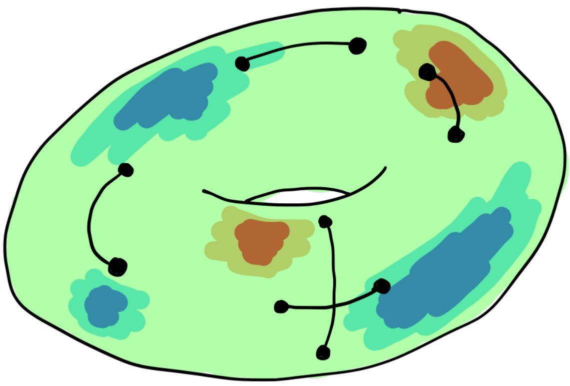
Metric Tomography with Sobolev Gradients

Candidate Number: 4643J
Supervisor: Dr. David Al-Attar



Submitted in partial fulfillment of the requirements for the degree of *Master of Advanced Studies*

Abstract

In underdetermined inverse problems such as ray-tracing tomography, finite-parametrisation of the model and regularisation of the misfit functional are often used to force a unique solution. However, these measures are unsatisfactory because they are arbitrarily chosen. An alternative method is to use gradient-based optimisation to find a local minimum of the misfit, where the first order gradient of the misfit is required for the scheme and the second order can be used to quantify the constraints on the obtained solution or to potentially improve the scheme. The adjoint method is an effective way of calculating these gradients. We show that, in order to obtain a sensible solution through this scheme, we must work in an appropriate Sobolev space, the space of square-integrable functions with additional continuity constraints. We apply these ideas to the metric tomographic problem, the problem of finding an unknown metric on a manifold given a finite number of geodesics, which is an analogous problem to ray-tracing tomography of surface waves. We numerically implement this method for the linearised version of the problem on a 2-dimensional torus. Additionally, we present the first and second adjoints for the general problem, which can be used to calculate the first and second order gradients of the misfit function.

Contents

1 Introduction

Tomography, imaging a medium through penetrating waves or signals, is an important method in many disciplines. Particularly in seismology, it is responsible for disproving the two-layered convection hypothesis by revealing that slabs sink into the lower mantle [1], and for discovering 'weather patterns' due to convection and magnetic activity inside the sun [2]. Tomography depends on the principle that observations derived from a medium should rely on the internal structure and therefore should contain information about it. If we consider the forward problem to be determining what observations would be made given some internal structure, then tomography is the inverse problem of determining the internal structure given some observations. This process is also often called 'inversion'.

In this article, we focus on the *metric tomographic problem*, the problem of determining an unknown metric tensor of a manifold given a finite set of end points and lengths of geodesics. This sketched in Fig. 1. This problem is of interest as it turns out that it is analogous to delay-time anisotropic ray-tomography from surface waves on a sphere [4].

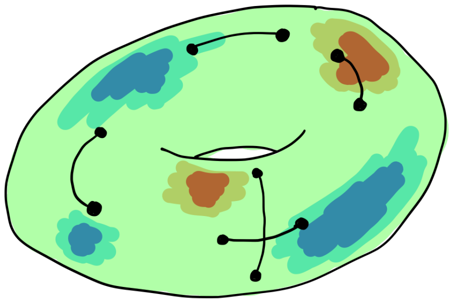


Figure 1: A sketch of the metric tomography problem. Patches of color indicate areas of different velocities, lines indicate geodesics.

Inversions are most often done through minimising the least-squares misfit $\hat{J}$ between observation (the data d) and synthesized observations $a(m)$, given some internal structure (the model m),

$$\hat{J}[m] = \frac{1}{2} \langle a(m) - d, a(m) - d \rangle ,$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product. However, most inverse problems only have a few observations, whereas internal structure, often described as a field valued at every point in space, is an infinite-dimensional parameter. This makes the problem severely underdetermined. The two conventional ways of tackling these problems, especially in ray tomography, are to a) finite-parametrise the model, expanding it into basis functions such that we fit to only a finite number of parameters, and b) regularise the misfit, adding a term $\langle Bm, m \rangle$ to the misfit, where B is a positive-definite matrix. These measures enforce a unique solution to the inverse problem [13]. However, this method has been shown to affect the resulting solution (eg. [3]), and is unsatisfactory because we arbitrarily assume certain features of the solution, without any physical justifications.

An alternative method is to use gradient-based optimisation schemes, where, roughly, we update some initial model guess in the negative direction of the gradient until an optimised model that locally minimises the misfit is found. Only the first order gradient of the misfit is used for most schemes, but second order gradients can potentially improve the scheme or quantify the *constraint* of the optimal model [9]. The *adjoint method* is an effective way of calculating these gradients. To implement these schemes, we must work in a Hilbert space, such that the misfit perturbation due to the model perturbation can be written as the inner product of some *kernel* and the model perturbation. Then, appropriate amounts of the kernel can be added to/subtracted from the model.

In this article, we propose that in order to find a sensible optimised model that is pointwise convergent and is differentiable up to a desired order, we must work in *Sobolev spaces*, rather than the conventional square-integrable function space (L^2). We numerically implement the scheme for a linearised version of the metric tomographic problem on a 2-dimensional torus, and highlight that as resolution increases, the kernels and the optimised models converge to a finite value in the appropriate Sobolev space, but diverges in L^2 space. Knowing that the metric tomographic problem can be posed in a sensible function space and therefore be solved in gradient-based optimisation schemes, we present novel calculations of the first and second adjoint problems for the general, non-linear metric tomographic problem, and discuss how it can be implemented in gradient-based optimisation schemes.

The article is organised as follows. Sec. 2 motivates the metric tomography problem from surface wave equations. Sec. 3 discusses inverse problems in general, the *adjoint method* as a means of calculating gradients effectively, and argue why Sobolev spaces are appropriate spaces to pose certain inverse problems. Then, Sec. 4 will formulate and numerically implement a linearised version of the problem on a 2-dimensional torus. In Sec. 5, we present the calculation results of the first and second order adjoints.

2 The Forward Problem

2.1 Wave Propagation in Manifolds

On a Riemannian manifold $\mathcal{M}$, the wave equation is

$$\frac{\partial^2 u}{\partial \tau^2} + \Delta u = 0,$$

where $u : \mathcal{M} \times \mathbb{R} \to \mathbb{R}$ is the wavefield, a function of points $x \in \mathcal{M}$ and time τ, and the Laplacian operator $\Delta := -\nabla_\mu \nabla^\mu$.

In coordinate notation, this can be written as

$$\frac{\partial^2 u}{\partial \tau^2} - \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x^\mu} \left(\sqrt{|g|} g^{\mu\nu} \frac{\partial u}{\partial x^\nu} \right) = 0,$$

where $g_{\mu\nu}$ is the metric tensor defined on the manifold $\mathcal{M}$, and $|g|$ is its determinant, and Einstein summation convention is used for Greek indices throughout the article.

By analogy to ray-tracing tomography on an elastic body (see appendix A.1), for non-homogeneous metrics, we approximate the solution to first order for each angular frequency ω as

$$u(x, \tau) \approx a(x) e^{i\omega(\tau - T(x))}.$$

Substituting it into the wave equation, to leading order,

$$g^{\mu\nu} p_\mu p_\nu = 1,$$

where we defined $p_\mu := \frac{\partial T(x)}{\partial x^\mu}$. Since this quantity is conserved, we employ the method of characteristics, defining the Hamiltonian as

$$H(x, p) := \frac{1}{2} g^{\mu\nu}(x) p_\mu p_\nu$$

with the Equation of Motion (EoM)

$$\frac{dx^\mu}{dt} = \frac{\partial H}{\partial p_\mu}; \quad \frac{dp_\mu}{dt} = -\frac{\partial H}{\partial x^\mu},$$

where t is some generating parameter.

In the next subsection, we will see that these expressions are the same as the EoM for a geodesic on a manifold. Therefore in the remainder of the article, we study the analogous geodesic problem with the identification of geodesic lengths with travel time data, and the sets of point on the manifold as locations of seismometers and sources.

2.2 The Geodesic Forward Problem

On a Riemannian manifold $\mathcal{M}$, the length l of a curve $x(t)$ with parameter $t \in [0,1]$ is given by

$$l = \int_0^1 \sqrt{g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu}\, dt,$$

where a dot represents differentiation under the parameter t, $\dot{} := \frac{d}{dt}$. A geodesic is defined as a curve such that its length is extremised. Since the square root is a monotonic function, $x(t)$ that extremises l also extremises the action S, defined as

$$S = \int_0^1 g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu dt = \int_0^1 L[x, \dot{x}]\, dt,$$

where $L[x, \dot{x}]$ is the Lagrangian. We use the variational principle to minimise the action, yielding the Euler-Lagrangian equation,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^\mu}\right) - \frac{\partial L}{\partial x^\mu} = g_{\mu\rho}\ddot{x}^\rho - \frac{1}{2}\left(\partial_\mu g_{\rho\sigma}\right)\dot{x}^\rho \dot{x}^\sigma = 0.$$

This EoM can be phrased in the Hamiltonian formulation. Defining canonical momenta p such that

$$p_\mu := \frac{\partial L}{\partial \dot{x}^\mu} = g_{\mu\nu}(x)\dot{x}^\nu,$$

and the Hamiltonian H such that

$$H[x,p] := p_\mu \dot{x}^\mu - L,$$

we find that

$$H[x,p] = g^{\mu\nu}(x) p_\mu p_\nu.$$

By varying the Lagrangian and Hamiltonian with respect to x^μ, p_μ, the equations of motion can be phrased as [6]

$$\frac{d\dot{x}^\mu}{dt} = \frac{\partial H}{\partial p_\mu} \quad \text{and} \quad \frac{d\dot{p}_\mu}{dt} = -\frac{\partial H}{\partial x^\mu},$$

identical in form with the ray-tracing problem. Furthermore, it can be shown that the Hamiltonian H is conserved along the geodesic. Consider imposing the starting condition of the geodesic, $\boldsymbol{x}_0, \boldsymbol{p}_0$. An interesting consequence of the choice of parametrisation t to be from 0 to 1 is that the length of the geodesic is equal to the norm of the starting momenta, since

$$l = \int_0^1 \sqrt{g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu}\, dt = \int_0^1 \sqrt{g^{\mu\nu}(x) p_\mu p_\nu}\, dt = \int_0^1 \sqrt{2H}\, dt.$$

Since H is conserved along the geodesic path, we set it to its starting value:

$$l = \sqrt{2H} = \sqrt{\langle \boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}(0), \boldsymbol{p}(0)\rangle} = \|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}\,, \tag{2.1}$$

where $[\boldsymbol{g}(\boldsymbol{x})]_{\mu\nu} := g_{\mu\nu}(x)$, $[\boldsymbol{g}^{-1}(\boldsymbol{x})]^{\mu\nu} := g^{\mu\nu}(x)$, $[\boldsymbol{p}]_\mu = p_\mu$.

We cast the EoM in even more concise fashion using a $2n$-dimensional vector $\boldsymbol{z}$, as

$$\frac{d\boldsymbol{z}}{dt} = \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}), \quad H = \langle \boldsymbol{g}^{-1}(\boldsymbol{x})\boldsymbol{p}, \boldsymbol{p}\rangle = \boldsymbol{p}^T \boldsymbol{g}^{-1}(\boldsymbol{x})\boldsymbol{p},$$

where

$$\mathbb{J} := \begin{bmatrix} \mathbf{0}_n & \mathbf{1}_n \\ -\mathbf{1}_n & \mathbf{0}_n \end{bmatrix}, \quad \boldsymbol{z}(t) := \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{p}(t) \end{bmatrix}.$$

Then, provided some initial condition

$$\boldsymbol{z}(0) = \begin{bmatrix} \boldsymbol{x}(0) \\ \boldsymbol{p}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{p}_0 \end{bmatrix},$$

and the metric $\boldsymbol{g}(\boldsymbol{x})$, the forward problem is to find the forward solution $\boldsymbol{x}(t)$. This can be readily implemented using a numerical ODE solver.

This statement of the forward problem is known as the strong form. When using a variational method, it is sometimes useful to pose the problem in the weak form, constructed as follows: we take the Euclidean inner product of the EoM with some test function $\boldsymbol{y}(t)$, then integrate it over the domain of $\boldsymbol{z}(t)$. The solution that holds for *any* test function $\boldsymbol{y}$ is known as the weak solution, and the problem can be posed as:

$$\int_0^1 \left\langle \frac{d\boldsymbol{z}}{dt}(t) - \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}\left(\boldsymbol{z}(t)\right), \boldsymbol{y}(t) \right\rangle dt = 0. \tag{2.2}$$

We take the analogous steps for the initial value condition for test value $\boldsymbol{y}_0$ to find

$$\langle \boldsymbol{z}(0) - \boldsymbol{z}_0, \boldsymbol{y}_0 \rangle = 0. \tag{2.3}$$

3 Inverse Problems

Given the forward problem, we can pose the inverse problem as follows.

Consider a geodesically complete manifold $\mathcal{M}$ equipped with an unknown metric tensor $\boldsymbol{g}(\boldsymbol{x})$, but that we have been given a finite number of collections of two points on the manifold and their geodesic distance between them, $\left\{ \left(\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_1^{(i)}, l^{(i)} \right) \right\}_{i=1}^{N_{\text{geo}}}$. Then, what inferences can we make about the metric, $\boldsymbol{g}(\boldsymbol{x})$?

3.1 Introduction to Inverse Problems

We first discuss a general inverse problem. For simplicity, we first consider a deterministic problem without data errors.

We can think of a forward problem as follows. Assuming that a physical theory (which requires additional model parameters), describes a physical system, given an input model parameter $m \in \mathfrak{M}$ we can output synthesized data $d \in \mathfrak{D}$, using some forward mapping operator $a : \mathfrak{M} \to \mathfrak{D}$, as

$$a(m) = d,$$

where $\mathfrak{M}$ is the model space and $\mathfrak{D}$ is the data space. Then, the inverse problem is to make quantitative inferences about the model parameters m, given input data d.

In the case of the article, the model parameter is the metric, $\boldsymbol{g}(\boldsymbol{x})$. This is a multi-component, continuously varying function, member of an infinite-dimensional function space. Meanwhile, the data, the lengths of the geodesics, $\{l^{(i)}\}_i$, are a set of numbers, member of a finite-dimensional space. Therefore, even if there exists a solution, it cannot be unique. This is a feature of most inverse problems - they are severely underdetermined.

To attempt to solve the inverse problem, we define a misfit $\hat{J}$, such as the least-squares misfit, as follows:

$$\hat{J}[m] := \frac{1}{2} \langle a(m) - d, a(m) - d \rangle,$$

where $\langle \cdot, \cdot \rangle$ is some appropriate inner product, here on the data space. Then, we try to find the model parameter m that minimises the misfit. Defining the functional derivative of J with respect to m, $D_m \hat{J}$, through

$$\hat{J}[m + \delta m] - \hat{J}[m] \equiv \left\langle D_m \hat{J}, \delta m \right\rangle + \mathcal{O}(\delta m^2), \tag{3.1}$$

in the vicinity of an optimising parameter $m_{\min}$ such that $D_m \hat{J}(m_{\min}) = 0$, the misfit can be linearised as

$$\hat{J}[m_{\min} + m] = \hat{J}[m_{\min}] + \frac{1}{2}\Big\langle \langle D_m a(m_{\min}), m\rangle - d, \langle D_m a(m_{\min}), m\rangle - d \Big\rangle + \mathcal{O}(m^3). \tag{3.2}$$

Note that we assumed that both $\mathfrak{M}$, $\mathfrak{D}$ are Hilbert spaces with appropriate inner products.

Assuming that the operator $A_{\min} : m \mapsto \langle D_m a(m_{\min}), m\rangle$, is a continuous, linear functional, we now consider the linear problem of the continuous, linear forward mapping operator $A \in \hom(\mathfrak{M}, \mathfrak{D})$, such that

$$Am = d.$$

Such linearisation is often appropriate for the context of seismic tomography, where the variation from the known model m_0 (such as the spherically symmetric model for the case of the solid earth) is small, eg. the velocity variation from the spherical model is only $\pm 2\%$ [12].

We define the null space, or alternatively the kernel of a linear operator A, $\ker(A)$, as

$$\ker(A) := \{m_{\text{null}} \in \mathfrak{M} \mid Am_{\text{null}} = 0\}. \tag{3.3}$$

This is the subspace of $\mathfrak{M}$ such that adding components of the subspace to the model does not affect the synthesized data at all, and therefore the misfit also.

Suppose we found a solution m such that $Am = d$. Then, for any $m_{\text{null}} \in \ker(A)$, $(m + m_{\text{null}})$ is also a solution as $A(m + m_{\text{null}}) = Am + Am_{\text{null}} = Am = d$. Therefore, assuming that a solution exists, it is only unique if the null space is trivial, i.e. if $\ker(A) = \{0\}$. For $\dim(\mathfrak{M}) > \dim(\mathfrak{D})$, such as the geodesic problem, the null space is not trivial, which means that there are an infinite number of solutions.

