

Uncovering Hidden Structure in Data and the Physical World for Seismic Tomography and Beyond

Thesis by
Angela Fang Gao

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended August 15, 2025

© 2026

Angela Fang Gao
ORCID: 0000-0001-8574-8728

All rights reserved

ACKNOWLEDGEMENTS

To my advisor, Dr. Katie Bouman, thank you for giving me the opportunity to do my Ph.D. at Caltech and be one of the first students in the lab. I am so grateful for the freedom to explore whatever I found interesting. I have grown so much under your guidance and mentorship, and I am looking forward to all of the amazing things that the Computational Cameras Lab continues to do.

To my committee members: Yisong Yue, thank you for always being a sounding board for projects and directions. Zachary Ross, thank you for providing me with so many opportunities that shaped the trajectory of my PhD. Joel Tropp, thank you for being a great teacher during the beginning of my PhD and also at the end.

I would like to thank all of my past educators that have helped get me here. This would not have been possible if I had not gotten into Caltech in the first place. Thanks to Adam Finklestein, Szymon Rusinkiewicz, and Aswin Sankaranarayanan for your patience and guidance when I was an undergrad. I want to thank Byron Yu and Pulkit Grover for teaching 18-290, Introduction to Signals and Systems. I know how hard it is to teach that class, but it was taught with the perfect amount of rigor that excited me about the field. To Mrs. Reichmann, my BC calculus teacher, and Mr. Heckman, my 4th and 5th grade math teacher, for nurturing my love of mathematical problem solving. To Ms. Kamm for being the wackiest physics teacher who made everything fun.

Thanks to wonderful undergrads that I had the pleasure of working with, especially Yvette Lin and Ayush Varshney.

Also, thank you to the Caltech CMS staff, especially Sydney Garstang and Monica Nolasco.

To my Caltech community, thanks for making Caltech feel like home. To the Computational Cameras lab (Berthy Feng, He Sun, Ray Wu, Aviad Levis, Brandon Zhao, Oscar Leong, Heri Rajaoberison, Bingliang Zhang, Fred Wang), thanks for making research so fun. I have learned so much from all of you. Special thanks to the Perona lab (Sara Beery, Eli Cole, Joe Marino, Tony Zhang, Neehar Kondapaneni, Mason McGill, Serim Ryou, Matteo Ronchi, Jennifer Sun) and early Yue Crew members (Jeremy Bernstein, Uriah Israel) for making me feel welcome when the Bouman lab was nascent. Lab beers was the highlight of my week as we were getting out of COVID. And to the growing AIML labs, Suzanne Stathatos, Laure

Delisle, Rogerio Guimaraes, Kevin Le, Raphi Kang, Michael Hobley, Madison Van Horn, Markus Marks, Geeling Chau, Yue Song, Chris Yeh, Sabera Talukder, Jiawei Zhao, Zongyi Li, for all the eye-opening discussions.

To my SeismoLab collaborators and friends, Jack Wilding, Caifeng Zou, Valeria Villa, Rajani Shrestha, Ettore Biondi, Yan Yang, Thea Ragon, Jorge Castellanos, thanks for being my lab away from lab. I appreciate all of you putting up with me asking questions about the most obvious things like at what timescales a solid becomes a liquid and whether a desk would ever be a liquid. Despite some unfortunate weather at the freezing CNLS conference in Santa Fe to the heat wave at SCEC 2024, we always found ways to have fun.

To 2478, thanks for giving me something to believe in. To the Caltech Fiber Arts club, thanks for being a warm cozy place to look forward to during my last year of grad school. To my morph buddies, Ana Duarte and Emily Fourney, I'm going to miss all of the fun workouts. To the rest of my Caltech friends that I haven't mentioned yet, Ryan Cosner, Andreas Butler, Apurva Badithela, Alex Ayling, Francesca Li, Tim Buelles, Anna Ding, Roberto Trevino, thanks for all the lovely dinners, game nights, and holidays together.

To the grad cafe addicts (Ani Shridhar, Ethan Tseng, Mary Jin, Rebekah Zhao, and Yongyi Zhao), thanks for being like older siblings to me, mentoring me through college, helping me get into grad school, and commiserating throughout. We have come so far. To Quad++ (Ivy Li, Jessica Jeon, Nishad Maskara, Akhil Dondapati), thanks for always making me laugh. And to other lifelong friends, Brianna Arscott Grant, Mehar Sawhney, Mallika Khanna, Jennifer Jo, Ricky Gupta, Arushi Patel for all of the experiences that have helped me get out of the Caltech bubble.

It is also necessary for me to thank the 2011 Virginia earthquake. Without that, I am not sure I would have been interested at all in seismology.

To the Singer and Chan families, thanks for being so curious and excited about what I'm doing. I'm so lucky to have you as my family.

To my parents, thank you for always supporting me and believing in me no matter what.

Finally, to my husband, Zach, thank you for being ready to celebrate every success and commiserate with me after every rejection. I could not have done this without your support.

ABSTRACT

Reliable and calibrated imaging strategies are essential for tackling challenging real-world imaging problems. Developing these strategies involves interdisciplinary approaches that translate inverse problem tooling into avenues for real scientific insight. However, raw scientific measurements present many challenges that modern imaging techniques do not adequately address. For instance, many state-of-the-art imaging methods struggle with seismic imaging problems due to several factors including non-linear forward models, uncalibrated noise, sensitivity to initial conditions, and limited angular observations. This thesis aims to overcome these challenges by presenting methods that image under uncertainty, leverage unconventional sources of regularization, and reshape the way we solve imaging inverse problems, ultimately enabling new scientific discoveries.

PUBLISHED CONTENT AND CONTRIBUTIONS

- [1] Angela F. Gao, Ethan N. Epperly, Ayush Varshney, Joel Tropp, and Katherine L. Bouman, “Eikolift: Convexifying eikonal tomography through inpainting time-of-flight,” *In submission to Advances in Neural Information Processing Systems*, 2025,
A.F.G. participated in the conception of the project, developed the algorithm, performed the experiment, and led in writing the manuscript.
- [2] Angela F. Gao, John D. Wilding, Ettore Biondi, Katherine L. Bouman, and Zachary E. Ross, “High-resolution eikonal imaging and uncertainty quantification of the Kilauea Caldera,” *In submission to Journal of Geophysical Research*, 2025,
A.F.G. participated in the conception of the project, developed the algorithm, performed the experiment execution, and led in writing the manuscript.
- [3] Théa Ragon, Angela F. Gao, and Zachary E. Ross, “DeepGEM-EGF: A Bayesian strategy for joint estimates of source-time functions and empirical Green’s functions,” *JGR Solid Earth*, 2025. doi: <https://doi.org/10.1029/2025JB031260>,
A.F.G. participated in the conception of the project, helped with modifying the original algorithm, and participated in writing the manuscript.
- [4] Yvette Y. Lin, Angela F. Gao, and Katherine L. Bouman, “Imaging an evolving black hole by leveraging shared structure,” *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2475–2479, 2024. doi: <https://doi.org/10.1109/ICASSP48485.2024.10445896>,
A.F.G. participated in the conception of the project, developed the algorithm, and assisted in writing the manuscript.
- [5] Angela F. Gao, Oscar Leong, He Sun, and Katherine L. Bouman, “Image reconstruction without explicit priors,” *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: <https://doi.org/10.1109/ICASSP49357.2023.10096515>,
A.F.G. participated in developing the algorithm, executing the IGM experiments, and co-led writing the manuscript.
- [6] Oscar Leong, Angela F. Gao, He Sun, and Katherine L. Bouman, “Discovering structure from corruption for unsupervised image reconstruction,” *IEEE Transactions on Computational Imaging*, vol. 9, pp. 992–1005, 2023. doi: <https://doi.org/10.1109/TCI.2023.3325752>,
A.F.G. participated in developing the algorithm, executing the IGM experiments, and co-led in writing the manuscript.
- [7] Yan Yang, Angela F. Gao, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Rapid seismic waveform modeling and inversion with

neural operators,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. doi: <https://doi.org/10.1109/TGRS.2023.3264210>,

A.F.G. participated in developing the algorithm and in writing the manuscript.

- [8] Angela F. Gao, Jorge Castellanos, Yisong Yue, Zachary E. Ross, and Katherine L. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021,

A.F.G. participated in the conception of the project, developed the algorithm, performed the experiments, and led writing the manuscript.

- [9] Yan Yang, Angela F. Gao, Jorge C. Castellanos, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Seismic wave propagation and inversion with neural operators,” *The Seismic Record*, vol. 1, no. 3, pp. 126–134, 2021. doi: <https://doi.org/10.1785/0320210026>,

A.F.G. participated in developing the algorithm and in writing the manuscript.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Published Content and Contributions	vi
Table of Contents	vii
List of Illustrations	x
List of Tables	xxix
Chapter I: Introduction	1
Chapter II: Background	5
2.1 Inverse problems	5
2.2 Seismic imaging	8
Chapter III: Imaging with uncertain forward models	12
3.1 Introduction	12
3.2 Background and related work	13
3.3 Methods	17
3.4 Case study: blind seismic tomography	18
3.5 Case study: blind deconvolution	25
3.6 Case study: blind deconvolution for estimation of earthquake rupture process using empirical Green's function (EGF)	27
3.7 Conclusions	35
Chapter IV: Imaging with limited priors	39
4.1 Introduction	39
4.2 Background and related work	42
4.3 Approach	45
4.4 Experimental results	54
4.5 Theory for linear IGMs	67
4.6 Conclusion	69
Chapter V: High resolution eikonal imaging of Kilauea	75
5.1 Introduction	75
5.2 Methodological Background	79
5.3 Method	82
5.4 Synthetic imaging	84
5.5 Imaging the Kilauea volcano	89
5.6 Discussion	93
5.7 Conclusions	95
Chapter VI: A convex relaxation for eikonal imaging	103
6.1 Introduction	103
6.2 Background	105
6.3 Approach	107
6.4 Results	111

6.5 Results with real data	116
6.6 Conclusions	117
Chapter VII: Surrogate models for wave imaging	123
7.1 Introduction	123
7.2 Preliminaries	125
7.3 Methods	126
7.4 Experimental results for acoustic wave propagation and inversion . .	128
7.5 Experimental results for elastic wave propagation and inversion . . .	133
7.6 Conclusions	137
Chapter VIII: Conclusion	142
Appendix A: Imaging with limited priors: supplementary materials	144
A.1 Experimental details	144
Appendix B: High resolution eikonal imaging of Kilauea: supplementary materials	146
B.1 Additional results	146
B.2 Theoretical perspective of gradient blurring for linear inverse problems	147
Appendix C: A convex relaxation for eikonal imaging: supplementary materials	154
C.1 Derivation of EikoLift	154
C.2 Block diagonal formulation	155
C.3 Relaxing the data-fit and Lagrangian	156
C.4 Additional experiments	158

LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
2.1 Measurement acquisition and inverse problems. The full measurement acquisition and inversion scheme is described by this diagram where the unobserved source of interest is $\mathbf{x}$ but what is observed is $\mathbf{y}$. This measurement process is governed by some system, often understood using physics. To recover the unobserved source $\mathbf{x}$, we use an inversion algorithm, which often involves priors or regularizers, to recover a reconstruction $\hat{\mathbf{x}}$	5
3.1 DeepGEM applied to blind seismic tomography. (a) A simulated cross section of the Earth's interior (velocity structure), along with the locations of receivers on the surface (red triangles) that collect measurements. (b) The time it takes for a wave traveling from a source below the surface to reach the specified receiver is visualized for each location in the region of interest. The overlaid dots represent the true locations of simulated earthquakes and indicate the measured travel times that constrain optimization. (c) The subsurface velocity reconstruction obtained using a baseline approach optimized with the help of a seismologist. Note that the bright anomaly is missing from this reconstruction. Overlaid dots represent the inferred earthquake locations. (d) DeepGEM reconstructed subsurface velocity and inferred earthquake locations. Note that DeepGEM is able to accurately recover the gradient of the velocity field as well as partially recover the central anomaly.	14

- 3.2 Image priors and source posterior visualization.** (Left) A visualization of the homogeneous, gradient, and layer models used in $p(\theta)$'s $\mathcal{L}_\theta$ term. (Right) Visualizations used to describe source configurations and earthquake source posteriors. (a) Visualizations of the true and initialized source locations are plotted as stars (true) connected to circles, which indicate the expected source locations according to $p(x)$. Note that the expected source locations deviate significantly from the true locations. (b) Visualizations of the learned posterior distribution, $q_\phi(x|y, \theta)$, for each source are plotted as colored histograms and overlaid with stars (of the same color) indicating the true source location. 19
- 3.3 DeepGEM reconstructions significantly outperform baselines, and improve with more sources and better $\mathcal{L}_\theta$ prior models.** Each row corresponds to the reconstructions obtained from a single noisy observation of travel times, where the true Earth velocity is shown in Fig. 3.1(a). Results shown are simulated using 20 surface receivers and a varying number of sources (9, 25, and 49) in a uniform grid. Columns 3-6 show DeepGEM results obtained using different $\mathcal{L}_\theta$ priors. Note that results improve as the $\mathcal{L}_\theta$ prior becomes closer to the true velocity structure, and as the number of sources increases. As a reference, column 2 shows the velocity reconstruction obtained using DeepGEM under fixed, perfectly known source locations. Columns 7-8 show results obtained by the baseline approaches. The velocity reconstruction MSE is included in the top right of each reconstruction. DeepGEM substantially outperforms both straight ray and $\text{MAP}_{\theta,x}$ baselines. Note that DeepGEM also sometimes outperforms the second column, where the source locations are known; this is due to the fact that DeepGEM is sometimes able to absorb measurement error due to its overparameterization. See the supplemental material for more examples. 21

- 3.4 **DeepGEM consistently recovers prominent features across various earthquake source configurations.** Row 2 shows DeepGEM results obtained using different random source configurations, where the true Earth velocity is shown in Fig. 3.1(a). As a reference, row 1 shows the velocity reconstruction obtained using DeepGEM under fixed, perfectly known source locations. As can be seen by the table, both velocity and localization error decrease with an increasing number of sources in the region of interest. Each mean and standard deviation is computed using 5 random realizations of true source configurations. 22
- 3.5 **DeepGEM recovery with a single receiver.** Velocity reconstructions shown in columns 3 and 5 demonstrate that DeepGEM is able to learn some of the true velocity features (see Fig. 3.1(a)), even when limited to measurements from a single receiver. However, these reconstructions show clear signs of overfitting to the measurement data. This is demonstrated by observing that the DeepGEM reconstruction with perfectly known source locations is significantly better with a single receiver when no measurement error is included on the measurements (comparing columns 1 and 2). Note that the learned posterior in column 4 appears to be non-Gaussian, but not multimodal. 23
- 3.6 **Performance of DeepGEM recovery on random velocity fields.** Ten random velocity fields were drawn from a GRF-based distribution and used to simulate travel time measurements with 20 receivers and 100 randomly placed sources. (left) Reconstructions obtained for 3 of these configurations are shown. (right) A table lists the mean and standard deviation of velocity and source error obtained across the ten models, each recovered using different $\mathcal{L}_\theta$ priors. 24
- 3.7 **DeepGEM results are comparable to inversion with known blur kernels for simple Fashion MNIST images.** Blurry measured images (column 1) are generated using the true sharp image and blur kernel (shown in column 2). The deconvolved images obtained using the true blur kernel are shown in column 3. Reconstructed deconvolved images and corresponding inferred blur kernel from DeepGEM are shown in column 4. 26

- 3.8 **Synthetic deconvolution test based on realistic sythetic data setup.** (a-f) Inferred (colored) and target (black) source time functions for synthetic tests designed with waveforms (P arrivals) recorded for four neighbor $M \sim 2$ events that occurred during the Cahuilla swarm. The mean of the source time functions is colored from the misfit between inferred and target EGFs; the standard deviation is shaded. (h) Mean inferred (colored) and target (black) EGF at station TOR. (g) Stations locations. 30
- 3.9 **Improved reconstruction quality as EGF quality increases.** Mean inferred (colored) and target (black) Source-Time Functions (STFs) and Empirical Green's Functions for toy models designed with waveforms (P arrivals only) recorded for four (a,b) or eight (c-f) neighboring $M \sim 2$ events that occurred during the Cahuilla swarm. In (e,f) the additional four waveforms are closer to the target than in (c,d). Standard deviation (1σ) is shaded. EGFs are colored by the root mean squared error (RMSE) between data and average predictions normalized for each component. 31
- 3.10 **Applying DeepGEM-EGF to the Cahuilla swarm.** Apparent STFs inferred for the 2018 Mw 4.41 mainshock of the Cahuilla swarm, assuming the same initial EGFs as in Fig. 3.8. (a) Using DeepGEM-EGF or (b) with the approach of [43]. (d-h) Comparison of the results of the two different approaches at a few stations, colored from their data misfit (cross-correlation between data and predictions). Note that σ estimates for the Landweber approach derive from the variability of multiples EGFs and their components, and not from posterior PDFs as for DeepGEM-EGF results. (i) Data fit at station TOR from the DeepGEM-EGF STF. (j) Map with stations locations. (k) Relocated catalog for the swarm, from [42]. 33

- 4.1 **Method overview.** In this chapter, we tackle ill-posed image reconstruction problems when a traditional image prior is not available or cannot be derived from example images. The key idea of our work is that image reconstruction is possible when one has access to only corrupted measurement examples and the underlying images share common, low-dimensional structure. We explore such ideas in this work in the following two ways: **Left:** Given a single measurement example (with the corresponding forward model) and a set of candidate image generation models (IGMs), we use our proposed criterion to select the best IGM from the measurements alone and then recover an image posterior under that IGM. The criterion we use for model selection is an approximation of the evidence lower bound (i.e., ELBOProxy in Eq. (4.3)) (see Sec. 4.3). **Right:** Building off of this, we consider the setting where we must infer the IGM from noisy measurements alone (see Sec. 4.3). In this case, we solve N inverse problems simultaneously. The inputs of this method are N measurement examples with their known forward models, and the outputs are a single inferred IGM and N latent embeddings that, when combined, lead to N image reconstruction posteriors. 40
- 4.2 We propose to solve a collection of N ill-posed inverse problems by exploiting the common, low-dimensional structure of the underlying images. Given a set of N measurement examples $\{y^{(i)}\}_{i=1}^N$ from N different underlying images, we propose to model each image posterior as the output of a shared IGM with a low-dimensional latent space. In particular, each posterior is approximated by $G_{\theta\#}q_{\phi^{(i)}}$, the push-forward of $q_{\phi^{(i)}}$ through G_{θ} , where G_{θ} is the shared, common generator to all N examples and $q_{\phi^{(i)}}$ is a low-dimensional, variational distribution particular to the i -th example. The parameters, θ and $\{\phi^{(i)}\}_{i=1}^N$, are colored in blue and are jointly inferred. That is, such parameters are inferred specifically for the measurements $\{y^{(i)}\}_{i=1}^N$. The loss we use is the negative ELBOProxy, which is denoted by $\mathcal{L}$ and given by Eq. (4.5). Note that there is no notion of a training set and test set as we aim to solve the inverse problems jointly from all available measurements. 46

- 4.3 We consider two inverse problems: denoising and phase retrieval. Top: the two topmost rows correspond to the ground-truth image x_c and the noisy measurements y_c . Center: in each column, we show the means of the distribution induced by the push-forward of G_j and each latent distribution $z \sim q_{\phi_j}$ for $j \in \{0, \dots, 9\}$. Bottom: each column of the array corresponds to the negative ELBOProxy achieved by each model in reconstructing the images. Here, lower is better. Boxes highlighted in green correspond to the best negative ELBOProxy values in each column. In all these examples, the correct model was chosen. 52
- 4.4 **Improvement with an increasing number of noisy observations.** We demonstrate our method of inferring an IGM to perform denoising for an increasing number of noisy MNIST images (5, 75, and 150 images from left to right). We showcase results on three randomly selected examples that appear in each collection of inverse problems. In each panel, we include the ground-truth, noisy measurements, mean of the posterior, and standard deviation of the posterior. We also include the residual error divided by the empirical standard deviation for $N = 150$. On the far right, we visualize reconstructions using an IGM trained on the full clean MNIST 8's class (6000 images). We observe that the mean reconstructions and standard deviations from our low-data IGMs become more similar to the full-data IGM with increasing data. Our residual errors are largely within 3 standard deviations. 53
- 4.5 **Denoising 95 images of celebrity A.** We demonstrate our method described in 4.3 using 95 noisy images of a celebrity. Here we show the underlying image (row 1), noisy measurements (row 2), mean reconstruction (row 3), and residual error (row 4) for a subset of the 95 different noisy images. Our reconstructions are much less noisy than the measurements and recover sharp features that are hard to discern in the noisy images. We visualize the residual error normalized by the empirical standard deviation, which indicates errors are largely within 3 standard deviations (refer to the colorbar in Fig. 4.4). Note that no explicit spatial-domain prior/regularizer was used in denoising. . . . 53

- 4.6 **Denoising baseline comparisons.** We compare to various baselines (AmbientGAN, Deep Image Prior (DIP), and regularized maximum likelihood using TV (TV-RML) with weight λ), and we report the average PSNR across all 95 reconstructions. We show both early stopping and full training results using DIP. Our method exhibits higher PSNR than all other baselines. We also include results for baselines that require fine-tuning to demonstrate sensitivity to subjective stopping conditions. 55
- 4.7 **Multi-noise denoising.** We demonstrate our method described in Section 4.3 to perform denoising on measurement collections experiencing different noise levels. For each experiment, we use 75 measurement examples, which are defined by $y^{(i)} = x^{(i)} + \eta$ where $\eta \in \{\eta_1, \eta_2, \eta_3\}$ and $\eta_i \sim \mathcal{N}(0, \sigma_i^2 I)$. In Experiment 1, we use noisy measurement examples that have additive noise with standard deviations of 0.01, 0.1, and 0.5. In Experiment 2, we use noisy measurement examples that have additive noise with standard deviations of 0.2, 0.3, and 0.5. We visualize the true underlying images, the measurement used for each experiment, and the mean of the image reconstruction posterior. Most of the reconstructions recover the primary features of the true image. However, in Experiment 1, the reconstructions of the low SNR measurements exhibit bias and do not match the true images well. This is likely because in Experiment 1, the high SNR measurements influence the inferred IGM more strongly than low SNR measurements, leading to biased reconstructions for the reconstructions highlighted in the red box. . . . 56
- 4.8 **Phase retrieval from MNIST 8’s.** We demonstrate our method described in 4.3 to perform phase retrieval on 150 images. For the Fourier phase retrieval setting, we show examples from the two observed modes of the posterior. For the Gaussian phase retrieval setting, we show the mean and standard deviation of our reconstructions. 57

- 4.9 **Visualization of the intrinsic resolution of the EHT compressed sensing measurements.** The EHT measures sparse spatial frequencies of the image (i.e., components of the image’s Fourier transform). In order to generate the underlying image (c), all frequencies in the entire domain of (a) must be used. Restricting spatial frequencies to the ones in (a) and (b)’s green circle generates the target (d), where (b) is a zoom in of (a). The EHT samples a subset of the interior of the green circle, indicated by the sparse black samples in (b). Naively recovering an image using only these frequencies results in the *dirty image* (e), which is computed by $A^H y$. The 2D spatial Fourier frequency coverage represented with (u, v) positions is referred to as the UV coverage. 57
- 4.10 **Recovering a 60-day video of the M87 black hole with the EHT telescope array under a compressed sensing forward model.** Left: We demonstrate our method described in 4.3 using simulated measurements from 60 frames of an evolving black hole target with the forward model described in Fig. 4.9. Here we show the underlying images, dirty images ($A^H y$, see Fig. 4.9), and mean reconstruction, respectively. Additionally, we show the unwrapped space $\times$ time image, which is taken along the overlaid white ring illustrated in the day 1 underlying image. The bright-spot’s temporal trajectory of our reconstruction matches that of the truth. Right: We compare our method to various baselines methods. Our results are much sharper and exhibit less artifacts than AmbientGAN and TV-RML with weight λ . Note that we include results using AmbientGAN with both default parameters and fine-tuned parameters. 58
- 4.11 **Visualization of the EHT forward model of Sgr A* over time.** The spatial frequency space coordinates of Sgr A* from an ngEHT [57] array measured over a course of one night is shown on the top. Snapshots of the coordinates that are measured at different times are shown on the bottom. Note that the location and sparsity of the measurement coordinates change over time. 61

- 4.12 **Recovering a video of the Sgr A* black hole with a futuristic telescope array over the course of a night under a compressed sensing forward model.** We demonstrate our method described in 4.3 to perform video reconstruction on 60 frames of a video, where the measurements for each frame are generated by different forward models given by imaging Sgr A*. Here we show the underlying image, the dirty image ($A^H y$), and the mean of the reconstructed posterior. Additionally, we show the unwrapped space $\times$ time image, which is taken along the white ring illustrated in the 4:00 UTC underlying image. The bright-spot's temporal trajectory of our reconstruction matches that of the truth. The measurement noise is consistent with a black hole having a flux of 2 Janskys. 62
- 4.13 **Our reconstructions of selected frames from a video of M87*.** We show the ground truth, dirty image, target image (see Fig. 4.9), and empirical mean and standard deviation of our reconstructed image distribution for selected frames from the 60-frame M87* video. Additionally, we show the unwrapped space vs. time image, which is taken counter-clockwise along the ring; the trajectory of the bright spot in our reconstruction matches the true image. Our method reconstructs the primary features of the true underlying images while also reconstructing some high frequency features. 65
- 4.14 **Baseline comparisons.** We show results from the following baseline methods: regularized maximum likelihood with maximum entropy (MEM-RML), total variation (TV-RML), and total-squared-variation (TSV-RML), Deep Image Prior (DIP) [48], and AmbientGAN [35]. We show the average peak signal to noise ratio (PSNR) and normalized cross correlation (NXCorr) compared to the target images (see Fig. 4.13). Our method exhibits fewer artifacts and has the highest average PSNR and NXCorr, although PSNR is not an ideal metric due to the shift ambiguities in phase retrieval. 66

- 5.1 Source-receiver geometry and recovered 3D magmatic system from a subdomain of the Island of Hawai'i.** For our real data study and realistic-geometry synthetic tests, we consider the sources and receivers in a small region (magenta dashed rectangle) of the Island of Hawai'i that includes the Kilauea caldera. The black rectangle in the inset map shows the location of the area plotted in the main plotted map. We use a total of 18 receivers (magenta triangles) and 45,310 earthquake events (dots colored by depth). These events were located using the method described in [1], but with modifications discussed in Sec. 5.5. We visualize the cross sections A-A' and B-B' used for later figures (blue lines). The bottom figure depicts our reconstruction of the 3D magmatic system that underlies Kilauea with long-period (LP) volcanic events are highlighted and located at the bottom of a large magmatic volume. 78
- 5.2 Noise level mismatch.** We consider the impact of misspecifying the noise level b_P of our measurement model $\log p(y|V_P)$ when solving for V_P . We consider the effect of mismatch in assumed level of noise (e.g. over or under estimating). We find that when the noise is less than expected, the structures are similar but the amplitude of recovery is lower. When the noise is much higher than expected, the structures are obfuscated due to overfitting. We find it necessary to do early stopping to prevent overfitting to noise especially when the data noise is larger than expected. These are cross-sections taken along A-A'. . . 86
- 5.3 Physically motivated synthetic anomaly tests.** We perform synthetic tests to explore the sensitivity of our expected source-receiver geometry to various $\sim 10\%$ slower velocity anomalies in the subsurface. For various candidate anomalies, we show the following for A-A' (shown in Fig. 5.1): the true velocity model used to generate the synthetic travel times, the relative change to the initial model, and the mean of the inferred posterior and relative change to the initial model. Note, we plot these cross sections masking out the topography. 87

- 5.4 **Checkerboard tests** We perform inversion on the 1D model from [2] modulated with various checkerboards of 20% increase or decrease with a side length of s . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the mean and standard deviation of our reconstruction and relative change between the 1D model and our reconstruction for the A-A' cross-section. We include z-scores to quantify the accuracy of our uncertainty quantification. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage. Interestingly, the estimated uncertainty for $s = 4$ km is higher in low velocity regions such as the one at around 6 km depth and between 6 and 10 km in X , which is expected behavior. We include the full results for both V_P and V_S in the appendix. 88
- 5.5 **Eikonal tomography of Kilauea.** We show cross-sections of results from eikonal tomography of Kilauea as described in Sec. 5.4. We show the estimated mean and uncertainty from our approach along with the results from [2]. The darkened part in our estimated mean is given by the resolvability index described in 5.10 where the lighter region is resolvable and darker region is not resolvable for a resolution of 4 km. Our tomography shows evidence of a high V_P anomaly at around 5 to 12 km depth between 3 to 9 km along A-A'. As described in Sec. 5.3, we use Gaussian smoothing of the data-likelihood's gradient to prevent overfitting and over-interpretation of heterogeneous parts of our model. For reference, A is 19.44°N, 155.35°W, A' is 19.34°N, 155.15°W, and B is 19.37°N, 155.30°W and B' is 19.44°N, 155.25°W. 91
- 5.6 **High V_P/V_S regions under Kilauea.** We visualize the isosurfaces of high V_P/V_S regions beneath the Kilauea caldera with two views of the same volume. This high V_P/V_S structure lies between the long period events and the surface expression of the caldera. We estimate the volume of the magma column as 117 km³ using the region with a $V_P/V_S \geq 2.1$, which is visualized with the second outermost isosurface. Sources are colored as a function of depth. . . . 92

- 5.7 We show two cross sections of V_P/V_S for [2] and the mean of our reconstructed posterior as well as additional depth slices for the mean of our reconstructed posterior. The masked region in our estimated mean is given by the resolvability index described in 5.10. We observe a much broader range of V_P and V_S than in the one observed in [2] because we are not as dampened and invert for them separately. Our reconstruction shows evidence of a high V_P/V_S chamber underlying the Kilauea caldera. Additional depth slices are included in Fig. B.7. 92
- 5.8 **Melt fraction computation with uncertainty.** We estimate the melt fraction by applying a self-consistent scheme (SCS) effective medium approach. Assuming textural equilibrium for basaltic magma with an aspect ratio of pores and cracks to be 0.2 [3], we estimate a melt fraction of $15 \pm 2.7\%$ at 11 km depth. We compute an average V_S , referred to as μ , and uncertainty, referred to as σ . The solid line represents the melt fraction prediction with μ and the shaded region represents 1 standard deviation σ from μ . For reference, A is 19.44°N , 155.35°W , A' is 19.34°N , 155.15°W , and B is 19.37°N , 155.30°W and B' is 19.44°N , 155.25°W 93
- 6.1 **Our approach.** Our method uses time-of-flight measurements $\{T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)})\}_{\forall i,j}$ between sources $\{\mathbf{x}_s^{(j)}\}_{j=1}^{n_s}$ and receivers $\{\mathbf{x}_r^{(i)}\}_{i=1}^{n_r}$ where $n_s = 20$ and $n_r = 2$, and outputs the reconstructed velocity model $\hat{V}$ following the method outlined in Sec. 6.3. The input measurements and source-receiver locations are visualized as partially observed matrices. Each matrix corresponds to measurements between a single receiver (indicated by the cyan triangle) and all observed sources. The shaded region represents areas with missing information and each square of the unshaded region represents a single measurement. These potentially noisy measurements are then used in our method, which consists of two parts: 1. inpainting in the measurement space as described in Sec. 6.3 and 2. velocity reconstruction from the inpainted measurements as in Sec. 6.3. From the reconstructed velocities, the optimal one is selected, as highlighted in yellow. 107

- 6.2 **Noise-free imaging.** We generate synthetic measurements between 2 receivers and 20 sources using the ground truth velocity models (top row). The corresponding reconstructions (bottom row) are from solving Eq. 6.6 and selecting the field with the minimal trace. Our method is able to reconstruct the regions with source-receiver coverage fairly well. We use the same sensor and receiver setup for each synthetic experiment where the yellow stars are sources and cyan triangles are receivers. 112
- 6.3 **EikoLift on noisy measurements observed from only one surface of the domain.** We reconstruct velocity models from noisy synthetic measurements with a SNR of 25 dB. We use 4 receivers and 20 sources where the yellow stars are sources and cyan triangles are receivers. The columns from left to right are: ground truth velocity model, reconstruction with the adjoint [10] method and TV prior, reconstruction with a ToF-PINN [39], reconstruction from EikoLift using Eq. 6.6, and reconstructions EikoLift using Eq. 6.7 with two separate criteria: A. minimizing the trace and B. minimizing measurement error. Mean squared error (MSE) is included for each reconstruction. 113
- 6.4 **Increasing sensor coverage improves reconstruction.** We show reconstructions with increasingly better receiver coverage to illustrate improved performance as receivers increase from 2 to 4. We find that this is true for the adjoint as well as our specialized EikoLift-B setting using Eq. 6.7 with minimizing measurement error as the selection criteria. The EikoLift results using Eq. 6.6 are overly smooth due to the high ε_{data} needed to prevent infeasibility. We use between 2 to 4 receivers and 20 sources where the black stars are sources and cyan triangles are receivers. The top row corresponds to 2 receivers, the middle with 3 receivers, and the bottom with 4 receivers. Mean squared error (MSE) is included for each reconstruction. 114

6.5	Applying EikoLift to real data We show a 2D slice from a reconstruction using EikoLift on real seismic data from the Kilauea caldera with 5 receivers and 100 events. This reconstruction uses a subset of the data used to image in Chapter 5. Here, we show the central cross section A-A' as described in 5.1. We include the relative change from the 1D model used in [11], which is used to initialize the algorithm in Chapter 5.	117
7.1	Our approach applying FNO to the 2D acoustic wave equation. The inputs to the FNO model are the velocity model with dimensions $d \times d \times 1$ and the source location, indicated by the white star. The input velocity model is lifted to a higher dimensional space with size $d \times d \times w$ using a neural network. Then, we apply four Fourier operator layers, and finally project back to the target wavefield dimensions of $d \times d \times N$ using a neural network. The bottom panel shows details of the Fourier layer architecture: we define v to be the input. On top: We apply a Fourier transform F to v , then apply a linear transformation R to the lower Fourier modes, filtering out higher modes. Then we apply an inverse Fourier transform F^{-1} . On the bottom: we apply a local linear transform W to v	125
7.4	Model generalization experiments. (a) the source-receiver locations with receivers in blue and source in red; (b) a velocity model with a homogeneous background of 3 km/s and a 5% square anomaly; (c) waveform simulated with SEM (black) and FNO (red). (d-f) same as (a-c), but for an input velocity model with 2x finer resolution than trained on. These experiments show that the model is not just memorizing the solutions, but is able to generalize to entirely new conditions.	128
7.2	Examples of two validation wavefield simulations from the trained FNO model. (a) the source-receiver locations with receivers in blue and source in red; (b) the velocity structure; (c) waveforms simulated with SEM (black) and FNO (red). (d-f) same as (a-c), but for a different velocity model. The relative ℓ_2 loss of the two examples are 0.180 and 0.363, respectively, which are representative of the entire validation data set with an average ℓ_2 loss of 0.273. We demonstrate that the FNO simulation results are able to capture both the major arrivals as well as some reflections.	129

7.3	Model performance as a function of the number of training samples. (a) Training and validation loss curves as a function of different numbers of training samples. (b) Example waveform fitting of a single training example from models trained with varying number of training examples. (c) Example waveform fitting of a single validation example from models trained with varying number of training examples. The numbers to the right of each waveform shows the relative ℓ_2 misfit. This shows that the model trained on 4800 samples is able to capture the reflections, whereas the model trained on smaller number of samples does not generalize to reflections in the validation example.	130
7.5	Example of a full waveform inversion using FNO. (a)(c)(e) True velocity models with source locations indicated by red circles and receivers placed at every node of the 64×64 grid ($10 \text{ km} \times 10 \text{ km}$ region). (b) Reconstruction using SEM and adjoint tomography. (d)(f) Reconstruction using FNO as the forward model and automatic differentiation to compute gradients. No regularization is used for these experiments.	131
7.6	Model generalization experiments 1: random fields of V_s and V_p/V_s model with 4-times roughness of the training data. Relative L2 loss of the U-NO simulation is 0.182.	134
7.7	Output evaluation. Distribution of relative L2 loss (top) and correlation coefficient (bottom) between the U-NO predictions and ground truth. From left to right, the columns are corresponding to the experiments in Figure 3, 5, 6. The red and black dashed lines mark the mean and standard deviation of the histograms.	135
7.8	Full waveform inversion. The inversion for a random subpanel from the Marmousi model. 14 sources are placed in a ring shape and receivers are placed at every node of the 64×64 grid. From left to right, the columns represent true velocity model, initial model for inversion, inverted model without regularization, and inverted model with damping and smoothing. The top and bottom rows are V_s and V_p models, respectively.	136

- B.1 Checkerboard tests** We perform inversion on the 1D model from [2] modulated with various checkerboards of 20% increase or decrease for both V_P and V_S . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the reconstruction and relative change for A-A'. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage for both V_P and V_S . We include z-scores to quantify the accuracy of our uncertainty quantification. We find that the uncertainty is underestimated near the sharp interfaces of the checkerboard as well as in the high frequency regions that are below the resolvability of our method. 147
- B.2 Checkerboard tests without early stopping** We perform inversion on the 1D model from [2] modulated with various checkerboards of 20% increase or decrease for both V_P and V_S . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the reconstruction and relative change for A-A'. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage for both V_P and V_S 148
- B.3 Physically motivated synthetic anomaly tests.** We perform synthetic tests to explore the sensitivity of our expected source-receiver geometry to various $\sim 10\%$ low velocity anomalies in the subsurface. For various candidate anomalies, we show the following for A-A': the true velocity model used to generate the synthetic travel times, the relative change to the initial model, and the mean of the inferred posterior and relative change to the initial model for both an $k = 250$ and $k = 500$. Note, we plot these cross sections masking out the topography. 149
- B.4 V_P/V_S ratio synthetic resolvability test.** We include two high V_P/V_S ratio anomalies within the initial velocity models to evaluate the ability (left panel). The right panel shows the result of applying MAP estimation process to the synthetic data to resolve the input anomalies. 149

B.5	Gradient smoothing. We visualize the reconstructions along A-A' for different values of k_{stop} , k_{max} , and k values for the gradient smoothing as described by Eq. 5.9. We find that the optimal set of parameters is $k_{stop} = 250$, $k_{max} = 500$, and $k = 250$, which is highlighted in green. We include the ground truth velocity model for reference.	150
B.6	Noise model mismatch We consider the impact of misspecifying the noise model η and/or measurement model $\log p(y x)$. We consider the effect of mismatch in the assumed noise model distribution (e.g. Gaussian vs. Laplacian noise). For the Gaussian distribution, we assume a standard deviation $\sigma_P = .11$ and $\sigma_S = 0.31$. For the Laplacian distribution, we assume $b_P = 0.08$ and $b_S = 0.15$. The noise model and data-fit model highlighted in green is the one that we expect our real data to match. Truth is in Fig. B.5.	150
B.7	All depth slices of V_P/V_S. Here we show all depth slices of V_P/V_S from $z = 0$ km to $z = 17$ km. Refer to the color map in Fig. 5.7 or Fig. B.10. The lighter region indicates resolvability with a checkerboard of 4 km.	151
B.8	Reconstructions at various SSST steps. Here we show reconstructions at various steps in the SSST localization procedure along the A-A' cross section. We find that the recovered structures are stable across these different steps.	151
B.9	1D profiles. Various depth slices across different lat-lon locations. The black line is the 1D model from [4] that we use for our initial model.	152
B.10	Travel time misfit distributions Here we show the distributions of travel time misfits for both the initial velocity model and the recovered posterior mean.	153
C.1	Block diagonal formulation. This is a schematic of how we implement the block diagonal representation as described in Sec. 3.1 of the main text and in Appendix C.2.	156

- C.2 Toy example with homogeneous ground truth.** We demonstrate our method on a simple velocity model with a homogeneous velocity of 6 km/s. For this toy example, we choose to use a homogeneous velocity model because there is no generalization error between $\text{span}(\phi)$ and the ToF fields $T^{(i)}$. We test our method using Eq. 8 from the main text on noise-free measurements generated by the Fast Marching Method as well as noisy measurements with i.i.d additive Gaussian noise with an SNR of 25. In the noise-free case, we find that the weight-vectors $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$, described in Sec. 3.2, fit the ideal reconstruction of a one hot vector where the only non-zero value corresponds to $T_{0,x_r^{(i)}}$. This holds when we increase the number of basis vectors for each $M^{(i)}$ to be $K = 6$. In the noisy measurement setting, we also find that the weight-vectors $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ fit the ideal reconstruction and we are able to recover the homogeneous velocity. We find that we need to increase ε_{data} when our measurements are noisy. 159
- C.3 Ablations for Eq. 8 for noisy measurements.** Here we show reconstructions using various values of ε_{PDE} and ε_{data} used in Eq. 8. We use noisy measurements with a SNR of 25. When ε_{data} is substantially less than the expected measurement noise, the optimization yields unexpected behavior. When ε_{data} is too large such that it is on the order of the signal, it yields overly smooth reconstructions. $\varepsilon_{PDE} = 0.1$ and $\varepsilon_{data} = 1$ corresponds to the parameters used for the results in the main text. 160
- C.4 Ablations for Eq. 9 for noisy measurements.** Here we show reconstructions using various values of ε_{PDE} and λ used in Eq. 9. We use noisy measurements with a SNR of 25. The reconstruction under-fits the measurements when λ is too small and overfits when λ is too large. $\varepsilon_{PDE} = 0.1$ and $\lambda = 100$ corresponds to the parameters used for the results in the main text. 161
- C.5 With increasing noise.** We show results across various SNR values. A SNR of 25 corresponds to $\sigma = 0.1$. In all cases, increasing the SNR improves reconstruction quality. 162

C.6 Change in reconstruction with increasing K.	
We show results from optimizing Eq. 9 from the main text using various values of K . $K = 6$ corresponds to the same parameter setup as in the figures from the main text. When K is too small, the reconstructions are overly smooth, and when K is too large, the reconstruction overfits to noise.	163
C.7 Full coverage	
We show full sensor coverage reconstructions for the adjoint method with TV prior, PINN, and our method using Eq. 9 from the main text with different selection criteria. Mean squared error (MSE) is included for each reconstruction.	164

LIST OF TABLES

<i>Number</i>	<i>Page</i>
4.1 Quantitative comparisons to baselines. We show the mean PSNR between the reconstructions of different methods (ours, AmbientGAN [35], DIP [48], and TV-RML) with either the true underlying image or the target in the case of black hole compressed sensing. The highest PSNR for each set of measurements is in bold. We fix the hyperparameters of the baselines across each forward model, empirically selecting them for good performance on each type of problem. Note that the range in performance depends on the choice of hyperparameters. For denoising, the DIP results corresponds to 100 and 300 epochs for the fewer and many epochs baselines, respectively, and the lower and higher λ 's are $1e3$ and $1e4$, respectively. For black hole compressed sensing, the DIP results corresponds to 1000 and 3000 epochs for the fewer and many epochs baselines, respectively, and the lower and higher λ 's are $1e2$ and $1e3$, respectively. Visual examples of celeb B. are shown in the supplemental materials.	59
4.2 Phase Retrieval PSNRs for different measurement operators. Each column corresponds to the type of measurements that the method was given (a total of $N = 150$). In the case of "Both", 75 Gaussian measurements and 75 Fourier measurements were given. We then show the average PSNR for our reconstructions given the specific measurement operator (either Fourier or Gaussian). Note that when given both Gaussian and Fourier measurement examples our method exhibits a higher PSNR on the recovered images from Fourier examples than when given only Fourier measurements. Additionally, there is only a slight decrease in performance on the Gaussian measurement reconstructions as compared to given entirely Gaussian examples.	60

- 6.1 We include quantitative comparisons across 20 examples for: 1. the adjoint method w/ TV regularization based on [10], 2. physics-informed neural network (PINN) based on [39], 3. our method using Eq. 6.6 selecting the minimum trace, 4. our method using Eq. 6.7 selecting the minimum trace, and 5. our method using Eq. 6.7 with the minimal measurement error. We include the mean and standard deviation of the $\text{MSE}(\uparrow)$ for both $n_r = 4$ and $n_r = 12$ with $\text{SNR} = 25$ dB and $n_r = 12$ for noise-free measurements. 111

Chapter 1

INTRODUCTION

Physical systems often reveal hidden structure through the way they respond to disturbances. While a single disturbance may only uncover fragmented information, many observations taken together can reveal a more complete picture. For example, mundane, naturally occurring vibrations from sources such as ocean tides and human activity can be used to recover images of Earth’s subsurface structure [1], [2]. This principle of harnessing passive measurements underlies many fundamental problems in science and engineering, where the goal is to use observations to recover hidden information about the underlying physical system that generated them. These problems are known as *inverse problems*, and they span applications ranging from GPS localization to detecting dark matter through the gravitational lensing of light.

In the age of deep learning, solving inverse problems has been simplified through data-driven supervised learning approaches. These supervised learning methods leverage vast amounts of paired training data, mapping measurements to reconstructed sources in the case of inverse problems. Training supervised models to solve inverse problems requires extensive collections of measurement-source pairs, which presents a fundamental challenge for many scientific applications. Paired datasets are often either too sparse to train with or prohibitively expensive to acquire at scale. Consider, for instance, that obtaining even a single ground truth image of the Earth’s subsurface structure remains impossible with current technology. Without real paired data, the remaining option is to resort to training with synthetic data. This approach introduces its own complications, as generating realistic synthetic source distributions that capture the complexity and variability of natural phenomena can be remarkably difficult.

Despite the widespread success of supervised learning, scientists often cannot solve inverse problems from data alone. Under these constraints, solutions must incorporate prior knowledge about the underlying system, including how the measurements are generated (i.e., physical equations) and the characteristics we expect the unobserved structure to possess. Accurate knowledge of both components is critical for identifying unobservable sources. Imaging systems designed in tandem with this domain knowledge often have the most success. In data-scarce settings, leveraging

the fundamental structure of natural systems enables our understanding of invisible phenomena.

Solving inverse problems requires knowing the structure of several key ingredients, including models of the measurement process, assumptions about the underlying system, and algorithms for inference. Each of these components represents an important piece of the puzzle and must be carefully designed and understood. When there is uncertainty in the model of the measurement process or lack of clarity in the assumptions of the underlying system, the inverse problem becomes significantly more difficult. This becomes further complicated when there are multiple inverse problems that must be solved simultaneously, where each problem has weak individual constraints requiring that it be simultaneously resolved.

Such challenging scenarios demand new algorithmic approaches, as traditional methods prove insufficient when faced with these compounded uncertainties. This raises a fundamental question: how can we solve inverse problems when the available information seems so limited? This thesis aims to offer methods that can solve inverse problems under missing information, operate under non-linear forward models, and offer more efficient methods for solving non-linear imaging inverse problems. This thesis demonstrates that hidden forms of regularization, which emerge from unconventional approaches to solving inverse problems, can often provide the additional constraints needed. While the methods discussed in this thesis focus primarily on *imaging* inverse problems, many of the underlying frameworks are general and can be adapted to other domains.

Using earthquakes to interrogate the earth

Seismic imaging and inversion presents a rich set of inverse problems, making it an excellent testbed for developing new methods. Several key challenges in seismic inverse problems have motivated much of this thesis work. These challenges include multiple interdependent inverse problems, challenging source-receiver geometry, and computationally expensive simulations [3].

Several fundamental inverse problems in seismology are tightly coupled, meaning that solving for one quantity inherently requires knowledge of the others. These inverse problems include tomography, hypocenter localization, event characterization, and source (or earthquake) characterization. More specifically, tomography involves imaging the Earth's subsurface structure. Hypocenter localization involves identifying the precise 3D location and origin time of an earthquake, the accuracy

of which depends on having a reliable tomographic model. Event characterization involves describing the temporal nature of a rupture, while source characterization involves describing the directionality of seismic forces involved, distinguishing between sources that are explosive, strike-slip, or subduction. Since these problems are deeply interdependent yet intractable to solve simultaneously, they are often solved individually under strong assumptions about the remaining quantities. These assumptions are necessary, but can introduce systematic bias. The work presented in Chapter 3 considers simultaneously solving tomography and hypocenter localization as well as event characterization and tomography through empirical Green's functions.

Many seismic studies utilize surface sensors to record naturally occurring earthquakes in order to probe the structure and dynamics of the Earth's interior. However, sensors are limited to the surface or near-surface for a variety of reasons including feasibility and cost. This restriction in geometry means that measurements are recorded from a narrow range of angles, resulting in a limited-angle inverse problem. This difficulty is further compounded by the geometry of natural seismicity. Earthquakes are not distributed uniformly, but instead are often clustered around specific geological features such as fault zones or magmatic systems. This clustering results in a highly non-uniform source-receiver geometry. Limited-angle inverse problems are tricky to solve because they are sensitive to initial conditions. Small perturbations in the initial model can lead to vastly different solutions. The approach in Chapter 6 aims to address this through a convex approach, which is invariant to the initialization.

Simulating the physics of various seismic phenomena such as wave propagation is extremely compute intensive, yet accurate forward modeling is a critical component for solving inverse problems. Simplifications are often applied to make inverse problems more tractable, but such approximations can introduce errors that propagate through the inversion and degrade the quality of the final results. These computational constraints motivate the need for alternatives to accelerate the forward simulations themselves. One such alternative is through surrogate models, including neural network surrogates which enable access to efficient computation of gradients, which is necessary for gradient-based approaches. Chapters 6 and 7 discuss alternative approaches to solving forward simulation and inversion without the need to use slow traditional solvers.

These outlined challenges motivate the approaches discussed in Chapter 3, Chapter

5, Chapter 6, and Chapter 7.

Black hole imaging

Imaging a black hole is as challenging as taking a photo of a donut on the moon with a conventional camera [4]. The footprint of a black hole on the night sky is so vanishingly small that a single telescope on Earth is unable to resolve it. Achieving the required resolving power requires an Earth-sized telescope. To produce an Earth-sized telescope, astronomers have linked up a series of telescopes and use a technique called Very Long Baseline Interferometry (VLBI) to effectively synthesize an aperture as big as the Earth itself. By harnessing the natural rotation of the Earth to increase the number of measurements, the imaging problem becomes more tractable. However, imaging a black hole is still highly ill-posed and requires using image priors. In Chapter 4, I discuss an approach to reconstruct frames of videos of black holes by leveraging shared structure across frames despite the corrupted measurements. We apply this approach to synthetic version of both M87* and Sagittarius A*.

References

- [1] K. Wapenaar, “Retrieving the Elastodynamic Green’s Function of an Arbitrary Inhomogeneous Medium by Cross Correlation,” *Physical review letters*, vol. 93, no. 25, p. 254 301, 2004.
- [2] G. Bensen, M. Ritzwoller, M. Barmin, *et al.*, “Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements,” *Geophysical journal international*, vol. 169, no. 3, pp. 1239–1260, 2007.
- [3] P. M. Shearer, *Introduction to seismology*. Cambridge university press, 2019.
- [4] Event Horizon Telescope Collaboration, *Astronomers reveal first image of the black hole at the heart of our galaxy*, <https://eventhorizontelescope.org/blog/astronomers-reveal-first-image-black-hole-heart-our-galaxy>, Accessed July 25, 2025, Event Horizon Telescope, May 2022.

Chapter 2

BACKGROUND

2.1 Inverse problems

In the inverse problems that will be discussed in this thesis, the *indirectly observed* sources $\mathbf{x} \in \mathbb{R}^{D_x}$ and *observed* measurements $\mathbf{y} \in \mathbb{R}^{D_y}$ are related via a forward model f :

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon. \quad (2.1)$$

$f(\cdot)$, referred to as the *forward model*, represents deterministic system behaviors, such as the physics of light or acoustic propagation, and ε represents stochastic variability of the measurement process. Often, ε is assumed to be of the form of zero-mean Gaussian noise where $\varepsilon \sim \mathcal{N}(0, \sigma I)$ for $\sigma \in \mathbb{R}^n$. For many parts of this thesis, $\mathbf{x}$ is a 2D image with $D_x = n \times n$, but for some section of this thesis, $\mathbf{x}$ may be a 1D time series, a 3D volume, or a less structured quantity. In many inverse problems, $D_y \ll D_x$, but this is not always the case, as in some seismic imaging applications like in Chapter 5.

The goal when solving inverse problems is to find a reconstruction $\hat{\mathbf{x}} \approx \mathbf{x}$ (i.e., close to the true solution $\mathbf{x}$) from the measurements $\mathbf{y}$. A natural way to find $\hat{\mathbf{x}}$ is to formulate the inverse map $f^{inv}(\cdot)$ and get a reconstruction $\hat{\mathbf{x}} = f^{inv}(\mathbf{y})$. However,

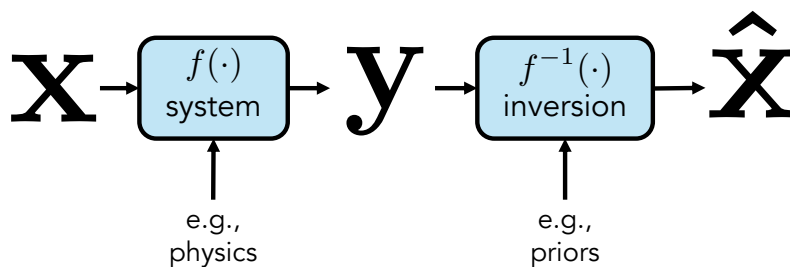


Figure 2.1: **Measurement acquisition and inverse problems.** The full measurement acquisition and inversion scheme is described by this diagram where the unobserved source of interest is $\mathbf{x}$ but what is observed is $\mathbf{y}$. This measurement process is governed by some system, often understood using physics. To recover the unobserved source $\mathbf{x}$, we use an inversion algorithm, which often involves priors or regularizers, to recover a reconstruction $\hat{\mathbf{x}}$.

f^{inv} is often not well defined. This occurs when the problem is *ill-posed*, which arises from a variety of reasons including sparse measurements, poor measurement coverage, or a non-linear forward model. Moreover, stochastic noise cannot be directly incorporated into solving $\hat{\mathbf{x}} = f^{inv}(\mathbf{y})$. This is described in Fig. 2.1.

Forward models

A forward model $f(\cdot)$ describes how measurements $\mathbf{y}$ are generated from unknown sources $\mathbf{x}$; in consumer photography, this is governed by the way a camera takes a photo, and in seismic tomography, the forward model is governed by wave physics. For Bayesian inference, the forward model is described using $p(\mathbf{y}|\mathbf{x})$. Describing $\log p(\mathbf{y}|\mathbf{x})$ is often assumed to be the easiest part of solving imaging inverse problems; $f(\cdot)$ is assumed to be a linear operator with i.i.d. mean-zero Gaussian noise, thus resulting in $\log p(\mathbf{y}|\mathbf{x}) \propto \frac{1}{\sigma^2} \|\mathbf{y} - A\mathbf{x}\|_2^2$.

However, this is more challenging to characterize in real-world scientific imaging where noise is poorly characterized and the forward model can be non-linear [1], [2]. Moreover, there are often parameters of the forward model itself that might not be well known; locating an earthquake requires knowledge of the Earth’s 3D subsurface material composition, which is poorly characterized [2]. Approximations of the forward model, noise, and forward model parameters are often used to make characterizing $p(\mathbf{y}|\mathbf{x})$ easier. In this thesis, I will discuss how we approach non-linear imaging problems, including when the forward model is poorly characterized.

Priors

In ill-posed problems, priors are used to inject our beliefs of what we believe to be a good reconstruction. Traditionally, this meant describing an image using a parametric equation, such as the popular total variation prior [3], but modern approaches using deep generative models have been used as data driven priors [4]. Like with any deep learning model, deep data-driven priors exhibit poor generalization to out of distribution data. Thus, having the right training dataset for the data-driven prior is important. To combat this, there are many lines of work that aim to tackle implicit priors instead of explicit priors [5]–[8]. In this thesis, I leverage various forms of implicit priors to solve imaging problems without good priors.

Bayesian inference

To solve for our unobserved source $\mathbf{x}$, we want to use both the forward model and prior to inform our recovery. Given a particular forward model and prior construction, we can utilize Bayesian inference to find the *maximum a posteriori* (MAP) solution, which is the minimizer of the following optimization problem:

$$\begin{aligned}
 \hat{\mathbf{x}} &= \arg \min_{\mathbf{x}} -\log p(\mathbf{x}|\mathbf{y}) \\
 &= \arg \min_{\mathbf{x}} -[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}) - \log p(\mathbf{y})] \\
 &= \arg \min_{\mathbf{x}} -[\underbrace{\log p(\mathbf{y}|\mathbf{x})}_{\text{data likelihood}} + \underbrace{\log p(\mathbf{x})}_{\text{prior}}].
 \end{aligned} \tag{2.2}$$

In many cases, the regularizer that we want to use is not inherently probabilistic (i.e., it does not have a particular probability density function $p(\mathbf{x})$) and is of the form $\mathcal{R}(\mathbf{x})$. Since the regularizer is uncalibrated in that the magnitude needed might be unknown, an additional hyperparameter λ is often used. This becomes a regularized maximum likelihood (RML) problem and is described by:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} -[\underbrace{\log p(\mathbf{y}|\mathbf{x})}_{\text{data likelihood}} + \underbrace{\lambda \mathcal{R}(\mathbf{x})}_{\text{regularizer}}]. \tag{2.3}$$

Variational inference

Posterior estimation is a useful tool in ill-posed imaging problems because it provides estimates of uncertainty for image reconstruction. Two popular classes of posterior estimation methods are Markov Chain Monte Carlo (MCMC) [9] and variational inference (VI) [10]. Although MCMC is exact and asymptotically converges to the true posterior, it is computationally demanding to reach the true solution, which is hard to validate. Moreover, it scales poorly to high dimensional posterior distributions and can be prohibitively expensive computationally even for 2D imaging problems[11]. On the other hand, VI is much more efficient at estimating the posterior by constraining estimation over a tractable class of parameterized candidate distributions, but approximates the distribution through this tractable candidate set, making it less exact.

