,j)}$, $\Omega_{(i,j-1/2)}$ and $\Omega_{(i,j+1/2)}$.

Using the above notation, we write the steps of the algorithm as:

$$
m_{x,(k,i-1/2,j)} = \begin{cases} \frac{\mu_{(k,i-1,j)}+\mu_{(k,i-1,j)}}{\mu_{(k,i,j)}+\mu_{(k,i-1,j)}+2\tau_1} \left(\tau_1 + \nabla_x \Phi_{(k,i-1/2,j)} + m_{x,(k,i-1/2,j)}\right) & \text{if } i = 1, \dots, n_x - 1 \\ 0 & \text{if } i = 0, n_x \end{cases}
$$

$$
m_{y,(k,i,j-1/2)} = \begin{cases} \frac{\mu_{(k,i,j)}+\mu_{(k,i,j-1)}}{\mu_{(k,i,j)}+\mu_{(k,i,j-1)}+2\tau_1} \left(\tau_1 + \nabla_y \Phi_{(k,i,j-1/2)} + m_{y,(k,i,j-1/2)}\right) & \text{if } j = 1, \dots, n_y - 1 \\ 0 & \text{if } j = 0, n_y \end{cases}
$$

$$
\begin{aligned}
\mu_{(k,i,j)} &= \mathrm{root}^+(1, -(\tau_1 * \partial_t \Phi_{(k,i,j)} + \mu_{(k,i,j)}), 0, \\
&\quad \frac{-\tau_1}{8} \left((m_{(k,i+1/2,j)} + m_{(k,i-1/2,j)})^2 + (m_{(k,i,j+1/2)} + m_{(k,i,j-1/2)})^2\right)) \\
f_{(k)} &= \frac{\alpha}{\alpha + \tau_1} \left(\tau_1 + \sum_i \sum_j \Phi_{(k,i,j)} \Delta x \Delta y + f_{(k)}\right) \\
\Phi_{(k,i,j)} &= \tau_2 * \left(\partial_t \tilde{\mu}_{(k,i,j)} + \nabla \cdot \tilde{m}_{(k,i,j)} - \tilde{f}_{(k)}\right) + \Phi_{(k,i,j)}
\end{aligned}
$$

where

$$\begin{aligned}
\nabla_x \Phi_{(k,i-1/2,j)} &= \frac{\Phi_{(k,i,j)} - \Phi_{(k,i-1,j)}}{\Delta x} \\
\nabla_y \Phi_{(k,i,j-1/2)} &= \frac{\Phi_{(k,i,j)} - \Phi_{(k,i,j-1)}}{\Delta y}; \\
\partial_t \Phi_{(k,i,j)} &= \begin{cases} \frac{1}{\Delta t}\left(\frac{\Phi_{(1,i,j)}}{2} + \Phi_{(0,i,j)}\right) & \text{if } k = 0 \\ \frac{1}{\Delta t}\left(\frac{\Phi_{(2,i,j)}}{2} - \Phi_{0,i,j}\right) & \text{if } k = 1 \\ \frac{1}{2\Delta t}\left(\Phi_{(k+1,i,j)} - \Phi_{(k-1,i,j)}\right) & \text{if } 1 < k < n_t - 2 \\ \frac{1}{\Delta t}\left(\Phi_{(n_t-1,i,j)} - \frac{\Phi_{(n_t-3,i,j)}}{2}\right) & \text{if } k = n_t - 2 \\ \frac{1}{\Delta t}\left(-\Phi_{(n_t-1,i,j)} - \frac{\Phi_{(n_t-2,i,j)}}{2}\right) & \text{if } k = n_t - 1 \end{cases} \\
\text{root}^+(a,b,c,d) &= \text{ the largest real solution to } ax^3 + bx^2 + cx + d = 0 \\
\partial_t \mu_{(k,i,j)} &= \begin{cases} \frac{1}{\Delta t}\left(\mu_{(1,i,j)} - \mu_{(0,i,j)}\right) & \text{if } k = 0 \\ \frac{1}{2\Delta t}\left(\mu_{(k+1,i,j)} - \mu_{(k-1,i,j)}\right) & \text{if } 0 < k < n_t - 1 \\ \frac{1}{\Delta t}\left(\mu_{(n_t-1,i,j)} - \mu_{(n_t-2,i,j)}\right) & \text{if } k = n_t - 1 \end{cases} \\
\nabla \cdot m_{(k,i,j)} &= \frac{m_{x,(k,i+1/2,j)} - m_{x,(k,i-1/2,j)}}{\Delta x} + \frac{m_{y,(k,i,j+1/2)} - m_{y,(k,i,j-1/2)}}{\Delta y}.
\end{aligned}$$