We define the adjoint operator of A, $A^\dagger \in \hom(\mathfrak{D}, \mathfrak{M})$, that maps from the data space to the model space such that

$$\langle Am, d\rangle_{\mathfrak{D}} \equiv \langle m, A^\dagger d\rangle_{\mathfrak{M}} \tag{3.4}$$

for all $m \in \mathfrak{M}$, $d \in \mathfrak{D}$.

We define the image of an operator $A^\dagger \in \hom(\mathfrak{D}, \mathfrak{M})$, $\operatorname{im}(A^\dagger)$, as

$$\operatorname{im}(A^\dagger) := \left\{A^\dagger d \in \mathfrak{M} \mid d \in \mathfrak{D}\right\}. \tag{3.5}$$

This is the the subspace of $\mathfrak{M}$ that $A^\dagger$ can map to, given any element $d \in \mathfrak{D}$.

It can be shown that the model space $\mathfrak{M}$ can be split into two subspaces as [8]:

$$\mathfrak{M} = \ker(A) \oplus \operatorname{im}(A^\dagger), \tag{3.6}$$

where the two components are orthogonal to each other.

3.2 Gradient Based Optimisation

As discussed in the introduction, the challenge of severely-underdertermined inverse problems is usually tackled by finite parametrisation and regularisation, somewhat unsatisfactory measures. An alternative method to find a solution is through the method of gradient-based optimisation. The crudest method is the steepest-descent method (SD). Schematically, we: a) calculate the gradient of the misfit with respect to the model, $D_m \hat{J}$ at some initial guess m_{start}, b) update the guess in the direction negative to gradient $m \mapsto m' = m - \alpha D_m \hat{J}(m)$, where scalar $\alpha > 0$ is chosen through a line search such that it minimises the misfit in that direction, and c) repeat until we find $m_{\min}$ such that $D_m \hat{J}(m_{\min}) = 0$, to some tolerance. We can understand $m_{\min}$ as the local minimum of the misfit.

Other gradient-based optimisation schemes choose a different descent direction, and may converge faster than SD. The non-linear conjugate gradient method (NLCG) uses information about the previous gradient [14]. The limited-memory BFGS method (LBFGS) is a pseudo-Newton method that uses an estimate of the second order gradient by storing some number of the previous calculated gradients [15]. Both of these methods are currently used in seismic tomography. The full Newton method requires the full information of the second order gradient of the misfit (in the Hessian), and has the potential to reduce the number of iterations needed [9]. These methods are sketched in Fig. 2.

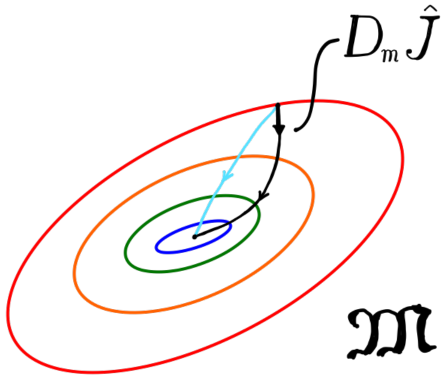


Figure 2: A sketch of gradient based optimisation. The steepest descent method, (black), may be superceded by methods that estimate the curvature of the misfit (light blue).

If the optimising model is found, the second order gradients can be used to constrain the solution in certain directions. At the minimum, we can expand the misfit around the minima as (3.2), to find

$$D_m \hat{J}|_{m_{\min}} = A^{\dagger}_{\min} \left(A_{\min} m - d\right) + \mathcal{O}(m^3), \tag{3.7}$$

to approximate the Hessian

$$D_m D_m \hat{J}|_{m_{\min}} = A^{\dagger}_{\min} A_{\min} + \mathcal{O}(m^3). \tag{3.8}$$

We now use a result from linear algebra, singular value decomposition: any real matrix A can be decomposed into $A = L^{\dagger} \Lambda R$, where $\Lambda = \mathrm{diag}(\lambda_1, ...)$, and L, R are orthogonal matrices, and $(\dagger)$ denote the hermitian conjugate here. For l_i: the ith column of L, r_i: the ith column of R, the following identities hold:

$$A^{\dagger} l_i = \lambda_i r_i, \tag{3.9}$$

$$A r_i = \lambda_i l_i. \tag{3.10}$$

Since L, R are orthogonal matrices, the set $\{l_i\}_i$ an $\{r_i\}_i$ are orthonormal sets of vectors. They are eigenvectors of $AA^{\dagger}$ and $A^{\dagger}A$ respectively:

$$AA^{\dagger} l_i = \lambda_i^2 l_i, \tag{3.11}$$

$$A^{\dagger} A r_i = \lambda_i^2 r_i. \tag{3.12}$$

This result generalises to continuous linear operators, with the analogue of Hermitian conjugate being the adjoint, with the same notation $(\dagger)$. Importantly, $A_{\min} A^{\dagger}_{\min}$ is just a $\dim(\mathfrak{D}) \times \dim(\mathfrak{D})$ matrix acting on the data space, so its eigenvalues $\{\lambda_i^2\}_i$ and eigenvectors $\{l_i\}_i$ can be found easily, using algorithms such as the Jacobi Algorithm [10]. From this, the eigendirections for $A^{\dagger}_{\min} A_{\min}$ can be found. From (3.8), $A^{\dagger}_{\min} A_{\min}$ is the first-order term of the second order functional derivative of $\hat{J}[m_{\min}]$, therefore the eigenvalues $\{\lambda_i^2\}_i$ describe the local curvature/sensitivity of the misfit $\hat{J}$ in model directions $\{r_i\}_i$. This is sketched in Fig. 3.

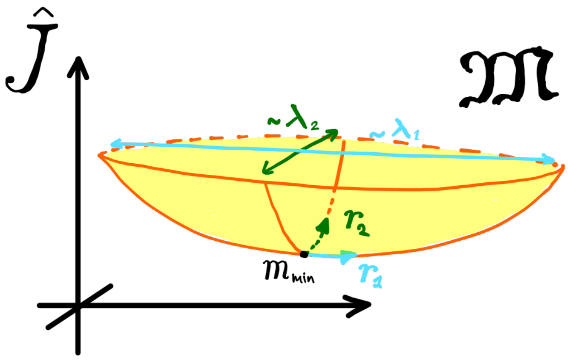


Figure 3: Sketch of the misfit around its local minima. The eigenvalues describe the sensitivity of the misfit along model eigendirections, whose reciprocal is proportional to the error along that direction.

Around $m_{\min}$, a change in model parameter by $\alpha_i r_i$ (α_i is some constant) will result in change in synthesized data as $A\alpha_i r_i = \lambda_i \alpha_i l_i$. Given some errors in the data, $d \pm \epsilon$, where we assume that ϵ has 0 mean and is Gaussian distributed with covariance matrix $C = \mathbb{E}[\epsilon \otimes \epsilon]$, where $\mathbb{E}[\cdot]$ denotes the expectation value and $\otimes$ denotes a tensor product. Then, it can be shown that $\mathbb{E}[\|\epsilon\|] = \sqrt{\operatorname{tr} C}$ [5].

Therefore, we can estimate the error in the model direction r_i by setting the synthesized data perturbation from model direction r_i equal to the error magnitude of the data, as

$$\|A\alpha_i r_i\| = \|\lambda_i \alpha_i l_i\| = \lambda_i |\alpha_i| = \mathbb{E}[\|\epsilon\|] = \sqrt{\operatorname{tr} C}, \tag{3.13}$$

thereby obtaining the estimate of the model error,

$$|\alpha_i| = \frac{\sqrt{\operatorname{tr} C}}{\lambda_i}. \tag{3.14}$$

Consider the linear problem with misfit $\hat{J} = \frac{1}{2}\langle Am - d, Am - d\rangle$. Following the prescription (3.1), first order gradient is exactly

$$D_m \hat{J}(m) = A^\dagger (Am - d). \tag{3.15}$$

Since the optimisation algorithm will update the model by adding some component of the gradient of the misfit, we can see from (3.15) that it will only add components in $\operatorname{im}(A^\dagger)$, as illustrated in Fig. 4.

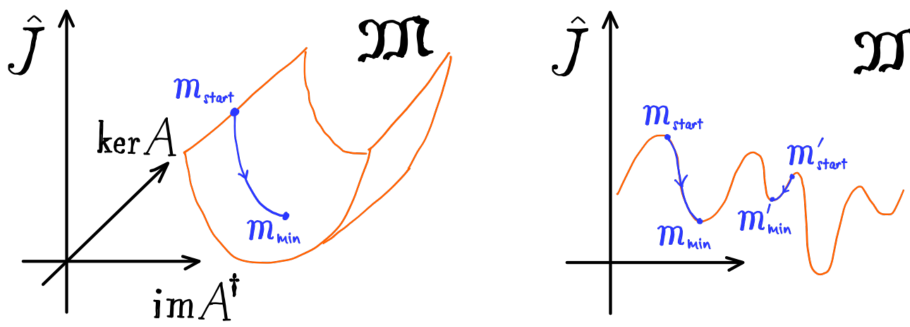


Figure 4: Left: sketch of linear gradient-based optimisation; the only components of $\operatorname{im}(A^\dagger)$ *are added during the optimisation process. Right: non-linear gradient-based optimisation, which may not give a solution that globally minimises the misfit.*

For a linear problem, there is a unique projection of the global minimum to $\operatorname{im}(A^\dagger)$ [13]. Therefore model error estimate (3.14) is accurate in that it effectively constrains the model in the $\operatorname{im}(A^\dagger)$ subspace of the model space $\mathfrak{M}$. For a non-linear problem however, there is no guarantee that the model solution is at a global minimum, as there may be multiple points where the gradient vanishes. One may wish to explore the model space to find a global minima. However, in the case in which the model space is far too large (or infinte) to explore the full model space, this is impractical or impossible, and is one of the open problems in non-linear inverse problems - how do we find the global minima of an infinite-dimensional model space? Therefore (3.14) only constrains the model around the local minimum. These ideas are illustrated in Fig. 4.

3.3 Adjoint Methods

The adjoint method can be used to effectively calculate gradient of the misfit with respect to model m, such that each gradient calculation only requires one solution to the forward problem and another closely related forward problem called the adjoint problem. We describe the method schematically here.

Consider an objective functional $J[m, u]$, dependent on u, the forward variables (ex. path of the geodesic $\boldsymbol{x}(t)$), and m the model parameters. Since the forward variables are ultimately dependent on the model parameters, we define the reduced objective functional $\hat{J}[m] = J[m, u(m)]$. We look to find the derivatives of $\hat{J}$.

If the dependence of u on m is enforced by a forward problem $b(u, m) = 0$ (such as the EoM), instead of calculating $D_m \hat{J}$ directly, we can construct a Lagrangian

$$L[u, u^{\dagger}, m] = J(m, u) + \langle b(u, m), u^{\dagger} \rangle.$$

We make use of the Lagrangian Multiplier Theorem (LMT), which states that

$$D_m \hat{J}[m] = D_m L[u, u^{\dagger}, m] \quad \text{if} \quad u, u^{\dagger} \ \text{ such that } \quad D_u L = 0, D_{u^{\dagger}} L = 0.$$

By inspection, from the condition $D_{u^{\dagger}} L = 0$, we retrieve the solution to the weak form of the forward problem. Meanwhile the condition $D_u L = 0$ creates another forward problem, called the adjoint problem.

The adjoint method can be used to evaluate the the second order gradient of the reduced objective functional $\hat{J}$ along some fixed model perturbation direction $m^{\sharp}$. We define another objective functional

$$J_1(m, u, u^{\dagger}) = \left\langle D_m L, m^{\sharp} \right\rangle.$$

To enforce the forward equations and adjoint equations we define

$$L_1(m, u, u^{\dagger}, u^{\sharp}, u^{\dagger\sharp}) := J_1(m, u, u^{\dagger}) + \left\langle D_u, L, u^{\sharp} \right\rangle + \left\langle D_u^{\dagger}, L, u^{\dagger\sharp} \right\rangle,$$

where we introduced the second-order adjoint variables $u^{\sharp}, u^{\dagger\sharp}$. We note that the gradient of the second order reduced functional can be used to evaluate to the second order gradient of the original reduced objective functional $\hat{J}$ along direction $m^{\sharp}$:

$$D_m \hat{J}_1 = \left\langle D_m D_m \hat{J}, m^{\sharp} \right\rangle.$$

We can find this quantity again by using LMT:

$$D_m \hat{J}_1 = D_m L_1 \quad \text{if} \quad D_u L_1 = D_{u^{\dagger}} L_1 = D_{u^{\sharp}} L_1 = D_{u^{\dagger\sharp}} L_1 = 0.$$

3.4 Derivation: The Geodesic Shooting Problem

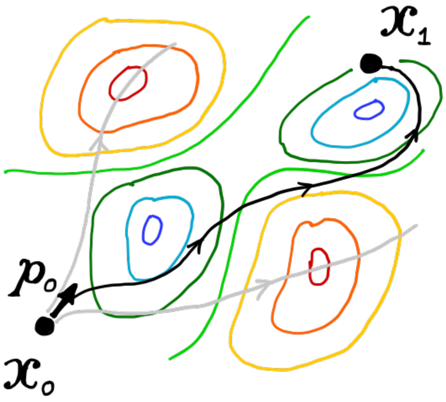


Figure 5: A sketch of the geodesic shooting problem, where we vary the initial momentum or 'shooting parameter' of the geodesic until we find one that reaches the desired point.

To illustrate these concepts, we consider the geodesic shooting problem, where, on a manifold with a known metric $\boldsymbol{g}(\boldsymbol{x})$, we fix the start and end points of the geodesic, $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, and find $\boldsymbol{p}_0$ such that the geodesic connects $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$ in unit parameter time $t_1 - t_0 = 1$. This is sketched in Fig. 5.

We cast this as an optimisation problem. We define the misfit

$$J[\boldsymbol{z}] = \frac{1}{2}\Big\langle \boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1, \boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1 \Big\rangle,$$

where we utilised the projection matrix to x-coordinates,

$$\boldsymbol{\pi}_x = \begin{bmatrix} \mathbf{1}_n & \mathbf{0}_n \end{bmatrix}.$$

To impose EoM (2.2) and initial value condition (2.3) in their weak form, we write down the Lagrangian

$$L[\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{y}_0, \boldsymbol{p}_0] = \frac{1}{2}\Big\langle \boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1, \boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1 \Big\rangle + \int_0^1 \Big\langle \boldsymbol{y}, \frac{d\boldsymbol{z}}{dt} - \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}) \Big\rangle \, dt + \Big\langle \boldsymbol{y}_0, \boldsymbol{z}(0) - \boldsymbol{z}_0 \Big\rangle, \tag{3.16}$$

where we introduced test function and value $\boldsymbol{y}(t)$ and $\boldsymbol{y}_0$. We define the reduced objective functional $\hat{J}[\boldsymbol{p}_0] = J[\boldsymbol{z}(\boldsymbol{p}_0)]$, where $\boldsymbol{z}(t)$ adheres to the EoM and initial condition. From LMT,

$$D_{\boldsymbol{p}_0} \hat{J}[\boldsymbol{p}_0] = D_{\boldsymbol{p}_0} L[\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{p}_0, \boldsymbol{y}_0],$$

if

$$D_{\boldsymbol{z}} L = \mathbf{0}, \quad D_{\boldsymbol{y}} L = \mathbf{0}, \quad D_{\boldsymbol{y}_0} L = \mathbf{0}. \tag{3.17}$$

By inspection of (3.16),

$$\Big\langle D_{\boldsymbol{y}} L, \delta \boldsymbol{y} \Big\rangle = \int_0^1 \Big\langle \frac{d\boldsymbol{z}}{dt} - \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}), \delta \boldsymbol{y} \Big\rangle dt, \tag{3.18}$$

$$\Big\langle D_{\boldsymbol{y}_0} L, \delta \boldsymbol{z} \Big\rangle = \Big\langle \boldsymbol{z}(0) - \boldsymbol{z}_0, \delta \boldsymbol{y}_0 \Big\rangle. \tag{3.19}$$

These must vanish for all $\delta \boldsymbol{y}$ and $\delta \boldsymbol{y}_0$, so the left hand side of the Euclidean inner prouct must vanish: we retrieve the initial value constraint and EoM. We can similarly vary (3.16) with respect to $\boldsymbol{z}$, integrating by parts to get the boundary conditions, using the adjoint x-projection matrix $\boldsymbol{\pi}_x^* = [\mathbf{1}_n \;\; \mathbf{0}_n]^T$, to find

$$\Big\langle D_{\boldsymbol{z}} L, \delta \boldsymbol{z} \Big\rangle = \Big\langle \boldsymbol{y}_0 - \boldsymbol{y}(0), \delta \boldsymbol{z}(0) \Big\rangle + \Big\langle \boldsymbol{\pi}_x^*(\boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1) + \boldsymbol{y}(1), \delta \boldsymbol{z}(1) \Big\rangle - \int_0^1 \Big\langle \frac{d\boldsymbol{y}}{dt} - \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}} \mathbb{J} \boldsymbol{y}, \delta \boldsymbol{z} \Big\rangle, \tag{3.20}$$

Which must also vanish for all $\delta \boldsymbol{z}$ by (3.17). This gives the adjoint EoM, initial and terminal value constraints. Finally,

$$\Big\langle D_{\boldsymbol{p}_0} L, \delta \boldsymbol{p}_0 \Big\rangle = \Big\langle -\boldsymbol{\pi}_p \boldsymbol{y}_0, \delta \boldsymbol{p}_0 \Big\rangle, \tag{3.21}$$

where the projection matrix to the p-coordinates is

$$\boldsymbol{\pi}_p = \begin{bmatrix} \mathbf{0}_n & \mathbf{1}_n \end{bmatrix},$$

which, substituting the other variables, will give us the gradient of the misfit $\hat{J}$. The optimisation algorithm then would follow the steps as follows.