VI poses the Bayesian posterior inference problem as one of optimization, in which the target distribution is approximated with a parameterized candidate distribution [11]. This candidate distribution is the one that best fits the target distribution and is a member of a pre-specified class of distributions, referred to as the *varia-*

tional class. It is found through optimization by minimizing a loss (e.g., Kullback Leibler divergence [12]) between distributions through $\arg \min_{\theta} D_{\text{KL}}(q_{\theta}||p)$.

Following from [11],

$$\begin{aligned}
\theta^* &= \arg \min_{\theta \in \mathbb{P}} D_{\text{KL}}(q_{\theta}(\mathbf{x})||p(\mathbf{x}|y)) \\
&= \arg \min_{\theta \in \mathbb{P}} E_{\mathbf{x} \sim q_{\theta}(\mathbf{x})} [\log q_{\theta}(\mathbf{x}) - \log p(\mathbf{x}|\mathbf{y})] \\
&= \arg \min_{\theta \in \mathbb{P}} E_{\mathbf{x} \sim q_{\theta}(\mathbf{x})} [\log q_{\theta}(\mathbf{x}) - \log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) \\
&\quad + \log p(\mathbf{y})] \\
&= \arg \min_{\theta \in \mathbb{P}} E_{\mathbf{x} \sim q_{\theta}(\mathbf{x})} [\underbrace{\log q_{\theta}(\mathbf{x})}_{\text{neg log entropy}} - \underbrace{\log p(\mathbf{y}|\mathbf{x})}_{\text{log data likelihood}} - \underbrace{\log p(\mathbf{x})}_{\text{log prior}}]. \tag{2.4}
\end{aligned}$$

The choice of variational class is extremely important to well-estimate the posterior, but having more complex variational classes can make the variational inference problem less tractable. There is a trade-off that must be considered between the expressivity of the variational distribution and the efficiency of the optimization problem.

2.2 Seismic imaging

This section focuses on describing the forward and inverse problems in *travel time tomography*, which uses the time-of-flight information of an earthquake to a sensor on the surface. This background is necessary to understand the following chapter so this thesis: Chapters 3, 5, and 6. Chapter 7 covers the extension to the full waveform, which shares some similarities in mechanics to travel time tomography.

Eikonal equation and forward simulation

The speed at which a signal travels through a material is dictated by the material's underlying heterogeneous properties. This relationship is governed by the eikonal equation, which relates the velocity model $V \in C(\mathbb{R}^d)$ to the time-of-flight field $T_V \in C^1(\mathbb{R}^d \times \mathbb{R}^d)$. The velocity field $V(\mathbf{x})$ is positive and stores the speed at which a wave propagates through each point $\mathbf{x}$ in the medium, and the time-of-flight field $T_V(\mathbf{x}_s, \mathbf{x}_r)$ is nonnegative and specifies the time a disturbance at source point $\mathbf{x}_s$ is observed at receiver point $\mathbf{x}_r$. These fields are related by the *eikonal equation*:

$$||\nabla_{\mathbf{x}_r} T_V(\mathbf{x}_s, \mathbf{x}_r)||_2^2 = \frac{1}{V(\mathbf{x}_r)^2}, \quad T_V(\mathbf{x}, \mathbf{x}) = 0. \tag{2.5}$$

Here, d is the dimension of the medium. If we are given a fully sampled time-of-flight field, we can compute the velocity model by solving the eikonal equation (Eq. 2.5) via a finite-difference scheme.

The forward simulation problem is to determine the travel-time field $T_V(\mathbf{x}_s, \mathbf{x}_r)$ between specified points $\mathbf{x}_s$ and $\mathbf{x}_r$ from a given velocity field V . Different V 's will produce different times of flight $T_V(\mathbf{x}_s, \mathbf{x}_r)$, which is why we denote it as T_V . This forward model is typically computed through finite difference methods, the most common one being the Fast Marching Method (FMM) [13].

Reciprocity: The operator $T(\mathbf{x}_s, \mathbf{x}_r)$ is symmetric, that is, $T(\mathbf{x}_s, \mathbf{x}_r) = T(\mathbf{x}_r, \mathbf{x}_s)$ [2]. Thus, everything discussed above, and everything discussed in later sections, still holds if we swap $\mathbf{x}_s$ with $\mathbf{x}_r$.

Travel-time tomography and adjoint methods

The inverse problem, which involves reconstructing the velocity model from sparse travel-time observations, is a highly non-convex tomography problem. Here, the goal is to solve for the velocity model $V(\mathbf{x})$ from a sparse set of noisy observations $T_{obs}(\mathbf{x}_s, \mathbf{x}_r)$ between all receivers $\mathcal{R} = \{\mathbf{x}_r^{(i)} \in \mathbb{R}^d\}_{i=1}^{n_r}$ and sources $\mathcal{S} = \{\mathbf{x}_s^{(j)} \in \mathbb{R}^d\}_{j=1}^{n_s}$. We refer to this problem as (TTT), but this problem is also known as *time-of-flight (ToF) imaging* within certain domains.

TTT aims to find the velocity model that best fits the observed sparse set of noisy data. Assuming the additive measurement noise is modeled as $\eta_{ij} \sim \mathcal{N}(0, \sigma_{ij})$, TTT can be solved by the following optimization problem:

$$\hat{V} = \arg \min_V \sum_{i \in [n_r], j \in [n_s]} \frac{1}{\sigma_{ij}^2} (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - T_V(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}))^2 \quad (2.6)$$

where $T_V(\mathbf{x}_s, \mathbf{x}_r)$, induced by the velocity model V , is the simulated travel-time between $\mathbf{x}_s$ and $\mathbf{x}_r$ and $T_{obs}(\mathbf{x}_s, \mathbf{x}_r)$ is the measured travel-time at the sensor. Simulating $T_V(\mathbf{x}_s, \mathbf{x}_r)$ requires the use of a forward simulator such as the Fast Marching Method (FMM).

TTT can be solved using an adjoint method applied to the eikonal equation [14]–[17]. Although the eikonal-based adjoint method obeys the physics exactly, optimizing V is extremely challenging; the eikonal-based adjoint method is sensitive to the initialization of V and each gradient step can be computationally slow, especially for large-scale high-resolution reconstructions which can be on the order of minutes for each step. Especially in cases with many gradient steps, such as posterior

estimation and other sampling-based approaches, using adjoint eikonal methods can be computationally prohibitive. A much faster alternative solves an inverse ray-tracing problem [18]–[24]. Inverse ray-tracing methods are often preferred over eikonal imaging due to their efficiency, but make approximations that yield less accurate reconstructions.

References

- [1] T. E. H. T. Collaboration, “First m87 event horizon telescope results. iv. imaging the central supermassive black hole,” *The Astrophysical Journal Letters*, vol. 875, no. 1, p. L4, 2019.
- [2] P. M. Shearer, *Introduction to seismology*. Cambridge university press, 2019.
- [3] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259–268, 1992.
- [4] B. T. Feng, J. Smith, M. Rubinstein, H. Chang, K. L. Bouman, and W. T. Freeman, “Score-based diffusion models as principled priors for inverse imaging,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 520–10 531.
- [5] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.
- [6] R. Heckel and P. Hand, “Deep decoder: Concise image representations from untrained non-convolutional networks,” *International Conference on Learning Representations (ICLR)*, 2019.
- [7] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, “Rare: Image reconstruction using deep priors learned without groundtruth,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1088–1099, 2020.
- [8] J. Batson and L. Royer, “Noise2self: Blind denoising by self-supervision,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 524–533.
- [9] M. Betancourt and M. Girolami, “Hamiltonian monte carlo for hierarchical models,” *Current trends in Bayesian methodology with applications*, vol. 79, no. 30, pp. 2–4, 2015.
- [10] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [11] H. Sun and K. L. Bouman, “Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2628–2637.

- [12] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [13] J. A. Sethian, “Fast marching methods,” *SIAM review*, vol. 41, no. 2, pp. 199–235, 1999.
- [14] E. Biondi, W. Zhu, J. Li, E. F. Williams, and Z. Zhan, “An upper-crust lid over the long valley magma chamber,” *Science Advances*, vol. 9, no. 42, eadi9878, 2023.
- [15] R.-E. Plessix, “A review of the adjoint-state method for computing the gradient of a functional with geophysical applications,” *Geophysical Journal International*, vol. 167, no. 2, pp. 495–503, 2006.
- [16] S. Li, A. Vladimirovsky, and S. Fomel, “First-break traveltimes tomography with the double-square-root eikonal equation,” *Geophysics*, vol. 78, no. 6, U89–U101, 2013.
- [17] E. Biondi, G. Barnier, R. G. Clapp, F. Picetti, and S. Farris, “An object-oriented optimization framework for large-scale inverse problems,” *Computers & Geosciences*, vol. 154, p. 104790, 2021.
- [18] J. R. Evans, D. Eberhart-Phillips, and C. Thurber, “User’s manual for simulps12 for imaging vp and vp/vs; a derivative of the "thurber" tomographic inversion simul3 for local earthquakes and explosions,” US Geological Survey, Tech. Rep., 1994.
- [19] H. Zhang and C. H. Thurber, “Double-difference tomography: The method and its application to the hayward fault, california,” *Bulletin of the Seismological Society of America*, vol. 93, no. 5, pp. 1875–1889, 2003.
- [20] K. Koketsu and S. Sekine, “Pseudo-bending method for three-dimensional seismic ray tracing in a spherical earth with discontinuities,” *Geophysical Journal International*, vol. 132, no. 2, pp. 339–346, 1998.
- [21] X. Zou, F. Vandenberghe, B. Wang, *et al.*, “A ray-tracing operator and its adjoint for the use of gps/met refraction angle measurements,” *Journal of Geophysical Research: Atmospheres*, vol. 104, no. D18, pp. 22301–22318, 1999.
- [22] F. Anis, Y. Lou, A. Conjusteau, *et al.*, “Investigation of the adjoint-state method for ultrasound computed tomography: A numerical and experimental study,” in *Photons Plus Ultrasound: Imaging and Sensing 2014*, SPIE, vol. 8943, 2014, pp. 452–457.
- [23] A. Chakrabarti, “Learning sensor multiplexing design through back-propagation,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [24] È. Campeau-Poirier, L. Perreault-Levasseur, A. Coogan, and Y. Hezaveh, “Time delay cosmography with a neural ratio estimator,” *arXiv preprint arXiv:2309.16063*, 2023.

Chapter 3

IMAGING WITH UNCERTAIN FORWARD MODELS

This chapter is adapted from previously published work [1], [2].

Abstract

Typically, inversion algorithms assume that a forward model, which relates a source to its resulting measurements, is known and fixed. Using collected indirect measurements and the forward model, the goal becomes to recover the source. When the forward model is unknown, or imperfect, artifacts due to model mismatch occur in the recovery of the source. In this chapter, we study the problem of blind inversion: solving an inverse problem with unknown or imperfect knowledge of the forward model parameters. We propose DeepGEM, a variational Expectation-Maximization (EM) framework that can be used to solve for the unknown parameters of the forward model in an unsupervised manner. DeepGEM makes use of a normalizing flow generative network to efficiently capture complex posterior distributions, which leads to more accurate evaluation of the source’s posterior distribution used in EM. We showcase the effectiveness of our DeepGEM approach by achieving strong performance on the challenging problem of blind seismic tomography, where we significantly outperform the standard method used in seismology. We also demonstrate the generality of DeepGEM by applying it to a simple case of blind deconvolution.

3.1 Introduction

Physics-based inversion methods typically recover an unknown source from indirect measurements by assuming that the source and measurements are related via a known forward model [3]–[5]. For example, non-blind deconvolution algorithms often assume that a measured blurry image is related to its true sharp image via a known spatially-invariant blur kernel [6]; and traditional seismic inversion methods assume that the spatially-varying velocity of the Earth’s interior is known *a priori* when solving for an earthquake’s hypocenter [5]. However, these “known” forward models are generally idealized and ignore intricacies of the systems that are either hard to capture or simply unknown. Inversion algorithms with forward *model mismatch* result in biased reconstructions. For instance, bias is regularly seen in

non-blind deconvolution results, where reconstruction artifacts are often present due to the use of an incorrect blur kernel [6].

In order to reduce the effects of model mismatch, in this chapter we tackle the problem of *blind inversion*: solving an inverse problem without knowledge of the underlying forward model parameters. When considering how learning can help, a natural first idea might be to learn a direct map from measurements to the desired source via supervised learning. However, such a “model-free” approach is generally not practical, due to the lack of available ground truth training data. For example, the blur kernel caused by handheld camera shake cannot be reproduced to get a training set of sharp-blurry pairs to train a deconvolution approach. In seismic tomography, synthetic earthquakes cannot be placed densely throughout the interior of the Earth to measure the ground response as a function of hypocenter location. An alternative approach, which we adopt, is to develop unsupervised methods that treat the true forward model as something that is unobserved and must be inferred. An additional consideration is that we must solve for the true forward model from a single dataset, without knowledge of the true source that produces the measurements

In this chapter, we propose Deep Generalized Expectation-Maximization (DeepGEM) for solving blind inverse problems. Using the indirect measurements as input, DeepGEM jointly estimates the source and forward model that together produce the observed measurements. DeepGEM is a variational inference based framework that makes use of deep learning machinery to easily capture and optimize complex probabilistic distributions that cannot be easily integrated in analytic Expectation-Maximization (EM) solutions. Our proposed framework is generic and can be applied to blind inversion problems described by differentiable forward models. In Section 3.4 we showcase the effectiveness of our DeepGEM approach on the challenging problem of blind seismic tomography, where we significantly outperform methods used in seismology. We also demonstrate the generality of DeepGEM by applying it to blind deconvolution in Section 3.5.

3.2 Background and related work

The joint optimization of a forward model with source recovery is a very challenging ill-posed problem, leading to many possible solutions that are hard to disambiguate. For example, in blind deconvolution the blurry image observed can be equivalently explained by a sharp source image convolved with an extended kernel or a blurry source image convolved with a impulse kernel; to prefer one solution over another,

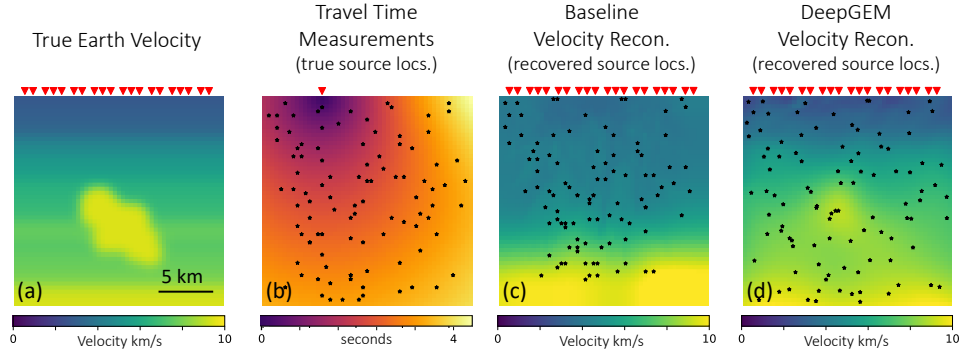


Figure 3.1: **DeepGEM applied to blind seismic tomography.** (a) A simulated cross section of the Earth’s interior (velocity structure), along with the locations of receivers on the surface (red triangles) that collect measurements. (b) The time it takes for a wave traveling from a source below the surface to reach the specified receiver is visualized for each location in the region of interest. The overlaid dots represent the true locations of simulated earthquakes and indicate the measured travel times that constrain optimization. (c) The subsurface velocity reconstruction obtained using a baseline approach optimized with the help of a seismologist. Note that the bright anomaly is missing from this reconstruction. Overlaid dots represent the inferred earthquake locations. (d) DeepGEM reconstructed subsurface velocity and inferred earthquake locations. Note that DeepGEM is able to accurately recover the gradient of the velocity field as well as partially recover the central anomaly.

additional information, such as image priors, must be considered. Previous work on inversion in poorly characterized systems focused on limited contexts, such as spatially-invariant blind deconvolution [7]–[10] and CT with a simple rotational error [11]. These methods tend to be highly specialized to each application domain, and cannot be easily generalized. In contrast, we propose a flexible model-based Bayesian framework that can be applied across multiple differentiable blind inversion problems.

Model-based Bayesian inversion

In model-based inversion, unobserved sources x and observed measurements y are related through a forward model: $y = f(x)$. When the parameters of the underlying model are unknown, one can parameterize the forward model as $f_\theta(x)$ and then solve for the true model parameters θ^* . A common approach is to solve a *maximum a posteriori* (MAP) objective: either $\text{MAP}_{\theta,x}$ or MAP_θ .

$\text{MAP}_{\theta,x}$ solves for the optimal point estimate of the pair $\{\hat{\theta}, \hat{x}\}$ that maximizes a

joint objective:

$$\{\hat{\theta}, \hat{x}\} = \arg \max_{\theta, x} [\log p(\theta, x|y)] = \arg \max_{\theta, x} [\log p(y|\theta, x) + \log p(\theta) + \log p(x)] . \quad (3.1)$$

In practice, formal probabilistic definitions of $p(\theta)$ and $p(x)$ are often unknown and replaced with regularization terms (e.g., total variation, sparsity) [7], [8]. Although $\text{MAP}_{\theta, x}$ provides a straightforward approach to solve for θ , the energy landscape of Eq. 3.1 is typically poorly behaved due to the ill-posed nature of the problem [8]. As a result, optimization is likely to get stuck in a (bad) local minimum.

MAP $_{\theta}$ attempts to smooth the energy landscape by reducing the number of parameters that must be optimized. This is done by solving for parameters of the forward model, θ , that perform best under the full volume of possible x interpretations:

$$\hat{\theta} = \arg \max_{\theta} [\log p(\theta|y)] = \arg \max_{\theta} \left[\log \left(\int_x p(y|\theta, x)p(x)dx \right) + \log p(\theta) \right] . \quad (3.2)$$

Since this marginalization integral is often intractable, Expectation-Maximization (EM) algorithms have long been used for solving MAP_{θ} efficiently [12]. EM is an iterative algorithm that alternates between: “E”-Step) calculating the posterior of x conditioned on the current estimated forward model parameters $\theta^{(t-1)}$; and M-Step) updating θ to maximize the expected value of the log likelihood:

$$\theta^{(t)} = \arg \max_{\theta} [\mathbb{E}_{x \sim p(x|y, \theta^{(t-1)})} [\log p(y|\theta, x)] + \log p(\theta)] . \quad (3.3)$$

The advantage of MAP_{θ} over $\text{MAP}_{\theta, x}$ optimization for the blind deconvolution problem was described in [8] and its success demonstrated via EM optimization in [7]. However, it is important to note that evaluating the expectation in of Eq. (3.3) over complex distributions is often intractable. For instance, the authors of [7] were forced to restrict the posterior distribution to a Gaussian distribution. Alternatively, stochastic EM methods [13], [14] bypasses the need to evaluate the expectation directly, approximating it by sampling the distribution. In this chapter, we solve MAP_{θ} using complex distributions parameterized by deep neural networks.

Forward model parameterization Model-based inversion requires that the parametric form of $f_{\theta}(x)$ is well matched with the true forward model, which is not always known. Alternatively, neural networks can be used to approximate the forward model, where θ represents the network weights. This setup is flexible, in that it

can be used to approximate arbitrarily complex forward models, with the downside that it is often not interpretable when the parameters have no physical meaning. In this work, we develop and make use of interpretable, physically-motivated deep neural networks to parameterize $f_\theta(\cdot)$ for the problems of blind seismic tomography and blind deconvolution.

Deep variational distributions

We are interested in solving blind inverse problems using the EM algorithm to optimize the MAP_θ objective. Optimizing in this fashion requires the use of the posterior distribution $p(x|y, \theta^{(t-1)})$ in evaluating Eq. 3.3. As inverse problems are often ill-posed, we expect that the posterior distribution of source x conditioned on the forward model parameters θ is likely to be complex and perhaps even multi-modal. Therefore, it is important to be able to parameterize a flexible family of distributions to best estimate this conditional posterior.

Deep Probabilistic Imaging (DPI) [15] uses a normalizing flow-based generative model, $G_\phi(\cdot)$, to solve for the uncertainty of an unknown source x given a fixed forward model $f(x)$ and measurements y . DPI solves the variational objective:

$$\hat{\phi} = \arg \min_{\phi} \text{KL}(q_\phi(x) || p(x|y)) = \arg \min_{\phi} E_{x \sim q_\phi(x)} [-\log p(y|x) - \log p(x) + \log q_\phi(x)], \quad (3.4)$$

where $q_\phi(x)$ is the implicit distribution defined by $G_\phi(z)$ for $z \in \mathcal{R}^{|x|} \sim \mathcal{N}(0, 1)^1$, and $\log p(y|x) \propto \|y - f(x)\|_2 + c$ when there exists *i.i.d.* Gaussian noise on the measurements, y . After inference, $q_\phi(x)$ can be efficiently sampled by evaluating $G_\phi(z)$ for a Gaussian sample z .

The DPI variational objective is equivalent to the Variational Autoencoder [16], [17] objective, except with a fixed decoder, $f(x)$. In practice, vanilla VAEs constrain the posterior to be a Gaussian distribution, relying on the reparameterization trick for tractable optimization. Alternatively, DPI uses flow-based networks to efficiently sample and directly compute $q_\phi(x)$ [18]–[20]. DPI’s use of a flow-based network allows for complicated and multi-modal posterior distributions constrained only by the space of possible invertible network architectures. Our proposed Deep-GEM approach utilizes similar tools to model flexible distributions over x , while simultaneously learning the forward model parameters θ .

¹ $|x|$ is the dimension of x

3.3 Methods

We propose a deep variational EM approach (DeepGEM) that optimizes the MAP_θ objective in Eq. 3.2 to recover the parameters of a forward model $f_\theta(x)$ using only a single static set of measurements y . Once learned, the updated forward model can then be used to estimate the posterior distribution of the unknown source, x . DeepGEM iterates between two stages that are inspired by the standard EM algorithm: (1) an ‘‘E’’-step that learns a variational distribution, $q_{\phi^{(t)}}(x)$, to approximate the posterior distribution of x given the current forward model parameters $\theta^{(t-1)}$, and (2) an M-step (refer to Eq. 3.3) that solves for $\theta^{(t)}$ that maximizes the expected value of the log likelihood function of θ , with respect to the posterior distribution $q_{\phi^{(t)}}(x)$ estimated in the prior ‘‘E’’-step. Each step is alternated and solved to convergence. Since this is an EM algorithm (up to the variational approximation), our method inherits properties of EM, including convergence to a local minimum of Eq. 3.2.

‘‘Expectation’’ step (‘‘E’’-step)

The goal of DeepGEM’s ‘‘E’’-step is to solve for the posterior distribution $p(x|y, \theta^{(t-1)})$ that facilitates optimizing Eq. 3.3. Because this posterior distribution can be very complex, and even multi-modal, we propose to use a flexible variational approach to learn the parameters ϕ of an approximate posterior distribution $q_\phi(x)$. The variational distribution $q_\phi(x)$ can then be used to evaluate Eq. 3.3 via efficient sampling.

Using DPI we solve for a flexible variational distribution, $q_\phi(x)$ that well approximates the posterior distribution $p(x|y, \theta^{(t-1)})$. DPI uses a normalizing flow network, $G_\phi(z)$, with input $z \in \mathbb{R}^{|x|}$, where $x = G_\phi(z) \sim q_\phi(x)$ when $z \sim \mathcal{N}(0, \mathbb{1})$. Normalizing flow networks allow for exact computation of the log-likelihood $\log q_\phi(x)$ needed to solve

$$\begin{aligned} \phi^{(t)} &= \arg \min_{\phi} \text{KL}(q_\phi(x) || p(x|y, \theta^{(t-1)})) \\ &\approx \arg \min_{\phi} \frac{1}{N} \sum_{n=1}^N [-\log p(y|\theta^{(t-1)}, x_n) - \log p(x_n) + \log q_\phi(x_n)] \\ &\text{for } x_n = G_\phi(z_n), \quad z_n \sim \mathcal{N}(0, \mathbb{1}), \end{aligned} \tag{3.5}$$

(as derived from Eq. 5.6) for a batch size of N where $\log p(x)$ is a prior on the source and $\log p(y|x, \theta^{(t)})$ is the data likelihood. When assuming the measurements y experience *i.i.d* additive Gaussian noise with standard deviation σ_y , $\log p(y|\theta^{(t)}, x_n) = \frac{1}{2\sigma_y^2} \|y - f_{\theta^{(t)}}(x_n)\|^2 + c$.

Maximization step (M-step)

The goal of DeepGEM’s M-step is to use the parameterized approximate posterior distribution, $q_{\phi^{(t)}}(x)$, from the prior “E”-step to update θ , the parameters of the unknown forward model $f_{\theta}(\cdot)$. This is achieved by sampling from the learned normalizing flow network, $G_{\phi^{(t)}}(\cdot)$, to stochastically solve Eq. 3.3 :

$$\theta^{(t)} \approx \arg \max_{\theta} \left[\frac{1}{N} \sum_{n=1}^N [\log p(y|\theta, x_n)] + \log p(\theta) \right] \quad \text{for } x_n = G_{\phi^{(t)}}(z_n), z_n \sim \mathcal{N}(0, \mathbb{I}), \quad (3.6)$$

where $p(\theta)$ is a prior on the forward model. This prior can be used to encourage the forward model parameters to remain close to an initial model $\tilde{\theta}$ by defining $\log p(\theta) \propto \|\theta - \tilde{\theta}\|_2 + c$.

3.4 Case study: blind seismic tomography

Two fundamental seismic inverse problems are spatially localizing an earthquake’s hypocenter (also referred to as source localization) and tomographic reconstruction of the Earth’s subsurface [5]. These problems are interconnected: source localization relies on knowing how fast waves propagate through different regions of the Earth’s interior, referred to as the Earth’s *velocity* structure. However, in standard seismological practice, due to difficulties in solving these problems jointly, they are generally treated separately: source localization is performed initially using oversimplified velocity models [5], and then the tomography problem is performed by taking those best-fitting hypocenters as ground truth [21]. This approach typically results in the need to over-regularize the inverse problem by smoothing out high-frequency information [22], and can only be improved by carefully incorporating other forms of information such as waveform-derived quantities [23]–[26] or by performing costly experiments such as controlled explosions. In contrast, we demonstrate our DeepGEM approach on blind seismic tomography, solving for the subsurface velocity (parameterized by θ) when the source hypocenters, x , are unknown. Measurements, y , used to constrain the inverse problem are the time it takes for the first wave to propagate from its source to a receiver on the Earth’s surface, referred to as a *travel time* measurement (refer to Fig. 3.1).

Seismic tomography background

Physics of earthquake source localization: The earthquake source location, also called the hypocenter, is the location where the earthquake nucleates [5]. The source location can be triangulated using travel times from multiple receivers. However,

when there are very few receivers (< 3 in 3D, < 2 in 2D), the source localization is ill-posed and there exists multiple equally optimal solutions.

Physics of travel time tomography: Travel time tomography is a method for reconstructing the velocity structure using the absolute arrival times of earthquake waves from the earthquake to the receiver [5]. Given perfect knowledge of the earthquake locations x and receivers r at every position in the ground, the exact velocity can be computed by solving the Eikonal equation $V(r) = \|\nabla_r T(x, r)\|_2^{-1}$, where $T(x, r)$ is the travel time from an earthquake to a receiver. There exists an analytical solution to the Eikonal equation; however, seismologists often use simplifications, such as straight ray tomography, for efficiency [22].

Deep learning for travel time tomography: EikoNet [27] implicitly solves the Eikonal equation [28] by learning a mapping from a source-receiver pair (x, r) to its associated travel time, learning θ such that $f_\theta(x, r) \approx T(x, r)$ where $T(x, r)$ is the true travel time. The velocity structure can then be extracted from the learned EikoNet by solving the Eikonal equation through automatic differentiation. In particular, the computed velocity $V(s)$ at location s is $\|\nabla_s f_\theta(x, s)\|_2^{-1}$. Note that this computed velocity depends on the source location x . Ideally, the velocity should be invariant to the source location; however, in practice, this is only true when EikoNet is trained with densely-sampled (x, r) pairs. Additionally, the travel time from a source to a receiver, $T(x, r)$, should be identical to the travel time from a receiver to a source, $T(r, x)$, but remains unconstrained by EikoNet.

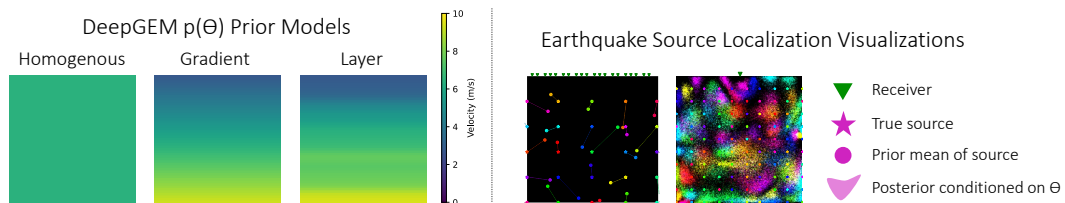


Figure 3.2: **Image priors and source posterior visualization.** (Left) A visualization of the homogeneous, gradient, and layer models used in $p(\theta)$'s $\mathcal{L}_\theta$ term. (Right) Visualizations used to describe source configurations and earthquake source posteriors. (a) Visualizations of the true and initialized source locations are plotted as stars (true) connected to circles, which indicate the expected source locations according to $p(x)$. Note that the expected source locations deviate significantly from the true locations. (b) Visualizations of the learned posterior distribution, $q_\phi(x|y, \theta)$, for each source are plotted as colored histograms and overlaid with stars (of the same color) indicating the true source location.

DeepGEM setup for blind seismic tomography

For blind seismic tomography, we parameterize the forward model using a modification of EikoNet, $f_\theta(x, r)$, with unknown source location x and known receiver location r as inputs and $y \approx T(x, r)$, the absolute travel time between x and r , as output. In order to solve Eqs. 3.5 and 3.6 for an updated forward model, we must define priors $p(x)$ and $p(\theta)$. The prior over source locations, $p(x)$, is often well defined, typically a Gaussian distribution $\mathcal{N}(\bar{x}, \sigma_x)$ with a standard deviation of $\sigma_x \sim 2$ km and mean $\bar{x}$.

We construct a prior over the forward model, $p(\theta)$, that encourages EikoNet to learn a velocity close to $\tilde{V}(s)$. Additionally, as discussed above, there are constraints specific to seismic tomography that are not explicitly enforced through EikoNet’s architecture: (1) velocity reconstruction invariance with respect to the source location, and (2) travel time symmetry between sources and receivers.

We augment the prior $p(\theta)$ to include these constraints, implemented as soft constraints:

$$\log p(\theta) = \lambda_\theta \underbrace{\sum_{\substack{r \in \mathcal{R}, \\ s \in \mathcal{S}}} \|\tilde{V}(s) - V_r(s)\|_2}_{\mathcal{L}_\theta} + \lambda_V \underbrace{\sum_{\substack{r_i, r_j \in \mathcal{R}, \\ s \in \mathcal{S}}} \|V_{r_i}(s) - V_{r_j}(s)\|_2}_{\mathcal{L}_V} + \lambda_T \underbrace{\sum_{\substack{r \in \mathcal{R}, \\ s \in \mathcal{S}}} \|T(s, r) - T(r, s)\|_2}_{\mathcal{L}_T}. \quad (3.7)$$

The velocity constraint is represented through $\mathcal{L}_V$ and travel time constraint through $\mathcal{L}_T$, with corresponding hyperparameters λ_V and λ_T . $\mathcal{S}$ is a set of points sampled uniformly from the region of interest, $\mathcal{R}$ is the set of all receiver locations, and $V_r(s) = \|\nabla_s f_\theta(r, s)\|_2^{-1}$.

Implementation details: In our experiments we define a realistic prior over unknown source locations as $p(x) = \mathcal{N}(\bar{x}, \sigma_x)$, where $\bar{x} \sim \mathcal{N}(x, \sigma_x)$ and $\sigma_x = 2$ km. We assume the measurements y are sampled from a Gaussian distribution with mean $\tilde{y}$ – true travel times computed using the package `eikonal_fm` [28], [29] – and a realistic standard deviation of $\sigma_y = 0.2$ seconds. To simulate real world passive tomography, we assume receivers are located only along the surface (the top edge of the 2D image) of the region of interest, which is 20 km $\times$ 20 km, and sources can be anywhere within the region of interest. Before DeepGEM optimization, we first initialize θ using maximum likelihood estimation assuming $\bar{x}$ is the true source location.

The posterior distribution of the source locations, x , is estimated using a Real-NVP network $G_\phi(\cdot)$ with 16 affine coupling layers. An updated EikoNet (described in

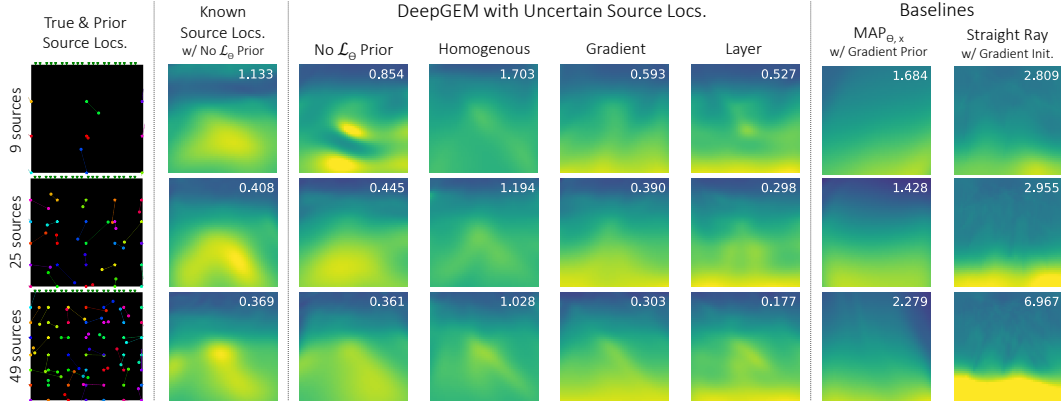


Figure 3.3: DeepGEM reconstructions significantly outperform baselines, and improve with more sources and better $\mathcal{L}_\theta$ prior models. Each row corresponds to the reconstructions obtained from a single noisy observation of travel times, where the true Earth velocity is shown in Fig. 3.1(a). Results shown are simulated using 20 surface receivers and a varying number of sources (9, 25, and 49) in a uniform grid. Columns 3-6 show DeepGEM results obtained using different $\mathcal{L}_\theta$ priors. Note that results improve as the $\mathcal{L}_\theta$ prior becomes closer to the true velocity structure, and as the number of sources increases. As a reference, column 2 shows the velocity reconstruction obtained using DeepGEM under fixed, perfectly known source locations. Columns 7-8 show results obtained by the baseline approaches. The velocity reconstruction MSE is included in the top right of each reconstruction. DeepGEM substantially outperforms both straight ray and MAP $_{\theta,x}$ baselines. Note that DeepGEM also sometimes outperforms the second column, where the source locations are known; this is due to the fact that DeepGEM is sometimes able to absorb measurement error due to its overparameterization. See the supplemental material for more examples.

the supplemental material) has been modified to parameterize $f_\theta(x)$. This EikoNet is pretrained with samples from the prior $p(x)$ as input and the simulated travel time measurements as output. We use Adam as the optimizer [30] with a batch size of 32 and an E-step learning rate of 1e-3 and M-step learning rate of 5e-5. Hyperparameters $(\lambda_T, \lambda_V, \lambda_\theta)$ were empirically chosen by inspecting the loss on a grid search over hyperparameters. Results presented have been run with 10 EM iterations, each with 800 “E”-step epochs and 2000 M-step epochs. Each DeepGEM model takes ~ 6 hours on a NVIDIA Quadro RTX 5000.

Results

Comparison to Baseline Approaches: We compare results from DeepGEM to results obtained using a baseline run by a seismologist. This iterative baseline

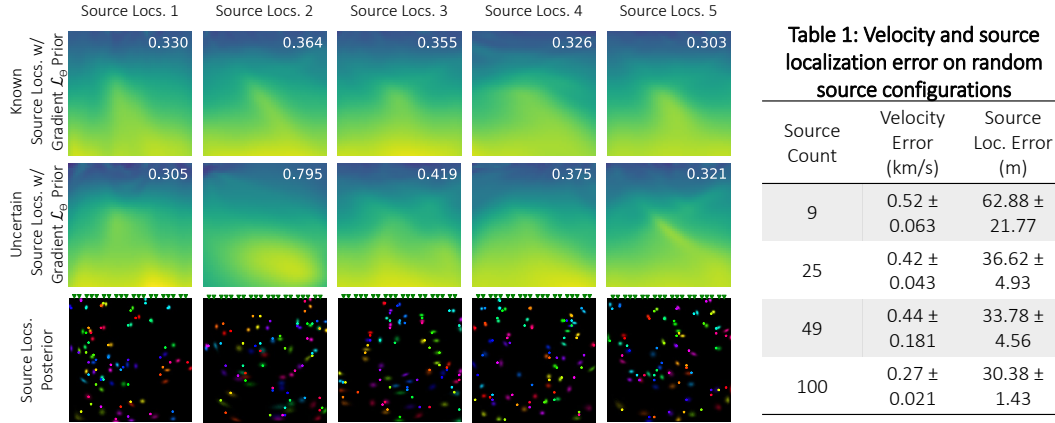


Figure 3.4: **DeepGEM consistently recovers prominent features across various earthquake source configurations.** Row 2 shows DeepGEM results obtained using different random source configurations, where the true Earth velocity is shown in Fig. 3.1(a). As a reference, row 1 shows the velocity reconstruction obtained using DeepGEM under fixed, perfectly known source locations. As can be seen by the table, both velocity and localization error decrease with an increasing number of sources in the region of interest. Each mean and standard deviation is computed using 5 random realizations of true source configurations.

alternates between source inversion and straight ray tomography, and is the standard approach used for blind tomography. Further detail on this baseline is provided in the supplemental material. The gradient model shown in Fig. 3.2 is used to perform the initial source inversion for this baseline; nonetheless, we find that the solution quickly diverges. Therefore, we rely on the expertise of the domain expert to decide when to terminate the optimization. As seen in Fig. 3.3, DeepGEM consistently outperforms this human-in-the-loop baseline across all source configurations. In Fig. 3.3 we also compare with a $\text{MAP}_{\theta,x}$ solution. $\text{MAP}_{\theta,x}$ is consistently outperformed by the DeepGEM MAP_{θ} approach across all source configurations. Additional results highlighting the sensitivity of $\text{MAP}_{\theta,x}$ to hyperparameters and initialization are shown in the supplemental material.

Sensitivity to $\mathcal{L}_{\theta}$ Prior Choice: We evaluate DeepGEM’s recovery of the true velocity structure shown in Fig. 3.1 using one of three different $\mathcal{L}_{\theta}$ priors shown in Fig. 3.2: homogeneous, gradient, and layer, as well as $\mathcal{L}_{\theta} = 0$. The homogeneous model takes on value of 6.419 km/s, the average velocity value of the true velocity structure. The gradient captures the increasing velocity as depth increases, and the layer model represents the true model without the added anomaly. As shown in

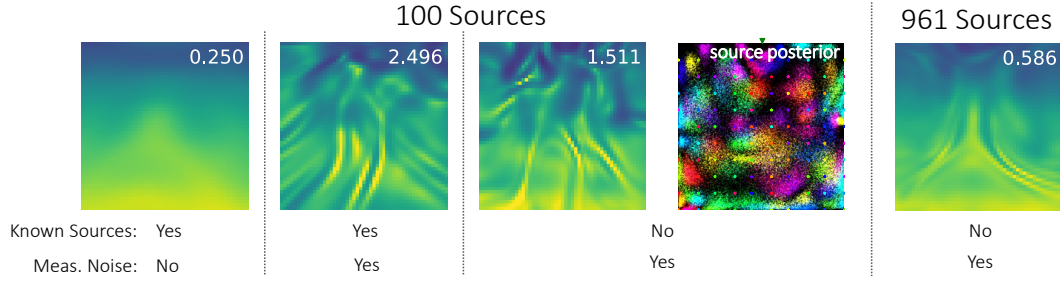


Figure 3.5: **DeepGEM recovery with a single receiver.** Velocity reconstructions shown in columns 3 and 5 demonstrate that DeepGEM is able to learn some of the true velocity features (see Fig. 3.1(a)), even when limited to measurements from a single receiver. However, these reconstructions show clear signs of overfitting to the measurement data. This is demonstrated by observing that the DeepGEM reconstruction with perfectly known source locations is significantly better with a single receiver when no measurement error is included on the measurements (comparing columns 1 and 2). Note that the learned posterior in column 4 appears to be non-Gaussian, but not multimodal.

Fig. 3.3, the mean squared error tends to decrease with the gradient and layer $\mathcal{L}_\theta$ prior models, which are closer to the true velocity structure.

Sensitivity to Source Configuration: We evaluate results with sources that are both uniformly and randomly spaced throughout the region of interest. Improved performance is expected when the number of sources is increased and/or when sources are well distributed. To better understand DeepGEM’s performance we introduce an ablation, where the velocity structure is learned by training $f_\theta(\cdot)$ with access to perfect source locations x (i.e., $p(x)$ is a delta function). As is shown in Fig 3.3, even when training with true source locations, the anomaly is not well resolved until ~ 49 sources.

Fig. 3.3 shows one realization of DeepGEM results obtained from different counts of uniformly spaced sources. As expected, the MSE between the reconstructed and true velocity structure tends to decrease as the number of sources increases. Fig. 3.4 shows results obtained using five randomly generated source configurations, with 49 sources each. These results demonstrate that, although the reconstructed velocity structure is somewhat sensitive to the underlying source configuration, the primary features of the true velocity can still be recovered in all cases. Velocity and source localization error obtained for different random configurations are shown in the accompanying table.

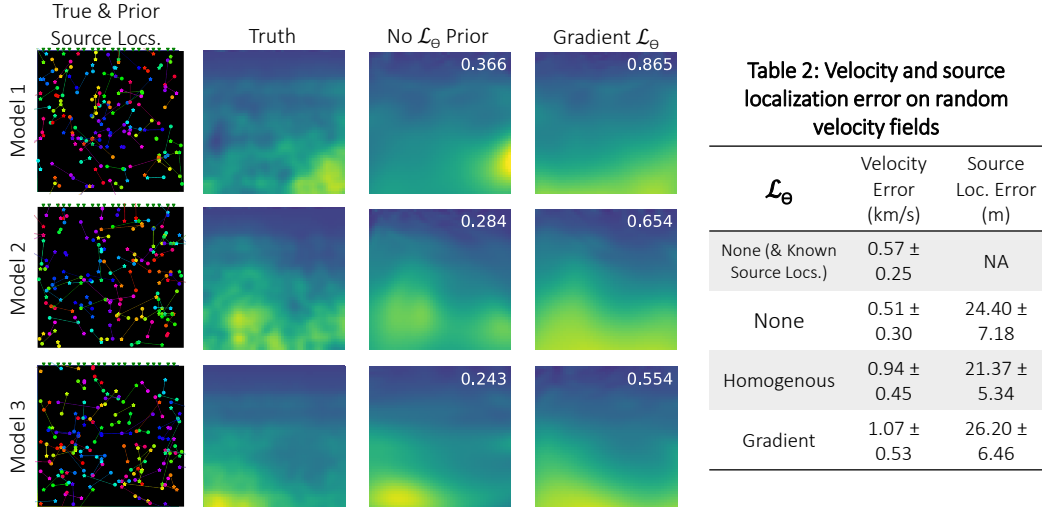


Figure 3.6: **Performance of DeepGEM recovery on random velocity fields.** Ten random velocity fields were drawn from a GRF-based distribution and used to simulate travel time measurements with 20 receivers and 100 randomly placed sources. (left) Reconstructions obtained for 3 of these configurations are shown. (right) A table lists the mean and standard deviation of velocity and source error obtained across the ten models, each recovered using different $\mathcal{L}_\theta$ priors.

Sensitivity to Number of Receivers: In the case of a single receiver, there exists an entire ring of source locations that result in the same travel time measurement. Fig. 3.5 contains results from this challenging one-receiver setting using DeepGEM with/without measurement noise and with/without known source locations. Since there is only one receiver, the velocity model is able to easily overfit. However, perhaps surprisingly, artifacts are more severe when source locations are perfectly constrained. These artifacts are caused by the velocity model overfitting to noise in the travel time measurements, and are substantially reduced when noise-free travel time measurements are used. Note that the recovered source location posterior $q_\phi(x)$ obtained by the “E”-step is non-Gaussian.

Sensitivity to Velocity Structure: In Fig. 3.6, DeepGEM is tested on randomly generated velocity fields, each generated from a Gaussian random field (GRF) described in the supplemental material. These results show that DeepGEM works well at recovering the primary features across a variety of different velocity structures. As compared to when $\mathcal{L}_\theta = 0$, the $\mathcal{L}_\theta$ gradient prior biases the reconstruction towards smoother velocity structures. The accompanying table contains error statistics for ten different randomly generated velocity fields.

3.5 Case study: blind deconvolution

We apply DeepGEM to the problem of blind deconvolution to further demonstrate the generality of our approach. Blind deconvolution is a classic ill-posed imaging problem that aims to reconstruct a sharp image from a blurry image with an unknown PSF [7]–[9], [31]–[36]. Blurry images, caused by handheld camera shake, can be modeled using a single spatially-invariant blur kernel:

$$y = x * k_\theta + \varepsilon \quad \text{for } \varepsilon \sim \mathcal{N}(0, \sigma), \quad (3.8)$$

where y is the blurry image, $*$ represents a 2D convolution, x is the true sharp image, k_θ is the spatially invariant blur kernel, and ε is additive Gaussian noise.

DeepGEM setup for blind deconvolution

For blind deconvolution, we parameterize the forward model $f_\theta(\cdot)$ using a deep network consisting of multiple convolution layers without non-linear activation, as proposed in [37]. Multiple convolutional layers without activation simply over-parameterizes a linear blur kernel, which has been empirically shown to produce multiple good minima that are easier to converge to. To ensure the blur kernel is non-negative and volume preserving, we use a Softmax layer and normalize the kernel. For an n layer network with weights θ_i for $i = 1, \dots, n$, the resulting parameterized forward model is:

$$f_\theta(x) = x * \hat{k}_\theta = x * \left[\frac{\text{Softmax}(\theta_1 * \theta_2 * \dots * \theta_n)}{\|\text{Softmax}(\theta_1 * \theta_2 * \dots * \theta_n)\|_1} \right]. \quad (3.9)$$

Implementation details: We demonstrate DeepGEM using simple Total Variation (TV) regularization in place of $\log p(x)$. We assume a Gaussian prior on the noise on the blurry measurements where $\varepsilon \sim \mathcal{N}(0, 0.01)$ as well as a sparsity prior on the reconstructed kernel through an $\ell_{0.8}$ soft constraint. The posterior distribution of the sharp image, x , is estimated using a Real-NVP network $G_\phi(\cdot)$ with 16 affine coupling layers. We use 5 convolution layers to parameterize k_θ . We use Adam as the optimizer [30] with a batch size of 64 and an E-step learning rate of 5e-4 and M-step learning rate of 1e-4. Hyperparameters, weights used for sparsity and TV priors, were empirically chosen by a grid search over hyperparameters. Results presented have been run with 10 EM iterations, each with 400 “E”-step epochs and 400 M-step epochs, which takes ~ 1 hour on a NVIDIA Tesla V100.

Results

In Fig. 3.7 we show results from DeepGEM on three different Fashion-MNIST [38] images. The Fashion MNIST images are 64×64 pixels in size and each blur kernel

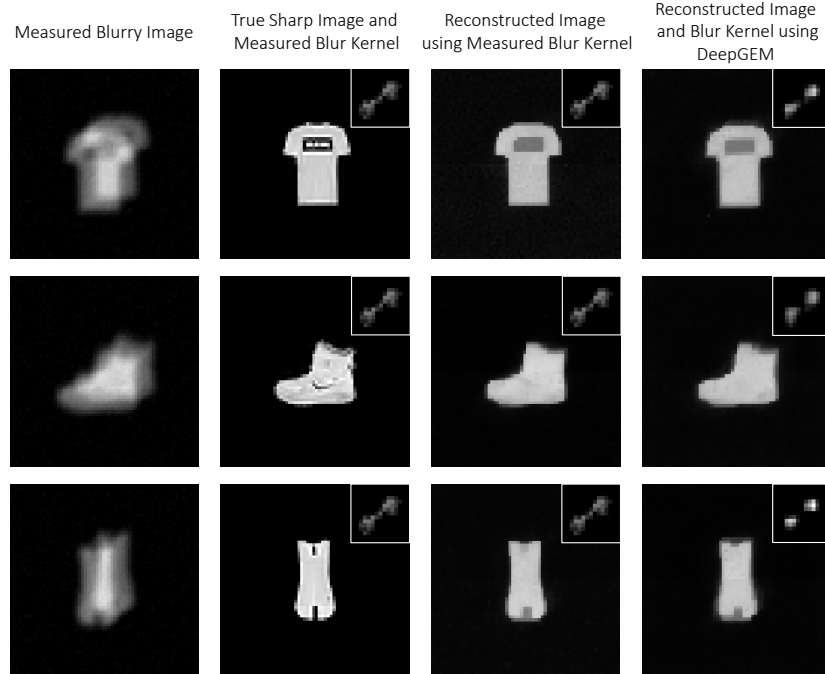


Figure 3.7: **DeepGEM results are comparable to inversion with known blur kernels for simple Fashion MNIST images.** Blurry measured images (column 1) are generated using the true sharp image and blur kernel (shown in column 2). The deconvolved images obtained using the true blur kernel are shown in column 3. Reconstructed deconvolved images and corresponding inferred blur kernel from DeepGEM are shown in column 4.

is contained within 15×15 pixels. We choose to use images from Fashion-MNIST because they are compatible with the TV prior used in place of $\log p(x)$. In contrast, other natural images such as images from [8] (which were originally curated for deconvolution) are not compatible with this handcrafted prior. While this highlights that handcrafted priors such as TV are often insufficient, we choose to leave this to future work. Please refer to the supplemental material for more details about the performance of a TV prior on different types of images.

The blurry images in the first column of Fig. 3.7 exhibit artifacts from the blur kernel in the form of repeated features, caused by the bi-modal blur kernel. The images recovered via DeepGEM are much sharper and are comparable to the images reconstructed with the known blur kernel. Additionally, DeepGEM reconstructs blur kernels (i.e., k_θ) that are similar to the true kernel shape, with two lobes along the same diagonal. Note that even when there is perfect knowledge about the shape of the blur kernel, detailed structures on the images are not recovered due to the TV

prior.

3.6 Case study: blind deconvolution for estimation of earthquake rupture process using empirical Green’s function (EGF)

Whether an earthquake is a quick impulse or a complex, prolonged event with many bursts, can tell us information about the fault behavior, rupture mechanics, and conditions of stress. Scientists use multiple sensors across an entire fault region to identify the earthquake’s rupture process, yielding conclusions about the fault itself. The earthquake record is used to understand an earthquake’s rupture process. This record is the convolution of source radiation, path propagation and site effects, and instrument response. Isolating the source component requires solving an ill-posed inverse problem. Whether the instability of inferred source parameters arises from varying properties of the source, or from approximations we introduce in solving the problem, remains an open question. Such approximations often derive from limited knowledge of the forward problem. The Empirical Green’s function (EGF) approach offers a partial remedy by approximating the forward response of larger events using the records of small events. Indeed, the choice of the *best* small event drastically influences the properties estimated for the larger earthquake. Discriminating variability in source properties from epistemic uncertainties, stemming from the forward problem or other modeling assumptions, requires us to reliably account for, and propagate, any bias or trade-off introduced in the problem. Thus, we use a variant of DeepGEM tailored for the EGF approach to provide reliable and probabilistic estimates of source parameters (here, we consider the source-time function or STF), and their posterior uncertainty, in the time domain. We jointly solve for the best EGF using one or a few small events as *prior* EGF. We demonstrate, with toy models and various applications to mainshocks of Mw ranging from ~ 4 to 6.3, the potential of DeepGEM-EGF to disentangle the variability of the seismic source from biases introduced by modeling assumptions.

DeepGEM setup: solving for the Source-Time Function with the E-step

The expectation step (i.e E-step) aims to approximate the posterior distribution of our target source time function (STF) $\mathbf{m}$ knowing the data $\mathbf{d}$ using a normalizing flow-based generative model as described in Sec. 3.3. We infer the posterior distribution $q_\phi(\mathbf{m})$ using variational inference where the class of normalizing flow-based neural networks defines our variational distribution. Once we have solved for the approximate posterior distribution, $q_\phi(\mathbf{m})$, we can sample from it to solve the

M-step.

We parameterize $q_\phi(\mathbf{m})$ with a normalizing flow F_ϕ such that $q_\phi(\mathbf{m}) = p(F_\phi(z))$. At the i -th E-step, we optimize the weights $\phi^{(i)}$ of our $F_{\phi^{(i)}}$ that best approximates our posterior distribution $p(\mathbf{m}|\mathbf{d}, \theta^{(i-1)})$ through the following objective:

$$\begin{aligned} \phi^{(i)} &= \arg \min_{\phi} \text{KL} (q_\phi(\mathbf{m}) \parallel p(\mathbf{m}|\mathbf{d}, \theta^{(i-1)})) \\ &\approx \arg \min_{\phi} \frac{1}{N} \sum_{n=1}^N [-\log p(\mathbf{d} | \theta^{(i-1)}, \mathbf{m}_n) - \log p(\mathbf{m}_n) + \beta \log q_\phi(\mathbf{m}_n)], \end{aligned} \quad (3.10)$$

for $\mathbf{m}_n = F_\phi(z_n)$, $z_n \sim \mathcal{N}(0, 1)$, a batch size of N , the prior on the STF $\log p(\mathbf{m})$, the data likelihood $\log p(\mathbf{d}|\mathbf{m}, \theta^{i-1})$, and KL denotes the Kullback-Leibler divergence. Note that the approximation comes from evaluation of an expectation using Monte Carlo sampling. An additional hyperparameter β is proposed in DPI to control the entropy of the generative model’s posterior samples. We use a Real-NVP network [39] for F_ϕ with $L=32$ affine coupling layers. We use Adam [40] as the optimizer with a batch size of 1024.

We define a realistic prior over the normalized amplitude of the STF as $p(\mathbf{m}) \sim \mathcal{N}(\overline{\mathbf{m}}, \sigma_m)$. By default, $\overline{\mathbf{m}}$ is a Gaussian-shaped STF with a σ_m of 5% in amplitude. The default Gaussian-shaped $\overline{\mathbf{m}}$ is centered on half of the preset STF length, and has a width equal to 10% of the total preset STF duration. By default, the weight of this prior is small and its nature has a very limited impact on the results. The assumed σ_m has a larger impact, in particular on the posterior standard deviation, as it conditions the exploration of the solution space.

Because the weight of the prior STF is small, we augment $p(\mathbf{m})$ with a few specific constraints. We want the STF to be close to zero on its boundaries, and we enforce sparsity with a ℓ_1 norm. With sparsity regularization, we avoid overestimating the complexity of the STF due to noise and data over-fitting, while allowing for sharp changes in the solution. We also impose a total variation regularization. Default hyperparameters (weights for the specific constraints) were empirically chosen by inspecting the loss and the fit to synthetic tests on a grid search.

DeepGEM setup: optimizing the EGFs for the forward model for the M-step

The maximization step (i.e. M-step) relies on estimates of the approximate posterior $q_\phi(\mathbf{m})$ from the preceding E-step to update θ , the parameters of the unknown forward model, from the initially assumed set of empirical Green’s function(s) (EGF(s)). We define G_θ as the convolution between model parameters $\mathbf{m}$ and a linear kernel k_θ

of parameters θ . The kernel is a deep network consisting in multiple convolution layers without non-linear activation, as proposed in [41] and described in Sec. 3.5:

$$G_\theta(\mathbf{m}) = \mathbf{m} * k_\theta = \mathbf{m} * \frac{\theta_1 * \theta_2 * \dots * \theta_K}{\|\theta_1 * \theta_2 * \dots * \theta_K\|_\infty}, \quad (3.11)$$

for a K -layer network. During the i -th M-step, we optimize the weights $\theta^{(i)}$ to find the *maximum a posteriori* (MAP) of the posterior distribution of parameters θ knowing the data $\mathbf{d}$:

$$\begin{aligned} \theta^{(i)} &= \arg \max_{\theta} \log p(\theta \mid \mathbf{d}) \\ &\approx \arg \max_{\theta} \left[\frac{1}{N} \sum_{n=1}^N [\log p(\mathbf{d} \mid \theta, \mathbf{m}_n)] + \log p(\theta) \right], \end{aligned} \quad (3.12)$$

for $\mathbf{m}_n = F_{\phi^{(i)}}(\mathbf{z}_n)$, $\mathbf{z}_n \sim \mathcal{N}(0, 1)$. We use $K=7$ convolution layers to parameterize G_θ . We use Adam as the optimizer.

