Training Distribution Optimization in the Space of Probability Measures

Nicholas H. Nelsen

MIT / Cornell / UT Austin

Email: nnelsen@cornell.edu
Web: nicholashnelsen.com

Work supported by:
NSF DMS-2402036 (MSPRF)
Cornell A&S Klarman Fellowship
Simons Foundation

⋆ ⋆ ⋆

Data Assimilation and Inverse Problems for Digital Twins
IMSI, Chicago

October 2025

Outline


Nicolas Guerra, Yunan Yang

Nicolas Guerra, Nicholas H. Nelsen, and Yunan Yang

Learning where to learn: Training data distribution optimization for scientific machine learning

Submitted May 2025, revised Oct 2025, arXiv:2505.21626 cs.LG

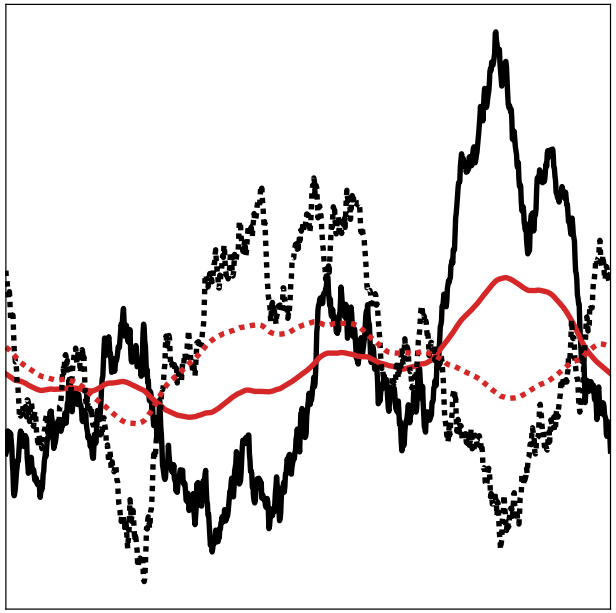
Nicholas H. Nelsen and Yunan Yang

Operator learning meets inverse problems: A probabilistic perspective

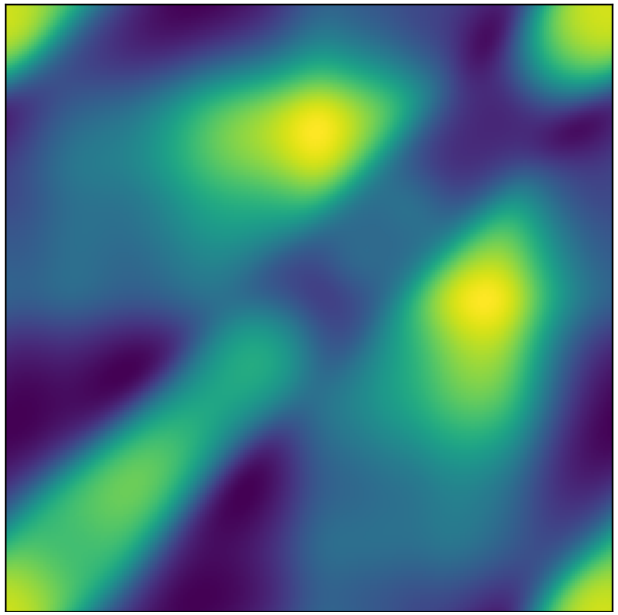
Submitted Aug 2025, arXiv:2508.20207 math.NA

Outline

EIT Inverse Problem: Likelihood Shift in Forward Data



(a) $\{(g_m, \Lambda_\gamma g_m)\}_{m=1}^2$



(b) Λ_γ : Neumann $\mapsto$ Dirichlet

(c) γ

$$\begin{cases} \nabla \cdot (\gamma \nabla \phi) = 0 & \text{in} \quad \Omega \\ \gamma \dfrac{\partial \phi}{\partial \mathsf{n}} = g & \text{on} \quad \partial\Omega \end{cases}$$

$$y_m = \Lambda_\gamma g_m + \eta_m \,, \quad \text{where} \quad g_m \sim \rho$$

Neural Inverse Operators[1] and Covariate Shift

Ideal continuum measurements: $Y = \Lambda_\gamma$ (practically inaccessible)