1. Choose some value of $\boldsymbol{p}_0$.
2. For this value of $\boldsymbol{p}_0$, solve the forward problem (3.18) to find $\boldsymbol{z}(t)$, and therefore $\boldsymbol{z}(1)$ also.
3. Solve the adjoint problem (3.20) (now with the terminal value $\boldsymbol{y}(1)$ known) to find $\boldsymbol{y}(t)$ and therefore $\boldsymbol{y}(0)$ also, and therefore $\boldsymbol{y}_0 = \boldsymbol{y}(0)$ can be found.
4. Finally, from (3.21) and LMT, $D_{\boldsymbol{p}_0} L = D_{\boldsymbol{p}_0} \hat{J}$ can be determined.
5. Update $\boldsymbol{p}_0$ in the negative direction of the gradient using some scheme; repeat steps 1-4 until misfit is minimised, i.e. $D_{\boldsymbol{p}_0} L = D_{\boldsymbol{p}_0} \hat{J} = 0$.

However, we note that the adjoint problem is a terminal value problem rather than an initial value problem. To turn it into an initial value problem, we define the adjoint variable $\boldsymbol{z}^\dagger$ such that

$$\boldsymbol{y}(t) =: \boldsymbol{z}^\dagger(1-t) = (\mathcal{P}\boldsymbol{z}^\dagger)(t),$$

where

$$\mathcal{P} : f(t) \mapsto f(1-t) \tag{3.22}$$

is the time reversal operator. Note that (†) does not refer to the Hermitian conjugate; it is just the adjoint label. The adjoint problem then becomes an initial value problem, with

$$\langle D_{\boldsymbol{z}} L, \delta \boldsymbol{z} \rangle = \Big\langle \boldsymbol{z}_0^\dagger - \boldsymbol{z}^\dagger(1), \delta \boldsymbol{z}(0) \Big\rangle + \Big\langle \boldsymbol{\pi}_x^* (\boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1) + \boldsymbol{z}^\dagger(0), \delta \boldsymbol{z}(1) \Big\rangle \tag{3.23}$$

$$+ \int_0^1 \Big\langle \frac{d\boldsymbol{z}^\dagger}{dt} + \Big\{ \mathcal{P} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{z}(t)) \mathbb{J} \mathcal{P} \Big\} \boldsymbol{z}^\dagger, \mathcal{P} \delta \boldsymbol{z} \Big\rangle, \tag{3.24}$$

where we used the anticommutation relation $\{\frac{\partial}{\partial t}, \mathcal{P}\} = 0$, and the relation $\int_0^1 \langle \cdot, \cdot \rangle dt = \int_0^1 \langle \mathcal{P} \cdot, \mathcal{P} \cdot \rangle dt$. We also relabel $\boldsymbol{y}_0 \mapsto \boldsymbol{z}_0^\dagger$. Then, order in which the variables are specified is:

$$\boldsymbol{p}_0 \to \boldsymbol{z} \to \boldsymbol{z}^\dagger \to \boldsymbol{z}_0^\dagger \to D_{\boldsymbol{p}_0} \hat{J}.$$

3.5 Sobolev Spaces

To implement a gradient-based optimisation where model space is a function space, we must work in an inner product space such that the terms $\langle D_m L, \delta m \rangle$ can be written as an inner product of the perturbation with some kernel R, $\langle D_m L, \delta m \rangle = (R, \delta m)$. Then we can add components of the functional gradients to the prior model, $m \mapsto m + \alpha R$. Some of the necessary conditions are:

1. This space must be a Hilbert space (the space must be equipped with an inner product).
2. We require this space to have the properties that the physical theory requires, such as pointwise convergence, and some level of differentiability. However, we do not want to impose other properties if possible.

3. The linear functional $\delta m \mapsto \langle D_m L, \delta m \rangle$ must be contained in the dual of this space.

Sobolev spaces can fulfill these three requirements. We discuss these ideas for a n-dimensional torus $\mathbb{T}^n$ of coordinate length 2π, and numerically implement the scheme for the linearised problem of torus dimension $n = 2$. However, the results that follows, namely Sobolev Embedding Theorem and Riesz Representation Theorem, can be extended to other manifolds [16].

3.5.1 Fourier Series

The set of square integrable functions, $L^2(\mathbb{T}^n)$, is a function space such that

$$L^2(\mathbb{T}^n) := \left\{ f : \mathbb{T}^n \mapsto \mathbb{C} \;\middle|\; \int_{\mathbb{T}^n} |f(\boldsymbol{\theta})|^2 d^n\boldsymbol{\theta} < \infty \right\},$$

equipped with an inner product $(\cdot\,,\,\cdot)_{L^2}$, such that $\forall\, f, g \in L^2(\mathbb{T}^n)$,

$$(f, g)_{L^2} := \int_{\mathbb{T}^n} f(\boldsymbol{\theta})\overline{g(\boldsymbol{\theta})}\, \frac{d^n\boldsymbol{\theta}}{(2\pi)^n},$$

where the line over a symbol denotes complex conjugation. As it has an inner product defined, it is a Hilbert space. Define the Fourier transform operator $\mathcal{F}$ as

$$\mathcal{F} : f \mapsto \tilde{f} \quad \text{such that} \quad \tilde{f}_{\boldsymbol{k}} = \int_{\mathbb{T}^n} f(\boldsymbol{\theta}) e^{-i\langle \boldsymbol{k}, \boldsymbol{\theta}\rangle}\, \frac{d^n\boldsymbol{\theta}}{(2\pi)^n}.$$

It can be shown that $\mathcal{F} : L^2(\mathbb{T}^n) \to \ell^2(\mathbb{Z}^n)$, where $\ell^2(\mathbb{Z}^n)$ is a sequence space on the n-dimensional integers such that

$$l^2(\mathbb{T}^n) := \left\{ u : \mathbb{Z}^n \to \mathbb{C} \;\middle|\; \sum_{\boldsymbol{k}\in\mathbb{Z}^n} |u_{\boldsymbol{k}}|^2 < \infty \right\},$$

equipped with an inner product $(\cdot\,,\,\cdot)_{l^2}$ such that $\forall\, u, v \in \ell^2(\mathbb{Z}^n)$,

$$(u, v)_{l^2} := \sum_{\boldsymbol{k}\in\mathbb{Z}^n} u_{\boldsymbol{k}} \overline{v_{\boldsymbol{k}}}. \tag{3.25}$$

Similarly, define the inverse Fourier transform operator as:

$$\mathcal{F}^* : u \mapsto (\mathcal{F}^* u) \quad \text{such that} \quad (\mathcal{F}^* u)(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} u_{\boldsymbol{k}} e^{+i\langle \boldsymbol{k}, \boldsymbol{\theta}\rangle}.$$

It can be shown that $\mathcal{F}^*\mathcal{F} = \mathrm{id}$ on $L^2(\mathbb{T}^n)$, that there is an isomorphism between $L^2(\mathbb{T}^n)$ and $\ell^2(\mathbb{Z}^n)$, and that the identity $(\mathcal{F}f, u)_{L^2} = (f, \mathcal{F}^* u)_{\ell^2}$ holds. Therefore, any function in $L^2(\mathbb{T}^n)$ can be written as its Fourier series. For brevity, we refer to [7] for a full proof.

Importantly, it is not guaranteed that a function in $L^2(\mathbb{T}^n)$ will be pointwise convergent or continuous. This is because the L^2 norm is invariant under a shift of a single point.

3.5.2 Sobolev Embedding Theorem

We define Sobolev spaces as follows:

$$H^s_\lambda := \left\{ f \in L^2(\mathbb{T}^n) \mid \left\{ \langle \boldsymbol{k} \rangle^s_\lambda \tilde{f}_{\boldsymbol{k}} \right\}_{\boldsymbol{k}} \in \ell^2(\mathbb{Z}^n) \right\},$$

or, alternatively, defining $\Delta := -\nabla^2$,

$$H^s_\lambda := \left\{ f \in L^2(\mathbb{T}^n) \mid \left(1 + \lambda^2\Delta\right)^{s/2} f \in L^2(\mathbb{T}^n) \right\},$$

equipped with the inner product $(\cdot,\cdot)_{H^s_\lambda}$ such that $\forall\, f, g \in H^s_\lambda(\mathbb{T}^n)$,

$$(f,g)_{H^s_\lambda} := \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \langle \boldsymbol{k}\rangle_\lambda^{2s} \tilde{f}_{\boldsymbol{k}} \overline{\tilde{g}_{\boldsymbol{k}}}, \tag{3.26}$$

where

$$\langle \boldsymbol{k}\rangle_\lambda := \sqrt{1+\lambda^2\boldsymbol{k}^2}, \tag{3.27}$$

and λ is some characteristic length parameter, which will be discussed later.

The Sobolev Embedding Theorem states that functions in $H^s_\lambda(\mathbb{T}^n)$ are differentiable up to l orders of if $s > \frac{n}{2}+l$, that is [17]

$$H^s_\lambda(\mathbb{T}^n) \subseteq C^l(\mathbb{Z}^n) \quad \text{if} \quad s > \frac{n}{2}+l.$$

Therefore we can choose an appropriate s given the dimension of space n and desired differentiability l.

3.5.3 Dual Spaces

A dual space of a function space is the set of all continuous, linear *functionals* on that space. Conventionally, we denote it by a dash $(')$ after the function space we are interested in.

Consider a function $f \in C^0(\mathbb{T}^n)$, and a linear functional f' such that $C^0(\mathbb{T}^n)' \ni f' : f \mapsto f'[f]$. We now extend the Fourier transform to linear functionals. For any $f' \in C^0(\mathbb{T}^n)'$, we define its Fourier transform to be

$$\mathcal{F} : f' \mapsto \tilde{f}' \quad \text{such that} \quad \tilde{f}'_{\boldsymbol{k}} = f'\left[e^{-i\langle \boldsymbol{k},\cdot\rangle}\right],$$

where we omit the argument of the complex exponential to write denote it as the entire function without supplied arguments. We note that the formal Fourier series is not necessarily convergent.

For any $f \in L^2(\mathbb{T}^n)$, $f' \in L^2(\mathbb{T}^n)'$, using the linear property of f',

$$f'[f] = f'\left[\sum_{\boldsymbol{k}\in\mathbb{Z}^n} \tilde{f}_{\boldsymbol{k}} e^{+i\langle \boldsymbol{k},\cdot\rangle}\right] = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \tilde{f}_{\boldsymbol{k}} f'\left[e^{+i\langle \boldsymbol{k},\cdot\rangle}\right] = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \tilde{f}_{\boldsymbol{k}} \tilde{f}'_{-\boldsymbol{k}}. \tag{3.28}$$

3.5.4 Riesz Representation Theorem

We work in $H^s_\lambda(\mathbb{T}^n)$. Consider $f \in H^s_\lambda(\mathbb{T}^n)$, $f' \in H^s_\lambda(\mathbb{T}^n)'$. Then, by (3.28),

$$f'[f] = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \tilde{f}_{\boldsymbol{k}} \tilde{f}'_{-\boldsymbol{k}} = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \langle \boldsymbol{k}\rangle_\lambda^{2s} \tilde{f}_{\boldsymbol{k}} \left(\langle \boldsymbol{k}\rangle_\lambda^{-2s} \tilde{f}'_{-\boldsymbol{k}}\right) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \langle \boldsymbol{k}\rangle_\lambda^{2s} \tilde{f}_{\boldsymbol{k}} \overline{\tilde{R}_{\boldsymbol{k}}},$$

where we formally defined the Fourier coefficient of the representation R of f' as

$$\tilde{R}_{\boldsymbol{k}} := \langle \boldsymbol{k}\rangle_\lambda^{-2s} \overline{\tilde{f}'_{-\boldsymbol{k}}}. \tag{3.29}$$

By (3.26), this is in the form of the Sobolev inner product

$$f'[f] = (f, R)_{H^s_\lambda},$$

given that R is also a member of the Sobolev space, that is if $\left\{\langle \boldsymbol{k}\rangle_\lambda^{s} \tilde{R}_{\boldsymbol{k}}\right\}_{\boldsymbol{k}} \in \ell^2(\mathbb{Z}^n) \leftrightarrow \left\{\langle \boldsymbol{k}\rangle_\lambda^{-s} \tilde{f}'_{\boldsymbol{k}}\right\}_{\boldsymbol{k}} \in \ell^2(\mathbb{Z}^n)$, that is if $f' \in H^{-s}_\lambda(\mathbb{T}^n)$. Therefore we can identify the isomorphism

$$H^s(\mathbb{T}^n)' \cong H^{-s}(\mathbb{T}^n).$$

3.5.5 Representations of Distributions on Sobolev Spaces

Upon calculation of the first adjoint for them metric tomographic problem, we will find that the functional gradient terms will have the form such as

$$\langle D_{\boldsymbol{g}}\hat{J}, \delta\boldsymbol{g}\rangle = \int_0^1 \langle F(\boldsymbol{x}(t)), \delta\boldsymbol{g}(\boldsymbol{x}(t))\rangle \, dt,$$

where F is some function. If we were to choose to work in an L^2 space, this can be presented in a form of an L^2 inner product, to retrieve

$$= \int_{\mathbb{T}^n} \langle \delta_\gamma F(\boldsymbol{x}), \delta\boldsymbol{g}(\boldsymbol{x})\rangle \, d^n\boldsymbol{x} \overset{\text{in } L^2}{\Longrightarrow} D_{\boldsymbol{g}}\hat{J} = \delta_\gamma F(\boldsymbol{x}), \tag{3.30}$$

where δ_γ is a delta function along the curve $\boldsymbol{x}(t)$. We note that, since we will be iteratively adding components of gradient to update the model, if we work in L^2, we will *not* get a point-wise convergent solution. To find solutions with point-wise convergence and differentiability up to some order, we must work in a Sobolev space. We identify that the linear functional $\delta\boldsymbol{g} \mapsto \int_{\mathbb{T}^n} \langle \delta_\gamma F(\boldsymbol{x}), \delta\boldsymbol{g}(\boldsymbol{x})\rangle \, d^n\boldsymbol{x}$ is a member of $C^0(\mathbb{T}^n)'$. From the Sobolev embedding theorem, $H^s_\lambda(\mathbb{T}^n) \subseteq C^0(\mathbb{Z}^n)$ given $s > \frac{n}{2}$. Therefore $C^0(\mathbb{Z}^n)' \subseteq H^s_\lambda(\mathbb{T}^n)' \cong H^{-s}_\lambda(\mathbb{T}^n)$ and so this linear functional has a valid representation in any Sobolev space such that $s > n/2$, and its representation can be constructed from its Fourier coefficients using the prescription (3.29) as

$$\tilde{R}_{\boldsymbol{k}} := \langle \boldsymbol{k}\rangle_\lambda^{-2s} \overline{\left(\int_0^1 F(\boldsymbol{x}(t))\, e^{-i\langle -\boldsymbol{k}, \boldsymbol{x}(t)\rangle} dt\right)}. \tag{3.31}$$

4 Linear metric tomography

4.1 Linearisation of the Forward Problem

We consider the simple, linearly perturbative problem of a flat, isotropic metric with a small, varying distortion field. In an isotropic medium, the metric is

$$g_{\mu\nu} = \frac{1}{c(x)^2}\delta_{\mu\nu},$$

where $c(x)$ is the velocity field.

On such a metric, consider a geodesic $x(t)$, connecting two coordinates $x(0) = x_0$ and $x(1) = x_1$. Its length is

$$l = \int_0^1 \sqrt{g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu} \, dt = \int_0^1 \frac{1}{c(x)} \|\dot{\boldsymbol{x}}\|_{\mathbb{R}^n} \, dt,$$

where $\|\cdot\|_{\mathbb{R}^n}$ denotes a Euclidean norm.