Note that the unusual boundary conditions of $\partial_t \Phi$ arise from the need to satisfy

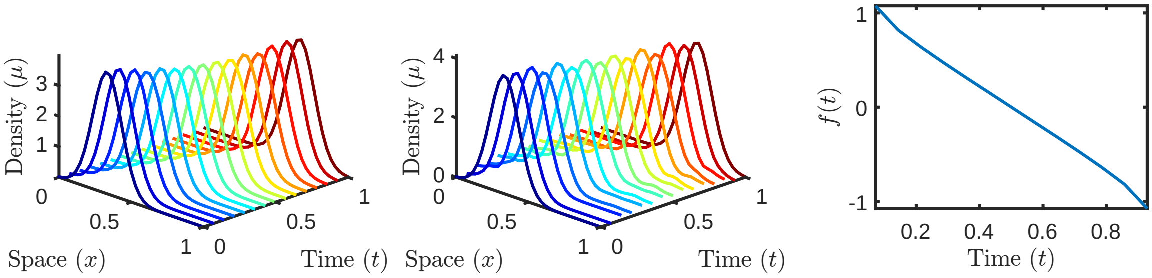
$$\sum_k \Phi_{(k,i,j)} \partial_t \mu_{(k,i,j)} \Delta t = -\sum_k \partial_t \Phi_{(k,i,j)} \mu_{(k,i,j)} \Delta t \quad \forall i, j.$$

4.2.3 Numerical Experiments

Now we present our numerical results. The first two experiments are in one dimension, and the rest are in two. The numerical parameters for our experiments are given in Table 4.1.

Parameter	Value	Parameter	Value
Discretization		**Optimization**	
n_t	15	Iterations	200,000
n_x	35	τ_1	10^{-3}
n_y	35	τ_2	10^{-1}
		α	100

Table 4.1: Numerical parameters for our experiments. Note that for our one dimensional experiments, n_y has no value.



(a) $\mu(t, x, y)$

(b) $\mu(t, x, y)$

(c) $f(t)$

Figure 4.2: A plot of (A) $W_2(\mu_0, \mu_1)$, (B) $UW_2(\mu_0, \mu_1)$ and (C) $f(t)$ in the unbalanced case.

4.3 Experimental Results

4.3.1 Experiment 1

Here we consider the problem where μ_0 and μ_1 are both one dimensional Gaussians of equal integral, $\Omega = [0, 1]$ and

$$\mu_0 = N\left(x; \frac{1}{3}, 0.1\right)$$

$$\mu_1 = N\left(x; \frac{2}{3}, 0.1\right)$$

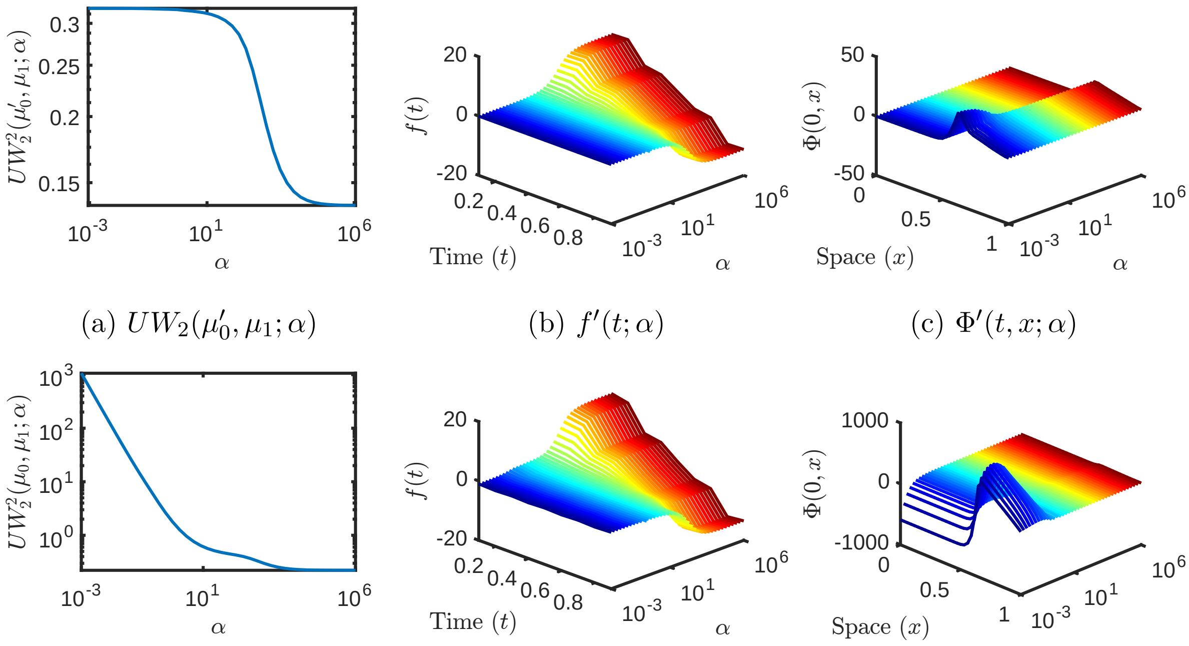
$$N(x; \mu_x, \sigma^2) = Ce^{\frac{(x-\mu_x)^2}{2\sigma^2}} \text{ where } C \text{ is such that } \int_\Omega N(x; \mu_x, \sigma^2)dx = 1$$

