

Robust Parameter Inversion Using Stochastic Estimates

Drayton William Munster

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

Eric De Sturler, Chair
Serkan Gugercin
Christopher A. Beattie
Matthias Chung

December 16, 2019
Blacksburg, Virginia

Keywords: Model Reduction, Trace Estimation, Trust Regions, Optimization
Copyright 2019, Drayton William Munster

Robust Parameter Inversion Using Stochastic Estimates

Drayton William Munster

(ABSTRACT)

For parameter inversion problems governed by systems of partial differential equations, such as those arising in Diffuse Optical Tomography (DOT), even the cost of repeated objective function evaluation can be overwhelming. Despite the linear (in the state variable) nature of the DOT problem, the nonlinear parameter inversion process is dominated by the computational burden of solving a large linear system for each source and frequency. To compute the Jacobian for use in Newton-type methods, an adjoint solve is required for each detector and frequency. When a three-dimensional tomography problem may have nearly 1000 sources and detectors, the computational cost of an optimization routine is a large burden. While techniques from model order reduction can partially alleviate the computational cost, obtaining error bounds in parameter space is typically not feasible. In this work, we examine two different remedies based on stochastic estimates of the objective function.

In the first manuscript, we focus on maximizing the efficiency of using stochastic estimates by replacing our objective function with a surrogate objective function computed from a reduced order model (ROM). We use as few as a single sample to detect a misfit between the full-order and surrogate objective functions. Once a sufficiently large difference is detected, it is necessary to update the ROM to reduce the error. We propose a new technique for improving the ROM with very few large linear solutions. Using this techniques, we observe a reduction of up to 98% in the number of large linear solutions for a three-dimensional tomography problem.

In the second manuscript, we focus on establishing a robust algorithm. We propose a new trust region framework that replaces the objective function evaluations with stochastic estimates of the improvement factor and the misfit between the model and objective function gradients. If these estimates satisfy a fixed multiplicative error bound with a high, but fixed, probability, we show that this framework converges almost surely to a stationary point of the objective function. We derive suitable bounds for the DOT problem and present results illustrating the robust nature of these estimates with only 10 samples per iteration.

Robust Parameter Inversion Using Stochastic Estimates

Drayton William Munster

(GENERAL AUDIENCE ABSTRACT)

For problems such as medical imaging, the process of reconstructing the state of a system from measurement data can be very expensive to compute. The ever increasing need for high accuracy requires very large models to be used. Reducing the computational burden by replacing the model with a specially constructed smaller model is an established and effective technique. However, it can be difficult to determine how well the smaller model matches the original model. In this thesis, we examine two techniques for estimating the quality of a smaller model based on randomized combinations of sources and detectors.

The first technique focuses on reducing the computational cost as much as possible. With the equivalent of a single randomized source, we show that this estimate is an effective measure of the model quality. Coupled with a new technique for improving the smaller model, we demonstrate a highly efficient and robust method.

The second technique prioritizes robustness in its algorithm. The algorithm uses these randomized combinations to estimate how the observations change for different system states. If these estimates are accurate with a high probability, we show that this leads to a method that always finds a minimum misfit between predicted values and the observed data.

Acknowledgments

I would like to begin by thanking Eric de Sturler, my advisor. He took a itinerant graduate student and offered them an academic home. Between the time spent proof-reading my work, his flexibility and understanding of my constantly varying schedule, and his encouragement, this thesis simply would not exist without him.

I would also like to thank Lizette Zietsman and Jeffery Borggaard. Their support and guidance during my formative undergraduate and graduate years was key in forming the mathematician I am today.

The feedback and guidance from my committee members, Serkan Gugercin, Christopher Beattie, and Matthias Chung were instrumental in forming this thesis, for which I am grateful.

Finally, I would like to thank my family. The encouragement I received from them were what allowed me to pursue my research in the first place. Especially in the process of writing this thesis, their patience and encouragement kept me going.

This material is partially based upon work supported by the National Science Foundation under Grant No. 1720305. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Contents

List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Attribution	2
2 Robust Parameter Inversion Using Adaptive Reduced Order Models	3
2.1 Introduction	4
2.2 Background	5
2.2.1 The DOT Problem	5
2.2.2 Interpolatory Model Reduction	7
2.2.3 Trust region methods	9
2.2.4 Trace Estimation	11
2.3 Estimating ROM Accuracy and Guiding Updates	11
2.4 Basis Updates	12
2.4.1 Full Interpolatory Updates	13
2.4.2 Residual Driven Updates	13
2.5 Probabilistic Estimates	16
2.6 Numerical Experiments	18
2.6.1 2D Example	18
2.6.2 3D	22
2.7 Conclusions and Future Work	25
2.8 Acknowledgements	27
3 Trust Regions with Stochastic Bounds	28

3.1	Introduction	29
3.2	Trust Regions with Stochastic Bounds	29
3.3	Stochastic Processes	31
3.3.1	Definitions	31
3.3.2	Stochastic Processes	32
3.4	Convergence Analysis	34
3.4.1	Trust Region Radius Convergence	35
3.4.2	$\liminf$ Convergence	46
3.4.3	$\lim$ Convergence	47
3.5	Projection-Based Model Order Reduction	50
3.5.1	Linear Time-Invariant Systems	50
3.5.2	Projection-Based Model Order Reduction	51
3.6	Estimates for the DOT Problem	52
3.6.1	Hutchinson Trace Estimation	52
3.6.2	Estimates for $f(x_k) - f(x_k + s_k)$	53
3.6.3	Estimates of $\ \nabla m_\ell(x_k) - \nabla f(x_k)\ $	56
3.6.4	Another Approach to $\ \nabla m_\ell(x_k) - \nabla f(x_k)\ $	66
3.7	Numerical Results	68
3.7.1	Two-Dimensional DOT Problem	68
3.7.2	Three-Dimensional DOT Problem	71
3.8	Conclusions	71
4	Conclusions	73
	Bibliography	74

List of Figures

2.1	2D and 3D Geometries	5
2.2	Improvement in Error Bound at First Rejected Step in 2D - 1 Point Experiment.	16
2.3	Improvement in Relative Residual Norms	17
2.4	True Image for the 2D Model and Initial Condition	19
2.5	Convergence behavior of 3-point ROM without updates and 3-point ROM with Interpolatory Updates for a representative realization of the 2D experiment	21
2.6	Convergence history of the 1-point ROM with Residual-Driven updates for a representative realization of the 2D experiment	23
2.7	2D Reconstruction Results	24
2.8	True Image for the 3D Model and Initial Condition	25
2.9	3D Reconstruction Results	26
3.1	$\mathbf{P} [T_{\mathbf{C}}^H \cdot T_{\mathbf{B}}^H \geq (1 - \epsilon)^2 \cdot T_{\mathbf{B}} \cdot T_{\mathbf{C}}] \geq \mathbf{P} [\{T_{\mathbf{B}}^H \geq (1 - \epsilon)T_{\mathbf{B}}\} \cap \{T_{\mathbf{C}}^H \geq (1 - \epsilon)T_{\mathbf{C}}\}]$	58
3.2	Convergence History of the Currently Accepted Step for a Two-Dimensional DOT Problem with Two Incorrectly Accepted Steps	69
3.3	PDF-Normalized Histogram of the Ratio between Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem	70

List of Tables

2.1	Number of Large Linear Solves and Relative Residual Error for the 2D Problem	22
2.2	Number of Large Linear Solves for the 3D Problem	27
3.1	Scaling of the factor ϕ for the sufficient bound of Theorem 3.20 with $\alpha_L = 4/5, \alpha_U = 4$	56
3.2	Percentile Values of the Ratio between Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem	70
3.3	Outcome Counts for Accepting or Rejecting the Candidate Step for the Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem	71
3.4	Outcome Counts for Accepting or Rejecting the Candidate Step for the Estimated and Actual Improvement Factors for the Three-Dimensional Tomography Problem	71

Chapter 1

Introduction

In this work, we focus on the reconstruction of a system’s unobservable parameters given a set of (noisy) measurements. This is known as parameter inversion or an inverse problem. To do so, we apply a mathematical model (henceforth, the forward model) to the parameterized state. We then compare the predicted measurements with the empirical data. To find a set of parameters that optimally match (or within some tolerance) the empirical data, an optimization routine may require many evaluations of the forward model. When the forward model is a large-scale, discretized approximation of a partial differential equation (PDE), these repeated function evaluations can pose a large computational burden. The use of Newton-type optimizers requires the computation of the Jacobian, which requires an additional set of systems for each evaluation.

In particular, we focus on the Diffuse Optical Tomography (DOT) problem in this thesis. While the DOT problem is linear in the state variables, evaluation of the full model involves the solution of a large linear system for each source and each frequency of the light source. Jacobian evaluations require an adjoint solve for each detector and each frequency. A typical three-dimensional tomography problem can involve nearly 1000 sources and detectors, thus requiring 1000-2000 large linear systems to be solved per optimizer iteration. This cost can be reduced using model order reduction to produce a surrogate model of significantly smaller order. However, determining the accuracy of this surrogate model can be very difficult for most problems. In this thesis, we introduce two new techniques for estimating the accuracy of the surrogate model.

In Chapter 2, we focus on an efficient method for estimating the difference between the objective function and its surrogate. We do this by employing trace estimation which is equivalent to using a random linear combination of sources. With even as few as a single sample per frequency, we show that this technique is effective for the DOT problem. Once a sufficiently large mismatch is detected, it becomes necessary to improve the reduced order model (ROM). We propose a new technique for updating the reduced order model that significantly reduces the interpolation error with $O(1-10)$ solutions to large linear systems per frequency. By starting with a very small initial ROM, estimating the interpolation error, and employing these cheap updates, we demonstrate a robust and extremely efficient parameter inversion method.

In Chapter 3, we focus on constructing a robust framework for optimization. We propose a novel trust region framework that completely removes the need for any (full-order) objective

function and Jacobian evaluations. Instead, we use stochastic estimates to estimate the mismatch between the model and objective function gradients as well as estimates of the change in objective function. In section 3.4, we show that, under the assumption of a fixed multiplicative bound on the estimate error with a fixed probability, that this framework converges with probability one. This is an improvement over methods that require a sequence of increasing probabilities [24] or accuracy relative to the trust region radius [19]. We then derive a series of suitable estimates for the DOT problem in Section 3.6, some with rigorous bounds on a sufficient number of samples requires, some with higher performance. We apply this framework to the DOT problem and observe an extremely robust parameter inversion method.

1.1 Attribution

Eric de Sturler is listed as a co-author of the manuscript titled *Robust Parameter Inversion Using Adaptive Reduced Order Models*. Section 2.2.3 is a background subsection on trust region methods that was primarily authored by Eric de Sturler. The remainder of the article, including all of the novel work, is the work of Drayton Munster.

The manuscript titled *Trust Regions with Stochastic Bounds* is entirely authored by Drayton Munster.

Chapter 2

Robust Parameter Inversion Using Adaptive Reduced Order Models

Nonlinear parametric inverse problems appear in many applications and are typically very expensive to solve, especially if they involve many measurements. These problems pose huge computational challenges as evaluating the objective function or misfit requires the solution of a large number of parameterized partial differential equations, typically one per source term. Newton-type algorithms, which may be required for fast convergence, typically require the additional solution of a large number of adjoint problems.

The use of parametric model reduction may substantially alleviate this problem. In [de Sturler, E., Gugercin, S., Kilmer, M. E., Chaturantabut, S., Beattie, C., & O’Connell, M. (2015). Nonlinear Parametric Inversion Using Interpolatory Model Reduction. *SIAM Journal on Scientific Computing*, 37(3)], interpolatory model reduction was successfully used to drastically speed up inversion for Diffuse Optical Tomography (DOT). However, when using model reduction in high dimensional parameter spaces, obtaining error bounds in parameter space is typically intractable. In this paper, we propose to use stochastic estimates to remedy this problem. At the cost of one (randomized) full-scale linear solve per optimization step we obtain a robust algorithm. Moreover, since we can now update the model when needed, this robustness allows us to further reduce the order of the reduced order model and hence the cost of computing and using it, further decreasing the cost of inversion. We also propose a method to update the model reduction basis that reduces the number of large linear solves required by 46%-98% compared to the fixed reduced-order model. We demonstrate that this leads to a highly efficient and robust inversion method.

2.1 Introduction

Nonlinear parameter inversion involves finding a set of parameters that minimizes the difference (or misfit) between the output of a parametric forward model and measured data. This minimization requires many evaluations of the forward model. When the forward model is described by discretized partial differential equations (PDEs), the number of large linear solves may be computationally intractable. To reduce this cost, the forward model can be approximated with a Reduced-Order Model (ROM). Using interpolatory, projection-based ROMs for Diffuse Optical Tomography (DOT) has been investigated in [22], yielding a very efficient method. However, these results do not provide an a posteriori error bound without evaluating the misfit using the Full Order Model (FOM). Certain classes of problems (for example, coercive parabolic and elliptical PDEs [5, 33]) and ROMs admit a posteriori error bounds in parameter space. When these bounds are not available, a globally certified reduced basis can be constructed with the “greedy POD” method [18, 33] in an “offline-online” approach, where a global basis is constructed beforehand (“offline”) and reused for particular problems (“online”). This “offline” construction is typically also computationally intractable for high-dimensional parameter spaces since the reduced basis must be sufficiently accurate across the entire parameter space. In [29], the author examines the use of a multi-level strategy to defer the most expensive evaluations for later iterations when the current iterate should be close to the optimal point. To reduce the costs of this “offline” phase and the costs of using a larger ROM, various adaptive strategies have been examined. In [1], the authors use an adaptive POD basis to approximate an optimal control. This basis is iteratively updated by computing an optimal control for the ROM and adding the (full-order) state corresponding to that control to the snapshot basis. The iteration stops when subsequent controls differ by less than a given tolerance. The interpolation of POD bases is another offline strategy, as examined in [6, 17] (and references therein). In [32], the authors incorporate new data during the online phase by expanding the DEIM basis with low-rank updates chosen to reduce the approximation error at a randomly selected subset of components.

In this paper, instead, we investigate the use of stochastic techniques for trace estimation to efficiently estimate the error at the current step and provide a robust optimization technique. A significant reduction in the number of large linear solves is achieved by using a small initial projection basis and using these estimates to indicate when the bases should be expanded. To further reduce the number of large linear solves, we propose an update scheme inspired by [31] to minimize the number of additional vectors used to extend the projection basis.

The paper is organized as follows. Section 2.2 introduces the DOT problem, interpolatory model reduction, trust region methods, and trace estimation techniques. We introduce our proposed algorithm using stochastic estimates to guide ROM updates in section 2.3. Section 2.4 discusses two possible choices for efficiently updating the projection basis to improve the accuracy of the ROM. While this paper focuses on efficiency, section 2.5 outlines the number of samples necessary for probabilistic guarantees of the accuracy of these estimates

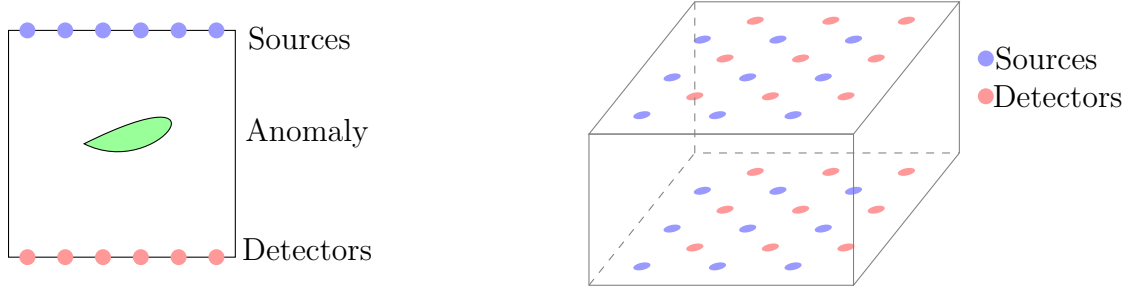


Figure 2.1: 2D and 3D Geometries

for problems that require more robustness. We explore some of the choices with numerical experiments and their performance implications in section 2.6. Finally, we provide conclusions and discuss future work in section 2.7.

2.2 Background

We first describe the forward model for DOT and express this problem in the system theoretic notation used in the remainder of the paper. This section borrows heavily from [22]. Next, we give background on the construction of reduced frequency response functions via interpolatory model reduction. The third subsection provides background on estimating the trace of a matrix and its application to estimating Frobenius norms.

2.2.1 The DOT Problem

We use a diffusion model [8] for the photon flux, $\eta(\mathbf{z}, t)$, driven by an input light source $g(\mathbf{z}, t)$. In practice, the DOT problem is posed in the frequency domain. Here, we pose the problem in the time domain to motivate the use of parametric inversion techniques from a system theoretic perspective. Light is transmitted from one of n_{src} physically stationary sources, so we write $g(\mathbf{z}, t) = b_j(\mathbf{z})u_j(t)$ for the source locations $b_j(\mathbf{z})$, $j = 1, \dots, n_{\text{src}}$. Observations are made using an array of n_{det} detectors located along the boundary. See Figure 2.1 for examples used in section 2.6. Let $o_i(t)$ denote the observed flux at detector i and time t . With the above notation, we model the diffusion and absorption of light by the following

time-dependent partial differential equation,

$$\frac{1}{\nu} \frac{\partial}{\partial t} \eta(\mathbf{z}, t) = \nabla \cdot (D(\mathbf{z}) \nabla \eta(\mathbf{z}, t)) - \mu(\mathbf{z}) \eta(\mathbf{z}, t) + b_j(\mathbf{z}) u_j(t), \text{ for } \mathbf{z} \in \mathcal{X} \quad (2.1)$$

$$0 = \eta(\mathbf{z}, t) + 2aD(\mathbf{z}) \frac{\partial}{\partial \xi} \eta(\mathbf{z}, t), \text{ for } \mathbf{z} \in \partial\mathcal{X} \setminus \partial\mathcal{X}_{\pm} \quad (2.2)$$

$$0 = \eta(\mathbf{z}, t), \text{ for } \mathbf{z} \in \partial\mathcal{X}_{\pm} \quad (2.3)$$

$$o_i(t) = \int_{\partial\mathcal{X}} c_i(\mathbf{z}) \eta(\mathbf{z}, t) d\mathbf{z}. \quad (2.4)$$

Here, $\mathbf{z}$ refers to a spatial location in our image domain $\mathcal{X}$, $\partial\mathcal{X}_{\pm}$ refer to the top and bottom of the image domain (respectively), a is a constant defining the diffusive boundary reflection, $D(\mathbf{z})$ and $\mu(\mathbf{z})$ refer to diffusion and absorption fields (respectively), ξ refers to the outward unit normal on the boundary, and ν is the speed of light in the medium.

The inverse problem associated with DOT is to reconstruct $D(\mathbf{z})$ and $\mu(\mathbf{z})$, the diffusion and absorption fields (respectively), given a set of observations $\mathbf{o}(t)$ made by illuminating the domain with a variety of source signals $\mathbf{u}(t)$. In this paper, we assume that the diffusion field $D(\mathbf{z})$ is known (a common assumption in DOT breast tissue imaging). This leaves the reconstruction of the absorption field, $\mu(\mathbf{z})$. We represent this field with a finite number of (unknown) parameters, $\mathbf{x} = [x_1, \dots, x_{n_p}]^T$. The choice of parameterization is crucial for physiologically relevant solutions to the inverse problem. In this setting, $\mu(\mathbf{z})$ is well-approximated by a piecewise constant, two-valued function. As for many inverse problems, the naive approach is ill-posed and extremely sensitive to noise in the measurements. To efficiently describe complex geometries with sharp boundaries and provide regularization, we use parametric level sets (PaLS), developed in [2] and used in the context of DOT imaging in [2, 22].

The spatial discretization of eqs. (2.1) to (2.4) yields the system of differential algebraic equations,

$$\frac{1}{\nu} \mathbf{E} \dot{\mathbf{y}}(t; \mathbf{x}) = -\mathbf{A}(\mathbf{x}) \mathbf{y}(t; \mathbf{x}) + \mathbf{B} \mathbf{u}(t), \quad (2.5)$$

$$\mathbf{o}(t; \mathbf{x}) = \mathbf{C}^T \mathbf{y}(t; \mathbf{x}), \quad (2.6)$$

where $\mathbf{y}$ denotes the discretized photon flux, $\mathbf{o} = [o_1, \dots, o_{n_{\text{det}}}]^T$ is the vector of detector outputs, $\mathbf{C}^T \mathbf{y}$ approximates eq. (2.4) via quadrature, the columns of $\mathbf{B}$ represent discretizations of the sources $b_j(\mathbf{z})$, $\mathbf{A}(\mathbf{x}) = \mathbf{A}_0 + \mathbf{A}_1(\mathbf{x})$, where $\mathbf{A}_0$ and $\mathbf{A}_1$ are discretizations of the diffusion and absorption terms, respectively. $\mathbf{E}$ includes the discretization of boundary terms such as the Robin condition eq. (2.2) and is singular as a result. Let $\hat{\mathbf{y}}(\omega; \mathbf{x})$, $\hat{\mathbf{o}}(\omega; \mathbf{x})$, $\hat{\mathbf{u}}(\omega; \mathbf{x})$ denote the Fourier transform of $\mathbf{y}(t; \mathbf{x})$, $\mathbf{o}(t; \mathbf{x})$, and $\mathbf{u}(t; \mathbf{x})$, respectively. Taking the Fourier transform of eqs. (2.5) and (2.6) yields

$$\hat{\mathbf{o}}(\omega; \mathbf{x}) = \Psi(\omega; \mathbf{x}) \hat{\mathbf{u}}(\omega; \mathbf{x}), \text{ where} \quad (2.7)$$

$$\Psi(\omega; \mathbf{x}) = \mathbf{C}^T (\mathbf{K}(\omega; \mathbf{x}))^{-1} \mathbf{B}, \quad (2.8)$$

where $\omega \in \mathbb{R}$ and $\mathbf{K}(\omega; \mathbf{x}) = \frac{i\omega}{\nu} \mathbf{E} - \mathbf{A}(\mathbf{x})$. $\Psi(\omega; \mathbf{x})$ is the *frequency response function* of the dynamical system eq. (2.5).

For a given frequency ω_j and input source i , we denote the predicted observations by the forward model $\hat{\mathbf{o}}_i(\omega_j; \mathbf{x})$. For a given parameter vector $\mathbf{x}$, the predicted observations for all n_{src} input sources and n_ω frequencies are given by

$$\mathbf{O}(\mathbf{x}) = [\hat{\mathbf{o}}_1(\omega_1; \mathbf{x}), \dots, \hat{\mathbf{o}}_{n_{\text{src}}}(\omega_1; \mathbf{x}), \hat{\mathbf{o}}_1(\omega_2; \mathbf{x}), \dots, \hat{\mathbf{o}}_{n_{\text{src}}}(\omega_{n_\omega}; \mathbf{x})], \quad (2.9)$$

where $\mathbf{O}(\mathbf{x}) \in \mathbb{C}^{n_{\text{det}} \times n_{\text{src}} \cdot n_\omega}$. In our case, $\hat{\mathbf{u}}_i(\omega_j) = \mathbf{e}_i$, representing the excitation of source i with a pure frequency ω_j . The evaluation of $\mathbf{O}(\mathbf{x})$ then reduces to the evaluation of the frequency response function for each frequency

$$\mathbf{O}(\mathbf{x}) = [\Psi(\omega_1; \mathbf{x}), \dots, \Psi(\omega_{n_\omega}; \mathbf{x})]. \quad (2.10)$$

Given the empirical data matrix of observations, $\mathbf{D}$, the optimization problem that must be solved is

$$\min_{\mathbf{x} \in \mathbb{R}^{n_p}} \|\mathbf{R}(\mathbf{x})\|_F = \min_{\mathbf{x} \in \mathbb{R}^{n_p}} \|\mathbf{O}(\mathbf{x}) - \mathbf{D}\|_F. \quad (2.11)$$

We assume $\mathbf{D}$ contains additive noise at a known noise level. Since PaLS regularizes the problem [2], no further regularization is necessary and we terminate the optimization when $\|\mathbf{R}(\mathbf{x})\|$ is at the noise level (or slightly above).

Although the number of systems is independent of the number of parameters, each objective function evaluation requires $\min(n_{\text{det}}, n_{\text{src}}) \cdot n_\omega$ large linear solves. Since we use Newton-type methods to minimize eq. (2.11), it is also necessary to compute the Jacobian of the objective function. Differentiation of $\Psi(\omega_k; \mathbf{x})$ with respect to parameter p_ℓ yields

$$\frac{\partial}{\partial x_\ell} \Psi(\omega_k; \mathbf{x}) = -\mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left(\frac{\partial}{\partial x_\ell} \mathbf{A}(\mathbf{x}) \right) (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B}. \quad (2.12)$$

Since $\frac{\partial}{\partial x_\ell} \mathbf{A}(\mathbf{x})$ is diagonal and inexpensive to compute in our case, the bulk of the computational cost is in solving the systems $\mathbf{K}(\omega_k; \mathbf{x}) \mathbf{X} = \mathbf{B}$ and $(\mathbf{K}(\omega_k; \mathbf{x}))^T \mathbf{Y} = \mathbf{C}$. Note that one of these solutions is already available from the objective function evaluation (Equation (2.10)).

2.2.2 Interpolatory Model Reduction

Since solving eq. (2.11) is dominated by the cost of computing eq. (2.10) and eq. (2.12), we build a surrogate frequency response function, $\Psi_R(\omega; \mathbf{x})$, that maintains a high-fidelity approximation to $\Psi(\omega; \mathbf{x})$ using techniques from projection based parametric model reduction [7, 15, 22].

Assuming the state, $\mathbf{y}$, evolves near the n_r -dimensional subspace $\text{Range}(\mathbf{V})$ for some appropriately chosen $\mathbf{V} \in \mathbb{C}^{n \times n_r}$, i.e. $\mathbf{y}(t; \mathbf{x}) \approx \mathbf{V} \mathbf{y}_R(t; \mathbf{x})$ and given a full-rank $\mathbf{W} \in \mathbb{C}^{n \times n_r}$, we

enforce the Petrov-Galerkin condition

$$\mathbf{W}^T \left(\frac{1}{\nu} \mathbf{E} \mathbf{V} \dot{\mathbf{y}}_{\mathbf{R}}(t; \mathbf{x}) + \mathbf{A}(\mathbf{x}) \mathbf{V} \mathbf{y}_{\mathbf{R}}(t; \mathbf{x}) - \mathbf{B} \mathbf{u}(t) \right) = 0, \quad \mathbf{o}(t; \mathbf{x}) = \mathbf{C}^T \mathbf{V} \mathbf{y}_{\mathbf{R}}(t; \mathbf{x})$$

to yield the reduced system

$$\frac{1}{\nu} \mathbf{E}_{\mathbf{R}} \dot{\mathbf{y}}_{\mathbf{R}}(t; \mathbf{x}) = -\mathbf{A}_{\mathbf{R}}(\mathbf{x}) \mathbf{y}_{\mathbf{R}}(t; \mathbf{x}) + \mathbf{B}_{\mathbf{R}} \mathbf{u}(t), \quad \mathbf{o}_{\mathbf{R}}(t; \mathbf{x}) = \mathbf{C}_{\mathbf{R}}^T \mathbf{y}_{\mathbf{R}}(t; \mathbf{x}) \quad (2.13)$$

and associated reduced frequency response function

$$\Psi_{\mathbf{R}}(\omega; \mathbf{x}) = \mathbf{C}_{\mathbf{R}}^T \left(\frac{i\omega}{\nu} \mathbf{E}_{\mathbf{R}} - \mathbf{A}_{\mathbf{R}}(\mathbf{x}) \right)^{-1} \mathbf{B}_{\mathbf{R}}, \quad (2.14)$$

where $\mathbf{E}_{\mathbf{R}} = \mathbf{W}^T \mathbf{E} \mathbf{V}$, $\mathbf{A}_{\mathbf{R}} = \mathbf{W}^T \mathbf{A} \mathbf{V}$, $\mathbf{B}_{\mathbf{R}} = \mathbf{W}^T \mathbf{B}$, and $\mathbf{C}_{\mathbf{R}} = \mathbf{V}^T \mathbf{C}$. We similarly define $\mathbf{K}_{\mathbf{R}}(\omega; \mathbf{x}) = \frac{i\omega}{\nu} \mathbf{E}_{\mathbf{R}} - \mathbf{A}_{\mathbf{R}}(\mathbf{x})$.