Now, consider a flat background metric c_0, perturbed by $c_1(\boldsymbol{x}) \ll c_0$,

$$c(x) = c_0 \to c'(x) = c_0 + c_1(x).$$

Upon this perturbation, the geodesic length will change with terms as

$$l \to l' = l + \begin{pmatrix}\text{metric perturbed,} \\ \text{same path}\end{pmatrix} + \begin{pmatrix}\text{same metric,} \\ \text{perturbed path}\end{pmatrix} + \dots .$$

Fermat's principle states that the first order contribution to the change in length of the geodesic is from the metric perturbation only [13]. The contribution from only the metric perturbation can be found as

$$l' = \int_0^1 \frac{1}{c_0 + c_1(\boldsymbol{x})} \|\dot{\boldsymbol{x}}\|_{\mathbb{R}^n} \, dt = \int_0^1 \frac{\|\dot{\boldsymbol{x}}\|_{\mathbb{R}^n}}{c_0} \left(1 - \frac{c_1(x)}{c_0} + \mathcal{O}\left([c_1/c_0]^2\right)\right) dt.$$

So, to linear order, the perturbation to the geodesic length from the perturbation to the metric is

$$\delta l = -\int_0^1 \frac{\|\dot{\boldsymbol{x}}\|_{\mathbb{R}^n}}{c_0^2} c_1(x)\, dt,$$

where $\dot{\boldsymbol{x}}(t)$ is the solution on the unperturbed metric. Noting that on the unperturbed metric,

$$\boldsymbol{p} = \frac{\partial L}{\partial \dot{\boldsymbol{x}}} = \boldsymbol{g}(x)\dot{\boldsymbol{x}} = \frac{\dot{\boldsymbol{x}}}{c_0^2},$$

the perturbation of geodesic length then is

$$\delta l = -\int_0^1 \|\boldsymbol{p}(t)\|_{\mathbb{R}^n}\, c_1(\boldsymbol{x}(t))\, dt = -\int_0^1 \|\boldsymbol{p}_0\|_{\mathbb{R}^n}\, c_1(\boldsymbol{x}(t))\, dt,$$

where both $\boldsymbol{x}(t)$ and $\boldsymbol{p}(t)$ are the solutions to the unperturbed problem.

The unperturbed problem can be solved using the EoM. Substituting the unperturbed metric into the Hamiltonian, we find that

$$H = \frac{1}{2} c_0^2 \, \|\boldsymbol{p}\|_{\mathbb{R}^n}^2 .$$

From Hamilton's equations, for each geodesic, we can find that

$$\frac{d\boldsymbol{x}}{dt} = c_0^2 \boldsymbol{p} = c_0^2 \boldsymbol{p}_0 \quad ; \quad \frac{d\boldsymbol{p}}{dt} = \boldsymbol{0},$$

with initial conditions $\boldsymbol{x}(0) = \boldsymbol{x}_0$, $\boldsymbol{p}(0) = \boldsymbol{p}_0$. We can simultaneously find the perturbation on the geodesic length along with the EoM by simultaneously solving

$$\frac{dl}{dt} = -\|\boldsymbol{p}_0\|_{\mathbb{R}^n}\, c_1(\boldsymbol{x}(t)),$$

from $t \in [0, 1]$ with initial value $\delta l(0) = 0$, for each of N_{geo} geodesics.

From these equations, we can solve the linearised forward problem, solving for unperturbed EoM and the perturbed geodesic length.

4.2 Linearised Inverse Problem

Working in an Sobolev space, we can cast the action of the linear functional on c_1, $c_1 \mapsto -\int_0^1 \|\boldsymbol{p}_0\|_{\mathbb{R}^n}\, c_1(\boldsymbol{x}(t))\, dt$, as an inner product between c_1 and representation (also known as kernel) R. We construct the kernel R of the linear functional from its Fourier series. Following the prescription (3.29), similarly to (3.31), its Fourier coefficients will be given by

$$\tilde{R}_{\boldsymbol{k}} = -\frac{1}{\left(1 + \lambda^2 \boldsymbol{k}^2\right)^s} \int_0^1 \|\boldsymbol{p}_0\|_{\mathbb{R}^n}\, e^{-i\langle -\boldsymbol{k}\cdot\boldsymbol{x}(t)\rangle}\, dt,$$

from which we can construc the kernel as

$$R(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{Z}^n} \tilde{R}_{\boldsymbol{k}} e^{+i\boldsymbol{k}\cdot\boldsymbol{x}}.$$

Then,

$$\delta l = (c_1, R)_{H^s} .$$

For each geodesic, this must hold. We define the data vector

$$d = \sum_{i=1}^{N_{geo}} \delta l^{(i)} e_i,$$

where the coefficients are given by N_{geo} perturbed lengths, and e_i is a unit vector in data space, $\mathbb{R}^{N_{\text{geo}}}$.

We can deduce the linear forward operator to be $A = \sum_i e_i \otimes R^{(i)}$, since

$$d = \sum_i \left(c_1, R^{(i)}\right)_{H^s} e_i = \left(\sum_i e_i \otimes R^{(i)}\right) c_1,$$

where we used the tensor product notation $(a \otimes b)c := (c, a)b$.

We define the misfit

$$\hat{J}(c_1) := \frac{1}{2}\left\langle Ac_1 - d, Ac_1 - d\right\rangle.$$

Note that, because the model parameters do not change the EoM to first order, we do not require adjoint methods, as the kernels $\left\{R^{(i)}\right\}_i$ are fixed. The first order gradients can be found to be

$$D_{c_1}\hat{J}(c_1) = A^\dagger \left(Ac_1 - d\right),$$

where the adjoint operator $A^\dagger = \sum_i R^{(i)} \otimes e_i$.

A choice that we must make is the starting guess of the model m_{start}. We note that iteration process only adds or subtracts model components $\in \text{im}(A^\dagger)$. To make sure that the final model only has information required by the data, a sensible choice of an initial condition one such that m_{start} is the zero element in the $\ker(A)$ subspace, $c_1(\boldsymbol{x}) = 0$. This way, the optimal model will only contain information given by the data.

The second order gradient can be found by varying $\langle D_{c_1}\hat{J}(c_1), c_1^\sharp\rangle$, from which we find that

$$D_{c_1} D_{c_1} \hat{J} = A^\dagger A,$$

which, for the special linear case, is independent of c_1.

4.3 Numerical Implementation

The forward problem and the inverse problem, described in Secs. 4.1 , 4.2 were implemented in the Fortran 90 programming language, for the torus of dimension $n = 2$. To implement the scheme, parametrisation of the velocity perturbation c_1 was discretised into a lattice of linear size N_{lat}, $c_1(\boldsymbol{x}) \mapsto c_1(x_i, y_j)$. In between these lattice points, the function was approximated from its Fourier series, where the Fourier coefficients, (k_x, k_y), spanned $\{-N_{\text{lat}}/2, ..., -1, 0, 1, ..., N_{\text{lat}}/2\}$. The program randomly generated a model m_{gen}, and generated data d_{gen} from solving the forward problem. Then, gradient-based optimisation schemes were used to solve the inverse problem to obtain an optimising model, $c_{1,\text{min}}$. More specifically,

1. The background velocity, c_0 was set.

2. The resolution, N_{lat}, was specified. Then, a perturbation field was created from by generating numbers using a pseudo-random number generator on a lattice of set size $N_{\text{pert}} < N_{\text{lat}}$ between set values $[-c_{1,\text{max}}, c_{1,\text{max}}]$ such that $c_{1,\text{max}} \ll c_0$. It was then interpolated by its Fourier series approximation on the lattice of size N_{lat}. This provided the generated model, $c_{1,\text{gen}}$.

3. The number of geodesics, N_{geo} was specified. Then, two-point pairs, $(\boldsymbol{x}_0, \boldsymbol{x}_1)$, were also generated between $[0, 2\pi]$ for each coordinate using a pseudo-random number generator. Next, the required starting momenta $\boldsymbol{p}_0$ for each geodesic on the background metric was found. Finally, the forward equations of motions were solved the length perturbations were found through solving a Runge-Kutta ODE solver, generating synthetic data $d_{\text{gen}} = \sum_i \delta l^{(i)} e_i$.

4. Given inputs for Sobolev space, s and λ, N_{geo} kernels/representations for the linear functional $c_1 \mapsto \delta l(c_1)$ were constructed through their Fourier coefficients.

5. Using only the informtion from synthetic data d_{gen}, the LBFGS gradient-based optimisation scheme was implemented given some initial model guess c_{start} to find $c_{1,\min}$ that minimised the misfit.

For simplicity, we chose the background velocity value $c_0 = 1$. The perturbation amplitude was chosen to be $c_{1,\max} = 0.05$, with perturbation resolution $N_{\text{pert}} = 2^3$, leading to a length scale of perturbation $L_{\text{hetero}} = 2\pi/N_{\text{pert}} = 0.785$. The velocity perturbation parametrisation was discretised into lattices of linear number $N_{\text{lat}} = 2^5$. Powers of two were chosen as natural number for N_{lat} as the Fast Fourier Transform algorithm worked by iteratively seperating the coefficents into two [19]. The pseudorandom numbers were generated using a seed. The Sobolev number was chosen to be $s = 2 > 2/2 + 0$, to admit continuous solutions. An example of a pseudorandomly generated velocity perturbation and geodesics, used as the generated model for the discussion below, with $N_{\text{geo}} = 100$, seed $= 2342234$ is shown on Fig 6.

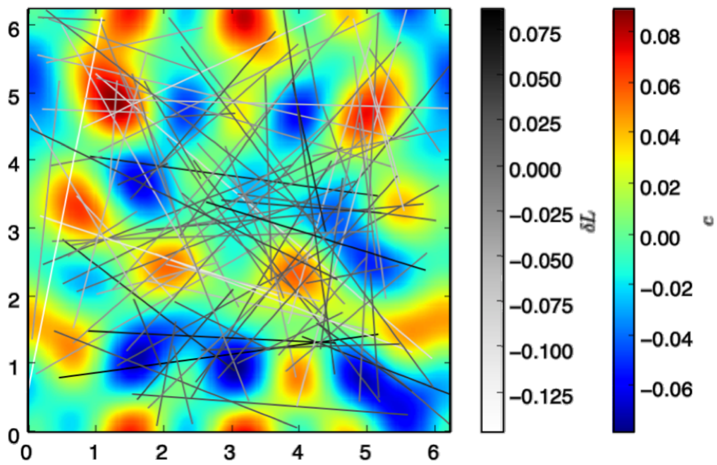


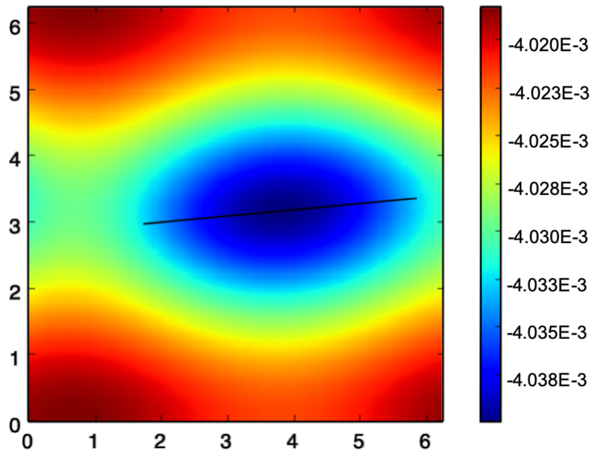
Figure 6: Pseudorandomly generated velocity field, with its value indicated in color, with generated geodesics with the pertrubed lengths in greyscale. They are generated with length scale of heterogeneity determined by N_{pert}.

4.4 Discussion

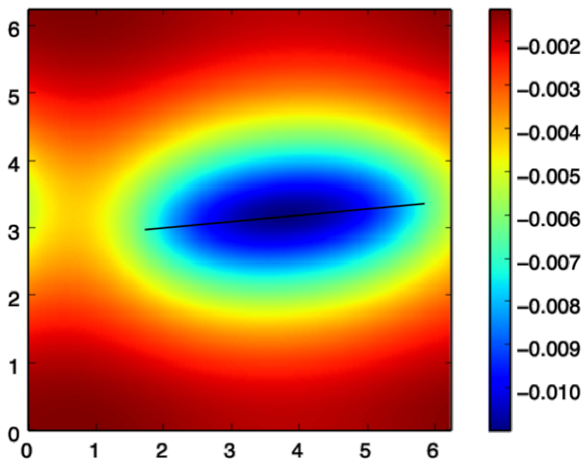
A keen reader may point out that by discretising c_1, we have essentially finitely-parameterised it. However, since we are merely sampling an underlying continuous function, for a real inverse problem, given a desired spatial resolution Δx, we could choose discretisation N_{lat} of c_1 up to a point where further increase in N_{lat} would not affect the kernels up to the resolution Δx. Since we would be working in an appropriate Sobolev space, the model solutions will converge as we increase lattice points. This process does not impose any physical properties of the solution by choosing a lattice N_{lat}.

We have yet discussed the role of characteristic length scale λ. Since the final solution is built up from a superposition of kernels, we can set and interpret λ as the expected length scale of heterogeneity/variation in the metric. We note that as λ tends to zero, the Sobolev inner product (3.26) tends to the ℓ^2 (equivalently L^2) inner product, (3.25). Therefore we expect that as λ decreases kernels should become more sharply peaked, and therefore expect to require greater N_{lat} to converge to satisfactory discrete representation. At $\lambda = 0$, equivalent to working in L^2, the kernel is proportional to δ_γ as (3.30), and would not converge even if infinitely many lattice points were added. These ideas are illustrated in Fig. 9. Representative kernels at different characteristic length scales, $\lambda_{\text{big}} = 0.99 \times 2\pi$, $\lambda_{\text{opt}} = L_{\text{hetero}} = 0.785$, $\lambda_{\text{small}} = 0.001 \times 2\pi$ are plotted on Fig. 7, which agrees with the discussion above. The optimal models for λ_{big}, λ_{opt} and λ_{small} are presented in Fig. 8. Comparing to the generated model Fig. 6, optimal model for $\lambda_{\text{small}} = 0.01 \times 2\pi$ fails to recreate the generated model as the kernels have support only in a small neighbourhood around the geodesic paths. At $\lambda = \lambda_{\text{opt}} = L_{\text{hetero}}$, we find a good match with the generated model, and when it is increased further $\lambda = \lambda_{\text{big}} = 0.99 \times 2\pi$, we find that the shorter-wavelength variation of the generated model are not captured.

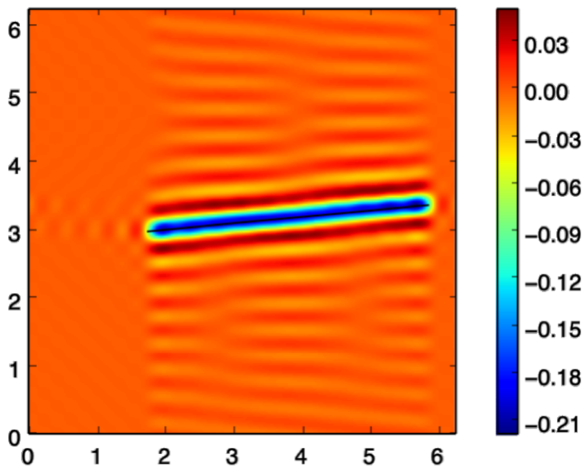
Using the methods outlined in Sec. 3.2, the eigendirections of the second order gradient, $A^\dagger A$ was found for characteristic length scale λ_{opt}. In Figs. 10a, 10b, we plot the the eigendirection with the highest eigenvalue$= 38.59$ (most constrained, or least error) and the lowest eigenvalue$= 0.19$ (least constrained, or most error). As expected, we can see that the center region with high density of geodesics is well constrained, whereas regions with only a few geodesics are not well constrained. Any variation that is orthogonal to all the kernels under the Sobolev inner product would lie in the null space of A and would be completely unconstrained by the data. Using a Gram-Schmidt process on the set of kernels $\{R^{(i)}\}_i$, a set of orthonormal basis functions of $\mathrm{im}(A^\dagger)$, $\{O_i\}_i$, can be found. Then, the projection of any c_1 on this subspace, $\mathrm{proj}_{\mathrm{im}(A^\dagger)}(c_1) = \sum_i (c_1, O_i)_{H^s_\lambda} O_i$, can be subtracted from c_1 to find the null-space component of c_1, $\mathrm{proj}_{\ker(A)}(c_1) = c_1 - \mathrm{proj}_{\mathrm{im}(A^\dagger)}(c_1)$. This was done for a constant perturbation of $c_{1,\mathrm{cst}}(\boldsymbol{x}) = 1$, and is shown in Fig. 10c. As expected, the projection is non-zero in regions with no geodesics, which is mostly edges of the coordinates.