As we already have a good guess for our forward model, we use a prior $\log p(\theta)$ that encourages the forward model to stay close to the initially assumed set of EGF(s). We define $\log p(\theta)$ as a weighted sum of mean absolute error ($\mathcal{L}_\phi$), ℓ_2 -norm, and dynamic time warping norm ($\mathcal{L}_{DTW}$) between the updated and initial EGF(s) k_{θ_0} :

$$\log p(\theta) = -\lambda_\phi \underbrace{\sum_{i=1}^H |k_{\theta_0} - k_{\theta^{(i)}}|}_{\mathcal{L}_\phi} - \lambda_2 \underbrace{\sum_{i=1}^H (k_{\theta_0} - k_{\theta^{(i)}})^2}_{\ell_2} - \lambda_{DTW} \underbrace{DTW_{0.1}(k_{\theta^{(i)}}, k_{\theta_0})}_{\mathcal{L}_{DTW}}, \quad (3.13)$$

with H the length of $k_{\theta^{(i)}}$.

When considering multiple candidate EGFs, we parameterize k_θ as an array of independent deep networks. During the i -th M-step and for each EGF e , weights $\theta_e^{(i)}$ are optimized. We augment $\log p(\theta)$ with a prior $\mathcal{L}_{\text{multi}}$ that encourages the inferred EGFs to converge towards a single EGF $k_{\theta_{\text{best}}^{(i)}}$ that minimizes $\alpha_e^{(i)}$, the misfit between data and predictions:

$$\log p(\theta) = -\mathcal{L}_\phi - \ell_2 - \mathcal{L}_{DTW} - \overbrace{\lambda_{\text{multi}} \sum_e \alpha_e^{(i)} \sum_{i=1}^H (k_{\theta_e^{(i)}} - k_{\theta_{\text{best}}^{(i)}})^2}_{\mathcal{L}_{\text{multi}}}, \quad (3.14)$$

$$\text{with } k_{\theta_{\text{best}}^{(i)}} = \arg \min_e \alpha_e^{(i)} \propto \sum_{i=1}^H (\mathbf{m}^{(i)} * k_{\theta_e^{(i)}} - \mathbf{d})^2. \quad (3.15)$$

Additionally, MSE loss for each EGF can be weighted with a user-defined parameter that can reflect the quality of each *a priori* selected EGF. Weights λ_ϕ , λ_2 , λ_{DTW} and λ_{multi} were empirically chosen with a grid search on synthetic tests.

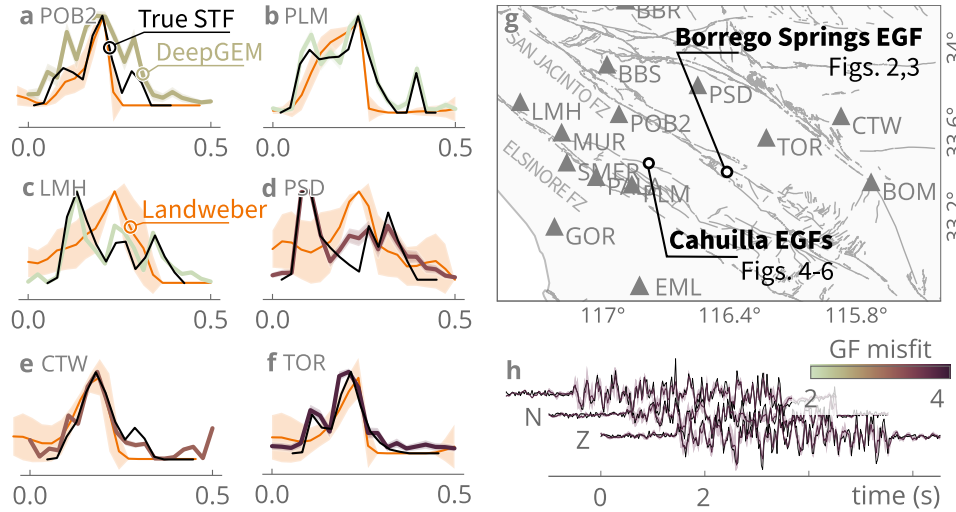


Figure 3.8: Synthetic deconvolution test based on realistic synthetic data setup. (a-f) Inferred (colored) and target (black) source time functions for synthetic tests designed with waveforms (P arrivals) recorded for four neighbor $M \sim 2$ events that occurred during the Cahuilla swarm. The mean of the source time functions is colored from the misfit between inferred and target EGFs; the standard deviation is shaded. (h) Mean inferred (colored) and target (black) EGF at station TOR. (g) Stations locations.

The 2016-2019 Cahuilla swarm, CA, USA

The Cahuilla swarm outlined a complex but well-defined fault structure between the San Jacinto and Elsinore fault zones (Fig. 3.8). [42] produced a seismicity catalog of more than 22,000 events with M_w ranging from 0.7 to 4.4, all of which seemingly affecting the same non-planar fault geometry. The main event (M_w 4.4) was ~ 5 km deep and likely caused drastic changes in the evolution of the swarm by allowing fluids to circulate at shallower depth. We take advantage of this detailed seismic catalog, with many small events, to investigate how the forward model assumptions that we make affect our knowledge of the source parameters. Along with evaluating the robustness of our approach and the effect of an ad-hoc selection of EGFs on the inference of source parameters, we aim to better characterize the source process of this puzzling event.

We use broadband seismic data recorded in Southern California for the 2018 M_w 4.41 mainshock and several $M \sim 2$ -2.5 events that occurred during the Cahuilla swarm. We use stations located within 200 km of the mainshock. We use P or S arrival waveforms, bandpass filtered between 0.1 and 40 Hz and decimated to 40 Hz. For S wave arrivals, our time window ends 13 sec after S arrivals. We select 1 to 4 initial EGFs, with a $2 < M_w < 2.5$, that are selected using the cross-correlation of

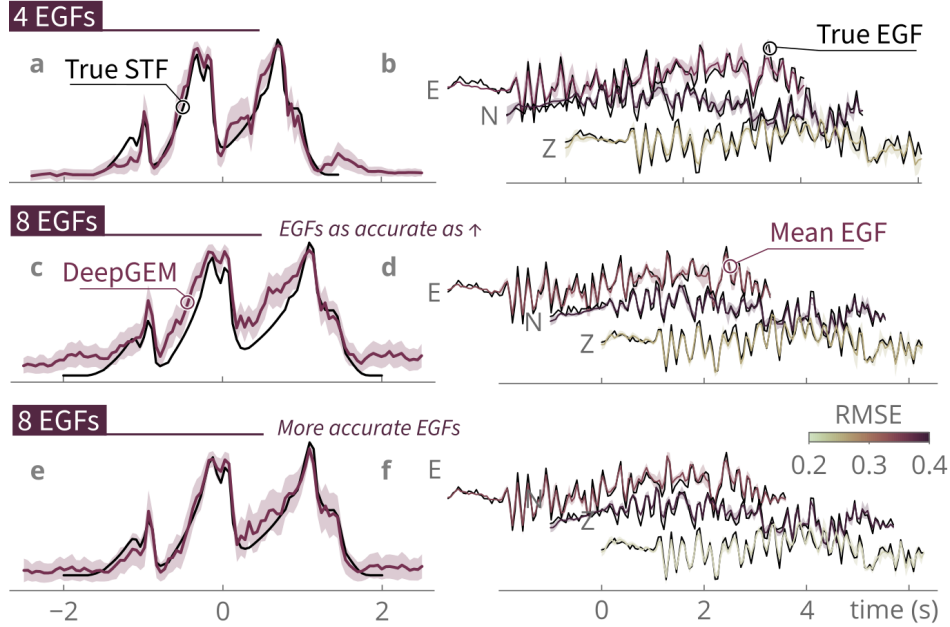


Figure 3.9: **Improved reconstruction quality as EGF quality increases.** Mean inferred (colored) and target (black) Source-Time Functions (STFs) and Empirical Green's Functions for toy models designed with waveforms (P arrivals only) recorded for four (a,b) or eight (c-f) neighboring $M \sim 2$ events that occurred during the Cahuilla swarm. In (e,f) the additional four waveforms are closer to the target than in (c,d). Standard deviation (1σ) is shaded. EGFs are colored by the root mean squared error (RMSE) between data and average predictions normalized for each component.

their waveforms with the mainshock waveforms at each station. The chosen EGFs are at a distance of less than 1 km from the mainshock, after the relocated catalog from [42]. The number of assumed EGFs depends on their SNR that should be larger than 2. This amounts to a total of 69 events.

Synthetic tests with multiple EGFs

We first design synthetic tests that differ from the ones previously presented: the target EGF, or the one convolved with synthetic STFs to obtain synthetic mainshock waveforms, is now defined as a random weighted sum of up to 4 prior EGFs. The synthetic STF is multi-peaked (3 pulses), with a duration of 0.35 to 5 sec to simulate a M4 event. For our measurements, we use waveforms that include the P arrivals.

The results of these tests (Fig. 3.8) show that target STFs are well estimated, with some misfit in amplitude for some peaks. Misfit between data and predictions is generally low. In general, the inferred EGF which shows the lower data misfit

is the one whose prior was the closest to the target EGF, which is an expected behavior. However, EGFs with an incorrect prior also show a relatively good data fit. Target EGFs are usually well within posterior uncertainties of mean posterior EGFs (Fig. 3.8h). From our comparisons with results of the Landweber approach, we can identify two potentially systematic artifacts in baseline STFs: more than three peaks cannot be imaged, and STFs are often inaccurately imaged as highly asymmetric, with an incorrect sharp change in amplitude.

For station TOR, we perform a few other tests to analyze the impact of the number of a priori assumed EGFs on our inference. Target STF is more complex, with 7 pulses, and with a duration of 4 sec. Waveforms are decimated to 20 Hz. We use 4 prior EGFs in the control test (Fig. 3.9a,b) and use 4 additional prior EGFs in two other tests. In one case (Fig. 3.9c,d), the additional four waveforms are random linear combinations of the initial four prior EGFs: in other words, the additional waveforms do not contain more information than the initial four. In the last case, the additional four waveforms are closer to the target EGF: they are equal to the target EGF to which 10% of the random linear combination is added (Fig. 3.9e,f). As expected, if additional EGFs do not add information content, the inferred model is not improved. In contrast, adding informative EGFs (i.e. more accurate EGFs) improves the quality of the inferred STF. Adding information only very slightly improves the fit to the data. It is therefore more efficient to assume a few well selected prior EGFs than multiple poorly constrained priors. These tests further confirm the robustness of DeepGEM-EGF, and its capacity to use good information.

Results

We first estimate STFs from the set of prior EGFs we selected based on cross-correlation. Then, we analyze the effect of the choice of prior EGFs on our inference. We compare all results to Landweber STFs. STFs inferred with all approaches and priors agree well with our preferred results (Fig. 3.10). Apparent STFs duration varies between 0.2 (station LMH, Fig. 3.10d) and 0.5 sec (PSD, Fig. 3.10f). To the N-NE of the Cahuilla swarm (stations BBR, PSD to TOR), apparent STFs are long in duration, with an energetic onset and a long decrease in amplitude. To the South (station EML), apparent STFs are wide and more symmetric in shape. In contrast, STFs are narrow and symmetric to the N-NW (station LMH). The variation in apparent STFs shapes could reflect an along-strike directivity of the Mw 4.4 event towards the northern portion of the fault.

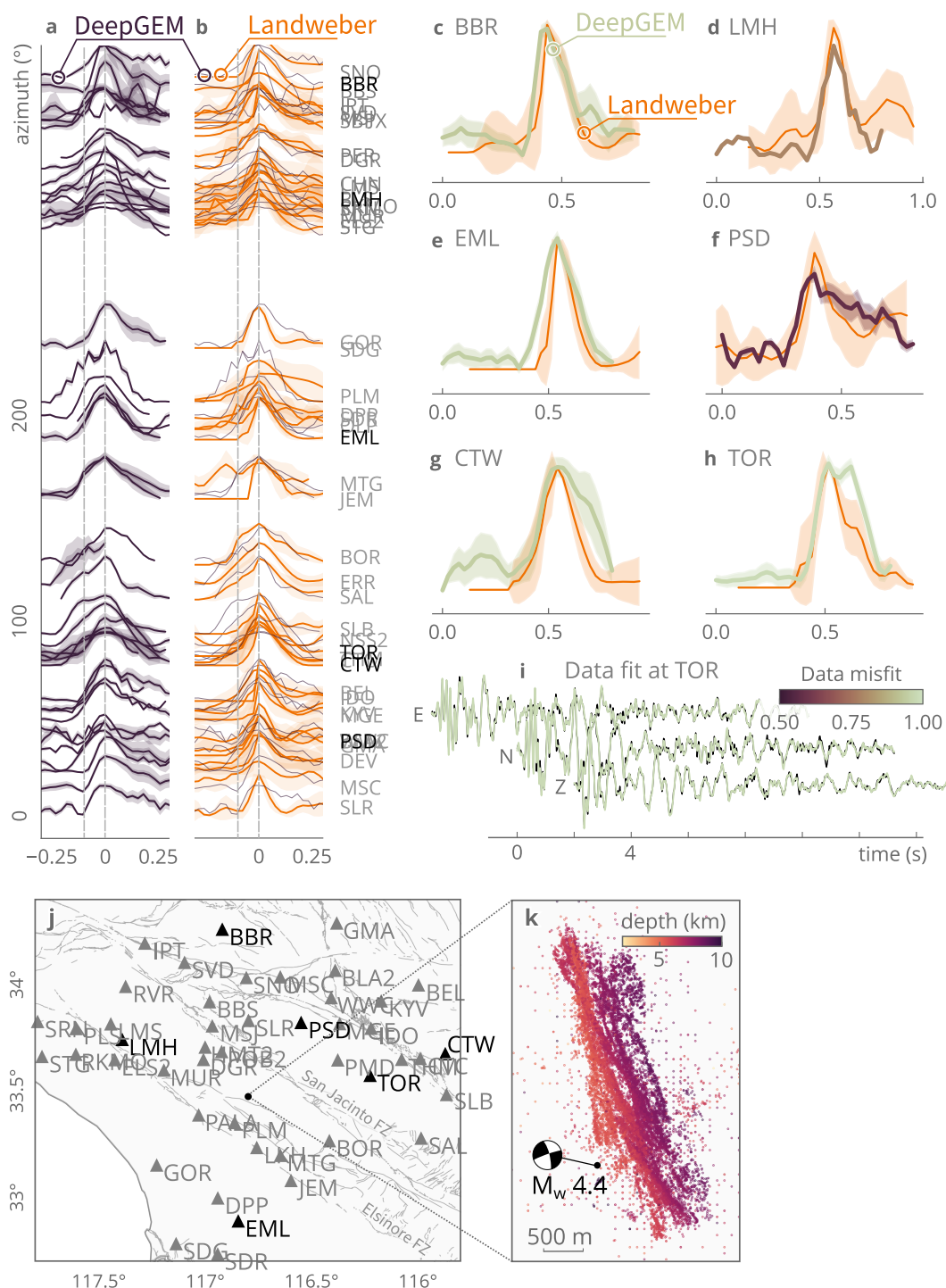


Figure 3.10: Applying DeepGEM-EGF to the Cahuilla swarm. Apparent STFs inferred for the 2018 Mw 4.41 mainshock of the Cahuilla swarm, assuming the same initial EGFs as in Fig. 3.8. (a) Using DeepGEM-EGF or (b) with the approach of [43]. (d-h) Comparison of the results of the two different approaches at a few stations, colored from their data misfit (cross-correlation between data and predictions). Note that σ estimates for the Landweber approach derive from the variability of multiples EGFs and their components, and not from posterior PDFs as for DeepGEM-EGF results. (i) Data fit at station TOR from the DeepGEM-EGF STF. (j) Map with stations locations. (k) Relocated catalog for the swarm, from [42].

STFs inferred with the Landweber approach agree with DeepGEM-EGF estimates to a certain degree. For some stations, estimates are similar (stations GOR or BOR for instance, 220° and 120° respectively). For other stations, baseline results often under-estimate STF durations by introducing a delayed sharp increase in amplitude (see station EML for instance): this corresponds to the artifacts observed with our synthetic tests. While STFs at neighbor stations are identical when estimated with DeepGEM-EGF (JEM and MTG at 170°), they can differ strongly when inferred with the Landweber approach.

We also compare apparent STFs inferred assuming various sets of prior EGFs. This time, we decimate waveforms to 20 Hz to decrease computation time. Our preferred results (Fig. 3.10) have been estimated with priors that minimize cross-correlation, using waveforms with P arrivals. We test 3 other sets of EGFs used as both an initialization and prior: (set 1) using S arrivals; (set 2) minimizing distance; (set 3) minimizing cross-correlation at one station. In set (3), selected EGFs are less than 1 km away from the mainshock and maximize the cross-correlation with the mainshock waveforms at station PLM. This choice of station is based on the quality of the waveforms and the deconvolution with set (2). We use identical hyperparameters for all tests. We expect the STFs inferred with set (3) to be the worst estimates, as EGFs are incorrectly constrained with a single station.

For all sets of EGFs, inferred STFs share similar first-order characteristics: width, shape and azimuthal variations. STFs estimated with set (2) show a good coherence with our preferred results. As expected, STFs estimated with set (3), while consistent between neighbor stations, show many high frequency peaks, which could reflect their relatively poor quality. Small posterior uncertainty on both STFs and EGFs might suggest that over-constraining EGFs to an incorrect prior induced the model to fall into a local minimum.

As for any inverse problem, we cannot conclude on which STF estimates are the closest to reality. However, because of the similarity of the results for neighbor stations, and the stability of DeepGEM-EGF estimates even for various sets of priors, we can affirm that our results are robust. In complement, the slight variability of our estimates made for various sets of priors suggests that DeepGEM-EGF remains sensitive to the prior selection of EGFs.

3.7 Conclusions

In this chapter, I present DeepGEM, a deep probabilistic framework for tackling blind inverse problems through estimation of the forward model. DeepGEM achieves strong performance in the task of joint seismic tomography and earthquake source localization, substantially outperforming standard approaches currently being used in seismology on synthetic data. The proposed framework is flexible and can be applied to different applications that require estimation or fine tuning of forward model parameters. We demonstrate this flexibility by also applying the approach to a simple, but challenging, blind image deconvolution problem. We extend this to solving a seismic deconvolution problem. Our results highlight the benefit of blending physically sound model-based techniques with learning machinery for blind inverse problems.

References

- [1] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.
- [2] Théa Ragon, Angela F. Gao, and Zachary E. Ross, “DeepGEM-EGF: A Bayesian strategy for joint estimates of source-time functions and empirical Green’s functions,” *JGR Solid Earth*, 2025. doi: <https://doi.org/10.1029/2025JB031260>,
- [3] J. A. Fessler, “Model-based image reconstruction for mri,” *IEEE signal processing magazine*, vol. 27, no. 4, pp. 81–89, 2010.
- [4] C. Che, F. Luan, S. Zhao, K. Bala, and I. Gkioulekas, “Inverse transport networks,” *arXiv preprint arXiv:1809.10820*, 2018.
- [5] P. M. Shearer, *Introduction to seismology*. Cambridge university press, 2019.
- [6] S. Cho, Jue Wang, and S. Lee, “Handling outliers in non-blind image deconvolution,” in *2011 International Conference on Computer Vision*, 2011, pp. 495–502.
- [7] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Efficient marginal likelihood optimization in blind deconvolution,” in *CVPR 2011*, 2011, pp. 2657–2664. doi: [10.1109/CVPR.2011.5995308](https://doi.org/10.1109/CVPR.2011.5995308).
- [8] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Understanding and evaluating blind deconvolution algorithms,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 1964–1971.
- [9] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” in *ACM SIGGRAPH 2006*

- Papers*, ser. SIGGRAPH '06, Boston, Massachusetts: Association for Computing Machinery, 2006, pp. 787–794, ISBN: 1595933646. DOI: 10.1145/1179352.1141956. [Online]. Available: <https://doi.org/10.1145/1179352.1141956>.
- [10] S. Cho and S. Lee, “Fast motion deblurring,” in *ACM SIGGRAPH Asia 2009 Papers*, ser. SIGGRAPH Asia '09, Yokohama, Japan: Association for Computing Machinery, 2009, ISBN: 9781605588582. DOI: 10.1145/1661412.1618491. [Online]. Available: <https://doi.org/10.1145/1661412.1618491>.
 - [11] M. Xie, Y. Sun, J. Liu, B. Wohlberg, and U. S. Kamilov, *Joint reconstruction and calibration using regularization by denoising*, 2020. arXiv: 2011.13391 [eess.IV].
 - [12] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
 - [13] O. Cappé and E. Moulines, “On-line expectation-maximization algorithm for latent data models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, Jun. 2009, ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2009.00698.x. [Online]. Available: <http://dx.doi.org/10.1111/j.1467-9868.2009.00698.x>.
 - [14] J. Chen, J. Zhu, Y. W. Teh, and T. Zhang, “Stochastic expectation maximization with variance reduction,” in *NeurIPS*, 2018, pp. 7978–7988.
 - [15] H. Sun and K. L. Bouman, “Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2628–2637.
 - [16] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
 - [17] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *International conference on machine learning*, PMLR, 2014, pp. 1278–1286.
 - [18] L. Dinh, D. Krueger, and Y. Bengio, *Nice: Non-linear independent components estimation*, 2015. arXiv: 1410.8516 [cs.LG].
 - [19] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density estimation using real nvp*, 2017. arXiv: 1605.08803 [cs.LG].
 - [20] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” in *Advances in neural information processing systems*, 2018, pp. 10 215–10 224.

- [21] N. Rawlinson, S. Pozgay, and S. Fishwick, “Seismic tomography: A window into deep earth,” *Physics of the Earth and Planetary Interiors*, vol. 178, no. 3-4, pp. 101–135, 2010.
- [22] M. P. Barmin, M. H. Ritzwoller, and A. L. Levshin, “A fast and reliable method for surface wave tomography,” in *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Surface Waves*. Basel: Birkhäuser Basel, 2001, pp. 1351–1375, ISBN: 978-3-0348-8264-4. DOI: 10.1007/978-3-0348-8264-4_3. [Online]. Available: https://doi.org/10.1007/978-3-0348-8264-4_3.
- [23] A. Fichtner, B. L. Kennett, H. Igel, and H.-P. Bunge, “Theoretical background for continental-and global-scale full-waveform inversion in the time–frequency domain,” *Geophysical Journal International*, vol. 175, no. 2, pp. 665–685, 2008.
- [24] J. Luo and R.-S. Wu, “Seismic envelope inversion: Reduction of local minima and noise resistance,” *Geophysical Prospecting*, vol. 63, no. 3, pp. 597–614, 2015.
- [25] E. Bozdağ, J. Trampert, and J. Tromp, “Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements,” *Geophysical Journal International*, vol. 185, no. 2, pp. 845–870, 2011.
- [26] Y. Luo and G. T. Schuster, “Wave-equation travelttime inversion,” *Geophysics*, vol. 56, no. 5, pp. 645–653, 1991.
- [27] J. D. Smith, K. Azizzadenesheli, and Z. E. Ross, *Eikonet: Solving the eikonal equation with deep neural networks*, 2020. arXiv: 2004.00361 [physics.comp-ph].
- [28] E. Treister and E. Haber, “A fast marching algorithm for the factored eikonal equation,” *Journal of Computational Physics*, vol. 324, pp. 210–225, Nov. 2016. DOI: 10.1016/j.jcp.2016.08.012.
- [29] J. A. Sethian, “Fast marching methods,” *SIAM Review*, vol. 41, no. 2, pp. 199–235, 1999. DOI: 10.1137/S0036144598347059. eprint: <https://doi.org/10.1137/S0036144598347059>. [Online]. Available: <https://doi.org/10.1137/S0036144598347059>.
- [30] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into deep learning*, 2021. arXiv: 2106.11342 [cs.LG].
- [31] D. Kundur and D. Hatzinakos, “Blind image deconvolution,” *IEEE Signal Processing Magazine*, vol. 13, no. 3, pp. 43–64, 1996.
- [32] D. Perrone and P. Favaro, “Total variation blind deconvolution: The devil is in the details,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2014, pp. 2909–2916. DOI: 10.1109/CVPR.2014.372. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.372>.

- [33] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, “Deblurgan: Blind motion deblurring using conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183–8192.
- [34] A. Chakrabarti, “A neural approach to blind motion deblurring,” in *European conference on computer vision*, Springer, 2016, pp. 221–235.
- [35] R. Aljadaany, D. K. Pal, and M. Savvides, “Douglas-rachford networks: Learning both the image prior and data fidelity terms for blind image deconvolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 235–10 244.
- [36] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo, “Neural blind deconvolution using deep priors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3341–3350.
- [37] S. Bell-Kligler, A. Shocher, and M. Irani, “Blind super-resolution kernel estimation using an internal-gan,” *CoRR*, vol. abs/1909.06581, 2019. arXiv: 1909.06581. [Online]. Available: <http://arxiv.org/abs/1909.06581>.
- [38] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [39] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *arXiv preprint arXiv:1605.08803*, 2016.
- [40] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [41] S. Bell-Kligler, A. Shocher, and M. Irani, *Blind super-resolution kernel estimation using an internal-gan*, 2020. arXiv: 1909.06581 [cs.CV].
- [42] Z. E. Ross, E. S. Cochran, D. T. Trugman, and J. D. Smith, “3D fault architecture controls the dynamism of earthquake swarms,” *Science*, vol. 368, no. 6497, pp. 1357–1361, 2020, issn: 0036-8075, 1095-9203. doi: 10.1126/science.abb0779.
- [43] M. Bertero, D. Bindi, P. Boccacci, M. Cattaneo, C. Eva, and V. Lanza, “Application of the projected Landweber method to the estimation of the source time function in seismology,” *Inverse Problems*, vol. 13, no. 2, p. 465, 1997, issn: 0266-5611. doi: 10.1088/0266-5611/13/2/017.

Chapter 4

IMAGING WITH LIMITED PRIORS

This chapter is adapted from previously published work [1]–[3].

Abstract

We consider solving ill-posed imaging inverse problems without access to an image prior or ground-truth examples. An overarching challenge in these inverse problems is that an infinite number of images, including many that are implausible, are consistent with the observed measurements. Thus, image priors are required to reduce the space of possible solutions to more desirable reconstructions. However, in many applications it is difficult or potentially impossible to obtain example images to construct an image prior. Hence inaccurate priors are often used, which inevitably result in biased solutions. Rather than solving an inverse problem using priors that encode the spatial structure of any one image, we propose to solve a set of inverse problems jointly by incorporating prior constraints on the collective structure of the underlying images. The key assumption of our work is that the underlying images we aim to reconstruct share common, low-dimensional structure. We show that such a set of inverse problems can be solved simultaneously without the use of a spatial image prior by instead inferring a shared image generator with a low-dimensional latent space. The parameters of the generator and latent embeddings are found by maximizing a proxy for the Evidence Lower Bound (ELBO). Once identified, the generator and latent embeddings can be combined to provide reconstructed images for each inverse problem. The framework we propose can handle general forward model corruptions, and we show that measurements derived from only a small number of ground-truth images (≤ 150) are sufficient for image reconstruction. We demonstrate our approach on a variety of convex and non-convex inverse problems, including denoising, phase retrieval, and black hole video reconstruction.

4.1 Introduction

In imaging inverse problems, the goal is to recover an underlying image from corrupted measurements, where the measurements and image are related via an understood forward model: $y = f(x) + \eta$. Here, y are measurements, x is the underlying image, f is a known forward model, and η is noise. Such problems are ubiquitous and include denoising [4], [5], super-resolution [6], compressed sensing

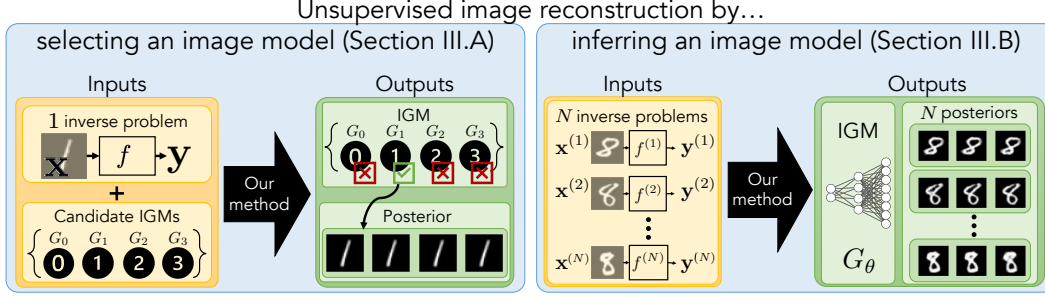


Figure 4.1: **Method overview.** In this chapter, we tackle ill-posed image reconstruction problems when a traditional image prior is not available or cannot be derived from example images. The key idea of our work is that image reconstruction is possible when one has access to only corrupted measurement examples and the underlying images share common, low-dimensional structure. We explore such ideas in this work in the following two ways: **Left:** Given a single measurement example (with the corresponding forward model) and a set of candidate image generation models (IGMs), we use our proposed criterion to select the best IGM from the measurements alone and then recover an image posterior under that IGM. The criterion we use for model selection is an approximation of the evidence lower bound (i.e., ELBOProxy in Eq. (4.3)) (see Sec. 4.3). **Right:** Building off of this, we consider the setting where we must infer the IGM from noisy measurements alone (see Sec. 4.3). In this case, we solve N inverse problems simultaneously. The inputs of this method are N measurement examples with their known forward models, and the outputs are a single inferred IGM and N latent embeddings that, when combined, lead to N image reconstruction posteriors.

[7], [8], phase retrieval [9], and deconvolution [10]. Due to corruption by the forward model and noise, these problems are often ill-posed: there are many images that are consistent with the observed measurements, including ones that are implausible.

To combat the ill-posedness in imaging problems, solving for an image traditionally requires imposing additional structural assumptions to reduce the space of possible solutions. We encode these assumptions in an *image generation model (IGM)*, whose goal is to capture the desired properties of an image’s spatial structure. IGMs are general; they encompass probabilistic *spatial-domain* priors (e.g., that encourage smoothness or sparsity), but also include deep image generators that are not necessarily probabilistic but are trained to primarily sample a certain class of images.

In order to define an IGM, it is necessary to have knowledge of the underlying image’s structure. If images similar to the underlying image are available, then an

IGM can be learned directly [11]–[13]. However, an abundance of clean images is not available for many scientific imaging modalities (e.g., geophysical imaging and astronomical imaging). Collecting images in these domains can be extremely invasive, time-consuming, expensive, or even impossible. For instance, how should we define an IGM for black hole imaging without having ever seen a direct image of a black hole or knowing what one should look like? Moreover, classical approaches that utilize hand-crafted IGMs, such as total variation [14] or sparsity in a wavelet basis [15], are prone to human bias [16].

In this work, we show how one can solve a set of ill-posed image reconstruction tasks in an unsupervised fashion, i.e., without prior information about an image’s spatial structure or access to clean, example images. The key insight of our work is that knowledge of common structure across multiple diverse images can be sufficient regularization alone. In particular, suppose we have access to a collection of noisy measurements $\{y^{(i)}\}_{i=1}^N$ that are observed through (potentially different) forward models $y^{(i)} := f^{(i)}(x^{(i)}) + \eta^{(i)}$. The core assumption we make is that the different underlying images $\{x^{(i)}\}_{i=1}^N$ are drawn from the same distribution (unknown *a priori*) and share common, low-dimensional structure. Thus, our “prior” is not at the spatial-level, but rather exploits the *collective structure* of the underlying images. This assumption is satisfied in a number of applications where there is no access to an abundance of clean images. For instance, although we might not know what a black hole looks like, we might expect it to be similar in appearance over time. We show that under this assumption, the image reconstruction posteriors $p(x|y^{(i)})$ can be learned jointly from a small number of examples $\{y^{(i)}\}_{i=1}^N$ due to the common, low-dimensional structure of the collection $\{x^{(i)}\}_{i=1}^N$. Specifically, our main result is that one can capitalize on this common structure by jointly solving for 1) a shared image generator G_θ and 2) N low-dimensional latent distributions $q_{\phi^{(i)}}$, such that the distribution induced by the push-forward of $q_{\phi^{(i)}}$ through G_θ approximately captures the image reconstruction posterior $p(x|y^{(i)})$ for each measurement example $i \in [N]$.

Our Contributions

We outline the main contributions of our work, which extends our prior work presented in [2]:

1. We solve a collection of ill-posed inverse problems without prior knowledge of an image’s spatial structure by exploiting the common, low-dimensional structure shared across images. This common structure is exploited when

inferring a shared IGM with a low-dimensional latent space.

2. To infer this IGM, we define a loss inspired by the evidence lower bound (ELBO). We motivate this loss by showing how it aids in unsupervised image reconstruction by helping select one IGM from a collection of candidate IGMs using a single measurement example.
3. We apply our approach to convex and non-convex inverse problems, such as denoising, black hole compressed sensing, and phase retrieval. We establish that we can solve inverse problems without spatial-level priors and demonstrate good performance with only a small number of independent measurement examples (e.g., ≤ 150).
4. We theoretically analyze the inferred IGM in linear inverse problems under a linear image model to show that in this setting the inferred IGM performs dimensionality reduction akin to PCA on the collection of measurements.

4.2 Background and related work

We now discuss related literature in model selection and learning-based IGMs. In order to highlight our key contributions, we emphasize the following assumptions in our framework:

1. We do not have access to a set of images from the same distribution as the underlying images.
2. We only have access to a collection of measurement examples, where each example comes from a different underlying image. The number of examples N is small, e.g., $N \leq 150$.
3. For each underlying image $x^{(i)}$ we wish to reconstruct, we only have access to a single measurement example $y^{(i)} = f^{(i)}(x^{(i)}) + \eta^{(i)}$. That is, we do not have multiple observations of the same underlying image. Note each $f^{(i)}$ can be potentially different.

Model selection

Model selection techniques seek to choose a model that best explains data by balancing performance and model complexity. In supervised learning problems with sufficiently large amounts of data, this can be achieved simply by evaluating the performance of different candidate models using reserved test data [17]. However,

in image reconstruction or other inverse problems with limited data, one cannot afford to hold out data. In these cases, model selection is commonly conducted using probabilistic metrics. The simplest probabilistic metric used for linear model selection is adjusted R^2 [18]. It re-weights the goodness-of-fit by the number of linear model parameters, helping reject high-dimensional parameters that do not improve the data fitting accuracy. Similar metrics in nonlinear model selection are Bayesian Information Criterion (BIC) [19] and Akaike Information Criterion (AIC) [20]. AIC and BIC compute different weighted summations of a model’s log-likelihood and complexity, offering different trade-offs between bias and variance to identify the best model for a given dataset.

In our work, we consider the use of the ELBO as a model selection criterion. In [21], [22], the use of the ELBO as a model selection criterion is theoretically analyzed and rates of convergence for variational posterior estimation are shown. Additionally, [23] proposes a generalized class of evidence lower bounds leveraging an extension of the evidence score. In [24], the ELBO is used for model selection to select a few, discrete parameters modeling a physical system (e.g., parameters that govern the orbit of an exoplanet). A significant difference in our context, however, is that we use the ELBO as a model selection criterion in a *high-dimensional* imaging context, and we optimize the ELBO over a continuous space of possible parameters.

Learning IGMs

With access to a large corpus of example images, it is possible to directly learn an IGM to help solve inverse problems. Seminal work along these lines utilizing generative networks showcased that a pre-trained Generative Adversarial Network (GAN) can be used as an IGM in the problem of compressed sensing [13]. To solve the inverse problem, the GAN was used to constrain the search space for inversion. This approach was shown to outperform sparsity-based techniques with 5-10x fewer measurements. Since then, this idea has been expanded to other inverse problems, including denoising [25], super-resolution [26], magnetic resonance imaging (MRI) [27], [28], and phase retrieval [29], [30]. However, the biggest downside to this approach is the requirement of a large dataset of example images similar to the underlying image, which is often difficult or impossible to obtain in practice. Hence, we consider approaches that are able to directly solve inverse problems without example images.

Methods that aim to learn an IGM from only noisy measurements have been proposed. The main four distinctions between our work and these methods are that these

works either: 1) require multiple independent observations of the *same* underlying image, 2) can only be applied to certain inverse problems, 3) require significantly more observations (either through more observations of each underlying image or by observing more underlying images), or 4) require significant hyperparameter tuning based on knowledge of example images.

Noise2Noise (N2N) [31] learns to denoise images by training on a collection of noisy, independent observations of the *same* image. To do so, N2N learns a neural network Φ_θ whose goal is to map between noisy images y and denoised images x . Since it has no denoised image examples to supervise training, it instead employs a loss that maps between noisy examples of the same underlying image. This objective is as follows:

$$\arg \min_{\theta} \sum_{i=1}^N \mathbb{E}_{y_1^{(i)} \sim \mathcal{Y}_1} \mathbb{E}_{y_2^{(i)} \sim \mathcal{Y}_2} [L(\Phi_\theta(y_1^{(i)}), y_2^{(i)})], \quad (4.1)$$

where $y_j^{(i)}$ corresponds to a noisy observation of the i -th underlying image $x^{(i)}$, and $\mathcal{Y}_j$ is a distribution of noisy images where $\mathbb{E}_{y \sim \mathcal{Y}_j} [y] = x$. This N2N objective requires at least two observations of the same image and is limited by the assumption that the expected value of multiple observations of a single image is the underlying image. Thus, N2N is only applicable to denoising problems where the forward model is the identity matrix with independent noise on each pixel. Additionally, in practice N2N requires thousands of underlying images (i.e., $N = O(1000)$) to perform well. Thus, N2N's main distinctions with our work are distinctions 1), 2), and 3).

Regularization by Artifact Removal (RARE) [32] generalizes N2N to perform image reconstruction from measurements under linear forward models. That is, the objective in Eq. (4.1) is modified to include a pseudo-inverse. Nonetheless, multiple observations of the same underlying image are required, such that $\mathbb{E}_{y \sim \mathcal{Y}} [A^\dagger y] = x$ for the pseudo-inverse matrix $A^\dagger$. Thus, RARE suffers from the same limiting distinctions as N2N (i.e., 1), 2), and 3)).

Noise2Void [33] and **Noise2Self** [34] assume that the image can be partitioned such that the measurement noise in one subset of the partition is independent conditioned on the measurements in the other subset. This is true for denoising, but not applicable to general forward models. For example, in black hole and MRI compressed sensing, it is not true that the measurement noise can be independently partitioned since each measurement is a linear combination of all pixels. While this makes Noise2Void and Noise2Self more restrictive in the corruptions they can handle compared to RARE,

they also don’t require multiple observations of the same underlying image. Hence the main differences between these works and our own are distinctions 2) and 3).

AmbientGAN [35] and other similar approaches based on GANs [36] and Variational Autoencoders (VAEs) [37], [38] have been proposed to learn an IGM directly from noisy measurements. For instance, AmbientGAN aims to learn a generator whose images lead to simulated measurements that are indistinguishable from the observed measurements; this generator can subsequently be used as a prior to solve inverse problems. However, AmbientGAN requires many measurement examples (on the order of 10,000) to produce a high quality generator. We corroborate this with experiments in Section 4.4 to show that they require many independent observations and/or fine tuning of learning parameters to achieve good performance. Thus, the main distinctions between AmbientGAN and our work are 3) and 4).

Deep Image Prior (DIP) [39] uses a convolutional neural network as an implicit “prior”. DIP has shown strong performance across a variety of inverse problems to perform image reconstruction without explicit probabilistic priors. However, it is prone to overfitting and requires selecting a specific stopping criterion. While this works well when example images exist, selecting this stopping condition from noisy measurements alone introduces significant human bias that can negatively impact results. Thus, the main distinction between DIP and our work is 4).

We would also like to highlight additional work done to improve certain aspects of the DIP. While the original DIP method used a U-Net architecture [40], other works such as the Deep Decoder [41] and ConvDecoder [42] used a decoder-like architecture that progressively grows a low-dimensional random tensor to a high-dimensional image. When underparameterized, such architectures have been shown to avoid overfitting and the need for early stopping. Other works mitigating early stopping include [43], which takes a Bayesian perspective to the DIP by using Langevin dynamics to perform posterior inference over the weights to improve performance and show this does not lead to overfitting.

4.3 Approach

In this work, we propose to solve a set of inverse problems without prior access to an IGM by assuming that the set of underlying images have common, low-dimensional structure. We motivate the use of optimizing the ELBO to infer an IGM by showing that it is a good criterion for generative model *selection* in Section 4.3. Then, by optimizing the ELBO, we show in Section 4.3 that one can directly infer an IGM from corrupted measurements alone by parameterizing the image model as a deep

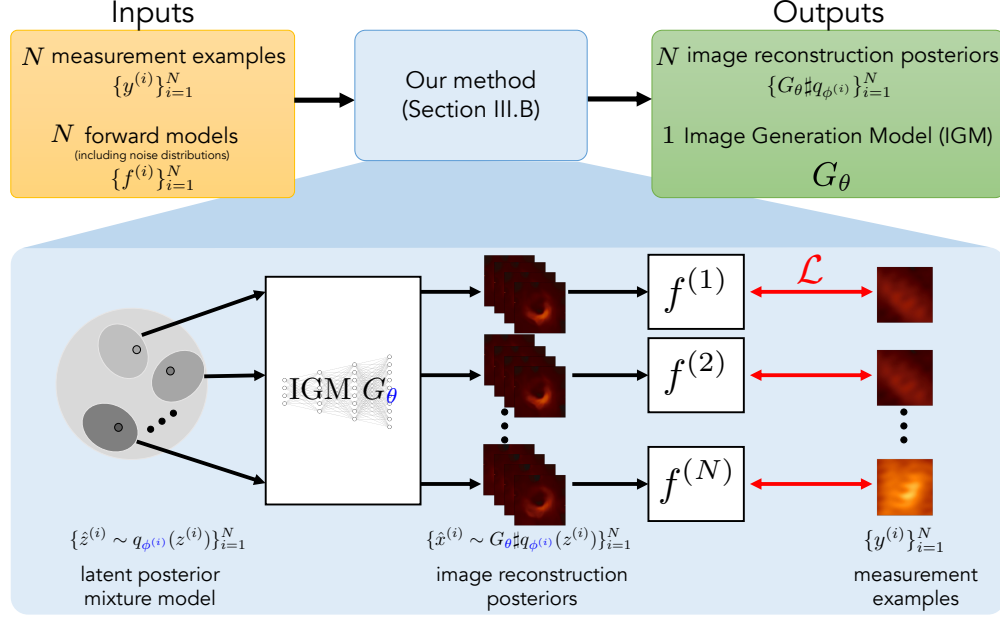


Figure 4.2: We propose to solve a collection of N ill-posed inverse problems by exploiting the common, low-dimensional structure of the underlying images. Given a set of N measurement examples $\{y^{(i)}\}_{i=1}^N$ from N different underlying images, we propose to model each image posterior as the output of a shared IGM with a low-dimensional latent space. In particular, each posterior is approximated by $G_{\theta} \# q_{\phi^{(i)}}$, the push-forward of $q_{\phi^{(i)}}$ through G_{θ} , where G_{θ} is the shared, common generator to all N examples and $q_{\phi^{(i)}}$ is a low-dimensional, variational distribution particular to the i -th example. The parameters, θ and $\{\phi^{(i)}\}_{i=1}^N$, are colored in blue and are jointly inferred. That is, such parameters are inferred specifically for the measurements $\{y^{(i)}\}_{i=1}^N$. The loss we use is the negative ELBOProxy, which is denoted by $\mathcal{L}$ and given by Eq. (4.5). Note that there is no notion of a training set and test set as we aim to solve the inverse problems jointly from all available measurements.

generative network with a low-dimensional latent distribution. The IGM network weights are shared across all images, capitalizing on the common structure present in the data, while the parameters of each latent distribution are learned jointly with the generator to model the image posteriors for each measurement example.

Motivation for ELBO as a model selection criterion

In order to accurately *infer* an IGM, we motivate the use of the ELBO as a loss by showing that it provides a principled criterion for *selecting* an IGM to use as a prior model. Suppose we are given noisy measurements from a single image: $y = f(x) + \eta$. In order to reconstruct the image x , we traditionally first require an IGM G that captures the distribution x was sampled from. A natural approach would

be to find or select the model G that maximizes the model posterior distribution $p(G|y) \propto p(y|G)p(G)$. That is, conditioned on the noisy measurements, find the IGM of highest likelihood. Unfortunately computing $p(y|G)$ is intractable, as it requires marginalizing and integrating over all x encompassed by the IGM G . However, we show that this quantity can be well approximated using the ELBO.

To motivate our discussion, we first consider estimating the image posterior $p(x|y, G)$ by learning the parameters ϕ of a variational distribution $h_\phi(x)$. Observe that the definition of the KL-divergence followed by an application of Bayes' theorem gives

$$\begin{aligned} D_{\text{KL}}(h_\phi(x) \parallel p(x|y, G)) &:= \mathbb{E}_{x \sim h_\phi(x)} \left[\log \frac{h_\phi(x)}{p(x|y, G)} \right] \\ &= \mathbb{E}_{x \sim h_\phi(x)} \left[\log \frac{h_\phi(x)p(y|G)}{p(y|x, G)p(x|G)} \right] \\ &= -\mathbb{E}_{x \sim h_\phi(x)} [\log p(y|x, G) + \log p(x|G) - \log h_\phi(x)] \\ &\quad + \log p(y|G). \end{aligned}$$

The ELBO of an IGM G given measurements y under variational distribution h_ϕ is defined by

$$\begin{aligned} \text{ELBO}(G, h_\phi; y) &:= \mathbb{E}_{x \sim h_\phi(x)} [\log p(y|x, G) \\ &\quad + \log p(x|G) - \log h_\phi(x)]. \end{aligned} \tag{4.2}$$

Rearranging the previous equation, we see that by the non-negativity of the KL-divergence that

$$\begin{aligned} \log p(y|G) &= D_{\text{KL}}(h_\phi(x) \parallel p(x|y, G)) \\ &\quad + \text{ELBO}(G, h_\phi; y) \\ &\geq \text{ELBO}(G, h_\phi; y). \end{aligned}$$

Thus, we can lower bound the model posterior as

$$\log p(G|y) \geq \text{ELBO}(G, h_\phi; y) + \log p(G) - \log p(y).$$

Note that $\log p(y)$ is independent of the parameters of interest, ϕ . If the variational distribution $h_\phi(x)$ is a good approximation to the posterior $p(x|y, G)$, $D_{\text{KL}} \approx 0$. Thus, maximizing $\log p(G|y)$ with respect to G is approximately equivalent to maximizing $\text{ELBO}(G, h_\phi; y) + \log p(G)$.

Each term in the ELBO objective encourages certain properties of the IGM G . In particular, the first term in the ELBO, $\mathbb{E}_{x \sim h_\phi(x)} [\log p(y|x, G)]$, requires that

G should lead to an image estimate that is consistent with our measurements y . The second term, $\mathbb{E}_{x \sim h_\phi(x)}[\log p(x|G)]$, encourages images sampled from $h_\phi(x)$ to have high likelihood under our model G . The final term is the entropy term, $\mathbb{E}_{x \sim h_\phi(x)}[-\log h_\phi(x)]$, which encourages a G that leads to “fatter” minima that are less sensitive to small changes in likely images x under G .

ELBOProxy

Some IGMs are explicit, which allows for direct computation of $\log p(x|G)$. For example, if our IGM models x as isotropic Gaussian with variance λ , then $-\log p(x|G) \propto \lambda^{-1}\|x\|_2^2$. In this case, we can optimize the ELBO defined in Equation (4.2) directly and then perform model selection. However, an important class of IGMs that we are interested in are those given by deep generative networks. Such IGMs are not probabilistic in the usual Bayesian interpretation of a prior, but instead implicitly enforce structure in the data. A key characteristic of many generative network architectures (e.g., VAEs and GANs) that we leverage is that they generate high-dimensional images from low-dimensional latent representations. Bottlenecking helps the network learn global characteristics of the underlying image distribution while also respecting the low intrinsic dimensionality of natural images. However, this means that we can only compute $\log p(x|G)$ directly if we have an injective map [44]. This architectural requirement limits the expressivity of the network.

We instead consider a proxy of the ELBO that is especially helpful for deep generative networks. That is, suppose our IGM is of the form $x = G(z)$. Introducing a variational family for our latent representations $z \sim q_\phi(z)$ and choosing a latent prior distribution $\log p_Z(z|G)$, we arrive at the following proxy of the ELBO:

$$\begin{aligned} \text{ELBOProxy}(G, q_\phi; y) &:= \mathbb{E}_{z \sim q_\phi(z)}[\log p(y|G(z)) \\ &\quad + \log p_Z(z|G) - \log q_\phi(z)]. \end{aligned} \quad (4.3)$$

In our experiments, we chose $p_Z(z|G)$ to be an isotropic Gaussian prior. This is a common choice in many generative modeling frameworks and has shown to be a good choice of prior in the latent space.

To motivate this proxy, it is instructive to consider the case where our variational distribution $h_\phi := G\#q_\phi$ is the push-forward of a latent distribution q_ϕ through an injective or invertible function G . To be precise, recall the following definition of the push-forward measure.

Definition 1. Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a measurable function and suppose p is a distribution (or, more generally, a measure) on $\mathbb{R}^k$. Then the push-forward measure $\mu := G\#p$ is the measure on $\mathbb{R}^n$ that satisfies the following: for all Borel sets A of $\mathbb{R}^n$, $\mu(A) = p(G^{-1}(A))$ where $G^{-1}(A)$ denotes the preimage of A with respect to G .

The push-forward measure essentially characterizes how a distribution p changes when passed through a function G . It follows from the definition of the push-forward that $x \sim G\#q_\phi$ if and only if $x = G(z)$ where $z \sim q_\phi$. In the case $h_\phi = G\#q_\phi$ for an injective function G , the ELBO and ELBOProxy are equivalent, as shown in the following proposition:

Proposition 1. Suppose $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is continuously differentiable and injective. For two probability distributions p_Z and q_ϕ on $\mathbb{R}^k$, define the measures $p(\cdot|G) = G\#p_Z$ and $h_\phi = G\#q_\phi$. Then

$$\text{ELBO}(G, h_\phi; y) = \text{ELBOProxy}(G, q_\phi; y) \quad \forall y \in \mathbb{R}^m.$$

Proof. It suffices to show

$$\begin{aligned} \mathbb{E}_{x \sim h_\phi(x)} [\log p(x|G) - \log h_\phi(x)] \\ = \mathbb{E}_{z \sim q_\phi(z)} [\log p_Z(z|G) - \log q_\phi(z)] \end{aligned}$$

Let $J_G(z) \in \mathbb{R}^{n \times k}$ denote the Jacobian of G at an input $z \in \mathbb{R}^k$. Since G is injective and continuously differentiable with $p(\cdot|G) = G\#p_Z$, we can compute the likelihood of any point $x \in \text{range}(G)$ [44] via

$$\begin{aligned} \log p(x|G) &= \log p_Z(G^\dagger(x)|G) \\ &\quad - \frac{1}{2} \log |\det[J_G(G^\dagger(x))^T J_G(G^\dagger(x))]| \end{aligned}$$

where $G^\dagger$ is the inverse of G on its range. This is essentially the classical change-of-variables formula specialized to the case when G is injective and we wish to access likelihoods on the range of G . Note that this equation is only valid for points in the range of the injective function G . Likewise, since $h_\phi = G\#q_\phi$, we can compute the entropy of h_ϕ for any point $x \in \text{range}(G)$ as

$$\begin{aligned} \log h_\phi(x) &= \log q_\phi(G^\dagger(x)) \\ &\quad - \frac{1}{2} \log |\det[J_G(G^\dagger(x))^T J_G(G^\dagger(x))]|. \end{aligned}$$

Now observe that for $x \sim h_\phi$, $x \in \text{range}(G)$ as h_ϕ is the push-forward of q_ϕ through G . Thus, for $x \sim h_\phi$, we have that

$$\begin{aligned} \log p(x|G) - \log h_\phi(x) \\ = \log p_Z(G^\dagger(x)|G) - \log q_\phi(G^\dagger(x)). \end{aligned}$$

By the definition of the push-forward measure, we have that $x \sim h_\phi$ implies $x = G(z)$ for some $z \sim q_\phi$. Using our previous formulas, we can compute the expectation over the difference $\log p(x|G) - \log h_\phi(x)$ with respect to h_ϕ as

$$\begin{aligned} \mathbb{E}_{x \sim h_\phi(x)} [\log p(x|G) - \log h_\phi(x)] \\ = \mathbb{E}_{x \sim h_\phi(x)} [\log p_Z(G^\dagger(x)|G) - \log q_\phi(G^\dagger(x))] \\ = \mathbb{E}_{z \sim q_\phi(z)} [\log p_Z(G^\dagger(G(z))|G) - \log q_\phi(G^\dagger(G(z)))] \\ = \mathbb{E}_{z \sim q_\phi(z)} [\log p_Z(z|G) - \log q_\phi(z)]. \end{aligned}$$

□

An important consequence of this result is that for injective generators G , the inverse of G (on its range) is not required for computing the ELBO. In this case, the ELBOProxy is in fact equivalent to the ELBO. While not all generators G will be injective, quality generators are largely injective over high likelihood image samples. In Section 4.3 and Fig. 4.3, we experimentally show that this proxy can aid in selecting potentially non-injective generative networks from corrupted measurements.

Toy example

To illustrate the use of the ELBOProxy as a model selection criterion, we conduct the following experiment that asks whether the ELBOProxy can identify the best model from a given set of image generation models. For this experiment, we use the MNIST dataset [45] and consider two inverse problems: denoising and phase retrieval. We train a generative model G_c on each class $c \in \{0, 1, 2, \dots, 9\}$ using the clean MNIST images directly. Hence, G_c generates images from class c via $G_c(z)$ where $z \sim \mathcal{N}(0, I)$. Then, given noisy measurements y_c from a single image from class c , we ask whether the generative model G_c from the appropriate class would achieve the best ELBOProxy. Each G_c is the decoder of a VAE with a low-dimensional latent space, with no architectural constraints to ensure injectivity. For denoising, our measurements are $y_c = x_c + \eta_c$ where $\eta_c \sim \mathcal{N}(0, \sigma^2 I)$ and

$\sigma = \sqrt{0.5}$. For phase retrieval, $y_c = |\mathcal{F}(x_c)| + \eta_c$ where $\mathcal{F}$ is the Fourier transform and $\eta_c \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = \sqrt{0.05}$.

We construct 10×10 arrays for each problem, where in the i -th row and j -th column, we compute the negative ELBOProxy obtained by using model G_{i-1} to reconstruct images from class $j - 1$. We calculate $\text{ELBOProxy}(G_c, q_{\phi_c}; y_c)$ by parameterizing q_{ϕ_c} with a Normalizing Flow [46] and optimizing network weights ϕ_c to maximize (4.3). The expectation in the ELBOProxy is approximated via Monte Carlo sampling. Results from the first 5 classes are shown in Fig. 4.3 and the full arrays are shown in the supplemental materials. We note that all of the correct models are chosen in both denoising and phase retrieval. We also note some interesting cases where the ELBOProxy values are similar for certain cases, such as when recovering the 3 or 4 image. For example, when denoising the 4 image, both G_4 and G_9 achieve comparable ELBOProxy values. By carefully inspecting the noisy image of 4, one can see that both models are reasonable given the structure of the noise.

Simultaneously solving many inverse problems

As the previous section illustrates, the ELBOProxy provides a good criterion for choosing an appropriate IGM from noisy measurements. Here, we consider the task of directly inferring the IGM from a collection of measurement examples $y^{(i)} = f^{(i)}(x^{(i)}) + \eta^{(i)}$ for $i \in [N]$, where the parameters are found by optimizing the ELBOProxy. The key assumption we make is that common, low-dimensional structure is shared across the underlying images $\{x^{(i)}\}_{i=1}^N$. We propose to find a *shared* generator G_θ with weights θ along with latent distributions $q_{\phi^{(i)}}$ that can be used to reconstruct the full posterior of each image $x^{(i)}$ from its corresponding measurement example $y^{(i)}$. This approach is illustrated in Fig. 4.2. Having the generator be shared across all images helps capture their common collective structure. Each forward model corruption, however, likely induces its own complicated image posteriors. Hence, we assign each measurement example $y^{(i)}$ its own latent distribution to capture the differences in their posteriors. Note that because we optimize a proxy of the ELBO, the inferred distribution may not necessarily be the true image posterior, but it still captures a distribution of images that fit to the observed measurements.

Inference approach More explicitly, given a collection of measurement examples $\{y^{(i)}\}_{i=1}^N$, we jointly infer a generator G_θ and a set of variational distributions

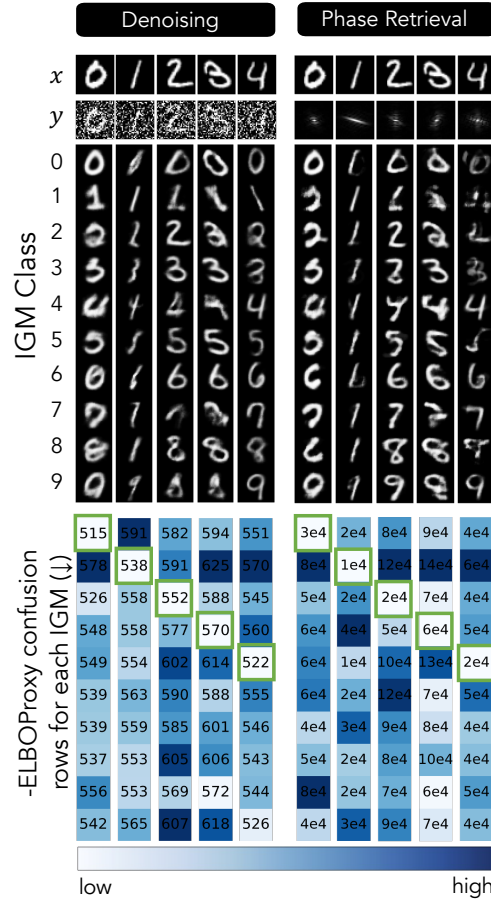


Figure 4.3: We consider two inverse problems: denoising and phase retrieval. Top: the two topmost rows correspond to the ground-truth image x_c and the noisy measurements y_c . Center: in each column, we show the means of the distribution induced by the push-forward of G_j and each latent distribution $z \sim q_{\phi_j}$ for $j \in \{0, \dots, 9\}$. Bottom: each column of the array corresponds to the negative ELBOProxy achieved by each model in reconstructing the images. Here, lower is better. Boxes highlighted in green correspond to the best negative ELBOProxy values in each column. In all these examples, the correct model was chosen.