For an accurate ROM, we must choose appropriate $\mathbf{V}$ and $\mathbf{W}$ to satisfy desired fidelity requirements. A natural choice comes from the following result in interpolatory model reduction [12],

Theorem 2.1. *Suppose $\mathbf{A}(\mathbf{x})$ is continuously differentiable in a neighborhood of $\mathbf{x}_0 \in \mathbb{R}^{n_p}$. Let $\omega \in \mathbb{R}$ and suppose both $\mathbf{K}(\omega; \mathbf{x}_0)$ and $\mathbf{K}_{\mathbf{R}}(\omega; \mathbf{x}_0)$ be invertible. If $(\mathbf{K}(\omega; \mathbf{x}_0))^{-1} \mathbf{B} \subset \text{Range}(\mathbf{V})$ and $(\mathbf{K}(\omega; \mathbf{x}_0))^{-T} \mathbf{C} \subset \text{Range}(\mathbf{W})$, then the reduced model, $\Psi_{\mathbf{R}}$, satisfies*

$$\begin{aligned} \Psi(\omega; \mathbf{x}_0) &= \Psi_{\mathbf{R}}(\omega; \mathbf{x}_0), \\ \nabla_{\omega} \Psi(\omega; \mathbf{x}_0) &= \nabla_{\omega} \Psi_{\mathbf{R}}(\omega; \mathbf{x}_0), \quad \text{and} \\ \nabla_{\mathbf{x}} \Psi(\omega; \mathbf{x}_0) &= \nabla_{\mathbf{x}} \Psi_{\mathbf{R}}(\omega; \mathbf{x}_0). \end{aligned}$$

So, both function and Jacobian evaluations exactly match at any point where these conditions are met. This gives us a recipe for constructing the projection and test spaces. For a set of interpolation points, $\pi_1, \dots, \pi_k$, we construct $\mathbf{V}$ from a (numerically stable) basis for

$$\text{Range} \left([(\mathbf{K}(\omega_1; \pi_1))^{-1} \mathbf{B}, \dots, (\mathbf{K}(\omega_1; \pi_k))^{-1} \mathbf{B}, \dots, (\mathbf{K}(\omega_{n_{\omega}}; \pi_k))^{-1} \mathbf{B}] \right)$$

and we construct $\mathbf{W}$ from a (numerically stable) basis for

$$\text{Range} \left(\left[(\mathbf{K}(\omega_1; \pi_1))^{-T} \mathbf{C}, \dots, (\mathbf{K}(\omega_1; \pi_k))^{-T} \mathbf{C}, \dots, (\mathbf{K}(\omega_{n_{\omega}}; \pi_k))^{-T} \mathbf{C} \right] \right).$$

In general, where these conditions are not satisfied, there is a difference between the reduced-order and the full-order model's function evaluation. The following theorem, a minor extension from [13, Theorem 3.1], relates the error in the reduced frequency response function to linear system residuals for any vector in $\text{Range}(\mathbf{B})$ (the notation has been modified to fit

this paper). This theorem motivates a very efficient approach to control the error between reduced order model and full order model, discussed in section 2.4.2.

Given two subspaces of $\mathbb{C}^n$, say $\mathcal{M}$ and $\mathcal{N}$, we can define the (sine of the) angle¹ between these subspaces, $\Theta(\mathcal{M}, \mathcal{N}) \in [0, \frac{\pi}{2}]$, as

$$\sup_{\mathbf{x} \in \mathcal{M}} \inf_{\mathbf{y} \in \mathcal{N}} \frac{\|\mathbf{y} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sin \Theta(\mathcal{M}, \mathcal{N}).$$

Theorem 2.2. *Given the full order frequency response function, $\Psi(\omega, \mathbf{x})$ and reduced order frequency response function, $\Psi_R(\omega, \mathbf{x})$, the tangential interpolation error at $(\omega, \mathbf{x})$ in the (arbitrary) direction $\mathbf{z} \in \mathbb{R}^{n_{\text{src}}}$ is given by*

$$\|\Psi_R(\omega, \mathbf{x})\mathbf{z} - \Psi(\omega, \mathbf{x})\mathbf{z}\|_2 \leq \left\| \mathbf{C}^T (\mathbf{K}(\omega; \mathbf{x}))^{-1} \right\|_2 \frac{\sin \Theta(\mathcal{C}(\omega, \mathbf{x}), \mathcal{W})}{\cos \Theta(\widehat{\mathcal{B}}(\omega, \mathbf{x}), \mathcal{W})} \|\boldsymbol{\eta}\|_2, \quad (2.15)$$

where²

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{K}(\omega; \mathbf{x})\mathbf{V}\mathbf{t} - \mathbf{B}\mathbf{z} \quad \text{with} \\ \mathbf{t} &= \underset{\tilde{\mathbf{t}} \in \mathbb{R}^{n_r}}{\operatorname{argmin}} \left\| \mathbf{K}(\omega; \mathbf{x})\mathbf{V}\tilde{\mathbf{t}} - \mathbf{B}\mathbf{z} \right\|_2, \end{aligned}$$

$$\widehat{\mathcal{B}}(\omega, \mathbf{x}) = \operatorname{Range}(\mathbf{K}(\omega; \mathbf{x})\mathbf{V}), \mathcal{C}(\omega, \mathbf{x}) = \left(\operatorname{Kernel}(\mathbf{C}^T (\mathbf{K}(\omega; \mathbf{x}))^{-1}) \right)^\perp, \text{ and } \mathcal{W} = \operatorname{Range}(\mathbf{W}).$$

Proof. Note that $\left(\frac{i\omega}{\nu}\mathbf{E} - \mathbf{A}(\mathbf{x})\right)\mathbf{V}\mathbf{t} = \mathbf{B}\mathbf{z} + \boldsymbol{\eta} \in \widehat{\mathcal{B}}(\omega, \mathbf{x})$. The rest of the proof follows identically from the proof of [13, Theorem 3.1]. $\square$

In practice, it is not efficient to compute $\left\| \mathbf{C}^T (\mathbf{K}(\omega; \mathbf{x}))^{-1} \right\|_2$ or the angle between $\mathcal{C}(\omega, \mathbf{x})$ and $\mathcal{W}$ since these are equivalent in work to a full-order function evaluation. However, Theorem 2.2 shows that the error in the function evaluation is bounded by the residuals norms (up to an unknown constant). Hence, to reduce the interpolation error at a point to a given tolerance at minimum cost, that is, with the smallest possible basis extension, we should include the directions with largest tangential interpolation error. We discuss this further in section 2.4.2.

2.2.3 Trust region methods

Robust Newton-type methods use a local model of the objective function to compute a candidate update. This is most commonly done using a line search or a trust region. Here, we

¹Note that this definition is not symmetric unless $\dim \mathcal{M} = \dim \mathcal{N}$.

²Equivalently, $\boldsymbol{\eta}$ is the smallest residual for $\mathbf{K}(\omega; \mathbf{x})\mathbf{x} = \mathbf{B}\mathbf{z}$ with the constraint $\mathbf{x} \in \operatorname{Range}(\mathbf{V})$.

focus on trust region methods [20] and a local model based on the Gauss-Newton approximation. Let $F(\mathbf{x}) = \frac{1}{2} \|\mathbf{R}(\mathbf{x})\|_F^2$ be the objective function, where $\mathbf{R}(\mathbf{x})$ is the nonlinear residual from (2.11). Furthermore, let $\mathbf{J}(\mathbf{x})$ be the Jacobian of the nonlinear residual. Then, given a current approximation $\mathbf{x}_c$, corresponding residual $\mathbf{r}_c$, and Jacobian $\mathbf{J}_c$, the local (Gauss-Newton) model is given by

$$m_{\text{GN}}(\mathbf{x}) = \frac{1}{2} (\mathbf{r}_c^T \mathbf{r}_c) + \mathbf{r}_c^T \mathbf{J}_c (\mathbf{x} - \mathbf{x}_c) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_c)^T \mathbf{J}_c^T \mathbf{J}_c (\mathbf{x} - \mathbf{x}_c). \quad (2.16)$$

To compute a candidate update, this local model is (approximately) minimized over a neighborhood of $\mathbf{x}_c$. We use the regularized trust region method TREGS [23] which uses a truncated SVD combined with a GCV-like condition to compute a candidate update $\mathbf{s} = \mathbf{x} - \mathbf{x}_c$. Trust-region methods subsequently accept the candidate update if the improvement of the objective function is larger than a chosen sufficient improvement predicted by the trust region model, in our case,

$$F(\mathbf{x}_c) - F(\mathbf{x}_c + \mathbf{s}) \geq \rho_k (m_{\text{GN}}(\mathbf{x}_c) - m_{\text{GN}}(\mathbf{x}_c + \mathbf{s})), \quad (2.17)$$

for a chosen parameter ρ_k [20]. To avoid expensive function and Jacobian evaluations, we replace the objective function and its derivatives in the Gauss-Newton model by approximations using an interpolatory ROM [16, 22]. This drastically reduces the computational cost [22] but introduces a number of complications.

First, in general, our parametric ROM does not satisfy the usual requirement in trust-region methods that the Gauss-Newton model and its gradient are exact at the current parameter point. This problem can be remedied by using so-called conditional models [20, Chapter 9][37]. Using conditional models, convergence can be proved if (1) the first order accuracy at the current point can be restored (exact or to sufficient accuracy) in a finite number of model improvements, and (2) for a sufficiently small trust region size, steps that satisfy eq. (2.17) can be guaranteed. Condition (2) is a standard condition for trust-region methods and is satisfied by TREGS. Condition (1) can be satisfied in a single step by adding the current point as an interpolation point to the ROM. While this is always possible, adding an interpolation point is quite expensive, as it requires the solution of a large linear system for each source and frequency combination and an adjoint (transpose) solve for each detector and frequency combination.

However, to determine the need for conditional model improvement requires either (1) a full function evaluation or (2) a global error bound on the ROM approximation [20, 37]. For interpolatory model reduction in a high dimensional parameter space, the cost of providing an error bound over the entire parameter region of interest is, in general, prohibitive [18]. We propose to use, instead, randomized estimates of the objective function to decide when to improve the ROM; we discuss this further in sections 2.3 and 2.4.

2.2.4 Trace Estimation

Recall that for $A \in \mathbb{R}^{m \times n}$,

$$\text{tr}(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_F^2.$$

Therefore, we can estimate the Frobenius norm of $\mathbf{A}$ by estimating the trace of $\mathbf{A}^T \mathbf{A}$.

Let $\mathbf{s} \in \mathbb{R}^n$ be a vector with independent, identically distributed entries with mean zero and unit variance. Then, we have

$$\begin{aligned} \mathbb{E} [\|\mathbf{A}\mathbf{s}\|_2^2] &= \mathbb{E} [\mathbf{s}^T \mathbf{A}^T \mathbf{A} \mathbf{s}] = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{A}^T \mathbf{A})_{ij} \mathbb{E} [s_i s_j] \\ &= \sum_{i=1}^n (\mathbf{A}^T \mathbf{A})_{ii} \mathbb{E} [s_i^2] + \sum_{\substack{i=1, j=1 \\ i \neq j}}^n (\mathbf{A}^T \mathbf{A})_{ij} \mathbb{E} [s_i s_j]. \end{aligned}$$

Since s_i and s_j are independent and have mean zero, $\mathbb{E} [s_i s_j] = \mathbb{E} [s_i] \mathbb{E} [s_j] = 0$. Since s_i also has unit variance, $\mathbb{E} [s_i^2] = 1$. Thus,

$$\mathbb{E} [\|\mathbf{A}\mathbf{s}\|_2^2] = \text{tr}(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_F^2.$$

By selecting s_i from the Rademacher distribution $\left[P(s_i = x) = \begin{cases} 1/2 & x = 1 \\ 1/2 & x = -1 \end{cases} \right]$, we obtain the Hutchinson trace estimator which minimizes the variance of the sample [28].

2.3 Estimating ROM Accuracy and Guiding Updates

Since the objective function eq. (2.11) involves the Frobenius norm, we can use the Hutchinson trace estimator to estimate the objective function. This offers significant computational savings over the full function evaluation. Consider the evaluation of the following estimate. Given a sample $\mathbf{s} \in \mathbb{R}^{n_{\text{src}}}$ drawn from the Rademacher distribution,

$$\begin{aligned} (\Psi(\omega_j; \mathbf{x}) - \mathbf{D}_j) \mathbf{s} &= (\mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} \mathbf{B} - \mathbf{D}_j) \mathbf{s} \\ &= \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} (\mathbf{B}\mathbf{s}) - \mathbf{D}_j \mathbf{s}. \end{aligned}$$

Whereas the (full-order) objective function evaluation requires solving $\min(n_{\text{det}}, n_{\text{src}})$ large linear systems per frequency, this trace estimate requires the solution of only 1 large system per sample and per frequency. While other methods use this estimate directly in the optimization [9, 27], we instead use it only to assess the quality of our reduced order model. In particular, we only need the estimate to indicate whether there is a significant difference between the reduced-order objective function and the full-order objective function evaluation.

Equipped with an efficient method to estimate the difference between the full-order and reduced-order objective function evaluations, we combine the ROM with our optimization routine, and we update the ROM if our estimates suggest a significant difference. For an iteration index k , let $F_R^{(k)}(\mathbf{x})$ denote the objective function computed using the reduced order model generated from $\mathbf{V}^{(k)}$ and $\mathbf{W}^{(k)}$, $F_E(\mathbf{x})$ denote the trace estimate of $F(\mathbf{x})$, and $d(F_R^{(k)}(\mathbf{x}), F_E(\mathbf{x}))$ denote our metric for model quality. Then given a starting point $\mathbf{x}_0$, initial projection bases $\mathbf{V}^{(0)}, \mathbf{W}^{(0)}$, optimizer tolerance tol_o , model quality tolerance tol_q , and an update strategy (discussed in section 2.4), we may apply Algorithm 1.

Algorithm 1: ROM-Based Optimization with Estimate-Driven Updates

```

1.1  $\mathbf{x}_c \leftarrow \mathbf{x}_0, f_c \leftarrow F_R^{(0)}(\mathbf{x}_c), k \leftarrow 0$ 
1.2 while  $f_c \geq \text{tol}_o$  do
1.3   Compute trust region candidate point,  $\mathbf{x}_p$ , using  $F_R^{(k)}$ .
1.4   Set  $f_p \leftarrow F_R^{(k)}(\mathbf{x}_p), f_{\text{est}} \leftarrow F_E(\mathbf{x}_p)$ .
1.5   if  $d(f_p, f_{\text{est}}) \geq \text{tol}_q$  then
1.6     Reject  $\mathbf{x}_p$ 
1.7     Update  $\mathbf{V}^{(k+1)}, \mathbf{W}^{(k+1)} \leftarrow \text{update}(\mathbf{V}^{(k)}, \mathbf{W}^{(k)}, \mathbf{x}_c, \mathbf{x}_p)$ 
1.8     Set  $f_c \leftarrow F_R^{(k+1)}(\mathbf{x}_c), k \leftarrow k + 1$ 
1.9   else
1.10    if Trust Region Algorithm Accepts  $\mathbf{x}_p$  then
1.11     Set  $\mathbf{x}_c \leftarrow \mathbf{x}_p, f_c \leftarrow f_p$ 

```

In this paper: tol_o is set to 1.1 times the noise level (based on the discrepancy principle), TREGS [23] is used as the optimization algorithm, and the reduced order model is rejected when $F_E(\mathbf{x})/F_R^{(k)}(\mathbf{x}) \geq 10$ at the proposed point.

2.4 Basis Updates

When $d(f_p, f_{\text{est}}) \geq \text{tol}_q$, it is necessary to update the projection bases. Below, we examine two possibilities: (1) using the interpolatory conditions from theorem 2.1 to make the reduced-order model exact at either the proposed or current point; (2) coupling the residual-based error bounds from theorem 2.2 with techniques from subspace recycling [3, 4, 31] to add directions corresponding to large residuals (in the sense of Theorem 2.2). We discuss performance comparisons in section 2.6.

2.4.1 Full Interpolatory Updates

When the objective function estimate suggests that the reduced and full order models differ significantly, we can use theorem 2.1 to make the reduced order model match the full order model at that point. This has the cost of evaluating $(\mathbf{K}(\omega; \mathbf{x}))^{-1} \mathbf{B}$ and $(\mathbf{K}(\omega; \mathbf{x}))^{-T} \mathbf{C}$ for all frequencies ω . In terms of large system solves, this at least the cost of 2 (full-order) function evaluations. In exchange, the reduced order model will match the full order model to first order at the given point. If we use the current point to update, this guarantees that the optimizer will choose the same step direction as the full order model.

As discussed in section 2.2.3, the conditional model framework can be used to *suggest* convergence. While interpolatory updates can be used to make the reduced-order model exact in a single step, the error estimates available through section 2.2.4 are only probabilistic. We believe the “relaxed first-order condition” from [37] can be extended to a probabilistic setting, and we will explore this in a future paper.

While theorem 2.1 suggests adding (in general) $n_\omega \cdot (n_{\text{det}} + n_{\text{src}})$ vectors when an update is needed, our experiments suggest that these full updates are too expensive and contain a significant amount of redundant information. For example, in our 3D experiments (section 2.6.2), theorem 2.1 implies that $1 \cdot (961 + 961) = 1922$ vectors (and thus large linear solves) are sufficient for each update. However, at the first rejection point, only 170 basis vectors are added after a rank revealing factorization, while the remainder are (numerically) linearly dependent. This indicates that there is a much smaller bases expansion that will produce a reduced order model of essentially the same accuracy.

2.4.2 Residual Driven Updates

To reduce the cost of updating the model, we turn to theorem 2.2 for motivation. This theorem indicates that reducing the norm of the residual in a given direction will reduce the interpolation error in that direction. More directly, we bound the interpolation error in $\mathbf{R}(\mathbf{x})$ as follows.

Theorem 2.3. *Let $\mathbf{T}_j = \mathbf{K}(\omega_j; \mathbf{x}) \mathbf{V} (\mathbf{K}_R(\omega_j; \mathbf{x}))^{-1} \mathbf{W}^T$, i.e. the projection onto $\text{Range}(\mathbf{K}(\omega_j; \mathbf{x}) \mathbf{V})$ along $\mathbf{W}$, and define $\boldsymbol{\eta}_j = (\mathbf{I} - \mathbf{T}_j) \mathbf{B}$. Define $\mathbf{H} = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_{n_\omega}]$. Then*

$$| \|\mathbf{R}(\mathbf{x})\|_F - \|\mathbf{R}_R(\mathbf{x})\|_F | \leq \kappa(\mathbf{x}) \|\mathbf{H}\|_F,$$

where κ does not depend on $\mathbf{V}, \mathbf{W}$.

Proof.

$$\begin{aligned}
| \|\mathbf{R}(\mathbf{x})\|_F - \|\mathbf{R}_R(\mathbf{x})\|_F |^2 &= | \|\mathbf{O}(\mathbf{x}) - \mathbf{D}\|_F - \|\mathbf{O}_R(\mathbf{x}) - \mathbf{D}\|_F |^2 \\
&\leq \|\mathbf{O}(\mathbf{x}) - \mathbf{O}_R(\mathbf{x})\|_F^2 \\
&= \sum_{j=1}^{n_\omega} \left\| \mathbf{C}^T [(\mathbf{K}(\omega_j; \mathbf{x}))^{-1} - \mathbf{V}(\mathbf{K}_R(\omega_j; \mathbf{x}))^{-1} \mathbf{W}^T] \mathbf{B} \right\|_F^2 \\
&= \sum_{j=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} [\mathbf{I} - \mathbf{T}_j] \mathbf{B} \right\|_F^2 \\
&= \sum_{j=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} \boldsymbol{\eta}_j \right\|_F^2 \\
&\leq \sum_{j=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} \right\|_F^2 \|\boldsymbol{\eta}_j\|_F^2 \\
&\leq \left(\max_{j=1, \dots, n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} \right\|_F \right)^2 \sum_{j=1}^{n_\omega} \|\boldsymbol{\eta}_j\|_F^2 \\
&= \left(\max_{j=1, \dots, n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_j; \mathbf{x}))^{-1} \right\|_F \right)^2 \|\mathbf{H}\|_F^2 \\
&= (\kappa(\mathbf{x}) \|\mathbf{H}\|_F)^2.
\end{aligned}$$

Taking square roots yields the desired bound. $\square$

While evaluating the condition number, $\kappa(\mathbf{x})$, is equivalent in cost to evaluating the full order model (and thus undesirable), we use the objective function estimates from section 2.3 to estimate the reduction in $\|\mathbf{H}\|_F$ necessary. Since $\kappa(\mathbf{x})$ does not depend on $\mathbf{V}$ or $\mathbf{W}$, we are free to expand these projection bases to reduce the residual norm, $\|\mathbf{H}\|_F$.

For each ω_j , we compute the (thin) QR factorization $\mathbf{K}(\omega_j; \mathbf{x})\mathbf{V} = \mathbf{Q}_j \hat{\mathbf{R}}_j$. Thus, $\mathbf{Q}_j \mathbf{Q}_j^T$ is the orthogonal projector onto $\text{Range}(\mathbf{K}(\omega_j; \mathbf{x})\mathbf{V})$. From this, we have (dropping the index j and $\mathbf{x}$ dependence for brevity):

$$\begin{aligned}
\boldsymbol{\eta} &= \mathbf{Q}\mathbf{Q}^T \boldsymbol{\eta} + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \boldsymbol{\eta} \\
&= \mathbf{Q}\mathbf{Q}^T (\mathbf{I} - \mathbf{T}) \mathbf{B} + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) (\mathbf{I} - \mathbf{T}) \mathbf{B} \\
&= \underbrace{(\mathbf{Q}\mathbf{Q}^T - \mathbf{T}) \mathbf{B}}_{\subset \text{Range}(\mathbf{K}\mathbf{V})} + \underbrace{(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}}_{\subset \text{Range}(\mathbf{K}\mathbf{V})^\perp}.
\end{aligned}$$

The first term in the above decomposition represents the difference between the skew projector $\mathbf{K}\mathbf{V}(\mathbf{K}_R)^{-1} \mathbf{W}^T$ and the orthogonal projector $\mathbf{K}\mathbf{V} \hat{\mathbf{R}}^{-1} \mathbf{Q}^T = \mathbf{Q}\mathbf{Q}^T$. Since this term is already in $\text{Range}(\mathbf{K}\mathbf{V})$, we focus on $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}$ and in particular, we add vectors to $\mathbf{V}$ from $\mathbf{K}^{-1} \text{Range}((\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B})$.

To minimize the number of basis vectors added (and thus the number of large linear solves), we use the singular value decomposition of $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}$ to determine the most significant components. Let $\mathbf{U} \mathbf{Y}^T = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}$ be a singular value decomposition. We expand $\mathbf{V}$ with $[\mathbf{K}^{-1}\mathbf{u}_1, \dots, \mathbf{K}^{-1}\mathbf{u}_r]$ for $r \ll n_{\text{src}}$ that sufficiently reduces $\|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{B}\|$. The Schmidt-Eckart-Young Theorem[26] shows that $\sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{y}_i^T$ is the optimal rank- r approximation in both the Frobenius and spectral norms. The relative error in this low-rank approximation in the Frobenius norm is given by $\sqrt{\frac{\sum_{i=r+1}^{n_{\text{src}}} \sigma_i^2}{\sum_{i=1}^{n_{\text{src}}} \sigma_i^2}}$. If we choose r such that this relative error is less than some ϵ (e.g. $\epsilon = \frac{1}{10}, \frac{1}{20}, \frac{1}{100}$, etc.) and compute the QR factorization $\mathbf{K}[\mathbf{V}, \mathbf{K}^{-1}\mathbf{u}_1, \dots, \mathbf{K}^{-1}\mathbf{u}_r] = \tilde{\mathbf{Q}}\tilde{\mathbf{R}}$, then it follows that

$$\left\| \left[\mathbf{I} - \tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T \right] \mathbf{B} \right\|_F \leq \epsilon \left\| \left[\mathbf{I} - \mathbf{Q}\mathbf{Q}^T \right] \mathbf{B} \right\|_F \leq \epsilon \|\boldsymbol{\eta}\|_F.$$

We expect³ to see a similar factor ϵ improvement in $\|\boldsymbol{\eta}_j\|_F$ and hence in the interpolation error between the reduced-order and full-order models. In our experience, this is the case. This procedure is repeated for each ω_i to compute the updates to $\mathbf{V}$.

Theorem 2.4. Let $\boldsymbol{\xi}_j = [\mathbf{I} - \mathbf{V}(\mathbf{K}_R(\omega_j; \mathbf{x}))^{-1} \mathbf{W}^T \mathbf{K}(\omega_j; \mathbf{x})]^T \mathbf{C}$ and define $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n_\omega}]$. Then

$$\left| \|\mathbf{R}(\mathbf{x})\|_F - \|\mathbf{R}_R(\mathbf{x})\|_F \right| \leq \mu(\mathbf{x}) \|\boldsymbol{\xi}\|_F,$$

where μ does not depend on $\mathbf{V}, \mathbf{W}$.

The proof is nearly identical to that of theorem 2.3 and a similar decomposition in terms of $\text{Range}(\mathbf{K}^T \mathbf{W})$ and its orthogonal complement yield an update scheme for $\mathbf{W}$.

Example

To illustrate the usefulness of this approach, we consider the first rejected step in the 2D 1-point ROM experiment in section 2.6. As Figure 2.3 shows, the relative residual norm for each of the 32 columns of $\mathbf{B}$ (blue circles) is $O(10^{-4})$. However, the relative error in the objective function, $\frac{\|F_R^{(0)}(\mathbf{x}_p) - F(\mathbf{x}_p)\|}{\|F(\mathbf{x}_p)\|}$, is approximately 11.7%. The singular values of $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}$ decay rapidly as shown in Figure 2.2a. Therefore, the residual norm can be sufficiently decreased with the addition of very few singular vectors. As seen in Figure 2.2b, only two linear solves are needed for a 90% reduction in relative error, 3 linear solves for a 95% reduction, and 5 linear solves for a 99% reduction.

With the addition of the three singular vectors corresponding to the largest three singular values of $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T) \mathbf{B}$ (requiring three large linear solves), the relative residual norms are

³Note that this is not guaranteed. The error term corresponding to the skew projection may increase in norm. However, the second term dominates the residual norm in our experiments. An analysis of how much the skew projection term can deteriorate is future work.

significantly reduced (orange circles in Figure 2.3) and the relative error in the objective function is reduced to approximately 0.2%.

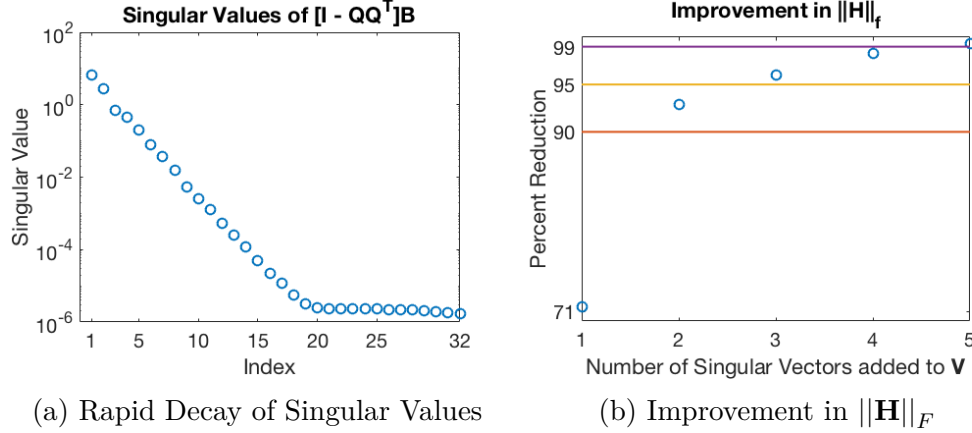


Figure 2.2: Improvement in Error Bound at First Rejected Step in 2D - 1 Point Experiment.

2.5 Probabilistic Estimates

Since each sample of the stochastic estimates of section 2.3 requires the solution of one large linear system per frequency, selecting an appropriate number of samples is key to the efficiency and robustness of our proposed technique. A small number of samples increases the variance of the estimate which could trigger unnecessary updates or fail to detect a large difference between the reduced- and full-order objective functions. A large number of samples makes the algorithm more robust but requires solving more large linear systems. A detailed discussion on this efficiency-robustness trade-off is beyond the scope of this paper, but using the analysis outlined in [10] and improved upon in [34], we provide rigorous bounds on the number of samples required to detect a sufficiently large under-estimate of the objective function with a given probability.

We focus our attention on under-estimates of the objective function because a sufficiently large over-estimate will simply trigger an update to the model. While this would impact the efficiency of our method, over-estimates do not degrade the robustness. In contrast, successive under-estimates of the objective function could prevent updates and degrade our algorithm to traditional ROM-based parameter inversion. Therefore, we want to explore the number of samples required to make large under-estimates sufficiently unlikely.

To bound the number of samples required, we use an intermediate result from [34] which says that given $N \geq 6\epsilon^{-2} \log(1/\delta)$ samples of the Hutchinson estimator, denoted tr_H^N , $P(\text{tr}_H^N(\mathbf{A}) \leq (1 - \epsilon) \text{tr}(\mathbf{A})) < \delta$. That is, the probability of under-estimating the trace of $\mathbf{A}$ by a relative factor of ϵ is less than δ .