(a) $\lambda_{\text{big}} = 0.99 \times 2\pi$

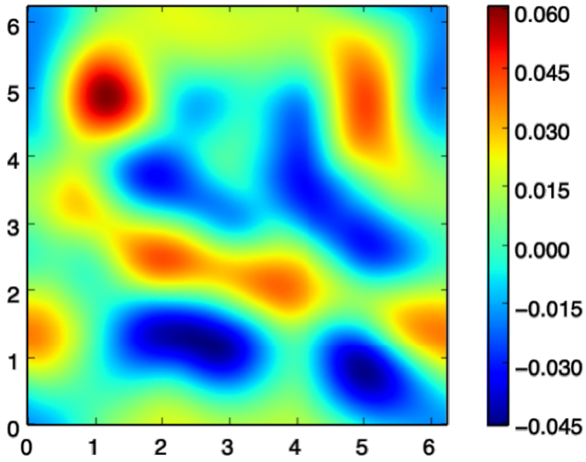


(b) $\lambda = \lambda_{\text{opt}} = L_{\text{hetero}} = 0.785$

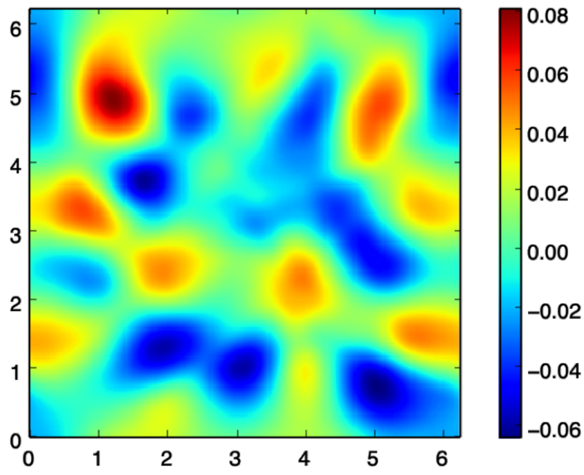


(c) $\lambda = \lambda_{\text{small}} = 0.001 \times 2\pi$

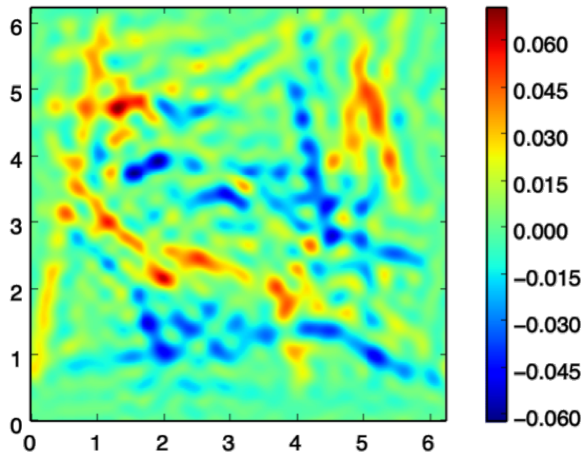
Figure 7: Sample kernels (color) superposed with the corresponding geodesic (black line) for various characteristic length scales. In particular, the ringing artifacts of low characteristic lengths are due to aliasing of the higher Fourier components, and is indicative that the kernel has not converged - higher N_{lat} is needed to remove the effects.



(a) $\lambda = \lambda_{\text{big}} = 0.99 \times 2\pi$

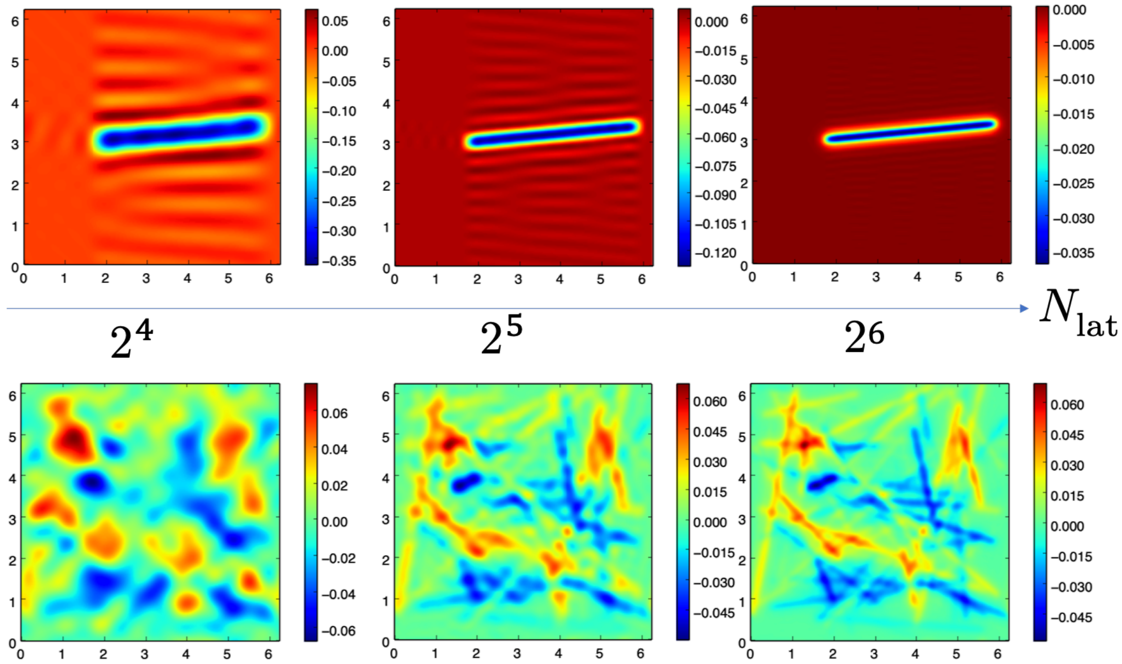


(b) $\lambda = \lambda_{\text{opt}} = L_{\text{hetero}} = 0.785$

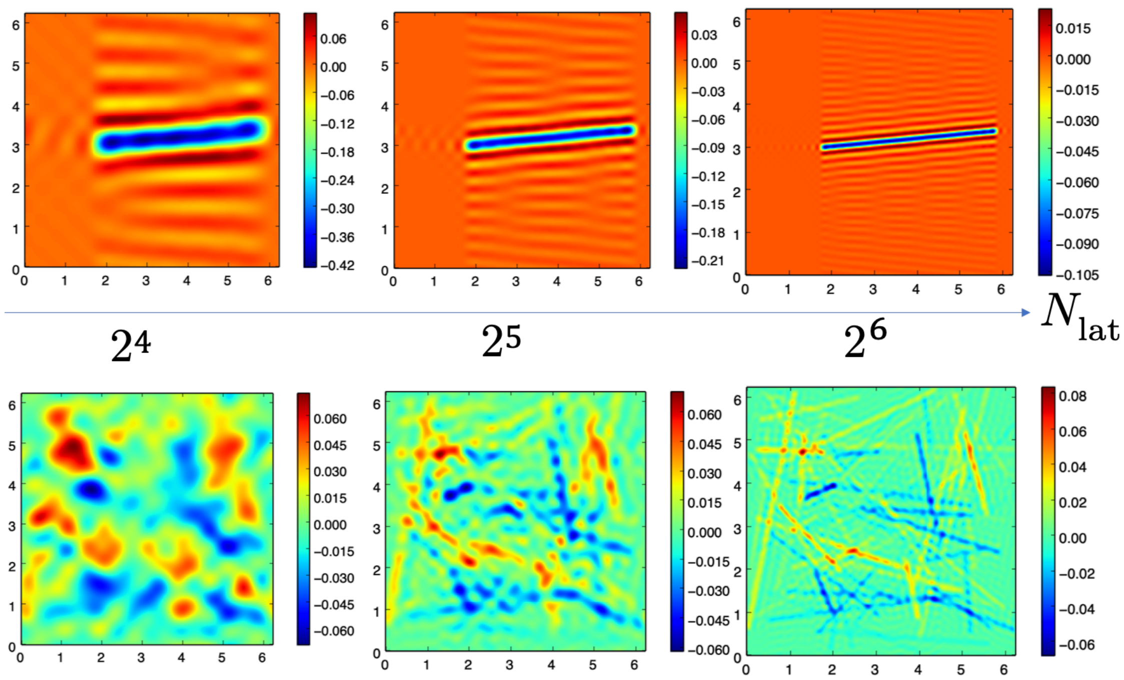


(c) $\lambda = \lambda_{\text{small}} = 0.001 \times 2\pi$

Figure 8: Optimal models for various characteristic length scales. Comparing to Fig. 6, we see that λ_{opt} provides the best match to the generated model compared to the other extremes. This is a result of the fact that the solutions are constructed as a superposition of kernels.

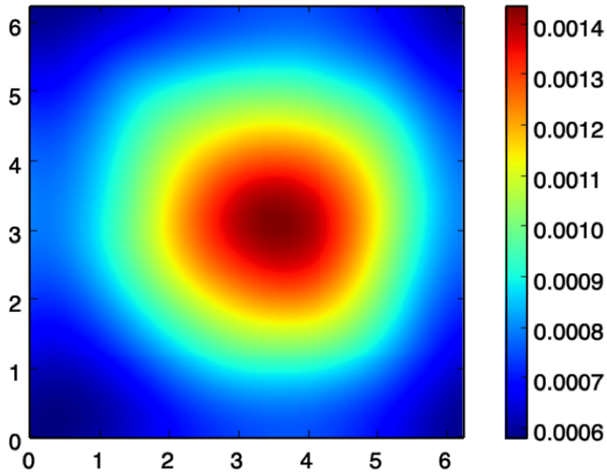


(a) $\lambda = \lambda_{\mathrm{smallish}} = 0.01 \times 2\pi$

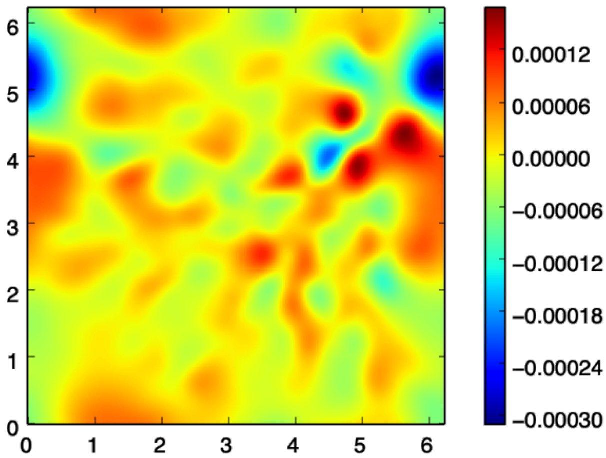


(b) $\lambda = \lambda_{\mathrm{zero}} = 0$

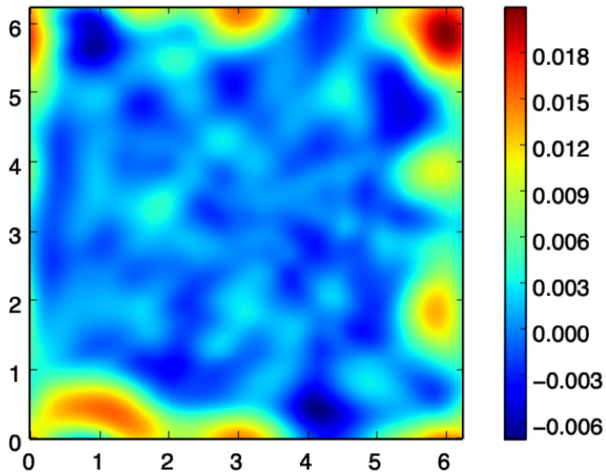
Figure 9: Figures illustrating the convergence properties of kernels and optimal model for $\lambda = \lambda_{\mathrm{smallish}} = 0.01 \times 2\pi$ *vs.* $\lambda = \lambda_{\mathrm{zero}} = 0$ *(latter equivalent to working in* L^2*). We note that for small* $N_{\mathrm{lat}} = 2^4$*, the kernels for* $\lambda_{\mathrm{smallish}}, \lambda_{\mathrm{zero}}$ *look similar and so do their optimal models. However, as* N_{lat} *increases,* $\lambda_{\mathrm{smallish}}$ *case converges and aliasing decreases, while for* λ_{zero} *case, it does not, as indicated by the increasing scale of the graphs. This clearly indicates that working in an* L^2 *space is inappropriate for the metric tomography problem.*



(a) Most constrained direction, with eigenvalue= 38.59*, corresponding to least error.*



(b) Least constrained direction, with eigenvalue= 0.19*, corresponding to least error.*



(c) Projection of $c_1(\boldsymbol{x}) = 1$ *onto* $\ker(A)$.

Figure 10: These figures illustrate the constraint of the optimised model. (a), (b) are eigendirections of the second-order gradient of the misfit, whose reciprocal eigenvalue is proportional to the error of the model along that direction, see (3.14)*. They are members of* $\mathrm{im}(A^\dagger)$ *- optimal model is constrained in these directions by the curvature of the misfit functional, to varying degrees. (c) is a member of* $\ker(A)$*, which means that the optimal model is completely unconstrained in this direction.*

5 Non-linear metric tomography

We now look at the non-linear version of the problem, where we do not assume that the deviation from some initial metric $\boldsymbol{g}_{\text{start}}$ is small, such that the path of the geodesics now vary as we vary the metric guess. Our data is a collection of points $\{(\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_1^{(i)})\}_{i=1}^{N_{geo}}$ and lengths of the geodesics $\{l^{(i)}\}_i$, from which we wish to find an unknown metric $\boldsymbol{g}(\boldsymbol{x})$. The EoM also requires us to specify $\{\boldsymbol{p}_0^{(i)}\}_i$ such that the path $\boldsymbol{x}^{(i)}(t)$ is a geodesic satisfying $\boldsymbol{x}^{(i)}(0) = \boldsymbol{x}_0^{(i)}$ and $\boldsymbol{x}^{(i)}(1) = \boldsymbol{x}_1^{(i)}$ for all $i \in \{1, ..., N_{geo}\}$.

We note that this case is an extension of the geodesic shooting problem; the optimisation problem must solve for both the starting momenta $\{\boldsymbol{p}_0^{(i)}\}_i$ and the metric $\boldsymbol{g}$. This is an example of combined source and structure inversion, which has shown to reduce errors compared to structure inversion from a pre-determined source parameters from a pre-existing model [11]. In our case, a separate source and structure inversion would correspond to pre-determining $\{\boldsymbol{p}_0\}$ on a preexisting model, and optimising only for the metric.

Noting the result (2.1) that $l = \|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}$, we can write down the Lagrangian for the adjoint method

$$\begin{aligned} L[m,u,u^\dagger] = {} & \overbrace{\underbrace{\frac{\lambda_1}{2}\sum_i \left\| l^{(i)} - \left\|\boldsymbol{p}_0^{(i)}\right\|_{\mathcal{T}^*\mathcal{M}} \right\|_{\mathbb{R}}^2}_{\text{a}} + \underbrace{\frac{\lambda_2}{2}\sum_i \left\| \boldsymbol{\pi}_x \boldsymbol{z}^{(i)}(1) - \boldsymbol{x}_1^{(i)} \right\|_{\mathbb{R}^n}^2}_{\text{b}}}^{J\left[\boldsymbol{g}, \{\boldsymbol{p}_0^{(i)}, \boldsymbol{z}^{(i)}\}_i\right]} \\ & + \underbrace{\sum_i \int_0^1 \left\langle \frac{d\boldsymbol{z}^{(i)}}{dt} - \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}\left(\boldsymbol{z}^{(i)}\right), \mathcal{P}\boldsymbol{z}^{\dagger(i)} \right\rangle dt}_{\text{c}} + \underbrace{\sum_i \left\langle \boldsymbol{z}^{(i)}(0) - \boldsymbol{z}_0^{(i)}, \boldsymbol{z}_0^{\dagger(i)} \right\rangle}_{\text{d}}, \end{aligned} \tag{5.1}$$

where $\{\boldsymbol{z}^{(i)}\}_i$ are the forward variables, $\{\boldsymbol{z}^{\dagger(i)}\}_i$ are the forward adjoint fields, $\{\boldsymbol{z}_0^{(i)}\}_i$ are the forward variable initial conditions (of which we only vary the starting momenta $\{\boldsymbol{p}_0^{(i)}\}_i$), $\{\boldsymbol{z}_0^{\dagger(i)}\}_i$ are the initial value adjoint variables, and $\boldsymbol{g}$ is the metric tensor. We clarify that the model parameters are

$$m = \left(\left\{\boldsymbol{p}_0^{(i)}\right\}_i, \boldsymbol{g}\right),$$

the forward variables are

$$u = \left\{\boldsymbol{z}^{(i)}\right\}_i,$$

the adjoint variables are

$$u^\dagger = \left\{\boldsymbol{z}^{\dagger(i)}, \boldsymbol{z}_0^{\dagger(i)}\right\}_i,$$

and the data, unaltered during optimisation, are

$$d = \left\{\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_1^{(i)}, l^{(i)}\right\}_i,$$

λ_1, λ_2 are weighting factors to prioritise between the two misfits, and $\mathcal{P} : f(t) \mapsto f(1-t)$ is the time reversal operator.

a is a new misfit term on the length of the geodesic, term b is the terminal value misfit as shown in the geodesic shooting problem. Together, Terms a, b define the objective functional that we want to minimise, $J\left[\boldsymbol{g}, \{\boldsymbol{p}_0^{(i)}, \boldsymbol{z}^{(i)}\}_i\right]$, under the constraints that the forward variables obey their EoMs and initial values, imposed by c, d respectively. A keen reader may point out that misfit term a is redundant, as we could parameterise $\{\boldsymbol{p}_0^{(i)}\}_i$ with $n-1$ variables, since its norm is known. However, this form is easier to manipulate in the case that there are data errors for $\{l^{(i)}\}_i$.

5.1 First Order Adjoint Calculations

For readability, we suppress the geodesic index (i), with sums $\sum_i$ to indicate whether it is summed over the geodesics. All equations hold for any i. From LMT,

$$D_m \hat{J} = D_m L \quad \leftrightarrow \quad D_{\boldsymbol{p}_0} \hat{J} = D_{\boldsymbol{p}_0} L, \quad D_{\boldsymbol{g}} \hat{J} = D_{\boldsymbol{g}} L,$$

given the forward and adjoint variables obey the conditions

$$D_{\boldsymbol{z}} L = 0, \quad D_{\boldsymbol{z}^\dagger} L = 0, \quad D_{\boldsymbol{z}_0^\dagger} L = 0. \tag{5.2}$$

We now present the expressions for these functional derivatives. For detailed techniques and derivation, see appendix A.2.