We plot the results in Figure 4.2. In this case the input densities are balanced and so $W_2(\mu_0, \mu_1)$ and $UW_2(\mu_0, \mu_1)$ appear similar. Indeed $UW_2(\mu_0, \mu_1) = 0.055$ and $W_2(\mu_0, \mu_1) =$

0.056.

Note that even in this simple case the behavior of $f(t)$ is nuanced. In this case, μ_0 and μ_1 are smooth, of equal integral and $W_2(\mu_0, \mu_1)$ is given by a simple analytical formula, and $f(t)$ is not identically zero. Integrating the constraint in Eq. 4.7 in space and time yields $|\Omega| \int_{[0,1]} f(t)dt = \int_\Omega \mu_1 dx - \int_\Omega \mu_0 dx$, and so for balanced inputs $\int_{[0,1]} f(t)dt = 0$, but experiment 1 shows that $f \not\equiv 0$.

4.3.2 Experiment 2



(a) $UW_2(\mu_0', \mu_1; \alpha)$ (b) $f'(t; \alpha)$ (c) $\Phi'(t, x; \alpha)$

(d) $UW_2(\mu_0, \mu_1; \alpha)$ (e) $f(t; \alpha)$ (f) $\Phi(t, x; \alpha)$

Figure 4.3: A plot of the asymptotic behavior of UW_2 in α with balanced and unbalanced inputs. Balanced: (A) $UW_2(\mu_0', \mu_1; \alpha)$, (B) $f'(t; \alpha)$, (C) $\Phi'(t, x; \alpha)$, and unbalanced: (D) $UW_2(\mu_0, \mu_1; \alpha)$, (E) $f(t; \alpha)$, (F) $\Phi(t, x; \alpha)$.

Again consider $\Omega = [0, 1]$, however in this experiment we analyse the asymptotic behavior of $UW_2(\mu_0, \mu_1)$ as a function of α and $\alpha \to 0$ and $\alpha \to \infty$. Here