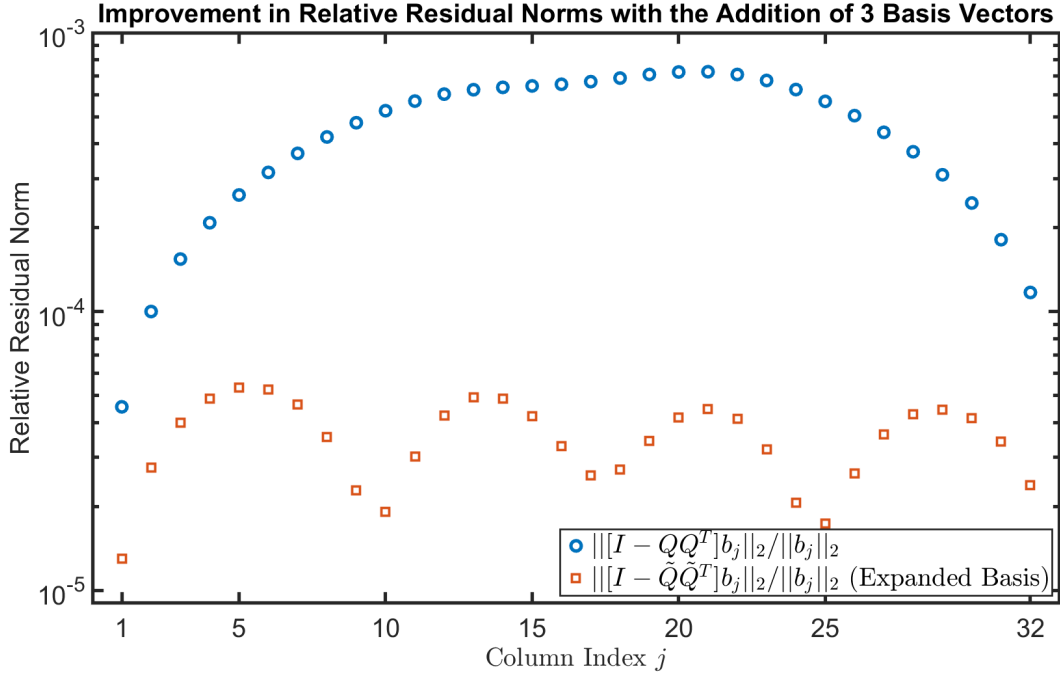


Figure 2.3: Improvement in Relative Residual Norms

Suppose we want to reject the reduced order model when $F_R^{(k)}(\mathbf{x}) > \alpha F(\mathbf{x})$ for some $\alpha > 1$ with high probability and our rejection threshold is set to $\frac{\alpha}{K}$ for some $K > 1$, that is, we reject the reduced order model if $F_E(\mathbf{x}) > \frac{\alpha}{K} F_R^{(k)}(\mathbf{x})$. Let $\beta = \frac{F(\mathbf{x})}{F_R^{(k)}(\mathbf{x})}$. Then our model is rejected if

$$F_E(\mathbf{x}) \geq \frac{\alpha}{K} F_R^{(k)}(\mathbf{x}) = \frac{\alpha}{K\beta} F(\mathbf{x}) = \left(1 - \left(1 - \frac{\alpha}{K\beta}\right)\right) F(\mathbf{x}).$$

This will occur with probability at least $1 - \delta$ given $N \geq 6\epsilon^{-2} \log(1/\delta)$ samples where $\epsilon = \left(1 - \frac{\alpha}{K\beta}\right)$. While β is not typically available to us without a full function evaluation, we are interested in detecting cases where $\beta \geq \alpha$. In such cases, the minimum ϵ occurs at $\alpha = \beta$ with $\epsilon = \left(1 - \frac{1}{K}\right)$.

By choosing K sufficiently large, we may detect such cases while requiring a modest number of samples. As an example, if $K = 2$, then only $\lceil \frac{6}{(1-\frac{1}{2})^2} \log(2) \rceil = 17$ samples are required to guarantee a rejection probability, δ , of at least $1/2$. Increasing the number of samples to 34 yields a rejection probability of at least $3/4$. Thus, by reducing the rejection threshold by a factor $\frac{1}{K}$ to provide a buffer and using a modest number of samples, we obtain strong lower bounds on the rejection probabilities.

2.6 Numerical Experiments

To demonstrate the value of our proposed techniques, we present a series of numerical experiments. To quantify the cost of each method, the experiments are evaluated by the number of large (i.e., the size of the full-order model) linear solves needed to drive the residual norm to 1.1 times the noise level. While there are computational costs for the update procedures (such as an SVD for the updates of section 2.4.2), these involve linear algebra on systems multiple orders of magnitude smaller than the full-order model (e.g., $O(10^3)$ vs $O(10^5)$). Therefore, the dominant computational cost will be the large linear solves. Since the full-order systems require iterative solvers to be efficiently solved, the computational cost is (roughly) linear in the number of large solves.

We examine reconstructions for a two-dimensional and a three-dimensional tomography problem. For each, we compare the cost of reconstruction using the full-order model, using a reduced order model constructed by interpolating the first three full-order optimization steps (henceforth labeled 3-point), and using a reduced order model interpolating only the initial condition (henceforth labeled 1-point). For the reduced order models, we explore the accuracy and cost without updates, and we compare the resulting efficiency of the (exact) interpolatory updates of section 2.4.1 with the residual-driven updates of section 2.4.2.

These 1-point models are of particular interest because the same initial condition is (typically) used for each reconstruction. Thus, the solutions of $(\mathbf{K}(\omega; \mathbf{x}))^{-1} \mathbf{B}$ and $(\mathbf{K}(\omega; \mathbf{x}))^{-T} \mathbf{C}$ from this initial condition can be computed off-line and reused for any reconstruction based on the same mesh, source/detector location, and frequencies. This *significantly* reduces the number of large-scale solves, especially for large problems. The number of large linear solves required beyond the off-line solves is henceforth referred to as the amortized cost.

The numerical experiments are constructed as follows. Synthetic data is generated by constructing a target anomaly⁴ in the 0-1 pixel basis. Background pixels are mapped to a background absorption level, μ_{out} , while pixels belonging to the anomaly are mapped to a higher absorption value, μ_{in} . A small inhomogeneity is added to the anomaly and the background and then measurements are then generated by evaluating the forward problem and adding white noise.

For these experiments, we use a symmetric, one-sided projection, i.e. $\mathbf{V} = \mathbf{W}$, and a single frequency, corresponding to $\omega = 0$.

2.6.1 2D Example

The 2D problem is evaluated on a 201×201 mesh, yielding $n = 40,401$ degrees of freedom for the large system. The standard centered finite difference scheme is used to discretize

⁴These anomalies are not exactly reconstructable in the PaLS basis, avoiding the “inverse crime” of using the same model for data generation and reconstruction.

eqs. (2.1) to (2.4). We use 32 sources and 32 detectors arranged as shown in Figure 2.1. The true image and initial condition are shown in Figure 2.4. We use 25 compactly-supported radial basis functions (CSRBF) for our PaLS reconstruction, yielding 100 (four per CSRBF) parameters to be estimated. The absorption values, μ_{out} and μ_{in} , are set to 0.005 and 0.15, respectively. White noise is added to the measurements at a level of 1 permille. To quantify the impact of the random nature of the trace estimates of section 2.5, each experiment is repeated 103 times⁵ and the minimum, 25th, 50th, 75th percentiles, and maximum number of large linear solves are reported in table 2.1. Figure 2.7 shows the reconstructions for a representative noise realization. Figure 2.5 details the convergence history of the 3-point ROM without updates and the 3-point ROM with the (exact) interpolatory updates of section 2.4.1.

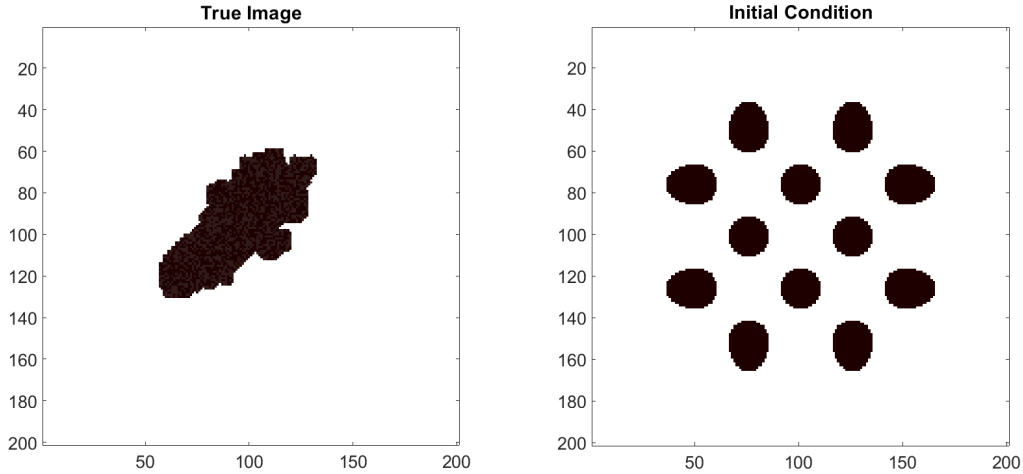


Figure 2.4: True Image for the 2D Model and Initial Condition

Choice of Update Location

As noted in section 2.4, there is a choice of updating using the currently accepted point, $\mathbf{x}_c$, the proposed step $\mathbf{x}_p$, or both if the current step is rejected.

⁵A number of the form $4k + 3$ was chosen to ensure that the median, 25th, and 75th percentiles were elements of the set.

After the model is updated, algorithm 1 returns to $\mathbf{x}_c$ to compute a new function evaluation and Jacobian. If interpolatory updates (section 2.4.1) are used to update at $\mathbf{x}_c$, then this function evaluation and Jacobian will be exact (identical to those from the full-order model) according to theorem 2.1. This guarantees that the trust region search direction for the reduced-order model is identical to the direction chosen for the full-order model at that point.

However, table 2.1 demonstrates that using updates at $\mathbf{x}_c$, in all but one setup, requires more large solves in the median minimization than using updates at $\mathbf{x}_p$. We conjecture that this occurs for the following reason. In order for $\mathbf{x}_c$ to have been accepted previously, the reduced-order model evaluated there, $F_R^{(k)}(\mathbf{x}_c)$, is likely to be sufficiently close to the estimate $F_E(\mathbf{x}_c)$. Similarly, for an update to occur, the reduced-order model at the proposed point, $F_R^{(k)}(\mathbf{x}_p)$, and the estimate, $F_E(\mathbf{x}_p)$, are likely to be sufficiently different. Thus, we expect that the error between the full-order and reduced-order models is larger at $\mathbf{x}_p$ than at $\mathbf{x}_c$. This suggests that updating at $\mathbf{x}_p$ will yield more improvement in the reduced order model.

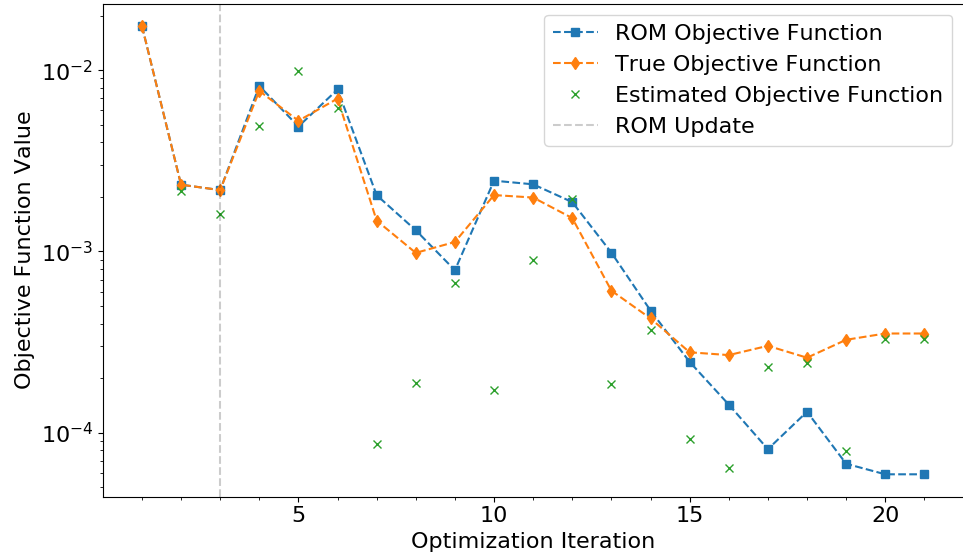
Since updating the reduced-order model at $\mathbf{x}_p$ required an equal number or fewer large linear solves in the majority of our experiments, with nearly identical quality results, we use updates at $\mathbf{x}_p$.

Results

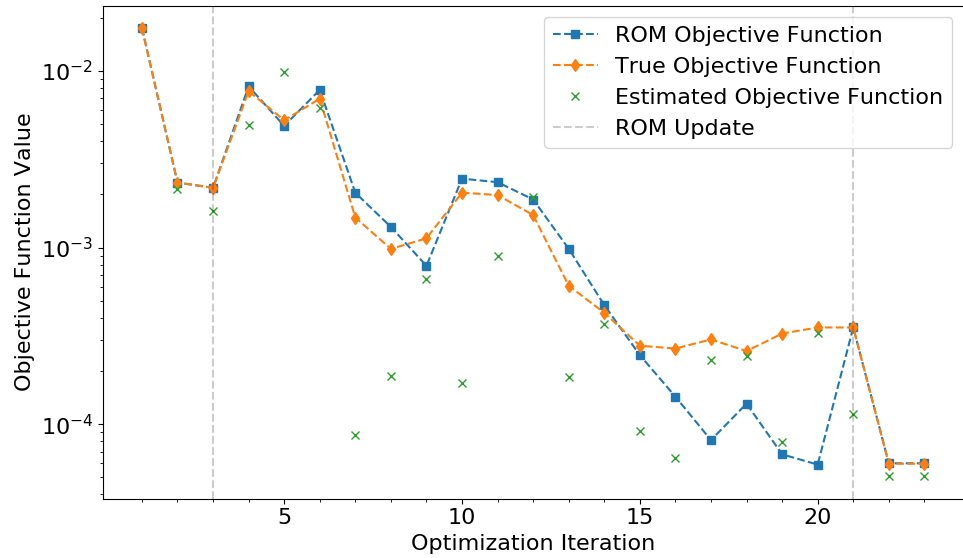
From table 2.1, we see that all the ROM-based methods result in a significant reduction in the number of large solves compared with the full order model. However, while the reconstruction using the 3-point ROM without updates (Figure 2.7) is visually similar to that generated using the full-order model, Figure 2.5a shows that the (full-order) objective function is approximately 10 times larger than the 3-point reduced objective function at the final iteration. As seen in Figure 2.5b, this discrepancy is detected by the probabilistic estimates at iteration 20, and the ROM is updated. For the remaining iterations, the ROM objective function closely matches the (full-order) objective function, and the optimizer terminates with the full-order objective function within the desired tolerance.

Thus, our error detection and update procedures are successful. However, the interpolatory update scheme incurs a non-trivial cost, approximately 45% more solves in the median case or nearly 68% more solves in the median amortized case. By using the residual-driven update scheme of section 2.4.2, the costs are reduced to approximately 14% more solves in the median case or approximately 21% more solves in the median amortized case. We note that these extra costs do provide an increase in robustness.

Despite using a significantly smaller initial reduced-order model, the 1-point models are able to produce reconstructions of a similar quality to the 3-point models (as seen in Figure 2.7) at the cost of multiple updates, but still with a significant reduction in number of large linear



(a) No Updates



(b) Interpolatory Updates

Figure 2.5: Convergence behavior of 3-point ROM without updates and 3-point ROM with Interpolatory Updates for a representative realization of the 2D experiment

Table 2.1: Number of Large Linear Solves and Relative Residual Error for the 2D Problem

Experiment Setup	Median Solves [Amortized]	Relative Error (%) $ F_R(\mathbf{x}) - F(\mathbf{x}) / F(\mathbf{x})$
Full Order Model	992 [928]	0
ROM - 3 Point - Interpolatory Updates at $\mathbf{x}_c$	279 [215]	2.5
ROM - 3 Point - Interpolatory Updates at $\mathbf{x}_p$	279 [215]	3.6
ROM - 3 Point - No Updates	192 [128]	1131.1
ROM - 3 Point - Residual-Driven Updates at $\mathbf{x}_c$	219 [155]	5.4
ROM - 3 Point - Residual-Driven Updates at $\mathbf{x}_p$	218 [154]	2.8
ROM - 1 Point - Interpolatory Updates at $\mathbf{x}_c$	302 [238]	13.9
ROM - 1 Point - Interpolatory Updates at $\mathbf{x}_p$	240 [176]	43.7
ROM - 1 Point - Residual-Driven Updates at $\mathbf{x}_c$	138 [74]	15.2
ROM - 1 Point - Residual-Driven Updates at $\mathbf{x}_p$	132 [68]	16.6

solves. A sample convergence history with three updates is seen in Figure 2.6. While using interpolatory update scheme with the 1-point models does require fewer large linear solves than the interpolatory update scheme used with the 3-point models, the residual-driven update scheme offers a significantly larger reduction compared with all other methods. The residual-driven update scheme required approximately 31% *fewer* large linear solutions in the median case and approximately 47% *fewer* solves in the median amortized case (that is, where the solutions from the initial conditions are re-used between optimizations).

2.6.2 3D

The 3D problem is evaluated on a $64 \times 64 \times 64$ mesh, yielding $n = 262,144$ degrees of freedom for the large system. The standard centered finite difference scheme is used to discretize eqs. (2.1) to (2.4). We use 961 sources and 961 detectors arranged as shown in Figure 2.1. The true image (before adding the random inhomogeneities) is shown in Figure 2.8. Note that this image is not exactly reconstructable using our choice of parameterization. We use 27 compactly-supported radial basis functions (CSRBF) for our PaLS reconstruction,

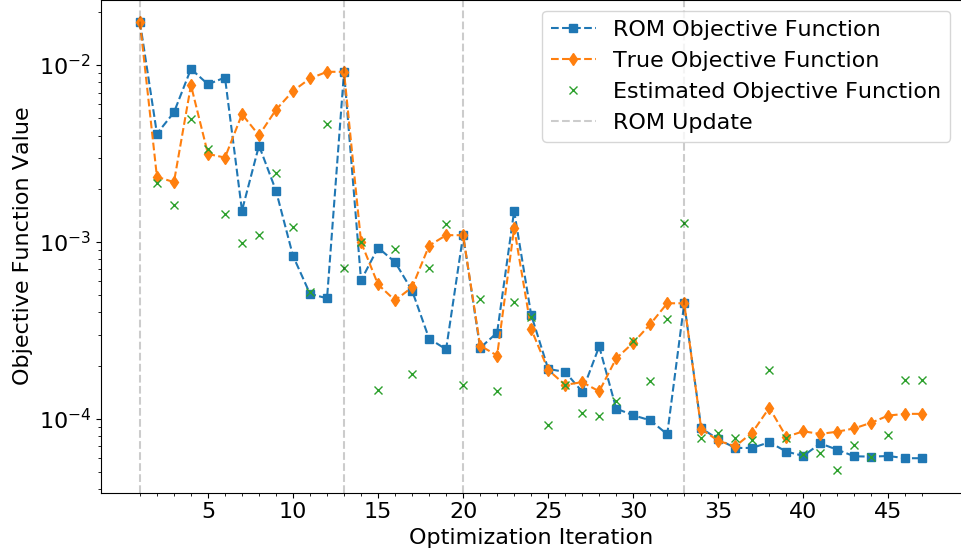


Figure 2.6: Convergence history of the 1-point ROM with Residual-Driven updates for a representative realization of the 2D experiment

yielding 135 (five per CSRBF) parameters. As in section 2.6.1, the absorption values, μ_{out} and μ_{in} , are set to 0.005 and 0.15, respectively. White noise is added to the measurements at a level of 1 permille. To quantify the impact of the random nature of the trace estimates of section 2.5, each experiment is repeated 15 times and the minimum, 25th, 50th, 75th percentiles, and maximum number of large linear solves are reported in table 2.2. Figure 2.9 shows the reconstructions for a representative noise realization⁶.

Results

The 3D results are very similar to the 2D results discussed in section 2.6.1, but the improvements are even more striking. The 3-point models do not require updates in most cases, resulting in a small increase in the number of large linear solutions required due to the objective function estimation. However, there are significant savings when using the 1-point reduced-order model. Even with interpolatory updates, there is a reduction of approximately 32% in the median case and a reduction of approximately 49% in the median amortized case. Using residual-driven updates results in a reduction of approximately 66% in the median case and a reduction of approximately 98% in the median amortized case. Thus, in the amortized case, our proposed technique requires a factor 50 fewer large linear solutions than the 3-point reduced-order model and results in a more than 300-fold reduction compared with the full-order model.

⁶Note that the anomaly is correctly localized in each case. Using additional frequencies may resolve the shape more finely.

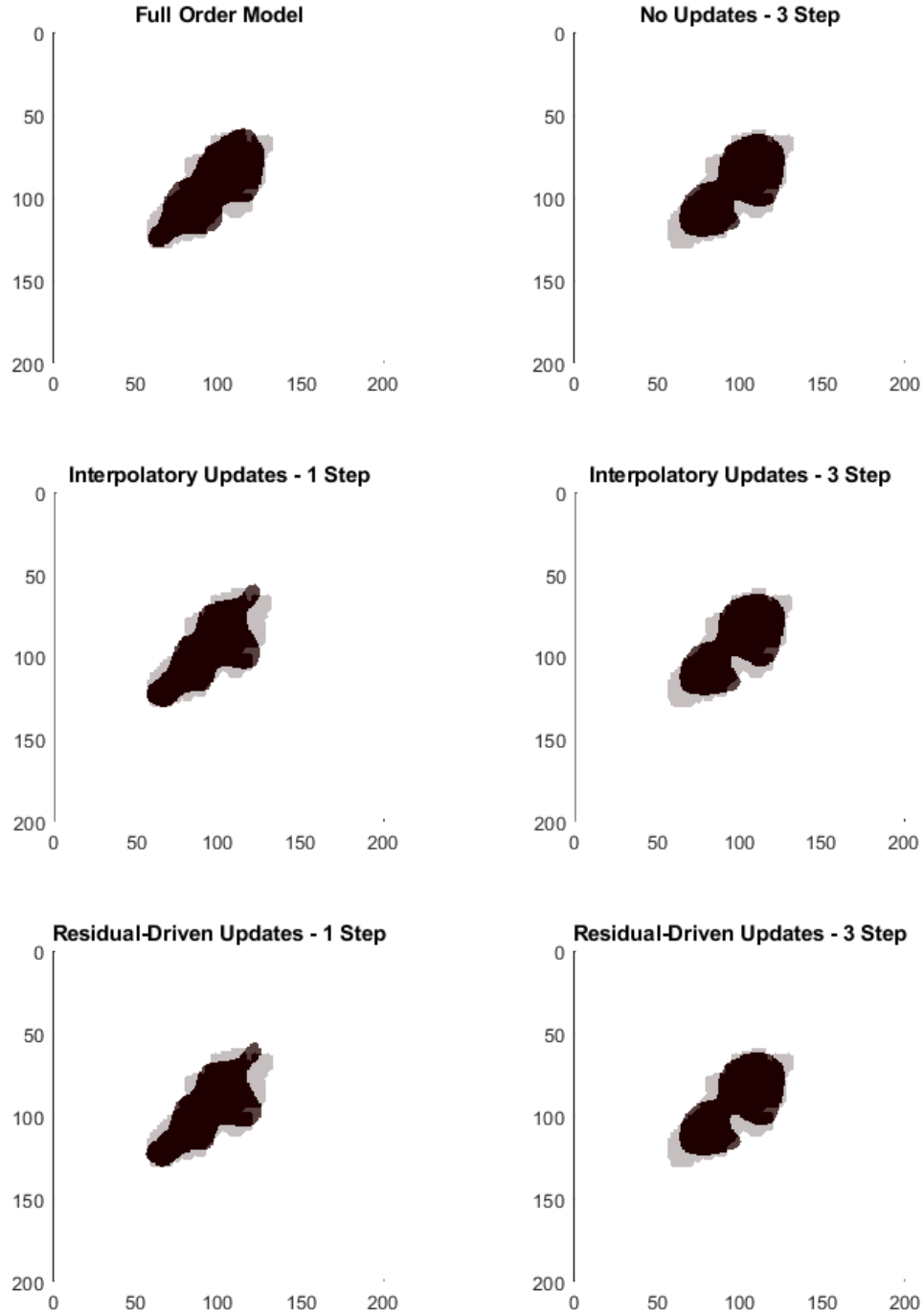
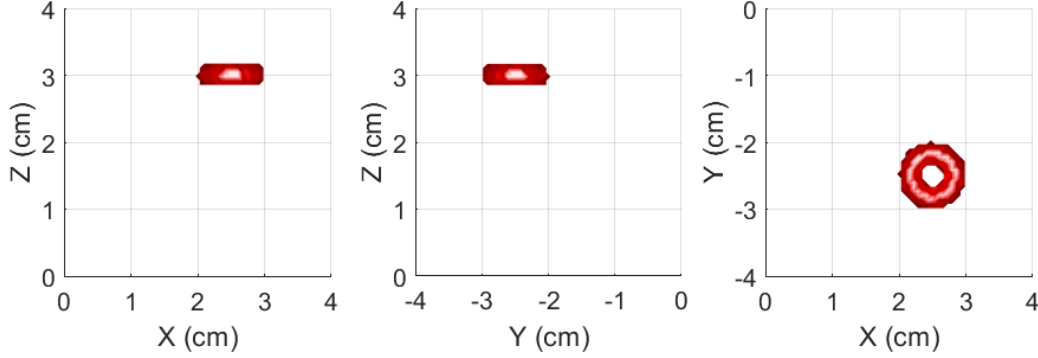
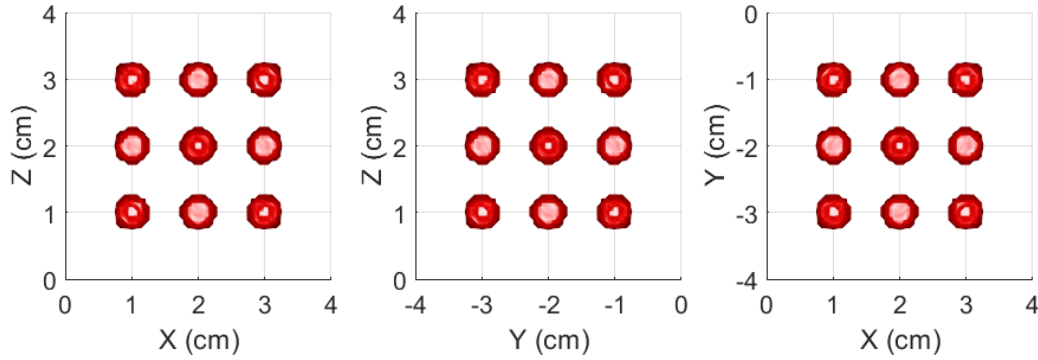


Figure 2.7: 2D Reconstruction Results



(a) Truth Image



(b) Initial Condition

Figure 2.8: True Image for the 3D Model and Initial Condition

2.7 Conclusions and Future Work

We have shown that the use of stochastic trace estimation enables the estimation of reduced order model quality for the DOT problem. This allows us to adaptively update the model during the optimization. With a modest cost increase, a larger initial reduced-order model can use these estimates for robustness. We have also presented an update scheme for projection-based model reduction that is able to significantly reduce the relative error in the objective function at the cost of just a handful of large linear solutions compared with an interpolatory update ($O(10)$ vs $O(n_{\text{det}} + n_{\text{src}})$). By starting with a lower order initial reduced-order model, using stochastic estimates to guide adaptivity, and updating using our residual-driven scheme, we obtain a method that is very efficient and robust.

The use of stochastic estimates and updates allows the user to choose a balance of efficiency

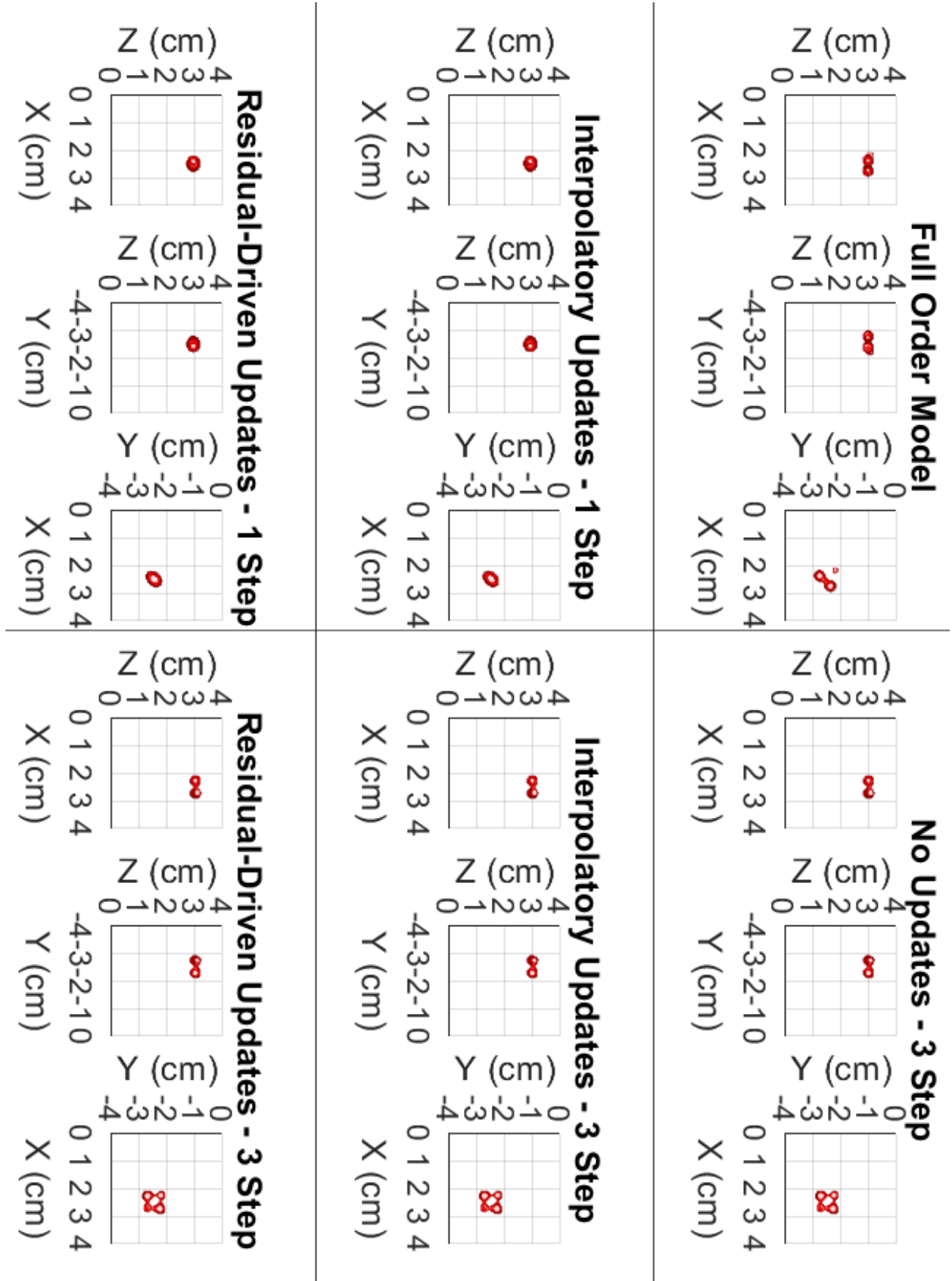


Figure 2.9: 3D Reconstruction Results

Table 2.2: Number of Large Linear Solves for the 3D Problem

Experiment Setup	Median Solves [Amortized]
Full Order Model	23064 [21142]
ROM - 3 Point - Interpolatory Updates at $\mathbf{x}_p$	5776 [3854]
ROM - 3 Point - No Updates	5766 [3844]
ROM - 3 Point - Residual-Driven Updates at $\mathbf{x}_p$	5776 [3854]
ROM - 1 Point - Interpolatory Updates at $\mathbf{x}_p$	3871 [1949]
ROM - 1 Point - Residual-Driven Updates at $\mathbf{x}_p$	1985 [63]

and robustness. For the purposes of this paper, we choose to focus on efficiency, but this choice may vary on a per-application basis. To reduce the cost of sampling, the stochastic estimates could be performed only at accepted (by the trust-region) steps instead at every trial point. This may require more accurate estimates and hence requires balancing between fewer sampling locations and more samples per location. In section 2.6.1, experimental evidence suggests updating at $\mathbf{x}_p$ was usually more efficient than updating at $\mathbf{x}_c$, but further analysis is necessary. Furthermore, extending the projection bases using solutions from both points could reduce the number of updates required at the cost of more expensive updates. In our work, we use the same number of samples for each objective function estimate. Robustness could be improved by increasing the number of samples taken (and thus reducing the variance of the estimate) as the optimizer approaches convergence.