By varying the first order adjoint variables in (5.1), we get

$$\left\langle D_{\boldsymbol{z}^\dagger} L, \delta \boldsymbol{z}^\dagger \right\rangle = \int_0^1 \left\langle \frac{d\boldsymbol{z}}{dt}(t) - \mathbb{J}\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}(t)), \left(\mathcal{P}\delta \boldsymbol{z}^\dagger\right)(t) \right\rangle dt, \tag{5.3}$$

$$\left\langle D_{\boldsymbol{z}_0^\dagger} L, \delta \boldsymbol{z}_0^\dagger \right\rangle = \left\langle \boldsymbol{z}(0) - \boldsymbol{z}_0, \delta \boldsymbol{z}_0^\dagger \right\rangle, \tag{5.4}$$

which, together with the conditions (5.2), these must hold for any $\delta \boldsymbol{z}^\dagger$, $\delta \boldsymbol{z}_0^\dagger$, we retrieve the forward problem and its initial conditions.

By varying the Lagrangian (5.1) with respect to the forward variable, we find that

$$\left\langle D_{\boldsymbol{z}} L, \delta \boldsymbol{z} \right\rangle = \left\langle \lambda_2 \boldsymbol{\pi}_x^* \left(\boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1\right) + \boldsymbol{z}^\dagger(0), \delta \boldsymbol{z}(1) \right\rangle \tag{5.5}$$

$$+ \left\langle \boldsymbol{z}_0^\dagger - \boldsymbol{z}^\dagger(1), \delta \boldsymbol{z}(0) \right\rangle \tag{5.6}$$

$$+ \int_0^1 \left\langle \frac{d\boldsymbol{z}^\dagger}{dt}(t) + \left\{ \mathcal{P} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{z}(t)) \mathbb{J} \mathcal{P} \right\} \boldsymbol{z}^\dagger(t), \left(\mathcal{P}\delta \boldsymbol{z}\right)(t) \right\rangle dt. \tag{5.7}$$

Thus, we obtain the adjoint equations together with its initial/terminal value conditions. Finally, we vary the Lagrangian with respect to the model parameters to find

$$\left\langle D_{\boldsymbol{p}_0} L, \delta \boldsymbol{p}_0 \right\rangle = \left\langle -\left(\lambda_1 \left[\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right] \boldsymbol{p}_0 \boldsymbol{g}^{-1}(\boldsymbol{x}_0)\right) + \boldsymbol{\pi}_p \boldsymbol{z}_0^\dagger \ , \ \delta \boldsymbol{p}_0 \right\rangle, \tag{5.8}$$

and

$$\begin{aligned} \left\langle \left(D_{\boldsymbol{g}} L\right)_{\mu\nu}, \delta g_{\mu\nu} \right\rangle = & \sum_i \left\langle \left[\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right] \left[\boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}_0\right]^\mu \left[\boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}_0\right]^\nu \ , \ \delta g_{\mu\nu}(\boldsymbol{x}_0) \right\rangle \\ & + \sum_i \int_0^1 \left\langle -\left[\left(\mathcal{P}\boldsymbol{z}^\dagger\right)(t)\right]^T \mathbb{J} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial g_{\mu\nu}}\left(\boldsymbol{z}(t)\right) \ , \ \delta g_{\mu\nu}\left(\boldsymbol{x}(t)\right) \right\rangle dt. \end{aligned} \tag{5.9}$$

As shown in Sec. 3.5.4 and implemented in the linear case in Sec. 4.4, For each iteration, (5.9) would be rewritten as an inner product between a kernel and and the metric perturbation in Sobolev space, so that the components of the functional derivatives can be added to the metric iteratively.

The optimisation scheme would be as follows.

1. We make a starting guess of $\{\boldsymbol{p}_0\}$ and $\boldsymbol{g}$.
2. We choose the initial value of forward variables according to (5.4), and solve the forward problem according to Eq (5.3), to obtain $\{\boldsymbol{z}(t)\}$ and also the terminal values $\{\boldsymbol{z}(1)\}$.

3. We choose initial value of the first order adjoint variables, $\{\boldsymbol{z}^\dagger(0)\}$, according to (5.5). We solve the adjoint problem (5.7) to find $\{\boldsymbol{z}^\dagger(1)\}$ and therefore $\{\boldsymbol{z}_0^\dagger\}$ by (5.6).
4. All the forward and adjoint variables are now solved. We can calculate the gradient with respect to the model parameters by substitution into (5.8) and (5.9). In particular, the gradient term for $\boldsymbol{g}$ should be recast as an inner product of $\delta\boldsymbol{g}$ with some representative/kernel R, in an appropriate Sobolev space.
5. We use some gradient descent method to update the model parameters and iterate until local minima is found, i.e. $D_m\hat{J} = 0$ to some tolerance.

The order in which variables are specified is

$$(\{\boldsymbol{p}_0\}, \boldsymbol{g}) \to \{\boldsymbol{z}\} \to \{\boldsymbol{z}^\dagger\} \to \{\boldsymbol{z}_0^\dagger\} \to \left(\{D_{\boldsymbol{p}_0}\hat{J}\}, D_{\boldsymbol{g}}\hat{J}\right).$$

5.2 Second Order Adjoint Calculations

As discussed in Sec. 3.2, the second order gradients can be used to quantify the the constraint on the solution of the model, or to potentially improve the optimisation routine. Motivated by this, we present the second order adjoints to obtain expressions of the second order gradients. See appendix A.3 for details.

As described in Sec. 3.3, we define the new objective functional given some model direction $m^\sharp$,

$$J_1[u, u^\dagger m \mid m^\sharp] := \left\langle D_m L, m^\sharp \right\rangle = \sum_i \left\langle D_{\boldsymbol{p}_0} L, \boldsymbol{p}_0^\sharp \right\rangle + \left\langle D_{\boldsymbol{g}} L, \boldsymbol{g}^\sharp \right\rangle. \tag{5.10}$$

To constrain the forward variables u and the adjoint variables $u^\dagger$ to their respective EoMs, we introduce the second order adjoint variables, $u^\sharp, u^{\dagger\sharp}$ as Lagrange multipliers, to write the second order Lagrangian as

$$\begin{aligned} L_1\left[m, u, u^\dagger, u^\sharp, u^{\dagger\sharp}\right] = & \underbrace{\sum_i \left\langle D_{\boldsymbol{p}_0} L, \boldsymbol{p}_0^\sharp \right\rangle}_{\alpha} + \underbrace{\left\langle D_{\boldsymbol{g}} L, \boldsymbol{g}^\sharp \right\rangle}_{\beta} \\ & + \underbrace{\sum_i \left\langle D_{\boldsymbol{z}} L, \boldsymbol{z}^\sharp \right\rangle}_{\gamma} + \underbrace{\sum_i \left\langle D_{\boldsymbol{z}^\dagger} L, \boldsymbol{z}^{\dagger\sharp} \right\rangle}_{\delta} + \underbrace{\sum_i \left\langle D_{\boldsymbol{z}_0^\dagger} L, \boldsymbol{z}_0^{\dagger\sharp} \right\rangle}_{\epsilon}, \end{aligned} \tag{5.11}$$

where model parameters $m = (\{\boldsymbol{p}_0\}, \boldsymbol{g})$, forward variables $u = \{\boldsymbol{z}\}$, first order adjoint variables $u^\dagger = \{\boldsymbol{z}^\dagger, \boldsymbol{z}_0^\dagger\}$, and the second order adjoint variables are $u^\sharp = \{\boldsymbol{z}^\sharp\}$ and $u^{\dagger\sharp} = \{\boldsymbol{z}^{\dagger\sharp}, \boldsymbol{z}_0^{\dagger\sharp}\}$, and

$$\alpha = -\sum_i \left\langle \lambda_1 \left[\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right] \boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}_0 + \boldsymbol{\pi}_p \boldsymbol{z}_0^\dagger \ , \ \boldsymbol{p}_0^\sharp \right\rangle, \tag{5.12}$$

$$\beta = \sum_i \left\langle \left[\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right] \left[\boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}_0\right]^\mu \left[\boldsymbol{g}^{-1}(\boldsymbol{x}_0)\boldsymbol{p}_0\right]^\nu \ , \ \boldsymbol{g}^\sharp(\boldsymbol{x}_0)_{\mu\nu} \right\rangle \tag{5.13}$$

$$-\sum_i \int_0^1 \left\langle \left[\left(\mathcal{P}\boldsymbol{z}^\dagger\right)(t)\right]^T \mathbb{J} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial g_{\mu\nu}} (\boldsymbol{z}(t)) \ , \ g^\sharp(\boldsymbol{x}(t)) \right\rangle dt, \tag{5.14}$$

$$\gamma = \sum_i \left\langle \lambda_2 \boldsymbol{\pi}_x^* \left(\boldsymbol{\pi}_x \boldsymbol{z}(1) - \boldsymbol{x}_1\right) + \boldsymbol{z}^\dagger(0), \boldsymbol{z}^\sharp(1) \right\rangle \tag{5.15}$$

$$+ \sum_i \left\langle \boldsymbol{z}_0^\dagger - \boldsymbol{z}^\dagger(1), \boldsymbol{z}^\sharp(0) \right\rangle \tag{5.16}$$

$$+ \sum_i \int_0^1 \left\langle \frac{d\boldsymbol{z}^\dagger}{dt}(t) + \left\{ \mathcal{P} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{z}(t)) \mathbb{J} \mathcal{P} \right\} \boldsymbol{z}^\dagger(t), \left(\mathcal{P} \boldsymbol{z}^\sharp\right)(t) \right\rangle dt, \tag{5.17}$$

$$\delta = \sum_i \int_0^1 \left\langle \frac{d\boldsymbol{z}}{dt}(t) - \mathbb{J} \frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}(t)), \left(\mathcal{P} \boldsymbol{z}^{\dagger\sharp}\right)(t) \right\rangle dt, \tag{5.18}$$

$$\epsilon = \sum_i \left\langle \boldsymbol{z}(0) - \boldsymbol{z}_0, \boldsymbol{z}_0^{\dagger\sharp} \right\rangle. \tag{5.19}$$

Then,

$$D_m \hat{J}_1 = \left\langle D_m D_m \hat{J}, m^\sharp \right\rangle = D_m L_1,$$

given the forward and 1st, 2nd adjoint variables obey the conditions

$$D_{\boldsymbol{z}} L_1 = 0, \quad D_{\boldsymbol{z}^\dagger} L_1 = 0, \quad D_{\boldsymbol{z}_0^\dagger} L_1 = 0, \quad D_{\boldsymbol{z}^\sharp} L_1 = 0, \quad D_{\boldsymbol{z}^{\dagger\sharp}} L_1 = 0, \quad D_{\boldsymbol{z}_0^{\dagger\sharp}} L_1 = 0. \tag{5.20}$$

By inspection, from conditions $D_{\boldsymbol{z}^\sharp} L_1 = 0$, $D_{\boldsymbol{z}^{\dagger\sharp}} L_1 = 0$, $D_{\boldsymbol{z}_0^{\dagger\sharp}} L_1 = 0$, we simply retrieve the forward EoM (5.3), the initial value condition (5.4) and the adjoint equations (5.5), (5.6), (5.7) respectively.

Varying L_1 with respect to $\boldsymbol{z}^\dagger$, $\boldsymbol{z}_0^\dagger$, we find that

$$\left\langle D_{\boldsymbol{z}^\dagger} L_1, \delta \boldsymbol{z}^\dagger \right\rangle = \int_0^1 \left\langle \frac{d\boldsymbol{z}^\sharp}{dt} - \mathbb{J} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{z}) \boldsymbol{z}^\sharp + \mathbb{J} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial g_{\mu\nu}}(\boldsymbol{z}) g_{\mu\nu}^\sharp, \mathcal{P} \delta \boldsymbol{z}^\dagger(t) \right\rangle dt, \tag{5.21}$$

$$\left\langle D_{\boldsymbol{z}_0^\dagger} L_1, \delta \boldsymbol{z}_0^\dagger \right\rangle = \left\langle \boldsymbol{z}^\sharp(0) - \boldsymbol{\pi}_p^* \boldsymbol{p}_0^\sharp, \delta \boldsymbol{z}_0^\dagger \right\rangle. \tag{5.22}$$

The EoM for variable $\boldsymbol{z}^\sharp$ has a source term dependent on $\boldsymbol{g}^\sharp$, with an initial condition dependent on $\boldsymbol{p}_0^\sharp$.

Next, varying L_1 with respect to $\boldsymbol{z}$, we find that

$$\left\langle D_{\boldsymbol{z}} L_1, \delta \boldsymbol{z} \right\rangle = \int_0^1 \left\langle \frac{d\boldsymbol{z}^{\dagger\sharp}}{dt} + \left\{ \mathcal{P} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{z}) \mathbb{J} \mathcal{P} \right\} \boldsymbol{z}^{\dagger\sharp} + \mathcal{P} g_{\mu\nu}^\sharp \frac{\partial^3 H}{\partial \boldsymbol{z} \partial g_{\mu\nu} \partial \boldsymbol{z}}(\boldsymbol{z}) \mathbb{J} \mathcal{P} \boldsymbol{z}^\dagger + \mathcal{P} \boldsymbol{z}^{\sharp T} \frac{\partial^3 H}{\partial \boldsymbol{z} \partial \boldsymbol{z} \partial \boldsymbol{z}} \mathbb{J} \mathcal{P} \boldsymbol{z}^\dagger, \mathcal{P} \delta \boldsymbol{z}^\dagger \right\rangle dt \tag{5.23}$$

$$+ \left\langle \boldsymbol{z}_0^{\dagger\sharp} - \boldsymbol{z}^{\dagger\sharp}(1), \delta \boldsymbol{z}(0) \right\rangle \tag{5.24}$$

$$+ \left\langle \boldsymbol{z}^{\dagger\sharp}(0) + \lambda_2 \boldsymbol{\pi}_x^* \boldsymbol{\pi}_x \boldsymbol{z}^\sharp(0), \delta \boldsymbol{z}(1) \right\rangle, \tag{5.25}$$

where we understand that the terms like

$$\boldsymbol{a}^T \frac{\partial^3 H}{\partial \boldsymbol{z} \partial \boldsymbol{z} \partial \boldsymbol{z}} \boldsymbol{b}$$

contracts the first and last components of $\partial^3 H / \partial \boldsymbol{z} \partial \boldsymbol{z} \partial \boldsymbol{z}$, and leaves the middle component free. The equations of motion for $\boldsymbol{z}^{\dagger\sharp}$ contains source terms from both $\boldsymbol{g}^\sharp$ and $\boldsymbol{z}^\sharp$.

Varying L_1 with respect to $\boldsymbol{p}_0$ we find

$$\begin{aligned}\left\langle D_{\boldsymbol{p}_0} L_1, \delta \boldsymbol{p}_0 \right\rangle = \Bigg\langle & \lambda_1 \left(\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) \left(\boldsymbol{g}^{-1}\boldsymbol{g}^{\sharp}\boldsymbol{g}^{-1}\boldsymbol{p}_0 - \boldsymbol{g}^{-1}\boldsymbol{p}_0^{\sharp} \right) \\ & + \frac{\lambda_1 l}{\|\boldsymbol{p}_0\|^2_{\mathcal{T}^*\mathcal{M}}} \left(2\boldsymbol{p}_0^T \boldsymbol{g}^{-1}\boldsymbol{p}_0^{\sharp} - \boldsymbol{p}_0^T\boldsymbol{g}^{-1}\boldsymbol{g}^{\sharp}\boldsymbol{g}^{-1}\boldsymbol{p}_0 \right) \boldsymbol{g}^{-1}\boldsymbol{p}_0 - \boldsymbol{\pi}_p \boldsymbol{z}_0^{\dagger\sharp},\ \delta\boldsymbol{p}_0 \Bigg\rangle,\end{aligned} \tag{5.26}$$

which requires the value $\boldsymbol{z}_0^{\dagger\sharp}$ to be calculated.