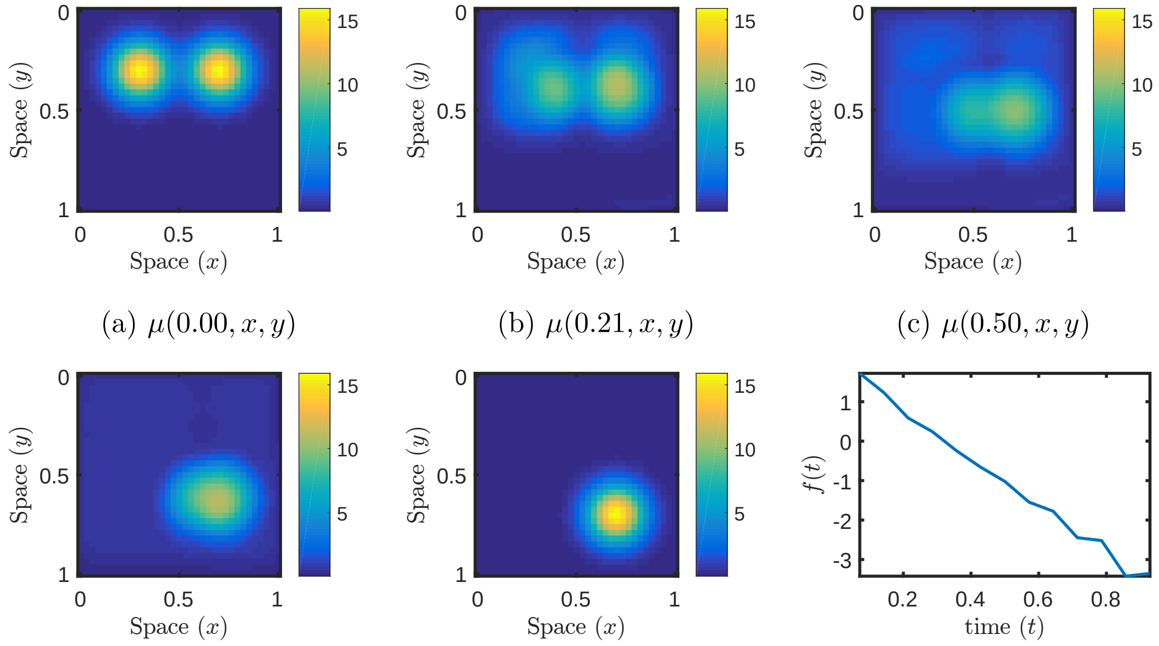
$$\begin{aligned}
\mu_0 &= N\left(x; 0, 0.1\right) + N\left(x; \frac{1}{3}, 0.1\right) \\
\mu_0' &= \frac{1}{2}\left(N\left(x; 0, 0.1\right) + N\left(x; \frac{1}{3}, 0.1\right)\right) \\
\mu_1 &= N\left(x; \frac{2}{3}, 0.1\right).
\end{aligned}$$

The balanced case refers to $UW_2(\mu_0', \mu_1)$, and the unbalanced refers to $UW_2(\mu_0, \mu_1)$. In both cases we compute the unnormalized Wasserstein distance. The results are given in Figure 4.3.

Figures 4.3a - 4.3c show that (at least numerically) $UW_2(\mu_0, \mu_1; \alpha), f(t, \alpha)$ and $\Phi(t, x; \alpha)$ converge as $\alpha \to 0^+$, $\alpha \to \infty$ when $\int_\Omega \mu_0 dx = \int_\Omega \mu_1 dx$. Further is seems plausible that for balanced inputs $UW_2(\mu_0, \mu_1; \alpha) \to W_2(\mu_0, \mu_1)$ as $\alpha \to 0^+$. For any α the μ, m and Φ from $W_2(\mu_0, \mu_1)$ along with $f(t) \equiv 0$ satisfy the constraint of Eq. 4.7. Formally sending $\alpha \to \infty$ causes $f(t)$ to 0.

Figures 4.3d - 4.3f illustrate the asymptotic behavior of $UW_2(\mu_0, \mu_1; \alpha)$ w.r.t. α when the inputs are unbalanced. In that case we (numerically) see that as $\alpha \to 0$, $f(t; \alpha)$ converges to a non-zero value, and both $UW_2(\mu_0, \mu_1; \alpha)$ and $\Phi(t, x; \alpha)$ diverge. This too is consistent with the formal argument that $UW_2(\mu_0, \mu_1; \alpha) \to W_2(\mu_0, \mu_1)$ as $\alpha \to 0^+$.

In a predecessor of this work [5] the authors solve for $W_2(\mu_1, \rho_2)$ using Lagrange multipliers in a similar formulation to Eq. 4.7. In their work the Lagrange multiplier $\Phi(t, x)$ is given up to an additive constant. If indeed $UW_2(\mu_0, \mu_1; \alpha) \to W_2(\mu_0, \mu_1)$ as $\alpha \to 0^+$ and $\Phi(t, x; \alpha)$ does converge then $\Phi(t, x; 0^+)$ is given uniquely (as a limit) and there is no issue of undetermined constants.



(a) $\mu(0.00, x, y)$

(b) $\mu(0.21, x, y)$

(c) $\mu(0.50, x, y)$

(d) $\mu(0.79, x, y)$

(e) $\mu(1.00, x, y)$

(f) $f(t)$

Figure 4.4: Plots of the $\mu(t, x, y)$ and $f(t)$ for $UW_2(\mu_0, \mu_1)$. (A) $\mu(0.00, x, y)$, (B) $\mu(0.21, x, y)$, (C) $\mu(0.50, x, y)$, (D) $\mu(0.79, x, y)$, (E) $\mu(1.00, x, y)$, (F) $f(t)$.