As discussed in section 2.2.3 and section 2.5, we believe that these estimates combined with the theory developed in [13] can be used to show convergence with high probability in an extension of the (trust region) conditional model framework. This will be explored in future work along with using the bounds of section 2.5 to extend the “error-aware” trust-region method in [37].

2.8 Acknowledgements

We thank Misha Kilmer for the use of the PaLS code [2] and Serkan Gugercin and Chris Beattie for many helpful discussions.

Chapter 3

Trust Regions with Stochastic Bounds

The expense of gradient calculations and repeated function evaluation can make high performance methods such as trust region methods unavailable for certain classes of problems, especially for parameter inversion for systems governed by systems of partial differential equations. We introduce a trust region algorithm that (1) replaces gradient computation by a stochastic estimate of the difference between the model and objective function gradients; (2) replaces the objective function evaluations by a stochastic estimate of the improvement ratio. We show that if these estimates are accurate to within a fixed multiplicative factor with a sufficiently high, fixed probability, then the algorithm converges to a stationary point of the objective function almost surely. Suitable bounds for the Diffuse Optical Tomography problem are derived and we demonstrate that this leads to a robust, high performance method that requires no complete full-order function or gradient evaluations.

3.1 Introduction

For most non-linear optimization tasks, the direct computation of optimal parameter values is not possible. Therefore, it is necessary to utilize an optimization framework that seeks to iteratively construct successively better approximations to the minimizer. To do so efficiently, most optimization techniques make use of derivatives either explicitly or indirectly. Among these methods, Newton’s method is a popular and high performance choice when second-order derivative information is available. However, the calculation of second-order derivatives is often computationally intractable. Thus, it often becomes necessary to approximate the objective function or some of its derivatives. Examples include the Gauss-Newton method (which approximates the Hessian for a special sub-class of optimization problems), general quasi-Newton methods (more general framework for Hessian approximation), and gradient descent methods (which use a linear approximation of the objective function). A popular choice is the trust region framework [20], which uses a surrogate model of the objective function over a region of parameter space where it is thought to be sufficiently well-approximated. Typically, this model is a quadratic Taylor series expansion about the center of the trust region. Since the quadratic model in the basic trust region framework still requires second-derivative information, many of the previously mentioned approximations can be used. This framework is sufficiently flexible enough to consider alternative models such as derivative-free variants that rely on samples of the objective function within the trust region [21], or even stochastic approximations/models [11, 19, 24].

In this paper, we present a flexible trust region framework that utilizes stochastic estimates of the gradient approximation error and of the improvement factor. In contrast to the STORM framework [19], the proposed framework does not require that the estimates are accurate relative to the trust region size, only that the estimates probabilistically satisfy a fixed multiplicative bound. Similarly, this framework is an improvement over methods that require that the estimates become increasingly more accurate [24]. This flexibility allows for a wide class of models and estimates to be used. In this paper, we will consider models generated by projection-based model order reduction [12, 14, 15, 22] equipped with Hutchinson trace estimates [28], previously used in the DOT context in [30].

3.2 Trust Regions with Stochastic Bounds

Algorithm 2 is a pseudo-code description of our proposed framework for the minimization of an unconstrained objective function, f . As is standard for the analysis of trust region methods, we impose only mild assumptions on f , that it is bounded below and that it has bounded second derivatives.

Assumption 3.1. There exists a lower bound $f_- > -\infty$ such that, for all x , $f(x) \geq f_-$.

Assumption 3.2. The objective function satisfies $f \in C^2(\chi)$ for the given domain χ . Furthermore, there exists a $\kappa_{\text{bhf}} > 1$ such that, for every step k ,

$$\max_{x \in B(x_k, \delta_k)} \|\nabla^2 f(x_k)\| \leq \kappa_{\text{bhf}} - 1,$$

where $\nabla^2 f$ denotes the Hessian of f .

We do not constrain the type of model used beyond some assumptions on continuity of the model and its derivatives and bounds thereof. In particular,

Assumption 3.3. Each model, m_ℓ , satisfies $m_\ell \in C^2(\chi)$ for the given domain χ . Furthermore, there exists a $\kappa_{\text{bhm}} > 1$ such that, for every step k and model m_ℓ ,

$$\max_{x \in B(x_k, \delta_k)} \|\nabla^2 m_\ell(x_k)\| \leq \kappa_{\text{bhm}} - 1.$$

At each iteration, k , we have a current point, x_k , and trust region radius, δ_k . The model, m_ℓ , is updated whenever the gradient error estimate, $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla f(x_k)\|)$, is sufficiently large. Since this update may occur multiple times at the start of an iteration or perhaps even not at all, we index the model separately with ℓ . The quadratic model that is typically used with trust regions can certainly be used, but we will abstain from assuming any particular structure so that we may consider using a reduced order model. If the norm of the model gradient is sufficiently large, a candidate step, s_k , is chosen that reduces the model by a sufficient amount.

In the basic trust region method [20], this step is accepted based on the computation of an improvement ratio. That is, the step is accepted if the ratio between the actual improvement in the objective function and the improvement in the model is sufficiently high. In our framework, we avoid the computation of the objective function by replacing the improvement factor with a stochastic estimate, $\psi_f(f(x_k) - f(x_k + s_k))$. If this estimated improvement ratio is sufficiently large, i.e. $\frac{\psi_f(f(x_k) - f(x_k + s_k))}{m_\ell(x_k) - m_\ell(x_k + s_k)} \geq (1 + \alpha_U)\eta_1$, we accept the step and increase our trust region radius. Otherwise, the step is rejected and we reduce the radius of the trust region.

To quantify what is meant by “sufficiently reduce” the model m_ℓ , we will use the fraction of the Cauchy decrease condition from the standard trust region method.

Assumption 3.4. For every step k , the step s_k satisfies

$$m_\ell(x_k) - m_\ell(x_k + s_k) \geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \min \left(\frac{\|\nabla m_\ell(x_k)\|}{\lambda_k}, \delta_k \right),$$

where $\lambda_k = 1 + \max_{x \in B(x_k, \delta_k)} \|\nabla^2 m_\ell(x_k)\|$ and $\kappa_{\text{fcd}} \in (0, 1)$.

Algorithm 2: Trust Regions with Stochastic Bounds

Input: x_0 : initial point; δ_0 : initial radius; $\alpha_U > 0$; $\eta_1 \in \left(0, \frac{1}{1+\alpha_U}\right)$; $\eta_2 > 0$; $\eta_3 > 0$; $\gamma > 1$;
 $\delta_{\max} > 0$

```

2.1 for  $k = 0, 1, 2, 3, \dots$  do
2.2   while  $\psi_{\nabla}(\|\nabla m_{\ell}(x_k) - \nabla f(x_k)\|) > \eta_3 \delta_k$  do
2.3      $\ell = \ell + 1$ ;  $m_{\ell+1} = \text{Update}(m_{\ell}, x_k)$ 
2.4   if  $\|\nabla m_{\ell}(x_k)\| < \eta_2 \delta_k$  then
2.5      $x_{k+1} = x_k$ ,  $\delta_{k+1} = \gamma^{-1} \delta_k$ .
2.6   else
2.7     Choose step  $s_k$  that sufficiently reduces the model  $m_{\ell}$  in  $B(x_k, \delta_k)$ .
2.8     Compute estimated acceptance ratio:  $\rho_k = \frac{\psi_f(f(x_k) - f(x_k + s_k))}{m_{\ell}(x_k) - m_{\ell}(x_k + s_k)}$ .
2.9     if  $\rho_k \geq (1 + \alpha_U)\eta_1$  then
2.10       $x_{k+1} = x_k + s_k$ ,  $\delta_{k+1} = \min(\gamma \delta_k, \delta_{\max})$ .
2.11     else
2.12       $x_{k+1} = x_k$ ,  $\delta_{k+1} = \gamma^{-1} \delta_k$ .
2.13       $\ell = \ell + 1$ ;  $m_{\ell+1} = \text{Update}(m_{\ell}, x_k)$ 

```

3.3 Stochastic Processes

3.3.1 Definitions

To describe the level of accuracy required of our estimates, we introduce the following definitions.

Definition 3.1. An estimate, $\psi(T)$, of a quantity T is said to be an α_U -accurate overestimate of T if

$$\psi(T)/T \leq 1 + \alpha_U.$$

Definition 3.2. An estimate, $\psi(T)$, of a quantity T is said to be an α_L -accurate underestimation of T if

$$\psi(T)/T \geq 1 - \alpha_L.$$

Definition 3.3. An estimate, $\psi(T)$, of a quantity T is said to be an $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimate of T if

$$1 - \alpha_L \leq \psi(T)/T \leq 1 + \alpha_U.$$

Note: An $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -estimate is also an $\begin{pmatrix} \alpha'_U \\ \alpha'_L \end{pmatrix}$ -estimate if $\alpha_L \leq \alpha'_L$ and $\alpha_U \leq \alpha'_U$.

Example 3.4. An $\begin{pmatrix} \alpha_U \\ 1 \end{pmatrix}$ -estimate of a quantity T must satisfy $\psi(T) \in \begin{cases} [0, (1 + \alpha_U)T] & T > 0 \\ [(1 + \alpha_U)T, 0] & T < 0 \end{cases}$.

3.3.2 Stochastic Processes

Since the models in this paper are updated based on random samples of various error measures, the models themselves are random variables. Let M_k denote the model used for the k -th iteration with realization $m_\ell = M_k(\omega)$. Similarly, the iterates $x_k = X_k(\omega)$, steps $s_k = S_k(\omega)$, and the trust region radii $\delta_k = \Delta_k(\omega)$ are random quantities with their respective realizations. Let Ψ_f^k denote the estimate of $f(X_k) - f(X_k + S_k)$ and Ψ_∇^k denote the estimate of $\|\nabla M_k(X_k) - \nabla f(X_k)\|$ with realizations $\Psi_f^k(\omega) = \psi_f^k$ and $\Psi_\nabla^k(\omega) = \psi_\nabla^k$, respectively. Then, Algorithm 2 results in a stochastic process $\{M_k, X_k, S_k, \Delta_k, \Psi_f^k, \Psi_\nabla^k\}$.

We can now define the notions of accuracy that we will require of our estimates in order to show convergence almost surely.

Definition 3.5. A sequence of estimates $\{\Psi_\nabla^k\}$ are called δ_∇ -probabilistically β -accurate with respect to the sequence $\{X_k\}$ if the events

$$\mathcal{I}_k = \{\Psi_\nabla^k \text{ is a } \beta\text{-accurate underestimate of } \|\nabla M_k(X_k) - \nabla f(X_k)\|\}$$

satisfy

$$\mathbf{P} \left[\mathcal{I}_k | \mathcal{F}_{k-1}^{M, \Psi} \right] \geq 1 - \delta_\nabla$$

where $\mathcal{F}_{k-1}^{M, \Psi}$ is the σ -algebra generated by $M_1, \dots, M_{k-1}$, $\Psi_\nabla^1, \dots, \Psi_\nabla^{k-1}$, and $\Psi_f^1, \dots, \Psi_f^{k-1}$.

Definition 3.6. A sequence of estimates $\{\Psi_f^k\}$ are called δ_f -probabilistically $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate with respect to the sequence $\{X_k, S_k\}$ if the events

$$\mathcal{J}_k = \left\{ \Psi_f^k \text{ is a } \begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}\text{-accurate estimate of } f(X_k) - f(X_k + S_k) \right\}$$

satisfy

$$\mathbf{P} \left[\mathcal{J}_k | \mathcal{F}_{k-1/2}^{M, \Psi} \right] \geq 1 - \delta_f$$

where $\mathcal{F}_{k-1/2}^{M, \Psi}$ is the σ -algebra generated by $M_1, \dots, M_k$, $\Psi_\nabla^1, \dots, \Psi_\nabla^k$, and $\Psi_f^1, \dots, \Psi_f^{k-1}$ (note that $\mathcal{F}_{k-1/2}^{M, \Psi}$ includes the σ -algebra generated by Ψ_∇^k while $\mathcal{F}_{k-1}^{M, \Psi}$ does not).

As in [11, 19], we will make use of the following result from martingale literature [25, Exercise 4.2.4].

Theorem 3.7 ([25]). *Let $\{X_n\}, n \geq 0$ be a sequence of random variables and a submartingale, that is,*

$$\mathbb{E} [X_{n+1} | \mathcal{F}_n^X] \geq X_n,$$

where $\mathcal{F}_n^X$ is the σ -algebra generated by $X_0, \dots, X_n$. If $\limsup_{n \rightarrow \infty} X_n < \infty$ and

$$\mathbb{E} \left[\limsup_{n \rightarrow \infty} (X_{n+1} - X_n)^+ \right] < \infty,$$

where $(f)^+ = \max(0, f)$, then X_n converges almost surely to a limit X with $\mathbb{E} [|X|] < \infty$.

This yields a corollary that will be useful for our analysis.

Corollary 3.8. *If $\{X_n\}, n \geq 0$ is a submartingale with $|X_{n+1} - X_n| \leq M$ for some $M \geq 0$, then*

$$\mathbf{P} \left[\left\{ \lim_{n \rightarrow \infty} X_n < \infty \right\} \cup \left\{ \limsup_{n \rightarrow \infty} X_n = \infty \right\} \right] = 1.$$

Proof. Since $|X_{n+1} - X_n| \leq M$, it follows that $(X_{n+1} - X_n)^+ \leq M$. Therefore, we have $\limsup_{n \rightarrow \infty} (X_{n+1} - X_n)^+ \leq M$ and thus $\mathbb{E} \left[\limsup_{n \rightarrow \infty} (X_{n+1} - X_n)^+ \right] \leq M$. If $\limsup_{n \rightarrow \infty} X_n < \infty$, Theorem 3.7 applies and $\lim_{n \rightarrow \infty} X_n < \infty$ almost surely. Thus,

$$\begin{aligned} 1 &= \mathbf{P} \left[\left\{ \limsup_{n \rightarrow \infty} X_n < \infty \right\} \cup \left\{ \limsup_{n \rightarrow \infty} X_n = \infty \right\} \right] \\ &= \mathbf{P} \left[\left\{ \lim_{n \rightarrow \infty} X_n < \infty \right\} \cup \left\{ \limsup_{n \rightarrow \infty} X_n = \infty \right\} \right], \end{aligned}$$

as desired. □

3.4 Convergence Analysis

Convergence of the Basic Trust Region Algorithm

For illustrative purposes, we first consider the basic trust region algorithm [20], as described by Algorithm 3, and its convergence framework.

Algorithm 3: Basic Trust Region [20]

Input: x_0 : initial point; δ_0 : initial radius; $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq \gamma_2 < 1$

```

3.1 for  $k = 0, 1, 2, 3, \dots$  do
3.2   Define a model  $m_k$  in  $B(x_k, \delta_k)$ 
3.3   Choose step  $s_k$  that sufficiently reduces the model  $m_k$  in  $B(x_k, \delta_k)$ .
3.4   Compute improvement ratio:  $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$ .
3.5   if  $\rho_k \geq \eta_1$  then
3.6     Set  $x_{k+1} = x_k + s_k$ ,  $\delta_{k+1} \in \begin{cases} [\delta_k, \infty) & \rho_k \geq \eta_2 \\ [\gamma_2 \delta_k, \delta_k] & \eta_1 \leq \rho_k < \eta_2 \end{cases}$ .
3.7   else
3.8     Set  $x_{k+1} = x_k$ ,  $\delta_{k+1} \in [\gamma_1 \delta_k, \gamma_2 \delta_k]$ .

```

To show convergence of Algorithm 3, Conn, et al., demonstrate that a sufficiently small (relative to the gradient norm) trust region radius yields a step that makes sufficient progress, is accepted, and expands the trust region. Assuming f is bounded below, they show infinitely many successful iterations requires $\liminf_{n \rightarrow \infty} \|\nabla f(x_k)\| = 0$. Finally, they show that if $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| \neq 0$, then $\|\nabla f(x_k)\|$ must go between ϵ and 2ϵ infinitely many times. Over these indices, the radii converge to zero, and thus the difference between iterates tends to zero. By continuity of the gradient, this implies the difference in $\nabla f(x_k)$ tends to zero, which contradicts the assumption.

Convergence of the TRSB Algorithm

Within the framework of the TRSB algorithm, we do not have access to the objective function, f , only to the estimates $\psi_f(f(x_k) - f(x_k + s_k))$ and $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla f(x_k)\|)$. When these estimates are sufficiently accurate (in the sense of Definitions 3.2 and 3.3), we can still show that a sufficiently small (relative to the model gradient norm) trust region radius yields an accepted step that makes sufficient progress and expands the trust region radius (Lemmas 3.10 and 3.12).

However, since we use stochastic estimates for the gradient error and the improvement factor, there is a possibility that the estimates are not sufficiently accurate for a particular iteration.

We show that if the improvement factor estimates are sufficiently accurate, an accepted step still yields sufficient progress and expands the trust region radius (Lemma 3.9).

By combining these results and adding conditions on the probabilities δ_∇ , δ_f (as defined in Definitions 3.5 and 3.6), we show that the trust region radii converge to zero almost surely (Theorem 3.13). Under these conditions, we show that $\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$ almost surely (Theorem 3.14). By considering a similar subsequence of iterates with $\|\nabla f(x_k)\|$ between ϵ and 2ϵ , we show that if $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| \neq 0$ with any positive probability, that the sum of trust region radii for these iterates is uniformly bounded away from 0, contradicting an auxiliary lemma (Lemma 3.15) on the summability of trust region radii for iterations with $\|\nabla f(x_k)\| \geq \epsilon$. Therefore, we have $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$ almost surely (Theorem 3.16).

We add the assumptions of the existence of a suitably accurate sequence of estimates for any values of the algorithmic parameters α_U , α_L , β and probabilities δ_∇ and δ_f . For the types of estimates explored in Section 3.6, this can always be done by increasing the number of samples.

Assumption 3.5. For a given $\alpha_U > 0$, $\alpha_L > 0$, and $0 \leq \delta_f < 1$, there exists a sequence of estimates, $\{\Psi_f^k\}$ such that Ψ_f^k are δ_f -probabilistically $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimates of $f(X_k) - f(X_k + S_k)$.

Assumption 3.6. For a given $0 \leq \beta < 1$, and $0 \leq \delta_\nabla < 1$, there exists a sequence of estimates, $\{\Psi_\nabla^k\}$ such that Ψ_∇^k are δ_∇ -probabilistically β -accurate underestimates of $\|\nabla M_k(X_k) - \nabla f(X_k)\|$.

3.4.1 Trust Region Radius Convergence

We begin with showing that, given sufficiently accurate estimates, an accepted step improves the objective function by an amount proportional to the trust region radius squared. We will use this to bound changes in the objective function and the trust region radius when the gradient estimate may not be sufficiently accurate.

Lemma 3.9. *If $\psi_f(f(x_k) - f(x_k + s_k))$ is an $\begin{pmatrix} \alpha_U \\ 1 \end{pmatrix}$ -accurate estimate and the step is accepted, then the improvement in f is bounded below by*

$$f(x_k) - f(x_k + s_k) \geq C_1 \delta_k^2 > 0,$$

where $C_1 = \eta_1 \eta_2 \kappa_{\text{fcd}} \min\left(\frac{\eta_2}{\kappa_{\text{bhm}}}, 1\right) > 0$.

Proof. If a step is accepted, then we have

$$(1 + \alpha_U) \eta_1 (m_\ell(x_k) - m_\ell(x_k + s_k)) \leq \psi_f(f(x_k) - f(x_k + s_k)).$$

Combining this with the assumption yields

$$(1 + \alpha_U) \eta_1 (m_\ell(x_k) - m_\ell(x_k + s_k)) \leq (1 + \alpha_U) (f(x_k) - f(x_k + s_k)).$$

Therefore, we have

$$\begin{aligned} f(x_k) - f(x_k + s_k) &\geq \eta_1 (m_\ell(x_k) - m_\ell(x_k + s_k)) \\ &\geq \eta_1 \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \min \left(\frac{\|\nabla m_\ell(x_k)\|}{\kappa_{\text{bhm}}}, \delta_k \right) \\ &\geq \eta_1 \kappa_{\text{fcd}} (\eta_2 \delta_k) \min \left(\frac{(\eta_2 \delta_k)}{\kappa_{\text{bhm}}}, \delta_k \right) \\ &= \underbrace{\eta_1 \eta_2 \kappa_{\text{fcd}} \min \left(\frac{\eta_2}{\kappa_{\text{bhm}}}, 1 \right)}_{C_1} \delta_k^2, \end{aligned}$$

where the second inequality follows from Assumption 3.4 and the final inequality follows from the model gradient norm condition (Line 4) in Algorithm 2. $\square$

We will show that a sufficiently small trust region radius leads to an improvement in the objective function proportional to the product of the objective function gradient and the trust region radius. To do so, we first show that a sufficiently small trust region radius leads to an improvement in the objective function proportional to the product of the model gradient and the trust region radius.

Lemma 3.10. *Let $\kappa_{\text{bh}} = \max(\kappa_{\text{bhm}}, \kappa_{\text{bhf}})$. Suppose $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla f(x_k)\|)$ is a β -accurate underestimate and δ_k satisfies*

$$\delta_k \leq \min \left(\frac{1}{\kappa_{\text{bh}}}, \frac{1}{2} \frac{(1 - \beta) \kappa_{\text{fcd}}}{\eta_3 + (1 - \beta) \kappa_{\text{bh}}} \right) \|\nabla m_\ell(x_k)\|.$$

Then the improvement in the trial step is bounded below by

$$f(x_k) - f(x_k + s_k) \geq \frac{\kappa_{\text{fcd}}}{2} \|\nabla m_\ell(x_k)\| \delta_k.$$

Proof. By Assumption 3.4, it follows that

$$\begin{aligned} m_\ell(x_k) - m_\ell(x_k + s_k) &\geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \min \left(\frac{\|\nabla m_\ell(x_k)\|}{\lambda_k}, \delta_k \right) \\ &= \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \delta_k. \end{aligned}$$

Since f and m_ℓ are scalar valued, we may write

$$\begin{aligned} f(x_k + s_k) &= f(x_k) + s_k^T \nabla f(x_k) + \frac{1}{2} s_k^T \nabla^2 f(\xi) s_k, \\ m_\ell(x_k + s_k) &= m_\ell(x_k) + s_k^T \nabla m_\ell(x_k) + \frac{1}{2} s_k^T \nabla^2 m_\ell(\nu) s_k, \end{aligned}$$

where $\xi = x_k + ts_k, \nu = x_k + rs_k$ for some $t, r \in (0, 1)$. By the first while loop in Algorithm 2, we have $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla f(x_k)\|) \leq \eta_3 \delta_k$ which, with the assumptions, implies

$$\|\nabla m_\ell(x_k) - \nabla f(x_k)\| \leq \frac{\eta_3}{1 - \beta} \delta_k.$$

Therefore,

$$\begin{aligned} f(x_k) - f(x_k + s_k) &= (f(x_k) - f(x_k + s_k)) - (m_\ell(x_k) - m_\ell(x_k + s_k)) + (m_\ell(x_k) - m_\ell(x_k + s_k)) \\ &= (m_\ell(x_k) - m_\ell(x_k + s_k)) + s_k^T (\nabla m_\ell(x_k) - \nabla f(x_k)) + \frac{1}{2} s_k^T (\nabla^2 m_\ell(\nu) - \nabla^2 f(\xi)) s_k \\ &\geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \delta_k - |s_k^T (\nabla m_\ell(x_k) - \nabla f(x_k))| - \left| \frac{1}{2} s_k^T (\nabla^2 m_\ell(\nu) - \nabla^2 f(\xi)) s_k \right|, \end{aligned}$$

where the inequality follows from $a \geq -|a|$ and the sufficient decrease condition (Assumption 3.4). Thus, we find

$$\begin{aligned} f(x_k) - f(x_k + s_k) &\geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \delta_k - \|s_k\| \|\nabla m_\ell(x_k) - \nabla f(x_k)\| - \frac{1}{2} \|s_k\|^2 \|\nabla^2 m_\ell(\nu) - \nabla^2 f(\xi)\| \\ &\geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \delta_k - \delta_k \left(\frac{\eta_3}{1 - \beta} \delta_k \right) - \frac{1}{2} \delta_k^2 (\|\nabla^2 f(\xi)\| + \|\nabla^2 m_\ell(\nu)\|) \\ &\geq \left(\kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| - \left(\frac{\eta_3}{1 - \beta} + \kappa_{\text{bh}} \right) \delta_k \right) \delta_k \\ &\geq \frac{1}{2} \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \delta_k. \end{aligned}$$

□

Lemma 3.11. Suppose $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla m_\ell(x_k)\|)$ is a β -accurate underestimate. Define $\widehat{\zeta} = \frac{\eta_3}{1 - \beta} + \max\left(\kappa_{\text{bh}}, \frac{2(\eta_3 + (1 - \beta)\kappa_{\text{bh}})}{(1 - \beta)\kappa_{\text{fcd}}}\right)$. If δ_k satisfies

$$\widehat{\zeta} \delta_k \leq \|\nabla f(x_k)\|,$$

then the improvement in the trial step is bounded by

$$f(x_k) - f(x_k + s_k) \geq C_2 \|\nabla f(x_k)\| \delta_k,$$

where

$$\begin{aligned} C_2 &= \frac{\kappa_{\text{fcd}}}{2} \left(1 - \frac{\eta_3}{(1 - \beta)\widehat{\zeta}} \right) \\ &= \frac{\kappa_{\text{fcd}}}{2} \max\left(\frac{(1 - \beta)\kappa_{\text{bh}}}{(1 - \beta)\kappa_{\text{bh}} + \eta_3}, \frac{2((1 - \beta)\kappa_{\text{bh}} + \eta_3)}{(\kappa_{\text{fcd}} + 2)\eta_3 + 2(1 - \beta)\kappa_{\text{bh}}} \right) \\ &> 0. \end{aligned}$$

Proof. As in Lemma 3.10, we have $\|\nabla m_\ell(x_k) - \nabla f(x_k)\| \leq \frac{\eta_3}{1-\beta} \delta_k$ which implies

$$\|\nabla m_\ell(x_k)\| \geq \|\nabla f(x_k)\| - \frac{\eta_3}{(1-\beta)} \delta_k. \quad (3.1)$$

By assumption, we have $\|\nabla f(x_k)\| \geq \max\left(\kappa_{\text{bh}} + \frac{\eta_3}{1-\beta}, \frac{2(\eta_3 + (1-\beta)\kappa_{\text{bh}})}{(1-\beta)\kappa_{\text{fcd}}} + \frac{\eta_3}{1-\beta}\right) \delta_k$, and thus,

$$\|\nabla m_\ell(x_k)\| \geq \max\left(\kappa_{\text{bh}}, \frac{2(\eta_3 + (1-\beta)\kappa_{\text{bh}})}{(1-\beta)\kappa_{\text{fcd}}}\right) \delta_k.$$

Hence, we may apply Lemma 3.10 which yields

$$f(x_k) - f(x_k + s_k) \geq \frac{\kappa_{\text{fcd}}}{2} \|\nabla m_\ell(x_k)\| \delta_k. \quad (3.2)$$

Also from Equation (3.1), we may conclude that

$$\begin{aligned} \|\nabla m_\ell(x_k)\| &\geq \|\nabla f(x_k)\| - \frac{\eta_3}{1-\beta} \left(\frac{1}{\widehat{\zeta}} \|\nabla f(x_k)\|\right) \\ &= \left(1 - \frac{\eta_3}{(1-\beta)\widehat{\zeta}}\right) \|\nabla f(x_k)\|. \end{aligned} \quad (3.3)$$

Combining Equations (3.2) and (3.3) yields the desired result. $\square$

Now we show that a successful iteration is guaranteed if the estimates are sufficiently accurate and the trust region radius is sufficiently small. This element is key to the convergence analysis, as a similar result in the standard trust region, to show that the method will eventually make progress.