Varying L_1 with respect to $\boldsymbol{g}$, we find

$$\begin{aligned}\left\langle D_{\boldsymbol{g}} L_1, \delta \boldsymbol{g} \right\rangle = & \sum_i \int_0^1 \left\langle \boldsymbol{z}^{\sharp T} \frac{\partial^3 H}{\partial \boldsymbol{z} \partial g_{\mu\nu} \partial \boldsymbol{z}}(z) \mathbb{J} \left(\mathcal{P}\boldsymbol{z}^{\dagger} \right) - \left(\mathcal{P}\boldsymbol{z}^{\dagger} \right)^T \mathbb{J} \frac{\partial^3 H}{\partial \boldsymbol{z} \partial g_{\rho\sigma} \partial g_{\mu\nu}} g^{\sharp}_{\rho\sigma} - \left(\mathcal{P}\boldsymbol{z}^{\dagger\sharp} \right)^T \mathbb{J} \frac{\partial^2 H}{\partial \boldsymbol{z} \partial g_{\mu\nu}}, \delta g_{\mu\nu} \right\rangle dt \\ & + \lambda_1 \sum_i \Bigg\langle \left(\frac{l}{\|\boldsymbol{p}_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) [\boldsymbol{g}^{-1}\boldsymbol{p}_0]^{\mu} [\boldsymbol{g}^{\sharp}\boldsymbol{p}_0 + \boldsymbol{g}^{-1}\boldsymbol{p}_0^{\sharp}]^{\nu} \\ & + \frac{l}{2\|\boldsymbol{p}_0\|^3_{\mathcal{T}^*\mathcal{M}}} \left(\frac{1}{2}\boldsymbol{p}_0^T\boldsymbol{g}^{-1}\boldsymbol{g}^{\sharp}\boldsymbol{g}^{-1}\boldsymbol{p}_0 - \boldsymbol{p}_0^T\boldsymbol{g}^{-1}\boldsymbol{p}_0^{\sharp} \right) [\boldsymbol{g}^{-1}\boldsymbol{p}_0]^{\mu} [\boldsymbol{g}^{-1}\boldsymbol{p}_0]^{\nu},\ \delta g_{\mu\nu} \Bigg\rangle.\end{aligned} \tag{5.27}$$

To calculate this quantity, $\boldsymbol{z}^{\sharp}$, $\boldsymbol{z}^{\dagger\sharp}$ must be known.

To calculate the second order derivative, the process is as follows.

1. Find $\{\boldsymbol{z}, \boldsymbol{z}^{\dagger}, \boldsymbol{z}_0^{\dagger}\}$ from the forward equations of motion ((5.3)), the initial value condition ((5.4)) and the adjoint equations (Eqs. (5.5), (5.6), (5.7)).
2. From Eqs. (5.21), (5.22), calculate $\{\boldsymbol{z}^{\sharp}\}$.
3. With $\{\boldsymbol{z}^{\sharp}\}$ found, the initial values $\{\boldsymbol{z}^{\dagger\sharp}(0)\}$ and the source terms for EoM of $\{\boldsymbol{z}^{\dagger\sharp}\}$ are specified, from (5.25) and (5.23) respectively. Solving these, we can find $\{\boldsymbol{z}_0^{\dagger\sharp}\}$ from (5.24).
4. All variables that fulfill (5.20) now specified. We can calculate calculate $D_m L_1 = D_m \hat{J}_1$ from Eqs. (5.26), (5.27).

Similar to the first-order adjoints, the order in which variables are specified is as follows:

$$(\{\boldsymbol{p}_0\}, \boldsymbol{g}) \to \{\boldsymbol{z}\} \to \{\boldsymbol{z}^{\dagger}\} \to \{\boldsymbol{z}_0^{\dagger}\} \to \{\boldsymbol{z}^{\sharp}\} \to \{\boldsymbol{z}^{\dagger\sharp}\} \to \{\boldsymbol{z}_0^{\dagger\sharp}\} \to \left(\left\{ \left\langle D_{\boldsymbol{p}_0} D_m \hat{J}, m^{\sharp} \right\rangle \right\}, \left\langle D_{\boldsymbol{g}} D_m \hat{J}, m^{\sharp} \right\rangle \right).$$

6 Outlook

In this article, we proposed that we can work in Sobolev spaces to calculate functional derivatives to be used in gradient-based optimisation, in the case that the kernels are not point-wise convergent posed in conventional function spaces such as L^2. We studied this idea on the metric tomographic problem, an analogue of ray-tracing of surface waves. We implemented the proposed optimisation scheme for the linear metric tomography problem on a $2d$ torus. Furthermore, we presented the first and second order adjoints for non-linear metric tomography on a manifold, where we find that the second order adjoint equations have source terms in their EoM. The first order adjoint equations can be used in gradient-based optimisation and the second order adjoint equations can be used in quantifying the constraint on the local minima of the output model.

The obvious next step would be to numerically implement gradient-based optimisation schemes for the non-linear tomographic problem, for other types of manifolds such as spheres or manifolds with boundaries, working in the appropriate Sobolev spaces. The general use of Sobolev gradients could be applied to any inverse tomographic problems such as full-waveform tomography where the kernel posed in L^2 is normally divergent. Furthermore, working in an appropriate Sobolev space, we could combine ray-tomography for high-frequency data, using the presented adjoints (Sec. 5), with full-waveform tomography for low-frequency data, into one optimisation. Such approached combining two data sources could improve images of the Earth's structure.

There are some open questions to consider for general non-linear inverse problems with a very large model parameter space, most importantly the question of how to find the global minima of a misfit function, in the case that exploring the full parameter space is impossible. The choice of length scale λ in the Sobolev space is also up for debate. We propose that it should be used for expected length scale of heterogeneity of the structure, but this fails to be of use when we have no a priori expectations of the structure.

Acknowledgements

I would like to thank my immensely helpful supervisor, David Al-Attar, for introducing me to the field of inverse problems. I would also like to thank Francis Syvret, Tal Gottesman, So Takao for helpful discussions.

References

[1] van der Hilst, R.D., Engdahl, E.R., Spakman, W., and Nolet, G. Tomographic imaging of subducted lithosphere below northwest Pacific island arcs. *Nature*, 353:3743, 1991. 5

[2] Kosovichev, A.G., Duvall, T.L., and Scherrer, P.H. Time-distance inversion methods and results. Sol. Phys., 192:159176, 2000. 7

[3] Boschi, Lapo, and Adam M. Dziewonski. "Highand lowresolution images of the Earth's mantle: Implications of different approaches to tomographic modeling." Journal of Geophysical Research: Solid Earth 104.B11 (1999): 25567-25594.

[4] Woodhouse, J. H. (1974). *Surface waves in a laterally varying layered structure*. Geophysical Journal International, 37(3), 461-490.

[5] Crawford O, Al-Attar D, Tromp J, Mitrovica JX. Forward and inverse modelling of post-seismic deformation. *Geophysical journal international*. 2016 Nov 3:ggw414.

[6] VI, Arnol'D. *Mathematical methods of classical mechanics.* Springer, 1989.

[7] Taylor, Michael E. "Partial Differential Equations I: Basic Theory (Applied Mathematical Sciences)." 2011.

[8] Yosida, K. (1980), *Functional Analysis, Grundlehren der Mathematischen Wissenschaften* (Fundamental Principles of Mathematical Sciences), vol. 123 (6th ed.), Berlin, New York: Springer-Verlag

[9] Fichtner, Andreas, and Jeannot Trampert. "Hessian kernels of seismic data functionals based upon adjoint techniques." Geophysical Journal International 185.2 (2011): 775-798.

[10] Saad, Yousef (2003). Iterative Methods for Sparse Linear Systems (2 ed.). SIAM. p. 414. ISBN 0898715342.

[11] Valentine, Andrew P., and John H. Woodhouse. "Reducing errors in seismic tomography: combined inversion for sources and structure." *Geophysical Journal International* 180.2 (2010): 847-857.

[12] French, Scott W., and Barbara Romanowicz. "Broad plumes rooted at the base of the Earth's mantle beneath major hotspots." Nature 525.7567 (2015): 95.

[13] Nolet, Guust. "A breviary of seismic tomography." A Breviary of Seismic Tomography, by Guust Nolet, Cambridge, UK: Cambridge University Press, 2008 (2008).

[14] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients", Comput. J. 7 (1964), 149154

[15] Liu, D. C.; Nocedal, J. (1989). "On the Limited Memory Method for Large Scale Optimization". Mathematical Programming B. 45 (3): 503528

[16] Aubin, Thierry. Nonlinear analysis on manifolds. Monge-Ampere equations. Vol. 252. Springer Science & Business Media, 1982.

[17] Neuberger, J. (2009). *Sobolev gradients and differential equations*. Springer Science & Business Media.

[18] Al-Attar, D. (2018). *Physics of the Earth as a Planet Lecture Notes*. Bullard Labs.

[19] Press, William H., et al. FORTRAN numerical recipes. Vol. 1. Cambridge England, 1996.

A Appendix

A.1 Wave Propagation in Elastic Media

The physics of wave-propagation is well understood for homogeneous media [18]. In linearised form, for homogeneous, simple, hyper-elastic media, the displacement field from equilibrium state, $\boldsymbol{u}(x)$ is given by

$$\boldsymbol{u}(\boldsymbol{x}) = f\left(t - \langle \boldsymbol{p}, \boldsymbol{x} \rangle\right) \boldsymbol{a},$$

where $\boldsymbol{x}$ is the coordinate in the reference body, t is time, $\boldsymbol{p}$ is the slowness vector, and $\boldsymbol{a}$ the polarisation vector, and we can understand $\langle \cdot\, , \cdot \rangle$ to be the Euclidean inner product, and f can be any well-behaved function, usually a wavepacket, with the condition

$$\Big(\boldsymbol{\Gamma}(\boldsymbol{p}) - \boldsymbol{1}\Big)\boldsymbol{a} = \boldsymbol{0},$$

where the $\boldsymbol{\Gamma}(\boldsymbol{p})$ is the Christoffel operator which encodes information about structure of the media. This can be solved through simple linear analysis; for a given $\boldsymbol{p}$, there is a solution if $\Gamma(\boldsymbol{p})$'s eigenvalues, $\lambda_k(\boldsymbol{p}) = 1$, and $\boldsymbol{a}$ is the eigenvector corresponding to the eigevnvalue $\lambda_k(\boldsymbol{p})$.

For heterogeneous, anisotropic bodies, for cases where the length scale of heterogeneity L_{hetero} is much greater than the length scale of the wavepacket, L_{packet}, $L_{\text{hetero}} \gg L_{\text{packet}}$, ray-tracing theory can be used to approximate wave propagation, where the solutions have the form

$$\boldsymbol{u}(\boldsymbol{x}, t) = \sum_{n=0}^{\infty} \left(\frac{1}{2\pi} \int_{\mathbb{R}} \frac{1}{(i\omega)^n} e^{i\omega(t - T(x))} d\omega \right) \boldsymbol{a}_n(x),$$

with leading term and one component of frequency given by

$$\boldsymbol{u}(x, t) \approx \boldsymbol{a}_0(x) e^{i\omega(t - T(x))},$$

where $T(x)$ is the travel time, and $\boldsymbol{a}_n$ is the n^{th} polarisation vector.

Defining the local slowness vector $\boldsymbol{p}(\boldsymbol{x}) := \frac{\partial T(\boldsymbol{x})}{\partial \boldsymbol{x}}$, the lowest order terms can be solved by the Eikonal equation:

$$\Big(\boldsymbol{\Gamma}(\boldsymbol{x}, \boldsymbol{p}) - \boldsymbol{1}\Big)\boldsymbol{a}_0(\boldsymbol{x}) = \boldsymbol{0},$$

where $\boldsymbol{\Gamma}(\boldsymbol{x}, \boldsymbol{p})$ is the local Christoffel operator.

This can be solved using the method of characteristics. We note that again, in order for there to be a solution, there must be some k such that $\lambda_k(x, p) = 1$. By continuity, all solutions must have this property. Defining the ray Hamiltonian H_k as

$$H_k(\boldsymbol{x}, \boldsymbol{p}) := \frac{1}{2}\lambda_k(\boldsymbol{x}, \boldsymbol{p}),$$

if H_k is written as a functional of two independent variables $\boldsymbol{x}, \boldsymbol{p}$,

$$H_k[\boldsymbol{x}, \boldsymbol{p}] = \frac{1}{2}.$$

Now we note that if we use Hamilton's canonical equations, H_k will be conserved, which means that $t_k = 1$. So, given some generating parameter t,

$$\frac{d\boldsymbol{x}}{dt} = \frac{\partial H_k}{\partial \boldsymbol{p}}; \quad \frac{d\boldsymbol{p}}{dt} = -\frac{\partial H_k}{\partial \boldsymbol{x}}.$$

Given some initial condition $\boldsymbol{p}_0, \boldsymbol{x}_0$ such that $\langle \boldsymbol{p}_0, \boldsymbol{x}_0 \rangle = t_0$, we can solve for the travel time $T(x)$, which turns out to be

$$\frac{d}{dt}T[\boldsymbol{x}(\sigma)] = 1 \implies T[\boldsymbol{x}(t)] = t_0 + t,$$

which suggests that the parameter t is simply the travel time itself.

A.2 Details of First Order Adjoint

For brevity, we drop the bolding of symbols and the geodesic indices (i). All equals signs are to first order. We stress that varying $z, z^\dagger, p_0, z_0^\dagger$ in an expression summing over geodesics $\sum_i$, picks out the element of the same geodesic index, thereby removing the summation. On the other hand varying g changes every element therefore does not remove the sum.

The following identities will be useful: the anticommutation relation,

$$\left\{ \frac{\partial}{\partial t}, \mathcal{P} \right\} = 0,$$

the invariance property,

$$\int_0^1 \langle \cdot, \cdot \rangle dt = \int_0^1 \langle \mathcal{P}\cdot, \mathcal{P}\cdot \rangle dt,$$

the involutory property,

$$\mathcal{P}^2 = 1,$$

and the transpose property of $\mathbb{J}$:

$$\mathbb{J}^T = -\mathbb{J}.$$

From (5.1), the first order Lagrangian is given by

$$L = \text{a} + \text{b} + \text{c} + \text{d},$$

where

$$\text{a} = \frac{\lambda_1}{2} \sum_i \left\| l - \|p_0\|_{\mathcal{T}^*\mathcal{M}} \right\|_{\mathbb{R}}^2,$$

$$\text{b} = \frac{\lambda_2}{2} \sum_i \left\| \pi_x z(1) - x_1 \right\|_{\mathbb{R}^n}^2,$$

$$\text{c} = \sum_i \int_0^1 \left\langle \frac{dz}{dt} - \mathbb{J}\frac{\partial H}{\partial z}(z), \mathcal{P}z^{\dagger(i)} \right\rangle dt,$$

$$\text{d} = \sum_i \left\langle z(0) - z_0, z_0^\dagger \right\rangle.$$

A.2.1 $D_{p_0} L$

The $\mathcal{T}^*\mathcal{M}$ norm of p_0 transforms, under the perturbation of p_0,

$$\begin{aligned}
\|p_0\|_{\mathcal{T}^*\mathcal{M}} = \sqrt{p_0^T g^{-1} p_0} \rightarrow \\
\|p_0 + \delta p_0\|_{\mathcal{T}^*\mathcal{M}} &= \sqrt{(p_0+\delta p_0)^T g^{-1}(p_0+\delta p_0)} = \sqrt{p_0^T g^{-1} p_0 + 2 p_0^T g^{-1}\delta p_0 + \mathcal{O}(2)} \\
&= \|p_0\|_{\mathcal{T}^*\mathcal{M}} \left(1 + \frac{2 p_0^T g^{-1}\delta p_0}{\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} + \mathcal{O}(2)\right)^{1/2} \\
&= \|p_0\|_{\mathcal{T}^*\mathcal{M}} + \frac{p_0^T g^{-1}\delta p_0}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} + \mathcal{O}(2).
\end{aligned}$$

a contains term

$$(l - \|p_0\|_{\mathcal{T}^*\mathcal{M}})^2 \rightarrow (l - \|p_0 + \delta p_0\|_{\mathcal{T}^*\mathcal{M}})^2 = (l - \|p_0\|_{\mathcal{T}^*\mathcal{M}})^2 - 2\left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right) p_0^T g^{-1}\delta p_0,$$

where only the p_0 of the same geodesic transforms. So

$$\left\langle D_{p_0} \text{a}, \delta p_0 \right\rangle = \left\langle -\lambda_1 \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right) g^{-1} p_0, \delta p_0 \right\rangle.$$

There is no p_0 dependence on b, c, therefore their gradients vanish. d contains the term

$$\left\langle z(0) - z_0, z_0^\dagger \right\rangle \rightarrow \left\langle z(0) - [z_0 + \pi_p^* \delta p_0], z_0^\dagger \right\rangle = \left\langle z(0) - z_0, z_0^\dagger \right\rangle + \left\langle -\pi_p^* z_0^\dagger, \delta p_0 \right\rangle,$$

therefore

$$\left\langle D_{p_0} \text{d}, \delta p_0 \right\rangle = \left\langle -\pi_p^* z_0^\dagger, \delta p_0 \right\rangle.$$

We add up the contributions to get the form seen in the main text.