4.3.3 Experiment 3

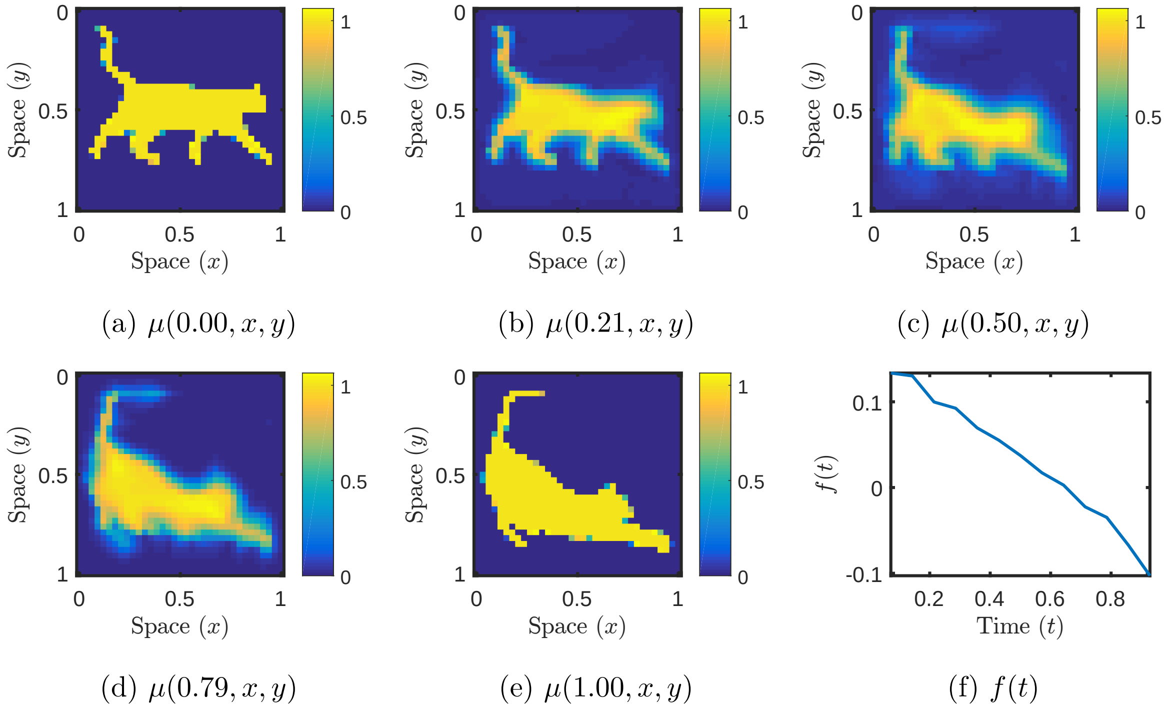
Now consider the two dimensional problem where $\Omega = [0, 1]^2$. In this case

$$\mu_0(x, y) = N(x, y; 0.3, 0.3, 0.1, 0.1) + N(x, y; 0.7, 0.3, 0.1, 0.1)$$

$$\mu_1(x, y) = N(x, y; 0.7, 0.7, 0.1, 0.1)$$

$$N(x, y; \mu_x, \mu_y, \sigma_1^2, \sigma_2^2) = Ce^{\frac{(x-\mu_x)^2}{2\sigma_1^2} + \frac{(y-\mu_y)^2}{2\sigma_2^2}},$$

where C is a normalization constant such that $\int_\Omega N(x, y; \mu_x, \mu_y, \sigma_1^2, \sigma_2^2) dxdy = 1$. The results from our experiments are shown in Figure 4.4. Note that although the mass of μ_0 is twice that of μ_1, the optimal $f(t)$ is not non-positive. Indeed from $t = 0$ to $t \approx \frac{1}{4}$, $f(t)$ is positive, before staying non-positive for the rest of the interval. This again illustrates that even in the case of gaussian movement the behavior of $f(t)$ is nuanced, and violates naive basic intuition.