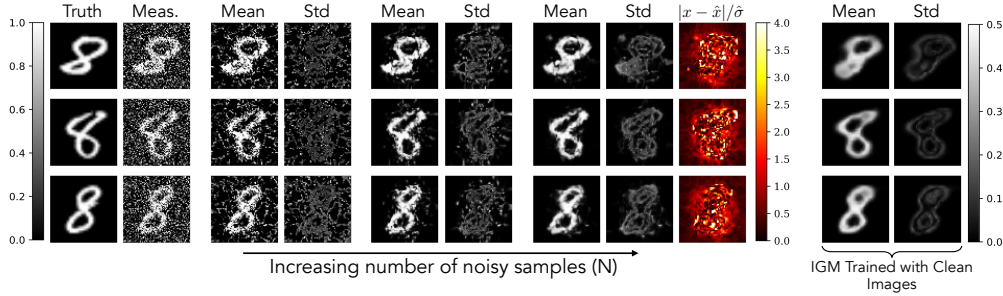


Figure 4.4: **Improvement with an increasing number of noisy observations.** We demonstrate our method of inferring an IGM to perform denoising for an increasing number of noisy MNIST images (5, 75, and 150 images from left to right). We showcase results on three randomly selected examples that appear in each collection of inverse problems. In each panel, we include the ground-truth, noisy measurements, mean of the posterior, and standard deviation of the posterior. We also include the residual error divided by the empirical standard deviation for $N = 150$. On the far right, we visualize reconstructions using an IGM trained on the full clean MNIST 8’s class (6000 images). We observe that the mean reconstructions and standard deviations from our low-data IGMs become more similar to the full-data IGM with increasing data. Our residual errors are largely within 3 standard deviations.

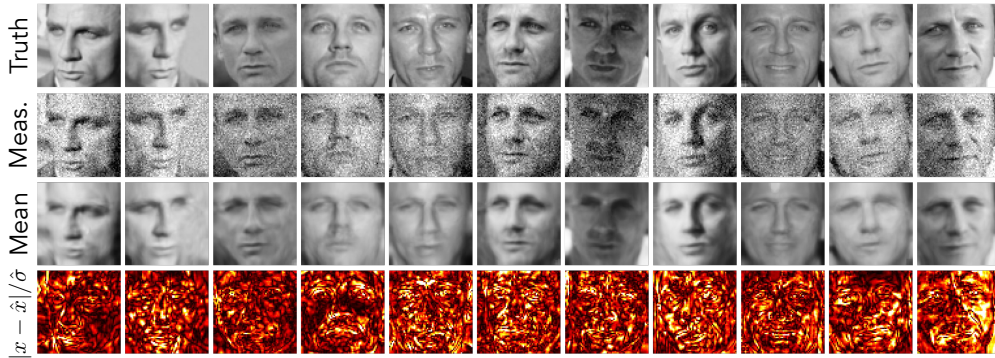


Figure 4.5: **Denoising 95 images of celebrity A.** We demonstrate our method described in 4.3 using 95 noisy images of a celebrity. Here we show the underlying image (row 1), noisy measurements (row 2), mean reconstruction (row 3), and residual error (row 4) for a subset of the 95 different noisy images. Our reconstructions are much less noisy than the measurements and recover sharp features that are hard to discern in the noisy images. We visualize the residual error normalized by the empirical standard deviation, which indicates errors are largely within 3 standard deviations (refer to the colorbar in Fig. 4.4). Note that no explicit spatial-domain prior/regularizer was used in denoising.

$\{q_{\phi^{(i)}}\}_{i=1}^N$ by optimizing a Monte Carlo estimate of the ELBOProxy from Eq. (4.3),

described by:

$$\{\hat{\theta}, \hat{\phi}^{(1)}, \dots, \hat{\phi}^{(N)}\} \in \arg \max_{\theta, \{\phi^{(i)}\}_{i=1}^N} \mathcal{L} \quad (4.4)$$

where

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \text{ELBOProxy}(G_{\theta}, q_{\phi^{(i)}}; y^{(i)}) + \log p(G_{\theta}). \quad (4.5)$$

In terms of choices for $\log p(G_{\theta})$, we can add additional regularization to promote particular properties of the IGM G_{θ} , such as having a small Lipschitz constant. Here, we consider having sparse neural network weights as a form of regularization and use dropout [47] during training to represent $\log p(G_{\theta})$.

Once a generator $G_{\hat{\theta}}$ and variational parameters $\hat{\phi}^{(i)}$ have been inferred, we solve the i -th inverse problem by simply sampling $\hat{x}^{(i)} = G_{\hat{\theta}}(\hat{z}^{(i)})$ where $\hat{z}^{(i)} \sim q_{\hat{\phi}^{(i)}}(z^{(i)})$ or computing an average $\bar{x}^{(i)} = \frac{1}{T} \sum_{t=1}^T G_{\hat{\theta}}(\hat{z}_t^{(i)})$. Producing samples for each inverse problem can help visualize the range of uncertainty under the learned IGM $G_{\hat{\theta}}$, while the expected value of the distribution empirically provides clearer estimates with better metrics in terms of PSNR or MSE. We report PSNR outputs in our subsequent experiments and also visualize the standard deviation of our reconstructions.

4.4 Experimental results

We now consider solving a set of inverse problems via the framework described in 4.3. For each of these experiments, we use a multivariate Gaussian distribution to parameterize each of the posterior distributions $q_{\phi^{(i)}}$ and a Deep Decoder [41] with 6 layers, 150 channels in each layer, a latent size of 40, and a dropout of 10^{-4} as the IGM. The multivariate Gaussian distributions are parameterized by means and covariance matrices $\{\mu^{(i)}, \Lambda^{(i)} = U_i U_i^T + \varepsilon I\}_{i=1}^N$, where εI with $\varepsilon = 10^{-3}$ is added to the covariance matrix to help with stability of the optimization. We choose to parameterize the latent distributions using Gaussians for memory considerations. Note that the same hyperparameters are used for all experiments demonstrating our proposed method.

In our experiments, we also compare to the following baseline methods: AmbientGAN [35], Deep Image Prior (DIP) [48], and regularized maximum likelihood using total variation (TV-RML). AmbientGAN is most similar to our setup, as it constructs an IGM directly from measurement examples; however, it doesn't aim to estimate image reconstruction posteriors, but instead aims to learn an IGM that

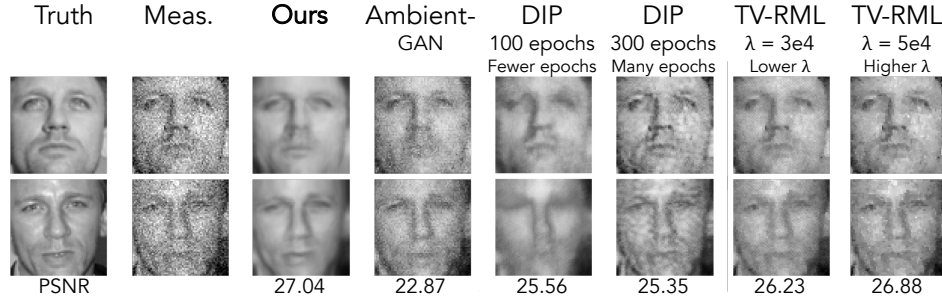


Figure 4.6: **Denoising baseline comparisons.** We compare to various baselines (AmbientGAN, Deep Image Prior (DIP), and regularized maximum likelihood using TV (TV-RML) with weight λ), and we report the average PSNR across all 95 reconstructions. We show both early stopping and full training results using DIP. Our method exhibits higher PSNR than all other baselines. We also include results for baselines that require fine-tuning to demonstrate sensitivity to subjective stopping conditions.

samples from the full underlying prior. TV-RML uses explicit total variation regularization, while DIP uses an implicit convolutional neural network “prior”. As we will show, all these baseline methods require fine-tuning hyperparameters to each set of measurements in order to produce their best results. All methods can also be applied to a variety of inverse problems, making them appropriate choices as baselines.

Denoising

We show results on denoising a collection of noisy images of 8’s from the MNIST dataset in Fig. 4.4 and denoising a collection of noisy images of a single face from the PubFig [49] dataset in Fig. 4.5. The measurements for both datasets are defined by $y = x + \eta$ where $\eta \sim \mathcal{N}(0, \sigma^2 I)$ with an SNR of ~ -3 dB for the MNIST digits and an SNR of ~ 15 dB for the faces. Our method is able to remove much of the added noise and recovers small scale features, even with only 10’s of observations. As shown in Fig. 4.4, the reconstructions achieved under the learned IGM improves as the number of independent observations increases. Our reconstructions also substantially outperform the baseline methods, as shown in Fig. 4.6. Unlike DIP, our method does not overfit and does not require early stopping. Our method does not exhibit noisy artifacts like those seen in all baselines methods, despite such methods being fine-tuned. We show quantitative comparisons in Table 4.1.

In Fig. 4.7 we show additional *multi-noise* denoising experiments where we have 75 noisy images, which have 3 different noise levels. More formally, $y^{(i)} = x^{(i)} + \eta$

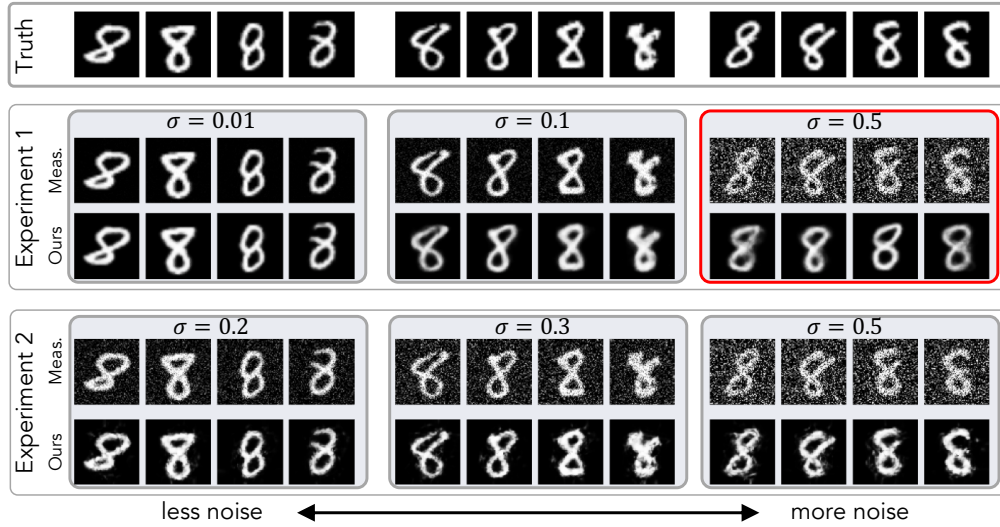


Figure 4.7: **Multi-noise denoising.** We demonstrate our method described in Section 4.3 to perform denoising on measurement collections experiencing different noise levels. For each experiment, we use 75 measurement examples, which are defined by $y^{(i)} = x^{(i)} + \eta$ where $\eta \in \{\eta_1, \eta_2, \eta_3\}$ and $\eta_i \sim \mathcal{N}(0, \sigma_i^2 I)$. In Experiment 1, we use noisy measurement examples that have additive noise with standard deviations of 0.01, 0.1, and 0.5. In Experiment 2, we use noisy measurement examples that have additive noise with standard deviations of 0.2, 0.3, and 0.5. We visualize the true underlying images, the measurement used for each experiment, and the mean of the image reconstruction posterior. Most of the reconstructions recover the primary features of the true image. However, in Experiment 1, the reconstructions of the low SNR measurements exhibit bias and do not match the true images well. This is likely because in Experiment 1, the high SNR measurements influence the inferred IGM more strongly than low SNR measurements, leading to biased reconstructions for the reconstructions highlighted in the red box.

where $\eta \in \{\eta_1, \eta_2, \eta_3\}$ and $\eta_i \sim \mathcal{N}(0, \sigma_i^2 I)$. In Experiment 1, the noise levels have a wide range, and we use standard deviations of $\{\sigma_1, \sigma_2, \sigma_3\} = \{0.01, 0.1, 0.5\}$. In Experiment 2, the noise levels are much closer together, and we use standard deviations of $\{\sigma_1, \sigma_2, \sigma_3\} = \{0.2, 0.3, 0.5\}$. When the SNRs are similar (as in Experiment 2), the reconstructions match the true underlying images well. However, when the measurements have a wide range of SNRs (i.e., Experiment 1), the reconstructions from low SNR measurements exhibit bias and poorly reconstruct the true underlying image, as shown in Fig.4.7. This is likely because the high SNR measurements influence the inferred IGM more strongly than the low SNR measurements. The full set of results are available in the supplemental materials.

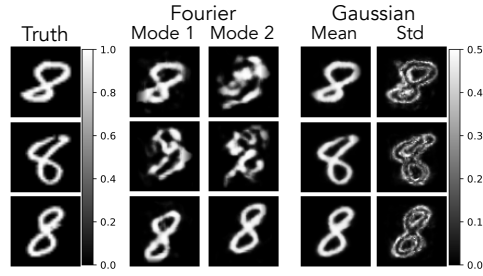


Figure 4.8: **Phase retrieval from MNIST 8's.** We demonstrate our method described in 4.3 to perform phase retrieval on 150 images. For the Fourier phase retrieval setting, we show examples from the two observed modes of the posterior. For the Gaussian phase retrieval setting, we show the mean and standard deviation of our reconstructions.

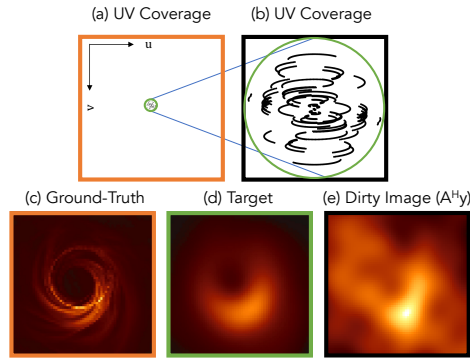


Figure 4.9: **Visualization of the intrinsic resolution of the EHT compressed sensing measurements.** The EHT measures sparse spatial frequencies of the image (i.e., components of the image's Fourier transform). In order to generate the underlying image (c), all frequencies in the entire domain of (a) must be used. Restricting spatial frequencies to the ones in (a) and (b)'s green circle generates the target (d), where (b) is a zoom in of (a). The EHT samples a subset of the interior of the green circle, indicated by the sparse black samples in (b). Naively recovering an image using only these frequencies results in the *dirty image* (e), which is computed by $A^H y$. The 2D spatial Fourier frequency coverage represented with (u, v) positions is referred to as the UV coverage.

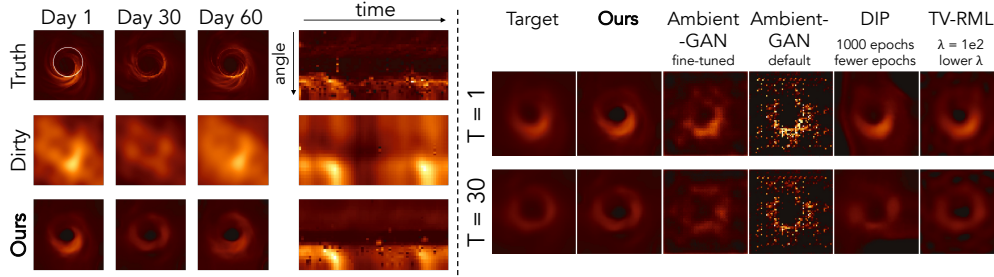


Figure 4.10: **Recovering a 60-day video of the M87 black hole with the EHT telescope array under a compressed sensing forward model.** Left: We demonstrate our method described in 4.3 using simulated measurements from 60 frames of an evolving black hole target with the forward model described in Fig. 4.9. Here we show the underlying images, dirty images ($A^H y$, see Fig. 4.9), and mean reconstruction, respectively. Additionally, we show the unwrapped space $\times$ time image, which is taken along the overlaid white ring illustrated in the day 1 underlying image. The bright-spot’s temporal trajectory of our reconstruction matches that of the truth. Right: We compare our method to various baselines methods. Our results are much sharper and exhibit less artifacts than AmbientGAN and TV-RML with weight λ . Note that we include results using AmbientGAN with both default parameters and fine-tuned parameters.

Phase retrieval

Here we consider solving non-convex inverse problems, and demonstrate our approach on phase retrieval. Our measurements are described by $y = |\mathcal{F}(x)| + \eta$ where $\mathcal{F}(\cdot)$ is a linear operator and $\eta \sim \mathcal{N}(0, \sigma^2 I)$. We consider two types of measurements, one where $\mathcal{F}(\cdot)$ is the Fourier transform and the other when $\mathcal{F}(\cdot)$ is an $m \times n$ complex Gaussian matrix with $m = \lceil 0.1n \rceil$. Since each measurement is the magnitude of complex linear measurements, there is an inherent phase ambiguity in the problem. Additionally, flipping and spatial-shifts are possible reconstructions when performing Fourier phase retrieval. Due to the severe ill-posedness of the problem, representing this complicated posterior that includes all spatial shifts is challenging. Thus, we incorporate an envelope (i.e., a centered rectangular mask) as the final layer of G_θ to encourage the reconstruction to be centered. Nonetheless, flipping and shifts are still possible within this enveloped region.

We show results from a set of $N = 150$ noisy phase retrieval measurements from the MNIST 8’s class with a SNR of ~ 52 dB. We consider three settings: 1) all measurements arise from a Gaussian measurement matrix, 2) all measurements arise from Fourier measurements, and 3) half of the measurements are Gaussian and the other half are Fourier. We show qualitative results for cases 1 and 2 in Fig. 4.8.

Dataset	N	Forward model $y = f(x) + \eta$	Ours	Baselines				
				AmbGAN fine-tuned	DIP fewer epochs	DIP many epochs	TV lower λ	TV higher λ
celeb. A	95	$y = x + \eta$	27.0	22.9	25.6	25.4	26.2	26.9
celeb. B	95	$y = x + \eta$	26.2	19.7	24.4	25.2	25.9	26.4
MNIST 8's	150	$y = x + \eta$	21.1	18.0	18.8	13.3	16.4	18.2
M87* (target)	60	$y = Ax + \eta$	29.3	25.7	29.0	28.6	24.3	25.9

Table 4.1: **Quantitative comparisons to baselines.** We show the mean PSNR between the reconstructions of different methods (ours, AmbientGAN [35], DIP [48], and TV-RML) with either the true underlying image or the target in the case of black hole compressed sensing. The highest PSNR for each set of measurements is in bold. We fix the hyperparameters of the baselines across each forward model, empirically selecting them for good performance on each type of problem. Note that the range in performance depends on the choice of hyperparameters. For denoising, the DIP results corresponds to 100 and 300 epochs for the fewer and many epochs baselines, respectively, and the lower and higher λ 's are $1e3$ and $1e4$, respectively. For black hole compressed sensing, the DIP results corresponds to 1000 and 3000 epochs for the fewer and many epochs baselines, respectively, and the lower and higher λ 's are $1e2$ and $1e3$, respectively. Visual examples of celeb B. are shown in the supplemental materials.

In the Gaussian case, we note that our mean reconstructions are nearly identical to the true digits and the standard deviations exhibit uncertainty in regions we would expect (e.g., around edges). In the Fourier case, our reconstructions have features similar to the digit 8, but contain artifacts. These artifacts are only present in the Fourier case due to additional ambiguities, which lead to a more complex posterior [50]. We also show the average PSNR of our reconstructions for each measurement model in Table 4.2. For more details on this experiment, please see Section ?? of the supplemental materials.

Idealized black hole imaging

We consider a real-world inverse problem for which ground-truth data would be impossible to obtain. In particular, we consider a compressed sensing problem inspired by astronomical imaging of black holes with the Event Horizon Telescope (EHT): suppose we are given access to N measurement examples of the form $y^{(i)} = A^{(i)}x^{(i)} + \eta^{(i)}$ where $A^{(i)} \in \mathbb{C}^{m \times n}$ is a low-rank compressed sensing matrix arising from interferometric telescope measurements and $\eta^{(i)}$ denotes noise with known properties (e.g., distributed as a zero-mean Gaussian with known covariance). The collection of images $\{x^{(i)}\}_{i=1}^N$ are snapshots of an evolving black hole target. This problem is ill-posed and requires the use of priors or regularizers to recover a reasonable image [51]. Moreover, it is impossible to directly acquire example

	Measurement Operator(s)		
	Gaussian	Fourier	Both
Gaussian	30.8	—	30.2
Fourier	—	13.6	19.4

Table 4.2: **Phase Retrieval PSNRs for different measurement operators.** Each column corresponds to the type of measurements that the method was given (a total of $N = 150$). In the case of “Both”, 75 Gaussian measurements and 75 Fourier measurements were given. We then show the average PSNR for our reconstructions given the specific measurement operator (either Fourier or Gaussian). Note that when given both Gaussian and Fourier measurement examples our method exhibits a higher PSNR on the recovered images from Fourier examples than when given only Fourier measurements. Additionally, there is only a slight decrease in performance on the Gaussian measurement reconstructions as compared to given entirely Gaussian examples.

images of black holes, so any pixel-level prior defined *a priori* will exhibit human bias. Recovering an image, or movie, of a black hole with as little human bias as possible is essential for both studying the astrophysics of black holes as well as testing fundamental physics [52], [53]. We show how our proposed method can be used to tackle this important problem. In particular, we leverage knowledge that, although the black hole evolves, it will not change drastically from minute-to-minute or day-to-day. We study two black hole targets: the black hole at the center of the Messier 87 galaxy (M87*) and the black hole at the center of the Milky Way galaxy – Sagattarius A* (Sgr A*).

Imaging M87* using the current EHT array We first consider reconstructing the black hole at the center of the Messier 87 galaxy, which does not evolve noticeably within the timescale of a single day. The underlying images are from a simulated 60 frame video with a single frame for each day. We show results on $N = 60$ frames from an evolving black hole target with a diameter of ~ 40 microarcseconds, as was identified as the diameter of M87 according to [54], [55] in Fig. 4.10. In particular, the measurements are given by $\{y^{(i)} = Ax^{(i)} + \eta^{(i)}\}_{i=1}^N$, where $x^{(i)}$ is the underlying image on day i , A is the forward model that represents the telescope array, which is static across different days, and the noise $\eta^{(i)} \sim \mathcal{N}(0, \Sigma)$ has a covariance of Σ , which is a diagonal matrix with realistic variances based on the telescope properties¹. Measurements are simulated from black hole images with a realistic flux of 1 Jansky [56]. We also visualize a reference “target” image, which

¹We leave the more challenging atmospheric noise that appears in measurements for future work.

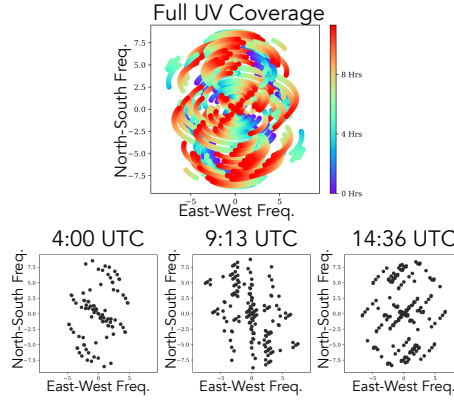


Figure 4.11: **Visualization of the EHT forward model of Sgr A* over time.** The spatial frequency space coordinates of Sgr A* from an ngEHT [57] array measured over a course of one night is shown on the top. Snapshots of the coordinates that are measured at different times are shown on the bottom. Note that the location and sparsity of the measurement coordinates change over time.

is the underlying image filtered with a low-pass filter that represents the maximum resolution achievable with the telescope array used to collect measurements – in this case the EHT array consisting of 11 telescopes (see Fig. 4.9 and Section ?? in the supplemental materials).

As seen in Fig. 4.10, our method is not only able to reconstruct the large scale features of the underlying image without any aliasing artifacts, but also achieves a level of super-resolution (see Table ?? in the supplemental materials). Our reconstructions also achieve higher super-resolution as compared to our baselines (i.e., AmbientGAN, TV-RML, and DPI) in Fig. 4.10 and do not exhibit artifacts evident in the reconstructions from these baselines. The two AmbientGAN settings were qualitatively chosen to show that the final result is sensitive to the choice of hyperparameters. The default AmbientGAN parameters produce poor results, and even with fine-tuning to best fit the underlying images (i.e., cheating with knowledge of the ground-truth), the results still exhibit substantial artifacts. We outperform the baselines in terms of PSNR when compared to the target image (see Table 4.1). Our results demonstrate that we are able to capture the varying temporal structure of the black hole, rather than just recovering a static image. It is important to note that there is no explicit temporal regularization introduced; the temporal smoothness is implicitly inferred by the constructed IGM.

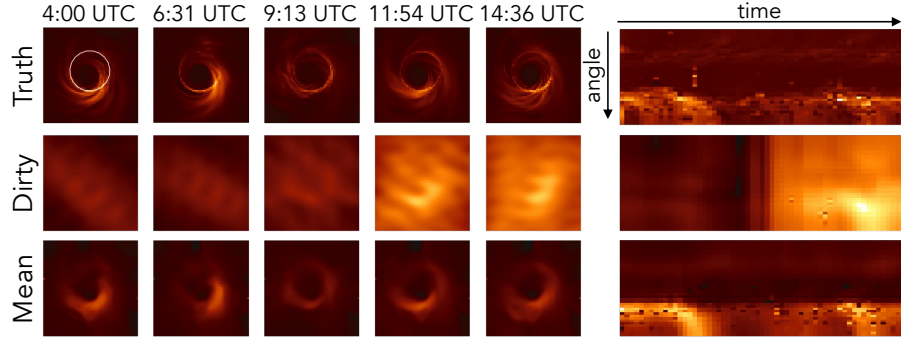


Figure 4.12: **Recovering a video of the Sgr A* black hole with a futuristic telescope array over the course of a night under a compressed sensing forward model.** We demonstrate our method described in 4.3 to perform video reconstruction on 60 frames of a video, where the measurements for each frame are generated by different forward models given by imaging Sgr A*. Here we show the underlying image, the dirty image ($A^H y$), and the mean of the reconstructed posterior. Additionally, we show the unwrapped space $\times$ time image, which is taken along the white ring illustrated in the 4:00 UTC underlying image. The bright-spot’s temporal trajectory of our reconstruction matches that of the truth. The measurement noise is consistent with a black hole having a flux of 2 Janskys.

Imaging Sgr A* from multiple forward models The framework we introduce can also be applied to situations in which the measurements themselves are induced by different forward models. In particular, the measurements $\{y^{(i)} = f^{(i)}(x^{(i)}) + \eta^{(i)}\}_{i=1}^N$ are given by an underlying image $x^{(i)}$, a forward model $f^{(i)}$ that is specific to that observation, and noise $\eta^{(i)}$ with known properties.

As an illustrative example, we consider the problem of reconstructing a video of the black hole at the center of the Milky Way – Sagittarius A* (Sgr A*). Unlike M87*, Sgr A* evolves on the order of minutes. Therefore, we can only consider that the black hole is static for only a short time when only a subset of telescopes are able to observe the black hole. This results in a different measurement forward model for each frame of the black hole “movie” [58]. In particular, the measurements are given by $\{y^{(i)} = A^{(i)}x^{(i)} + \eta^{(i)}\}_{i=1}^N$, where $x^{(i)}$ is the underlying image at time i , $A^{(i)}$ is the forward model that incorporates the telescope configuration at that time, and $\eta^{(i)} \sim \mathcal{N}(0, \Sigma^{(i)})$ is noise where $\Sigma^{(i)}$ is a diagonal matrix with realistic standard deviations derived from the telescopes’ physical properties. The measurement noise is consistent with a black hole with a flux of 2 Janskys. The measurement operator is illustrated in Fig. 4.11.

We show examples of reconstructing 60 frames of Sgr A* with a diameter of

~ 50 microarcseconds using measurements simulated from a proposed future next-generation EHT (ngEHT) [57] array, which consists of 23 telescopes. These results are shown in Fig. 4.12. Our reconstructions remove much of the aliasing artifacts evident in the dirty images and reconstructs the primary features of the underlying image without any form of temporal regularization. These results have high fidelity especially considering that the measurements are very sparsely sampled. Although these results come from simulated measurements that do not account for all forms of noise we expect to encounter, the high-quality movie reconstructions obtained without the use of a spatio-temporal prior show great promise towards scientific discovery that could come from a future telescope array paired with our proposed reconstruction approach.

Black hole imaging under atmospheric errors

In the previous section, we described a simpler black hole interferometry problem under no atmospheric noise, which allowed us to model the measurement generation process as a compressed sensing problem. In reality, the atmospheric error is not negligible. We consider a more realistic imaging problem where what we measure is a single 2D spatial Fourier frequency of the image x for each pair of telescopes a and b at time t . This measurement is called the *complex visibility* $F_{a,b}^t(x)$. This results in $\binom{S_t}{2}$ measurements for S_t observing telescopes at time t . For more details on how these measurements are acquired, see [59]. In Fig. 4.9, we visualize the measured frequency coverage, measurements of M87* for an EHT array with 11 telescopes, and the intrinsic resolution of the telescope array.

In reality, the complex visibility measurements made by the array of telescopes, such as the EHT, include different sources of noise. Specifically, the noisy measurements are characterized by $\Gamma_{a,b}^t = e^{i(\phi_a^t - \phi_b^t)} F_{a,b}^t(x) + \eta_{a,b}^t$, where $F_{a,b}^t$ is the ideal Fourier component measured by telescopes a, b at time t , $\eta_{a,b}^t$ is noise arising from Gaussian thermal noise [60], and ϕ_a and ϕ_b are phase errors arising from the inhomogeneity of the atmosphere [61]. These phase errors make the phase from raw visibility measurements unusable at mm and sub-mm wavelengths [61], rendering black hole imaging as a phase retrieval problem. However, when we consider a set of three telescopes a, b, c , in the triple product $\Gamma_{a,b}^t \Gamma_{b,c}^t \Gamma_{c,a}^t$, we have an identity property $e^{i(\phi_a - \phi_b)} e^{i(\phi_b - \phi_c)} e^{i(\phi_c - \phi_a)} = 1$. Hence, the product $\Gamma_{a,b} \Gamma_{b,c} \Gamma_{c,a}$, called the *bispectrum*, is invariant to atmospheric error [62]. This motivates the usage of the phase of the bispectrum, called the *closure phase*, as a constraint for the image reconstruction problem. Additionally, with calibration, the visibility amplitudes

$|\Gamma_{a,b}^t|$ can be well estimated [59], giving us a second set of constraints². Formally, we define our measurements for a single image as

$$\begin{aligned}
y &:= (y^{\text{amp.}}, y^{\text{clph.}}) \\
y^{\text{amp.}} &:= \{|\Gamma_{a,b}^t|\}_{(a,b) \in S_{2,t}^t} = \{|F_{a,b}^t(x)| + \eta_{a,b}^{\text{amp.},t}\}_{(a,b) \in S_{2,t}^t} \\
y^{\text{clph.}} &:= \{\angle(\Gamma_{a,b}^t \Gamma_{b,c}^t \Gamma_{c,a}^t)\}_{(a,b,c) \in S_{3,t}^t} \\
&= \{\angle(F_{a,b}^t(x) F_{b,c}^t(x) F_{c,a}^t(x)) + \eta_{a,b,c}^{\text{clph.},t}\}_{(a,b,c) \in S_{3,t}^t}
\end{aligned} \tag{4.6}$$

where a, b, c index telescopes, t is a time stamp from 0 to T , and $S_k^t = \binom{S^t}{k}$ ³. Following [59], [63] we treat the noise on $y^{\text{amp.}}$ and $y^{\text{clph.}}$ as Gaussian.

Modifications to objective: In the black hole phase-retrieval problem setting, each measurement y can be described as Eq. 4.6, which induces $\text{ELBOProxy}(G_\theta, q_{\phi^{(i)}}; \{y^{\text{amp.},(i)}, y^{\text{clph.},(i)}\})$. To combine these into a single objective, we control the relative strength between the visibility amplitude and closure phase data-fits with a hyperparameter α , resulting in the updated data-fit $\log p(y|G(z))$ from Eq. 4.3:

$$\log p(y|G(z)) = \log p(y^{\text{clph.}}|G_\theta(z)) + \alpha \log p(y^{\text{amp.}}|G_\theta(z))$$

Since phase retrieval problems have intrinsic phase ambiguities, spatial shifts and flips are possible reconstructions. Closure phases remove the flip ambiguity, but the spatial shift ambiguity still remains. Modeling such a multi-modal distribution is challenging, so we introduce a centering loss term to help with the optimization. The center loss is defined by $\mathcal{L}_{\text{center}} := \frac{1}{2N} \sum_{i=1}^N |(\text{Center}(x^{(i)}) - \text{COM}(x^{(i)}))|^2$ where $\text{Center}(x)$ ⁴ and $\text{COM}(x)$ ⁵ are the center point and the center of mass of the image x respectively. We use the hyperparameter β to control the strength of the centering loss, which we anneal from β to 0 as a function of epoch k (ε_k). Thus our final optimization objective is

$$\{\hat{\theta}, \hat{\phi}^{(1)}, \dots, \hat{\phi}^{(N)}\} = \arg \min_{\theta, \{\phi^{(i)}\}_{i=1}^N} \{\mathcal{L}_{\text{ELBO}} + \beta \varepsilon_k \mathcal{L}_{\text{center}}\}.$$

²Although the amplitudes can be largely calibrated, some error typically remains. As phase errors are much more challenging, we chose to make the simplifying assumption of fully calibrated amplitudes in this work.

³It is unnecessary to use all $|S_3^t|$ closure phase measurements. We use the minimum set S_3^t such that the set of all telescopes $S_t = \cup_{S_i \in S_3^t} S_i$.

⁴ $\text{Center}(x) := (\frac{1}{2}(M+1), \frac{1}{2}(M+1))$

⁵ $\text{COM}(x) := \frac{1}{\sum_{m,n=1}^M x_{mn}} \sum_{m,n=1}^M x_{mn}(m, n)$

Once the parameters have been inferred, $\hat{x}^{(i)}$ is found by sampling $\hat{z}^{(i)} \sim q_{\hat{\phi}^{(i)}}(z^{(i)})$ and computing $\hat{x}^{(i)} = G_{\theta}(\hat{z}^{(i)})$.

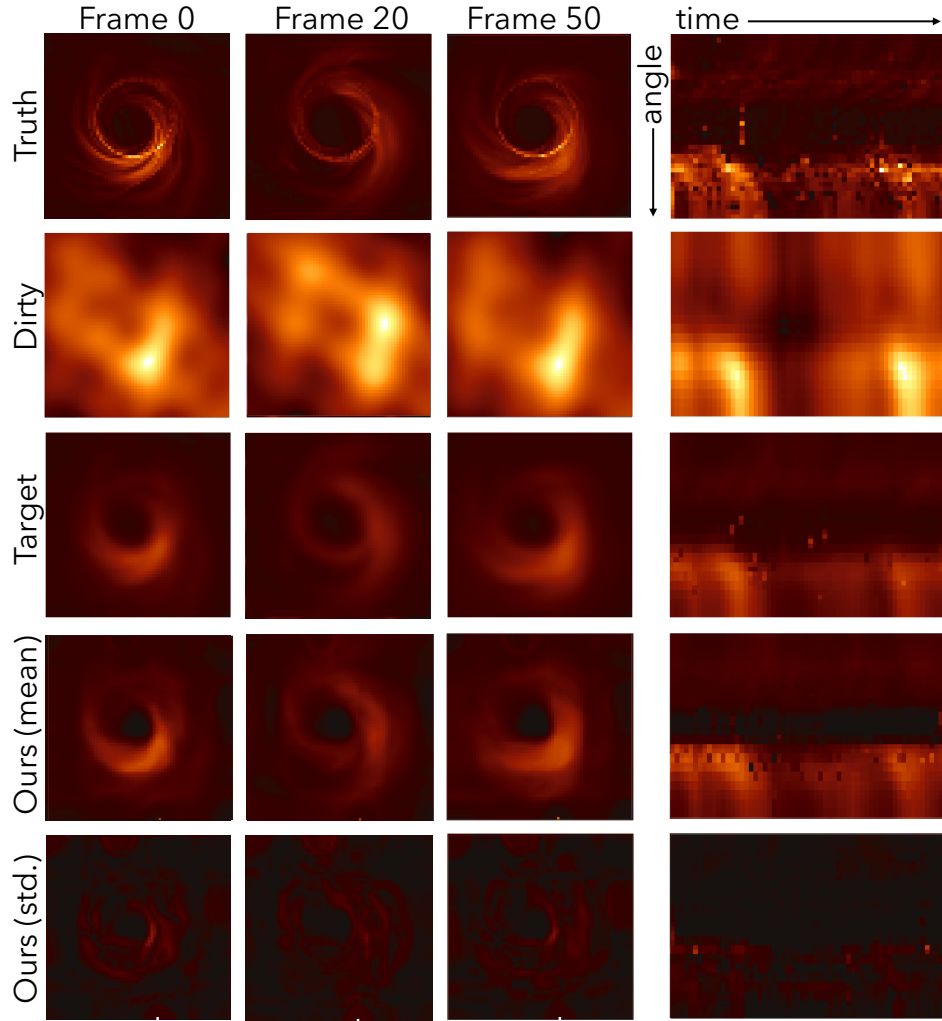


Figure 4.13: **Our reconstructions of selected frames from a video of M87*.** We show the ground truth, dirty image, target image (see Fig. 4.9), and empirical mean and standard deviation of our reconstructed image distribution for selected frames from the 60-frame M87* video. Additionally, we show the unwrapped space vs. time image, which is taken counter-clockwise along the ring; the trajectory of the bright spot in our reconstruction matches the true image. Our method reconstructs the primary features of the true underlying images while also reconstructing some high frequency features.

Imaging M87* using the current EHT array under atmosphere We show results of our reconstruction method on selected frames in Fig. 4.13. The target image is computed by blurring the ground-truth image to the intrinsic resolution of the EHT telescope array ($\sim 25 \mu\text{as}$), as shown in Fig. 4.9.d. The time $\times$ angle plots, which visualize the temporal trajectory of the ring, are created by plotting

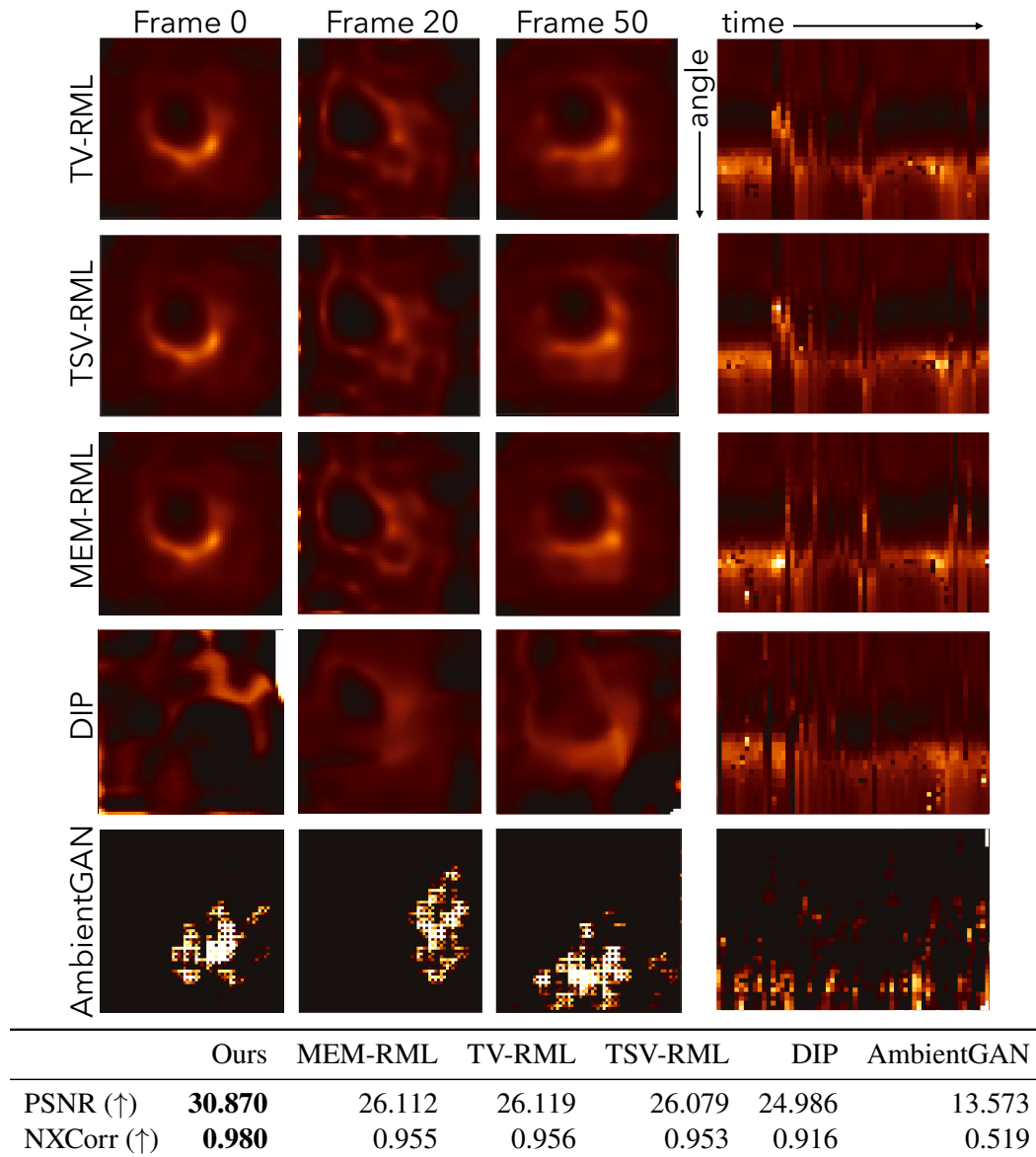


Figure 4.14: **Baseline comparisons.** We show results from the following baseline methods: regularized maximum likelihood with maximum entropy (MEM-RML), total variation (TV-RML), and total-squared-variation (TSV-RML), Deep Image Prior (DIP) [48], and AmbientGAN [35]. We show the average peak signal to noise ratio (PSNR) and normalized cross correlation (NXCorr) compared to the target images (see Fig. 4.13). Our method exhibits fewer artifacts and has the highest average PSNR and NXCorr, although PSNR is not an ideal metric due to the shift ambiguities in phase retrieval.

the intensity counter-clockwise for each of the 60 frames. Since we reconstruct not just a single image but an image distribution, we show images of the empirical mean and standard deviation. Our reconstructions are visually similar to the target and accurately reconstruct the primary features while reconstructing some high-

frequency features. Additionally, the time $\times$ angle plots of our reconstructions are similar to that of the target, indicating that our reconstruction accurately captures the dynamics of the black hole even without any temporal regularization. We find that our reconstructions best match the ground truth image at a $10 \mu\text{as}$ resolution, substantially smaller than the intrinsic $25 \mu\text{as}$ resolution of the telescope, implying that our approach also achieves a level of superresolution.

Baseline Comparisons We further quantify the quality of the reconstruction by comparing our reconstruction to baseline methods in Fig. 6.3. We used the official EHT published code in `eht-imaging` [64] to produce the regularized maximum likelihood (RML) baselines with the following regularizers: maximum entropy (MEM-RML) [65], total variation (TV-RML) [66], and total squared variation (TSV-RML) [67]. We also include the following baselines: 1) Deep Image Prior (DIP) [48], which uses a deep implicit prior that we modify to have a centering loss and 2) AmbientGAN [35], which learns the underlying data distribution through a generative model directly from noisy measurements.

We show the peak signal-to-noise ratio (PSNR) and normalized cross-correlation (NXCorr) for each method compared to the target image. Our method exhibits the highest PSNR and NXCorr, is visually more accurate in reconstructing the image features, and exhibits less artifacts than the other reconstruction methods. Unlike the RML baselines, our method is able to reconstruct the dynamics of the spiral structure in Frame 20. Moreover, our results have temporal consistency, substantially outperforming DIP even when using the same centering loss throughout DIP’s inference.

For our forward model dependent hyperparameters, we find that the choice of α has a substantial impact on the data-fit while the reconstructions are less sensitive to β since it is annealed quickly during optimization.

4.5 Theory for linear IGMs

We now introduce theoretical results on the inferred IGM for linear inverse problems. Specifically, we consider the case when the IGM G_θ is linear and the latent variational distributions are Gaussian. The goal of this section is to develop intuition for the inferred IGM in a simpler setting. While our results may not generalize to non-linear generators parameterized by deep neural networks, our results aim to provide an initial understanding on the inferred generator.

More concretely, suppose we are given N measurement examples of noisy linear

measurements of the form

$$y^{(i)} := Ax^{(i)} + \eta^{(i)}, \quad \eta^{(i)} \sim \mathcal{N}(0, \sigma^2 I)$$

where $A \in \mathbb{R}^{m \times n}$ with $m \leq n$ and $x^{(i)} \in \mathbb{R}^n$. We aim to infer $G_\theta \in \mathbb{R}^{n \times k}$ with $k < n$ and latent distributions $q_{\phi^{(i)}} = \mathcal{N}(\mu_i, U_i U_i^T)$ where $\phi^{(i)} = \{\mu_i, U_i\}$, $\mu_i \in \mathbb{R}^k$, and $U_i \in \mathbb{R}^{k \times k}$ by minimizing the negative ELBOProxy:

$$\begin{aligned} \mathcal{L}(G_\theta, \{\phi^{(i)}\}) &:= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{z \sim q_{\phi^{(i)}}(z)} [-\log p(y^{(i)} | G_\theta(z)) \\ &\quad - \log p_Z(z | G_\theta) + \log q_{\phi^{(i)}}(z)]. \end{aligned} \quad (4.7)$$

We characterize the generator G_* and latent parameters $\phi_*^{(i)} = \{\mu_i^*, U_i^*\}$ that are stationary points of (4.7). The result is proven in Section ??:

Theorem 2. Fix $\sigma > 0$ and let $y^{(i)} = Ax^{(i)} + \eta^{(i)} \in \mathbb{R}^n$ for $i \in [N]$ where $\eta^{(i)} \sim \mathcal{N}(0, \sigma^2 I)$ and $A \in \mathbb{R}^{m \times n}$ has $k \leq \text{rank}(A) = m \leq n$. Define the sample covariance of the measurements $Y_N := \frac{1}{N} \sum_{i=1}^N y^{(i)} (y^{(i)})^T$. Then, with probability 1, G_* , μ_i^* , and U_i^* that satisfy

$$\begin{aligned} AG_* &:= E_k (L_k - \sigma^2 I)^{1/2} R, \\ \mu_i^* &:= (G_*^T A^T A G_* + \sigma^2 I)^{-1} G_*^T A^T y^{(i)}, \text{ and} \\ U_i^* &:= \sigma (G_*^T A^T A G_* + \sigma^2 I)^{-1/2} \tilde{R}, \end{aligned}$$

for $i \in [N]$ are stationary points of the objective (4.7) over $\mathbb{R}^{n \times k} \times (\mathbb{R}^k \times \text{GL}(k, \mathbb{R}))^N$ where $\text{GL}(k, \mathbb{R})$ denotes the set of real invertible $k \times k$ matrices. Here, $E_k \in \mathbb{R}^{m \times k}$ denotes the matrix whose columns correspond to the top- $\min\{\text{rank}(Y_N), k\}$ eigenvectors of Y_N , $L_k \in \mathbb{R}^{k \times k}$ contains the corresponding eigenvalues of Y_N , and $R, \tilde{R} \in \mathbb{R}^{k \times k}$ are arbitrary orthogonal matrices. If $\text{rank}(Y_N) < k$, then for $\text{rank}(Y_N) < i \leq k$, the i -th column of E_k can be arbitrary and $L_{k,ii} = \sigma^2$.

The Theorem establishes the precise form of stationary points of the objective (4.7). In particular, it shows that this inferred IGM performs dimensionality reduction akin to PCA [68] on the collection of measurement examples. To gain further intuition about the Theorem, in Section ?? of the supplemental materials, we analyze this result in a context where the underlying images we wish to reconstruct explicitly lie in a low-dimensional subspace. We show that, in that setting, our estimator returns an approximation of the solution found via MAP estimation, which can only be

computed with complete prior knowledge of the underlying image structure. While the Theorem focused on linear IGMs, it would be interesting to theoretically analyze non-linear IGMs parameterized by deep networks. We leave this for future work.

4.6 Conclusion

In this work we showcased how one can solve a set of inverse problems without a pre-defined IGM (e.g., a traditional spatial image prior) by leveraging common structure present across a collection of diverse underlying images. We demonstrated that even with a small set of corrupted measurements, one can jointly solve these inverse problems by directly inferring an IGM that maximizes a proxy of the ELBO. We demonstrate our method on a number of convex and non-convex imaging problems, including the challenging problem of black hole video reconstruction from interferometric measurements. Overall, our work showcases the possibility of solving inverse problems in a completely unsupervised fashion, free from significant human bias typical of ill-posed image reconstruction. We believe our approach can aid in automatic discovery of novel structure from scientific measurements, potentially paving the way to new avenues of exploration.

References

- [1] Oscar Leong, Angela F. Gao, He Sun, and Katherine L. Bouman, “Discovering structure from corruption for unsupervised image reconstruction,” *IEEE Transactions on Computational Imaging*, vol. 9, pp. 992–1005, 2023. doi: <https://doi.org/10.1109/TCI.2023.3325752>,
- [2] A. F. Gao, O. Leong, H. Sun, and K. L. Bouman, “Image reconstruction without explicit priors,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.
- [3] Yvette Y. Lin, Angela F. Gao, and Katherine L. Bouman, “Imaging an evolving black hole by leveraging shared structure,” *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2475–2479, 2024. doi: <https://doi.org/10.1109/ICASSP48485.2024.10445896>,
- [4] H. C. Burger, C. J. Schuler, and S. Harmeling, “Image denoising: Can plain neural networks compete with bm3d?” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [5] L. Dyer, A. Parker, K. Paphiti, and J. Sanderson, “Lightsheet microscopy,” *Current Protocols*, vol. 2, no. 7, e448, 2022.

- [6] E. J. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *Communications on Pure and Applied Mathematics*, vol. 67(6), pp. 906–956, 2013.
- [7] E. J. Candès, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure and Applied Mathematics*, vol. 59(8), pp. 1207–1223, 2006.
- [8] D. Donoho, “For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics*, vol. 59(6), 2006.
- [9] A. Fannjiang and T. Strohmer, “The numerics of phase retrieval,” *Acta Numerica*, vol. 29, pp. 125–228, 2020.
- [10] X. Li, S. Ling, T. Strohmer, and K. Wei, “Rapid, robust, and reliable blind deconvolution via nonconvex optimization,” *arXiv preprint*, vol. arXiv:1606.04933, 2016.
- [11] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (red),” *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.
- [12] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, “Plug-and-play priors for model based reconstruction,” in *2013 IEEE Global Conference on Signal and Information Processing*, IEEE, 2013, pp. 945–948.
- [13] A. Bora, A. Jalal, E. Price, and A. Dimakis, “Compressed sensing using generative models,” *International Conference on Machine Learning (ICML)*, 2017.
- [14] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259–268, 1992.
- [15] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.
- [16] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Understanding and evaluating blind deconvolution algorithms,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1971, 2009.
- [17] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.
- [18] J. Miles, “R squared, adjusted r squared,” *Encyclopedia of Statistics in Behavioral Science*, 2005.
- [19] G. E. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6(2), pp. 461–464, 1978.
- [20] H. Akaike, “A new look at the statistical model identification,” *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

- [21] B.-E. Chérif-Abdellatif, “Consistency of elbo maximization for model selection,” *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, vol. 96, pp. 1–21, 2018.
- [22] B.-E. Chérif-Abdellatif and P. Alquier, “Consistency of variational bayes inference for estimation and model selection in mixtures,” *Electronic Journal of Statistics*, vol. 12(2), pp. 2995–3035, 2018.
- [23] C. Tao, L. Chen, R. Zhang, R. Henao, and L. Carin, “Variational inference and model selection with generalized evidence bounds,” *International Conference on Machine Learning (ICML)*, vol. 80, pp. 893–902, 2018.
- [24] H. Sun, K. L. Bouman, P. Tiede, J. J. Wang, S. Blunt, and D. Mawet, “Alpha-deep probabilistic inference (alpha-dpi): Efficient uncertainty quantification from exoplanet astrometry to black hole feature extraction,” *The Astrophysical Journal*, vol. 932(2), p. 99, 2022.
- [25] R. Heckel, W. Huang, P. Hand, and V. Voroninski, “Rate-optimal denoising with deep neural networks,” *Information and Inference: A Journal of the IMA*, 2020.
- [26] S. Menon, A. Damian, M. Hu, N. Ravi, and C. Rudin, “Pulse: Self-supervised photo upsampling via latent space exploration of generative models,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.
- [27] M. Alley, J. Y. Cheng, W. Dally, *et al.*, “Deep generative adversarial networks for compressed sensing (gancs) automates mri,” *arXiv preprint arXiv:1706.00051*, 2017.
- [28] Y. Song, L. Shen, L. Xing, and S. Ermon, “Solving inverse problems in medical imaging with score-based generative models,” *International Conference on Learning Representations (ICLR)*, 2022.
- [29] P. Hand, O. Leong, and V. Voroninski, “Phase retrieval under a generative prior,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [30] F. Shamshad and A. Ahmed, “Compressed sensing-based robust phase retrieval via deep generative priors,” *IEEE Sensors Journal*, vol. 21(2), pp. 2286–2298, 2017.
- [31] J. Lehtinen, J. Munkberg, J. Hasselgren, *et al.*, “Noise2noise: Learning image restoration without clean data,” *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 2965–2974, 2018.
- [32] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, “Rare: Image reconstruction using deep priors learned without groundtruth,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1088–1099, 2020.
- [33] A. Krull, T.-O. Buchholz, and F. Jug, “Noise2void - learning denoising from single noisy images,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019.

- [34] J. Batson and L. Royer, “Noise2self: Blind denoising by self-supervision,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 524–533.
- [35] A. Bora, E. Price, and A. G. Dimakis, “Ambientgan: Generative models from lossy measurements,” in *International conference on learning representations*, 2018.
- [36] M. Kabkab, P. Samangouei, and R. Chellappa, “Task-aware compressed sensing with generative adversarial networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [37] A. Olsen, Y. Hu, and V. Ganapati, “Data-driven computational imaging for scientific discovery,” in *NeurIPS AI for Science Workshop*, 2022.
- [38] R. Mendoza, M. Nguyen, J. W. Zhu, *et al.*, “A self-supervised approach to reconstruction in sparse x-ray computed tomography,” in *NeurIPS Machine Learning and the Physical Sciences Workshop*, 2022.
- [39] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” *International Journal of Computer Vision*, vol. 128, pp. 1867–1888, 2020.
- [40] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, pp. 234–241, 2015.
- [41] R. Heckel and P. Hand, “Deep decoder: Concise image representations from untrained non-convolutional networks,” *International Conference on Learning Representations (ICLR)*, 2019.
- [42] M. Z. Darestani and R. Heckel, “Accelerated mri with un-trained neural networks,” *IEEE Transactions of Computational Imaging*, vol. 7, pp. 724–733, 2021.
- [43] Z. Cheng, M. Gadelha, S. Maji, and D. Sheldon, “A bayesian perspective on the deep image prior,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5438–5446, 2019.
- [44] K. Kothari, A. Khorashadizadeh, M. de Hoop, and I. Dokmanić, “Trumpets: Injective flows for inference and inverse problems,” in *Uncertainty in Artificial Intelligence*, PMLR, 2021, pp. 1269–1278.
- [45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86(11), pp. 2278–2324, 1998.
- [46] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *arXiv preprint arXiv:1605.08803*, 2016.
- [47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

- [48] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.
- [49] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and simile classifiers for face verification,” in *2009 IEEE 12th international conference on computer vision*, IEEE, 2009, pp. 365–372.
- [50] K. Jaganathan, S. Oymak, and B. Hassibi, “Sparse phase retrieval: Convex algorithms and limitations,” *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 1022–1026, 2013.
- [51] T. E. H. T. Collaboration, “First m87 event horizon telescope results. iv. imaging the central supermassive black hole,” *The Astrophysical Journal Letters*, vol. 875, no. 1, p. L4, 2019.
- [52] T. E. H. T. Collaboration, “First m87 event horizon telescope results. i. the shadow of the supermassive black hole,” *The Astrophysical Journal Letters*, vol. 875, no. 1, p. L1, Apr. 2019. doi: 10.3847/2041-8213/ab0ec7.
- [53] T. E. H. T. Collaboration, “First sagittarius a* event horizon telescope results. i. the shadow of the supermassive black hole in the center of the milky way,” *The Astrophysical Journal Letters*, vol. 930, no. 2, p. L12, May 2022. doi: 10.3847/2041-8213/ac6674.
- [54] O. Porth, K. Chatterjee, R. Narayan, *et al.*, “The event horizon general relativistic magnetohydrodynamic code comparison project,” *The Astrophysical Journal Supplement Series*, vol. 243, no. 2, p. 26, 2019.
- [55] T. E. H. T. Collaboration, “First m87 event horizon telescope results. v. physical origin of the asymmetric ring,” *The Astrophysical Journal Letters*, vol. 875, no. 1, p. L5, 2019.
- [56] G. N. Wong, B. S. Prather, V. Dhruv, *et al.*, “PATOKA: Simulating Electromagnetic Observables of Black Hole Accretion,” *The Astrophysical Journal Supplement Series*, vol. 259, no. 2, p. 64, Apr. 2022. doi: 10.3847/1538-4365/ac582e. arXiv: 2202.11721 [astro-ph.HE].
- [57] A. W. Raymond, D. Palumbo, S. N. Paine, *et al.*, “Evaluation of new submillimeter vlbi sites for the event horizon telescope,” *The Astrophysical Journal Supplement Series*, vol. 253, no. 1, p. 5, 2021.
- [58] T. E. H. T. Collaboration, “First sagittarius a* event horizon telescope results. iii. imaging of the galactic center supermassive black hole,” *The Astrophysical Journal Letters*, vol. 930, no. 2, p. L14, 2022.
- [59] A. Thompson, J. Moran, and G. Swenson Jr., *Interferometry and Synthesis in Radio Astronomy, 3rd Edition*. Springer Cham, 2017. doi: 10.1007/978-3-319-44431-4.
- [60] *Synthesis Imaging in Radio Astronomy II*, vol. 180, Astronomical Society of the Pacific Conference Series, Jan. 1999.