Lemma 3.12. *Suppose $\psi_\nabla(\|\nabla m_\ell(x_k) - \nabla f(x_k)\|)$ is a β -accurate underestimate, $\psi_f(f(x_k) - f(x_k + s_k))$ is an $\left(\frac{\alpha_U}{\alpha_L}\right)$ -accurate estimate with $\alpha_L < 1 - \frac{2(1+\alpha_U)\eta_1}{\kappa_{\text{fcd}}}$, and δ_k satisfies*

$$\delta_k \leq \min\left(\frac{1}{\kappa_{\text{bh}}}, \frac{1}{2} \frac{(1-\beta)\kappa_{\text{fcd}}}{\eta_3 + (1-\beta)\kappa_{\text{bh}}}, \frac{\frac{(1-\alpha_L)\kappa_{\text{fcd}}}{(1+\alpha_U)\eta_1} - 2}{\kappa_{\text{bh}}}\right) \|\nabla m_\ell(x_k)\|.$$

Then the k -th iteration is accepted.

Proof. Note that our assumptions also meet those of Lemma 3.10 and thus

$$f(x_k) - f(x_k + s_k) \geq \frac{\kappa_{\text{fcd}}}{2} \|\nabla m_\ell(x_k)\| \delta_k > 0.$$

Thus, it follows that

$$\psi_f(f(x_k) - f(x_k + s_k)) \geq (1 - \alpha_L)(f(x_k) - f(x_k + s_k)) \geq (1 - \alpha_L) \frac{\kappa_{\text{fcd}}}{2} \|\nabla m_\ell(x_k)\| \delta_k.$$

By Taylor's theorem, we have that

$$\begin{aligned} m_\ell(x_k) - m_\ell(x_k + s_k) &= |m_\ell(x_k) - m_\ell(x_k + s_k)| \\ &= \left| s_k^T \nabla m_\ell(x_k) + \frac{1}{2} s_k^T \nabla^2 m_\ell(\nu) s_k \right| \\ &\leq \|s_k\| \|\nabla m_\ell(x_k)\| + \frac{1}{2} \|s_k\|^2 \|\nabla^2 m_\ell(\nu)\| \\ &\leq \|\nabla m_\ell(x_k)\| \delta_k + \frac{\kappa_{\text{bh}}}{2} \delta_k^2, \end{aligned}$$

where $\nu = x_k + r s_k$ for some $r \in (0, 1)$. Therefore, we see that

$$\begin{aligned} \rho &= \frac{\psi_f(f(x_k) - f(x_k + s_k))}{m_\ell(x_k) - m_\ell(x_k + s_k)} \\ &\geq \frac{(1 - \alpha_L) \frac{\kappa_{\text{fcd}}}{2} \|\nabla m_\ell(x_k)\| \delta_k}{\|\nabla m_\ell(x_k)\| \delta_k + \frac{\kappa_{\text{bh}}}{2} \delta_k^2} \\ &= \frac{(1 - \alpha_L) \kappa_{\text{fcd}}}{2 + \frac{\kappa_{\text{bh}}}{\|\nabla m_\ell(x_k)\|} \delta_k} \\ &\geq (1 + \alpha_U) \eta_1. \end{aligned}$$

□

Note: Requiring $\alpha_L < 1 - \frac{2(1 + \alpha_U)\eta_1}{\kappa_{\text{fcd}}}$ is not overly restrictive. For example, with $\alpha_U = 4$, $\eta_1 = 1/100$, and $\kappa_{\text{fcd}} = 1/2$, then $\alpha_L < 4/5$.

We are now prepared to show that the trust region radii almost surely go to zero.

Theorem 3.13. *Let Assumptions 3.1 - 3.6 be satisfied. Choose algorithmic parameters $\alpha_U > 0$, $\eta_1 \in (0, \frac{1}{1 + \alpha_U})$, $\eta_3 > 0$, and $\gamma > 1$. Define C_1 as in Lemma 3.9, $\hat{\zeta}$ and C_2 as in Lemma 3.11, and let*

$$\zeta = \max \left(\hat{\zeta}, \frac{\eta_3}{1 - \beta} + \frac{\frac{(1 - \alpha_L)\kappa_{\text{fcd}}}{2} - 2}{\kappa_{\text{bh}}} \right) \quad (3.4)$$

and $\nu \in (0, 1)$ such that

$$\frac{\nu}{1 - \nu} = \max \left(\frac{4\gamma^2}{C_2\zeta}, \frac{\gamma^2}{C_1}, \frac{\gamma^2}{\kappa_{\text{bh}}} \right) \quad (3.5)$$

Then, if δ_f and δ_∇ are chosen sufficiently close to 0 that they satisfy

$$\frac{(1 - \delta_f)(1 - \delta_\nabla) - \frac{1}{2}}{\delta_f \delta_\nabla} > \frac{3}{2C_2}, \quad (3.6)$$

$$\delta_f \delta_\nabla \leq \frac{1}{2 \left(1 + \frac{\gamma^2}{\gamma^2 - 1} \left(\frac{\nu}{1 - \nu} \left(\zeta + \frac{\kappa_{\text{bh}}}{2} \right) + \gamma^2 + 1 \right) \right)}, \quad (3.7)$$

and the algorithmic parameters η_2 and α_L are chosen to satisfy

$$\eta_2 \geq \max \left(\kappa_{\text{bh}}, \frac{C_2 \zeta}{4}, \frac{1 + \frac{\eta_3}{1 - \beta} + \kappa_{\text{bh}}}{\kappa_{\text{fcd}}} \right), \quad (3.8)$$

$$\alpha_L < 1 - \frac{2(1 + \alpha_U)\eta_1}{\kappa_{\text{fcd}}}, \quad (3.9)$$

the sequence of trust region radii, $\{\Delta_k\}$, satisfy $\sum_{k=0}^{\infty} \Delta_k^2 < \infty$ almost surely.

Proof. This proof is similar in structure to that given in [19] for their STORM algorithm, but some changes are necessary to fit the structure of our framework. We will similarly consider the random function

$$\Phi_k = \nu f(X_k) + (1 - \nu) \Delta_k^2 \quad (3.10)$$

and its realizations $\Phi_k(\omega) = \phi_k$. In particular, we will show that there exists a constant $\sigma > 0$ such that

$$\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi} \right] \leq -\sigma \Delta_k^2, \quad (3.11)$$

where $\mathcal{F}_{k-1}^{M, \Psi}$ is as defined in Definition 3.5. Therefore, since f is bounded below by Assumption 3.1 and $\Delta_k > 0$, it follows that Φ_k is bounded from below and, as in [19], we may conclude that $\sum_{k=0}^{\infty} \Delta_k^2 < \infty$ almost surely by summing over Equation (3.11) from $k = 1$ to ∞ and taking the expectation of both sides.

To show that Equation (3.11) holds on each iteration, we will examine the cases where $\nabla f(x_k)$ is large relative to the trust region radius (i.e. when $\|\nabla f(x_k)\| \geq \zeta \delta_k$) and when it is not. If Equation (3.11) holds in both cases, then it holds in general. For each case, we will examine four sub-cases corresponding to the possible values of the events $\mathcal{I}_k$ and $\mathcal{J}_k$ as defined in Definitions 3.5 and 3.6. For brevity, we will use the notation $!\mathcal{I}_k$ for when $\mathcal{I}_k$ does not hold and similarly for $\mathcal{J}_k$.

To begin, we note that either a step is accepted, or the trust region radius is reduced. Therefore, we have the following two options. When the step is not accepted, we have $x_{k+1} = x_k$ and

$$\phi_{k+1} - \phi_k = (1 - \nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2 < 0. \quad (3.12)$$

Whereas for an accepted step, we have $x_{k+1} = x_k + s_k$ and

$$\phi_{k+1} - \phi_k \leq \nu (f(x_{k+1}) - f(x_k)) + (1 - \nu) (\gamma^2 - 1) \delta_k^2, \quad (3.13)$$

where the inequality is due to the fact that we bound $\delta_{k+1} = \max(\gamma\delta_k, \delta_{\max})$.

Case 1: $[||\nabla f(x_k)|| \geq \zeta\delta_k]$

When $||\nabla f(x_k)||$ is sufficiently large, relative to the trust region radius, accurate estimates yield decreases in ϕ_k proportional to $||\nabla f(x_k)|| \delta_k$. However, poor estimates may increase ϕ_k by an amount also proportional to $||\nabla f(x_k)|| \delta_k$. This will place conditions on δ_f and δ_{∇} to ensure that such mis-steps occur sufficiently rarely that ϕ_k decreases in expectation.

Case 1(a): $[\mathcal{I}_k \text{ and } \mathcal{J}_k]$

With our choice of ζ , Lemma 3.11 applies and thus we have that

$$f(x_k + s_k) - f(x_k) \leq -C_2 ||f(x_k)|| \delta_k.$$

Furthermore, since $\psi_{\nabla}(||\nabla m_{\ell}(x_k) - \nabla f(x_k)||)$ is a β -accurate underestimate, we have

$$\begin{aligned} ||\nabla m_{\ell}(x_k)|| &\geq ||\nabla f(x_k)|| - \frac{\eta_3}{1 - \beta} \delta_k \\ &\geq \left(\zeta - \frac{\eta_3}{1 - \beta} \right) \delta_k. \end{aligned}$$

Therefore, Lemma 3.12 applies and the step is accepted. Using Equation (3.13), we have

$$\begin{aligned} \phi_{k+1} - \phi_k &\leq \nu(-C_2 ||\nabla f(x_k)|| \delta_k) + (1 - \nu)(\gamma^2 - 1) \delta_k^2 \\ &\leq (-\nu C_2 \zeta + (1 - \nu)(\gamma^2 - 1)) \delta_k^2. \end{aligned} \quad (3.14)$$

Since, by theorem assumption 3.5, ν satisfies $\frac{\nu}{1 - \nu} \geq \frac{4\gamma^2}{C_2\zeta} > \frac{\gamma^2 - 1}{C_2\zeta}$, it follows that $-\nu C_2 \zeta + (1 - \nu)(\gamma^2 - 1) < 0$.

Case 1(b): $[\mathcal{I}_k \text{ and } \neg \mathcal{J}_k]$

In this case, Lemma 3.11 still applies and thus

$$f(x_k + s_k) - f(x_k) \leq -C_2 ||f(x_k)|| \delta_k.$$

However, without the $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimate of the improvement, we are no longer guaranteed to accept the step. If the step was successful, we have the same change as in Case 1(a). If the step is (erroneously) rejected, Equation (3.12) applies. Since, by theorem assumption

3.5, we have $\frac{\nu}{1-\nu} \geq \frac{\gamma^2}{C_2\zeta}$, it follows that $-\nu C_2\zeta + (1-\nu)(\gamma^2 - 1) \leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right)$. Therefore, we have

$$\phi_{k+1} - \phi_k \leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2. \quad (3.15)$$

Case 1(c): [$\mathcal{I}_k$ and $\mathcal{J}_k$]

Given the sufficiently accurate estimates of the improvement factor, if the step is accepted, then Lemma 3.9 applies and we have

$$f(x_{k+1}) - f(x_k) \leq -C_1 \delta_k^2.$$

Since, by theorem assumption 3.5, we have $\frac{\nu}{1-\nu} \geq \frac{\gamma^2}{C_1} > \frac{\gamma^2 - \frac{1}{\gamma^2}}{C_1}$, it follows that

$$-\nu C_1 + (1-\nu)(\gamma^2 - 1) \leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right).$$

Therefore, regardless of the acceptance of the step, we have

$$\phi_{k+1} - \phi_k \leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2. \quad (3.16)$$

Case 1(d): [$\mathcal{I}_k$ and $\mathcal{J}_k$]

Without quality estimates of the gradient error or the change in f , the algorithm may accept a step that increases f . In particular, we find

$$\begin{aligned} f(x_k + s_k) - f(x_k) &= s_k^T \nabla f(x_k) + \frac{1}{2} s_k^T \nabla^2 f(\xi) s_k \\ &\leq \|s_k\| \|\nabla f(x_k)\| + \frac{1}{2} \|s_k\|^2 \|\nabla^2 f(\xi)\| \\ &\leq \|\nabla f(x_k)\| \delta_k + \frac{\kappa_{\text{bh}}}{2} \delta_k^2, \end{aligned}$$

where $\xi = x_k + ts_k$ for some $t \in (0, 1)$. Since $\zeta \geq \kappa_{\text{bh}} + \frac{\eta_3}{1-\beta}$, we have that $\|\nabla f(x_k)\| \geq \kappa_{\text{bh}} \delta_k$. Therefore, $f(x_k + s_k) - f(x_k) \leq \frac{3}{2} \|\nabla f(x_k)\| \delta_k$. Thus, for this case, we have

$$\phi_{k+1} - \phi_k \leq \nu \left(\frac{3}{2} \|\nabla f(x_k)\| \delta_k \right) + (1-\nu)(\gamma^2 - 1) \delta_k^2. \quad (3.17)$$

We are now prepared to evaluate $\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_K)\| \geq \zeta \Delta_k \} \right]$. For brevity, let $E_1 = \mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_K)\| \geq \zeta \Delta_k \} \right]$. Since Case 1(a) occurs with probability at least $(1 - \delta_{\nabla})(1 - \delta_f)$, Case 1(d) occurs with probability at most $\delta_{\nabla} \delta_f$, and Case 1(b) or Case 1(c) occur with probability at least $\delta_{\nabla}(1 - \delta_f) + (1 - \delta_{\nabla})\delta_f$, we may apply Equations (3.14), (3.15), (3.16) and (3.17) to see

$$\begin{aligned}
E_1 \leq & (1 - \delta_f)(1 - \delta_\nabla) [\nu(-C_2 \|\nabla f(X_k)\| \Delta_k) + (1 - \nu)(\gamma^2 - 1)\Delta_k^2] \\
& + (\delta_f(1 - \delta_\nabla) + (1 - \delta_f)\delta_\nabla) \left[(1 - \nu) \left(\frac{1}{\gamma^2} - 1 \right) \Delta_k^2 \right] \\
& + \delta_f \delta_\nabla \left[\nu \left(\frac{3}{2} \|\nabla f(X_k)\| \Delta_k \right) + (1 - \nu)(\gamma^2 - 1)\Delta_k^2 \right]. \tag{3.18}
\end{aligned}$$

Gathering terms, we may write this as

$$\begin{aligned}
E_1 \leq & \left(\frac{3}{2} \delta_f \delta_\nabla - (1 - \delta_f)(1 - \delta_\nabla) C_2 \right) \nu \|\nabla f(X_k)\| \Delta_k \\
& + \left((1 - \delta_f)(1 - \delta_\nabla) + \delta_f \delta_\nabla - \frac{1}{\gamma^2} ((1 - \delta_f)\delta_\nabla + \delta_f(1 - \delta_\nabla)) \right) (1 - \nu)(\gamma^2 - 1)\Delta_k^2.
\end{aligned}$$

Note that since

$$\begin{aligned}
& \left((1 - \delta_f)(1 - \delta_\nabla) + \delta_f \delta_\nabla - \frac{1}{\gamma^2} ((1 - \delta_f)\delta_\nabla + \delta_f(1 - \delta_\nabla)) \right) \\
& \leq ((1 - \delta_f)(1 - \delta_\nabla) + \delta_f \delta_\nabla + 1((1 - \delta_f)\delta_\nabla + \delta_f(1 - \delta_\nabla))) = 1,
\end{aligned}$$

it follows that

$$\begin{aligned}
\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_k)\| \geq \zeta \Delta_k \} \right] \leq & \left(\frac{3}{2} \delta_f \delta_\nabla - (1 - \delta_f)(1 - \delta_\nabla) C_2 \right) \nu \|\nabla f(X_k)\| \Delta_k \\
& + (1 - \nu)(\gamma^2 - 1)\Delta_k^2.
\end{aligned}$$

From our choice of $\frac{\nu}{1 - \nu} \geq \frac{4\gamma^2}{C_2\zeta}$, we see that

$$(1 - \nu)(\gamma^2 - 1) \leq (1 - \nu)\gamma^2 \leq \frac{C_2}{4}\nu\zeta.$$

Since $\|\nabla f(X_k)\| \geq \zeta \Delta_k$, this implies

$$(1 - \nu)(\gamma^2 - 1)\Delta_k^2 \leq \frac{C_2}{4}\nu\zeta\Delta_k^2 \leq \frac{C_2}{4}\nu\|\nabla f(X_k)\| \Delta_k.$$

Therefore, we may write

$$E_1 \leq \left(\frac{3}{2} \delta_f \delta_\nabla - (1 - \delta_f)(1 - \delta_\nabla) C_2 + \frac{C_2}{4} \right) \nu \|\nabla f(X_k)\| \Delta_k.$$

Since, in Equation (3.6), we assumed δ_f, δ_∇ satisfy $\frac{(1 - \delta_f)(1 - \delta_\nabla) - \frac{1}{2}}{\delta_f \delta_\nabla} > \frac{3}{2C_2}$, it follows that

$$\frac{3}{2} \delta_f \delta_\nabla < (1 - \delta_f)(1 - \delta_\nabla) C_2 - \left(\frac{1}{4} + \frac{1}{4} \right) C_2$$

which implies

$$\frac{3}{2}\delta_f\delta_\nabla - (1 - \delta_f)(1 - \delta_\nabla)C_2 + \frac{1}{4}C_2 < -\frac{1}{4}C_2.$$

Thus, by applying our case assumption, we conclude

$$\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_k)\| \geq \zeta \Delta_k \} \right] \leq -\frac{1}{4}C_2\nu \|\nabla f(X_k)\| \Delta_k \quad (3.19)$$

$$\begin{aligned} &\leq -\frac{1}{4}C_2\nu\zeta\Delta_k^2 \\ &\leq -\gamma^2(1 - \nu)\Delta_k^2. \end{aligned} \quad (3.20)$$

Case 2: $\|\nabla f(x_k)\| < \zeta\delta_k$

When $\|\nabla f(x_k)\|$ is small relative to the trust region radius δ_k , we may no longer rely on our earlier lemmas that guarantee progress except for Lemma 3.9. Hence, many of the following arguments will reduce to showing that for accepted steps, the change in ϕ is bounded by Equation (3.12). In this regime, when both the gradient and improvement estimates are poor, the increase in ϕ is bounded proportional to δ_k^2 . Thus we can choose δ_f, δ_∇ to be sufficiently small so that this increase is small relative to $(1 - \nu)(\frac{1}{\gamma^2} - 1)\delta_k$ and ϕ decreases in expectation.

Case 2(a): $[\mathcal{I}_k \text{ and } \mathcal{J}_k]$

As in Case 1(c), since the improvement factor estimate is sufficiently accurate, if the step is accepted, the change in ϕ is bounded by the decrease given by Equation (3.12). Thus,

$$\phi_{k+1} - \phi_k \leq (1 - \nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2. \quad (3.21)$$

Case 2(b): $[\mathcal{I}_k \text{ and } !\mathcal{J}_k]$

If the step, s_k , is not accepted, then Equation (3.12) applies. If the step is accepted, then, by Line 4 in Algorithm 2, we must have

$$\|m_\ell(x_k)\| \geq \eta_2\delta_k.$$

Thus, by Assumption 3.4, we have

$$\begin{aligned} m_\ell(x_k) - m_\ell(x_k + s_k) &\geq \kappa_{\text{fcd}} \|\nabla m_\ell(x_k)\| \min \left(\frac{\|\nabla m_\ell(x_k)\|}{\kappa_{\text{bh}}}, \delta_k \right) \\ &\geq \kappa_{\text{fcd}} \eta_2 \delta_k^2 \min \left(\frac{\eta_2}{\kappa_{\text{bh}}}, 1 \right) \\ &= \kappa_{\text{fcd}} \eta_2 \delta_k^2. \end{aligned}$$

Therefore, may use Taylor's theorem to yield

$$\begin{aligned}
f(x_k + s_k) - f(x_k) &= (f(x_k + s_k) - f(x_k)) - (m_\ell(x_k + s_k) - m_\ell(x_k)) + (m_\ell(x_k + s_k) - m_\ell(x_k)) \\
&= (m_\ell(x_k + s_k) - m_\ell(x_k)) + \left(s_k^T (\nabla f(x_k) - \nabla m_\ell(x_k)) + \frac{1}{2} s_k^T (\nabla^2 f(\xi) - \nabla^2 m(\psi)) s_k \right) \\
&\leq -\kappa_{\text{fcd}} \eta_2 \delta_k^2 + \frac{\eta_3}{1-\beta} \delta_k^2 + \kappa_{\text{bh}} \delta_k^2 \\
&= \left(-\kappa_{\text{fcd}} \eta_2 + \frac{\eta_3}{1-\beta} + \kappa_{\text{bh}} \right) \delta_k^2.
\end{aligned}$$

where $\xi = x_k + ts_k, \psi = x_k + rs_k$ for some $r, t \in (0, 1)$. Since, in Equation (3.8), we chose $\eta_2 \geq \frac{1 + \frac{\eta_3}{1-\beta} + \kappa_{\text{bh}}}{\kappa_{\text{fcd}}}$, it follows that

$$f(x_k + s_k) - f(x_k) \leq -\delta_k^2.$$

Furthermore, since we chose $\frac{\nu}{1-\nu} \geq \gamma^2$, it follows that

$$\begin{aligned}
\phi_{k+1} - \phi_k &\leq (-\nu + (1-\nu)(\gamma^2 - 1)) \delta_k^2 \\
&\leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2.
\end{aligned} \tag{3.22}$$

Case 2(c): $[\mathcal{I}_k$ and $\mathcal{J}_k]$

As in Case 1(c), since the improvement factor estimate is sufficiently accurate, if the step is accepted, the change in ϕ is bounded by the decrease given by Equation (3.12). Thus,

$$\phi_{k+1} - \phi_k \leq (1-\nu) \left(\frac{1}{\gamma^2} - 1 \right) \delta_k^2. \tag{3.23}$$

Case 2(d): $[\mathcal{I}_k$ and $\mathcal{J}_k]$

Similar to Case 1(d), we may bound $f(x_k + s_k) - f(x_k) \leq \|\nabla f(x_k)\| \delta_k + \frac{\kappa_{\text{bh}}}{2} \delta_k^2$. By our case assumption, we have $\|\nabla f(x_k)\| \leq \zeta \delta_k$ and thus

$$\phi_{k+1} - \phi_k \leq \left(\nu \left(\zeta + \frac{\kappa_{\text{bh}}}{2} \right) + (1-\nu)(\gamma^2 - 1) \right) \delta_k^2. \tag{3.24}$$

We are now prepared to evaluate $\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_K)\| < \zeta \Delta_k \} \right]$. For brevity, let $E_2 = \mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_K)\| < \zeta \Delta_k \} \right]$. Since Case 2(a) occurs with probability at least $(1 - \delta_\nabla)(1 - \delta_f)$, Case 2(d) occurs with probability at most $\delta_\nabla \delta_f$, and Case 2(b) or Case 2(c) occur with probability at least $\delta_\nabla(1 - \delta_f) + (1 - \delta_\nabla)\delta_f$, we may apply Equations (3.21), (3.22), (3.23) and (3.24) to see

$$\begin{aligned}
E_2 &\leq ((1 - \delta_f)(1 - \delta_\nabla) + \delta_f(1 - \delta_\nabla) + (1 - \delta_f)\delta_\nabla) \left[(1 - \nu) \left(\frac{1}{\gamma^2} - 1 \right) \Delta_k^2 \right] \\
&\quad + \delta_f \delta_\nabla \left[\nu \left(\zeta + \frac{\kappa_{\text{bh}}}{2} \right) + (1 - \nu)(\gamma^2 - 1) \Delta_k^2 \right] \\
&= \left((1 - \delta_f \delta_\nabla)(1 - \nu) \left(\frac{1}{\gamma^2} - 1 \right) + \delta_f \delta_\nabla \left(\nu \left(\zeta + \frac{\kappa_{\text{bh}}}{2} \right) + (1 - \nu)(\gamma^2 - 1) \right) \right) \Delta_k^2.
\end{aligned}$$

Since, in Equation (3.7), we chose

$$\delta_f \delta_\nabla \leq \frac{1}{2 \left(1 + \frac{\gamma^2}{\gamma^2 - 1} \left(\frac{\nu}{1 - \nu} \left(\zeta + \frac{\kappa_{\text{bh}}}{2} \right) + \gamma^2 + 1 \right) \right)},$$

we have

$$\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi}, \{ \|\nabla f(X_k)\| < \zeta \Delta_k \} \right] \leq -\frac{1}{2}(1 - \nu) \left(1 - \frac{1}{\gamma^2} \right) \Delta_k^2. \quad (3.25)$$

Since $\gamma^2 \geq \frac{1}{2} \left(1 - \frac{1}{\gamma^2} \right)$, it follows that we may combine Equations (3.20) and (3.25) to yield

$$\mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi} \right] \leq -\frac{1}{2}(1 - \nu) \left(1 - \frac{1}{\gamma^2} \right) \Delta_k^2, \quad (3.26)$$

thus satisfying Equation (3.11) with $\sigma = \frac{1}{2}(1 - \nu) \left(1 - \frac{1}{\gamma^2} \right)$ and completing the proof. $\square$

3.4.2 liminf Convergence

As in [11, 19], the assumptions and conditions of the previous theorem immediately yield that

$$\liminf_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$$

with probability 1. The main idea of the proof is to assume the contrary, that is, that $\|\nabla f(X_k)\|$ is bounded away from zero. Since $\Delta_k \rightarrow 0$ almost surely, there exists an iteration such that, by Lemma 3.12, all subsequent iterations are accepted when the estimates are accurate. This is used to generate a contradiction by forming a submartingale from $\log_\gamma(\Delta_k)$ and showing that this must become arbitrarily large infinitely often with probability 1. Since the radii converge to zero, this cannot be.

Theorem 3.14. *Let Assumptions 3.1 - 3.6 and the conditions of Theorem 3.13 be satisfied. Then*

$$\mathbf{P} \left[\left\{ \liminf_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0 \right\} \right] = 1.$$

Proof. Since the assumptions of Theorem 3.13 are satisfied, we will condition on the almost sure event $\Delta_k \rightarrow 0$. Suppose the conclusion does not hold, that is, there exists an $\epsilon > 0$ such that $\|\nabla f(X_k)\| \geq \epsilon$ with some positive probability. Define

$$\widehat{\delta} = \min \left(\frac{\epsilon}{\zeta}, \frac{\delta_{\max}}{\gamma} \right).$$

Let $\{\widetilde{x}_k, \widetilde{\delta}_k\}$ be realizations of $\{X_k, \Delta_k\}$ for which $\|\nabla f(X_k)\| \geq \epsilon$ for all $k \geq 0$. Then, since $\widetilde{\delta}_k \rightarrow 0$ due to our conditioning, there exists a k_0 such that, for all $k \geq k_0$, $\widetilde{\delta}_k \leq \widehat{\delta}$. Define random variables with $x_n = \widetilde{x}_{n+k_0}$, $\delta_n = \widetilde{\delta}_{n+k_0}$.

Define a random variable with realizations $R_n(\omega) = r_n = \log_\gamma \left(\frac{\delta_n}{\widehat{\delta}} \right)$. Thus, if an iteration is successful, we have $\delta_{n+1} = \gamma \delta_n$ and thus $r_{n+1} = r_n + 1$. Similarly, if the iteration is unsuccessful, $\delta_{n+1} = \gamma^{-1} \delta_n$ and thus $r_{n+1} = r_n - 1$. Therefore, $|r_{n+1} - r_n| = 1$. Note that for our given realization, we have that $r_n \leq 0$ for all $n \geq 0$.

To show that R_n is a submartingale, recall the events $\mathcal{I}_k$ and $\mathcal{J}_k$. For a given $n \geq 0$, it follows that $\mathbf{P}[\{\mathcal{I}_{n+k_0} \cap \mathcal{J}_{n+k_0}\}] \geq (1 - \delta_\nabla)(1 - \delta_f)$. In such a case, Lemma 3.12 applies and the iteration is successful and thus $r_{n+1} = r_n + 1$. In all other cases, which occur with probability at most $1 - (1 - \delta_\nabla)(1 - \delta_f)$, we have $r_{n+1} \geq r_n - 1$. Therefore, letting $\mathcal{F}_{n-1}^{\mathcal{I}, \mathcal{J}}$ be the σ -algebra generated by $\mathcal{I}_{k_0}, \dots, \mathcal{I}_{k_0+n-1}$ and $\mathcal{J}_{k_0}, \dots, \mathcal{J}_{k_0+n-1}$, we have that

$$\begin{aligned} \mathbb{E} \left[r_{n+1} | \mathcal{F}_{n-1}^{\mathcal{I}, \mathcal{J}} \right] &\geq [(1 - \delta_\nabla)(1 - \delta_f)] (r_n + 1) + [1 - (1 - \delta_\nabla)(1 - \delta_f)] (r_n - 1) \\ &= r_n + (2(1 - \delta_\nabla)(1 - \delta_f) - 1) \\ &\geq r_n, \end{aligned}$$

since $(1 - \delta_\nabla)(1 - \delta_f) \geq 1/2$ is an implication of Equation (3.6).