(a) $\mu(0.00, x, y)$

(b) $\mu(0.21, x, y)$

(c) $\mu(0.50, x, y)$

(d) $\mu(0.79, x, y)$

(e) $\mu(1.00, x, y)$

(f) $f(t)$

Figure 4.5: Plots of the $\mu(t, x, y)$ and $f(t)$ for $UW_2(\mu_0, \mu_1)$. (A) $\mu(0.00, x, y)$, (B) $\mu(0.21, x, y)$, (C) $\mu(0.50, x, y)$, (D) $\mu(0.79, x, y)$, (E) $\mu(1.00, x, y)$, (F) $f(t)$.

4.3.4 Experiment 4

Consider again the two dimensional problem, however this time we choose μ_0 and μ_1 to be the cats in [38]. Our results are summarized in Figure 4.5. This illustrates that our new method can be used as a general purpose OT solver for unbalanced inputs, and so can be used to interpolate between two functions.

CHAPTER 5

Conclusion

In this thesis we have studied inverse problem from a theoretical and practical perspective. We have shown that the work done in [42] can be improved both in quality of reconstruction and speed of reconstruction by careful choice of regularizer as well as algorithm for solving the resulting inverse problem.

We have also demonstrated that in inverse problems forward operator correctness is paramount, and even a modest error in modeling said operator can lead to catastrophic error in reconstruction. To address this problem we have developed a new tool called the *s*tructure. We prove some new results concerning the treatment of noise by the Earth Mover's Distance (EMD). Further, consistent with these theoretical results, we perform numerical experiments and show that the structure is able to distinguish between error in the modeling of a forward operator, and noise in the signal of an inverse problem. Therefore the structure of the residual of an inverse procedure can be used as a proxy for the correctness of the forward operator used.

We also do numerical experiments that concern model linear forward operators. On these problems the structure of the residual is indeed minimized when the correct forward operator is used. The L^1 or L^2 norms of the residual are also minimized around the correct forward operator, the structure, however, is more localized and has better contrast around the minimum. Further, we observe that the degree to which the inverse problem is overdetermined and degree of regularization is critical to the success of the procedure. The more over determined the problem, the more useful the structure. This is borne out by the analysis in the case of linear regularization, as well as the numerical results on more sophisticated problems.

Finally we also propose and solve an unnormalized optimal transport problem, specifically we develop the Unnormalized Wasserstein-1 and 2 distances. We show that the proposed distances are well defined and easily numerically computed. Our generalization is parameterized by a positive scalar α and reduces to the original Wasserstein-p distance when $\alpha \to 0^+$. For $\alpha > 0$ our generalization is a smooth extension of the Wasserstein-p distance but crutially does not require that the input arguments have equal integral.

The Wasserstein-1 and 2 distances are both common in both pure and applied math ([41, 36, 12, 1]) and to either one must do some bespoke, often heuristic, preprocessing step to normalize the input data. We believe that our Unnormalized Wasserstein-1 and 2 distances can be used as a drop-in replacement for both this heuristic step and the subsequent calculation and so can unify many of these desperate approaches.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

[2] Simon R Arridge. Optical tomography in medical imaging. Inverse problems, 15(2):R41, 1999.

[3] Zhaojun Bai and James W Demmel. Computing the generalized singular value decomposition. SIAM Journal on Scientific Computing, 14(6):1464–1486, 1993.

[4] Stephen Becker. Lbfgsb (l-bfgs-b) mex wrapper, 2012–2015.

[5] 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.

[6] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. IMA Journal of Applied Mathematics, 6(1):76–90, 1970.

[7] Moustafa T Chahine. Inverse problems in radiative transfer: Determination of atmospheric parameters. Journal of the Atmospheric Sciences, 27(6):960–967, 1970.

[8] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[9] Tony F Chan and Jianhong Jackie Shen. Image processing and analysis: variational, PDE, wavelet, and stochastic methods, volume 94. Siam, 2005.

[10] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. Communications on pure and applied mathematics, 41(7):909–996, 1988.

[11] Bjorn Engquist and Brittany D Froese. Application of the wasserstein metric to seismic signals. arXiv preprint arXiv:1311.4581, 2013.

[12] Bjorn Engquist, Brittany D Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. arXiv preprint arXiv:1602.01540, 2016.

[13] Lawrence C Evans. Partial differential equations and monge-kantorovich mass transfer. Current developments in mathematics, 1997(1):65–126, 1997.

[14] Lawrence C Evans and Wilfrid Gangbo. Differential equations methods for the Monge-Kantorovich mass transfer problem, volume 653. American Mathematical Soc., 1999.