Finite measurements: $Y^M := \{(g_m, y_m)\}_{m=1}^{M(\gamma)}$

$$y_m = \Lambda_\gamma g_m \,, \quad \text{where} \quad g_m \sim \rho$$

[1] R. Molinaro, Y. Yang, B. Engquist, S. Mishra '23; N. Guerra, N., Y. Yang '25; N., Y. Yang '25

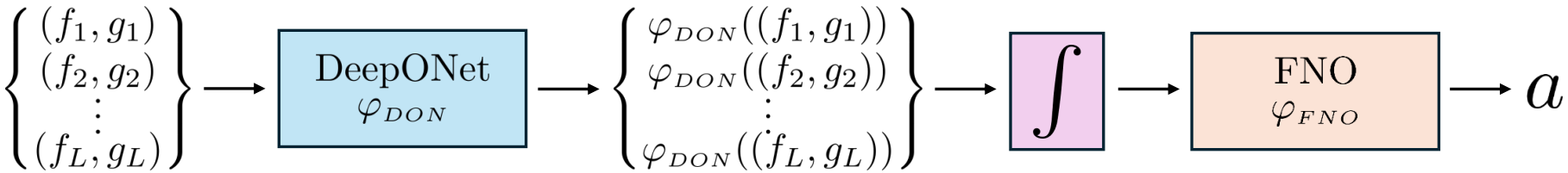
Neural Inverse Operators[1] and Covariate Shift

Ideal continuum measurements: $Y = \Lambda_\gamma$ (practically inaccessible)

Finite measurements: $Y^M := \{(g_m, y_m)\}_{m=1}^{M(\gamma)}$

$$y_m = \Lambda_\gamma g_m \,, \quad \text{where} \quad g_m \sim \rho$$

Neural operators on the space of probability measures



[1] R. Molinaro, Y. Yang, B. Engquist, S. Mishra '23; N. Guerra, N., Y. Yang '25; N., Y. Yang '25

Neural Inverse Operators[1] and Covariate Shift

Ideal continuum measurements: $Y = \Lambda_\gamma$ (practically inaccessible)

Finite measurements: $Y^M := \{(g_m, y_m)\}_{m=1}^{M(\gamma)}$

$$y_m = \Lambda_\gamma g_m \,, \quad \text{where} \quad g_m \sim \rho$$

Neural operators on the space of probability measures

$$\Psi_{\mathrm{NIO}} \colon \mathscr{P}(L^2(\partial\Omega) \times L^2(\partial\Omega)) \to L^2(\Omega)$$

$$\frac{1}{M}\sum_{m=1}^{M} \delta_{(g_m, y_m)} \mapsto \varphi_{\mathrm{FNO}}\left(\frac{1}{M}\sum_{m=1}^{M} \varphi_{\mathrm{DON}}\big((g_m, y_m)\big)\right).$$

[1] R. Molinaro, Y. Yang, B. Engquist, S. Mishra '23; N. Guerra, N., Y. Yang '25; N., Y. Yang '25

Neural Inverse Operators[1] and Covariate Shift

Ideal continuum measurements: $Y = \Lambda_\gamma$ (practically inaccessible)