Thus R_n is a submartingale with bounded increments and thus Corollary 3.8 applies. Since $R_{n+1} - R_n = \pm 1$, it follows that R_n can have no finite limit. Thus, Corollary 3.8 implies that $\limsup_{n \rightarrow \infty} R_n = \infty$ almost surely. In particular, R_n must be positive infinitely often with probability 1. Thus, realizations with $R_n(\omega) < 0$ for all $n \geq 0$ must occur with probability 0. This contradicts our assumption that such a sequence occurs with positive probability and the result follows. $\square$

3.4.3 lim Convergence

To show that $\|\nabla f(X_k)\| \rightarrow 0$ almost surely, we will need the following auxiliary lemma adapted from [19].

Lemma 3.15. *Let the assumptions Theorem 3.13 hold. For a given $\epsilon > 0$, define the random sequence $K_\epsilon = \{k \geq 0 : \|\nabla f(X_k)\| \geq \epsilon\}$. Then, with probability 1,*

$$\sum_{k \in K_\epsilon} \Delta_k < \infty.$$

Proof. Since the assumptions of Theorem 3.13 are satisfied, it follows that $\Delta_k \rightarrow 0$ almost surely. Therefore, we may condition on this event. For a given realization $\Delta_k(\omega) = \delta_k$, there exists a k_0 such that $\delta_k \leq \epsilon/\zeta$. Define the random variable $K_0(\omega) = k_0$ and let K be the (random) sequence of indices such that $K = \{k \in K_\epsilon | k \geq K_0\}$. For any $k \in K$, we have that $\|\nabla f(X_k)\| \geq \epsilon \geq \zeta \Delta_k$. Therefore, Case 1 of Theorem 3.13 applies and we may apply Equation (3.19). That is,

$$\begin{aligned} \mathbb{E} \left[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M, \Psi} \right] &\leq -\frac{1}{4} C_2 \nu \|\nabla f(X_k)\| \Delta_k \\ &\leq -\frac{1}{4} C_2 \nu \epsilon \Delta_k. \end{aligned}$$

Since Φ_k is bounded from below, we may sum both sides over $k \in K$ to conclude that $\sum_{k \in K} \Delta_k < \infty$. Since K_0 is almost surely finite and $K_\epsilon \subseteq K \cup \{k \in \mathbb{N} : k \leq K_0\}$, the result follows. $\square$

By using an adaptation of the proof found in [11], we are ready to show $\|\nabla f(X_k)\| \rightarrow 0$ almost surely. The main idea of the proof is to show that if, with any positive probability, there exists an $\epsilon > 0$ with $\limsup_{k \rightarrow \infty} \|\nabla f(X_k)\| \geq \epsilon$, then such a realization would have $\sum_{k \in K_\epsilon} \delta_k = \infty$. However, Lemma 3.15 implies that such events are only possible with probability zero.

Theorem 3.16. *Let the assumptions of Theorem 3.13 be satisfied. Then*

$$\mathbf{P} \left[\left\{ \lim_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0 \right\} \right] = 1.$$

Proof. Since the assumptions of Theorem 3.13 are satisfied, we may condition on the almost sure events $\Delta_k \rightarrow 0$ and $\liminf_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$. Let $\epsilon > 0$ and suppose that $\limsup_{k \rightarrow \infty} \|\nabla f(X_k)\| > 2\epsilon$ with positive probability. That is, with positive probability, the event $\|\nabla f(X_k)\| > 2\epsilon$ occurs infinitely often. Let K_ϵ be defined as in Lemma 3.15.

As in [11], we define a pair of integers (W', W'') as an *ascent pair* if $0 < W' < W''$, $\|\nabla f(X_{W'})\| \leq \epsilon$, $\|\nabla f(X_{W'+1})\| > \epsilon$, $\|\nabla f(X_{W''})\| > 2\epsilon$, and $\epsilon < \|\nabla f(X_w)\| \leq 2\epsilon$ for all $w \in (W', W'')$. Since $\liminf_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$ and $\|\nabla f(X_k)\| > 2\epsilon$ occurs infinitely often, there must be infinitely many distinct ascent pairs. For each pair, it follows $\{W' + 1, \dots, W''\} \subseteq K_\epsilon$. Note that this implies $\sum_{i=1}^{\infty} \sum_{j=W'_i+1}^{W''_i-1} \Delta_j \leq \sum_{k \in K_\epsilon} \Delta_k$. Let $\{(W'_\ell, W''_\ell)\}_{\ell=1}^{\infty}$ be a sequence

of distinct ascent pairs such that $\Delta_{W'_\ell} < \frac{\epsilon}{2\kappa_{\text{bh}}}$ for all $\ell \geq 0$. This is always possible by choosing i sufficiently large since we conditioned on $\Delta_k \rightarrow 0$.

Let us now consider realizations $x_k, \delta_k, \{(w'_\ell, w''_\ell)\}_\ell^\infty, k_\epsilon$ of $X_k, \Delta_k, \{(W'_\ell, W''_\ell)\}_{\ell=1}^\infty, K_\epsilon$ (respectively) for which $\|\nabla f(x_k)\| > \epsilon$ for $k \in k_\epsilon$. For each ℓ , we have

$$\begin{aligned} \epsilon &< \left| \|\nabla f(x_{w'_\ell})\| - \|\nabla f(x_{w''_\ell})\| \right| \\ &\leq \sum_{j=w'_\ell}^{w''_\ell-1} \left| \|\nabla f(x_j)\| - \|\nabla f(x_{j+1})\| \right| \\ &\leq \sum_{j=w'_\ell}^{w''_\ell-1} \|\nabla f(x_j) - \nabla f(x_{j+1})\| \end{aligned}$$

by the triangle inequality. It is a straight-forward extension of [35, Theorem 5.19] to show that

$$\|\nabla f(x_j) - \nabla f(x_{j+1})\| \leq \|x_j - x_{j+1}\| \kappa_{\text{bh}}.$$

Therefore, we have

$$\begin{aligned} \epsilon &< \kappa_{\text{bh}} \sum_{j=w'_\ell}^{w''_\ell-1} \|x_j - x_{j+1}\| \\ &\leq \kappa_{\text{bh}} \delta_{w'_\ell} + \kappa_{\text{bh}} \sum_{j=w'_\ell+1}^{w''_\ell-1} \delta_j \\ &\leq \epsilon/2 + \kappa_{\text{bh}} \sum_{j=w'_\ell+1}^{w''_\ell-1} \delta_j. \end{aligned}$$

This implies that $0 < \frac{\epsilon}{2\kappa_{\text{bh}}} < \sum_{j=w'_\ell}^{w''_\ell-1} \delta_j$. Since the sum over each ascent pair is uniformly

bounded away from zero and there are infinitely many such pairs, it follows that this implies

$\sum_{\ell=1}^{\infty} \sum_{j=w'_\ell+1}^{w''_\ell-1} \delta_j = \infty$ and hence $\sum_{k \in k_\epsilon} \delta_j = \infty$. However, Lemma 3.15 implies that this is only

occurs with probability zero. This contradicts our assumption that the event $\|\nabla f(X_k)\| > 2\epsilon$ occurs infinitely often with positive probability. Therefore, we have $\limsup_{k \rightarrow \infty} \|\nabla f(X_k)\| < 2\epsilon$ with probability 1. Since $\epsilon > 0$ was arbitrary, we must have $\limsup_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$ with probability 1 and therefore,

$$\mathbf{P} \left[\left\{ \lim_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0 \right\} \right] = 1.$$

□

3.5 Projection-Based Model Order Reduction

3.5.1 Linear Time-Invariant Systems

For many problems governed by parametric linear time-invariant systems, the spatial discretization of the governing equations results in a system of differential-algebraic equations (DAEs),

$$\begin{aligned}\mathbf{E}(\mathbf{x})\dot{\mathbf{y}}(t; \mathbf{x}) &= -\mathbf{A}(\mathbf{x})\mathbf{y}(t; \mathbf{x}) + \mathbf{B}(\mathbf{x})\mathbf{u}(t), \\ \mathbf{o}(t; \mathbf{x}) &= \mathbf{C}(\mathbf{x})^T \mathbf{y}(t; \mathbf{x}).\end{aligned}$$

where $\mathbf{y}$ is a discretization of the system state, t is the time variable, $\mathbf{x}$ is a set of parameters, $\mathbf{u}$ is a set of controls, and $\mathbf{o}$ is the set of observables. The matrices $\mathbf{E}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are discretizations of the system dynamics.

The Diffuse Optical Tomography Problem

As an illustrative example, we will consider the Diffuse Optical Tomography (DOT) problem in this paper. The parameters represent a Parametric Level Set (PaLS) parameterization of the anomaly, used in this context in [2, 22]. $\mathbf{A}$ represents the diffusion and absorption of photons. $\mathbf{B}$ represents the discretized set of n_{src} light sources. $\mathbf{C}^T \mathbf{y}$ represents the approximation of the n_{det} detectors via quadrature. In this context, the matrices $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{C}$ are independent of the parameterization and thus we will focus our attention on the system

$$\mathbf{E}\dot{\mathbf{y}}(t; \mathbf{x}) = -\mathbf{A}(\mathbf{x})\mathbf{y}(t; \mathbf{x}) + \mathbf{B}\mathbf{u}(t), \quad (3.27)$$

$$\mathbf{o}(t; \mathbf{x}) = \mathbf{C}^T \mathbf{y}(t; \mathbf{x}), \quad (3.28)$$

The extension of the topics to follow to systems with more parameter dependence is straightforward, but we will restrict our attention to systems of this form for brevity.

Through the use of the Fourier transform, we can replace this system of DAEs with a linear system for each frequency of the light. Let $\hat{\mathbf{y}}(\omega; \mathbf{x})$, $\hat{\mathbf{o}}(\omega; \mathbf{x})$, $\hat{\mathbf{u}}(\omega; \mathbf{x})$ denote the Fourier transform of $\mathbf{y}(t; \mathbf{x})$, $\mathbf{o}(t; \mathbf{x})$, and $\mathbf{u}(t; \mathbf{x})$, respectively. Taking the Fourier transform of Equations (3.27) and (3.28) yields

$$\hat{\mathbf{o}}(\omega; \mathbf{x}) = \Psi(\omega; \mathbf{x})\hat{\mathbf{u}}(\omega; \mathbf{x}), \quad \text{where} \quad (3.29)$$

$$\Psi(\omega; \mathbf{x}) = \mathbf{C}^T (\mathbf{K}(\omega; \mathbf{x}))^{-1} \mathbf{B}, \quad (3.30)$$

where $\omega \in \mathbb{R}$ and $\mathbf{K}(\omega; \mathbf{x}) = \frac{i\omega}{\nu} \mathbf{E} - \mathbf{A}(\mathbf{x})$, $i = \sqrt{-1}$, $\nu > 0$ is the speed of light in the medium. $\Psi(\omega; \mathbf{x})$ is called the *frequency response function* of the dynamical system given by Equations (3.27) and (3.28).

For a given frequency ω_j and input source i , we denote the predicted observations by the forward model $\widehat{\mathbf{o}}_i(\omega_j; \mathbf{x})$. We use the control input $\widehat{\mathbf{u}}_i(\omega_j) = \mathbf{e}_i$, representing the excitation of source i with a pure frequency ω_j . For a given parameter vector $\mathbf{x}$, computation of all detector values reduces to the evaluation of the frequency response function,

$$\mathbf{O}(\mathbf{x}) = [\Psi(\omega_1; \mathbf{x}), \dots, \Psi(\omega_{n_\omega}; \mathbf{x})]. \quad (3.31)$$

where $\mathbf{O}(\mathbf{x}) \in \mathbb{C}^{n_{\text{det}} \times n_{\text{src}} \cdot n_\omega}$.

Given a set of observations, $\mathbf{D}$, the parameter inversion problem is thus

$$\min_{\mathbf{x} \in \mathbb{R}^{n_p}} \|\mathbf{R}(\mathbf{x})\|_F^2 = \min_{\mathbf{x} \in \mathbb{R}^{n_p}} \|\mathbf{O}(\mathbf{x}) - \mathbf{D}\|_F^2. \quad (3.32)$$

We assume $\mathbf{D}$ contains additive noise at a known noise level. Since PaLS has a regularization effect [2], no further regularization is necessary.

For use in Newton-type methods, the differentiation of $\Psi(\omega_k; \mathbf{x})$ with respect to parameter x_ℓ is necessary. We can see that this Jacobian is given by

$$\frac{\partial}{\partial x_\ell} \Psi(\omega_k; \mathbf{x}) = -\mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left(\frac{\partial}{\partial x_\ell} \mathbf{K}(\mathbf{x}) \right) (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B}. \quad (3.33)$$

The computation of the Jacobian is dominated in solving the systems $\mathbf{K}(\omega_k; \mathbf{x})\mathbf{X} = \mathbf{B}$ and $(\mathbf{K}(\omega_k; \mathbf{x}))^T \mathbf{Y} = \mathbf{C}$ since $\frac{\partial}{\partial \mathbf{x}_\ell} \mathbf{K}(\mathbf{x})$ is diagonal and inexpensive to compute.

3.5.2 Projection-Based Model Order Reduction

Rather than a general quadratic model as is common in most trust region methods, we will use techniques from projection-based model order reduction [15, 22] to build a reduced transfer function, Ψ_R .

We assume that the state, $\mathbf{y}$, is well approximated by a n_r -dimensional subspace $\text{Range}(\mathbf{V})$ for some appropriately chosen $\mathbf{V} \in \mathbb{C}^{n \times n_r}$. Given a full-rank $\mathbf{W} \in \mathbb{C}^{n \times n_r}$, we enforce the Petrov-Galerkin condition

$$\mathbf{W}^T \left(\frac{1}{\nu} \mathbf{E} \mathbf{V} \dot{\mathbf{y}}_R(t; \mathbf{x}) + \mathbf{A}(\mathbf{x}) \mathbf{V} \mathbf{y}_R(t; \mathbf{x}) - \mathbf{B} \mathbf{u}(t) \right) = 0, \quad \mathbf{o}(t; \mathbf{x}) = \mathbf{C}^T \mathbf{V} \mathbf{y}_R(t; \mathbf{x})$$

to yield the reduced system

$$\frac{1}{\nu} \mathbf{E}_R \dot{\mathbf{y}}_R(t; \mathbf{x}) = -\mathbf{A}_R(\mathbf{x}) \mathbf{y}_R(t; \mathbf{x}) + \mathbf{B}_R \mathbf{u}(t), \quad \mathbf{o}_R(t; \mathbf{x}) = \mathbf{C}_R^T \mathbf{y}_R(t; \mathbf{x}) \quad (3.34)$$

and associated reduced frequency response function

$$\Psi_R(\omega; \mathbf{x}) = \mathbf{C}_R^T \left(\frac{i\omega}{\nu} \mathbf{E}_R - \mathbf{A}_R(\mathbf{x}) \right)^{-1} \mathbf{B}_R, \quad (3.35)$$

where $\mathbf{E}_R = \mathbf{W}^T \mathbf{E} \mathbf{V}$, $\mathbf{A}_R = \mathbf{W}^T \mathbf{A} \mathbf{V}$, $\mathbf{B}_R = \mathbf{W}^T \mathbf{B}$, and $\mathbf{C}_R = \mathbf{V}^T \mathbf{C}$. We similarly define $\mathbf{K}_R(\omega; \mathbf{x}) = \frac{i\omega}{\nu} \mathbf{E}_R - \mathbf{A}_R(\mathbf{x})$ and $\mathbf{O}_R(\mathbf{x}) = [\Psi_R(\omega_1; \mathbf{x}), \dots, \Psi_R(\omega_{n_\omega}; \mathbf{x})]$. We then can construct our model

$$m_\ell(\mathbf{x}) = \|\mathbf{O}_R(\mathbf{x}) - \mathbf{D}\|_F^2. \quad (3.36)$$

For more information on the selection of effective $\mathbf{V}$, $\mathbf{W}$, we refer interested readers to [12, 13, 15, 22]. For updating projection-based reduced order models, we will use the “residual-driven” update scheme from [30].

3.6 Estimates for the DOT Problem

In order to use Algorithm 2 and the framework developed in Section 3.2, it is necessary to obtain estimates of the error in our gradient approximation, $\|\nabla f(x_k) - \nabla m_\ell(x_k)\|$ and the improvement factor $f(x_k) - f(x_k + s_k)$ where the objective function, f is given by Equation (3.32) and our model, m_ℓ , is given by Equation (3.36). In both cases, we will make use of trace estimation techniques since, for $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\text{tr}(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_F^2$. We will make use of this directly for our gradient estimates and exploit the linearity of the trace operator for our improvement factor since $\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 = \text{tr}(\mathbf{A}^T \mathbf{A}) - \text{tr}(\mathbf{B}^T \mathbf{B}) = \text{tr}(\mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B})$.

3.6.1 Hutchinson Trace Estimation

Let $\mathbf{s} \in \mathbb{R}^n$ be a vector with independent, identically distributed entries from the Rademacher distribution $\left[\mathbf{P}[s_i = x] = \begin{cases} 1/2 & x = 1 \\ 1/2 & x = -1 \end{cases} \right]$. Then, it follows that

$$\begin{aligned} \mathbb{E}[\|\mathbf{A}\mathbf{s}\|_2^2] &= \mathbb{E}[\mathbf{s}^T \mathbf{A}^T \mathbf{A} \mathbf{s}] = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{A}^T \mathbf{A})_{ij} \mathbb{E}[s_i s_j] \\ &= \sum_{i=1}^n (\mathbf{A}^T \mathbf{A})_{ii} \mathbb{E}[s_i^2] + \sum_{\substack{i=1, j=1 \\ i \neq j}}^n (\mathbf{A}^T \mathbf{A})_{ij} \mathbb{E}[s_i s_j]. \end{aligned}$$

Since s_i and s_j are independent and have mean zero, $\mathbb{E}[s_i s_j] = \mathbb{E}[s_i] \mathbb{E}[s_j] = 0$. Since s_i also has unit variance, $\mathbb{E}[s_i^2] = 1$. Thus,

$$\mathbb{E}[\|\mathbf{A}\mathbf{s}\|_2^2] = \text{tr}(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_F^2.$$

Any distribution which has mean zero and unit variance may be used for the entries of $\mathbf{s}$, but the Rademacher distribution yields the minimum variance for the trace estimate [28]. This variance is given by the following lemma.

Lemma 3.17. *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Define tr_H^N to be a Hutchinson trace estimator with N samples, i.e. $\text{tr}_H^N(\mathbf{A}) = \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i^T \mathbf{A} \mathbf{w}_i$ where the entries in $\mathbf{w}_i$ are independently drawn from the Rademacher distribution. Then*

$$\text{Var}(\text{tr}_H^N(\mathbf{A})) = \frac{2(\|\mathbf{A}\|_F^2 - \sum_{i=1}^n A_{ii}^2)}{N}.$$

Proof. Since the $\mathbf{w}_i$ are independent (hence, uncorrelated) and identically distributed, it follows that

$$\text{Var}(\text{tr}_H^N(\mathbf{A})) = \frac{1}{N} \text{Var}(\text{tr}_H^1(\mathbf{A})).$$

From [28], $\text{Var}(\mathbf{w}_i^T \mathbf{A} \mathbf{w}_i) = 2(\|\mathbf{A}\|_F^2 - \sum_{i=1}^n A_{ii}^2)$. The result then follows. $\square$

3.6.2 Estimates for $f(x_k) - f(x_k + s_k)$

Equipped with the variance of this trace estimate for an arbitrary symmetric matrix, we turn to Cantelli's inequality to provide a sufficient number of samples for upper or lower multiplicative bounds.

Lemma 3.18. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be symmetric, positive semi-definite matrices with $\text{tr}(\mathbf{A} - \mathbf{B}) \neq 0$. Then, for $\delta \in (0, 1]$ and $\epsilon > 0$, if N satisfies $N \geq \frac{2(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2)(1 - \delta)}{\text{tr}(\mathbf{A} - \mathbf{B})^2 \delta \epsilon^2}$, then*

$$\mathbf{P} \left[\frac{\text{tr}_H^N(\mathbf{A} - \mathbf{B})}{\text{tr}(\mathbf{A} - \mathbf{B})} \geq (1 - \epsilon) \right] \geq 1 - \delta.$$

Proof. Let $\lambda = -\epsilon \text{tr}(\mathbf{A} - \mathbf{B})$. By Cantelli's inequality, it follows that

$$\begin{aligned} \mathbf{P} [\text{tr}_H^N(\mathbf{A} - \mathbf{B}) - \text{tr}(\mathbf{A} - \mathbf{B}) \geq \lambda] &= \mathbf{P} [\text{tr}_H^N(\mathbf{A} - \mathbf{B}) \geq (1 - \epsilon) \text{tr}(\mathbf{A} - \mathbf{B})] \\ &\begin{cases} \geq 1 - \frac{\text{Var}(\text{tr}_H^N(\mathbf{A} - \mathbf{B}))}{\text{Var}(\text{tr}_H^N(\mathbf{A} - \mathbf{B})) + \lambda^2} & \lambda < 0 \\ \leq \frac{\text{Var}(\text{tr}_H^N(\mathbf{A} - \mathbf{B}))}{\text{Var}(\text{tr}_H^N(\mathbf{A} - \mathbf{B})) + \lambda^2} & \lambda > 0 \end{cases}. \end{aligned}$$

Suppose $\text{tr}(\mathbf{A} - \mathbf{B}) > 0$. Therefore, $\lambda < 0$. Applying Lemma 3.17, we find that

$$\begin{aligned} \mathbf{P} \left[\frac{\text{tr}_H^N(\mathbf{A} - \mathbf{B})}{\text{tr}(\mathbf{A} - \mathbf{B})} \geq (1 - \epsilon) \right] &= \mathbf{P} [\text{tr}_H^N(\mathbf{A} - \mathbf{B}) \geq (1 - \epsilon) \text{tr}(\mathbf{A} - \mathbf{B})] \\ &\geq 1 - \frac{2(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2) / N}{2(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2) / N + \epsilon^2 \text{tr}(\mathbf{A} - \mathbf{B})^2}. \end{aligned}$$

Now suppose $\text{tr}(\mathbf{A} - \mathbf{B}) < 0$ and thus $\lambda > 0$. Then

$$\begin{aligned} \mathbf{P} \left[\frac{\text{tr}_H^N(\mathbf{A} - \mathbf{B})}{\text{tr}(\mathbf{A} - \mathbf{B})} \geq (1 - \epsilon) \right] &= \mathbf{P} [\text{tr}_H^N(\mathbf{A} - \mathbf{B}) \leq (1 - \epsilon) \text{tr}(\mathbf{A} - \mathbf{B})] \\ &= 1 - \mathbf{P} [\text{tr}_H^N(\mathbf{A} - \mathbf{B}) \geq (1 - \epsilon) \text{tr}(\mathbf{A} - \mathbf{B})] \\ &\geq 1 - \frac{2 \left(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2 \right) / N}{2 \left(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2 \right) / N + \epsilon^2 \text{tr}(\mathbf{A} - \mathbf{B})^2}. \end{aligned}$$

If N satisfies the hypotheses, then the result follows. $\square$

Lemma 3.19. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be symmetric, positive semi-definite matrices with $\text{tr}(\mathbf{A} - \mathbf{B}) \neq 0$. Then, for $\delta \in (0, 1]$ and $\epsilon > 0$, if N satisfies $N \geq \frac{2 \left(\|\mathbf{A} - \mathbf{B}\|_F^2 - \sum_{i=1}^n (A_{ii} - B_{ii})^2 \right) (1 - \delta)}{\text{tr}(\mathbf{A} - \mathbf{B})^2 \delta \epsilon^2}$, then*

$$\mathbf{P} \left[\frac{\text{tr}_H^N(\mathbf{A} - \mathbf{B})}{\text{tr}(\mathbf{A} - \mathbf{B})} \leq (1 + \epsilon) \right] \geq 1 - \delta.$$

Proof. The proof is nearly identical to that of Lemma 3.18 with the $\lambda = \epsilon \text{tr}(\mathbf{A} - \mathbf{B})$ and swapping the cases. $\square$

We are now equipped to find a sufficient number of samples required to be a δ_f -probabalistically $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimate.

Theorem 3.20. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ with $\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 \neq 0$. Define ϕ by*

$$\phi = \frac{1 - \delta_f}{\delta_f} \left(\frac{\alpha_L^2 + \alpha_U^2 + \sqrt{(\alpha_L^2 + \alpha_U^2)^2 + 4\alpha_L^2\alpha_U^2 \left(\frac{\delta_f(2 - \delta_f)}{(1 - \delta_f)^2} \right)}}{2\alpha_L^2\alpha_U^2} \right).$$

If $N > 0$ satisfies

$$N \geq \left(\frac{2 \left(\|\mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B}\|_F^2 - \sum_{i=1}^n ((\mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B})_{ii})^2 \right)}{\text{tr}(\mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B})^2} \right) \phi,$$

then $\text{tr}_H^N(\mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B})$ is a δ_f -probabilistically $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimate of $\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2$.

Proof. Without loss of generality, let $\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 > 0$. For brevity, let $\mathbf{C} = \mathbf{A}^T \mathbf{A} - \mathbf{B}^T \mathbf{B}$. We begin by recalling the definition of a δ_f -probabilistically $\begin{pmatrix} \alpha_U \\ \alpha_L \end{pmatrix}$ -accurate estimate. We seek a sufficient bound on N to ensure that, with probability at least $1 - \delta_f$, $1 - \alpha_L \leq \frac{\text{tr}_H^N(\mathbf{C})}{\text{tr}(\mathbf{C})} \leq 1 + \alpha_U$. Using the relations derived in the proofs of Lemmas 3.18 and 3.19 and the union bound, see that this occurs with probability,

$$\begin{aligned} \mathbf{P} \left[1 - \alpha_L \leq \frac{\text{tr}_H^N(\mathbf{C})}{\text{tr}(\mathbf{C})} \leq 1 + \alpha_U \right] &= 1 - \mathbf{P} \left[\left\{ 1 - \alpha_L \geq \frac{\text{tr}_H^N(\mathbf{C})}{\text{tr}(\mathbf{C})} \right\} \parallel \left\{ \frac{\text{tr}_H^N(\mathbf{C})}{\text{tr}(\mathbf{C})} \geq 1 + \alpha_U \right\} \right] \\ &\geq 1 - \left(\frac{\sigma^2/N}{\sigma^2/N + \alpha_L^2 \text{tr}(\mathbf{C})^2} + \frac{\sigma^2/N}{\sigma^2/N + \alpha_U^2 \text{tr}(\mathbf{C})^2} \right), \end{aligned}$$

where $\sigma^2 = 2(\|\mathbf{C}\|_F^2 - \sum_{i=1}^n C_{ii}^2)$ is the variance of a single estimate, as per Lemma 3.17. We now turn our attention to the parenthetical quantity. If this is less than δ_f , we have shown the result. If $\sigma^2 = 0$, each estimate is exact with probability 1, and we're done. Otherwise, we may write

$$\frac{\sigma^2/N}{\sigma^2/N + \alpha_L^2 \text{tr}(\mathbf{C})^2} + \frac{\sigma^2/N}{\sigma^2/N + \alpha_U^2 \text{tr}(\mathbf{C})^2} = \frac{1}{1 + \psi \alpha_L^2} + \frac{1}{1 + \psi \alpha_U^2},$$

where $\psi = \frac{N \text{tr}(\mathbf{C})^2}{\sigma^2}$. To satisfy our requirements, we seek a lower bound for ψ such that

$$\frac{1}{1 + \psi \alpha_L^2} + \frac{1}{1 + \psi \alpha_U^2} \leq \delta_f.$$

This can be rewritten as a quadratic (in ψ) polynomial inequality that is satisfied for $\psi \geq \phi$. Solving for N yields $N \geq \frac{\phi \sigma^2}{\text{tr}(\mathbf{C})^2}$, which is the stated bound. $\square$

Note: We could have used the results of Lemmas 3.18 and 3.19 and the union bound directly to yield $N \geq \frac{2(\|\mathbf{C}\|_F^2 - \sum_{i=1}^n (C_{ii})^2)(2 - \delta)}{\text{tr}(\mathbf{C})^2 \delta \min(\alpha_L, \alpha_U)^2}$. However, assuming that $\alpha_U \gg \alpha_L$, this is significantly larger than the bound derived in Theorem 3.20.

Note: At first glance, it seems that the matrix-dependant quantity scales like $\frac{1}{\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2}$, which would be disastrous since that difference may be arbitrarily small, especially for smaller trust regions. In practice, this is typically balanced by the numerator. In our experience, for the DOT problem, we have not observed a value larger than 2 for the matrix-dependent factor.

Example 3.21. For $\alpha_L = 4/5, \alpha_U = 4$, Table 3.1 contains values of $\lceil \phi \rceil$ for varying δ_f .