- [61] J. D. Monnier and R. J. Allen, “Radio and optical interferometry: Basic observing techniques and data analysis,” in *Planets, Stars and Stellar Systems: Volume 2: Astronomical Techniques, Software, and Data*, T. D. Oswalt and H. E. Bond, Eds. Dordrecht: Springer Netherlands, 2013, pp. 325–373, ISBN: 978-94-007-5618-2. DOI: 10.1007/978-94-007-5618-2_7. [Online]. Available: https://doi.org/10.1007/978-94-007-5618-2_7.
- [62] R. Jennison, “A phase sensitive interferometer technique for the measurement of the Fourier transforms of spatial brightness distributions of small angular extent,” *MNRAS*, vol. 118, p. 276, Jan. 1958. DOI: 10.1093/mnras/118.3.276.
- [63] K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doeleman, and W. T. Freeman, *Computational imaging for VLBI image reconstruction*, 2016. arXiv: 1512.01413 [astro-ph.IM].
- [64] A. A. Chael, M. D. Johnson, K. L. Bouman, L. L. Blackburn, K. Akiyama, and R. Narayan, “Interferometric Imaging Directly with Closure Phases and Closure Amplitudes,” *The Astrophysical Journal*, vol. 857, no. 1, 23, p. 23, Apr. 2018. DOI: 10.3847/1538-4357/aab6a8. arXiv: 1803.07088.
- [65] J. Skilling and R. K. Bryan, “Maximum entropy image reconstruction: general algorithm,” *Monthly Notices of the Royal Astronomical Society*, vol. 211, no. 1, pp. 111–124, Nov. 1984, ISSN: 0035-8711. DOI: 10.1093/mnras/211.1.111. [Online]. Available: <https://doi.org/10.1093/mnras/211.1.111>.
- [66] C. Bouman and K. Sauer, “A generalized Gaussian image model for edge-preserving MAP estimation,” *IEEE Transactions on Image Processing*, vol. 2, no. 3, pp. 296–310, 1993. DOI: 10.1109/83.236536.
- [67] K. Kuramochi, K. Akiyama, S. Ikeda, *et al.*, “Superresolution interferometric imaging with sparse modeling using total squared variation: Application to imaging the black hole shadow,” *The Astrophysical Journal*, vol. 858, no. 1, p. 56, 2018.
- [68] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61(3), pp. 611–622, 1999.

HIGH RESOLUTION EIKONAL IMAGING OF KILAUEA

Abstract

Images of the Earth’s interior can provide us with insight into the underlying properties of the Earth, such as how seismic activity might emerge and the interplay between seismic and volcanic activity. Understanding these systems requires reliable high-resolution images to understand mechanisms and estimate physical quantities. However, reliable images are often difficult to obtain due to the non-linear nature of seismic wave propagation and the ill-posedness of the related inverse problem. Reconstructions rely on good initial estimates as well as hand-crafted priors, which can ultimately bias solutions. In our work, we present a 3D reconstruction of Kilauea’s magmatic system at a previously unattained resolution. Our eikonal tomography procedure improves upon prior imaging results of Kilauea through increased resolution and per-pixel uncertainties estimated through variational inference. In particular, solving eikonal imaging using variational inference with stochastic gradient descent enables stable inversion and uncertainty quantification in the absence of strong prior knowledge of the velocity structure. Our work makes two key contributions: developing a stochastic eikonal tomography scheme with uncertainty quantification and illuminating the structure and melt quantity of the magmatic system that underlies Kilauea.

5.1 Introduction

Kilauea volcano in Hawai‘i is one of the most extensively monitored volcanoes on Earth. Since the onset of the decades-long Pu‘u ‘O‘o eruption in 1983, seismological, geodetic, and gravity investigations into the sustained eruptive activity at Kilauea have provided a rich understanding of the location of the volcano’s near-surface melt reservoirs, rift zone structure, and dike pathways and dynamics [1]–[4]. The summit of Kilauea is an approximately 3 km by 4 km caldera underlain by two magma chambers at depths of 1-2 and 3-5 km [5]. Two rift zones, the Southwest Rift Zone and the East Rift Zone, extend from Kilauea’s summit. Intrusive dikes originating at the summit travel through either of these two rift zones; the dike associated with the 2018 caldera collapse episode at Kilauea traveled tens of km down the East Rift Zone, breaching the surface at multiple locations [5]. The East Rift Zone

is underlain by a décollement at 9-12 km depth that marks the boundary between the Hawaiian volcanic pile and the paleoseafloor [6]; the décollement is capable of hosting large ($M \sim 7$) earthquakes [7] as well as slow slip [8]. Because of Kilauea's frequent eruptive activity and dense geophysical monitoring network, pre-eruptive patterns of deformation and seismicity originating within the upper kilometers of the subsurface can often be interpreted to assess the volcano's immediate eruptive potential [5], [8]–[10].

Our understanding of Kilauea's longer-term eruption cycle behavior, however, may be limited by our understanding of the lower-crustal and mantle structures that supply magma to the volcano. Kilauea is supplied with magma from a zone of melting at 80-100 km depth proximal to the Hawaiian hotspot [6]. Recent work has suggested that structures within Kilauea's mantle plumbing system can have an important influence on short-term fluctuations in magma supply rate to Kilauea, which may in turn influence its eruptive behavior on timescales of weeks to years [11]. During an observed “magma surge” episode in 2003-2007, the rate of magma supply to Kilauea increased by a factor of roughly two [12]; other work has suggested that such surges may be enabled by feedback loops between tectonic and magmatic processes within mantle magma-bearing structures [11]. Furthermore, temporal correlations between eruptions at Kilauea and earthquake swarms at >30 km depth have been interpreted to evidence an “open” fluid connection between the surface and the mantle through which pressure gradients can rapidly propagate [13]. While density anomalies constrained by gravity modeling [14] and concentrations of long-period seismicity [6], [15] suggest the existence of a columnar magma conduit extending downwards from Kilauea's shallow reservoirs, the extent, geometry, and compositional melt fraction of this structure is still unknown. Improving our understanding of the mid-to-lower crustal structures that supply magma to Kilauea would allow us to better understand the features that enable short-term magma surges to the volcano, as well as the correlations between surface unrest and mantle seismicity.

Numerous seismic tomography models have been developed to study subsurface structure at Kilauea [16]–[21] but there is still significant uncertainty about the overall magma storage beneath the Hawai'i volcanic system. Prior models typically resolve features along Kilauea's rift zones, corresponding to basaltic dikes, as well as anomalies at depths of 1-12 km beneath the caldera that are usually interpreted as corresponding to the deposition of ultramafic cumulates and a magma column beneath the summit [22], [23]. As the amount of high-quality first-arrival pick

data collected at Kilauea has grown, the resolution of resolved models has steadily increased (e.g., [21]). However, more detailed conclusions from such models are often hindered by challenges associated with the tomographic inverse problem.

It is well known that seismic tomography is ill-posed and non-unique [24], and these properties result from several factors including the sparsity of noisy observations and the non-linearity of the forward problem. Consequently, inversion results are sensitive to the initial values and rely on strong regularization (e.g., smoothing)[24]. Seismic imaging studies typically utilize dampening as a form of regularization, which relies on the assumption that the initial 1-D velocity model captures the first-order structure of the underlying region of interest (e.g., [25]). However, an incomplete or inaccurate initialization can cause the inversion to be trapped in local minima due to the non-linearity of the optimization problem; thus, solving for a single reconstruction may provide a limited or biased view of the structures that are actually constrained by the available travel-time data [26].

The uncharacterized uncertainty of the velocity, which results from the non-linear and non-unique tomography problem, limits our ability to draw robust inferences about the geometry and melt fraction of melt storage regions. For example, existing tomographic models of Kilauea generally resolve a single, undifferentiated velocity anomaly that is interpreted as a melt storage region below the caldera surface, although evidence from geodetic studies suggests the presence of multiple, distinct chambers [3], [12]. Without quantitative information about model uncertainty, it is difficult to assess these models' ability to differentiate between a single spatially distributed melt storage region or multiple neighboring chambers. Furthermore, placing confidence bounds on melt fraction values calculated from velocity values within imaged melt storage regions (e.g., [27]) requires an estimate of uncertainty in the model parameters, which is not available in typical tomographic workflows.

In this study, we develop an eikonal tomography procedure and apply it to produce a high-resolution tomographic image of the crust beneath the Kilauea summit. Our technique leverages advances in statistical inference to produce first-order uncertainty estimates through variational inference on recovered tomographic models. In particular, we use stochastic optimization through Monte Carlo sampling, which enables stable inversion in settings with weak or no priors. We validate our technique on realistic test cases and then apply our technique to travel-time data recorded at Kilauea volcano and identify a distinctly high- V_P/V_S region beneath the summit that we interpret as the magma-bearing column connecting the surface reservoirs

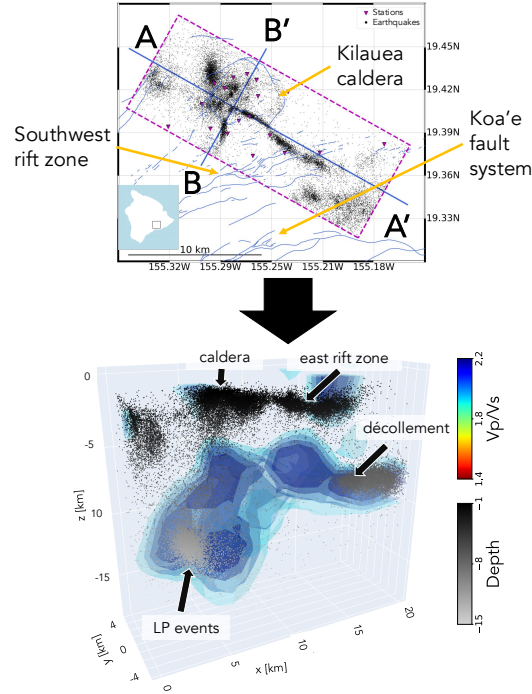


Figure 5.1: Source-receiver geometry and recovered 3D magmatic system from a subdomain of the Island of Hawai'i. For our real data study and realistic-geometry synthetic tests, we consider the sources and receivers in a small region (magenta dashed rectangle) of the Island of Hawai'i that includes the Kilauea caldera. The black rectangle in the inset map shows the location of the area plotted in the main plotted map. We use a total of 18 receivers (magenta triangles) and 45,310 earthquake events (dots colored by depth). These events were located using the method described in [28], but with modifications discussed in Sec. 5.5. We visualize the cross sections A-A' and B-B' used for later figures (blue lines). The bottom figure depicts our reconstruction of the 3D magmatic system that underlies Kilauea with long-period (LP) volcanic events are highlighted and located at the bottom of a large magmatic volume.

to magma supply from the mantle. We take advantage of our derived uncertainties to estimate the size of this column and constrain a melt fraction of roughly 15% within this column. Our results suggest the existence of persistent melt structures connecting Kilauea to the mantle that may enable feedback between the surface and deeper parts of the magma plumbing system.

5.2 Methodological Background

Travel time tomography (TTT)

Traditionally, seismic tomography using travel times has been solved using ray-based tomography techniques [25], [29], [30]. These methods aim to solve for a heterogeneous velocity model through modeling the ray trajectory, which are curved by the heterogeneous velocity model. However, continuously curved ray tracing is expensive, so in practice, ray-based tomography often uses pseudo-ray bending [31], which relaxes the curved ray tracing problem into a piecewise linear approximation to improve efficiency. Ray-theory-based tomography techniques [25], [29] are often very efficient, but they are only an approximation of continuous ray tracing, which is a weak approximation of the wave equation.

Imaging using the eikonal equation has recently become possible due to improvements in computation. The eikonal equation, which is an infinite frequency approximation of the wave equation, underlies the arrival times of earthquakes and governs the travel time along the shortest path between two points [32]. Let the travel time for a seismic wave be defined as T induced by velocity V . The eikonal equation relates the travel time to velocity and is defined by

$$\|\nabla T_{\mathbf{x}_s}(\mathbf{x})\|^2 = \left(\frac{1}{V(\mathbf{x})} \right)^2 \quad \text{with the boundary condition} \quad T_{\mathbf{x}_s}(\mathbf{x}_s) = 0 \quad (5.1)$$

where $T_{\mathbf{x}_s}(\mathbf{x}) \in C^1(\mathbb{R}^{2d})$ is the travel time field between source location $\mathbf{x}_s \in \mathbb{R}^d$ and receiver location $\mathbf{x} \in \mathbb{R}^d$ induced by the velocity model V and $V(\mathbf{x})$ is the velocity at a location $\mathbf{x}$. In other words, the eikonal equation states that the velocity $V(\mathbf{x})$ at a location $\mathbf{x}$ is inversely proportional to the norm of the gradient of the travel time field $T_{\mathbf{x}_s}(\mathbf{x})$ with respect to that location $\mathbf{x}$. For an earthquake's P and S waves, the associated arrival times are induced by separate velocity models, which we denote as V_P and V_S respectively.

It has become relatively common in recent years to solve seismic imaging problems with Bayesian inference [33], [34]. Here, prior knowledge about the velocity can be incorporated when available to help stabilize the inverse problem like in [35]. In the simplest setting, one can solve for the minimizer of the *maximum a posteriori* (MAP) objective. In particular, the MAP approach for TTT aims to find the velocity model V (e.g., V_P or V_S) that maximizes the following objective given travel time measurements y (e.g., P wave or S wave travel times)

$$\hat{V} = \arg \max_V \underbrace{\log p(y|V)}_{\text{log data likelihood}} + \underbrace{\log p(V)}_{\text{log prior}}. \quad (5.2)$$

Often, the travel time measurement from sources $\{\mathbf{x}_{s_i}\}_{i \in \mathcal{S}}$ and receivers $\{\mathbf{x}_{r_j}\}_{j \in \mathcal{R}}$ induced by a velocity model V are assumed to be of the form

$$\{y_{ij} = T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) + \varepsilon_{ij}\}_{i \in \mathcal{S}, j \in \mathcal{R}} \quad (5.3)$$

where $\mathcal{S}$ and $\mathcal{R}$ represent the set of sources and receivers respectively. The measurement noise ε_{ij} is commonly assumed to follow a Gaussian distribution $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij})$. This measurement noise model produces the data likelihood model $y_{ij} \sim \mathcal{N}(T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}), \sigma_{ij})$, which results in the log data likelihood, $\log p(y|V)$, taking the form of a weighed sum of squared error (SSE):

$$\hat{V} = \arg \max_V \underbrace{\sum_{ij} \frac{1}{\sigma_{ij}^2} (y_{ij} - T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}))^2}_{\text{log data likelihood}} + \underbrace{\log p(V)}_{\text{log prior}}. \quad (5.4)$$

Eikonal-based travel time calculations can be advantageous to those derived from traditional ray-tracing due to their greater stability in the presence of strong velocity heterogeneity [34], [36]–[41]. Historically, however, eikonal TTT has been less popular; this may be due to various factors including particular challenges of solving the eikonal equation itself [32], [42], added computation time and memory, and the limited availability of strategies for solving these inverse problems. Recently [40], [41] developed an adjoint-state approach for eikonal TTT that requires relatively low memory overhead, enabling the approach to scale to millions of source-receiver pairs. In our approach, we utilize this adjoint operator to update the inferred velocity model.

Variational inference in seismic imaging

Posterior estimation is a useful tool in ill-posed imaging problems because it provides estimates of uncertainty for image reconstruction. Two popular classes of posterior estimation methods are Markov Chain Monte Carlo (MCMC) [43] and variational inference (VI) [44]. Although MCMC is exact and asymptotically converges to the true posterior, it is computationally demanding to reach the true solution, which is hard to validate. Moreover, it scales poorly to high dimensional posterior distributions and can be prohibitively expensive computationally even for 2D imaging problems[45]. On the other hand, VI is much more efficient at estimating the posterior by constraining estimation over a tractable class of parameterized candidate distributions, but approximates the distribution through this tractable candidate set,

making it less exact. In our work, we focus on using VI for uncertainty quantification due to the high dimensional 3D imaging case that we consider.

VI poses the Bayesian posterior inference problem as one of optimization, in which the target distribution is approximated with a parameterized candidate distribution [45]. This candidate distribution is the one that best fits the target distribution and is a member of a pre-specified class of distributions, referred to as the *variational class*. It is found through optimization by minimizing a loss (e.g., Kullback Leibler divergence [46]) between distributions through $\arg \min_{\theta} D_{\text{KL}}(q_{\theta} || p)$. The choice of variational class is extremely important to well-estimate the posterior, but having more complex variational classes can make the variational inference problem less tractable. There is a trade-off that must be considered between the expressivity of the variational distribution and the efficiency of the optimization problem.

Prior works in seismic imaging performing uncertainty quantification for TTT using VI include [33], [47]. In particular, [47] looks at 2D posterior estimation using ADVI and Stein variational gradient descent (SVGD) while [33] uses normalizing flows as the variational distribution. There are other works that study uncertainty quantification for full-waveform inversion (FWI) using VI including [48], [49] which use SVGD for FWI.

Hypocenter locations

Earthquake hypocenters are almost always determined by solving an inverse problem and thus subject to error, which ultimately can impact the resulting tomography. However, accurate locations depends on accurate velocity models, creating significant trade-off between the two quantities [30]. For this reason, iterative methods that alternate between localization and tomography are popular to prevent bias from poor hypocenter locations [25], [29], [41]. However, accurately estimating the joint uncertainty in the retrieved hypocentral locations and tomographic models is challenging in this iterative framework because of non-convexity.

One way to circumvent the need for iterative methods is to better localize earthquakes initially by implicitly capturing the unmodeled 3D structure without systematically biasing the locations. A popular technique to do this is using source specific station correction terms (SSSTs) when locating events [50], [51]. Given the origin times (i.e. the time of earthquake rupture), which are solved for concurrently, the hypocenter

locations corrected with SSSTs are found by solving the following objective:

$$\begin{aligned} \{\hat{\mathbf{x}}_{s_i}\}_{i \in \mathcal{S}} &= \arg \min_{\substack{\{\mathbf{x}_{s_i}\}_{i \in \mathcal{S}} \\ \{\tau_{ij}\}_{(i,j) \in \mathcal{S} \times \mathcal{R}}} \sum_{i,j} \gamma_{ij} \| (T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) - \tau_{ij}) - y_{ij} \| \quad (5.5) \\ \text{s.t. } \tau_{ij} &= \text{Median}(\{\delta_{kj}\}_{kj}) \quad \text{where} \\ \delta_{kj} &= |T_{\mathbf{x}_{s_k}}(\mathbf{x}_{r_j}) - y_{kj}| \quad \text{for } \mathbf{x}_{s_k} \in \mathcal{B}_r(\mathbf{x}_{s_i}) \end{aligned}$$

where the SSST value, τ_{ij} , corresponding to each travel time measurement y_{ij} is found through iteratively shrinking the neighborhood $\mathcal{B}_r(\mathbf{x}_{s_i})$, a ball of radius r and γ_{ij} weighs each neighbor based on the quality of the measurement. Essentially, this implicitly incorporates the unmodeled 3D structure through estimating a local correction τ_{ij} and finds the optimal location based on this implicit structure.

5.3 Method

Our variational method performs high resolution eikonal tomography without the need to relocate hypocenters during the inversion procedure. Our approach aims to solve for the minimizer of the Kullback-Leibler (KL) divergence, a commonly used objective for VI, which aims to find the optimal recovered posterior distribution $q_\theta(V)$. In particular, our goal is to find the optimal set of parameters θ for our inferred posterior distribution $q_\theta(V)$ which belongs to some predefined variational distribution $\mathbb{P}$. The optimization problem we aim to solve is

$$\begin{aligned} \theta^* &= \arg \min_{\theta \in \mathbb{P}} D_{\text{KL}}(q_\theta(V) \| p(V|y)) \\ &= \arg \min_{\theta \in \mathbb{P}} E_{V \sim q_\theta(V)} [\log q_\theta(V) - \log p(V|y)] \\ &= \arg \min_{\theta \in \mathbb{P}} E_{V \sim q_\theta(V)} [\log q_\theta(V) - \log p(y|V) - \log p(V) \\ &\quad + \log p(y)] \\ &= \arg \min_{\theta \in \mathbb{P}} E_{V \sim q_\theta(V)} [\underbrace{\log q_\theta(V)}_{\text{neg log entropy}} - \underbrace{\log p(y|V)}_{\text{log data likelihood}} - \underbrace{\log p(V)}_{\text{log prior}}]. \quad (5.6) \end{aligned}$$

The log data likelihood takes on different forms depending on how the added noise is modeled. For instance, a very common assumption is that the measurements have i.i.d. Gaussian additive noise $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$ where the log data likelihood is $\log p(y_{ij}|V) = \frac{1}{\sigma^2} \|T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) - y_{ij}\|_2$ and the measurements are assumed to be of the form $y_{ij} \sim T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) + \varepsilon_{ij}$. Another setting that we consider is when the noise is i.i.d. Laplacian additive noise $\varepsilon_{ij} \sim \text{Laplace}(0, b)$ where the log data likelihood is $\log p(y_{ij}|V) = \frac{1}{b} \|T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) - y_{ij}\|_1$ and the measurements are of the form $y_{ij} \sim T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) + \varepsilon_{ij}$. In this work, we consider the setting where the prior,

$\log p(V)$, is implicit and consequently ignored in Eq. 5.6. In particular, we assume an implicit prior through gradient smoothing that is described in Sec. 5.3 and set $\log p(V)$ in 5.6 to be 0.

For our variational family, we use a multivariate log normal distribution of the form $V \sim \text{Lognormal}(\mu, \sigma I)$ where $\mu, \sigma \in \mathbb{R}^{n_x \times n_y \times n_z}$ for $V \in \mathbb{R}^{n_x \times n_y \times n_z}$. The log normal variational distribution ensures that the velocity is non-negative. Our inferred posterior distribution $q_\theta(V)$ is the distribution that minimizes the negative log data likelihood ($-\log p(y|V)$), minimizes the negative log prior ($-\log p(V)$), and maximizes the entropy ($-\log q_\theta(V)$) of the distribution, which encourages diversity of the samples of the inferred posterior.

We estimate the expectation in Eq. 5.6 using Monte Carlo by averaging over samples from the inferred posterior distribution $q_\theta(V)$ giving

$$\theta^* \approx \arg \min_{\theta \in \mathbb{P}} \frac{1}{N} \sum_{i=1}^N [\log q_\theta(V^{(i)}) - \log p(y|V^{(i)})] \quad \text{for } \log p(V^{(i)}) = 0. \quad (5.7)$$

We solve the optimization problem in Eq. 5.7 using gradient descent and initialize the velocity model to be the 1D model used for source localization in [28]. The uncertainty is computed through sampling the distribution and computing per-pixel standard deviations. The forward eikonal prediction is performed using the Pykonal solver developed by [52] and the gradient is computed via the adjoint state described by [41]. Unlike second order methods like BFGS, gradient based methods allow for stochastic gradients, which are induced from the Monte Carlo sampling procedure. For the optimizer, we use Adam [53], a popular optimizer used in machine learning which adaptively updates learning rates. We refer to $\log q_\theta(V^{(i)})$ as our *reconstructed posterior* or *inferred posterior* and show visualizations of the empirical mean and standard deviations of our reconstructed posterior distribution. We also refer to the empirical mean as the *mean reconstruction*.

Gradient smoothing

Many inverse problems are ill-posed and require the usage of priors to prevent overfitting to suboptimal solutions. In seismology, tomography methods typically employ Laplacian regularization to penalize roughness and avoid overfitting to noise.

An alternative regularization approach, designed for iterative inversion algorithms based on gradient descent, smooths the gradient of $\log p(y|V)$ from Eq. 5.6 spatially

to avoid overfitting. We adopt this latter approach in our study and spatially smooth the gradient of $\log p(y|V)$ annealing the degree of smoothing over the course of the optimization. We implement the gradient smoothing with a 3D isotropic Gaussian blur through the following parameter update rule,

$$V^{(k+1)} = V^{(k)} - \eta G(r^{(k)}) \frac{\delta \mathcal{L}}{\delta V} \quad (5.8)$$

where η is the learning rate (step length), G is a Gaussian spatial smoothing operator, $r^{(k)}$ is the radius of isotropic Gaussian blur that monotonically decreases over epochs (iterations) k , and $\mathcal{L}$ is the objective as defined in Eq. 5.4 or Eq. 5.7. Annealing the smoothing operator improves the optimization landscape by preventing overfitting early on in the optimization. This is equivalent to multi-scale inversion approaches under certain assumptions, and further analysis is included in the appendix.

We consider various ways to define the smoothing radius $r^{(k)}$ as a function of epoch k , each of which begins with relatively strong gradient smoothing and ends with relatively weak smoothing. In particular, we consider a piecewise function for the radius $r^{(k)}$ of the isotropic Gaussian blur that initially decays exponentially and then becomes constant after k_{stop} epochs:

$$r^{(k)} = \begin{cases} 5e^{-3k/k_{\text{max}}} & k < k_{\text{stop}} \\ 5e^{-3k_{\text{stop}}/k_{\text{max}}} & k \geq k_{\text{stop}} \end{cases}. \quad (5.9)$$

In our synthetic tests, we consider various values of k , k_{max} , and k_{stop} . For our imaging results with real data, we use $k = 250$, $k_{\text{max}} = 500$, and $k_{\text{stop}} = 250$, as explained in Sec. 5.4. Our smoothing parameter is heuristically chosen, guided by the checkerboard test, to mitigate data overfitting and location–velocity tradeoffs during the tomographic inversion. Each inversion takes around 4 hours on a single Nvidia Tesla V100-SXM2-32GB.

5.4 Synthetic imaging

In this section, we show results from synthetic tests to understand the intrinsic resolution of the imaging setup, evaluate the quality of the uncertainty quantification, and evaluate the effect of various algorithmic choices on the resulting tomography.

For our synthetic tests, we take the same catalog of 45,310 events as described in Sec. 5.5 and generate synthetic arrival times between all 18 receivers for a total of 815,580 P and S arrival time measurements each. We also assume the same measurement noise as in Sec. 5.5 where the observed travel time between source

x_{s_i} and receiver x_{r_j} is $y_{ij} = T_{\mathbf{x}_{s_i}}(\mathbf{x}_{r_j}) + \eta_{ij}$ for $\eta_{ij} \sim \text{Laplace}(0, b)$. The size of the domain used is approximately $25 \text{ km} \times 10 \text{ km} \times 20 \text{ km}$. Our reconstructions are performed on a uniform grid of 1 km^3 for both the forward and inversion modeling, which ensures sufficient computational accuracy while keeping the numerical errors below the observational picking uncertainties.

To understand the sensitivity of reconstruction to various algorithmic choices, we perform synthetic tests with the actual source-receiver geometry to inform hyperparameter choices when imaging with the real dataset. These main design choices that we consider are 1. gradient smoothing hyperparameters and 2. assumed measurement noise model. We also perform various resolution tests through checkerboard and anomaly tests to validate which regions we believe to have sufficient sensitivity.

Hyperparameter tests

Gradient smoothing In Fig. B.5, we consider various functions of decreasing the radius $r^{(k)}$ of isotropic Gaussian blur as described in Eq. 5.9. Although in most cases, the data fit is quite similar at the final point of the optimization, they have varying degrees of overfitting to measurement noise. We find that the result from early stopping with hyperparameters $k = 250$, $k_{\text{max}} = 500$, and $k_{\text{stop}} = 250$ seems to balance fitting the data while preventing overfitting to noise for both V_P and V_S .

Sensitivity to noise model One challenge with performing Bayesian tomography using real data is that it requires a forward model specifying the noise distribution, but the true measurement noise is difficult to characterize. This is due to the complex nature of instrumentation noise and arrival time picking errors, as well as the hypocenter algorithm itself. For a rough measurement noise model characterization, we use the data residuals from the seismicity catalog for the noise. We find that this is a Laplacian distribution as shown in 5.5. However, we could be incorrectly mischaracterizing the true noise model, so we test the effects of mischaracterized noise distributions. We do this by performing synthetic tests where we make different assumptions about the true added measurement noise model and the assumed noise model in Fig. B.6. We consider the measurement noise η_{ij} drawn either from Gaussian or Laplacian distribution and the assumed measurement models for the data likelihood $\log p(y|V)$ either to be Gaussian or Laplacian. From our synthetic tests in Fig. B.6, there are some differences in the reconstructions in terms of the size and amplitude of the recovered anomalies, but we do not find significant differences in the recovery of large features when these noise models are interchanged. We

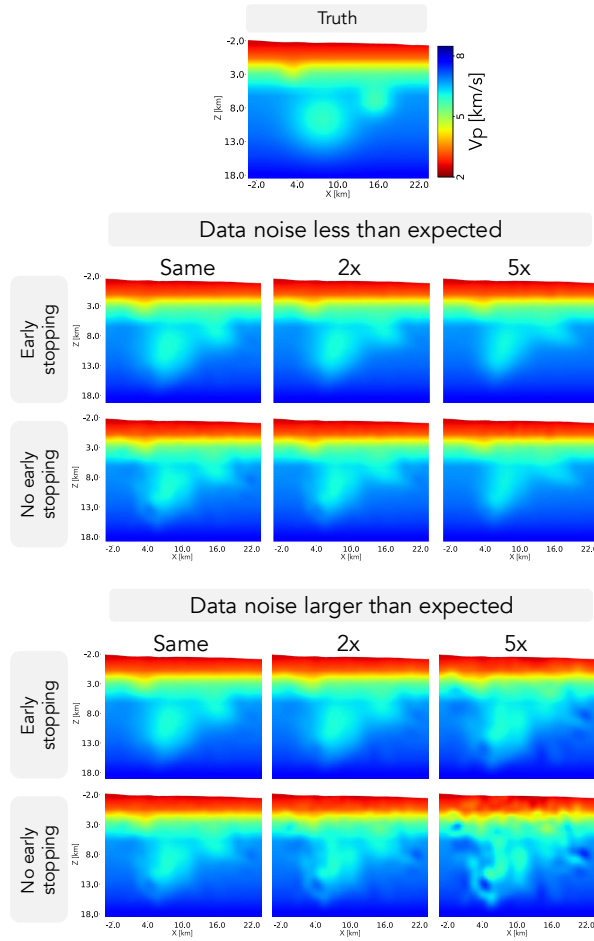


Figure 5.2: **Noise level mismatch.** We consider the impact of misspecifying the noise level b_P of our measurement model $\log p(y|V_P)$ when solving for V_P . We consider the effect of mismatch in assumed level of noise (e.g. over or under estimating). We find that when the noise is less than expected, the structures are similar but the amplitude of recovery is lower. When the noise is much higher than expected, the structures are obfuscated due to overfitting. We find it necessary to do early stopping to prevent overfitting to noise especially when the data noise is larger than expected. These are cross-sections taken along A-A'.

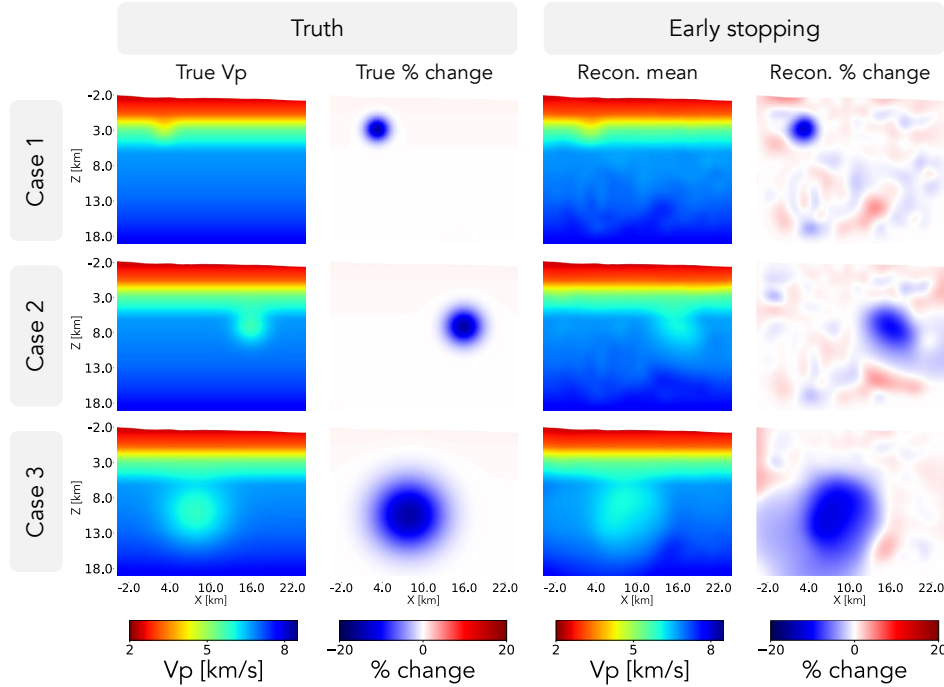


Figure 5.3: **Physically motivated synthetic anomaly tests.** We perform synthetic tests to explore the sensitivity of our expected source-receiver geometry to various $\sim 10\%$ slower velocity anomalies in the subsurface. For various candidate anomalies, we show the following for A-A' (shown in Fig. 5.1): the true velocity model used to generate the synthetic travel times, the relative change to the initial model, and the mean of the inferred posterior and relative change to the initial model. Note, we plot these cross sections masking out the topography.

perform additional tests where we deliberately underestimate and overestimate the amount of noise, as shown in Fig. 5.2. In the case where we underestimate the noise, we find that early stopping helps to prevent overfitting to the measurement noise. When we overestimate the noise, we find that the resulting anomalies that are reconstructed are of a slightly lower amplitude. However, in both cases, the main features are correctly localized.

Resolution tests

Anomaly tests In Figure 5.3, we perform various synthetic anomaly tests to identify how sensitive our source-receiver geometry is to the spatial and amplitude variations of potential anomalies. These anomalies are motivated by several questions of scientific interest about the subsurface plumbing structure at Kilauea. Case 1 considers small magma chambers located near the caldera, Case 2 considers medium-sized anomalies in the rift zone, and Case 3 considers deep magmatic

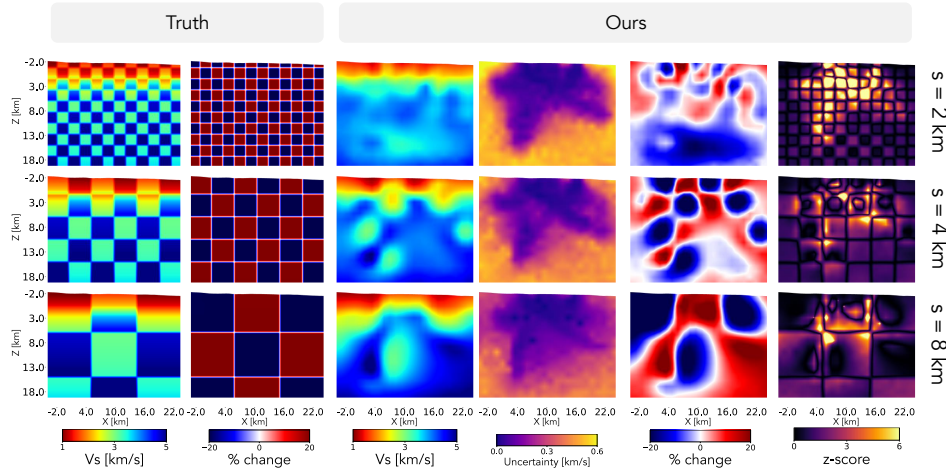


Figure 5.4: Checkerboard tests We perform inversion on the 1D model from [21] modulated with various checkerboards of 20% increase or decrease with a side length of s . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the mean and standard deviation of our reconstruction and relative change between the 1D model and our reconstruction for the A-A' cross-section. We include z-scores to quantify the accuracy of our uncertainty quantification. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage. Interestingly, the estimated uncertainty for $s = 4$ km is higher in low velocity regions such as the one at around 6 km depth and between 6 and 10 km in X , which is expected behavior. We include the full results for both V_P and V_S in the appendix.

systems underneath the Kilauea caldera. We find that our source-receiver geometry has sufficient coverage to localize and reconstruct Case 1, which is close to the surface where the majority of events are located. There are some artifacts in the reconstruction due to the synthetic noise added to the inverse problem. The anomaly in Case 2 is well identified, but not as well reconstructed and localized, due to the asymmetrical source-receiver geometry that results in some smearing artifacts. Again, there are artifacts due to the synthetic noise added to the inverse problem. In Case 3, our method recovers a large deep anomaly; however, the relative shape, location, and amplitude is not exactly correct and there are some smearing artifacts. In all cases, the anomalies are reconstructed with some slight perturbations due to the source-receiver geometry. These tests give confidence in being able to recover anomalies in these regions even though the exact shape and structure is sometimes smeared.

We also consider variations on Case 3 where the anomaly is much smaller in Fig. B.3. In Case 3.1, the anomaly is half the size of Case 3, and in Case 3.2, the anomaly is

a fourth of the size of Case 3. In Case 3.1, the anomaly is recovered but there are more artifacts, and in Case 3.2, the recovered anomaly appears to be on the order of the measurement noise artifacts. In all of these cases, the early stopping parameters, as mentioned in Sec. 5.4, help prevent overfitting.

Checkerboard tests In Figure 5.4, we test the inherent resolution of the source-receiver geometry by using various checkerboard patterns. Using the selected hyperparameters, as mentioned in Sec. 5.4, a vast majority of the region where we have earthquake-coverage is resolvable to a resolution of at least 8 km. In the primary regions of interest, we can achieve a resolution of 4 km. In Fig. B.1, we include z-scores and find that for $s = 4$ km and $s = 8$ km, the ground truth is predominantly within three to four standard deviations from the mean. Most of the high z-score regions are close to the boundaries of the checkerboard, which are sharp and hard to recover. For $s = 2$, we do not trust both the recovered structure and uncertainty because of a mismatch between the blur induced by the early stopping and the resolution of the underlying model we are trying to recover. For the complete result of V_P and V_S , see Fig. B.1.

To understand the regions for which we have sensitivity (for a certain resolution size), we use both the checkerboard test and the ray coverage map to produce a resolvability index map. We borrow the resolvability index introduced in [54] and used in [41]. R is defined as

$$R_s = \frac{\sum_{i \in \mathcal{D}} (\%_{\text{recon}} + \%_{\text{truth}})^2}{2 \sum_{i \in \mathcal{D}} (\%_{\text{recon}}^2 + \%_{\text{truth}}^2)}$$

$$s.t. \quad \%_{\text{recon}} = \frac{\hat{V}_c - V_0}{V_0}, \%_{\text{truth}} = \frac{V_c - V_0}{V_0} \quad (5.10)$$

where $\mathcal{D}$ is the a chosen domain size, $\hat{V}_c$ is the recovered volume with a s sized checkerboard, V_c is the ground truth with an s sized checkerboard, and V_0 is the 1D model [55]. The resolvability index R ranges from 0 to 1 where 0 represents perfectly opposite recovery, 1 represents perfect recovery, and 0.5 represents no recovery. We use $D = 0.6$ as in [41]. We define the domain $\mathcal{D}$ as slightly larger than the size of the checkerboard, and we use the checkerboard with grid size 4 km as shown in Fig 5.5.

5.5 Imaging the Kilauea volcano

In this section, we describe our imaging method, from raw data to final inversion result for V_P and V_S , velocities for the P and S waves respectively. Here we describe

the procedure of generating our catalog, define our measurement model, and show results from our eikonal tomography method. We find that we are able to perform reliable tomography without the need for relocation during the tomography process, unlike previous tomography approaches. By implicitly modeling the 3D structure when determining the initial hypocenters, we are able to estimate the velocity in one-shot, which allows for streamlined uncertainty quantification for the velocity model. The size of the domain used is approximately $25 \text{ km} \times 10 \text{ km} \times 20 \text{ km}$. Our reconstructions are performed on a uniform grid of 1 km^3 in size. This setup is consistent with our synthetic tests.

Data and preprocessing procedure

To generate a travel-time dataset, we use a modified version of the catalog-generating methodology presented in [28]. [28] built a catalog for south Hawai‘i spanning May 2018 to April 2022 by applying shrinking-box source-specific station terms (SSSTs) [50] to the P and S arrival times at the receivers, known as picks. These picks are identified by the automated phase-picking algorithm PhaseNet [56]. Inspired by recent work demonstrating that greater picking accuracy in volcanic environments can be achieved using phase-picking models trained on volcano-tectonic seismicity, we create a new catalog of picks by applying the volcano-trained model Volpick [57] to the same raw data used in [28]. We associate these new picks to existing picks from the [28] catalog. We then recalculate locations for the events of [28] using the new pick times, iteratively applying SSSTs to improve location quality as described in Sec. 5.2. Finally, we remove events with fewer than 10 picks to eliminate low-SNR events which are likely to suffer from greater picking error. For our study, we select events from a window encompassing the Kilauea caldera with depths between 16 km bsl and 1 km asl. This process results in a total of 45,310 events, which yields a total of 378,074 P arrival time measurements and 280,152 S arrival time measurements.

Estimation of measurement noise model

To perform the tomography, we must define the log data likelihood as described in Eq. 5.4 and Eq. 5.7. Since we model our measurements to have additive i.i.d noise, it is sufficient to simply define a model for the measurement noise. We estimate the measurement noise by approximating a distribution over the travel time residuals from locations given by the SSSTs. We find the parameters of the residuals from P and S picks separately and find that the P residuals are best

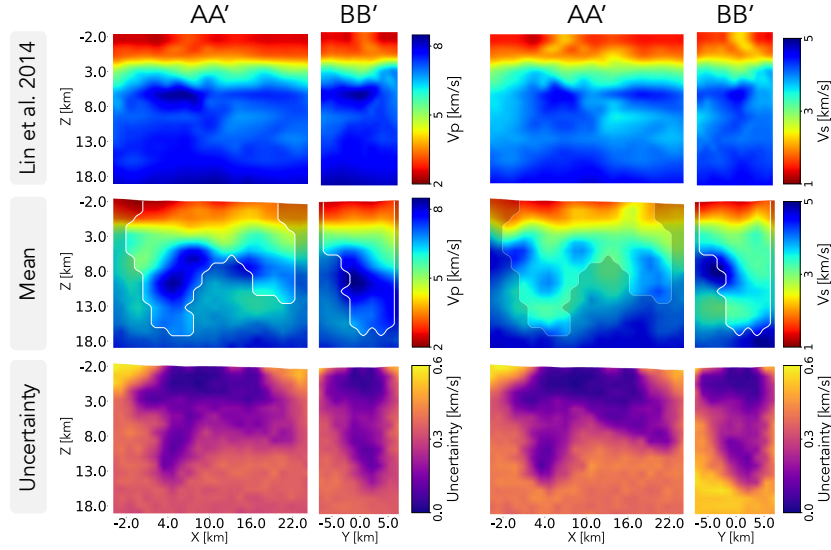


Figure 5.5: Eikonal tomography of Kilauea. We show cross-sections of results from eikonal tomography of Kilauea as described in Sec. 5.4. We show the estimated mean and uncertainty from our approach along with the results from [21]. The darkened part in our estimated mean is given by the resolvability index described in 5.10 where the lighter region is resolvable and darker region is not resolvable for a resolution of 4 km. Our tomography shows evidence of a high V_P anomaly at around 5 to 12 km depth between 3 to 9 km along A-A'. As described in Sec. 5.3, we use Gaussian smoothing of the data-likelihood's gradient to prevent overfitting and over-interpretation of heterogeneous parts of our model. For reference, A is 19.44°N , 155.35°W , A' is 19.34°N , 155.15°W , and B is 19.37°N , 155.30°W and B' is 19.44°N , 155.25°W .

approximated as $\eta_{ij} \sim \text{Laplace}(0, b_P := 0.08)$ and S residuals are best approximated as $\eta_{ij} \sim \text{Laplace}(0, b_S := 0.14)$. The resulting log data likelihood is described in Sec. 6.3, and we use this log data likelihood to solve Eq. 5.4 and Eq. 5.7.

Results

In Fig. 5.5 and Fig. 5.7, we show results from eikonal tomography of Kilauea using real measurements with hypocenters located using SSSTs as described in [28] and Sec. 5.2 and solve for V_P and V_S independently. Our P-velocity tomography (see Fig. 5.5) shows evidence of a high velocity anomaly at 8 km depth under the Kilauea caldera. This feature is consistent with a previous tomography study [21] which used the method of [25] to perform tomography using locations from the Hawaiian Volcano Observatory seismic catalog [HVO]. Similar features are resolved within the S-wave model, but the different amplitudes of the imaged anomalies enable the better definition of the structures obtained from V_P/V_S ratio, which is described in the

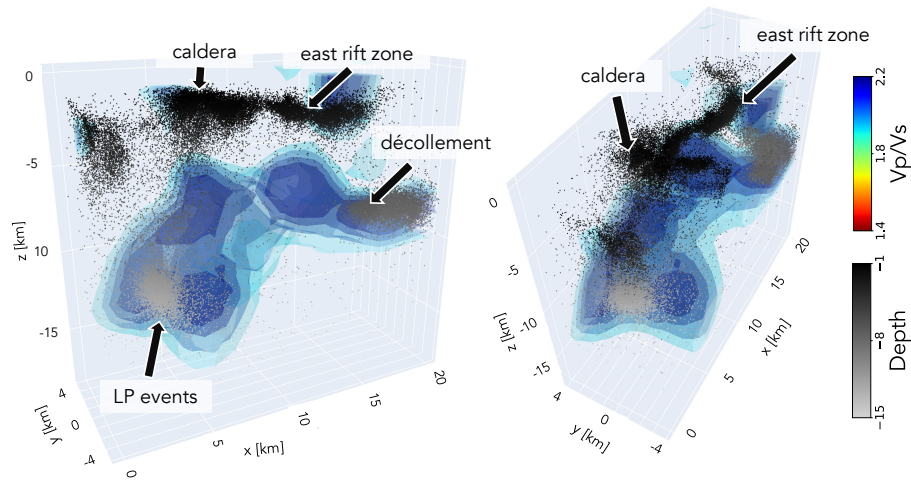


Figure 5.6: **High V_P/V_S regions under Kilauea.** We visualize the isosurfaces of high V_P/V_S regions beneath the Kilauea caldera with two views of the same volume. This high V_P/V_S structure lies between the long period events and the surface expression of the caldera. We estimate the volume of the magma column as 117 km^3 using the region with a $V_P/V_S \geq 2.1$, which is visualized with the second outermost isosurface. Sources are colored as a function of depth.

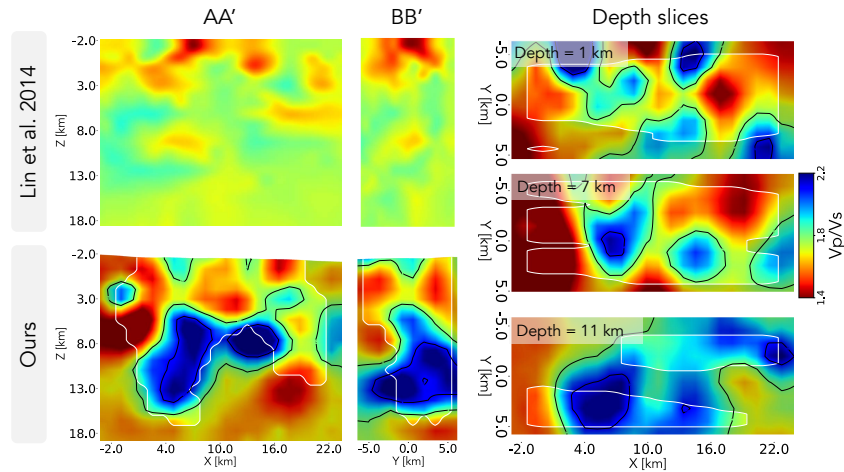


Figure 5.7: We show two cross sections of V_P/V_S for [21] and the mean of our reconstructed posterior as well as additional depth slices for the mean of our reconstructed posterior. The masked region in our estimated mean is given by the resolvability index described in 5.10. We observe a much broader range of V_P and V_S than in the one observed in [21] because we are not as damped and invert for them separately. Our reconstruction shows evidence of a high V_P/V_S chamber underlying the Kilauea caldera. Additional depth slices are included in Fig. B.7.

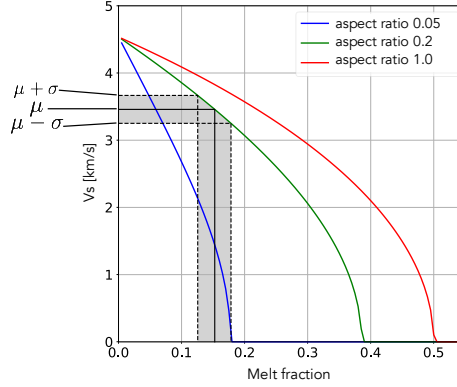


Figure 5.8: Melt fraction computation with uncertainty. We estimate the melt fraction by applying a self-consistent scheme (SCS) effective medium approach. Assuming textural equilibrium for basaltic magma with an aspect ratio of pores and cracks to be 0.2 [58], we estimate a melt fraction of $15 \pm 2.7\%$ at 11 km depth. We compute an average V_S , referred to as μ , and uncertainty, referred to as σ . The solid line represents the melt fraction prediction with μ and the shaded region represents 1 standard deviation σ from μ . For reference, A is 19.44°N , 155.35°W , A' is 19.34°N , 155.15°W , and B is 19.37°N , 155.30°W and B' is 19.44°N , 155.25°W .

next section. While our reconstructions are broadly similar to the features reported in [21], our more finely-spaced parametrization and dense event catalog enable a more detailed view of subcaldera velocity structure. Moreover, our method also includes the estimation of the velocity uncertainties in addition to the mean reconstruction, which can provide bounds to the accuracy of the reconstructed velocities. Although dividing independently inverted V_P and V_S models can amplify errors in the derived V_P/V_S ratios [21], we follow this approach to avoid potential bias introduced by joint inversions; as a result, our interpretations focus on the larger-scale anomalies that are robustly resolved. Synthetic checkerboard and anomaly tests (Figs. 5.3, 5.4, and B.4), which match the real data experimental setup, show that in these regions we are able to resolve features of this scale of 4 km resolution as from the checkerboard tests.

5.6 Discussion

In this section, we present a detailed interpretation of the subsurface structures revealed by our method. We begin by comparing the imaged features with those reported in previous studies and interpreting them within the geological framework of Kīlauea. Finally, we estimate the melt fractions in the subsurface regions we identify as likely to contain partial melt.

In Fig. 5.5, we show the mean and standard deviations of our inverted V_P and V_S models. In Fig. 5.7, we show V_P/V_S for both our method and [21]. We identify two regions marked by large V_P/V_S ratios. The high V_P/V_S region in the south flank zone around 155.2°W and 19.32°N and between 7-10 km depth is also found in [21]. Strikingly, we also observe high V_P/V_S values (>2.1) along a column extending from the caldera to our lowest resolvable model depth (~ 15 km) that we interpret as evidence of a high-melt fraction region. We also identify low V_P/V_S segments at shallow depths (>0 km asl) similar to features identified by [21], who hypothesized that these anomalies reflected the presence of water-saturated pores. Additionally, we identify a high-velocity contrast beneath the rift zone at approximately 5 km depth and between 10 km to 22 km along the x axis, consistent with teleseismic receiver function analyses [59]. Below this contrast, a high V_P/V_S anomaly centered at 8 km depth aligns with previous observations of magmatic underplating [59].

A prominent high V_P/V_S body, located beneath Kīlauea's crater (approximately 4 km on the x-axis in Fig. 5.7) and extending from 13 km to 4 km depth, coincides with the large magma reservoir previously inferred from seismicity and geological evidence [60], [61]. The lower extreme of this anomaly corresponds to the location of previously observed long-period earthquakes [28], [55], [62], suggesting that the imaged structure represents a deep magmatic pathway feeding the reservoir. To better illustrate the imaged structures, we present a 3D view of the V_P/V_S isosurface in Fig. 5.6. Along the east rift zone, deeper high V_P/V_S bodies align with seismicity along the decollement.

To estimate the volume of the deep magma column, we calculate a V_P/V_S isosurface of 2.1 and assume that volumes with V_P/V_S above this value are likely melt-bearing. This process results in an estimated total volume of 117 km^3 for the deep magma column. To determine the melt fraction, we apply a self-consistent scheme (SCS) effective medium approach [63], [64], assuming seismic velocities from [65] ($V_P = 7.955 \text{ km/s}$, $V_S = 4.536 \text{ km/s}$ at 1200 K). Fig. 5.8 displays melt fraction curves for different aspect ratios used in the SCS model. Since our tomography approach optimally fits the observed travel times while incorporating uncertainty bounds, we can quantify both the melt fraction and its associated uncertainty. Assuming textural equilibrium for basaltic magma with an aspect ratio of 0.2 [58], we estimate a melt fraction of $15 \pm 2.7\%$ at 11 km depth. Using the calibration curve from the SCS model, the total melt volume is estimated at $18.5 \pm 4.5 \text{ km}^3$. While this value is slightly higher than previous seismic tomography estimates [65], the overall

density of the picritic mush remains nearly unchanged at 3.2 g/cm^3 , assuming melt and olivine densities of 2.65 g/cm^3 and 3.3 g/cm^3 , respectively. As suggested in prior studies, melt likely separates from the crystal mush to ascend toward shallower depths [65].

Magma volumes within Kilauea’s summit reservoir have traditionally been estimated by using geophysical methods or estimating extrusive volumes; however, the magma volume estimates by these models can vary by orders of magnitude (e.g. [66]–[68]). Smaller estimates derived from petrological analyses indicate that these geophysical methods may have difficulty in distinguishing between a magma chamber’s molten core and the surrounding crystal-mush, higher-temperature rock [69]. In contrast, by reducing damping and performing a high-resolution inversion, our technique allows us to identify extremely high V_P/V_S regions beneath Kilauea that we can reliably interpret as melt-rich regions. Our results are also sensitive to melt-rich regions at depths beneath the summit reservoir and improve our understanding of the magma supply system at Kilauea. Importantly, our technique is also applicable to any volcano with sufficient seismicity. Magma supply processes are recognized as an important control on eruptive styles and timing at volcanoes [70]; the ability to characterize the structures responsible for supply and transport may enable new insight into eruptive hazard and eruption style at other volcano complexes, as well.

5.7 Conclusions

Our work recovers evidence of a magmatic system underneath Kilauea at the highest resolution to date using travel-time to retrieve high-resolution models of V_P and V_S structures. From our 3D tomography study, we estimate the volume of the deep (15–3 km bsl) magma column under Kilauea to be around 117 km^3 . Our eikonal tomography method bypasses the need for iterative earthquake relocations by leveraging SSST-based hypocenter corrections. In addition, we use variational inference and stochastic optimization to estimate the point-wise uncertainty of the inverted tomographic model. Our approach enables imaging high V_P/V_S regions in high seismicity volcanic regions, providing a pathway to better characterize complex magmatic systems and advance understanding of seismovolcanic processes.