Finite measurements: $Y^M := \{(g_m, y_m)\}_{m=1}^{M(\gamma)}$

$$y_m = \Lambda_\gamma g_m \,, \quad \text{where} \quad g_m \sim \rho$$

Neural operators on the space of probability measures

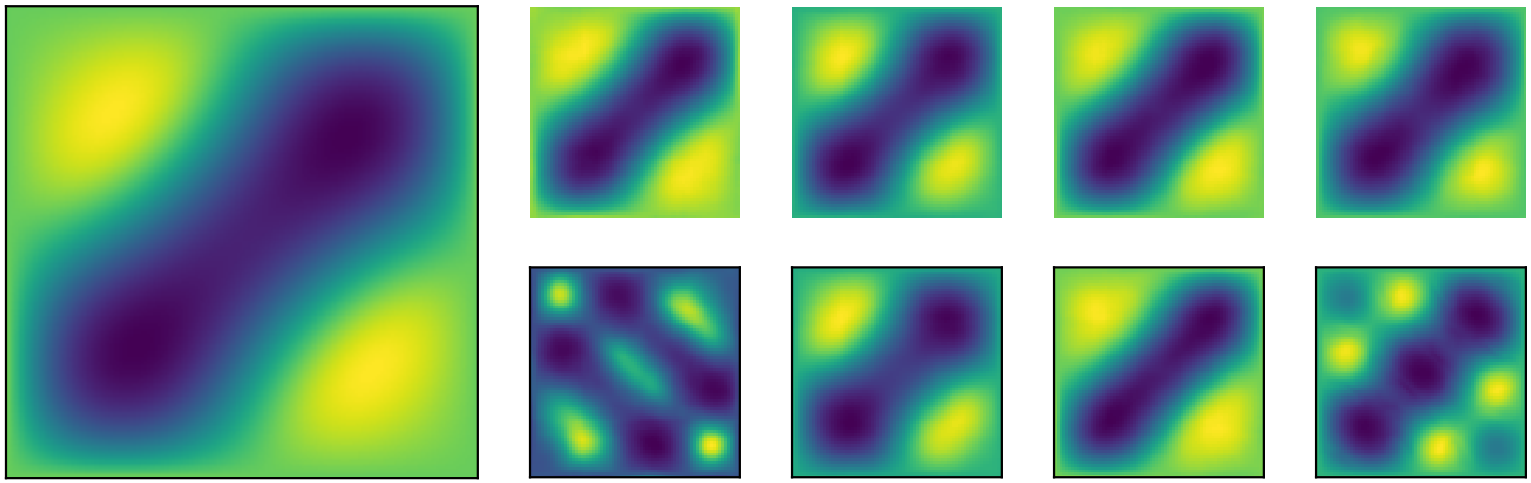
$$\Psi_{\mathrm{NIO}} \colon \mathscr{P}(L^2(\partial\Omega) \times L^2(\partial\Omega)) \to L^2(\Omega)$$

$$(\mathrm{Id}, \Lambda_\gamma)_\# \rho \mapsto \varphi_{\mathrm{FNO}} \left(\mathbb{E}_{g\sim\rho} \, \varphi_{\mathrm{DON}}\big((g, \Lambda_\gamma g)\big) \right)$$

Covariate Shift Problem: *What if new test input $Y^{M_{\mathrm{test}}}$ comes from $\rho_{\mathrm{test}} \neq \rho$?*

[1] R. Molinaro, Y. Yang, B. Engquist, S. Mishra '23; N. Guerra, N., Y. Yang '25; N., Y. Yang '25

Choice of Boundary Data Distribution Matters



(top row) in-distribution $\rho_{\text{train}} = \nu_i$; (bottom row) out-of-distribution ρ_{test}

Outline

A General SciML Setting

Regression: Find $\mathcal{G} \approx \mathcal{G}^\star : \mathcal{U} \to \mathcal{Y}$ from data

$$u_n \sim \nu \quad \text{and} \quad y_n = \mathcal{G}^\star(u_n) \qquad \text{(noisy } y_n \text{ also possible)}$$

Input primitive: For any ν, generate samples $u_n \sim \nu$

Output primitive: Query the black-box oracle $u \mapsto \mathcal{G}^\star(u)$

Goal

What distribution ν should training data be sampled from to best approximate $\mathcal{G}^\star$?

Warm Up: Test Errors and Covariate Shift

Test accuracy: For *fixed* $\nu' \in \mathscr{P}_2(\mathcal{U})$, define

$$\mathrm{Acc}(\mathcal{G}; \nu') := \mathbb{E}_{u' \sim \nu'} \|\mathcal{G}^\star(u') - \mathcal{G}(u')\|_{\mathcal{Y}}^2$$

Training data design: Free to choose training distribution ν for ERM on $u_n \sim \nu$:

$$\widehat{\mathcal{G}}^{(\nu)} := \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{G}^\star(u_n) - \mathcal{G}(u_n)\|_{\mathcal{Y}}^2$$

What ν minimizes $\mathrm{Acc}(\widehat{\mathcal{G}}^{(\nu)}; \nu')$?