	$\delta_f = 1/2$	$\delta_f = 1/4$	$\delta_f = 1/10$	$\delta_f = 1/20$	$\delta_f = 1/50$	$\delta_f = 1/100$
$\lceil \phi \rceil$	2	6	15	31	80	161

Table 3.1: Scaling of the factor ϕ for the sufficient bound of Theorem 3.20 with $\alpha_L = 4/5, \alpha_U = 4$

3.6.3 Estimates of $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|$

With $\Psi(\omega_k; \mathbf{x})$ defined as in Section 3.5.1, we may write

$$\begin{aligned} f(\mathbf{x}) &= \|\Psi(\omega_1; \mathbf{x}), \dots, \Psi(\omega_{n_\omega}; \mathbf{x}) - \mathbf{D}\|_F^2 \\ &= \|\text{vec}([\Psi(\omega_1; \mathbf{x}), \dots, \Psi(\omega_{n_\omega}; \mathbf{x})] - \mathbf{D})\|_2^2, \end{aligned}$$

where $\text{vec}(\mathbf{A})$ is the vectorization operator that stacks columns of $\mathbf{A}$. Define

$$\mathbf{R} = \text{vec}([\Psi(\omega_1; \mathbf{x}), \dots, \Psi(\omega_{n_\omega}; \mathbf{x})] - \mathbf{D})$$

and $\mathbf{J}$ as

$$\mathbf{J} = \left[\text{vec} \left(\frac{\partial \Psi(\omega_1; \mathbf{x})}{\partial \mathbf{x}_1} \right), \dots, \text{vec} \left(\frac{\partial \Psi(\omega_{n_\omega}; \mathbf{x})}{\partial \mathbf{x}_{n_p}} \right) \right],$$

where $\frac{\partial \Psi}{\partial \mathbf{x}_i}$ is given by Equation (3.33). Then we find $\nabla f(\mathbf{x}) = 2\mathbf{J}^T \mathbf{R}$. Letting $\mathbf{R}_R$ and $\mathbf{J}_R$ be defined analogously for Ψ_R , we have $\nabla m_\ell(\mathbf{x}) = 2\mathbf{J}_R^T \mathbf{R}_R$. Therefore,

$$\begin{aligned} \|\nabla m_\ell(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 &= 2 \|\mathbf{J}^T \mathbf{R} - \mathbf{J}_R^T \mathbf{R}_R\|_2 \\ &= 2 \|\mathbf{J}^T \mathbf{R} - \mathbf{J}^T \mathbf{R}_R + \mathbf{J}^T \mathbf{R}_R - \mathbf{J}_R^T \mathbf{R}_R\|_2 \\ &\leq 2 \|\mathbf{J}^T (\mathbf{R} - \mathbf{R}_R)\|_2 + 2 \|(\mathbf{J} - \mathbf{J}_R)^T \mathbf{R}_R\|_2 \\ &\leq 2 (\|\mathbf{J}\|_2 \|\mathbf{R} - \mathbf{R}_R\|_2 + \|\mathbf{J} - \mathbf{J}_R\|_2 \|\mathbf{R}_R\|_2) \\ &\leq 2 (\|\mathbf{J}\|_F \|\mathbf{R} - \mathbf{R}_R\|_2 + \|\mathbf{J} - \mathbf{J}_R\|_F \|\mathbf{R}_R\|_2) \\ &= 2 \left(\|\mathbf{J}\|_F \|\mathbf{R} - \mathbf{R}_R\|_2 + \|\mathbf{J} - \mathbf{J}_R\|_F \sqrt{m_\ell(\mathbf{x})} \right). \end{aligned} \tag{3.37}$$

Thus, we have three quantities to bound: $\|\mathbf{J}\|_F$, $\|\mathbf{R} - \mathbf{R}_R\|_2$, and $\|\mathbf{J} - \mathbf{J}_R\|_F$.

Bounding $\|\mathbf{J}\|_F$

To bound the first, we turn again to its definition:

$$\begin{aligned}
\|\mathbf{J}\|_F^2 &= \left\| \left[\text{vec} \left(\frac{\partial \Psi(\omega_1; \mathbf{x})}{\partial x_1} \right), \dots, \text{vec} \left(\frac{\partial \Psi(\omega_{n_\omega}; \mathbf{x})}{\partial x_{n_p}} \right) \right] \right\|_F^2 \\
&= \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \Psi(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2 \\
&= \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left(\frac{\partial}{\partial x_\ell} \mathbf{K}(\mathbf{x}) \right) (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \right\|_F^2 \\
&\leq \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \right\|_F^2 \cdot \left\| (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \right\|_F^2 \left\| \frac{\partial}{\partial x_\ell} \mathbf{K}(\mathbf{x}) \right\|_F^2 \\
&\leq \left(\max_{i=1, \dots, n_\omega} \left(\left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F \cdot \left\| (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B} \right\|_F \right) \right)^2 \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial}{\partial x_\ell} \mathbf{K}(\mathbf{x}) \right\|_F^2.
\end{aligned}$$

Let $\mu(\mathbf{x}) = \max_{i=1, \dots, n_\omega} \left(\left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F \cdot \left\| (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B} \right\|_F \right)$. Computing this at every parameter is undesirable since its cost is equivalent to a full order Jacobian evaluation. This leaves us with a number of options:

1. Estimate the norms. If the estimates from Section 3.6.2 are used, then we already have access to solutions to $(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} (\mathbf{B}\mathbf{s})$ where $\mathbf{s} \in \mathbb{R}^{n_{\text{src}}}$ has entries distributed according to the Rademacher distribution. Since $\text{tr}(\mathbf{A}^T \mathbf{A}) = \text{tr}(\mathbf{A} \mathbf{A}^T)$ for all matrices $\mathbf{A}$, we could divide the samples equally between solutions to $(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} (\mathbf{B}\mathbf{s})$ and $((\mathbf{C}\tilde{\mathbf{s}})^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1})$ where $\tilde{\mathbf{s}} \in \mathbb{R}^{n_{\text{det}}}$ with Rademacher-distributed entries. For no additional work, this would yield an equivalent estimate for the improvement factor and estimates of $\left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F \cdot \left\| (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B} \right\|_F$.

Lemma 3.22. *Let $T_{\mathbf{B}} = \left\| (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B} \right\|_F^2$, $T_{\mathbf{C}} = \left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F^2$, and $T_{\mathbf{B}}^H, T_{\mathbf{C}}^H$ be the Hutchinson trace estimates with N samples thereof. If $0 < \epsilon \leq 1$ and $N \geq 6\epsilon^{-2} \log(2/\delta)$, then $\sqrt{T_{\mathbf{B}}^H \cdot T_{\mathbf{C}}^H}$ is a δ -probabilistically ϵ -accurate underestimate of $\sqrt{T_{\mathbf{B}} \cdot T_{\mathbf{C}}}$.*

Proof. By an intermediate result in [34], it follows that N is sufficiently large for

$T_{\mathbf{B}}^H, T_{\mathbf{C}}^H$ to be $\frac{\delta}{2}$ -probabilistically ϵ -accurate underestimates. Thus,

$$\begin{aligned}
 \mathbf{P} \left[1 - \epsilon \leq \frac{\sqrt{T_{\mathbf{B}}^H \cdot T_{\mathbf{C}}^H}}{\sqrt{T_{\mathbf{B}} \cdot T_{\mathbf{C}}}} \right] &= \mathbf{P} \left[(1 - \epsilon)^2 \leq \frac{T_{\mathbf{B}}^H \cdot T_{\mathbf{C}}^H}{T_{\mathbf{B}} \cdot T_{\mathbf{C}}} \right] \\
 &\geq \mathbf{P} \left[\left\{ 1 - \epsilon \leq \frac{T_{\mathbf{B}}^H}{T_{\mathbf{B}}} \right\} \cap \left\{ 1 - \epsilon \leq \frac{T_{\mathbf{C}}^H}{T_{\mathbf{C}}} \right\} \right] \\
 &\geq \left(1 - \frac{\delta}{2} \right) + \left(1 - \frac{\delta}{2} \right) - 1 \\
 &= 1 - \delta.
 \end{aligned}$$

For intuition on the first inequality, see Figure 3.1. □

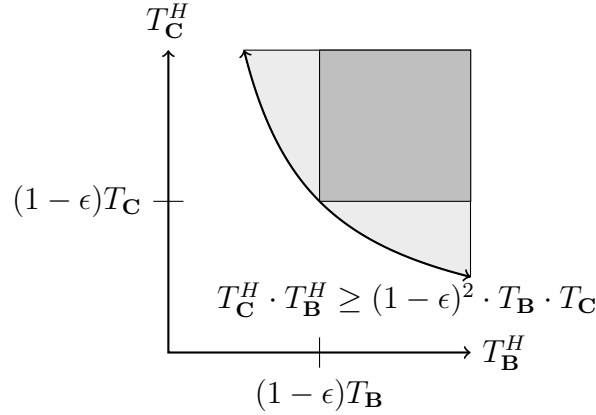


Figure 3.1: $\mathbf{P} [T_{\mathbf{C}}^H \cdot T_{\mathbf{B}}^H \geq (1 - \epsilon)^2 \cdot T_{\mathbf{B}} \cdot T_{\mathbf{C}}] \geq \mathbf{P} [\{T_{\mathbf{B}}^H \geq (1 - \epsilon)T_{\mathbf{B}}\} \cap \{T_{\mathbf{C}}^H \geq (1 - \epsilon)T_{\mathbf{C}}\}]$

By considering the case where every estimate is ϵ -accurate, we can bound $\mu(\mathbf{x})$.

Theorem 3.23. *If $0 < \epsilon \leq 1$ and $N \geq 6\epsilon^{-2} \log(2n_{\omega}/\delta)$, then*

$$\max_{i=1, \dots, n_{\omega}} \left(\sqrt{\text{tr}_H^N \left(\|(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B}\|_F^2 \right) \cdot \text{tr}_H^N \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F^2 \right)} \right)$$

is a δ -probabilistically ϵ -accurate underestimate of $\mu(\mathbf{x})$.

Proof. Let $\mu_i^H(\mathbf{x}) = \sqrt{\text{tr}_H^N \left(\|(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B}\|_F^2 \right) \cdot \text{tr}_H^N \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F^2 \right)}$. By Lemma 3.22, it follows that N is sufficiently large for each μ_i^H to be a $\frac{\delta}{n_{\omega}}$ -probabalistically ϵ -accurate underestimate. Therefore,

$$\begin{aligned}
\mathbf{P} \left[(1 - \epsilon) \mu(\mathbf{x}) \leq \max_{i=1, \dots, n_\omega} (\mu_i^H) \right] &\geq \mathbf{P} \left[\bigcap_{i=1}^{n_\omega} \{ (1 - \epsilon) \|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F \cdot \|(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B}\|_F \leq \mu_i^H \} \right] \\
&\geq \sum_{i=1}^{n_\omega} \left(1 - \frac{\delta}{n_\omega} \right) - (n_\omega - 1) \\
&= 1 - \delta.
\end{aligned}$$

□

2. Approximate $\mu(\mathbf{x})$ by $\mu(\mathbf{x}_0)$. Since we require the solutions to $\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}_0))^{-1}$ and $(\mathbf{K}(\omega_i; \mathbf{x}_0))^{-1} \mathbf{B}$ to construct the ROM, we have access to the exact value of $\mu(\mathbf{x}_0)$. In practice, for the DOT problem, we have observed very little change in $\mu(\mathbf{x})$ over the course of the optimization. For a simplified problem (just the discretized Laplacian), Sariaydin [36] showed that

$$\frac{\|(\mathbf{K}(\omega_i; \mathbf{x}_0))^{-1} \mathbf{B} - (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B}\|_F}{\|(\mathbf{K}(\omega_i; \mathbf{x}_0))^{-1} \mathbf{B}\|_F} \leq Ch,$$

where C is a constant and h is the mesh width. Since we observe similar behavior for our problem and for $\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F$, it is reasonable to use a value proportional to $\mu(\mathbf{x}_0)$ to bound $\mu(\mathbf{x})$.

Bounding $\|\mathbf{R} - \mathbf{R}_R\|_2$

To bound the difference in the full-order and reduced-order residuals (in the Gauss-Newton sense), we adapt the following theorem:

Theorem 3.24 ([30]). *Let $\mathbf{Q}_i = \mathbf{K}(\omega_i; \mathbf{x}) \mathbf{V} (\mathbf{K}_R(\omega_i; \mathbf{x}))^{-1} \mathbf{W}^T$. Then*

$$\|\mathbf{R} - \mathbf{R}_R\|_2^2 \leq \max_{i=1, \dots, n_\omega} \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F \right)^2 \sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i) \mathbf{B}\|_F^2.$$

Since computing $\mathbf{Q}_i$ (or more accurately, the action thereof) does not require any large-scale linear solves, computing the $\|(\mathbf{I} - \mathbf{Q}_i) \mathbf{B}\|_F^2$ terms is trivial. For the $\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F$ term, this bound inherits the same issues and solutions as in Section 3.6.3. For the estimate option, since only the $\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}$ term is involved, an equal number of samples yields an accurate estimate with significantly higher probability.

Theorem 3.25. *Let $\mu_C(\mathbf{x}) = \max_{i=1, \dots, n_\omega} (\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F)$ and*

$$\mu_{C,i}^H = \sqrt{\text{tr}_H^N \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F^2 \right)}.$$

If $N \geq 6\epsilon^{-2} \log(2n_\omega/\delta)$, then $\max_{i=1,\dots,n_\omega}(\mu_{C,i}^H)$ is a $\left(n_\omega \left(\frac{\delta}{2n_\omega}\right)^{-(2-\epsilon)^2}\right)$ -probabilistically $(2\epsilon - \epsilon^2)$ -accurate underestimate of $\mu_C(\mathbf{x})$.

Proof. Let $\hat{\epsilon} = 2\epsilon - \epsilon^2$. Since

$$\mathbf{P} \left[(1 - \epsilon) \left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F \leq \mu_{C,i}^H \right] = \mathbf{P} \left[(1 - \hat{\epsilon}) \leq \text{tr}_H^N \left(\left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F^2 \right) \right],$$

we use an intermediate result from [34] to see that this value of N is sufficient for $\mu_{C,i}^H$ to be a $\frac{\hat{\delta}}{n_\omega}$ -probabalistically $(2\epsilon - \epsilon^2)$ -accurate underestimate for any $\hat{\delta}$ that satisfies $6\epsilon^{-2} \log(2n_\omega/\delta) \geq 6\hat{\epsilon}^{-2} \log(n_\omega/\hat{\delta})$. In particular, this holds with equality with $\hat{\delta} = n_\omega \left(\frac{\delta}{2n_\omega}\right)^{-(2-\epsilon)^2}$. Therefore, we find

$$\begin{aligned} \mathbf{P} \left[(1 - \epsilon) \mu_C(\mathbf{x}) \leq \max_{i=1,\dots,n_\omega} (\mu_{C,i}^H) \right] &\geq \mathbf{P} \left[\bigcap_{i=1}^{n_\omega} \left\{ (1 - \epsilon) \left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F \leq \mu_{C,i}^H \right\} \right] \\ &= \mathbf{P} \left[\bigcap_{i=1}^{n_\omega} \left\{ (1 - \hat{\epsilon}) \left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F^2 \leq \text{tr}_H^N \left(\left\| \mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \right\|_F^2 \right) \right\} \right] \\ &\geq \sum_{i=1}^{n_\omega} \left(1 - \frac{\hat{\delta}}{n_\omega} \right) - (n_\omega - 1) \\ &= 1 - \hat{\delta}. \end{aligned}$$

□

Bounding $\|\mathbf{J} - \mathbf{J}_R\|_F$

Much like in Section 3.6.3, we will use the sub-multiplicative nature of the Frobenius norm to split off a factor of $\mu(\mathbf{x})$.

Lemma 3.26. Define $\mathbf{Q}_k = \mathbf{K}(\omega_k; \mathbf{x}) \mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T$ and $\mathbf{P}_k = \mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T \mathbf{K}(\omega_k; \mathbf{x})$. Then

$$\|\mathbf{J} - \mathbf{J}_R\|_F^2 \leq \mu^2(\mathbf{x}) \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2.$$

Proof. We begin by noting

$$\begin{aligned}
\frac{\partial \Psi(\omega_k; \mathbf{x})}{\partial x_\ell} - \frac{\partial \Psi_R(\omega_k; \mathbf{x})}{\partial x_\ell} &= \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \\
&\quad - \mathbf{C}^T \mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T \mathbf{B} \\
&= \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left[\frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \right. \\
&\quad \left. - (\mathbf{K}(\omega_k; \mathbf{x}) \mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T) \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} (\mathbf{V} (\mathbf{K}_R(\omega_k; \mathbf{x}))^{-1} \mathbf{W}^T \mathbf{K}(\omega_k; \mathbf{x})) \right] \\
&\quad \cdot (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \\
&= \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left[\frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right] (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B}
\end{aligned}$$

Therefore, we find

$$\begin{aligned}
\|\mathbf{J} - \mathbf{J}_R\|_F^2 &= \left\| \left[\text{vec} \left(\frac{\partial \Psi(\omega_1; \mathbf{x})}{\partial x_1} \right), \dots \right] - \left[\text{vec} \left(\frac{\partial \Psi_R(\omega_1; \mathbf{x})}{\partial x_1} \right), \dots \right] \right\|_F^2 \\
&= \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \Psi(\omega_k; \mathbf{x})}{\partial x_\ell} - \frac{\partial \Psi_R(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2 \\
&= \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \left[\frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right] (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \right\|_F^2 \\
&\leq \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \right\|_F^2 \cdot \left\| (\mathbf{K}(\omega_k; \mathbf{x}))^{-1} \mathbf{B} \right\|_F^2 \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2 \\
&\leq (\mu(\mathbf{x}))^2 \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2.
\end{aligned}$$

□

While the bound in Lemma 3.26 does not require any solutions to large-scale linear systems, $\mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k$ is a dense $n \times n$ matrix. Rather than actually form these matrices, we use the Hutchinson trace estimator.

Lemma 3.27. *For a given $0 < \hat{\epsilon} \leq 1$, and $0 < \hat{\delta} < 1$, if $M \in \mathbb{N}$ satisfies $M \geq 6(2\hat{\epsilon} - \hat{\epsilon}^2)^{-2} \log(n_p n_\omega / \hat{\delta})$, then*

$$d(\mathbf{x}) = \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \text{tr}_H^M \left(\left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2 \right)}$$

is a $\widehat{\delta}$ -probabalistically $\widehat{\epsilon}$ -accurate underestimate of $\sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2}$.

Proof. Let $d_{\ell,k} = \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2$. By application of the results from [34], M is sufficiently large so that $\text{tr}_H^M(d_{\ell,k})$ is a $\frac{\widehat{\delta}}{n_p n_\omega}$ -probabalistically $(2\widehat{\epsilon} - \widehat{\epsilon}^2)$ -accurate underestimate. Thus,

$$\begin{aligned}
1 - \widehat{\delta} &\leq 1 - \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \mathbf{P} \left[(1 - (2\widehat{\epsilon} - \widehat{\epsilon}^2)) d_{\ell,k} \geq \text{tr}_H^M(d_{\ell,k}) \right] \\
&\leq 1 - \mathbf{P} \left[\bigcup_{\ell=1}^{n_p} \bigcup_{k=1}^{n_\omega} \{ (1 - (2\widehat{\epsilon} - \widehat{\epsilon}^2)) d_{\ell,k} \geq \text{tr}_H^M(d_{\ell,k}) \} \right] \\
&= \mathbf{P} \left[\bigcap_{\ell=1}^{n_p} \bigcap_{k=1}^{n_\omega} \{ (1 - \widehat{\epsilon})^2 d_{\ell,k} \leq \text{tr}_H^M(d_{\ell,k}) \} \right] \\
&\leq \mathbf{P} \left[(1 - \widehat{\epsilon})^2 \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} d_{\ell,k} \leq \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \text{tr}_H^M(d_{\ell,k}) \right] \\
&= \mathbf{P} \left[(1 - \widehat{\epsilon}) \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} d_{\ell,k}} \leq \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \text{tr}_H^M(d_{\ell,k})} \right].
\end{aligned}$$

□

Estimates of $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|$

We are now prepared to bound $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|$. Let $\mu_{\mathbf{C}}(\mathbf{x}) = \max_{i=1,\dots,n_\omega} (\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F)$. By using the results of Sections 3.6.3 to 3.6.3, we may write

$$\begin{aligned}
\|\nabla m_\ell(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 &\leq 2 \left(\|\mathbf{J}\|_F \|\mathbf{R} - \mathbf{R}_R\|_2 + \|\mathbf{J} - \mathbf{J}_R\|_F \sqrt{m_\ell(\mathbf{x})} \right) \\
&\leq 2 \left(\left(\mu(\mathbf{x}) \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial}{\partial x_\ell} \mathbf{K}(\mathbf{x}) \right\|_F^2} \right) \left(\mu_{\mathbf{C}}(\mathbf{x}) \sqrt{\sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i) \mathbf{B}\|_F^2} \right) \right. \\
&\quad \left. + \left(\mu(\mathbf{x}) \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2} \right) \sqrt{m_\ell(\mathbf{x})} \right) \\
&= 2\mu(\mathbf{x}) \left(\mu_{\mathbf{C}}(\mathbf{x}) \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2} \cdot \sqrt{\sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i) \mathbf{B}\|_F^2} \right. \\
&\quad \left. + \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2} \sqrt{m_\ell(\mathbf{x})} \right).
\end{aligned}$$

Thus, depending on the chosen resolution of estimating or bounding μ , $\mu_{\mathbf{C}}$, we arrive at the following two theorems.

Theorem 3.28. *Let $0 < \epsilon, \delta < 1$. Define*

- $\hat{\epsilon} = \frac{1}{81} \epsilon (3 - \epsilon) (\epsilon + 3) (\epsilon + 6)$,
- $\mu^{H,N}(\mathbf{x}) = \max_{i=1,\dots,n_\omega} \left(\sqrt{\text{tr}_H^N \left(\|(\mathbf{K}(\omega_i; \mathbf{x}))^{-1} \mathbf{B}\|_F^2 \right) \cdot \text{tr}_H^N \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F^2 \right)} \right)$,
- $\mu_C^{H,N}(\mathbf{x}) = \max_{i=1,\dots,n_\omega} \sqrt{\text{tr}_H^N \left(\|\mathbf{C}^T (\mathbf{K}(\omega_i; \mathbf{x}))^{-1}\|_F^2 \right)}$,
- $d^{H,M}(\mathbf{x}) = \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \text{tr}_H^M \left(\left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2 \right)}$,
- and $d(\mathbf{x}) = \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2}$.

If $N \geq 6\hat{\epsilon}^{-2} \log(4n_\omega/\delta)$ and $M \geq 6\hat{\epsilon}^{-2} \log(4n_p n_\omega/\delta)$, then

$$2\mu^{H,N}(\mathbf{x}) \left(\mu_C^{H,N}(\mathbf{x}) \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2} \cdot \sqrt{\sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i) \mathbf{B}\|_F^2} + d^{H,M}(\mathbf{x}) \sqrt{m_\ell(\mathbf{x})} \right)$$

is a δ -probabalistically ϵ -accurate underestimate of $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|_2$.

Proof. For brevity, let $\kappa_1 = \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2} \cdot \sqrt{\sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i)\mathbf{B}\|_F^2}$ and $\kappa_2 = \sqrt{m_\ell(x_k)}$. Therefore, we find

$$\begin{aligned} \mathbf{P} \left[(1 - \epsilon) \|\nabla m_\ell(x_k) - \nabla f(x_k)\|_2 \leq 2\mu^{H,N}(\mathbf{x}) \left(\mu_C^{H,N}(\mathbf{x})\kappa_1 + d^{H,M}(\mathbf{x})\kappa_2 \right) \right] \\ \geq \mathbf{P} \left[(1 - \epsilon) 2\mu(\mathbf{x}) (\mu_C(\mathbf{x})\kappa_1 + d(\mathbf{x})\kappa_2) \leq 2\mu^{H,N}(\mathbf{x}) \left(\mu_C^{H,N}(\mathbf{x})\kappa_1 + d^{H,M}(\mathbf{x})\kappa_2 \right) \right] \\ \geq \mathbf{P} \left[\left\{ 2(1 - \epsilon)^\alpha \mu(\mathbf{x}) \leq 2\mu^{H,N}(\mathbf{x}) \right\} \cap \left\{ (1 - \epsilon)^{(1-\alpha)} (\mu_C(\mathbf{x})\kappa_1 + d(\mathbf{x})\kappa_2) \leq \mu_C^{H,N}(\mathbf{x})\kappa_1 + d^{H,M}(\mathbf{x})\kappa_2 \right\} \right] \\ \geq \underbrace{\mathbf{P} \left[(1 - \epsilon)^\alpha \mu(\mathbf{x}) \leq \mu^{H,N}(\mathbf{x}) \right]}_{(1)} + \underbrace{\mathbf{P} \left[(1 - \epsilon)^{(1-\alpha)} \mu_C(\mathbf{x}) \leq \mu_C^{H,N}(\mathbf{x}) \right]}_{(2)} + \underbrace{\mathbf{P} \left[(1 - \epsilon)^{(1-\alpha)} d(\mathbf{x}) \leq d^{H,M}(\mathbf{x}) \right]}_{(3)} - 2, \end{aligned}$$

for any $\alpha \in (0, 1)$. Higher values of α require tighter estimates from (2) and (3) while lower values require tighter estimates from (1). We defer choosing a value for α until later in the proof.

Since $(1 - \epsilon)^\alpha \leq 1 - (\alpha\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2)$ for our ranges of ϵ and α , it follows that

$$(1) \geq \mathbf{P} \left[1 - \left(\alpha\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2 \right) \leq \mu^{H,N}(\mathbf{x}) \right].$$

Thus, by Theorem 3.23, if $N \geq 6 \left(\alpha\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2 \right)^{-2} \log(4n_\omega/\delta)$, then $(1) \geq 1 - \frac{\delta}{2}$.

Similarly, we may bound $(1 - \epsilon)^{(1-\alpha)} \leq 1 - ((1 - \alpha)\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2)$ and thus

$$(2) \geq \mathbf{P} \left[\left(1 - \left((1 - \alpha)\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2 \right) \right) \mu_C(\mathbf{x}) \leq \mu_C^{H,N}(\mathbf{x}) \right]$$

and

$$(3) \geq \mathbf{P} \left[\left(1 - \left((1 - \alpha)\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2 \right) \right) d(\mathbf{x}) \leq d^{H,M}(\mathbf{x}) \right].$$

Let $\tilde{\epsilon} = (1 - \alpha)\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2$. Then, if $N \geq 6(2\tilde{\epsilon} - \tilde{\epsilon}^2)^{-2} \log(4n_\omega/\delta)$, $(2) \geq 1 - \frac{\delta}{4}$. By the same argument, if $M \geq 6(2\tilde{\epsilon} - \tilde{\epsilon}^2)^{-2} \log(4n_\omega/\delta)$, $(3) \geq 1 - \frac{\delta}{4}$. Therefore, if $N \geq 6 \max \left(\left(\alpha\epsilon + \frac{1}{2}\alpha(1 - \alpha)\epsilon^2 \right)^{-2}, (2\tilde{\epsilon} - \tilde{\epsilon}^2)^{-2} \right) \log(4n_\omega/\delta)$ and $M \geq 6(2\tilde{\epsilon} - \tilde{\epsilon}^2)^{-2} \log(4n_\omega/\delta)$,

$$\begin{aligned} \mathbf{P} \left[(1 - \epsilon) \|\nabla m_\ell(x_k) - \nabla f(x_k)\|_2 \leq 2\mu^{H,N}(\mathbf{x}) \left(\mu_C^{H,N}(\mathbf{x})\kappa_1 + d^{H,M}(\mathbf{x})\kappa_2 \right) \right] \\ \geq \underbrace{\mathbf{P} \left[(1 - \epsilon)^\alpha \mu(\mathbf{x}) \leq \mu^{H,N}(\mathbf{x}) \right]}_{(1)} + \underbrace{\mathbf{P} \left[(1 - \epsilon)^{(1-\alpha)} \mu_C(\mathbf{x}) \leq \mu_C^{H,N}(\mathbf{x}) \right]}_{(2)} + \underbrace{\mathbf{P} \left[(1 - \epsilon)^{(1-\alpha)} d(\mathbf{x}) \leq d^{H,M}(\mathbf{x}) \right]}_{(3)} - 2, \\ \geq \left(1 - \frac{\delta}{2} \right) + \left(1 - \frac{\delta}{4} \right) + \left(1 - \frac{\delta}{4} \right) - 2 \\ = 1 - \delta. \end{aligned}$$

While this holds for any value of $\alpha \in (0, 1)$, we seek to minimize (or at least approximately minimize) the number of samples N . This amounts to solving the quartic (in α) equation

$$\frac{1}{2}(1 - \alpha)\alpha\epsilon^2 + \alpha\epsilon = \frac{1}{4}(1 - \alpha)\epsilon(\alpha\epsilon + 2)(4 - (1 - \alpha)\epsilon(\alpha\epsilon + 2)).$$

While the positive, real root does have a closed form solution, the optimal value of α decreases monotonically with ϵ from a supremum of $2/3$ as $\epsilon \rightarrow 0$ to a minimum of ≈ 0.625 at $\epsilon = 1$. Choosing $\alpha = 2/3$ is very close to optimal and ensures that the $2\tilde{\epsilon} - \tilde{\epsilon}^2$ term is larger. The maximum difference occurs at $\epsilon = 1$ where it suggests that 3 samples is sufficient instead of 2. Thus, for brevity, we simply use the value $\alpha = 2/3$ which yields $\tilde{\epsilon} = \hat{\epsilon}$ as defined in the assumptions. $\square$

If, on the other hand, a bound for $\mu(\mathbf{x})$ and $\mu_C(\mathbf{x})$ is known, only a single estimate is involved and the proof is much more simple.