References

- [1] J.-L. Got and P. Okubo, “New insights into Kilauea’s volcano dynamics brought by large-scale relative relocation of microearthquakes,” in, *Jour-*

- nal of Geophysical Research: Solid Earth*, vol. 108, no. B7, 2002JB002060, Jul. 2003, ISSN: 0148-0227. DOI: 10.1029/2002JB002060. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2002JB002060> (visited on 06/06/2025).
- [2] O. Lengliné, Z. Duputel, and P. Okubo, “Tracking dike propagation leading to the 2018 Kīlauea eruption,” en, *Earth and Planetary Science Letters*, vol. 553, p. 116 653, Jan. 2021, ISSN: 0012821X. DOI: 10.1016/j.epsl.2020.116653. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0012821X20305975> (visited on 06/06/2025).
- [3] S. Baker and F. Amelung, “Top-down inflation and deflation at the summit of Kīlauea Volcano, Hawai‘i observed with InSAR,” en, *Journal of Geophysical Research: Solid Earth*, vol. 117, no. B12, 2011JB009123, Dec. 2012, ISSN: 0148-0227. DOI: 10.1029/2011JB009123. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2011JB009123> (visited on 06/06/2025).
- [4] D. J. Johnson, A. A. Eggers, M. Bagnardi, M. Battaglia, M. P. Poland, and A. Miklius, “Shallow magma accumulation at Kīlauea Volcano, Hawai‘i, revealed by microgravity surveys,” en, *Geology*, vol. 38, no. 12, pp. 1139–1142, Dec. 2010, ISSN: 1943-2682, 0091-7613. DOI: 10.1130/G31323.1. [Online]. Available: <http://pubs.geoscienceworld.org/geology/article/38/12/1139/130151/Shallow-magma-accumulation-at-K%C4%ABlauea-Volcano> (visited on 06/06/2025).
- [5] C. A. Neal, S. R. Brantley, L. Antolik, *et al.*, “The 2018 rift eruption and summit collapse of Kīlauea Volcano,” en, *Science*, vol. 363, no. 6425, pp. 367–374, Jan. 2019, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aav7046. [Online]. Available: <https://www.science.org/doi/10.1126/science.aav7046> (visited on 06/06/2025).
- [6] T. L. Wright and F. W. Klein, “Deep magma transport at Kīlauea volcano, Hawaii,” en, *Lithos*, vol. 87, no. 1-2, pp. 50–79, Mar. 2006, ISSN: 00244937. DOI: 10.1016/j.lithos.2005.05.004. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0024493705001350> (visited on 06/06/2025).
- [7] M. Ando, “The Hawaii Earthquake of November 29, 1975: Low dip angle faulting due to forceful injection of magma,” en, *Journal of Geophysical Research: Solid Earth*, vol. 84, no. B13, pp. 7616–7626, Dec. 1979, ISSN: 0148-0227. DOI: 10.1029/JB084iB13p07616. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/JB084iB13p07616> (visited on 09/22/2025).
- [8] E. K. Montgomery-Brown, D. K. Sinnett, M. Poland, *et al.*, “Geodetic evidence for an echelon dike emplacement and concurrent slow slip during the June 2007 intrusion and eruption at Kīlauea volcano, Hawaii,” en, *Journal of Geophysical Research: Solid Earth*, vol. 115, no. B7, 2009JB006658,

- Jul. 2010, ISSN: 0148-0227. DOI: 10.1029/2009JB006658. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2009JB006658> (visited on 06/06/2025).
- [9] K. R. Anderson, I. A. Johanson, M. R. Patrick, *et al.*, “Magma reservoir failure and the onset of caldera collapse at Kīlauea Volcano in 2018,” *en, Science*, vol. 366, no. 6470, eaaz1822, Dec. 2019, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz1822. [Online]. Available: <https://www.science.org/doi/10.1126/science.aaz1822> (visited on 06/06/2025).
- [10] F. W. Klein, R. Y. Koyanagi, J. S. Nakata, and W. R. Tanigawa, “The seismicity of Kilauea’s magma system,” in *Volcanism in Hawaii*, ser. U.S. Geological Survey Professional Paper 1350, R. Decker, T. Wright, and P. Stauffer, Eds., vol. 2, 1987, pp. 1019–1185.
- [11] J. D. Wilding and Z. E. Ross, “Insights on the state of stress in the mantle beneath Pahala, Hawai‘i,” *Volcanica*, vol. 7, no. 1, pp. 1–19, Jan. 2024, ISSN: 2610-3540. DOI: 10.30909/vol.07.01.0119. [Online]. Available: <https://www.jvolcanica.org/ojs/index.php/volcanica/article/view/239> (visited on 06/06/2025).
- [12] M. P. Poland, A. Miklius, A. Jeff Sutton, and C. R. Thornber, “A mantle-driven surge in magma supply to Kīlauea Volcano during 2003–2007,” *en, Nature Geoscience*, vol. 5, no. 4, pp. 295–300, Apr. 2012, ISSN: 1752-0894, 1752-0908. DOI: 10.1038/ngeo1426. [Online]. Available: <https://www.nature.com/articles/ngeo1426> (visited on 06/06/2025).
- [13] J. D. Wilding, W. Zhu, Z. E. Ross, and J. M. Jackson, “The magmatic web beneath Hawai‘i,” *en, Science*, vol. 379, no. 6631, pp. 462–468, Feb. 2023, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade5755. [Online]. Available: <https://www.science.org/doi/10.1126/science.ade5755> (visited on 06/06/2025).
- [14] R. P. Denlinger and A. Flinders, “Density structure of Kīlauea volcano: Implications for magma storage and transport,” *en, Geophysical Journal International*, vol. 237, no. 3, pp. 1339–1352, Apr. 2024, ISSN: 0956-540X, 1365-246X. DOI: 10.1093/gji/ggae101. [Online]. Available: <https://academic.oup.com/gji/article/237/3/1339/7633450> (visited on 06/06/2025).
- [15] J. Battaglia, J.-L. Got, and P. Okubo, “Location of long-period events below Kilauea Volcano using seismic amplitudes and accurate relative relocation,” *en, Journal of Geophysical Research: Solid Earth*, vol. 108, no. B12, 2003JB002517, Dec. 2003, ISSN: 0148-0227. DOI: 10.1029/2003JB002517. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2003JB002517> (visited on 06/06/2025).

- [16] W. L. Ellsworth and R. Y. Koyanagi, “Three-dimensional crust and mantle structure of Kilauea Volcano, Hawaii,” en, *Journal of Geophysical Research*, vol. 82, no. 33, pp. 5379–5394, Nov. 1977, ISSN: 01480227. DOI: 10.1029/JB082i033p05379. [Online]. Available: <http://doi.wiley.com/10.1029/JB082i033p05379> (visited on 06/06/2025).
- [17] C. H. Thurber, “Seismic Detection of the Summit Magma Complex of Kilauea Volcano, Hawaii,” en, *Science*, vol. 223, no. 4632, pp. 165–167, Jan. 1984, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.223.4632.165. [Online]. Available: <https://www.science.org/doi/10.1126/science.223.4632.165> (visited on 06/06/2025).
- [18] P. G. Okubo, H. M. Benz, and B. A. Chouet, “Imaging the crustal magma sources beneath Mauna Loa and Kilauea volcanoes, Hawaii,” en, *Geology*, vol. 25, no. 10, p. 867, 1997, ISSN: 0091-7613. DOI: 10.1130/0091-7613(1997)025<0867:ITCMSB>2.3.CO;2. [Online]. Available: <https://pubs.geoscienceworld.org/geology/article/25/10/867-870/187675> (visited on 06/06/2025).
- [19] S. Hansen, “Seismic Velocity and Attenuation Structure of the East Rift Zone and South Flank of Kilauea Volcano, Hawaii,” en, *Bulletin of the Seismological Society of America*, vol. 94, no. 4, pp. 1430–1440, Aug. 2004, ISSN: 0037-1106. DOI: 10.1785/012003154. [Online]. Available: <https://pubs.geoscienceworld.org/bssa/article/94/4/1430-1440/120996> (visited on 06/06/2025).
- [20] J. Park, J. Morgan, C. Zelt, P. Okubo, L. Peters, and N. Benesh, “Comparative velocity structure of active Hawaiian volcanoes from 3-D onshore–offshore seismic tomography,” en, *Earth and Planetary Science Letters*, vol. 259, no. 3-4, pp. 500–516, Jul. 2007, ISSN: 0012821X. DOI: 10.1016/j.epsl.2007.05.008. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0012821X07003019> (visited on 06/06/2025).
- [21] G. Lin, P. M. Shearer, R. S. Matoza, P. G. Okubo, and F. Amelung, “Three-dimensional seismic velocity structure of mauna loa and kilauea volcanoes in hawaii from local seismic tomography,” *Journal of Geophysical Research: Solid Earth*, vol. 119, no. 5, pp. 4377–4392, 2014.
- [22] D. P. Hill and J. J. Zucca, “Geophysical constraints on the structure of Kilauea and Mauna Loa volcanoes and some implications for seismomagmatic processes,” in *Volcanism in Hawaii*, ser. U.S. Geological Survey Professional Paper 1350, R. Decker, T. Wright, and P. Stauffer, Eds., vol. 2, 1987, pp. 903–917.
- [23] M. P. Ryan, “The mechanics and three-dimensional internal structure of active magmatic systems: Kilauea Volcano, Hawaii,” en, *Journal of Geophysical Research: Solid Earth*, vol. 93, no. B5, pp. 4213–4248, May 1988, ISSN: 0148-0227. DOI: 10.1029/JB093iB05p04213. [Online]. Available:

<https://agupubs.onlinelibrary.wiley.com/doi/10.1029/JB093iB05p04213> (visited on 06/06/2025).

- [24] N. Rawlinson, A. Fichtner, M. Sambridge, and M. K. Young, “Seismic Tomography and the Assessment of Uncertainty,” en, in *Advances in Geophysics*, vol. 55, Elsevier, 2014, pp. 1–76, ISBN: 978-0-12-800272-8. DOI: 10.1016/bs.agph.2014.08.001. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0065268714000028> (visited on 06/06/2025).
- [25] C. Thurber and D. Eberhart-Phillips, “Local earthquake tomography with flexible gridding,” *Computers & Geosciences*, vol. 25, no. 7, pp. 809–818, 1999.
- [26] A. Fichtner, A. Zunino, and L. Gebraad, “Hamiltonian Monte Carlo solution of tomographic inverse problems,” en, *Geophysical Journal International*, vol. 216, no. 2, pp. 1344–1363, Feb. 2019, ISSN: 0956-540X, 1365-246X. DOI: 10.1093/gji/ggy496. [Online]. Available: <https://academic.oup.com/gji/article/216/2/1344/5199200> (visited on 06/06/2025).
- [27] R. Maguire, B. Schmandt, J. Li, *et al.*, “Magma accumulation at depths of prior rhyolite storage beneath Yellowstone Caldera,” en, *Science*, vol. 378, no. 6623, pp. 1001–1004, Dec. 2022, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade0347. [Online]. Available: <https://www.science.org/doi/10.1126/science.ade0347> (visited on 06/06/2025).
- [28] J. D. Wilding, W. Zhu, Z. E. Ross, and J. M. Jackson, “The magmatic web beneath hawai ‘i,” *Science*, vol. 379, no. 6631, pp. 462–468, 2023.
- [29] H. Zhang and C. H. Thurber, “Double-difference tomography: The method and its application to the hayward fault, california,” *Bulletin of the Seismological Society of America*, vol. 93, no. 5, pp. 1875–1889, 2003.
- [30] H. Zhang and C. Thurber, “Development and applications of double-difference seismic tomography,” *pure and applied geophysics*, vol. 163, pp. 373–403, 2006.
- [31] J. Um and C. Thurber, “A fast algorithm for two-point seismic ray tracing,” *Bulletin of the Seismological Society of America*, vol. 77, no. 3, pp. 972–986, 1987.
- [32] N. Rawlinson and M. Sambridge, “The fast marching method: An effective tool for tomographic imaging and tracking multiple phases in complex layered media,” *Exploration Geophysics*, vol. 36, no. 4, pp. 341–350, 2005.
- [33] X. Zhao, A. Curtis, and X. Zhang, “Bayesian seismic tomography using normalizing flows,” *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, 2022.

- [34] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.
- [35] J. B. Muir and V. C. Tsai, “Geometric and level set tomography using ensemble kalman inversion,” *Geophysical Journal International*, vol. 220, no. 2, pp. 967–980, 2020.
- [36] J. Vidale *et al.*, “Finite-difference calculation of travel times,” *Bulletin of the seismological society of America*, vol. 78, no. 6, pp. 2062–2076, 1988.
- [37] J. E. Vidale, “Finite-difference calculation of traveltimes in three dimensions,” *Geophysics*, vol. 55, no. 5, pp. 521–526, 1990.
- [38] J. Hole, “Nonlinear high-resolution three-dimensional seismic travel time tomography,” *Journal of Geophysical Research: Solid Earth*, vol. 97, no. B5, pp. 6553–6562, 1992.
- [39] P. Tong, “Adjoint-state traveltime tomography: Eikonal equation-based methods and application to the Anza area in southern California,” *Journal of Geophysical Research: Solid Earth*, vol. 126, no. 5, e2021JB021818, 2021.
- [40] S. Leung and J. Qian, “An adjoint state method for three-dimensional transmission traveltime tomography using first-arrivals,” 2006.
- [41] E. Biondi, W. Zhu, J. Li, E. F. Williams, and Z. Zhan, “An upper-crust lid over the long valley magma chamber,” *Science Advances*, vol. 9, no. 42, eadi9878, 2023.
- [42] S. Fomel, S. Luo, and H. Zhao, “Fast sweeping method for the factored eikonal equation,” *Journal of Computational Physics*, vol. 228, no. 17, pp. 6440–6455, 2009.
- [43] M. Betancourt and M. Girolami, “Hamiltonian monte carlo for hierarchical models,” *Current trends in Bayesian methodology with applications*, vol. 79, no. 30, pp. 2–4, 2015.
- [44] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [45] H. Sun and K. L. Bouman, “Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2628–2637.
- [46] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [47] X. Zhang and A. Curtis, “Seismic tomography using variational inference methods,” *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 4, e2019JB018589, 2020.

- [48] X. Zhang and A. Curtis, “Variational full-waveform inversion,” *Geophysical Journal International*, vol. 222, no. 1, pp. 406–411, 2020.
- [49] X. Zhang, A. Lomas, M. Zhou, Y. Zheng, and A. Curtis, “3-d bayesian variational full waveform inversion,” *Geophysical Journal International*, vol. 234, no. 1, pp. 546–561, 2023.
- [50] K. Richards-Dinger and P. Shearer, “Earthquake locations in southern california obtained using source-specific station terms,” *Journal of Geophysical Research: Solid Earth*, vol. 105, no. B5, pp. 10 939–10 960, 2000.
- [51] G. Lin and P. Shearer, “Tests of relative earthquake location techniques using synthetic data,” *Journal of Geophysical Research: Solid Earth*, vol. 110, no. B4, 2005.
- [52] M. C. White, H. Fang, N. Nakata, and Y. Ben-Zion, “Pykonal: A python package for solving the eikonal equation in spherical and cartesian coordinates using the fast marching method,” *Seismological Research Letters*, vol. 91, no. 4, pp. 2378–2389, 2020.
- [53] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [54] H.-H. Huang, F.-C. Lin, B. Schmandt, J. Farrell, R. B. Smith, and V. C. Tsai, “The yellowstone magmatic system from the mantle plume to the upper crust,” *Science*, vol. 348, no. 6236, pp. 773–776, 2015.
- [55] T. L. Wright and F. W. Klein, “Deep magma transport at kilauea volcano, hawaii,” *Lithos*, vol. 87, no. 1-2, pp. 50–79, 2006.
- [56] W. Zhu and G. C. Beroza, “PhaseNet: A deep-neural-network-based seismic arrival-time picking method,” *Geophysical Journal International*, vol. 216, no. 1, pp. 261–273, 2019.
- [57] Y. Zhong and Y. J. Tan, “Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes,” *Geophysical Research Letters*, vol. 51, no. 12, e2024GL108438, 2024.
- [58] M. Holness, “Melt–solid dihedral angles of common minerals in natural rocks,” *Journal of Petrology*, vol. 47, no. 4, pp. 791–800, 2006.
- [59] G. M. Leahy, J. A. Collins, C. J. Wolfe, G. Laske, and S. C. Solomon, “Underplating of the hawaiian swell: Evidence from teleseismic receiver functions,” *Geophysical Journal International*, vol. 183, no. 1, pp. 313–329, 2010.
- [60] M. Ryan, “The elasticity and contractancy of hawaiian olivine tholeiite, and its role in the stability and structural evolution of sub-caldera magma reservoirs and rift systems. in volcanism in hawaii,” *US Geol. Surv. Prof. Pap.*, vol. 1350, pp. 1395–1447, 1987.

- [61] M. P. Ryan, “The mechanics and three-dimensional internal structure of active magmatic systems: Kilauea volcano, hawaii,” *Journal of Geophysical Research: Solid Earth*, vol. 93, no. B5, pp. 4213–4248, 1988.
- [62] R. S. Matoza, P. M. Shearer, and P. G. Okubo, “High-precision relocation of long-period events beneath the summit region of kilauea volcano, hawai ‘i, from 1986 to 2009,” *Geophysical Research Letters*, vol. 41, no. 10, pp. 3413–3421, 2014.
- [63] G. M. Mavko, “Velocity and attenuation in partially molten rocks,” *Journal of Geophysical Research: Solid Earth*, vol. 85, no. B10, pp. 5173–5189, 1980.
- [64] M. Paulatto, E. E. E. Hooft, K. Chrapkiewicz, B. Heath, D. R. Toomey, and J. V. Morgan, “Advances in seismic imaging of magma and crystal mush,” *Frontiers in Earth Science*, vol. 10, p. 970 131, 2022. doi: 10.3389/feart.2022.970131.
- [65] G. Lin, F. Amelung, Y. Lavallée, and P. G. Okubo, “Seismic evidence for a crustal magma reservoir beneath the upper east rift zone of kilauea volcano, hawaii,” *Geology*, vol. 42, no. 3, pp. 187–190, 2014.
- [66] D. J. Johnson, “Dynamics of magma storage in the summit reservoir of Kilauea Volcano, Hawaii,” en, *Journal of Geophysical Research: Solid Earth*, vol. 97, no. B2, pp. 1807–1820, Feb. 1992, ISSN: 0148-0227. doi: 10.1029/91JB02839. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/91JB02839> (visited on 06/13/2025).
- [67] R. W. Decker, “Dynamics of Hawaiian volcanoes: An overview,” in *Volcanism in Hawaii*, ser. U.S. Geological Survey Professional Paper 1350, R. Decker, T. Wright, and P. Stauffer, Eds., vol. 2, 1987, pp. 903–917.
- [68] F. W. Klein, “Patterns of historical eruptions at Hawaiian volcanoes,” en, *Journal of Volcanology and Geothermal Research*, vol. 12, no. 1, pp. 1–35, Mar. 1982, ISSN: 0377-0273. doi: 10.1016/0377-0273(82)90002-6. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0377027382900026> (visited on 07/11/2022).
- [69] A. J. Pietruszka and M. O. Garcia, “The size and shape of Kilauea Volcano’s summit magma storage reservoir: A geochemical probe,” en, *Earth and Planetary Science Letters*, vol. 167, no. 3-4, pp. 311–320, Apr. 1999, ISSN: 0012821X. doi: 10.1016/S0012-821X(99)00036-9. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0012821X99000369> (visited on 06/13/2025).
- [70] L. Caricchi, C. Annen, J. Blundy, G. Simpson, and V. Pinel, “Frequency and magnitude of volcanic eruptions controlled by magma injection and buoyancy,” en, *Nature Geoscience*, vol. 7, no. 2, pp. 126–130, Feb. 2014, ISSN: 1752-0894, 1752-0908. doi: 10.1038/ngeo2041. [Online]. Available: <https://www.nature.com/articles/ngeo2041> (visited on 06/13/2025).

A CONVEX RELAXATION FOR EIKONAL IMAGING

This chapter is adapted from previously published work [1].

Abstract

Eikonal tomography is an imaging inverse problem ubiquitous in many scientific domains. Solving eikonal tomography is challenging due to the problem’s non-linear and ill-posed nature when reconstructing the underlying image directly. In this chapter, we present an alternative perspective that transforms eikonal imaging from a nonlinear tomographic inverse problem in image space to an inpainting inverse problem in *measurement* space. Our proposed method, EikoLift, does this by solving a physics-constrained inpainting problem in measurement space, and then subsequently maps it back to image space. EikoLift takes advantage of tools from convex optimization, using a semidefinite program to solve the constrained inpainting problem. This convex formulation is grid-free and enables initialization-free inversion. On simulated examples inspired by seismic tomography, EikoLift outperforms both a traditional adjoint-based solver and a physics-informed neural network designed for eikonal imaging.

6.1 Introduction

Stimuli can reveal a material’s properties. The way a material interacts with stimuli can be measured, and the measurements can be used to understand material properties such as heterogeneous composition. For transparent or translucent materials, we can pass light through the material, and the way light deforms reveals the material’s properties. In opaque materials, we can measure how non-visible electromagnetic or acoustic signals are perturbed. We refer to this imaging process as refractive-index imaging (RII). RII aims to reconstruct unknown material composition by analyzing how the material responds to stimuli.

To reconstruct unknown materials, we often apply sources of perturbation (e.g., light, sound, displacement) and measure the response at various surface points. This includes using light from galaxies to identify the index of refraction of dark matter, ultrasound to identify internal body parts, and earthquakes to image and identify geological features such as faults or magma chambers. These imaging

problems require that waves propagate through the underlying object of interest and reach some external sensors around the surface of the object. Then, we use the surface observations to solve an imaging problem, which often takes the form of an optimization problem. Solving the imaging problem requires forward and backward simulation of the propagation, in this case wave physics. We call the class of RII problems that involve wave propagation, wave-field refractive-index imaging (WF-RII).

Modeling complex wave physics is computationally expensive. In seismic WF-RII, which uses surface measurements of earthquake waves to image the Earth, the forward simulation of wave propagation is so expensive that neural network surrogate models are preferred for tractability [2]–[4]. Surrogate models that approximate wave propagation are also often preferred in ultrasound computed tomography (CT) [5]. Supervised methods for WF-RII have also been proposed, but they do not explicitly obey the underlying physics [6], making it impossible to guarantee that the measurements actually match the proposed solution. For this reason, we do not consider end-to-end approaches as methods that solve WF-RII. WF-RII exhibits additional problems that make it challenging to solve: 1. run-times scaling polynomially with respect to the domain size; 2. ill-posedness, which results in sensitivity to initialization; and 3. challenges in optimization due a non-linear wave-physics forward model.

A faster alternative to using waveform information is to use the first arrival time of the wave-field to inform the recovery of the underlying materials. This form of imaging arises in many of the same scientific and industrial applications as WF-RII and has been used to understand cardiac function [7], image the ionosphere to correct for atmospheric turbulence in GPS and space imaging [8], [9], and travel time tomography (TTT) in seismology to estimate the composition of the Earth [10]–[15]. In these scenarios, refractive-index imaging uses information of the time-of-flight (ToF) from sources to receivers to reconstruct the composition of the resonating material. The relationship between the ToF and material composition is governed by the eikonal equation. This equation is much less computationally demanding to simulate as compared to wave physics. We refer to the class of imaging methods using ToF as ToF-RII.

Although ToF-RII is less computationally demanding than WF-RII, it still has some of the same fundamental issues as it involves an ill-posed inverse problem with a non-linear forward model. These non-linear inverse problems often scale poorly with

respect to the number of pixels or voxels; doubling the resolution is equivalent to an $8\times$ time increase in forward and inverse modeling [10]. This can be prohibitive when recovering high resolution domains even if the measurements lend themselves to a low-dimensional reconstruction. Traditional approaches often make approximations to speed up inference, especially for meter to kilometer scale domains. In some cases, these approximations are reasonable, such as in photoacoustic tomography [16], [17], but in other cases there is a substantial mismatch, such as when using pseudo-ray bending [18], [19] for seismic tomography where strong bending breaks the approximation. Thus, there have been multiple approaches using physics-informed neural networks that simulate ToF measurements, which speed up imaging while enforcing the underlying physics [20]–[22]. Even with these approximations and alternative approaches, problems associated with non-linearity and ill-posedness still exist.

We introduce EikoLift, a convex relaxation for solving refractive index imaging problems using time-of-flight measurements (ToF-RII). In particular, we reframe the inverse problem from a tomography problem in *target* space into a constrained inpainting problem in *measurement* space. We find that this approach addresses some fundamental limitations of traditional tomography methods in that it does not require initialization and compute times are independent of resolution size. Our approach yields an imaging algorithm that is more robust to initialization than conventional approaches. In simulated examples inspired by seismic tomography, EikoLift outperforms previous methods designed for eikonal imaging. While performance under realistic noise is more limited, EikoLift remains competitive and offers a promising framework for further developing robust, physics-informed ToF imaging techniques.

6.2 Background

Eikonal equation and forward simulation

Details on the eikonal equation and forward simulation are in Sec. 2.2.

Time-of-flight refractive-index imaging (ToF-RII) and adjoint methods

ToF-RII can be solved using an adjoint method applied to the eikonal equation [10], [23]–[25]. Although the eikonal-based adjoint method obeys the physics exactly, optimizing V is extremely challenging; the eikonal-based adjoint method is sensitive to the initialization of V and each gradient step can be computationally slow, especially for large-scale high-resolution reconstructions which can be on

the order of minutes for each step. Especially in cases with many gradient steps, such as posterior estimation and other sampling-based approaches, using adjoint eikonal methods can be computationally prohibitive. A much faster alternative solves an inverse ray-tracing problem [14], [15], [19], [26]–[29]. Inverse ray-tracing methods are often preferred over eikonal imaging due to their efficiency, but make approximations that yield less accurate reconstructions.

There is a related class of methods that use image-space measurements instead of ToF for RII, including as Schlieren imaging [30]–[32], intensity diffraction tomography [33]–[35], and others [36], [37]. These methods do not readily apply to ToF-RII because the measurements are of a different form. Another related class of methods use techniques from ToF-RII and WF-RII for imaging a surface rather than the medium itself, such as [38]. These methods involve the same physics, but the ultimate reconstructed object is different.

Details on ToF-RII (referred to as travel-time tomography) are in Sec. 2.2.

PINNs for ToF-RII

Forward simulation for predicting ToF is computationally expensive. Thus, physics-informed neural networks (PINNs), such as EikoNet [21] and others [20], [22], were introduced as a way to speed up the forward modeling process to not require ray tracing or use of a finite difference solver such as FMM. PINNs, which take the form of a neural network $F_\theta(\cdot)$ with parameters θ , are trained such that the output from the neural network obeys the underlying PDE. In the case of ToF, PINNs are trained to output $F_\theta(\mathbf{x}_s, \mathbf{x}_r) \approx T_V(\mathbf{x}_s, \mathbf{x}_r)$ without actually computing $T_V(\mathbf{x}_s, \mathbf{x}_r)$. This is done by requiring that the neural network satisfies

$$\|\nabla_{\mathbf{x}_r} F_\theta(\mathbf{x}_s, \mathbf{x}_r)\|_2^2 = \frac{1}{V(\mathbf{x}_r)^2}, \quad F_\theta(\mathbf{x}, \mathbf{x}) = 0 \quad (6.1)$$

for a specified velocity field V .

When applied to solving ToF-RII, EikoNet does not work out of the box and requires additional regularizers to ensure that intrinsic properties of the eikonal PDE are satisfied [39]. Although PINNs are able to act as an efficient surrogate forward model once they have been trained, they do not perform well when solving the inverse problem. When using PINNs for inverse problems, the observed measurements are sparse, resulting in weak supervision that does not guarantee that the resulting reconstructions are physically sound.

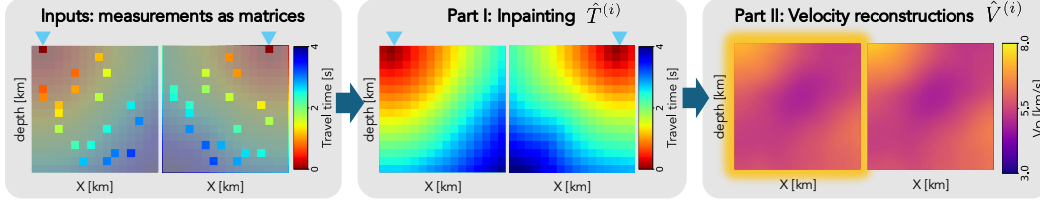


Figure 6.1: **Our approach.** Our method uses time-of-flight measurements $\{T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)})\}_{\forall i,j}$ between sources $\{\mathbf{x}_s^{(j)}\}_{j=1}^{n_s}$ and receivers $\{\mathbf{x}_r^{(i)}\}_{i=1}^{n_r}$ where $n_s = 20$ and $n_r = 2$, and outputs the reconstructed velocity model $\hat{V}$ following the method outlined in Sec. 6.3. The input measurements and source-receiver locations are visualized as partially observed matrices. Each matrix corresponds to measurements between a single receiver (indicated by the cyan triangle) and all observed sources. The shaded region represents areas with missing information and each square of the unshaded region represents a single measurement. These potentially noisy measurements are then used in our method, which consists of two parts: 1. inpainting in the measurement space as described in Sec. 6.3 and 2. velocity reconstruction from the inpainted measurements as in Sec. 6.3. From the reconstructed velocities, the optimal one is selected, as highlighted in yellow.

6.3 Approach

We propose EikoLift, an alternative approach to adjoint tomography which frames the problem as a constrained inpainting problem. EikoLift solves this problem in two steps: 1. performing the constrained inpainting problem in measurement space and 2. taking the inpainted measurements and casting them back to the image space to get the velocity reconstruction.

Part 1: Constrained inpainting in the measurement space

This framing allows us to recast a highly non-convex tomography problem into a convex problem by inpainting in the measurement space instead of directly solving for the target velocity model V . Note that unlike the velocity model, the ToF field $T(\mathbf{x}_s, \mathbf{x}_r)$ is defined for pairs of locations $\mathbf{x}_s$ and $\mathbf{x}_r$, so we are inpainting a continuous volume in $\mathbb{R}^{2d}$ ($2d$ since the ToF field is defined for $\mathbf{x}_s \in \mathbb{R}^d$ and $\mathbf{x}_r \in \mathbb{R}^d$). However, since the receiver geometry is sparse and often fixed in real-world problems, many points, and even planes, do not have any observations. So instead of inpainting a large poorly-constrained volume in $\mathbb{R}^{2d}$, we inpaint individual ToF fields $T(\mathbf{x}, \mathbf{x}_r^{(i)})$ for each receiver $\mathbf{x}_r^{(i)}$. Thus, our proposed optimization problem inpaints n_r ToF fields $\{\hat{T}^{(i)}(\mathbf{x})\}_{i=1}^{n_r}$ from our observations. Note that here we make a distinction between T_V and T where T_V is the ToF induced by a velocity field V and T

represents a ToF field recovered by our algorithm.

However, solving the inpainting problem alone is not enough; we must also obey the eikonal equation so that the velocity given by each of the n_r ToF fields is equal. Although the ToF fields are different, they map to the same velocity field V . Thus, EikoLift aims to solve the following objective:

$$\begin{aligned} \{\hat{T}^{(i)}(\mathbf{x})\}_{i=1}^{n_r} = \arg \min_{\{T^{(i)}(\mathbf{x})\}_{i=1}^{n_r}} & \sum_{i \in [n_r], j \in [n_s]} (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - T^{(i)}(\mathbf{x}_s^{(j)}))^2 \\ \text{s.t. } & \|\nabla T^{(i)}(\mathbf{x})\|_2^2 = \|\nabla T^{(h)}(\mathbf{x})\|_2^2 \quad \forall i, h, \mathbf{x}. \end{aligned} \quad (6.2)$$

Here, the objective contains the same data constraint as in Eq. 2.6. However, we update our decision variable to be the n_r ToF fields $\{T^{(i)}(\mathbf{x}) \in C^1(\mathbb{R}^d)\}_{i=1}^{n_r}$. Since the ToF fields are now unconstrained, we add an additional constraint where the ToF fields must satisfy the eikonal equation $V(\mathbf{x})^{-1/2} = \|\nabla T^{(i)}(\mathbf{x})\|_2^2 = \|\nabla T^{(h)}(\mathbf{x})\|_2^2$ for each i, h , and $\mathbf{x}$ as given by Eq. 2.5. Without this constraint, the recovered ToF fields would be inconsistent and not satisfy Eq. 2.5.

To convexify Eq. 6.2, we represent each ToF field $T^{(i)}$ as a linear combination of K basis functions $\phi_k \in C^1(\mathbb{R}^d)$:

$$T^{(i)}(\mathbf{x}) = \langle \alpha^{(i)}, \phi(\mathbf{x}) \rangle = \sum_{k=1}^K \alpha_k^{(i)} \phi_k(\mathbf{x}). \quad (6.3)$$

Here, $\alpha^{(i)} = (\alpha_1^{(i)}, \dots, \alpha_K^{(i)}) \in \mathbb{R}^K$ is a vector of expansion coefficients and $\phi = (\phi_1, \dots, \phi_K) : \mathbb{R}^d \rightarrow \mathbb{R}^K$ is a vector-valued function that collates the basis functions. We can reconstruct the velocity model from the ToF fields by solving Eq. 2.5 using Eq. 6.3, yielding:

$$\begin{aligned} \frac{1}{V(\mathbf{x})^2} = \|\nabla T^{(i)}(\mathbf{x})\|_2^2 &= \sum_{k=1}^d \left\langle \left(\sum_{j=1}^K \alpha_j^{(i)} \partial_k \phi_j(\mathbf{x}) \right), \left(\sum_{l=1}^K \alpha_l^{(i)} \partial_k \phi_l(\mathbf{x}) \right) \right\rangle \\ &= \langle \alpha^{(i)}, \nabla \phi(\mathbf{x}) \nabla \phi^T(\mathbf{x}) \alpha^{(i)} \rangle = \langle \alpha^{(i)} \alpha^{(i)T}, \nabla \phi(\mathbf{x}) \nabla \phi^T(\mathbf{x}) \rangle. \end{aligned} \quad (6.4)$$

Here, $\nabla \phi(\mathbf{x}) = (\partial_k \phi_j(\mathbf{x}))_{k \in [d], j \in [K]} \in \mathbb{R}^{K \times d}$ is the Jacobian matrix of ϕ , and we have used $\langle \cdot, \cdot \rangle$ for matrix arguments to denote the trace inner product $\langle A, B \rangle = \text{Trace}(A^T B)$.

In its current form, we are solving a non-linear physics-constrained imaging inpainting problem. The non-linear constraints limit us from access to a wide range of

methods that only apply to linear problems. To enable the use of special tooling for solving linear problems, we need to make these constraints linear. We do this by introducing a new matrix decision variable $M^{(i)} \succ 0$ that approximates the rank-1 version $\alpha^{(i)}\alpha^{(i)T}$. This relaxation makes the optimization problem easier to solve because we can use convex algorithms. Ideally $M^{(i)}$ would be exactly $\alpha^{(i)}\alpha^{(i)T}$, but by solving for $M^{(i)}$ directly rather than $\alpha^{(i)}$, it is not guaranteed to be perfectly rank-1. The best we can do is to find solutions $\hat{M}^{(i)}$ with minimal rank. Thus, we have turned this inpainting problem into a rank minimization problem:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} & \sum_i \text{Rank}(M^{(i)}) \\ \text{s.t.} & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(h)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, h, \mathbf{x} \\ & \quad (T_{obs}^2(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi(\mathbf{x}_s^{(j)})^T M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall \mathbf{x}_r^{(i)}, \mathbf{x}_s^{(j)} \\ & \quad M^{(i)} \succ 0 \quad \forall i. \end{aligned} \tag{6.5}$$

This rank minimization problem is still nonconvex and NP-hard. To deal with this obstruction, the standard approach [40]–[42] is to convexify the problem by using the trace in place of the rank, yielding the following semi-definite program (SDP), which is a convex problem:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} & \sum_i \text{Trace}(M^{(i)}) \\ \text{s.t.} & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(h)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, h, \mathbf{x} \\ & \quad (T_{obs}^2(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi(\mathbf{x}_s^{(j)})^T M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall \mathbf{x}_r^{(i)}, \mathbf{x}_s^{(j)} \\ & \quad M^{(i)} \succ 0 \quad \forall i. \end{aligned} \tag{6.6}$$

Full derivation is in the Appendix. We use `cvxpy` [43], [44] as our convex optimization solver, which uses a splitting conic solver SCS [45] and runs on CPU. To parallelize, we use GNU Parallel [46].

Block diagonal formulation: To reduce the number of parameters, we parameterize $M^{(i)} = I_{rect}^{(i)} A^{(i)} I_{rect}^{(i)T}$ to be a block diagonal matrix where it is non-zero only for entries that correspond to the bases related to $T^{(i)}$. Full implementation details are included in the Appendix.

A relaxation for real measurement noise

In real-world applications, the measurement noise can take large values with non-negligible probability. As such, having a hard constraint for the data-fit as in Eq. 6.6

can cause the problem to be infeasible when the measurement noise is large. A naive solution would be to make the tolerance ε_{data} substantially larger to allow for these improbable high-noise occurrences. However, there is no guarantee that ε_{data} is large enough, and having a large ε_{data} yields overly smooth reconstructions because perturbations from the underlying velocity model appear to be on the order of noise under this measurement model. Thus, we empirically find that it is better to replace these constraints with a penalty term in the objective with hyperparameter $\lambda > 0$:

$$\begin{aligned} \{\hat{A}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{A^{(i)}\}_{i=1}^{n_r}} & \underbrace{\sum_i \text{Trace}(A^{(i)})}_{\text{rank min. surrogate}} + \lambda \underbrace{\sum_i \|T_{obs}^2 - \phi^T M^{(i)} \phi\|_2^2}_{\text{data constraint}} \quad (6.7) \\ \text{s.t.} \quad & \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(k)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, k, \mathbf{x} \\ & M^{(i)} = I_{rect}^{(i)} A^{(i)} I_{rect}^{(i)T}, \quad A^{(i)} \succcurlyeq 0 \quad \forall i. \end{aligned}$$

Part 2: Recast to velocity reconstruction

Solving Eq. 6.6 (or Eq. 6.7) yields $\{\hat{M}^{(i)}\}_{i=1}^{n_r}$, which represent the inpainted ToF fields. Since solving Eq. 6.6 does not guarantee perfectly rank-1 solutions, we must find a “good” rank-1 approximation for all $\{\hat{M}^{(i)}\}_{i=1}^{n_r}$. We define this “good” rank-1 approximation as $\hat{\alpha}^{(i)}$. To find $\hat{\alpha}^{(i)}$, which are used to reconstruct the velocity $\hat{V}^{(i)}$, we take the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}^{(i)}$ of each $\hat{M}^{(i)}$ to get $\hat{\alpha}^{(i)}$. Thus, our reconstructed ToF fields $\hat{T}^{(i)}$ are:

$$\hat{T}^{(i)} = \langle \hat{\alpha}^{(i)}, \phi \rangle \quad (6.8)$$

and the reconstructed velocity fields $\hat{V}^{(i)}$ are

$$\hat{V}^{(i)}(\mathbf{x}) = \left(\langle \hat{\alpha}^{(i)}, \nabla \phi(\mathbf{x}) \nabla \phi^T(\mathbf{x}) \hat{\alpha}^{(i)} \rangle \right)^{1/2}. \quad (6.9)$$

Solving Eq. 6.8 and Eq. 6.9 yields n_r velocity reconstructions corresponding to n_r inpainted ToF fields. When there is no gap in generalization (i.e., there is no noise and the candidate $T^{(i)}$ is in the span of the basis) and the eikonal constraint is sufficiently strong in Eq. 6.6 or Eq. 6.7, all $\hat{V}^{(i)}$ should be equal so there is no need to select between possible reconstructions. In practice, this is not the case since real data is noisy and a finite basis will never fully span the space of all continuous functions $C(\mathbb{R}^d)$. Thus, it is important to design a criteria on how to select the best reconstruction, but designing such a criteria is unclear. We explore various options, including selecting the velocity reconstruction that fits the data the best or selecting the one where $M^{(i)}$ has the lowest rank. Selecting the one with the best data-fit

	Adjoint mean init.	PINN mean init.	PINN rand. init.	ours Eq. 6.6	ours Eq. 6.7 - A	ours Eq. 6.7 - B
mesh-free	×	✓	✓	✓	✓	✓
initialization-free	×	×	✓	✓	✓	✓
MSE $\sigma=0.1$ $n_r=4$	.63 $\pm$.60	.21 $\pm$.15	.24 $\pm$.16	.36 $\pm$.32	.35 $\pm$.21	.19 $\pm$.12
MSE $\sigma=0.1$ $n_r=12$	.28 $\pm$.24	.06 $\pm$.03	.13 $\pm$.11	.22 $\pm$.17	.38 $\pm$.32	.11 $\pm$.05
MSE $\sigma=0$ $n_r=12$	.20 $\pm$.08	.05 $\pm$.01	.15 $\pm$.12	.09 $\pm$.05	–	–

Table 6.1: We include quantitative comparisons across 20 examples for: 1. the adjoint method w/ TV regularization based on [10], 2. physics-informed neural network (PINN) based on [39], 3. our method using Eq. 6.6 selecting the minimum trace, 4. our method using Eq. 6.7 selecting the minimum trace, and 5. our method using Eq. 6.7 with the minimal measurement error. We include the mean and standard deviation of the MSE($\uparrow$) for both $n_r = 4$ and $n_r = 12$ with SNR = 25 dB and $n_r = 12$ for noise-free measurements.

can potentially be prone to overfitting the noise, but do not find this to be the case empirically.

Computational complexity

The complexity of traditional methods (e.g., eikonal-based adjoint tomography and inverse ray-tracing) is dependent on the domain size and resolution in addition to $\min(n_s, n_r)$, which is due to reciprocity as described in Sec. 6.2. On the other hand, EikoLift’s computational complexity is independent of the domain size and resolution. Although we use a conic solver, which has no strict time bound, EikoLift can also be solved using interior point methods which have a run-time complexity that is polynomial in m and n [47] where m is the number of constraints and $n \times n$ is the size of the decision variable matrices $M^{(i)}$. That is, the complexity scales with the number of observations ($\leq n_s n_r$), number of receivers n_r or sources n_s whichever is smaller, the data tolerance ε_{data} , and the number of points used for the eikonal constraint. In theory, EikoLift is discretization free when ϕ and $\nabla\phi$ are continuously defined. If ϕ and $\nabla\phi$ are not continuously defined, then there can be large memory costs associated with storing them. However, there is a memory-runtime trade-off if we choose to approximate $\nabla\phi$ at runtime using finite difference to reduce memory constraints.

6.4 Results

In this section, we consider eikonal imaging using EikoLift and present results for $d = 2$ for a domain of size 16 km $\times$ 16 km with variable numbers of sources and receivers. Although this chapter presents results for $d = 2$, our method handles

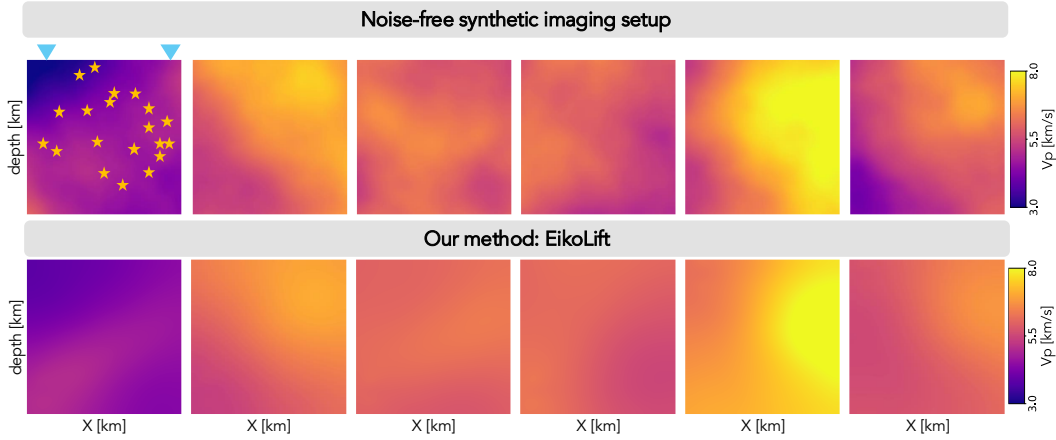


Figure 6.2: **Noise-free imaging.** We generate synthetic measurements between 2 receivers and 20 sources using the ground truth velocity models (top row). The corresponding reconstructions (bottom row) are from solving Eq. 6.6 and selecting the field with the minimal trace. Our method is able to reconstruct the regions with source-receiver coverage fairly well. We use the same sensor and receiver setup for each synthetic experiment where the yellow stars are sources and cyan triangles are receivers.

arbitrary d . For our basis functions ϕ , we use PCA on the ToF fields from 1000 samples of ground truth velocity fields. We use `pykonal` [48] to generate synthetic data. We consider both noise-free and noisy measurements that simulate real world seismic measurement acquisition. We model the noisy measurements to have additive i.i.d Gaussian noise $\eta_{ij} \sim \mathcal{N}(0, \sigma = 0.1)$ for all $i \in [n_r]$ and $j \in [n_s]$ such that $T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) = T_V(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) + \eta_{ij}$. $\sigma = 0.1$ seconds corresponds to a SNR of 25 dB, which is realistic or even conservative for seismic measurements.

Following from the block diagonal construction described in Sec. 6.3 and the Appendix, $\{\phi^{(i)}\}_{i=0}^{K-1}$ consists of the top K PCs from the ToF fields for receiver $i = 0$, $\{\phi^{(i)}\}_{i=K}^{2K-1}$ correspond to the top K PCs from the ToF fields for receiver $i = 1$, and so on. We know that ϕ well-approximates the ToF fields $T^{(i)}$ used in these examples because it is data-driven, and we use $K = 6$, which captures over 99.8% of the variance for each i . Since $\phi^{(i)}$'s are discrete functions, we use finite differences to compute $\nabla \phi$. We include example ground truth velocity model and basis functions $\phi^{(i)}$ with the partial derivatives $\partial_x \phi^{(i)}$ and $\partial_y \phi^{(i)}$ in the Appendix. Note that although we choose the basis functions to be data-driven for computational efficiency, this can also be done using continuous functions with analytic partial derivatives, a common choice being Chebyshev polynomials [49].

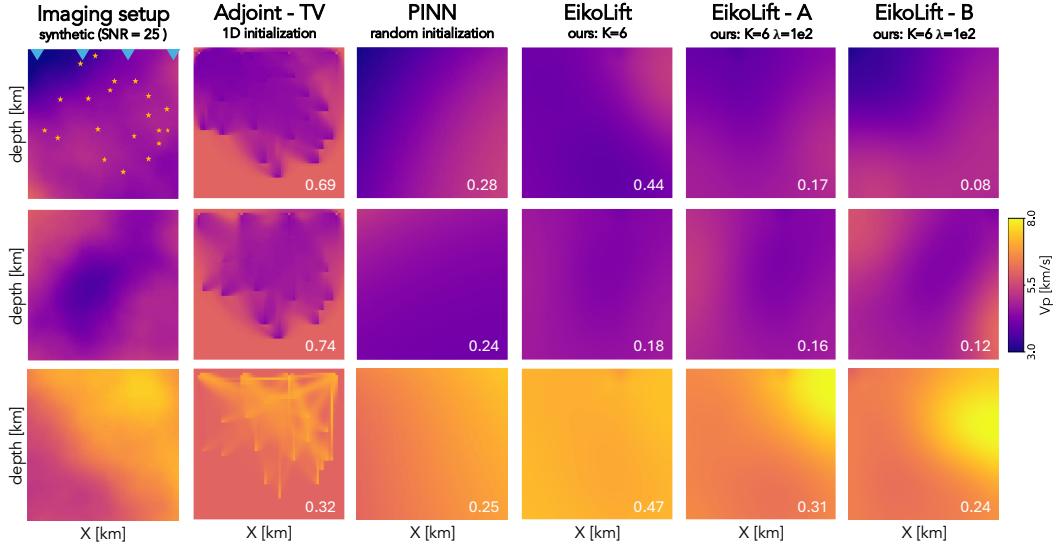


Figure 6.3: EikoLift on noisy measurements observed from only one surface of the domain. We reconstruct velocity models from noisy synthetic measurements with a SNR of 25 dB. We use 4 receivers and 20 sources where the yellow stars are sources and cyan triangles are receivers. The columns from left to right are: ground truth velocity model, reconstruction with the adjoint [10] method and TV prior, reconstruction with a ToF-PINN [39], reconstruction from EikoLift using Eq. 6.6, and reconstructions EikoLift using Eq. 6.7 with two separate criteria: A. minimizing the trace and B. minimizing measurement error. Mean squared error (MSE) is included for each reconstruction.

Baselines

In Fig. 6.3 and Fig. 6.4, we show comparisons between our method and two physics-based baselines: 1) eikonal-based adjoint tomography and 2) a physics-informed neural network (PINN).

Adjoint eikonal tomography: The adjoint eikonal tomography method based on [10] aims to solve the standard eikonal tomography problem described in Eq. 2.6 by using the adjoint method. This method obeys the underlying physics and is highly sensitive to initialization. We add additional regularization through a total variation (TV) prior with weight $\lambda = 0.1$ chosen to minimize MSE relative to the ground truth – yielding an optimistic, best-case result. We optimize using Adam [50] with a learning rate of $1e-2$.

PINN: We compare with a ToF-PINN, as described in Sec. 6.2. The ToF-PINN that we use for this baseline is based on the M-step of [39]. The ToF-PINN uses

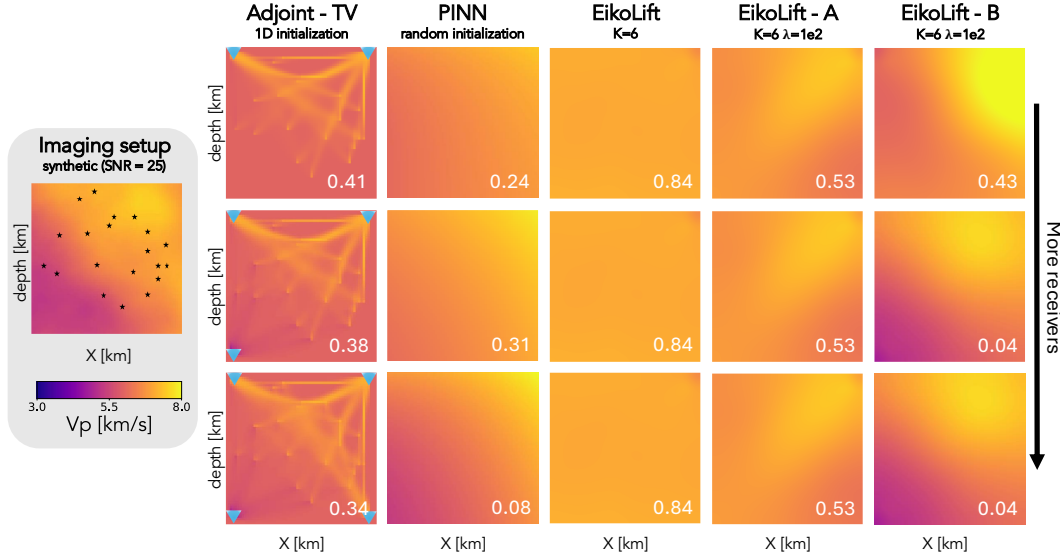


Figure 6.4: **Increasing sensor coverage improves reconstruction.** We show reconstructions with increasingly better receiver coverage to illustrate improved performance as receivers increase from 2 to 4. We find that this is true for the adjoint as well as our specialized EikoLift-B setting using Eq. 6.7 with minimizing measurement error as the selection criteria. The EikoLift results using Eq. 6.6 are overly smooth due to the high ε_{data} needed to prevent infeasibility. We use between 2 to 4 receivers and 20 sources where the black stars are sources and cyan triangles are receivers. The top row corresponds to 2 receivers, the middle with 3 receivers, and the bottom with 4 receivers. Mean squared error (MSE) is included for each reconstruction.

the eikonal PDE to learn a neural field that fits the observations. The objective is similar to the one described in Eq. 6.7, and the loss function consists of the data-fit and PDE constraint as a loss. Once the neural field is trained to fit the measurement data, the velocity model is then extracted from the PINN by solving Eq. 2.5 through automatic differentiation. We show results where the PINN is initialized to random but smooth ToF fields.

Results

In Fig. 6.2, we consider ToF-RII using noise-free measurements between 2 receivers and 20 sources. We simulate the measurements using FMM and solve the ToF-RII problem using Eq. 6.6 with $\varepsilon_{data} = 0.01$ and $\varepsilon_{PDE} = 0.1$. We visualize the reconstructed velocity field that minimizes the trace. Our method resolves the large scale features within the regions that have source coverage. The trace minimization criteria for this SDP works well when imaging using noise-free measurements.

Although there should be a small gap between the true ToF fields and $\text{span}(\phi)$ by construction, even in a noiseless setting, it is not guaranteed that the inpainted ToF fields are well described by $\text{span}(\phi)$ due to the convex relaxation of the rank-1 constraint. Nonetheless, the recovered solutions remain close to rank-1, indicating that the relaxation is effective in guiding the solution toward the intended structure.

In Fig. 6.3 and Fig. 6.4, we consider imaging with noisy measurements using both Eq. 6.6 and Eq. 6.7. In Fig. 6.3, we use 4 surface-only receivers and 20 sources to image the field of interest. We include results from the adjoint and PINN baselines as well as from EikoLift using Eq. 6.6 and Eq. 6.7, where EikoLift-A indicates selecting the $M^{(i)}$ with minimal trace and EikoLift-B indicates selecting the $M^{(i)}$ with minimal measurement error. We find noisy measurements require a looser data constraint ($\varepsilon_{data} = 0.5$) in Eq. 6.6 to avoid infeasibility, but this can lead to overly smooth results as shown by the EikoLift results. EikoLift-A and -B results that instead optimize Eq. 6.7 are more robust to measurement noise; for these results we use $\varepsilon_{PDE} = 0.1$ and $\lambda = 100$. See Appendix for ablations. We find that across all settings of EikoLift, our method yields a lower MSE than the adjoint, and EikoLift-B’s performance outperforms the PINN with random initialization. As we increase the receiver coverage, as in Fig. 6.4, we find that the performance improves with more receivers for the adjoint and EikoLift-B. However, we find that EikoLift-A struggles to select the optimal reconstruction when minimizing trace, and a limitation of EikoLift using Eq. 6.6 is overly smooth reconstructions due to needing a high ε_{data} to prevent infeasibility.

Baseline comparison In terms of image reconstruction quality, the adjoint has the worst performance. This is due to sensitivity to the initialization and weak regularization, unlike the explicit regularization through smooth PCA components. The adjoint is almost unnaturally heterogeneous due to the way the gradient is computed, which makes finding features challenging. The streaky artifacts are mitigated by the total-variation prior, but is insufficient even though the weight has been chosen optimally. If we increase the TV prior, the reconstruction does not deviate much from the initial homogeneous solution. We show tests varying TV in the Appendix.

Under the same assumption of no initialization, EikoNet-B outperforms the randomly initialized PINN, as shown in Table 6.1. This is expected, as the PINN is solving a non-convex optimization problem that can get stuck in a local minima

when not initialized close to the ground truth. Only when the PINN is initialized via pretraining to the correct mean value derived from the ground truth, the performance is comparable and sometimes better than EikoNet-B. While EikoLift and the PINN-based method can yield similar results when the PINN is carefully initialized, EikoLift offers a principled convex formulation that provides a clearer view of the optimization landscape. Since both approaches have similar objectives that balance data-fit and PDE constraints, EikoLift can serve as a reference point for interpreting or guiding the behavior of ToF-PINNs, especially when the PINN is made to optimize over the same basis functions.

Basis functions Since this method is convex, it is not biased by the initialization. However, EikoLift is dependent on the choice of basis $\{\phi^{(i)}\}_{i \in [n_r K]}$ that describes the ToF fields $T^{(i)}$. When using PCA to find the basis functions, selecting the right K is just as important as selecting the right basis. Having a smaller K reduces the runtime, but can lead to overly smooth reconstructions. Having a larger K can yield higher resolution reconstructions, but might overfit to noise. We include experiments for varying K in the Appendix.

6.5 Results with real data

In this section, we extend our framework to a 3D real-data TOF-RII problem. We consider the earthquake data described in Sec. 5.5 from Chapter 5. This catalog consists of earthquakes from the Island of Hawai‘i from May 2018 to April 2022. These events are automatically identified using a variant of PhaseNet [51] that is specifically trained on volcano data, called Volpick [52]. In particular, we study the events from a window that surrounds the Kilauea caldera and consider the V_P velocity, which is the velocity that governs the P-wave time-of-flight. Although this process yields 45,310 events, we only consider a small fraction of the events used in Sec. 5.5. We take a subset of 100 earthquakes and 5 receivers to demonstrate EikoLift on a 3D TOF-RII problem for V_P velocity inversion. For our imaging setup, we use EikoLift-B, which solves Eq. 6.7. We use $K = 10$ basis functions and set the hyperparameters $\varepsilon_{PDE} = 0.5$ and $\lambda = 100$. To generate the collection of $\{\phi^{(k)}\}_{k=1}^K$, we extract the top K PCs from synthetic datasets generated from two strategies: 1. Gaussian anomalies modulated on top of an initial 1D, and 2. randomly adding DCT components to the same initial 1D model where the weight on the DCT component scales with respect to the distance from the DC component. The 1D model we use matches that in Sec. 5.5. Although our data is only a subset of the full dataset, we

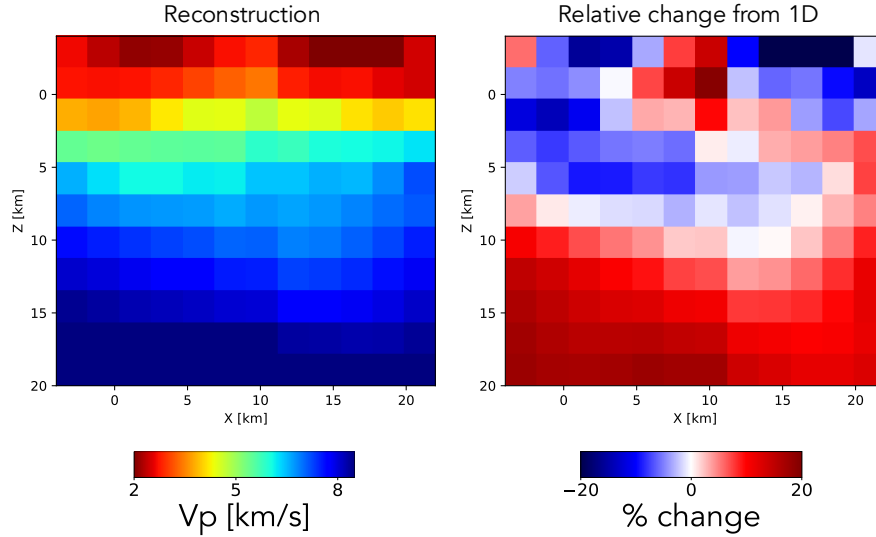


Figure 6.5: **Applying EikoLift to real data** We show a 2D slice from a reconstruction using EikoLift on real seismic data from the Kilauea caldera with 5 receivers and 100 events. This reconstruction uses a subset of the data used to image in Chapter 5. Here, we show the central cross section A-A' as described in 5.1. We include the relative change from the 1D model used in [11], which is used to initialize the algorithm in Chapter 5.

see the same high level features of a large low velocity region in the field of view with both strategies to define ϕ .