Warm Up: Test Errors and Covariate Shift

Test accuracy: For *fixed* $\nu' \in \mathscr{P}_2(\mathcal{U})$, define

$$\mathrm{Acc}(\mathcal{G}; \nu') := \mathbb{E}_{u' \sim \nu'} \|\mathcal{G}^\star(u') - \mathcal{G}(u')\|_{\mathcal{Y}}^2$$

Training data design: Free to choose training distribution ν for ERM on $u_n \sim \nu$:

$$\widehat{\mathcal{G}}^{(\nu)} := \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{G}^\star(u_n) - \mathcal{G}(u_n)\|_{\mathcal{Y}}^2$$

What ν minimizes $\mathrm{Acc}(\widehat{\mathcal{G}}^{(\nu)}; \nu')$? Fact: $\nu_{\mathrm{opt}} \neq \nu'$

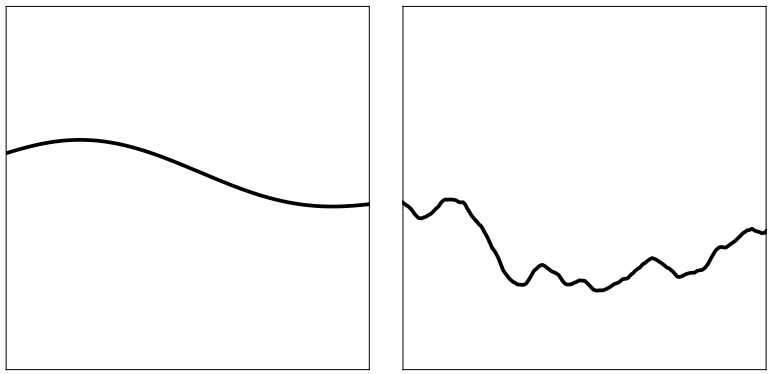
Training on the Test Distribution is *Suboptimal* with Finite Data

Christoffel sampling:[2] With finite data, minimizing an Acc upper bound gives

$$\nu_{\mathrm{opt}}(du) = \frac{1}{w_{\mathrm{opt}}(u)}\nu'(du)$$

(w_{opt} depends on both $\mathcal{H}$ and ν', and estimator is from weighted least squares ERM)

Linear operator learning:[3] ν should be less regular than ν', so $\mathrm{supp}(\nu') \subset \mathrm{supp}(\nu)$



(a) Smooth (b) Rough

[2] B. Adock '25 (review paper); A. Cohen and G. Migliorati '17
[3] M. de Hoop, N. Kovachki, N., A. Stuart '23

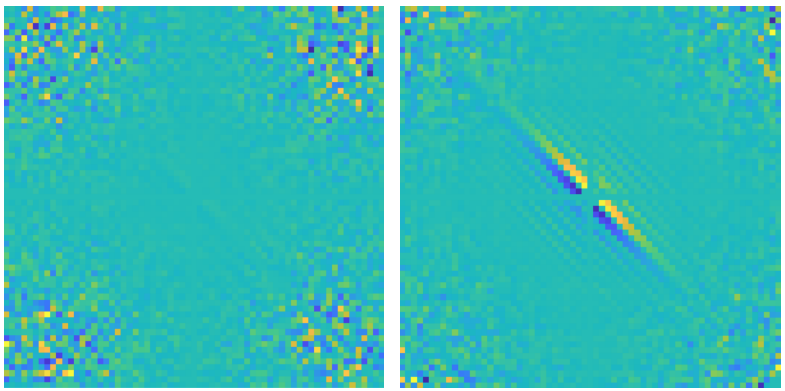
Training on the Test Distribution is *Suboptimal* with Finite Data

Christoffel sampling:[2] With finite data, minimizing an Acc upper bound gives

$$\nu_{\mathrm{opt}}(du) = \frac{1}{w_{\mathrm{opt}}(u)}\nu'(du)$$

(w_{opt} depends on both $\mathcal{H}$ and ν', and estimator is from weighted least squares ERM)

Linear operator learning:[3] ν should be less regular than ν', so $\mathrm{supp}(\nu') \subset \mathrm{supp}(\nu)$



(a) Learned (smooth) (b) Learned (rough)

(uncertainty sampling, ridge leverage scores, RPCholesky, etc)

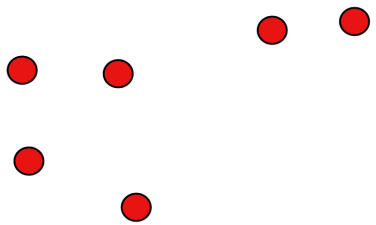
[2] B. Adock '25 (review paper); A. Cohen and G. Migliorati '17
[3] M. de Hoop, N. Kovachki, N., A. Stuart '23