A.2.2 $D_g L$

To vary the misfit with respect to $\boldsymbol{g}$, we requre the tranformation of $\boldsymbol{g}^{-1}$. We note that under $\boldsymbol{g} \rightarrow \boldsymbol{g} + \delta \boldsymbol{g}$, to first order

$$g g^{-1} = \mathbf{1} \rightarrow (g + \delta g)\left(g^{-1} + \delta\left(g^{-1}\right)\right) = \mathbf{1} \implies \delta\left(g^{-1}\right) = g^{-1} - g^{-1}\delta g g^{-1},$$

therefore the norm of p_0 transforms under perturbation of g as

$$\|p_0\|_{\mathcal{T}^*\mathcal{M}} = \sqrt{p_0^T g^{-1} p_0} \rightarrow \|p_0\|_{\mathcal{T}^*\mathcal{M}'} = \|p_0\|_{\mathcal{T}^*\mathcal{M}} - \frac{p_0^T g^{-1}\delta g g^{-1} p_0}{2\|p_0\|_{\mathcal{T}^*\mathcal{M}}}.$$

so the following term, contained in a, transforms as

$$(l - \|p_0\|_{\mathcal{T}^*\mathcal{M}})^2 \rightarrow (l - \|p_0\|_{\mathcal{T}^*\mathcal{M}'})^2 = (l - \|p_0\|_{\mathcal{T}^*\mathcal{M}})^2 + \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right) p_0^T g^{-1}\delta g g^{-1} p_0.$$

Therefore

$$\left\langle (D_g \text{b})_{\mu\nu}, \delta g_{\mu\nu} \right\rangle = \sum_i \left\langle \frac{\lambda_1}{2}\left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1\right)\left[g^{-1}p_0\right]^\mu \left[g^{-1}p_0\right]^\nu, \delta g_{\mu\nu} \right\rangle,$$

where g is evaluated at x_0. b has no dependence on g. We now look at c. By Taylor expanding $\partial H/\partial z$ around g,

$$\left\langle (D_g \text{c})_{\mu\nu}, \delta g_{\mu\nu} \right\rangle = \sum_i \int_0^1 \left\langle -\mathbb{J}\frac{\partial^2 H}{\partial z \partial g_{\mu\nu}}(z)\,\delta g_{\mu\nu}, \mathcal{P}z^\dagger \right\rangle dt = \sum_i \int_0^1 \left\langle -(\mathcal{P}z)^{\dagger T}\,\mathbb{J}\frac{\partial^2 H}{\partial z \partial g_{\mu\nu}}(z)\,, \delta g_{\mu\nu} \right\rangle dt.$$

d has no dependence on g. We sum up the contributions to get the form shown in the main text.

A.2.3 $D_z L$

a has no dependence on z. b contains the string

$$(\pi_x z(1) - x_1)^2 \to ([\pi_x z(1) - x_1] + \pi_x \delta z(1))^2 = (\pi_x z(1) - x_1)^2 + \langle 2\pi_x^*(\pi_x z(1) - x_1), \delta z(1)\rangle\,,$$

so

$$\left\langle D_z \text{b}, \delta z \right\rangle = \left\langle \lambda_2 \pi_x^* \left(\pi_x z(1) - x_1\right), \delta z(1) \right\rangle.$$

We now look at c. By varying z, we find that

$$\left\langle D_z \text{c}, \delta z \right\rangle = \int_0^1 \left\langle \frac{d(\delta z)}{dt} - \mathbb{J}\frac{\partial^2 H}{\partial z \partial z}(z)\,\delta z, \mathcal{P}z^\dagger \right\rangle dt.$$

The first term can be evaluated value conditions using integration by parts:

$$\begin{aligned}\int_0^1 \left\langle \frac{d(\delta z)}{dt}, \mathcal{P}z^\dagger \right\rangle dt &= \int_0^1 \frac{d}{dt}\left\langle \delta z, \mathcal{P}z^\dagger \right\rangle dt - \int_0^1 \frac{d}{dt}\left\langle \delta z, \frac{d\left(\mathcal{P}z^\dagger\right)}{dt} \right\rangle dt \\ &= \left\langle \left(\mathcal{P}z^\dagger\right)(1), \delta z(1) \right\rangle - \left\langle \left(\mathcal{P}z^\dagger\right)(0), \delta z(0) \right\rangle + \int_0^1 \left\langle \mathcal{P}\frac{dz^\dagger}{dt}, \delta z \right\rangle dt\end{aligned}$$

where we have used the anticommutation relation, while the second term can be massaged as:

$$\int_0^1 \left\langle -\mathbb{J}\frac{\partial^2 H}{\partial z \partial z}(z)\,\delta z, \mathcal{P}z^\dagger \right\rangle dt = \int_0^1 \left\langle \delta z, -\left[\frac{\partial^2 H}{\partial z \partial z}(z)\right]^T \mathbb{J}^T \mathcal{P}z^\dagger \right\rangle dt = \int_0^1 \left\langle \left\{\mathcal{P}\frac{\partial^2 H}{\partial z \partial z}(z)\,\mathbb{J}\mathcal{P}\right\} z^\dagger, \mathcal{P}\delta z \right\rangle dt,$$

using the invariance and involutory properties and transpose property. We combine the two expressions to obtain the gradient expresison for c:

$$\begin{aligned}\left\langle D_z \text{c}, \delta z \right\rangle &= \left\langle \left(\mathcal{P}z^\dagger\right)(1), \delta z(1) \right\rangle - \left\langle \left(\mathcal{P}z^\dagger\right)(0), \delta z(0) \right\rangle \\ &\quad + \int_0^1 \left\langle \frac{dz^\dagger}{dt} + \left\{\mathcal{P}\frac{\partial^2 H}{\partial z \partial z}(z)\,\mathbb{J}\mathcal{P}\right\} z^\dagger, \mathcal{P}\delta z \right\rangle dt.\end{aligned}$$

Finally, d transforms as following:

$$\left\langle D_z \text{d}, \delta z \right\rangle = \left\langle z_0^\dagger, \delta z(0) \right\rangle;$$

we combine the contributions to find the expression found in the main text.

A.3 Details for the Second Order Adjoint

As discussed in the main text, the second order Lagrangian (5.11) is given by

$$L_1 = \alpha + \beta + \gamma + \delta + \epsilon,$$

where

$$\begin{aligned}
\alpha &= -\sum_i \left\langle \lambda_1 \left[\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right] g^{-1} p_0 + \pi_p z_0^\dagger \ , \ p_0^\sharp \right\rangle, \\
\beta &= \beta 1 + \beta 2 \\
&= \sum_i \left\langle \left[\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right] \left[g^{-1} p_0 \right]^\mu \left[g^{-1} p_0 \right]^\nu \ , \ g^\sharp_{\mu\nu} \right\rangle \\
&- \sum_i \int_0^1 \left\langle \left[\mathcal{P} z^\dagger \right]^T \mathbb{J} \frac{\partial^2 H}{\partial z \partial g_{\mu\nu}} (z) \ , \ g^\sharp \right\rangle dt, \\
\gamma &= \gamma 1 + \gamma 2 + \gamma 3 \\
&= \sum_i \left\langle \lambda_2 \pi_x^* \left(\pi_x z(1) - x_1 \right) + z^\dagger(0), z^\sharp(1) \right\rangle \\
&+ \sum_i \left\langle z_0^\dagger - z^\dagger(1), z^\sharp(0) \right\rangle \\
&+ \sum_i \int_0^1 \left\langle \frac{dz^\dagger}{dt} + \left\{ \mathcal{P} \frac{\partial^2 H}{\partial z \partial z}(z) \mathbb{J} \mathcal{P} \right\} z^\dagger, \mathcal{P} z^\sharp \right\rangle dt, \\
\delta &= \sum_i \int_0^1 \left\langle \frac{dz}{dt} - \mathbb{J} \frac{\partial H}{\partial z}(z), \mathcal{P} z^{\dagger\sharp} \right\rangle dt, \\
\epsilon &= \sum_i \left\langle z(0) - z_0, z_0^{\dagger\sharp} \right\rangle.
\end{aligned}$$

A.3.1 $D_{p_0} L_1$

We look at α. We note that the following string contained in α, expanding binomially, transforms as

$$\begin{aligned}
\left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) p_0^T &\to \left\{ \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) - \frac{2 l p_0^T g^{-1} \delta p_0}{\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} \right\} \left(p_0^T + \delta p_0^T \right) \\
&= \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) p_0^T - \frac{2 l p_0^T g^{-1} \delta p_0}{\|p_0\|^3_{\mathcal{T}^*\mathcal{M}}} p_0^T + \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) \delta p_0^T.
\end{aligned}$$

With substitution and massaging of the equation, we find that

$$\left\langle D_{p_0} \alpha, \delta p_0 \right\rangle = \left\langle \lambda_1 \left\{ \frac{2 l p_0^T g^{-1} p_0^\sharp}{\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} g^{-1} p_0 - \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) g^{-1} p_0^\sharp \right\}, \delta p_0 \right\rangle.$$

We $\beta 1$ and collect orders of $\mathcal{O}(\delta p)$, where the expressions are similar to given before. We find

$$\beta 1' - \beta 1 = \frac{\lambda_1}{2} \left\{ - \frac{2 l p_0^T g^{-1} \delta p_0}{\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} p_0^T g^{-1} g^\sharp g^{-1} p_0 + \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) 2 p_0^T g^{-1} g^\sharp g^{-1} \delta p_0 \right\}$$

from which we can deduce that

$$\left\langle D_{p_0}\beta, \delta p_0 \right\rangle = \left\langle \lambda_1 \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) g^{-1} g^\sharp g^{-1} p_0 - \lambda_1 \frac{l p_0^T g^{-1} g^\sharp g^{-1} p_0}{\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} g^{-1} p_0, \delta p_0 \right\rangle.$$

$\beta 1$, γ, δ do not depend on p_0. Upon substitution, similar to the case of d for L, we can deduce that

$$\left\langle D_{p_0}\epsilon, \delta p_0 \right\rangle = \left\langle -\pi_p z_0^{\dagger\sharp}, \delta p_0 \right\rangle.$$

Combining the contributions, collecting the common factors, we find the expression given in the main text.

A.3.2 $D_g L_1$

Using the transformation property of g^{-1} as discussed before, binomially expanding, we can find that

$$\frac{1}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} \to \frac{1}{\|p_0\|_{\mathcal{T}^*\mathcal{M}'}} = \frac{1}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} \left(1 + \frac{p_0^T g^{-1} \delta g g^{-1} p_0}{2\|p_0\|^2_{\mathcal{T}^*\mathcal{M}}} \right);$$

From this, we can deduce that

$$\alpha' - \alpha = \lambda_1 \sum_i \left\{ \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) p_0^T g^{-1} \delta g g^{-1} p_0^\sharp - \frac{l p_0^T g^{-1} \delta g g^{-1} p_0}{2\|p_0\|^3_{\mathcal{T}^*\mathcal{M}}} p_0^T g^{-1} p_0^\sharp \right\},$$

from which we can deduce that

$$\left\langle D_g\alpha, \delta g \right\rangle = \lambda_1 \sum_i \left\langle \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) [g^{-1}p_0]^\mu [g^{-1}p_0^\sharp]^\nu - \frac{l p_0^T g^{-1} p_0^\sharp}{2\|p_0\|^3_{\mathcal{T}^*\mathcal{M}}} [g^{-1}p_0]^\mu [g^{-1}p_0]^\nu, \delta g_{\mu\nu} \right\rangle.$$

$\beta 1$ transforms similarly to α. Variation of $\beta 2$ with respect to g can be Taylor-expanded, yielding

$$\left\langle D_g\beta, \delta g \right\rangle = \lambda_1 \sum_i \left\langle \left(\frac{l}{\|p_0\|_{\mathcal{T}^*\mathcal{M}}} - 1 \right) [g^{-1}p_0]^\mu [g^\sharp p_0^\sharp]^\nu - \frac{l p_0^T g^{-1} g^\sharp g^{-1} p_0}{4\|p_0\|^3_{\mathcal{T}^*\mathcal{M}}} [g^{-1}p_0]^\mu [g^{-1}p_0]^\nu, \delta g_{\mu\nu} \right\rangle$$
$$- \sum_i \left\langle \left(\mathcal{P} z^\dagger\right)^T \mathbb{J} \frac{\partial^3 H}{\partial z \partial g_{\rho\sigma} g_{\mu\nu}} g_{\rho\sigma}, \delta g_{\mu\nu} \right\rangle dt.$$

$\gamma 1$, $\gamma 2$ are not dependent on g. To evaluate how $\gamma 3$ varies with g, it is useful to note the distributivity of $\mathcal{P}$:

$$\mathcal{P}AB := \mathcal{P}(AB) = (\mathcal{P}A)(\mathcal{P}B).$$

Therefore

$$\gamma 3' - \gamma 3 = \sum_i \int_0^1 \left(\mathcal{P} z^\sharp\right)^T \mathcal{P} \frac{\partial^3 H}{\partial z \partial g_{\mu\nu} \partial z}(z) \delta g_{\mu\nu} \mathbb{J} \mathcal{P} z^\dagger dt$$
$$= \left\langle D_g\gamma 3, \delta g \right\rangle = \sum_i \int_0^1 \left\langle \mathcal{P} z^{\sharp T} \frac{\partial^3 H}{\partial z \partial g_{\mu\nu} \partial z}(z) \mathbb{J} \mathcal{P} z^\dagger, \mathcal{P} \delta g_{\mu\nu} \right\rangle dt$$
$$= \sum_i \int_0^1 \left\langle z^{\sharp T} \frac{\partial^3 H}{\partial z \partial g_{\mu\nu} \partial z}(z) \mathbb{J} \mathcal{P} z^\dagger, \delta g_{\mu\nu} \right\rangle dt.$$

Similarly for δ, we find that

$$\left\langle D_g\delta, \delta g \right\rangle = \sum_i int_0^1 \left\langle -\mathbb{J} \frac{\partial^2 H}{\partial g_{\mu\nu} \partial z}(z) \mathcal{P} z^{\dagger\sharp}, \delta g_{\mu\nu} \right\rangle dt.$$

Finally, ϵ does not depend on g. We combine the contributions and get the result in the main text.

A.3.3 $D_z L_1$

Only $\beta 2$, $\gamma 2$, $\gamma 3$, δ, ϵ are dependent on z. All derivations integration by parts and Taylor expansion techniques similar to the first order adjoint; we present some of the steps below.

$$\beta 2' - \beta 2 = -\int_0^1 \left\langle \left(\mathcal{P}z^\dagger\right)^T \mathbb{J}\frac{\partial^3 H}{\partial z \partial g_{\mu\nu} \partial z}\delta z, g^\sharp_{\mu\nu} \right\rangle dt = \int_0^1 \left\langle \mathcal{P} g^\sharp_{\mu\nu} \frac{\partial^3 H}{\partial z \partial g_{\mu\nu} \partial z} \mathbb{J}\mathcal{P}z^\dagger \delta z, \right\rangle dt,$$

$$\gamma 2' - \gamma 2 = \left\langle \lambda_2 \pi_x^* \pi_x z^\sharp(0) \delta z(1) \right\rangle,$$

$$\gamma 3' - \gamma 3 = \int_0^1 \left\langle \mathcal{P}\frac{\partial^3 H}{\partial z \partial z_a \partial z}\delta z_a \mathbb{J}\mathcal{P}z^\dagger, \mathcal{P}z^\sharp \right\rangle dt = \int_0^1 \left\langle \mathcal{P}z^{\sharp T}\frac{\partial^3 H}{\partial z \partial z \partial z}\mathbb{J}\mathcal{P}z^\dagger, \mathcal{P}z \right\rangle dt,$$

$$\delta' - \delta = \left\langle \delta z(1), \left(\mathcal{P}z^{\dagger\sharp}\right)(1) \right\rangle - \left\langle \delta z(0), \left(\mathcal{P}z^{\dagger\sharp}\right)(0) \right\rangle + \int_0^1 \left\langle \frac{dz^{\dagger\sharp}}{dt} + \mathcal{P}\frac{\partial^2 H}{\partial z \partial z}\mathbb{J}\mathcal{P}z^{\dagger\sharp}, \mathcal{P}\delta z \right\rangle dt,$$

$$\epsilon' - \epsilon = \left\langle \delta z(0), z_0^{\dagger\sharp} \right\rangle.$$

A.3.4 $D_{z^\dagger} L_1$

Only $\beta 2$, γ are dependent on $z^\dagger$.

$$\beta 2' - \beta 2 = \int_0^1 \left\langle \mathbb{J}\frac{\partial^2 H}{\partial z \partial g_{\mu\nu}} g^\sharp_{\mu\nu}, \mathcal{P}\delta z^\dagger \right\rangle dt,$$

$$\gamma 1' - \gamma 1 = \left\langle -z^\sharp(0), \delta z^\dagger(1) \right\rangle; \quad \gamma 2' - \gamma 2 = \left\langle -z^\sharp(1), \delta z^\dagger(0) \right\rangle$$

$$\gamma 3' - \gamma 3 = \left\langle \left(\mathcal{P}z^\sharp\right)(1), \delta z^\dagger(1) \right\rangle - \left\langle \left(\mathcal{P}z^\sharp\right)(0), \delta z^\dagger(0) \right\rangle + \int_0^1 \left\langle \frac{dz^\sharp}{dt} - \mathbb{J}\frac{\partial^2 H}{\partial z \partial z} z^\sharp, \mathcal{P}\delta z^\dagger \right\rangle.$$

We note that the first two terms in the $\gamma 3$ cancels the $\gamma 1$, $\gamma 2$ expressions.

We sum up all the contributions to obtain the expression given in the main text.

A.3.5 $D_{z_0^\dagger} L_1$

Finally, we look at the variation of L_1 with respect to $z_0^\dagger$. Thankfully, this derivation is easy. Only α, $\gamma 1$ are dependent on $z_0^\dagger$.

$$\alpha' - \alpha = \left\langle -\pi_p^* p_0^\sharp, \delta z_0^\dagger \right\rangle,$$

and

$$\gamma 1' - \gamma 1 = \left\langle z^\sharp(0), \delta z_0^\dagger \right\rangle.$$

We sum up all the contributions to obtain the expression given in the main text.