In terms of computational complexity, extending to 3D results in a negligible added computational cost for solving the convex problem defined in Eq. 6.7, which describes EikoLift-B. However, it substantially increases the computational complexity especially for solving Eq. 6.6. Additionally, the velocity reconstruction is significantly slower in part because there is a complicated inner-product that needs to be computed for each grid point.

6.6 Conclusions

In this chapter, we present EikoLift, a convex framework for eikonal imaging. EikoLift enables the use of linear techniques by reformulating ToF-RII as a PDE-constrained inpainting problem in measurement space, rather than a non-convex inverse problem in image space. This leads to a more stable optimization procedure that is grid-free and does not require initialization, departing from the limitations of traditional methods. By shifting how we approach time-of-flight reconstruction, EikoLift opens new avenues for scientific imaging, with potential applications in

fields such as seismology.

References

- [1] Angela F. Gao, Ethan N. Epperly, Ayush Varshney, Joel Tropp, and Katherine L. Bouman, “Eikolift: Convexifying eikonal tomography through inpainting time-of-flight,” *In submission to Advances in Neural Information Processing Systems*, 2025,
- [2] Yan Yang, Angela F. Gao, Jorge C. Castellanos, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Seismic wave propagation and inversion with neural operators,” *The Seismic Record*, vol. 1, no. 3, pp. 126–134, 2021. doi: <https://doi.org/10.1785/0320210026>,
- [3] Yan Yang, Angela F. Gao, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Rapid seismic waveform modeling and inversion with neural operators,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. doi: <https://doi.org/10.1109/TGRS.2023.3264210>,
- [4] C. Zou, K. Azizzadenesheli, Z. E. Ross, and R. W. Clayton, “Deep neural helmholtz operators for 3-d elastic wave propagation and inversion,” *Geophysical Journal International*, vol. 239, no. 3, pp. 1469–1484, 2024.
- [5] Z. Zeng, Y. Zheng, Y. Zheng, Y. Li, Z. Shi, and H. Sun, “Neural born series operator for biomedical ultrasound computed tomography,” *arXiv preprint arXiv:2312.15575*, 2023.
- [6] L. Lozenski, H. Wang, F. Li, *et al.*, “Learned full waveform inversion incorporating task information for ultrasound computed tomography,” *IEEE Transactions on Computational Imaging*, vol. 10, pp. 69–82, 2024. doi: [10.1109/TCI.2024.3351529](https://doi.org/10.1109/TCI.2024.3351529).
- [7] T. Grandits, K. Gillette, A. Neic, *et al.*, “An inverse eikonal method for identifying ventricular activation sequences from epicardial activation maps,” *Journal of computational physics*, vol. 419, p. 109 700, 2020.
- [8] C. N. Mitchell and P. S. Spencer, “A three-dimensional time-dependent algorithm for ionospheric imaging using gps,” *Annals of Geophysics*, vol. 46, no. 4, pp. 687–696, 2003.
- [9] A. Settimi, A. Ippolito, C. Cesaroni, and C. Scotto, “Scientific review on the ionospheric absorption and research prospects of a complex eikonal model for one-layer ionosphere,” *International Journal of Geophysics*, vol. 2014, no. 1, p. 657 434, 2014.
- [10] E. Biondi, W. Zhu, J. Li, E. F. Williams, and Z. Zhan, “An upper-crust lid over the long valley magma chamber,” *Science Advances*, vol. 9, no. 42, eadi9878, 2023.

- [11] G. Lin, F. Amelung, Y. Lavallée, and P. G. Okubo, "Seismic evidence for a crustal magma reservoir beneath the upper east rift zone of kilauea volcano, hawaii," *Geology*, vol. 42, no. 3, pp. 187–190, 2014.
- [12] F.-C. Lin, M. H. Ritzwoller, and R. Snieder, "Eikonal tomography: Surface wave tomography by phase front tracking across a regional broad-band seismic array," *Geophysical Journal International*, vol. 177, no. 3, pp. 1091–1110, 2009.
- [13] P. M. Shearer, *Introduction to seismology*. Cambridge university press, 2019.
- [14] H. Zhang and C. H. Thurber, "Double-difference tomography: The method and its application to the hayward fault, california," *Bulletin of the Seismological Society of America*, vol. 93, no. 5, pp. 1875–1889, 2003.
- [15] J. R. Evans, D. Eberhart-Phillips, and C. Thurber, "User's manual for simulps12 for imaging vp and vp/vs; a derivative of the "thurber" tomographic inversion simul3 for local earthquakes and explosions," US Geological Survey, Tech. Rep., 1994.
- [16] S. Dey, S. Saha, B. T. Feng, *et al.*, "Score-based diffusion models for photoacoustic tomography image reconstruction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 2470–2474.
- [17] H. Liu, K. Wang, D. Peng, *et al.*, "Curve-driven-based acoustic inversion for photoacoustic tomography," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2546–2557, 2016.
- [18] J. Um and C. Thurber, "A fast algorithm for two-point seismic ray tracing," *Bulletin of the Seismological Society of America*, vol. 77, no. 3, pp. 972–986, 1987.
- [19] K. Koketsu and S. Sekine, "Pseudo-bending method for three-dimensional seismic ray tracing in a spherical earth with discontinuities," *Geophysical Journal International*, vol. 132, no. 2, pp. 339–346, 1998.
- [20] S. Grubas, A. Duchkov, and G. Loginov, "Neural eikonal solver: Improving accuracy of physics-informed neural networks for solving eikonal equation in case of caustics," *Journal of Computational Physics*, vol. 474, p. 111 789, 2023.
- [21] J. D. Smith, K. Azizzadenesheli, and Z. E. Ross, *Eikonet: Solving the eikonal equation with deep neural networks*, 2020. arXiv: 2004.00361 [physics.comp-ph].
- [22] U. bin Waheed, E. Haghighat, T. Alkhalifah, C. Song, and Q. Hao, "Pin-neik: Eikonal solution using physics-informed neural networks," *Computers & Geosciences*, vol. 155, p. 104 833, 2021.
- [23] R.-E. Plessix, "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications," *Geophysical Journal International*, vol. 167, no. 2, pp. 495–503, 2006.

- [24] S. Li, A. Vladimirsky, and S. Fomel, “First-break travelttime tomography with the double-square-root eikonal equation,” *Geophysics*, vol. 78, no. 6, U89–U101, 2013.
- [25] E. Biondi, G. Barnier, R. G. Clapp, F. Picetti, and S. Farris, “An object-oriented optimization framework for large-scale inverse problems,” *Computers & Geosciences*, vol. 154, p. 104 790, 2021.
- [26] X. Zou, F. Vandenberghe, B. Wang, *et al.*, “A ray-tracing operator and its adjoint for the use of gps/met refraction angle measurements,” *Journal of Geophysical Research: Atmospheres*, vol. 104, no. D18, pp. 22 301–22 318, 1999.
- [27] F. Anis, Y. Lou, A. Conjusteau, *et al.*, “Investigation of the adjoint-state method for ultrasound computed tomography: A numerical and experimental study,” in *Photons Plus Ultrasound: Imaging and Sensing 2014*, SPIE, vol. 8943, 2014, pp. 452–457.
- [28] A. Chakrabarti, “Learning sensor multiplexing design through back-propagation,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [29] È. Campeau-Poirier, L. Perreault-Levasseur, A. Coogan, and Y. Hezaveh, “Time delay cosmography with a neural ratio estimator,” *arXiv preprint arXiv:2309.16063*, 2023.
- [30] A. Teh, W. H. Ali, J. Rapp, and H. Mansour, “Indoor airflow imaging using physics-informed schlieren tomography,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [31] A. Mazumdar, “Principles and techniques of schlieren imaging systems,” *Columbia University Computer Science Technical Reports*, vol. 14, 2013.
- [32] E. Traldi, M. Boselli, E. Simoncelli, *et al.*, “Schlieren imaging: A powerful tool for atmospheric plasma diagnostic,” *EPJ Techniques and Instrumentation*, vol. 5, pp. 1–23, 2018.
- [33] R. Liu, Y. Sun, J. Zhu, L. Tian, and U. S. Kamilov, “Recovery of continuous 3d refractive index maps from discrete intensity-only measurements using neural fields,” *Nat. Mach. Intell.*, vol. 4, pp. 781–791, 2022.
- [34] Y. Sung, W. Choi, C. Fang-Yen, K. Badizadegan, R. R. Dasari, and M. S. Feld, “Optical diffraction tomography for high resolution live cell imaging,” *Optics express*, vol. 17, no. 1, pp. 266–277, 2009.
- [35] Y. Park, M. Diez-Silva, G. Popescu, *et al.*, “Refractive index maps and membrane dynamics of human red blood cells parasitized by plasmodium falciparum,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 37, pp. 13 730–13 735, 2008.

- [36] T. Xue, M. Rubinstein, N. Wadhwa, A. Levin, F. Durand, and W. T. Freeman, "Refraction wiggles for measuring fluid depth and velocity from video," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13*, Springer, 2014, pp. 767–782.
- [37] B. Zhao, A. Levis, L. Connor, P. P. Srinivasan, and K. L. Bouman, "Single view refractive index tomography with neural fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 358–25 367.
- [38] D. B. Lindell, G. Wetzstein, and M. O'Toole, "Wave-based non-line-of-sight imaging using fast fk migration," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–13, 2019.
- [39] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, "Deepgem: Generalized expectation-maximization for blind inversion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.
- [40] A. Chai, M. Moscoso, and G. Papanicolaou, "Array imaging using intensity-only measurements," *Inverse Problems*, vol. 27, no. 1, p. 015 005, 2010.
- [41] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.
- [42] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [43] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.
- [44] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.
- [45] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, "Conic optimization via operator splitting and homogeneous self-dual embedding," *Journal of Optimization Theory and Applications*, vol. 169, no. 3, pp. 1042–1068, Jun. 2016. [Online]. Available: <http://stanford.edu/~boyd/papers/scs.html>.
- [46] O. Tange, *Gnu parallel 20241222 ('bashar')*, GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them., Dec. 2024. doi: 10.5281/zenodo.14550073. [Online]. Available: <https://doi.org/10.5281/zenodo.14550073>.

- [47] H. Jiang, T. Kathuria, Y. T. Lee, S. Padmanabhan, and Z. Song, “A faster interior point method for semidefinite programming,” in *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*, IEEE, 2020, pp. 910–918.
- [48] M. C. White, H. Fang, N. Nakata, and Y. Ben-Zion, “Pykonal: A python package for solving the eikonal equation in spherical and cartesian coordinates using the fast marching method,” *Seismological Research Letters*, vol. 91, no. 4, pp. 2378–2389, 2020.
- [49] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. Chapman and Hall/CRC, 2002.
- [50] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [51] W. Zhu and G. C. Beroza, “PhaseNet: A deep-neural-network-based seismic arrival-time picking method,” en, *Geophysical Journal International*, vol. 216, no. 1, pp. 261–273, Jan. 2019, ISSN: 0956-540X. DOI: 10.1093/gji/ggy423. [Online]. Available: <https://academic.oup.com/gji/article/216/1/261/5129142> (visited on 11/04/2019).
- [52] Y. Zhong and Y. J. Tan, “Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes,” *Geophysical Research Letters*, vol. 51, no. 12, e2024GL108438, 2024.

SURROGATE MODELS FOR WAVE IMAGING

This chapter is adapted from previously published work [1], [2].

Abstract

Seismic wave propagation forms the basis for most aspects of seismological research, yet solving the wave equation is a major computational burden that inhibits the progress of research. This is exacerbated by the fact that new simulations must be performed whenever the velocity structure, source location, or source structure is perturbed. Here, we explore a framework for learning general solutions using a recently developed machine learning paradigm called Neural Operator. A trained Neural Operator can compute a solution in negligible time for any velocity structure or source location. We develop a scheme to train Neural Operators on an ensemble of waveform simulations performed with random velocity models and source locations. As Neural Operators are grid-free and continuously defined, it is possible to evaluate solutions on higher resolution velocity models than trained on, providing additional computational efficiency. We illustrate the method with the 2D acoustic and elastic wave equations and demonstrate the method's applicability to seismic tomography, using reverse-mode automatic differentiation to compute gradients of the wavefield with respect to the velocity structure. The developed procedure is nearly an order of magnitude faster than using conventional numerical methods for full waveform inversion.

7.1 Introduction

The simulation of seismic wave propagation through Earth's interior underlies most aspects of seismological research, from the simulation of strong ground shaking due to large earthquakes [3], [4], to imaging the subsurface velocity structure [5]–[9], to derivation of earthquake source properties [10]–[12]. The compute costs associated with these wavefield simulations are substantial, and for reasons of computational efficiency, 1D models are often used, even when better 3D velocity models are available. As a result, seismic wave simulations are often the limiting factor in the pace of geophysical research.

Recently, deep learning approaches have been explored with the goal of solving

various geophysical partial differential equations [13]–[16]. Beyond the goal of accelerating compute capabilities, these physics-informed neural networks may offer other advantages such as grid-independence, low memory overhead, differentiability, and on-demand solutions. These properties can enable deep learning being used to solve geophysical inverse problems [17]–[21], as a wider selection of algorithms and frameworks then are available for use, such as approximate Bayesian inference techniques like variational inference.

One of the major challenges associated with wave propagation is that a new simulation must be performed whenever the properties of the source or velocity structure are perturbed in some way. This alone substantially increases the necessary compute costs, making some problems prohibitively expensive even if they are mathematically or physically tractable. For the most part, these limitations have been accepted as an inevitable part of seismology, but now physics-informed machine learning approaches have started to offer some pathways for moving beyond this issue. For example, [13] use a deep neural network to solve the Eikonal equation for any source-receiver pair by taking these locations as input. This then can be exploited for hypocenter inversion by allowing for gradients of the travel time field to be computed with respect to the source location [20]. However, these models are relatively inefficient to train and even then, are only able to learn approximate solution operators to the differential equations.

The aforementioned limitations may seem surprising, but result from a basic attribute of neural networks that in fact makes them ill-suited for solving differential equations. Specifically, neural networks are designed for learning maps between two finite dimensional spaces, while learning a general solution operator for a differential equation requires the ability to map between two infinite dimensional spaces (i.e. function spaces). A paradigm for learning maps between function spaces was recently developed [22]–[24], and has been termed Neural Operator. The general idea behind these models is that they have shared parameters over all possible functions describing the initial conditions, which allows them operate on functions, even when the inputs are a numerically discretized representation of them.

Here, we explore the potential of Neural Operators in improving seismic wave propagation and inversion. We develop a framework for training Neural Operators on the 2D acoustic and elastic wave equations and show that this approach provides a suite of tantalizing new advantages over conventional numerical methods for seismic wave propagation. This study provides a proof of concept of this technology and its

application to seismology.

7.2 Preliminaries

Figures

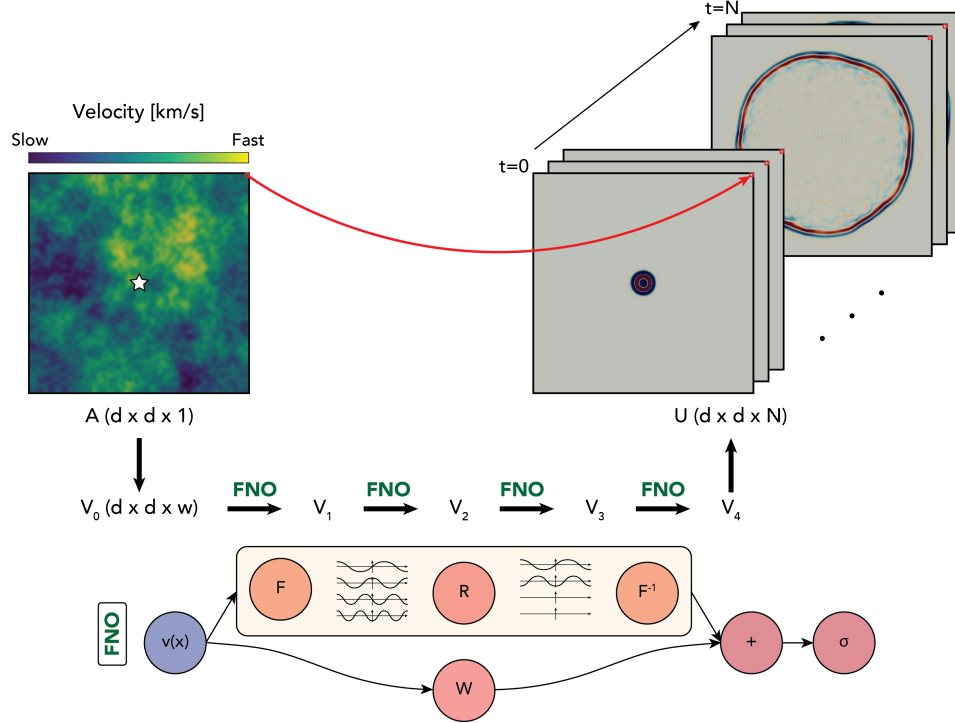


Figure 7.1: Our approach applying FNO to the 2D acoustic wave equation. The inputs to the FNO model are the velocity model with dimensions $d \times d \times 1$ and the source location, indicated by the white star. The input velocity model is lifted to a higher dimensional space with size $d \times d \times w$ using a neural network. Then, we apply four Fourier operator layers, and finally project back to the target wavefield dimensions of $d \times d \times N$ using a neural network. The bottom panel shows details of the Fourier layer architecture: we define v to be the input. On top: We apply a Fourier transform F to v , then apply a linear transformation R to the lower Fourier modes, filtering out higher modes. Then we apply an inverse Fourier transform F^{-1} . On the bottom: we apply a local linear transform W to v .

For a given function A and a Green's function G , let U denote the solution to a linear PDE, i.e., the solution operator,

$$U(x) = (\mathcal{L}A)(x) = \int G(x, y)A(y)dy$$

where $\mathcal{L}$ is the corresponding linear operator. For example, suppose that the PDE to be solved is the acoustic wave equation; then A could describe the velocity structure as well as the initial conditions. Neural operator generalizes this formulation to the nonlinear setting where a set of linear operators are sequentially applied to construct a general nonlinear solution operator. In its basic form, an ℓ -layered neural operator is constructed as follows:

$$U(x) = \mathcal{L}_\ell (\sigma(\mathcal{L}_{\ell-1} \dots \sigma(\mathcal{L}_1 V) \dots)) (x)$$

where $\mathcal{L}_i$ is such that for any function V , we have,

$$(\mathcal{L}_i V)(x) = W_i(x) + \int K_i(x, y) V(y) dy \quad (7.1)$$

Under this framework, $W_i(x)$ and $K_i(x, y)$ constitute the learnable components of the Neural Operator and allow for a solution to be produced for any prescribed function A . In a limited sense, Neural Operators can be viewed as generalized Green's functions.

7.3 Methods

We designed a framework that applies Neural Operators to the 2D acoustic and elastic wave equations. The basic idea for this procedure is outlined schematically in Figure 7.1. A specific type of Neural Operator, called a Fourier Neural Operator (FNO) [24] receives a velocity model specified on an arbitrary, possibly irregular mesh, along with the coordinates of a point source. One of the main features of FNO is that the major calculations are performed in the frequency domain, which allows the convolutions in Eq. 7.1 to be rapidly computed. The output of the FNO is the complete wavefield solution, which can be queried anywhere within the medium, regardless of whether the points lie on the input mesh.

The most basic component of the FNO is a Fourier block (Fig. 7.1), which first transforms an input function (V) to the Fourier domain. In the first layer of the network, V is equal to the initial conditions, A . Then, a kernel (K_i) is computed specifically for this function and is truncated at low order, before performing the integration via multiplication. Finally, the result is transformed back and a non-linear activation function is applied, which concludes the Fourier block. For this study, the architecture of the FNO contains 4 sequential Fourier blocks and applies a ReLU activation to the output of each (Fig. 7.1). We note that the truncation of the Fourier modes is performed on the function values after lifting them to a higher

dimensional space, rather than the raw input function, so that this does not lead to compression. We refer the interested readers to [24] for more mathematical details about the FNO.

Training dataset

As Gaussian random fields can represent all continuous functions, the purpose of these steps is to create a suite of simulations that span a range of possible conditions. We show later that they can even well-approximate strongly discontinuous velocity models. An example velocity model and simulation is shown in Figure 7.1.

Dataset for acoustic wave propagation We constructed a training dataset of 5,000 simulations to learn from by first generating random velocity models. We set up a 64×64 grid with 0.16 km spacing to define the velocity model. Then, we created 5,000 random velocity models by sampling random fields having a von Karman covariance function with the following parameters: Hurst exponent $\kappa = 0.5$, correlation length $\mathbf{a} = [a_x, a_y] = [0.16\text{km}, 0.16\text{km}]$, and $\varepsilon = 0.1$, $\mu = 3\text{km/s}$, and $\sigma = 0.15\text{km}$. Then to get the acoustic wavefield, for each of these velocity models, we apply a Ricker wavelet source at a random point, and solve the acoustic wave equation using a spectral-element method (SEM) [25]. It should be noted that there is a source grid used since this is a requirement of the spectral-element method.

Dataset for elastic wave propagation We define V_P in the same way as the velocity used for acoustic wave propagation and generate 20,000 examples. For the elastic wavefield, we also require V_P/V_S , which we set to have an average background of 1.732 perturbed by a smooth Gaussian random field with the following parameters: correlation length $\lambda=32$ grids, standard deviation $\sigma=2\%$ of the background. For simulating the elastic wavefield, we use [26] with the top boundary set with a free-surface boundary condition and the other three edges have absorbing boundary conditions. A total of 4-sec wavefield with a time step of 0.01 sec and a major frequency content up to 6 Hz is simulated.

Training

Given the simulation dataset, we can proceed to train a FNO model in a supervised capacity, where the goal is to learn a model that can reliably output a solution to the wave equation for arbitrary input conditions. The training is performed using batch gradient descent, where the parameters of the FNO are updated to minimize the error against the "true" solutions. For the acoustic FNO, a mean-square error

loss function is used. We use a batch size of 30 simulations and train the model for a maximum of 300 epochs. We use all but 200 of the simulations for training, and set aside the remainder for cross-validation of the model. The time required to train the model from scratch is approximately 18 hours using 6 NVIDIA Tesla V100 GPUs.

For the elastic FNO, a mixture of L1 and L2 losses is used. We use a batch size of 8 and train for 100 epochs. We use 90% of the simulations for training and the remainder for validation. This takes approximately 40 minutes per epoch on a single NVIDIA Tesla V100 GPU.

7.4 Experimental results for acoustic wave propagation and inversion

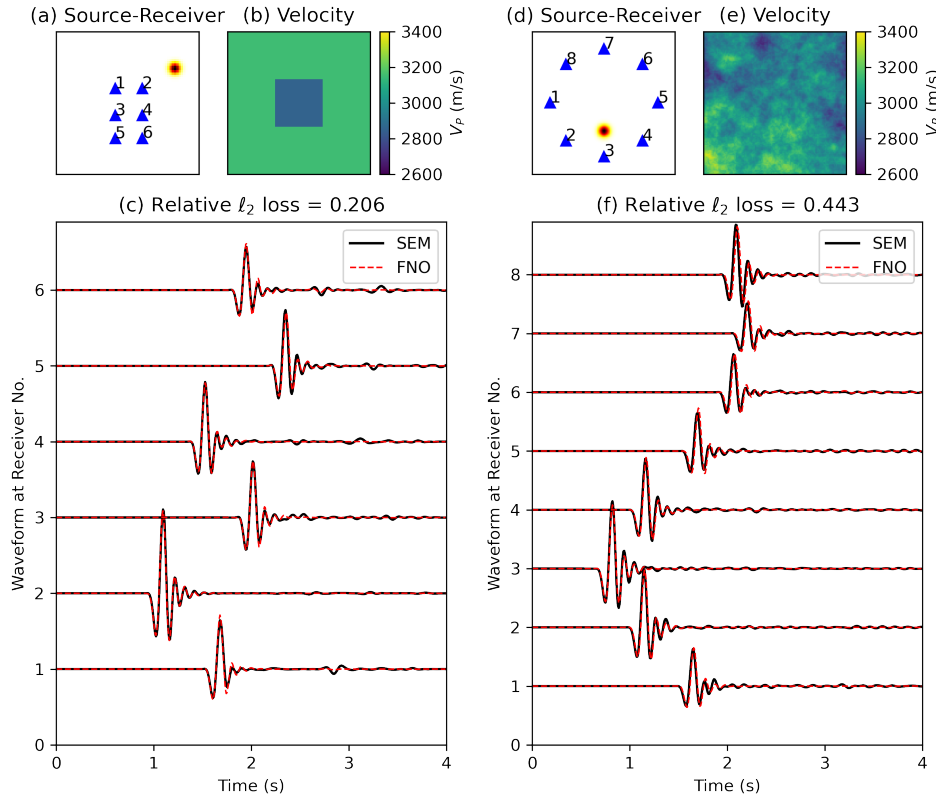


Figure 7.4: Model generalization experiments. (a) the source-receiver locations with receivers in blue and source in red; (b) a velocity model with a homogeneous background of 3 km/s and a 5% square anomaly; (c) waveform simulated with SEM (black) and FNO (red). (d-f) same as (a-c), but for an input velocity model with 2x finer resolution than trained on. These experiments show that the model is not just memorizing the solutions, but is able to generalize to entirely new conditions.

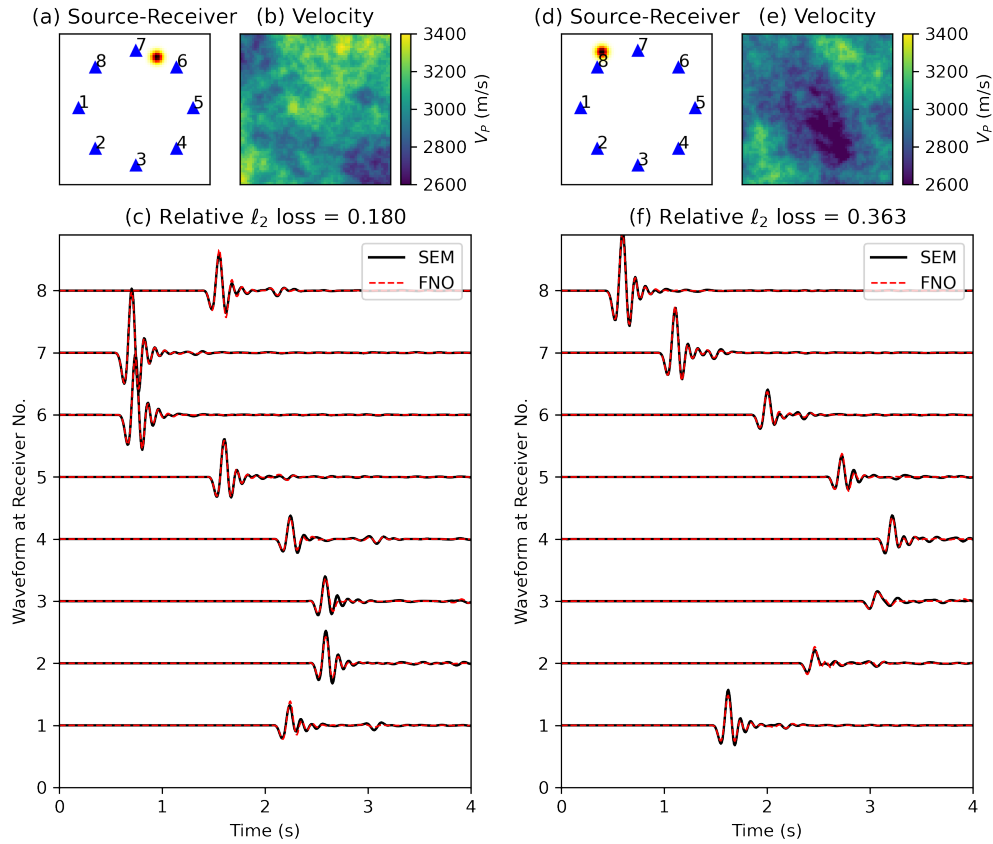


Figure 7.2: Examples of two validation wavefield simulations from the trained FNO model. (a) the source-receiver locations with receivers in blue and source in red; (b) the velocity structure; (c) waveforms simulated with SEM (black) and FNO (red). (d-f) same as (a-c), but for a different velocity model. The relative ℓ_2 loss of the two examples are 0.180 and 0.363, respectively, which are representative of the entire validation data set with an average ℓ_2 loss of 0.273. We demonstrate that the FNO simulation results are able to capture both the major arrivals as well as some reflections.

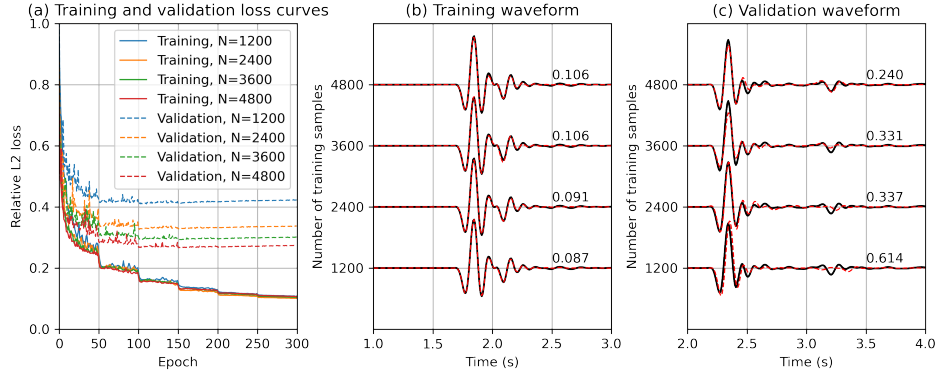


Figure 7.3: Model performance as a function of the number of training samples. (a) Training and validation loss curves as a function of different numbers of training samples. (b) Example waveform fitting of a single training example from models trained with varying number of training examples. (c) Example waveform fitting of a single validation example from models trained with varying number of training examples. The numbers to the right of each waveform shows the relative ℓ_2 misfit. This shows that the model trained on 4800 samples is able to capture the reflections, whereas the model trained on smaller number of samples does not generalize to reflections in the validation example.

Initial wavefield demonstration

Figure 7.2 shows two example wavefields corresponding to two different velocity models, each with a different source. The spectral element solution is shown alongside the wavefield predicted by the FNO for the 8 different receivers (blue triangles). For these examples, the input velocity model is 64×64 . The relative ℓ_2 loss of the FNO wavefields are 0.180 and 0.363. These examples are representative of the entire validation dataset, which has a loss of 0.273 relative to the spectral element solutions.

The number of simulations needed for training

Once fully trained, the FNO can evaluate a new solution in a fraction of a second, and thus the time to train the FNO will be the vast majority of the needed compute time. A primary concern about the computational demands of the FNO approach is therefore the number of simulations needed for training. Here, we examine how the number of training simulations influences the accuracy of the solution. We create a series of tests in which the number of training simulations is varied from 800 at the fewest, to 4800 at the most. The results are shown in Figure 7.3 where we show the FNO wavefield predictions for each dataset. Even with 1200 training samples, there

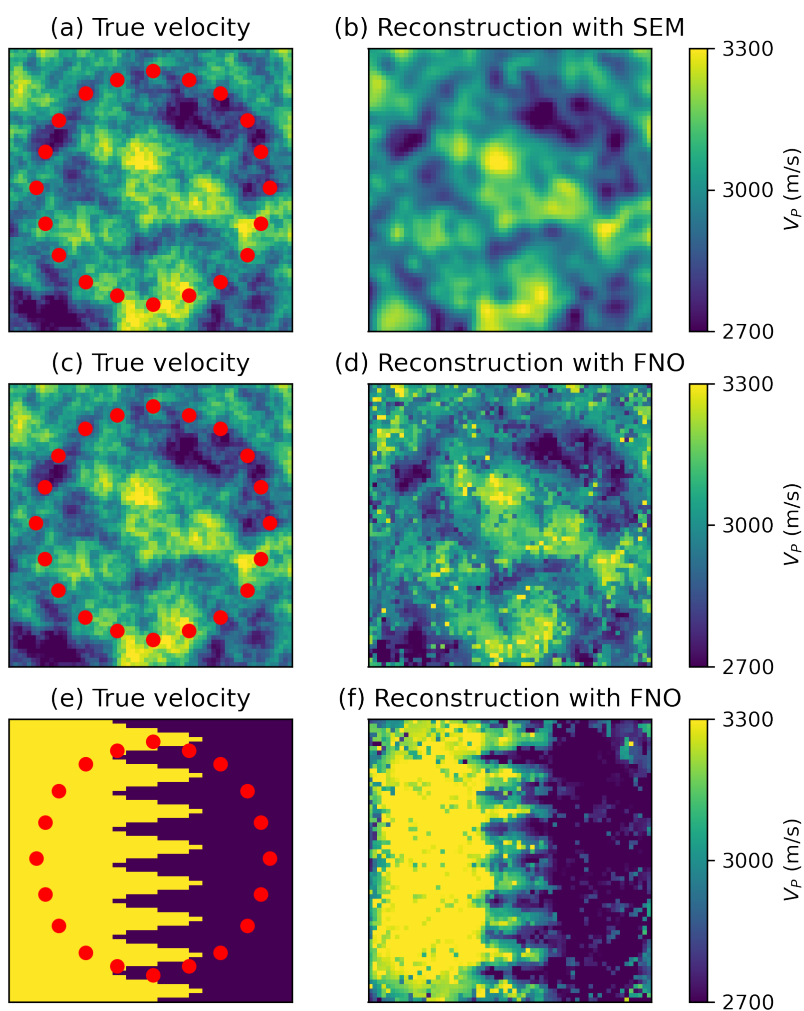


Figure 7.5: Example of a full waveform inversion using FNO. (a)(c)(e) True velocity models with source locations indicated by red circles and receivers placed at every node of the 64×64 grid ($10 \text{ km} \times 10 \text{ km}$ region). (b) Reconstruction using SEM and adjoint tomography. (d)(f) Reconstruction using FNO as the forward model and automatic differentiation to compute gradients. No regularization is used for these experiments.

is no indication that there is overfitting since the training waveform error is similar across different models (Figure 7.3a,b). Training using just 1200 simulations is able to predict the major arrival. Increasing number of training samples provides a better fit of the reflections (e.g. 3.2 sec in Figure 7.3c).

Generalization to arbitrary velocity models

The FNO was trained only on velocity models drawn from Gaussian random fields, and while this family of functions is broad, it does not include some types of functions that exist in the Earth, such as discontinuous functions. This rises the question of whether the FNO can still generalize well to these cases. Figure 7.4a-c shows an example of a predicted wavefield for a velocity model containing a constant velocity square embedded within a homogeneous medium. While the velocity model itself is rather simple, it is actually very far removed from the characteristics of the random fields that the FNO was trained on and represents a challenging example. We can see that the predicted wavefield does a very good job of approximating the wavefield compared to the ground truth. We believe the small residual errors can be reduced with better hyperparameter selection.

Generalization to higher resolution grids

FNO can be viewed in some sense as a method for learning generalized Green's functions valid for arbitrary boundary conditions. Since it is intrinsically learning a mapping between function spaces, the FNO is theoretically independent of the resolution at which the functions are discretized (this is only a requirement for evaluation on a computer). One important advantage of this is that the FNO can be trained on velocity models with a certain grid spacing, and then be evaluated on velocity models with a different grid spacing at inference time. Here, we are not simply talking about interpolating the wavefield after solving the PDE; but rather, the solutions to the PDE can actually be evaluated on a higher resolution velocity model with negligible extra compute cost. To demonstrate this, Figure 7.4d-f shows the FNO prediction for a random velocity field with 3x higher resolution 128x128 than the models used during training, alongside the spectral element solution. The FNO solution closely approximates the spectral element solution. Note that the velocity models with different meshes have the same roughness as the training data set. Resolving more rough structure with denser spacing can be achieved by training with many more GRFs with varying correlation length scales and variance.

Full waveform inversion with Neural Operators

One of the most useful applications of wavefield simulations is in inversion, to image the Earth’s interior. The adjoint-state method is an approach to efficiently compute the gradients of an objective function with respect to parameters of interest and can be used for seismic tomography (e.g. [5], [9]). Neural Operators are differentiable by design, which enables gradient computation with reverse-mode automatic differentiation. Automatic differentiation has been shown to be mathematically equivalent to the adjoint-state method [27]. This allows for the gradients of the wavefield to be determined with respect to the inputs (velocity model and source location).

Figure 7.5 demonstrates our FWI performance. For each case, we compute synthetic observations using the source distribution as shown (red circles), taking every point in the 64x64 grid as a receiver. The observations are noise-free for this experiment. Then, we perform tomography by starting with a homogeneous initial velocity model and forward propagating a wavefield with the FNO for each source. We calculate the loss $L = \sum_i \sum_j [u^{obs}(x_i, x_j) - u^{pred}(x_i, x_j)]^2$, and compute ∇L with automatic differentiation. The velocity model is then iteratively updated with gradient descent for 1000 iterations using the Adam optimizer [28] and a learning rate of 0.01. For comparison, Fig. 7.5ab shows the imaging result using SEM and adjoint-state method, with a relative ℓ_2 misfit between the inverted and true velocity model of 0.0289. Fig. 7.5cd shows the result for the same velocity structure using FNO and automatic differentiation, with a misfit of 0.0319. Fig. 7.5ef is designed to demonstrate sharp discontinuous changes with a short wavelength. The results demonstrate the remarkable capabilities of FNO to learn a general solution operator.

We note that our FWI approach does not require an adjoint wavefield to be computed, nor a cross-correlation; the gradients can be rapidly computed with GPUs using automatic differentiation. The rapid simulation makes it substantially more efficient than adjoint methods. For these experiment, 20 sources take ~ 1 second for one tomographic iteration including the costs of computing the forward model, while the spectral element method with adjoint methods takes ~ 100 seconds for one tomographic iteration. These time measurements are from using only a single NVIDIA Tesla V100 GPU.

7.5 Experimental results for elastic wave propagation and inversion

Generalizability to arbitrary velocity structure or discretization The U-NO model is trained on random velocity models generated with the von Karman corre-

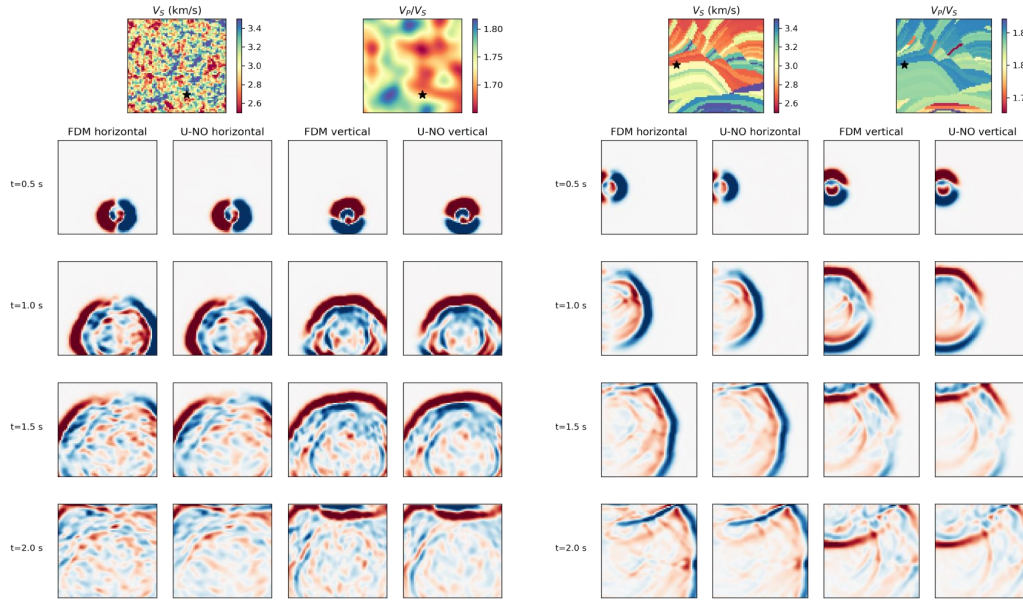


Figure 7.6: Model generalization experiments 1: random fields of V_s and V_p/V_s model with 4-times roughness of the training data. Relative L2 loss of the U-NO simulation is 0.182.

lation function, which can best mimic the Earth’s heterogeneous velocity distribution. We show by example that the U-NO model, although trained on random velocity models with parameters described above, is applicable to arbitrary velocity models. These outcomes are in fact expected from theoretical grounds because von Karman random fields are able to approximate any physical function to arbitrary accuracy.

One example is with velocity models from a von Karman-type random distribution, but with a different covariance function than the one used for the training data. We increase the roughness of the velocity structure by a factor of four by decreasing the correlation length of V_s and V_p/V_s to only one-fourth that of the training data. As shown in Figure 7.6, the wavefield snapshot has more coda than with the smoother models because of the scattering from increased heterogeneity. However, the coda waves are well modeled by U-NO when compared to the ground truth simulation by FDM.

The distribution of the relative L2-norm misfit and cross-correlation coefficient are plotted in Figure 7.7. In the case of the Marmousi model, the extended tail of the

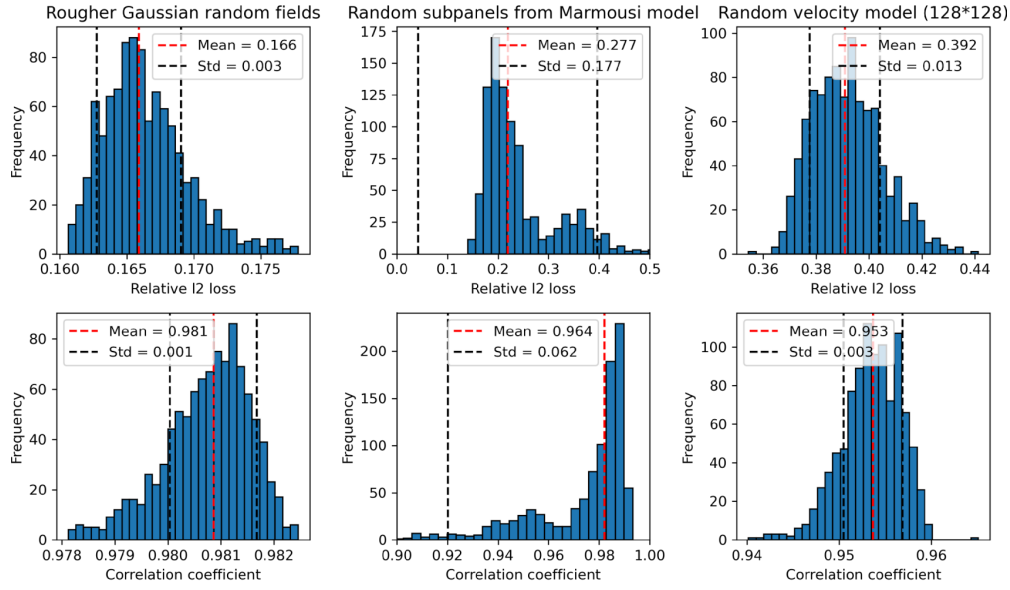


Figure 7.7: Output evaluation. Distribution of relative L2 loss (top) and correlation coefficient (bottom) between the U-NO predictions and ground truth. From left to right, the columns are corresponding to the experiments in Figure 3, 5, 6. The red and black dashed lines mark the mean and standard deviation of the histograms.

histogram is attributed to the model's imbalanced complexity. In general, however, we see a very high cross-correlation coefficient (>0.95) between U-NO prediction and ground truth, confirming its robust generalizability.

Full waveform inversion One of the most important applications of wavefield simulations is in full-waveform inversion (FWI), which uses the full recorded waveform to image the Earth's interior. The adjoint-state method is the traditional approach to compute the gradients of an objective function with respect to parameters of interest [29], [30]. Neural operators are differentiable by design, which enables gradient computation with reverse-mode automatic differentiation. It has been shown that automatic differentiation and the adjoint approach are mathematically equivalent [27]. Hence, the trained U-NO model allows for convenient FWI and the associated speed and accuracy should depend only on the forward modeling part.

We demonstrate the inversion performance using the velocity structure of random subpanels in the Marmousi model [31]. The synthetic waveform data are simulated

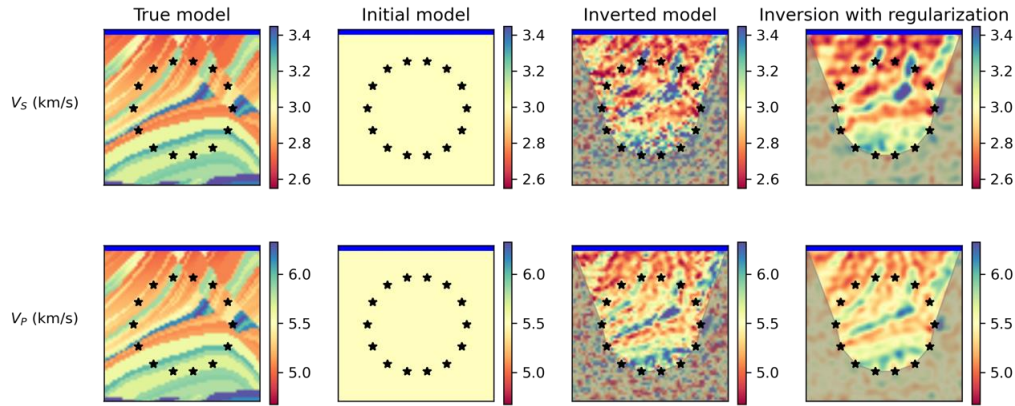


Figure 7.8: Full waveform inversion. The inversion for a random subpanel from the Marmousi model. 14 sources are placed in a ring shape and receivers are placed at every node of the 64×64 grid. From left to right, the columns represent true velocity model, initial model for inversion, inverted model without regularization, and inverted model with damping and smoothing. The top and bottom rows are Vs and Vp models, respectively.

with FDM [26] using 14 sources distributed in a ring shape. In Figure 7.8, we use the true source location, receivers on all 64×64 grids and noise-free waveform data; the goal here is not to demonstrate resolution, but rather the computational accuracy of the method. We then invert for Vp and Vs simultaneously by starting with homogeneous initial Vp and Vs models and forward propagating the wavefield with the U-NO for each source. The misfit is defined by the mean square error between the forward modeled and true wavefield. The gradient of the misfit with respect to Vp and Vs can be computed through automatic differentiation. Vp and Vs are then iteratively updated with gradient descent for 100 iterations using the Adam optimizer [32] with a learning rate of 0.01. Each iteration takes only about 1.4 sec by taking advantage of U-NO forward computation. The results in Figure 8 show a relative L2-norm misfit between the true and inverted model of only 3%. This successful inversion, in turn, further validates the accuracy of forward modeling with U-NO.

Besides the fact that the inversion target velocity model is quite different from the smooth random fields in the training dataset, this experiment itself is difficult due to conventional problems in full-waveform inversion, such as cycle skipping (multiple local maxima in the least-squares misfit function).

7.6 Conclusions

This chapter presents a prototype for applying Neural Operators to the 2D acoustic wave equation and extends this prototype to the 2D elastic wave equation. In our tests, we found that only a few thousand simulations were needed to train a FNO model, and from there, required negligible time to compute a new solution. Since FNO can be trained on lower resolution simulations and then generalize to higher resolution solutions once trained, this results in substantially faster computations than using traditional numerical methods at the full resolution.

One of the limitations of the approach is that the solutions are approximate, as seen in several of the figures. However, since this is a learning-based approach, the performance can be improved in the future by using a better model architecture, thorough tuning of hyperparameters, improving the size of the training dataset, using a more appropriate objective function, and various other factors. Additionally, as new developments within machine learning emerge in this area, they would be able to be incorporated. Thus, these performance metrics should only be viewed as a starting point. For some applications, the error may be enough of an issue and traditional numerical methods may be preferable; however for many other situations in geophysics, a reasonably accurate solution may be acceptable.

Among the most exciting benefits of our approach is that by training the FNO on random velocity models, the FNO is able to produce solutions for arbitrary velocity models. This is because FNO learns a general solution operator to the PDE, and not specifically the velocity model. This means that the model does not need to be retrained for each region. Thus, the approach offers the potential for a single FNO model to be used by the entire seismology community for any region of a similar size. While the initial cost of training a FNO and performing the training simulations may be expensive, it only needs to be done a single time for the community as a whole.

References

- [1] Yan Yang, Angela F. Gao, Jorge C. Castellanos, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Seismic wave propagation and inversion with neural operators,” *The Seismic Record*, vol. 1, no. 3, pp. 126–134, 2021. doi: <https://doi.org/10.1785/0320210026>,
- [2] Yan Yang, Angela F. Gao, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross, “Rapid seismic waveform modeling and inversion with neural operators,” *IEEE Transactions on Geoscience and Remote Sensing*,

- vol. 61, pp. 1–12, 2023. DOI: <https://doi.org/10.1109/TGRS.2023.3264210>,
- [3] A. J. Rodgers, N. A. Petersson, A. Pitarka, D. B. McCallen, B. Sjogreen, and N. Abrahamson, “Broadband (0–5 Hz) Fully Deterministic 3D Ground-Motion Simulations of a Magnitude 7.0 Hayward Fault Earthquake: Comparison with Empirical Ground-Motion Models and 3D Path and Site Effects from Source Normalized Intensities,” en, *Seismological Research Letters*, vol. 90, no. 3, pp. 1268–1284, May 2019, Publisher: GeoScienceWorld, ISSN: 0895-0695. DOI: 10.1785/0220180261. [Online]. Available: <https://pubs.geoscienceworld.org/ssa/srl/article/90/3/1268/568984/Broadband-0-5-Hz-Fully-Deterministic-3D-Ground> (visited on 11/03/2020).
 - [4] R. Graves and A. Pitarka, “Kinematic Ground-Motion Simulations on Rough Faults Including Effects of 3D Stochastic Velocity Perturbations,” en, *Bulletin of the Seismological Society of America*, vol. 106, no. 5, pp. 2136–2153, Oct. 2016, Publisher: GeoScienceWorld, ISSN: 0037-1106. DOI: 10.1785/0120160088. [Online]. Available: <https://pubs.geoscienceworld.org/ssa/bssa/article/106/5/2136/350935/Kinematic-Ground-Motion-Simulations-on-Rough> (visited on 11/03/2020).
 - [5] C. Tape, Q. Liu, A. Maggi, and J. Tromp, “Adjoint Tomography of the Southern California Crust,” en, *Science*, vol. 325, no. 5943, pp. 988–992, Aug. 2009, Publisher: American Association for the Advancement of Science Section: Report, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1175298. [Online]. Available: <https://science.sciencemag.org/content/325/5943/988> (visited on 07/01/2021).
 - [6] J. Virieux and S. Operto, “An overview of full-waveform inversion in exploration geophysics,” *GEOPHYSICS*, vol. 74, no. 6, WCC1–WCC26, Nov. 2009, Publisher: Society of Exploration Geophysicists, ISSN: 0016-8033. DOI: 10.1190/1.3238367. [Online]. Available: <https://library.seg.org/doi/full/10.1190/1.3238367> (visited on 08/06/2021).
 - [7] A. Fichtner, B. L. N. Kennett, H. Igel, and H.-P. Bunge, “Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods,” *Geophysical Journal International*, vol. 179, no. 3, pp. 1703–1725, Dec. 2009, ISSN: 0956-540X. DOI: 10.1111/j.1365-246X.2009.04368.x. [Online]. Available: <https://doi.org/10.1111/j.1365-246X.2009.04368.x> (visited on 08/06/2021).
 - [8] E.-J. Lee, P. Chen, T. H. Jordan, P. B. Maechling, M. A. M. Denolle, and G. C. Beroza, “Full-3-D tomography for crustal structure in Southern California based on the scattering-integral and the adjoint-wavefield methods,” en, *Journal of Geophysical Research: Solid Earth*, vol. 119, no. 8, pp. 6421–6451, 2014, eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JB011346>, ISSN: 2169-9356. DOI: 10.1002/2014JB011346. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JB011346>

//agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JB011346 (visited on 08/09/2021).

- [9] L. Gebraad, C. Boehm, and A. Fichtner, “Bayesian Elastic Full-Waveform Inversion Using Hamiltonian Monte Carlo,” en, *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 3, e2019JB018428, 2020, ISSN: 2169-9356. DOI: 10.1029/2019JB018428. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JB018428> (visited on 08/06/2021).
- [10] Z. Duputel, J. Jiang, R. Jolivet, *et al.*, “The Iquique earthquake sequence of April 2014: Bayesian modeling accounting for prediction uncertainty,” en, *Geophysical Research Letters*, vol. 42, no. 19, pp. 7949–7957, 2015, ISSN: 1944-8007. DOI: 10.1002/2015GL065402. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL065402> (visited on 11/01/2019).
- [11] L. Ye, T. Lay, H. Kanamori, and L. Rivera, “Rupture characteristics of major and great $m_w \geq 7.0$ megathrust earthquakes from 1990 to 2015: 2. Depth dependence,” en, *Journal of Geophysical Research: Solid Earth*, vol. 121, no. 2, 2015JB012427, Feb. 2016, ISSN: 2169-9356. DOI: 10.1002/2015JB012427. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.1002/2015JB012427/abstract> (visited on 10/05/2017).
- [12] X. Wang and Z. Zhan, “Moving from 1-D to 3-D velocity model: Automated waveform-based earthquake moment tensor inversion in the Los Angeles region,” *Geophysical Journal International*, vol. 220, no. 1, pp. 218–234, Jan. 2020, ISSN: 0956-540X. DOI: 10.1093/gji/ggz435. [Online]. Available: <https://doi.org/10.1093/gji/ggz435> (visited on 08/06/2021).
- [13] J. D. Smith, K. Azizzadenesheli, and Z. E. Ross, “EikoNet: Solving the Eikonal Equation With Deep Neural Networks,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2020, Conference Name: IEEE Transactions on Geoscience and Remote Sensing, ISSN: 1558-0644. DOI: 10.1109/TGRS.2020.3039165.
- [14] B. Moseley, A. Markham, and T. Nissen-Meyer, “Solving the wave equation with physics-informed deep learning,” *arXiv:2006.11894 [physics]*, Jun. 2020, arXiv: 2006.11894. [Online]. Available: <http://arxiv.org/abs/2006.11894> (visited on 08/06/2021).
- [15] B. Moseley, T. Nissen-Meyer, and A. Markham, “Deep learning for fast simulation of seismic waves in complex media,” English, *Solid Earth*, vol. 11, no. 4, pp. 1527–1549, Aug. 2020, Publisher: Copernicus GmbH, ISSN: 1869-9510. DOI: 10.5194/se-11-1527-2020. [Online]. Available: <https://se.copernicus.org/articles/11/1527/2020/> (visited on 08/06/2021).

- [16] B. Moseley, A. Markham, and T. Nissen-Meyer, “Finite Basis Physics-Informed Neural Networks (FBPINNs): A scalable domain decomposition approach for solving differential equations,” *arXiv:2107.07871 [physics]*, Jul. 2021, arXiv: 2107.07871. [Online]. Available: <http://arxiv.org/abs/2107.07871> (visited on 08/06/2021).
- [17] W. Zhu, K. Xu, E. Darve, B. Biondi, and G. C. Beroza, “Integrating Deep Neural Networks with Full-waveform Inversion: Reparametrization, Regularization, and Uncertainty Quantification,” *arXiv:2012.11149 [physics]*, Dec. 2020, arXiv: 2012.11149. [Online]. Available: <http://arxiv.org/abs/2012.11149> (visited on 08/06/2021).
- [18] W. Zhang and J. Gao, “Deep-Learning Full-Waveform Inversion Using Seismic Migration Images,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, 2021, Conference Name: IEEE Transactions on Geoscience and Remote Sensing, ISSN: 1558-0644. DOI: 10.1109/TGRS.2021.3062688.
- [19] C. Xiao, Y. Deng, and G. Wang, “Deep-Learning-Based Adjoint State Method: Methodology and Preliminary Application to Inverse Modeling,” en, *Water Resources Research*, vol. 57, no. 2, e2020WR027400, 2021, _eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR027400>, ISSN: 1944-7973. DOI: 10.1029/2020WR027400. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR027400> (visited on 08/09/2021).
- [20] J. D. Smith, Z. E. Ross, K. Azizzadenesheli, and J. B. Muir, “HypoSVI: Hypocenter inversion with Stein variational inference and Physics Informed Neural Networks,” *arXiv:2101.03271 [physics]*, Jan. 2021, arXiv: 2101.03271. [Online]. Available: <http://arxiv.org/abs/2101.03271> (visited on 08/06/2021).
- [21] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.
- [22] Z. Li, N. Kovachki, K. Azizzadenesheli, *et al.*, “Neural Operator: Graph Kernel Network for Partial Differential Equations,” *arXiv:2003.03485 [cs, math, stat]*, Mar. 2020, arXiv: 2003.03485. [Online]. Available: <http://arxiv.org/abs/2003.03485> (visited on 07/13/2021).
- [23] Z. Li, N. Kovachki, K. Azizzadenesheli, *et al.*, “Multipole Graph Neural Operator for Parametric Partial Differential Equations,” *arXiv:2006.09535 [cs, math, stat]*, Oct. 2020, arXiv: 2006.09535. [Online]. Available: <http://arxiv.org/abs/2006.09535> (visited on 07/13/2021).
- [24] Z. Li, N. Kovachki, K. Azizzadenesheli, *et al.*, “Fourier Neural Operator for Parametric Partial Differential Equations,” *arXiv:2010.08895 [cs, math]*, May 2021, arXiv: 2010.08895. [Online]. Available: <http://arxiv.org/abs/2010.08895> (visited on 06/04/2021).