[15] N Fedorczak, F Brochard, G Bonhomme, K Schneider, M Farge, P Monier-Garbet, et al. Tomographic reconstruction of tokamak plasma light emission from single image using wavelet-vaguelette decomposition. Nuclear Fusion, 52(1):013005, 2011.

[16] Roger Fletcher. A new approach to variable metric algorithms. The computer journal, 13(3):317–322, 1970.

[17] Anthony Freeman. Sar calibration: An overview. IEEE Transactions on Geoscience and Remote Sensing, 30(6):1107–1121, 1992.

[18] Wilfrid Gangbo, Wuchen Li, Stanley J. Osher, and Michael A. Puthawala. Unnormalized optimal transport. 2019.

[19] Donald Goldfarb. A family of variable-metric methods derived by variational means. Mathematics of computation, 24(109):23–26, 1970.

[20] Tom Goldstein and Stanley Osher. The split bregman method for l1-regularized problems. SIAM journal on imaging sciences, 2(2):323–343, 2009.

[21] Gene H Golub. Matrix computations. Johns Hopkins University Press, 1996.

[22] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. SIAM Journal on Matrix Analysis and Applications, 21(1):185–194, 1999.

[23] Per Christian Hansen. Regularization, gsvd and truncatedgsvd. BIT numerical mathematics, 29(3):491–504, 1989.

[24] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. SIAM review, 34(4):561–580, 1992.

[25] Per Christian Hansen and Dianne Prost O'Leary. The use of the l-curve in the regularization of discrete ill-posed problems. SIAM Journal on Scientific Computing, 14(6):1487–1503, 1993.

[26] Matt Jacobs, Flavien Léger, Wuchen Li, and Stanley Osher. Solving large-scale optimization problems with a convergence rate independent of grid size. arXiv preprint arXiv:1805.09453, 2018.

[27] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3):425–464, 2001.

[28] Andreas Kirsch. An introduction to the mathematical theory of inverse problems, volume 120. Springer Science & Business Media, 2011.

[29] Wuchen Li, Stanley Osher, and Wilfrid Gangbo. A fast algorithm for earth mover's distance based on optimal transport and l1 type regularization. arXiv preprint arXiv:1609.07092, 2016.

[30] Wuchen Li, Ernest K Ryu, Stanlet Osher, Wotao Yin, and Wolfred Gangbo. A parallel method for earth mover's distance. Journal of Scientific Computing, page 75(1), 2018.

[31] Stephane G Mallat. Multiresolution approximations and wavelet orthonormal bases of l^2(r). Transactions of the American mathematical society, 315(1):69–87, 1989.

[32] Dean S Oliver, Albert C Reynolds, and Ning Liu. Inverse theory for petroleum reservoir characterization and history matching. Cambridge University Press, 2008.

[33] Christopher C Paige and Michael A Saunders. Towards a generalized singular value decomposition. SIAM Journal on Numerical Analysis, 18(3):398–405, 1981.

[34] Ken Perlin. An image synthesizer. ACM Siggraph Computer Graphics, 19(3):287–296, 1985.

[35] Ken Perlin. Improving noise. In ACM Transactions on Graphics (TOG), volume 21, pages 681–682. ACM, 2002.

[36] Michael A. Puthawala, Cory D. Hauck, and Stanley J. Osher. Diagnosing forward operator error using optimal transport. 2019.

[37] 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.

[38] Ernest Ryu, Yongxin Chen, Wuchen Li, and Stanley Osher. Vector and matrix optimal mass transport: Theory, algorithm, and applications. arXiv, 2017.

[39] Uwe Schneider, Eros Pedroni, and Antony Lomax. The calibration of ct hounsfield units for radiotherapy treatment planning. Physics in Medicine & Biology, 41(1):111, 1996.

[40] David F Shanno. Conditioning of quasi-newton methods for function minimization. Mathematics of computation, 24(111):647–656, 1970.

[41] Cédric Villani. Optimal transport: old and new, volume 338. Springer Science & Business Media, 2008.

[42] Andreas Wingen, MW Shafer, Ezekial A Unterberg, Judith C Hill, and Donald L Hillis. Regularization of soft-x-ray imaging in the diii-d tokamak. Journal of Computational Physics, 289:83–95, 2015.

[43] 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.

[44] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Lbfgs-b: Fortran subroutines for large-scale bound constrained optimization. Report NAM-11, EECS Department, Northwestern University, 1994.