Theorem 3.29. *Let $\mu(\mathbf{x}) \leq \tilde{\mu}$ and $\mu_C(\mathbf{x}) \leq \tilde{\mu}_C$. As in Theorem 3.28, define*

$$d^{H,M}(\mathbf{x}) = \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \text{tr}_H^M \left(\left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right\|_F^2 \right)}.$$

If $M \geq 6(2\epsilon - \epsilon^2)^{-2} \log(n_p n_\omega / \delta)$, then

$$2\tilde{\mu} \left(\tilde{\mu}_C \sqrt{\sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \right\|_F^2} \cdot \sqrt{\sum_{i=1}^{n_\omega} \|(\mathbf{I} - \mathbf{Q}_i)\mathbf{B}\|_F^2} + d^{H,M}(\mathbf{x}) \sqrt{m_\ell(x_k)} \right)$$

is a δ -probabalistically ϵ -accurate underestimate of $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|_2$.

Proof. Define κ_1 and κ_2 as in Theorem 3.28. Since the result trivially holds for $\kappa_2 = 0$, assume $\kappa_2 > 0$. We begin by noting

$$\begin{aligned} & \mathbf{P} \left[(1 - \epsilon) \|\nabla m_\ell(x_k) - \nabla f(x_k)\|_2 \leq 2\tilde{\mu} (\tilde{\mu}_C \kappa_1 + d^{H,M}(\mathbf{x}) \kappa_2) \right] \\ & \geq \mathbf{P} \left[(1 - \epsilon) 2\tilde{\mu} (\tilde{\mu}_C \kappa_1 + d(\mathbf{x}) \kappa_2) \leq 2\tilde{\mu} (\tilde{\mu}_C \kappa_1 + d^{H,M}(\mathbf{x}) \kappa_2) \right] \\ & = \mathbf{P} \left[-\epsilon \tilde{\mu}_C \frac{\kappa_1}{\kappa_2} + (1 - \epsilon) d(\mathbf{x}) \leq d^{H,M}(\mathbf{x}) \right] \\ & \geq \mathbf{P} \left[(1 - \epsilon) d(\mathbf{x}) \leq d^{H,M}(\mathbf{x}) \right] \end{aligned}$$

and the result follows from a direct application of Lemma 3.27. $\square$

Example 3.30. Let us consider a specific example. For a two-dimensional tomography problem, using 25 radial basis functions, each with four parameters, yields a 100-variable parameterization of the problem, $n_p = 100$. If we use only the zero frequency, $n_\omega = 1$.

Choosing $\alpha_U = 4$, $\alpha_L = 11/15$, $\kappa_{\text{fcd}} = 1/2$, $\eta_1 = 1/100$, $\eta_3 = 1$, $\gamma = 2$, and leaving β free yields the following inequalities:

$$\frac{(1 - \delta_f)(1 - \delta_\nabla) - \frac{1}{2}}{\delta_f \delta_\nabla} \geq \frac{6}{1 - \frac{1}{5 + 4(1 - \beta)\kappa_{\text{bh}}}}, \quad (3.38)$$

$$\delta_f \delta_\nabla \leq \frac{1}{2 \left(\frac{23}{2} + \max \left(\frac{4}{\kappa_{\text{bh}}}, \frac{400(1 - \beta)}{2 - \beta + (1 - \beta)\kappa_{\text{bh}}} \right) \left(\frac{20}{3(1 - \beta)} + 6\kappa_{\text{bh}} \right) \right)}, \quad (3.39)$$

$$\eta_2 \geq 2 \left(1 + \frac{1}{1 - \beta} + \kappa_{\text{bh}} \right). \quad (3.40)$$

The right hand side of Equation (3.38) can be bounded above by $\frac{3(9 - 4\beta)}{2(2 - \beta)}$. If $0 \leq \beta \leq 8/9$, the right hand side of Equation (3.39) is bounded below by $\frac{3}{14446}$. At first glance, this lower bound may look unappealing since $\sqrt{\frac{3}{14446}} \approx 0.014$. However, if we are equipped with bounds that allow use of Theorem 3.29, then we can make δ_∇ very small without large computational cost. Indeed, since the scaling of M with respect to δ_∇ is $\log(1/\delta_\nabla)$, we see that even very small δ_∇ can be achieved with a modest number of samples M . In particular, if we choose $\beta = 8/9$, then $N \geq 7$ and $M \geq 75$ samples are sufficient to guarantee $\delta_f \leq 0.39$ and $\delta_\nabla \leq \frac{1}{2000}$ which satisfy the bounds in Equations (3.38) and (3.39) so long as κ_{bh} is finite.

Unfortunately, the one place we cannot make independent of κ_{bh} is in η_2 , which limits our trust region radius relative to the norm of the model gradient. This, in principle, would limit our maximum trust region size fairly severely. As mentioned in a remark in [19], it would be possible to introduce another quantity (proportional to η_2) that controls the step size, but this option is not explored in this paper. In practice, the algorithm is successful even without such large values of η_2 , but the our proof of convergence relies on a more limited trust region radius.

3.6.4 Another Approach to $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|$

While the estimates of the previous section enjoy rigorous bounds on the number of samples required, they can be quite pessimistic. This leads to poor performance as models are rebuilt more frequently. An alternative is inspired by the Random Simultaneous Sources and Detectors approach outlined in [36]. With this approach, we apply random vectors from both the left and the right to reduce the cost of estimating the Jacobian terms in Equation (3.37). First, we will show that this is an unbiased estimate.

Lemma 3.31. *Let $\mathbf{L} = (\mathbf{l}_1, \dots, \mathbf{l}_{n_L})$ and $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_{n_R})$ where each column vector $\mathbf{l}_i \in \mathbb{R}^{n_{\text{det}}}$ and $\mathbf{r}_j \in \mathbb{R}^{n_{\text{src}}}$ have entries that are independently drawn from the Rademacher distribution. Let $\mathbf{J} \in \mathbb{R}^{n_{\text{det}} \times n_{\text{src}}}$. Then,*

$$\mathbb{E} \left[\frac{1}{n_L \cdot n_R} \|\mathbf{L}^T \mathbf{J} \mathbf{R}\|_F^2 \right] = \|\mathbf{J}\|_F^2.$$

Proof. Since the entries of $\mathbf{L}$ and $\mathbf{R}$ are independent Rademacher-distributed variables, it follows that $\mathbb{E} [\mathbf{L} \mathbf{L}^T] = n_L \mathbf{I}_{n_L}$ and $\mathbb{E} [\mathbf{R} \mathbf{R}^T] = n_R \mathbf{I}_{n_R}$. Therefore, we can see that

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n_L \cdot n_R} \|\mathbf{L}^T \mathbf{J} \mathbf{R}\|_F^2 \right] &= \frac{1}{n_L \cdot n_R} \mathbb{E} [\text{tr} ((\mathbf{R}^T \mathbf{J}^T \mathbf{L}) (\mathbf{L}^T \mathbf{J} \mathbf{R}))] \\ &= \frac{1}{n_L \cdot n_R} \mathbb{E} [\text{tr} (\mathbf{R} \mathbf{R}^T \mathbf{J}^T \mathbf{L} \mathbf{L}^T \mathbf{J})] \\ &= \frac{1}{n_L \cdot n_R} \text{tr} (\mathbb{E} [\mathbf{R} \mathbf{R}^T] \mathbf{J}^T \mathbb{E} [\mathbf{L} \mathbf{L}^T] \mathbf{J}) \\ &= \text{tr} (\mathbf{J}^T \mathbf{J}) \\ &= \|\mathbf{J}\|_F^2. \end{aligned}$$

□

We now turn again to the Jacobian terms in Equation (3.37). Since

$$\|\mathbf{J}\|_F^2 = \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T \mathbf{K}^{-1}(\omega_k; \mathbf{x}) \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{K}^{-1}(\omega_k; \mathbf{x}) \mathbf{B} \right\|_F^2$$

and

$$\|\mathbf{J} - \mathbf{J}_R\|_F^2 = \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{C}^T \mathbf{K}^{-1}(\omega_k; \mathbf{x}) \left(\frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right) \mathbf{K}^{-1}(\omega_k; \mathbf{x}) \mathbf{B} \right\|_F^2,$$

we can estimate both of these terms by computing $\mathbf{L}_{k,\mathbf{x}}^S = (\mathbf{C} \mathbf{L})^T \mathbf{K}^{-1}(\omega_k; \mathbf{x})$ and $\mathbf{R}_{k,\mathbf{x}}^S = \mathbf{K}^{-1}(\omega_k; \mathbf{x}) \mathbf{B} \mathbf{R}$ and thus estimating

$$\|\mathbf{J}\|_F^2 \approx \frac{1}{n_L \cdot n_R} \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{L}_{k,\mathbf{x}}^S \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{R}_{k,\mathbf{x}}^S \right\|_F^2$$

and

$$\|\mathbf{J} - \mathbf{J}_R\|_F^2 \approx \frac{1}{n_L \cdot n_R} \sum_{\ell=1}^{n_p} \sum_{k=1}^{n_\omega} \left\| \mathbf{L}_{k,\mathbf{x}}^S \left(\frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} - \mathbf{Q}_k \frac{\partial \mathbf{K}(\omega_k; \mathbf{x})}{\partial x_\ell} \mathbf{P}_k \right) \mathbf{R}_{k,\mathbf{x}}^S \right\|_F^2.$$

In addition to the $n_L \cdot n_R$ samples of the Jacobian terms, we may also reuse $\mathbf{L}_{k,\mathbf{x}}^S$ and $\mathbf{R}_{k,\mathbf{x}}^S$ to compute $n_L + n_R$ samples of the improvement factor $f(x_k) - f(x_k + s_k)$, as discussed in Section 3.6.3, while only requiring additional large linear solves at $\mathbf{x} = x_k + s_k$. Since this method estimates the Jacobian terms directly, there is no need to estimate or bound the μ and $\mu_{\mathbf{C}}$ terms.

3.7 Numerical Results

To illustrate the effectiveness of the algorithm, we apply Algorithm 2 to the DOT problem using the two-sided estimates from Section 3.6.4 with the following parameters:

- $\eta_1 = 0.01$,
- $\eta_2 = 0$,
- $\eta_3 = 1$,
- $\gamma = 2$,
- $n_L = n_R = 5$,
- $\alpha_U = 4$.

We note that a much larger (at the very least least positive) value for η_2 is an element of the convergence analysis presented in Section 3.4. We aim to show here that this is often unnecessary in practice and that even relatively few samples still result in a robust method.

The numerical experiments are constructed as follows. The observations are generated by constructing a target anomaly in the 0-1 pixel basis. Measurements are then generated by evaluating the forward problem and adding white noise at a level of 5 permille. Based on the discrepancy principle, the optimization is terminated when the residual norm is less than 1.1 times the norm of the added noise. For these experiments, we use a symmetric, one-sided projection, i.e. $\mathbf{V} = \mathbf{W}$, and a single frequency, corresponding to $\omega = 0$. The initial condition is evaluated with the full-order model and the reduced-order model is constructed from that evaluation (note: since the same initial condition is used for each instance, this could be computed ahead of time and reused). The step algorithm from TREGS [23] is used to compute candidate steps.

These parameter values correspond to $5 \cdot 5 = 25$ samples per Jacobian estimate and $5 + 5 = 10$ samples per improvement estimate. Each step requires no more than $2 \cdot (5 + 5) = 20$ solutions to large linear systems (though the model improvement step may use more).

Since the focus of this paper is on the construction of a robust method, we focus our discussion on evaluating the quality of our improvement factor estimates. The quality of the gradient error estimate may impact the efficiency/performance of the method, but the improvement factor estimate is key to rejecting steps that would increase the objective function.

3.7.1 Two-Dimensional DOT Problem

The two-dimensional DOT problem is solved on with a grid size of 201×201 , yielding $n = 40,401$ state variables after discretization using the standard centered finite difference

formula. There are $n_{\text{src}} = n_{\text{det}} = 32$ sources and detectors. We use 25 compactly-supported radial basis functions (CSRBF) for our PaLS reconstruction, yielding 100 (four per CSRBF) parameters to be estimated. A total of 350 trials were run with the same anomaly but a different instance of noise added to the observation data.

To illustrate the robust nature of the algorithm, Figure 3.2 is the convergence history of the trial with the largest number of incorrectly accepted steps (2). At indices 14 and 22, the objective function increases due to the incorrectly accepted step. However, this is detected quickly, and the algorithm corrects these increases and converges to within the given numerical tolerance.

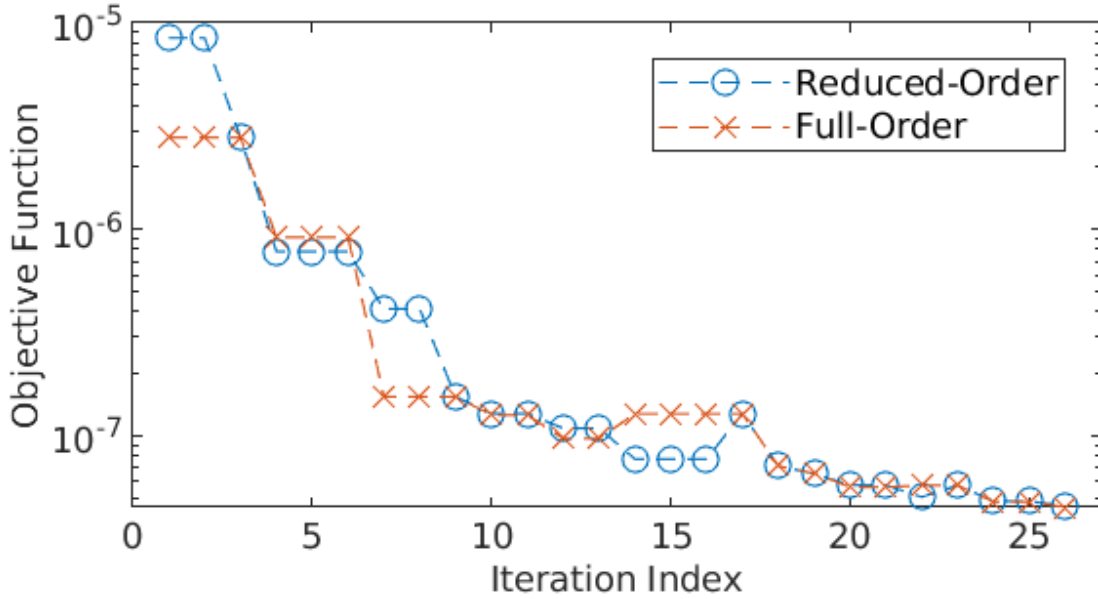


Figure 3.2: Convergence History of the Currently Accepted Step for a Two-Dimensional DOT Problem with Two Incorrectly Accepted Steps

To examine the quality of our improvement factor estimate with only 10 samples, we present a PDF-normalized histogram of the ratio between the estimated improvement factor, labeled ρ_{est} , and the actual improvement factor, labeled ρ_{act} , in Figure 3.3. A value of 1 would indicate an exact match between the estimate and the true value. Values less than zero indicate instances where the incorrect conclusion was drawn (i.e. the step was accepted when it should have been rejected, or vice versa). Of the total of 5698 estimates performed during the 350 trials, approximately 60 are not included in Figure 3.3 as they fall outside of the presented range.

As Figure 3.3 shows, the vast majority of samples are relatively close to the actual improvement value, though most are over-estimates. The 5th, 25th, 50th, 75th, and 95th percentiles are reported in Table 3.2. Since we used a value of $\alpha_U = 4$, we see that at least 90 percent of all samples were within a factor of $(0, 1 + \alpha_U)$. From Table 3.3, we see that the correct

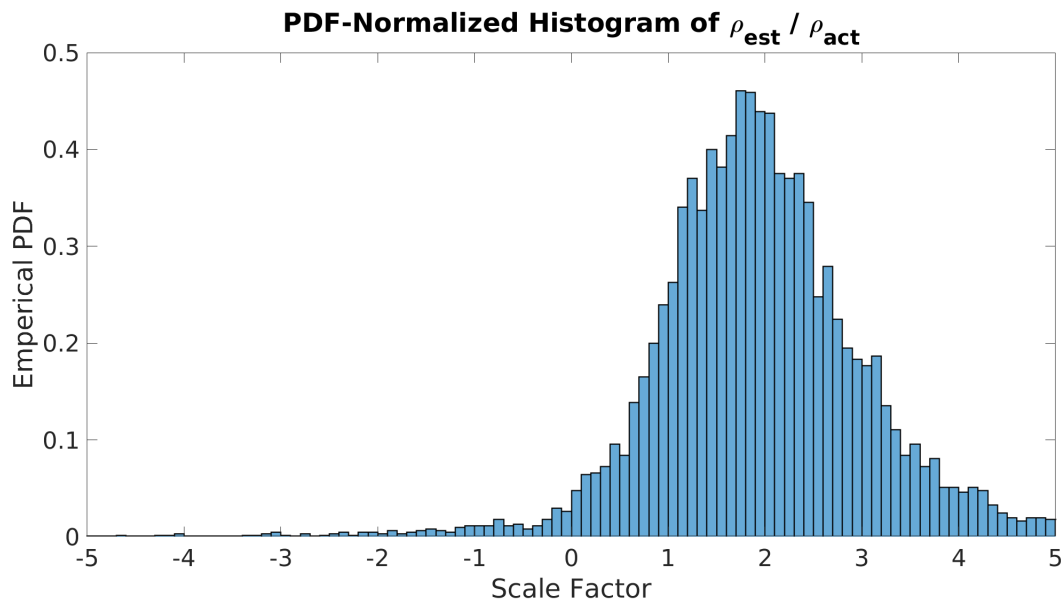


Figure 3.3: PDF-Normalized Histogram of the Ratio between Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem

Table 3.2: Percentile Values of the Ratio between Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem

Percentile	Scale Factor
5th	0.2802
25th	1.2963
50th	1.9072
75th	2.5805
95th	4.0967

decision is made in nearly 97% of cases and that the case where the step is erroneously accepted occurs in approximately 1.79% of cases. Therefore, for the two-dimensional tomography problem, the Trust Region with Stochastic Bounds equipped with the estimates of Sections 3.6.2 and 3.6.4 is a robust algorithm. Furthermore, while the total number of solutions to large linear systems is not reported, this method represents a decrease from 32 such solutions per step to a maximum of 20 per step, thus resulting in a minor reduction in the number of large linear systems to solve.

Table 3.3: Outcome Counts for Accepting or Rejecting the Candidate Step for the Estimated and Actual Improvement Factors for the Two-Dimensional Tomography Problem

Outcome	Count
Correctly Accepted	4059 (71.24%)
Incorrectly Accepted	102 (01.79%)
Incorrectly Rejected	90 (01.58%)
Correctly Rejected	1447 (25.39%)

3.7.2 Three-Dimensional DOT Problem

The three-dimensional DOT problem is solved on with a grid size of $64 \times 64 \times 64$, yielding $n = 262,144$ state variables after discretization using the standard centered finite difference formula. There are $n_{\text{src}} = n_{\text{det}} = 961$ sources and detectors. We use 27 compactly-supported radial basis functions (CSRBF) for our PaLS reconstruction, yielding 135 (five per CSRBF) parameters to be estimated. A total of 30 trials were run with the same anomaly but a different instance of noise added to the observation data.

As in Section 3.7.1, we use a total of 10 samples per estimate of the improvement factor. In this case, a decrease from 961 solutions per step to a maximum of 20 per step represents a very large decrease in the number of solutions required. While the number of trials is much more limited, we see in Table 3.4 that the method remains robust against erroneously accepting a step even with nearly a factor 50 reduction in the number of large linear systems solved per step.

Table 3.4: Outcome Counts for Accepting or Rejecting the Candidate Step for the Estimated and Actual Improvement Factors for the Three-Dimensional Tomography Problem

Outcome	Count
Correctly Accepted	103 (75.74%)
Incorrectly Accepted	1 (00.74%)
Incorrectly Rejected	17 (12.50%)
Correctly Rejected	15 (11.03%)

3.8 Conclusions

We have shown that replacing the improvement factor by stochastic estimates of the improvement factor and the gradient error results in a robust optimization method that is completely free of any full-order function or gradient evaluations. We derived rigorous bounds on the

number of samples needed to fulfill the assumptions of our convergence analysis for the Diffuse Optical Tomography problem as well as some more higher performance, but less rigorous estimates. By using as few as 10 samples per improvement estimate, we achieved a robust method independent of the system size as demonstrated by the success of both the two-dimensional and three-dimensional DOT problems.

While the algorithm presented here is presented in a very general setting, the specific stochastic bounds are very specific to the Linear Time-Invariant structure of the DOT problem. Deriving efficient estimates for other classes of problems remains an open problem. Since the estimate of the improvement factor implies the ability to estimate $f(x_k) - f(x_k + s_k)$ for an arbitrary s_k (within the trust region), one possibility is to estimate the gradient error by using the stochastic estimates to approximating a finite difference, though this scales poorly with the accuracy required and the number of parameters.

Chapter 4

Conclusions

We have proposed two efficient and robust techniques for using stochastic estimates to reduce the cost of parameter inversion for large-scale forward models.

In the first manuscript, we showed that as few as a single sample of the Hutchinson trace estimator was sufficient to reliably detect large interpolation errors when using a projection-based ROM. We also introduced a highly efficient “Residual-Driven” update scheme that requires far fewer solutions to large linear systems than techniques inspired by the interpolation conditions in [12]. By combining a reusable and very small initial ROM with our stochastic estimates and the Residual-Driven update scheme, we demonstrate a technique that reliably reconstructs the anomaly while only requiring 0.3% as many solutions to large linear systems in the median case for the three-dimensional tomography problem.

In the second manuscript, we showed that stochastic estimates of the improvement factor, $f(x_k) - f(x_k + s_k)$, and the model gradient mismatch, $\|\nabla m_\ell(x_k) - \nabla f(x_k)\|$, can be used to construct a general trust region framework that converges with probability one given mild assumptions on the probabilistic accuracy of the estimates. Suitable bounds for the DOT problem were derived based on using projection-based ROMs as the trust region model and numerical experiments suggest that just 10 samples per iteration can be sufficient for an extremely robust method, erroneously accepting a step in less than 2% of steps for the two-dimensional problem and less than 1% of steps in the three-dimensional problem.

While the examples used in this thesis apply specifically to the DOT problem, they extend naturally to other Linear Time-Invariant systems. The trust region framework developed in Section 3.2 is sufficiently general that it may be applied to other classes of problems, so long as they admit suitable stochastic bounds. The application of this framework to other problems remains open work.

Bibliography

- [1] K. Afanasiev and M. Hinze, *Adaptive Control Of A Wake Flow Using Proper Orthogonal Decomposition*, in *Shape Optimization And Optimal Design*, vol. 216, 2010, pp. 317–332.
- [2] A. Aghasi, M. Kilmer, and E. L. Miller, *Parametric Level Set Methods for Inverse Problems*, *SIAM Journal on Imaging Sciences*, 4 (2011), pp. 618–650.
- [3] K. Ahuja, P. Benner, E. de Sturler, and L. Feng, *Recycling BiCGSTAB with an Application to Parametric Model Order Reduction*, *SIAM Journal on Scientific Computing*, 37 (2015), pp. S429–S446.
- [4] K. Ahuja, E. de Sturler, S. Gugercin, and E. R. Chang, *Recycling BiCG with an Application to Model Reduction*, *SIAM Journal on Scientific Computing*, 34 (2012), pp. A1925–A1949.
- [5] A. Alla, M. Hinze, O. Lass, and S. Ulbrich, *A Certified Model Reduction Approach for robust optimal control with PDE constraints*, (2017).
- [6] D. Amsallem and C. Farhat, *Interpolation Method for Adapting Reduced-Order Models and Application to Aeroelasticity*, *AIAA Journal*, 46 (2008), pp. 1803–1813.
- [7] A. Antoulas, C. Beattie, and S. Gugercin, *Interpolatory Methods for Model Reduction*, SIAM, 2020.
- [8] S. R. Arridge, *Optical tomography in medical imaging*, *Inverse Problems*, 15 (1999), pp. R41–R93.
- [9] S. S. Aslan, E. de Sturler, and M. E. Kilmer, *Randomized Approach to Nonlinear Inversion Combining Random and Optimized Simultaneous Sources and Detectors*, *SIAM Journal on Scientific Computing*, 41 (2019), pp. B229–B249.
- [10] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, *Journal of the ACM*, 58 (2011), pp. 1–34.
- [11] A. S. Bandeira, K. Scheinberg, and L. N. Vicente, *Convergence of Trust-Region Methods Based on Probabilistic Models*, *SIAM Journal on Optimization*, 24 (2014), pp. 1238–1264.
- [12] U. Baur, C. Beattie, P. Benner, and S. Gugercin, *Interpolatory Projection Methods for Parameterized Model Reduction*, *SIAM Journal on Scientific Computing*, 33 (2011), pp. 2489–2518.

- [13] C. Beattie, S. Gugercin, and S. Wyatt, *Inexact solves in interpolatory model reduction*, Linear Algebra and its Applications, 436 (2012), pp. 2916–2943.
- [14] P. Benner and S. Grundel, *Model Order Reduction for a family of systems with application in parametric and uncertain systems*, mpi-magdeburg.mpg.de, (2014).
- [15] P. Benner, S. Gugercin, and K. Willcox, *A survey of projection-based model reduction methods for parametric dynamical systems*, SIAM Review, 57 (2015), pp. 1–49.
- [16] L. Borcea, V. Druskin, A. V. Mamonov, and M. Zaslavsky, *A model reduction approach to numerical inversion for a parabolic partial differential equation*, Inverse Problems, 30 (2014), p. 125011.
- [17] J. Borggaard, K. R. Pond, and L. Zietsman, *Parametric reduced order models using adaptive sampling and interpolation*, vol. 19, IFAC, 2014.
- [18] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Model Reduction for Large-Scale Systems with High-Dimensional Parametric Input Space*, SIAM Journal on Scientific Computing, 30 (2008), pp. 3270–3288.
- [19] R. Chen, M. Menickelly, and K. Scheinberg, *Stochastic optimization using a trust-region method and random models*, Mathematical Programming, 169 (2018), pp. 447–487.
- [20] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-region methods*, Book, MPS/SIAM S (2000), pp. xx+959.
- [21] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points*, SIAM Journal on Optimization, 20 (2009), pp. 387–415.
- [22] E. de Sturler, S. Gugercin, M. E. Kilmer, S. Chaturantabut, C. Beattie, and M. O’Connell, *Nonlinear Parametric Inversion Using Interpolatory Model Reduction*, SIAM Journal on Scientific Computing, 37 (2015), pp. B495–B517.
- [23] E. de Sturler and M. E. Kilmer, *A Regularized Gauss–Newton Trust Region Approach to Imaging in Diffuse Optical Tomography*, SIAM Journal on Scientific Computing, 33 (2011), pp. 3057–3086.
- [24] G. Deng and M. C. Ferris, *Variable-Number Sample-Path Optimization*, Mathematical Programming, 117 (2009), pp. 81–109.
- [25] R. Durrett, *Probability: Theory and Examples*, Cambridge University Press, 5 ed., 4 2019.
- [26] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

- [27] E. Haber, M. Chung, and F. Herrmann, *An Effective Method for Parameter Estimation with PDE Constraints with Multiple Right-Hand Sides*, SIAM Journal on Optimization, 22 (2012), pp. 739–757.
- [28] M. Hutchinson, *A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines*, Communications in Statistics - Simulation and Computation, 19 (1990), pp. 433–450.
- [29] B. Kragel, *Streamline Diffusion POD Models in Optimization*, PhD thesis, University of Trier, 2005.
- [30] D. Munster and E. De Sturler, *Robust Parameter Inversion Using Adaptive Reduced Order Models*, In Preparation, (2019).
- [31] M. O’Connell, M. E. Kilmer, E. de Sturler, and S. Gugercin, *Computing Reduced Order Models via Inner-Outer Krylov Recycling in Diffuse Optical Tomography*, SIAM Journal on Scientific Computing, 39 (2017), pp. B272–B297.
- [32] B. Peherstorfer and K. Willcox, *Online Adaptive Model Reduction for Nonlinear Systems via Low-Rank Updates*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2123–A2150.
- [33] E. Qian, M. Grepl, K. Veroy, and K. Willcox, *A Certified Trust Region Reduced Basis Approach to PDE-Constrained Optimization*, SIAM Journal on Scientific Computing, 39 (2017), pp. S434–S460.
- [34] F. Roosta-Khorasani and U. Ascher, *Improved Bounds on Sample Size for Implicit Matrix Trace Estimators*, Foundations of Computational Mathematics, 15 (2015), pp. 1187–1212.
- [35] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, Inc., New York, 3rd ed., 1987.
- [36] S. Sariaydin, *Randomization for Efficient Nonlinear Parametric Inversion*, PhD thesis, Virginia Polytechnic Institute and State University, 2018.
- [37] Y. Yue and K. Meerbergen, *Accelerating optimization of parametric linear systems by model order reduction*, SIAM Journal on Optimization, 23 (2013), pp. 1344–1370.