- [25] M. Afanasiev, C. Boehm, M. van Driel, *et al.*, “Modular and flexible spectral-element waveform modelling in two and three dimensions,” *Geophysical Journal International*, vol. 216, no. 3, pp. 1675–1692, Mar. 2019, issn: 0956-540X. doi: 10.1093/gji/ggy469. [Online]. Available: <https://doi.org/10.1093/gji/ggy469> (visited on 08/05/2021).
- [26] D. Li, D. Helmberger, R. W. Clayton, and D. Sun, “Global synthetic seismograms using a 2-d finite-difference method,” *Geophysical Journal International*, vol. 197, no. 2, pp. 1166–1183, 2014.
- [27] W. Zhu, K. Xu, E. Darve, and G. C. Beroza, “A general approach to seismic inversion with automatic differentiation,” *Computers & Geosciences*, vol. 151, p. 104751, 2021.
- [28] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Dec. 2014, arXiv: 1412.6980. [Online]. Available: <http://arxiv.org/abs/1412.6980> (visited on 08/08/2017).
- [29] C. Tape, Q. Liu, A. Maggi, and J. Tromp, “Adjoint tomography of the southern california crust,” *Science*, vol. 325, no. 5943, pp. 988–992, 2009.
- [30] A. Fichtner, B. L. Kennett, H. Igel, and H.-P. Bunge, “Full seismic waveform tomography for upper-mantle structure in the australasian region using adjoint methods,” *Geophysical Journal International*, vol. 179, no. 3, pp. 1703–1725, 2009.
- [31] R. Versteeg, “The marmousi experience: Velocity model determination on a synthetic complex data set,” *The Leading Edge*, vol. 13, no. 9, pp. 927–936, 1994.
- [32] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

Chapter 8

CONCLUSION

In this thesis, I have discussed new methods and techniques for solving inverse problems, primarily focusing on non-linear imaging inverse problems. These methods are designed in tandem with how scientists use and interpret the data. The key contributions are as follows:

Contributions

- An expectation maximization inspired method for solving blind inverse problems using probabilistic generative models, namely normalizing flows, to estimate the posterior.
- A technique for leveraging the shared structure across corrupted measurements to solve imaging inverse problems without Bayesian priors.
- Rich 3D maps identifying and quantifying regions of melt that underly the Kilauea caldera.
- A convex approach to solving the nonlinear eikonal tomography problem, which recasts the problem into a constrained inpainting in the measurement space.
- Mesh-free surrogate models for seismic wave propagation, enabling faster forward and backward simulation.

Directions for Future Work

The results of this thesis present several interesting directions for future work.

1. Building interpretability directly into imaging problems: For many scientific imaging problems, a single image is not enough to draw robust conclusions. A key question is how can we build interpretable uncertainty and use it to inform downstream scientific conclusions. For instance, per-pixel uncertainties on velocity have limited scientific utility, as discussed in Chapter 5. Building semantic uncertainty could unlock the ability to interpret our uncertainty, which in turn strengthen existing scientific conclusions or enable new insights.

2. *Multi-modal physics-guided inversion:* In self-driving, there is the problem of sensor fusion. Similarly, many scientific imaging problems have similar problems where different types of measurements highlight different structures. Traditional methods currently leverage information from a single measurement modality to probe at the unknown information. For instance, in Chapters 3, 6, and 5, we focus on imaging using just time-of-flight information. However, teleseismic waves can give rich information of large scale features, receiver-function information can give rich information of depths of sharp boundaries, and the full-wave information can give rich information on high frequency features. Using a single data source limits the ability to draw robust conclusions from the entire suite of observations. Instead, fusing these multi-modal data sources can give more confidence in results.

3. *Alternative strategies for forward and gradient modeling for non-linear inverse problems:* Our understanding of linear inverse problems is quite rich, but our insight into feasibility and quality from solving non-linear inverse problems remains limited. This is especially necessary as the field turns towards using surrogate forward models for faster inversion. How can we trust the reconstructions when we don't understand how it works? As shown in Chapter. 6, we show a convex relaxation of the eikonal imaging problem, which in its Lagrangian formulation, looks very similar imaging with a physics informed neural network where the neural network is a neural field. Alternative strategies, like EikoLift, enable the interrogation of modern techniques to gain insights into the behavior of learning based inversion schemes.

4. *Uncovering the structure of forward models through uncovering posterior invariance:* It is well known that for the well-studied imaging inverse problem of Fourier phase-retrieval, losing the phase information results in shift ambiguity in the reconstruction. As discussed in Chapter 4, we can add a loss to ensure that the center of the mass of the object is at the center of the image. Alternatively, we can infer the entire posterior by allowing for spatially shifted results. This is challenging to infer directly from data, but encoding these invariance apriori can be limiting in non-linear inverse problems where we do not know what this structure is exactly. Learning the structure in tandem with solving the inverse problem would improve the optimization landscape while improving the uncertainty in non-linear imaging modalities like seismic imaging.

Appendix A

IMAGING WITH LIMITED PRIORS: SUPPLEMENTARY MATERIALS

A.1 Experimental details

Model selection We parameterize the latent variational distributions $q_\phi(z)$ with normalizing flows l_ϕ . In particular, suppose our latent variables $z = l_\phi(s)$ where $s \sim \mathcal{N}(0, I)$. Then,

$$\log q_\phi(z) = \log \pi(s) - \log \left| \det \frac{dl_\phi(s)}{ds} \right|$$

where $\pi(\cdot)$ denotes the likelihood of the base distribution of s , which in this case is a Gaussian. For our prior on z given G_θ , we use an isotropic Gaussian prior. Then the ELBOProxy equation becomes

$$\begin{aligned} \text{ELBOProxy}(G_\theta, q_\phi; y) &:= \mathbb{E}_{z \sim l_\phi(z)} [\log p(y|G(z)) \\ &\quad + \log p(z|G) - \log q_\phi(z)] \\ &= \mathbb{E}_{s \sim \mathcal{N}(0, I)} \left[\log p(y|G_\theta(l_\phi(s))) \right. \\ &\quad + \frac{1}{2} \|l_\phi(s)\|_2^2 - \log \pi(s) \\ &\quad \left. + \log \left| \det \frac{dl_\phi(s)}{ds} \right| \right]. \end{aligned}$$

Since the expectation over $\log \pi(s)$ is constant with respect to the parameters ϕ , we arrive at

$$\begin{aligned} \text{ELBOProxy}(G_\theta, q_\phi; y) &= \mathbb{E}_{s \sim \mathcal{N}(0, I)} \left[\log p(y|G(l_\phi(s))) \right. \\ &\quad + \frac{1}{2} \|l_\phi(s)\|_2^2 \\ &\quad \left. + \log \left| \det \frac{dl_\phi(s)}{ds} \right| \right]. \end{aligned}$$

Learning the IGM We parameterize the posterior using Gaussian distributions due to memory constraints; the memory of the whole setup with Normalizing Flows using 75 images is around 5 times larger than the setup using a Gaussian variational family. Note that we do not notice overfitting so there is no need for early stopping.

We train for 99,000 epochs using the Adam [1] optimizer with a learning rate of $1e-4$, which takes around 54 hours for 75 images and 4 hours for 4 images on a single NVIDIA A100 GPU. We use a batch size of 20 samples per measurement i , and we perform batch gradient descent based on all measurements $i \in \{1, \dots, N\}$.

References

- [1] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

Appendix B

HIGH RESOLUTION EIKONAL IMAGING OF KILAUEA: SUPPLEMENTARY MATERIALS

B.1 Additional results

Full resolution tests

In Fig. B.1, we show the resolution tests using a checkerboard perturbation for both V_P and V_S . We find that they have roughly the same resolvability with around 4 km resolution everywhere that has sensitivity and up to 2 km resolution near the surface.

In Fig. B.3, we show resolvability tests on possible anomalies.

In Fig. B.4, we show the resolvability test for V_P/V_S ratio synthetic anomalies obtained by decreasing the V_S background velocity, assuming a 15% melt fraction and having a V_P/V_S ratio of 2.2. The MAP estimation process can reasonably resolve the input anomalies, while some vertical smearing is observable due to the source distribution used in the inversion process.

Optimization parameters

In Fig. B.5, we show the optimization hyperparameters that we tested before selecting the optimal set. We use $k = 250$, $k_{max} = 500$, and $k_{stop} = 250$ since it fits the data while not overfitting the noise. We include the normalized data likelihood $\log p(y|V) = |y - f(V)|/b$ for each reconstruction where b is the parameter of the Laplacian distribution.

Noise model misspecification

In Fig. 5.2, we show results on reconstruction V_P when we have misestimated the noise by a factor of 2 and 5. In all settings, we are able to resolve the anomalies, but when the noise is much higher than expected, there is substantial overfitting to noise. We find it necessary to do early stopping to prevent overfitting to noise, especially when the data noise is larger than expected.

We also consider a mismatch in the noise model itself and consider Laplacian and Gaussian distributions as shown in Fig. B.6.

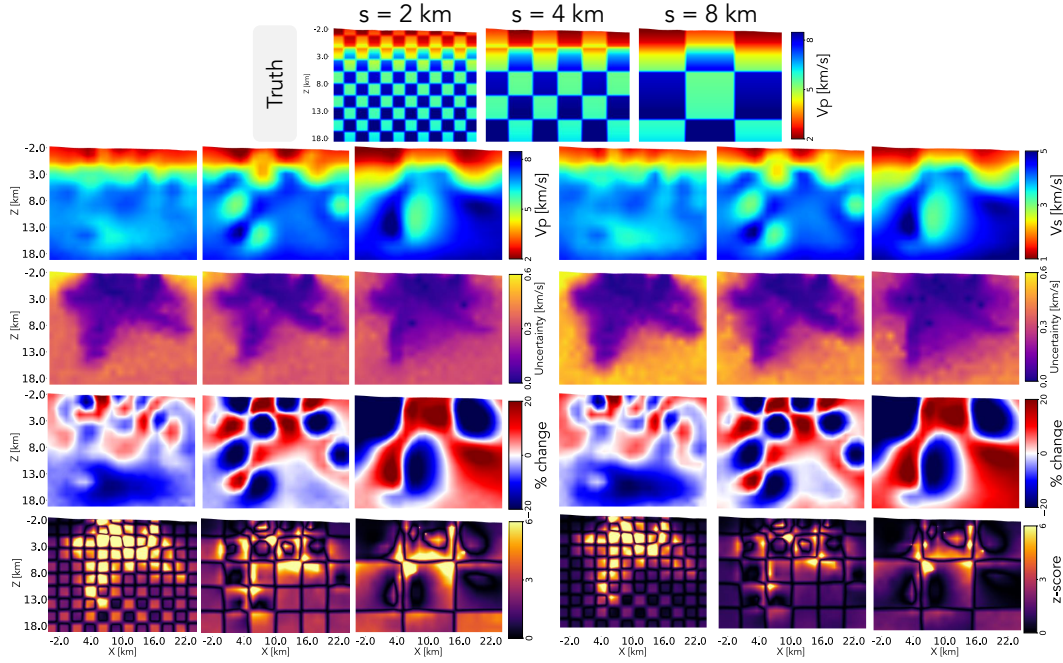


Figure B.1: **Checkerboard tests** We perform inversion on the 1D model from [1] modulated with various checkerboards of 20% increase or decrease for both V_P and V_S . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the reconstruction and relative change for A-A'. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage for both V_P and V_S . We include z-scores to quantify the accuracy of our uncertainty quantification. We find that the uncertainty is underestimated near the sharp interfaces of the checkerboard as well as in the high frequency regions that are below the resolvability of our method.

B.2 Theoretical perspective of gradient blurring for linear inverse problems

Gradient smoothing for linear inverse problems

We consider solving a linear inverse problem where $A \in \mathbb{R}^{n^2 \times n^2}$, $\mathbf{v} \in \mathbb{R}^{n^2 \times 1}$ is our target, $y \in \mathbb{R}^{n^2 \times 1}$ is our measurements, and $\mathcal{R}(\mathbf{v})$ is our regularizer by solving the following optimization problem:

$$\begin{aligned} \hat{\mathbf{v}} &= \arg \min_{\mathbf{v}} \mathcal{L} \\ &= \arg \min_{\mathbf{v}} ||y - A\mathbf{v}||_2^2 + \mathcal{R}(\mathbf{v}). \end{aligned} \quad (\text{B.1})$$

For simplicity, we assume that the regularizer is 0:

$$\hat{\mathbf{v}} = \arg \min_{\mathbf{v}} ||y - A\mathbf{v}||_2^2. \quad (\text{B.2})$$

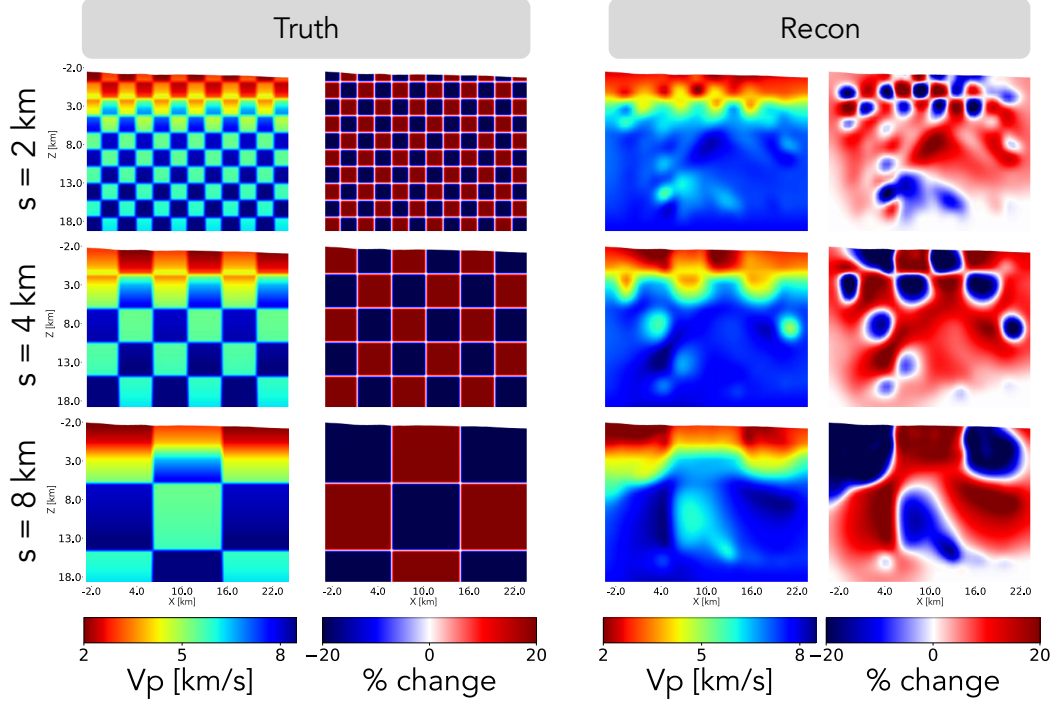


Figure B.2: **Checkerboard tests without early stopping** We perform inversion on the 1D model from [1] modulated with various checkerboards of 20% increase or decrease for both V_P and V_S . These checkerboard patterns have a size of 2 km, 4 km, and 8 km. We also show the reconstruction and relative change for A-A'. We are able to resolve around a resolution of 4 km in the upper half of the domain and at least a resolution of 8 km throughout the region with ray coverage for both V_P and V_S .

If we solve this using gradient descent, then the gradient is defined by:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = -A^T(y - A\mathbf{v}). \quad (\text{B.3})$$

We consider adding an isotropic spatial Gaussian blur G to the data-likelihood that can be decomposed as $G = G_{\text{ID}}G_{\text{ID}}^T \in \mathcal{R}^{n^2 \times n^2}$. G is self-adjoint. Applying G to the gradient gives us:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{v}} &= G [-A^T(y - A\mathbf{v})] \\ &= -GA^T(y - A\mathbf{v}). \end{aligned} \quad (\text{B.4})$$

Essentially, this is saying that the update from the data should be blurred because of our inductive biases (i.e. implicit smoothness prior on the gradient, potential artifacts of the gradient). Essentially, what this is doing is that it attenuates any high

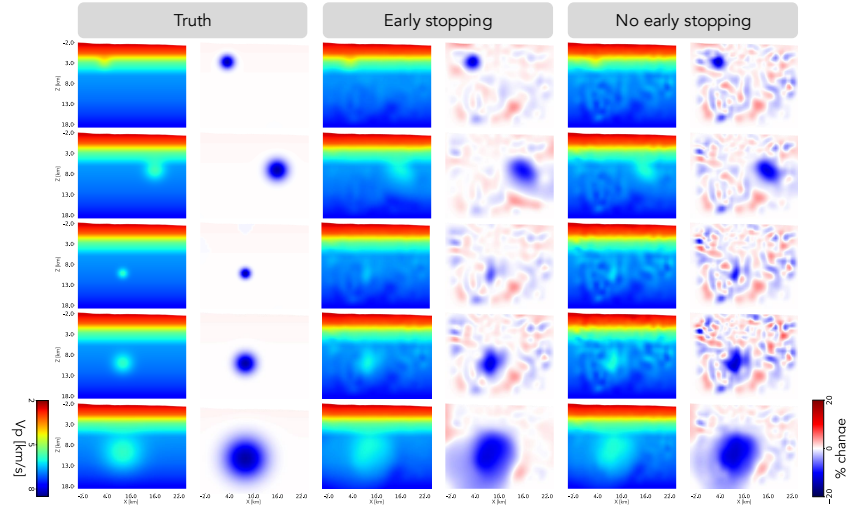


Figure B.3: **Physically motivated synthetic anomaly tests.** We perform synthetic tests to explore the sensitivity of our expected source-receiver geometry to various $\sim 10\%$ low velocity anomalies in the subsurface. For various candidate anomalies, we show the following for A-A': the true velocity model used to generate the synthetic travel times, the relative change to the initial model, and the mean of the inferred posterior and relative change to the initial model for both an $k = 250$ and $k = 500$. Note, we plot these cross sections masking out the topography.

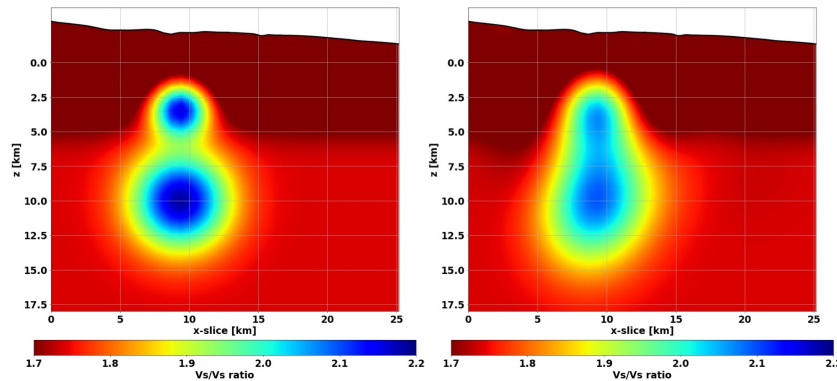


Figure B.4: **V_P/V_S ratio synthetic resolvability test.** We include two high V_P/V_S ratio anomalies within the initial velocity models to evaluate the ability (left panel). The right panel shows the result of applying MAP estimation process to the synthetic data to resolve the input anomalies.

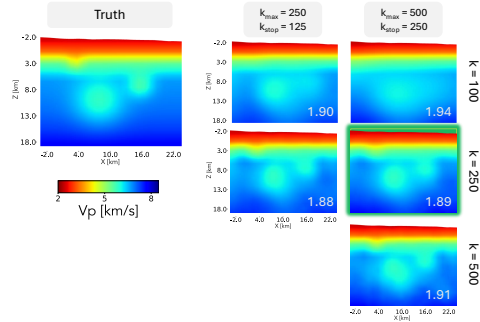


Figure B.5: **Gradient smoothing.** We visualize the reconstructions along A-A' for different values of k_{stop} , k_{max} , and k values for the gradient smoothing as described by Eq. 5.9. We find that the optimal set of parameters is $k_{stop} = 250$, $k_{max} = 500$, and $k = 250$, which is highlighted in green. We include the ground truth velocity model for reference.

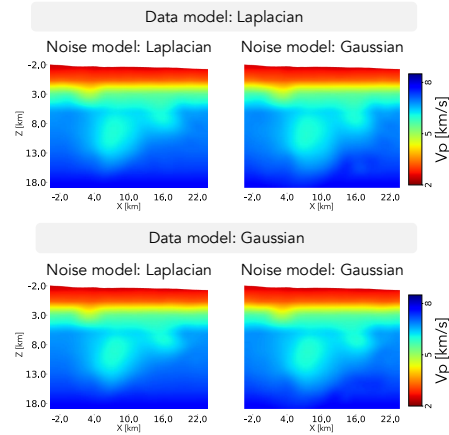


Figure B.6: **Noise model mismatch** We consider the impact of misspecifying the noise model η and/or measurement model $\log p(y|x)$. We consider the effect of mismatch in the assumed noise model distribution (e.g. Gaussian vs. Laplacian noise). For the Gaussian distribution, we assume a standard deviation $\sigma_P = .11$ and $\sigma_S = 0.31$. For the Laplacian distribution, we assume $b_P = 0.08$ and $b_S = 0.15$. The noise model and data-fit model highlighted in green is the one that we expect our real data to match. Truth is in Fig. B.5.

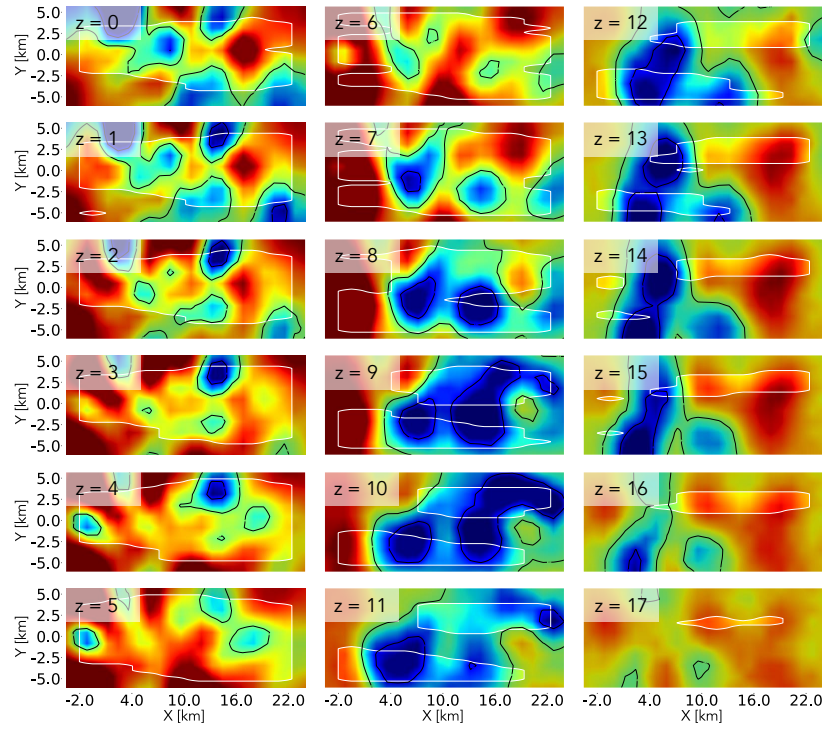


Figure B.7: **All depth slices of V_P/V_S .** Here we show all depth slices of V_P/V_S from $z = 0$ km to $z = 17$ km. Refer to the color map in Fig. 5.7 or Fig. B.10. The lighter region indicates resolvability with a checkerboard of 4 km.

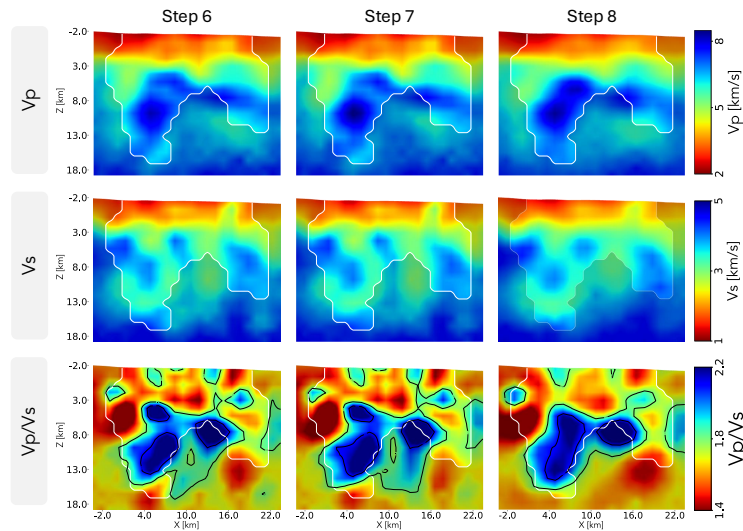


Figure B.8: **Reconstructions at various SSST steps.** Here we show reconstructions at various steps in the SSST localization procedure along the A-A' cross section. We find that the recovered structures are stable across these different steps.

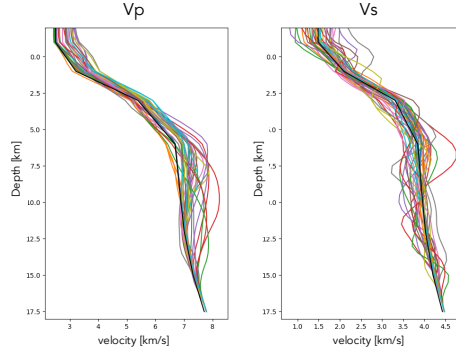


Figure B.9: **1D profiles.** Various depth slices across different lat-lon locations. The black line is the 1D model from [2] that we use for our initial model.

frequency content from the gradient of the data fit. Thus, this blurred gradient only updates the low frequencies of $\mathbf{v}$, forcing the high frequencies of the initialization to be constant. In this scheme, it is important to make sure that the frequency content of $\mathbf{v}$ matches the frequency content of the blurring operator G or else the reconstructed $\mathbf{v}$ will retain these high frequency features.

Hierarchical approaches for linear inverse problems

If instead we want to improve the reconstruction by employing a hierarchical approach, then this would result in the following optimization where $G\mathbf{v}$ represents the inversion at a specific level in the hierarchical approach:

$$\begin{aligned}\hat{\mathbf{v}} &= \arg \min_{\mathbf{v}} \mathcal{L} \\ &= \arg \min_{\mathbf{v}} \|y - AG\mathbf{v}\|_2^2\end{aligned}\tag{B.5}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = -G^T A^T (y - AG\mathbf{v}).\tag{B.6}$$

Note that G is self-adjoint (i.e. $G^T = G$). Here, the gradient appears to be different than that in Sec.B.2, but if $G\mathbf{v} = \mathbf{v}$, then they are equivalent. This only occurs when $\mathbf{v}$ is band-limited relative to the filter G .

References

- [1] G. Lin, P. M. Shearer, R. S. Matoza, P. G. Okubo, and F. Amelung, “Three-dimensional seismic velocity structure of mauna loa and kilauea volcanoes in hawaii from local seismic tomography,” *Journal of Geophysical Research: Solid Earth*, vol. 119, no. 5, pp. 4377–4392, 2014.

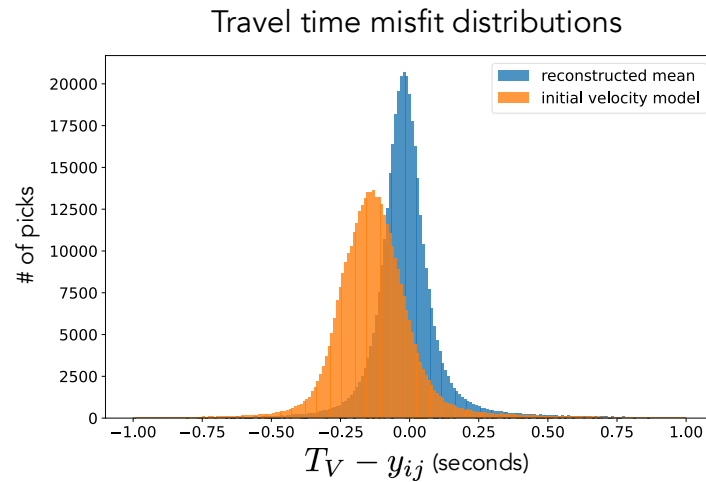


Figure B.10: **Travel time misfit distributions** Here we show the distributions of travel time misfits for both the initial velocity model and the recovered posterior mean.

- [2] T. L. Wright and F. W. Klein, “Deep magma transport at kilauea volcano, hawaii,” *Lithos*, vol. 87, no. 1-2, pp. 50–79, 2006.

Appendix C

A CONVEX RELAXATION FOR EIKONAL IMAGING: SUPPLEMENTARY MATERIALS

C.1 Derivation of EikoLift

We begin with Eq. 4 from the main text,

$$\begin{aligned} \{\hat{T}^{(i)}(x_s)\}_{i=1}^{n_r} = \arg \min_{\{T^{(i)}(\mathbf{x}_s^{(i)})\}_{i=1}^{n_r}} \sum_{i \in [n_r], j \in [n_s]} \left(T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - T^{(i)}(\mathbf{x}_s^{(j)}) \right)^2 \\ \text{s.t.} \quad \|\nabla T^{(i)}(\mathbf{x})\|_2^2 = \|\nabla T^{(k)}(\mathbf{x})\|_2^2 \quad \forall i, k, \mathbf{x}. \end{aligned} \quad (\text{C.1})$$

Substituting in $T^{(i)}(\mathbf{x}_s^{(j)}) = \alpha^{(i)T} \phi(\mathbf{x}_s^{(j)})$ where $\phi(\mathbf{x}_s^{(j)}) = (\phi_1(\mathbf{x}_s^{(j)}), \dots, \phi_K(\mathbf{x}_s^{(j)}))$:

$$\begin{aligned} \{\hat{\alpha}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{\alpha^{(i)}\}_{i=1}^{n_r}} \sum_{i \in [n_r], j \in [n_s]} \left(T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \alpha^{(i)T} \phi(\mathbf{x}_s^{(j)}) \right)^2 \\ \text{s.t.} \quad \langle \alpha^{(i)}, \nabla \phi \nabla \phi^T \alpha^{(i)} \rangle = \langle \alpha^{(k)}, \nabla \phi \nabla \phi^T \alpha^{(k)} \rangle \quad \forall i, k. \end{aligned} \quad (\text{C.2})$$

Here, $\nabla \phi(\mathbf{x}) = (\partial_k \phi_j(\mathbf{x}))_{k \in [d], j \in [K]} \in \mathbb{R}^{K \times d}$ is the Jacobian matrix of ϕ , and we have used $\langle \cdot, \cdot \rangle$ for matrix arguments to denote the trace inner product $\langle A, B \rangle = \text{Trace}(A^T B)$. We reorganize the eikonal constraint and modify the objective to use the travel times squared instead of travel times:

$$\begin{aligned} \{\hat{\alpha}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{\alpha^{(i)}\}_{i=1}^{n_r}} \sum_{i \in [n_r], j \in [n_s]} \left(T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)})^2 - \phi^T(\mathbf{x}_s^{(j)}) \alpha^{(i)} \alpha^{(i)T} \phi(\mathbf{x}_s^{(j)}) \right)^2 \\ \text{s.t.} \quad \langle \nabla \phi \nabla \phi^T, \alpha^{(i)} \alpha^{(i)T} \rangle = \langle \nabla \phi \nabla \phi^T, \alpha^{(j)} \alpha^{(j)T} \rangle \quad \forall i, j. \end{aligned} \quad (\text{C.3})$$

Substituting $M^{(i)} = \alpha^{(i)} \alpha^{(i)T}$:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} \sum_{i \in [n_r], j \in [n_s]} \left(T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)}) M^{(i)} \phi(\mathbf{x}_s^{(j)}) \right)^2 \\ \text{s.t.} \quad \langle \nabla \phi \nabla \phi^T, M^{(i)} \rangle = \langle \nabla \phi \nabla \phi^T, M^{(j)} \rangle \quad \forall i, j \\ M^{(i)} \succeq 0 \quad \forall i. \end{aligned} \quad (\text{C.4})$$

Since we want $M^{(i)} = \alpha^{(i)}\alpha^{(i)T}$ as described in Sec. 3.1, we want to find $M^{(i)}$ such that it is rank-1, so we change the problem to be a rank minimization problem and soften the eikonal constraint:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} & \sum_{i=1}^{n_r} \text{Rank}(M^{(i)}) \\ \text{s.t.} & \quad (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)})M^{(i)}\phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall i \in [n_r], j \in [n_s] \\ & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi^T(\mathbf{x}), M^{(i)} - M^{(j)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, j, \mathbf{x} \\ & \quad M^{(i)} \succeq 0 \quad \forall i. \end{aligned} \quad (\text{C.5})$$

However, rank minimization is non-convex and NP hard, so we relax it to be a trace minimization problem:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} & \sum_{i=1}^{n_r} \text{Trace}(M^{(i)}) \\ \text{s.t.} & \quad (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)})M^{(i)}\phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall i \in [n_r], j \in [n_s] \\ & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi^T(\mathbf{x}), M^{(i)} - M^{(j)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, j, \mathbf{x} \\ & \quad M^{(i)} \succeq 0 \quad \forall i. \end{aligned} \quad (\text{C.6})$$

Having a such a strong constraint for the PDE loss can yield infeasibility in the optimization problem. Thus, we weaken this to be an inequality enforced on random points $\mathbf{x}$:

$$\begin{aligned} \{\hat{M}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{M^{(i)}\}_{i=1}^{n_r}} & \sum_{i=1}^{n_r} \text{Trace}(M^{(i)}) \\ \text{s.t.} & \quad (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)})M^{(i)}\phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall i \in [n_r], j \in [n_s] \\ & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(k)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, k, \mathbf{x} \\ & \quad M^{(i)} \succeq 0 \quad \forall i. \end{aligned} \quad (\text{C.7})$$

C.2 Block diagonal formulation

To reduce the number of parameters, we parameterize $M^{(i)}$ to be a block diagonal matrix where it is non-zero only for entries that correspond to the bases related to $T^{(i)}$. Naively, if we had K basis functions for each of the n_r receivers, $M^{(i)} \in \mathbb{R}^{n_r K \times n_r K}$ would take the form of $M^{(i)} = I^{(i)} A^{(i)} I^{(i)T}$ where $A^{(i)} \in \mathbb{R}^{n_r K \times n_r K}$ and $I^{(i)} \in \mathbb{R}^{n_r K \times n_r K}$ is block diagonal with an identity matrix of size $I^{K \times K}$ from the

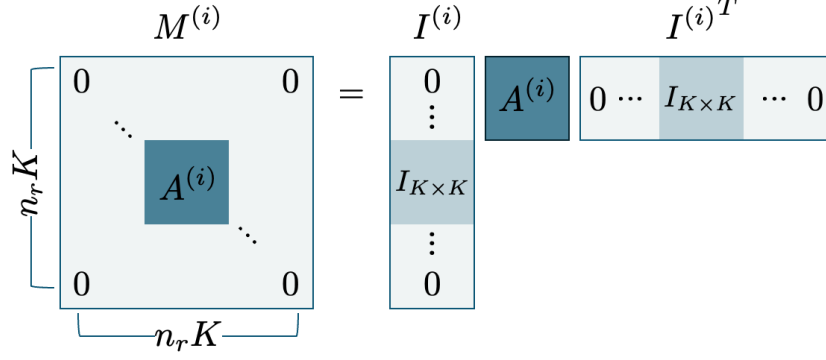


Figure C.1: **Block diagonal formulation.** This is a schematic of how we implement the block diagonal representation as described in Sec. 3.1 of the main text and in Appendix C.2.

$(i - 1)K$ to iK -th entries. That is,

$$I_{jl}^{(i)} = \begin{cases} 1, & \text{for } \forall j \in [(i - 1)K, iK], l = j \\ 0, & \text{otherwise.} \end{cases}$$

Since many of the entries in $A^{(i)}$ are irrelevant, we instead parameterize $M^{(i)} \in \mathbb{R}^{n_r K \times n_r K}$ as $M^{(i)} = I_{rect}^{(i)} A^{(i)} I_{rect}^{(i)T}$ where $A^{(i)} \in \mathbb{R}^{K \times K}$ and $I_{rect}^{(i)} \in \mathbb{R}^{n_r K \times K}$ with a $K \times K$ identity matrix starting at row iK . We visualize this in Fig. C.1. That is,

$$I_{rectjl}^{(i)} = \begin{cases} 1, & \text{for } \forall j \in [K], l = j + iK \\ 0, & \text{otherwise.} \end{cases}$$

The PSD constraint can now be on $A^{(i)}$ and the optimization now takes the form:

$$\begin{aligned} \{\hat{A}^{(i)}\}_{i=1}^{n_r} &= \arg \min_{\{A^{(i)}\}_{i=1}^{n_r}} \sum_i \text{Trace}(A^{(i)}) & (C.8) \\ \text{s.t.} \quad & \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(h)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, h, \mathbf{x} \\ & (T_{obs}^2(\mathbf{x}_r^{(i)}, \mathbf{x}_s^{(j)}) - \phi(\mathbf{x}_s^{(j)})^T M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall i, \mathbf{x}_s^{(j)} \\ & A^{(i)} \succcurlyeq 0 \quad \forall i \\ & M^{(i)} = I_{rect}^{(i)} A^{(i)} I_{rect}^{(i)T}. \end{aligned}$$

C.3 Relaxing the data-fit and Lagrangian

The convex optimization problem in Eq. C.7 (also Eq. 8 of the main text) has the nice property of optimality if it is a feasible problem, but it is computationally expensive to solve the full convex problem. Moreover, to guarantee feasibility when there is

measurement noise, ε_{data} needs to be quite large, which can yield overly smooth solutions.

We propose an alternative regime under measurement noise starting with Eq. C.7,

$$\begin{aligned} \{\hat{A}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{A^{(i)}\}_{i=1}^{n_r}} & \sum_i \text{Trace}(A^{(i)}) \\ \text{s.t.} & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(k)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, k, \mathbf{x} \\ & \quad (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)}) M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2 \leq \varepsilon_{data} \quad \forall i \in [n_r], j \in [n_s] \\ & \quad M^{(i)} = I^{(i)} A^{(i)} I^{(i)T} \quad A^{(i)} \succcurlyeq 0 \quad \forall i \end{aligned}$$

If we take the Lagrangian and introduce weight λ_{data} , we get:

$$\begin{aligned} \{\hat{A}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{A^{(i)}\}_{i=1}^{n_r}} & \underbrace{\sum_i \text{Trace}(A^{(i)})}_{\text{original objective}} + \lambda_{data} \underbrace{\sum_{i \in [n_r], j \in [n_s]} (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)}) M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2}_{\text{data constraint}} \\ \text{s.t.} & \quad \langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(k)} \rangle^2 \leq \varepsilon_{PDE} \quad \forall i, k, \mathbf{x} \\ & \quad M^{(i)} = I^{(i)} A^{(i)} I^{(i)T} \quad A^{(i)} \succcurlyeq 0 \quad \forall i \end{aligned} \quad (\text{C.9})$$

Full Lagrangian

If we take the full Lagrangian, remove the convexity requirement, and solve using gradient descent, then we can solve directly for $\{\alpha^{(i)}\}$. By construction $A^{(i)}$ is rank 1 for all i , so we can remove the trace minimization objective and get the following objective:

$$\begin{aligned} \{\hat{\alpha}^{(i)}\}_{i=1}^{n_r} = \arg \min_{\{\alpha^{(i)}\}_{i=1}^{n_r}} & \lambda_{data} \underbrace{\sum_{i \in [n_r], j \in [n_s]} (T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) - \phi^T(\mathbf{x}_s^{(j)}) M^{(i)} \phi(\mathbf{x}_s^{(j)}))^2}_{\text{data constraint}} \\ & + \lambda_{PDE} \underbrace{\langle \nabla \phi(\mathbf{x}) \nabla \phi(\mathbf{x})^T, M^{(i)} - M^{(k)} \rangle^2}_{\text{PDE constraint}} \quad (\text{C.10}) \\ \text{s.t.} & \quad M^{(i)} = I^{(i)} \alpha^{(i)} \alpha^{(i)T} I^{(i)T} \end{aligned}$$

Note that this is very similar to the regularizers for the ToF-PINN from [1]. However, a key distinction is that [1] leverages the implicit regularization of the neural field whereas EikoLift uses explicit regularization through the basis functions ϕ , which can yield substantially different solutions.

C.4 Additional experiments

Toy example

We consider a simple eikonal tomography problem where the ground truth medium is homogeneous with a velocity of 6 km/s. We simulate measurements of $n_r = 2$ subsurface sources at $n_s = 20$ surface receivers and use reciprocity. We consider $n_r K = 2$ basis functions that consist of distance functions $\phi^{(i)} = T_{0, \mathbf{x}_r^{(i)}}(\mathbf{x})$ where $T_{0, \mathbf{x}_r^{(i)}}(\mathbf{x}) := \|\mathbf{x}_r^{(i)} - \mathbf{x}\|_2$. As shown in Fig. C.2, we are able to demonstrate near perfect recovery of the true underlying velocity model in the case where our basis perfectly describes the true underlying model. Even with additional basis functions where $n_r K = 12$, we are able to obtain a near perfect recovery of the true velocity model. This is also true with added measurement noise with an SNR of 25, which corresponds to i.i.d Gaussian noise $\eta_{ij} \sim \mathcal{N}(0, 0.1 \text{ sec})$ for all $i \in [n_s]$ and $j \in [n_r]$ such that $T_{obs}(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) = T_v(\mathbf{x}_s^{(j)}, \mathbf{x}_r^{(i)}) + \eta_{ij}$.

Ablations

In Fig. C.3 and Fig. C.4, we show ablations for the various hyperparameter choices for Eq. 8 and Eq. 9 of the main text respectively using noisy measurements with an SNR of 25. In Fig. C.3, we show reconstructions using various values of ε_{PDE} and ε_{data} used in Eq. 8. We use $\varepsilon_{data} = 1$ and $\varepsilon_{PDE} = 0.1$ in the main thesis figures. In Fig. C.4, we show reconstructions using various values of ε_{PDE} and λ used in Eq. 9. We use $\lambda = 100$ and $\varepsilon_{PDE} = 0.1$ in the main thesis figures.

Varying measurement noise

In Fig. C.5, we show how the result change with various SNRs. For the Adjoint w/ TV, the TV weight, λ , is fixed for each setup. A SNR of 10 corresponds to $\sigma = 0.5$, SNR of 25 to a $\sigma = 0.1$, and SNR of 45 to a $\sigma = 0.01$. For the most part, there is an improvement as the SNR increases.

Varying the size of the basis

The choice of the number of basis vectors is a critical one that is hard to determine apriori. In Fig. C.6, we show how the result changes with an increasing number of basis functions. When $K = 2$, there are too few basis functions and the reconstruction is biased to being homogeneous. When $K = 12$, some of the reconstructions overfit to noise. $K = 6$ appears to balance reconstruction quality for most examples in this dataset. $K = 6$ is also the inflection point at which a majority of the variance is captured when performing PCA.

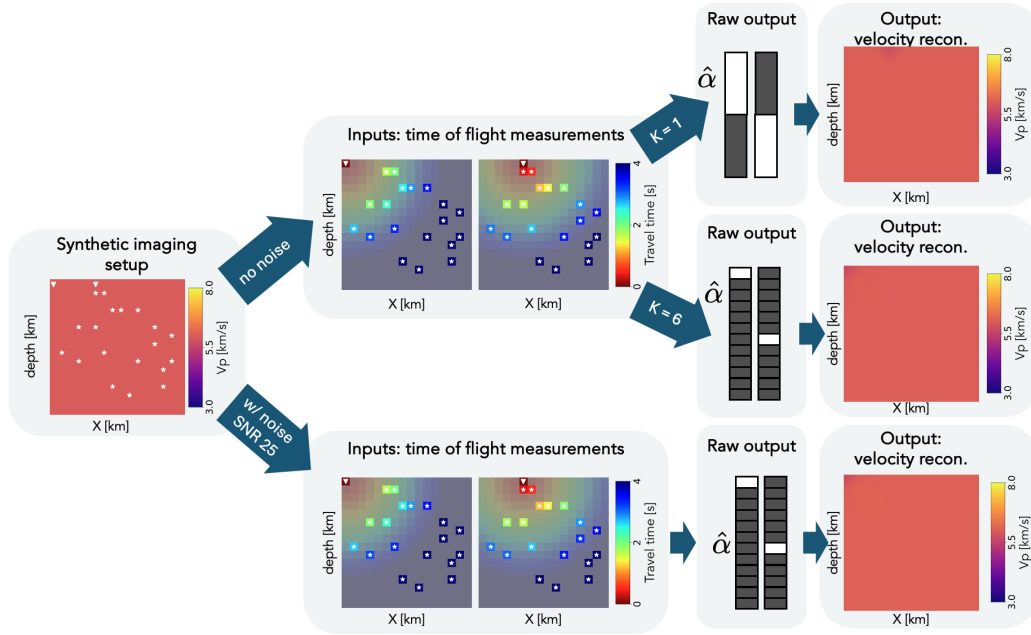


Figure C.2: Toy example with homogeneous ground truth. We demonstrate our method on a simple velocity model with a homogeneous velocity of 6 km/s. For this toy example, we choose to use a homogeneous velocity model because there is no generalization error between $\text{span}(\phi)$ and the ToF fields $T^{(i)}$. We test our method using Eq. 8 from the main text on noise-free measurements generated by the Fast Marching Method as well as noisy measurements with i.i.d additive Gaussian noise with an SNR of 25. In the noise-free case, we find that the weight-vectors $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$, described in Sec. 3.2, fit the ideal reconstruction of a one hot vector where the only non-zero value corresponds to $T_{0,x_r^{(i)}}$. This holds when we increase the number of basis vectors for each $M^{(i)}$ to be $K = 6$. In the noisy measurement setting, we also find that the weight-vectors $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ fit the ideal reconstruction and we are able to recover the homogeneous velocity. We find that we need to increase ε_{data} when our measurements are noisy.

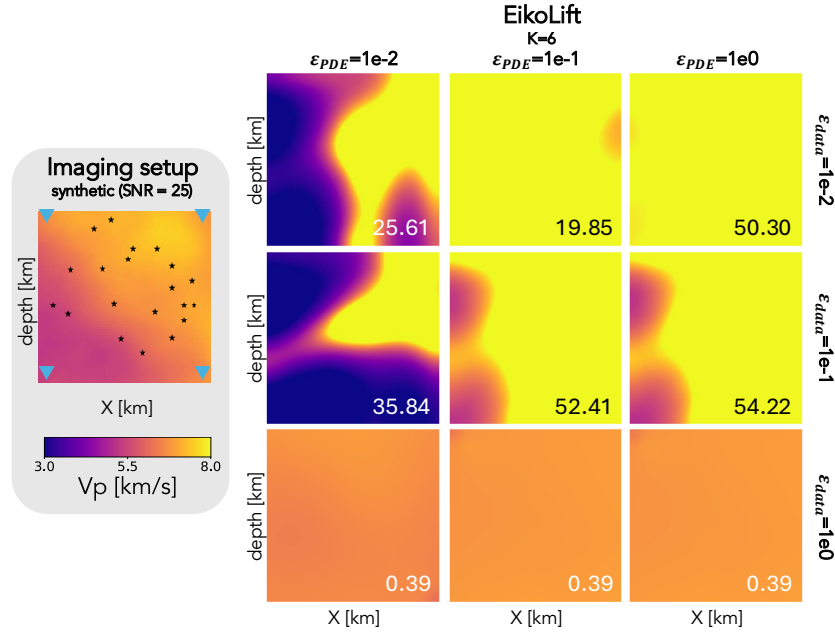


Figure C.3: **Ablations for Eq. 8 for noisy measurements.** Here we show reconstructions using various values of ϵ_{PDE} and ϵ_{data} used in Eq. 8. We use noisy measurements with a SNR of 25. When ϵ_{data} is substantially less than the expected measurement noise, the optimization yields unexpected behavior. When ϵ_{data} is too large such that it is on the order of the signal, it yields overly smooth reconstructions. $\epsilon_{PDE} = 0.1$ and $\epsilon_{data} = 1$ corresponds to the parameters used for the results in the main text.

References

- [1] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.

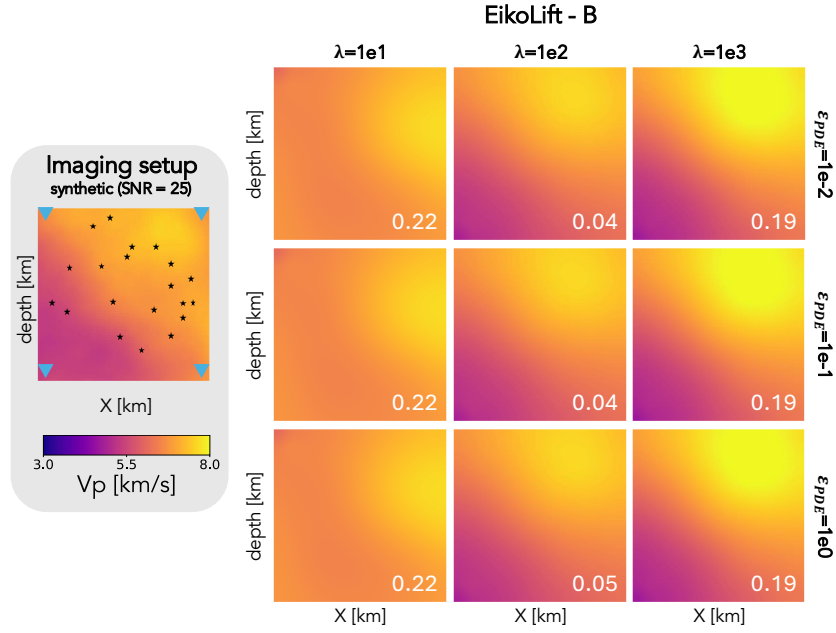


Figure C.4: **Ablations for Eq. 9 for noisy measurements.** Here we show reconstructions using various values of ϵ_{PDE} and λ used in Eq. 9. We use noisy measurements with a SNR of 25. The reconstruction under-fits the measurements when λ is too small and overfits when λ is too large. $\epsilon_{PDE} = 0.1$ and $\lambda = 100$ corresponds to the parameters used for the results in the main text.

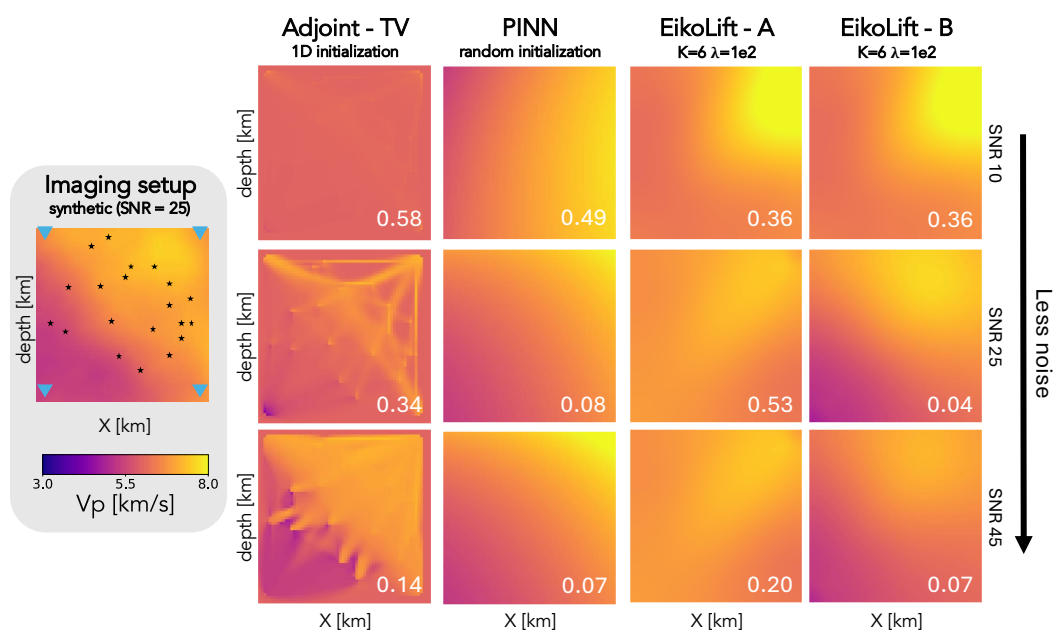


Figure C.5: **With increasing noise.** We show results across various SNR values. A SNR of 25 corresponds to $\sigma = 0.1$. In all cases, increasing the SNR improves reconstruction quality.

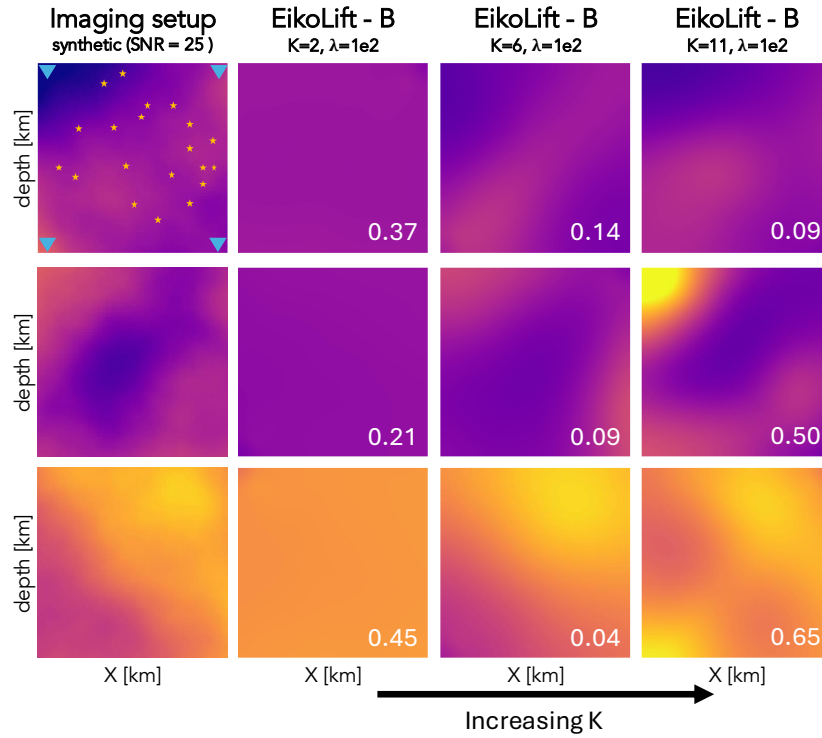


Figure C.6: **Change in reconstruction with increasing K .** We show results from optimizing Eq. 9 from the main text using various values of K . $K = 6$ corresponds to the same parameter setup as in the figures from the main text. When K is too small, the reconstructions are overly smooth, and when K is too large, the reconstruction overfits to noise.

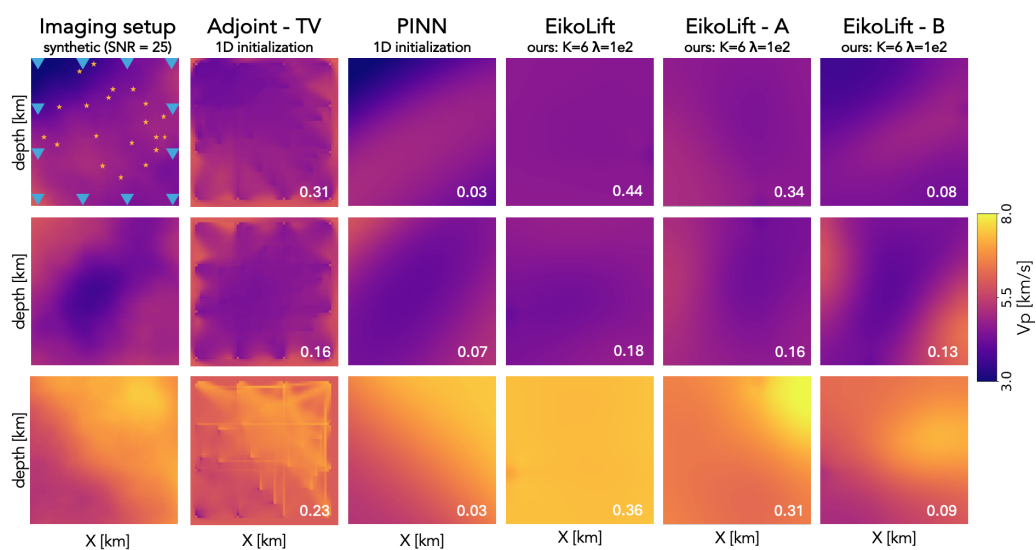


Figure C.7: **Full coverage** We show full sensor coverage reconstructions for the adjoint method with TV prior, PINN, and our method using Eq. 9 from the main text with different selection criteria. Mean squared error (MSE) is included for each reconstruction.