Today: Beyond One Task Via Meta Test Distributions

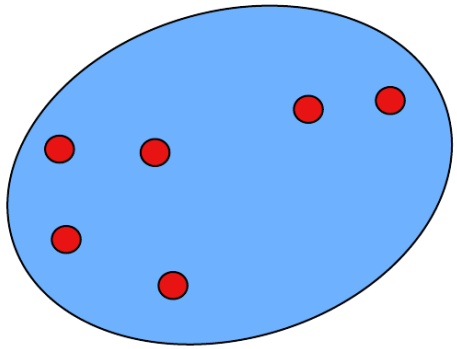
A probability measure over probability measures: $\mathbb{Q} \in \mathscr{P}(\mathscr{P}(\mathcal{U}))$

Example: A typical single task setup involves just one test distribution ν', so $\mathbb{Q} = \delta_{\nu'}$

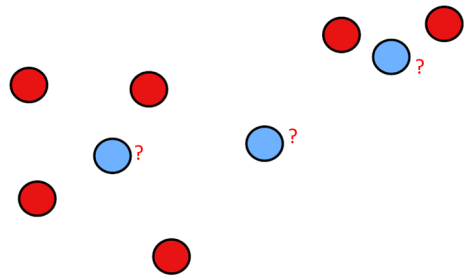
Example: 6 test distributions (e.g., different classes [airplane, chair, lamp, etc], experiments)



Family of Test Distributions $\mathbb{Q}$



Ideal: A broad training distribution (blue) covering the support of all test scenarios



In practice: constrained

$$\mathbb{Q} = \frac{1}{6} \sum_{k=1}^{6} \delta_{\nu'_k}$$

Average Out-of-Distribution Prediction Accuracy

$$\mathcal{E}_{\mathbb{Q}}(\mathcal{G}) := \mathbb{E}_{\nu' \sim \mathbb{Q}} \, \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^{\star}(u') - \mathcal{G}(u') \right\|_{\mathcal{Y}}^{2}$$

Average Out-of-Distribution Prediction Accuracy

$$\mathcal{E}_{\mathbb{Q}}(\mathcal{G}) := \mathbb{E}_{u' \sim \nu_{\mathbb{Q}}} \left\| \mathcal{G}^\star(u') - \mathcal{G}(u') \right\|_{\mathcal{Y}}^2, \quad \text{where} \quad \nu_{\mathbb{Q}} := \mathbb{E}_{\nu' \sim \mathbb{Q}} \, \nu'$$

Access model: Not $\mathbb{Q}$ itself, but K empirical measures $\{(\nu_k')^{(M_k)}\}_{k=1}^K$, where

$$(\nu_k')^{(M_k)} := \frac{1}{M_k} \sum_{m=1}^{M_k} \delta_{u_m^k}, \quad \text{where} \quad u_m^k \sim \nu_k' \sim \mathbb{Q}$$

(*fixed* data sizes K and $\{M_k\}_{k=1}^K$)

Now: *What ν minimizes $\mathcal{E}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)})$?*

Outline

Deriving Algorithms: An Optimize-Then-Empiricalize Approach

$(u_n \overset{\text{i.i.d.}}{\sim} \nu)$

Practice (finite validation, finite train)

$$\inf_{\nu} \left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{m=1}^{M_k} \left\| \mathcal{G}^\star(u_m^k) - \widehat{\mathcal{G}}_N^{(\nu)}(u_m^k) \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}_N^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{G}^\star(u_n) - \mathcal{G}(u_n) \right\|_{\mathcal{Y}}^2 \right\}$$

Deriving Algorithms: An Optimize-Then-Empiricalize Approach

$(u_n \overset{\text{i.i.d.}}{\sim} \nu)$

Practice (finite validation, finite train)

$$\inf_{\nu} \left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{m=1}^{M_k} \left\| \mathcal{G}^\star(u_m^k) - \widehat{\mathcal{G}}_N^{(\nu)}(u_m^k) \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}_N^{(\nu)} \in \arg\min_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{G}^\star(u_n) - \mathcal{G}(u_n) \right\|_{\mathcal{Y}}^2 \right\}$$

Ideal (infinite validation, finite train)

$$\inf_{\nu \in \mathscr{P}_2(\mathcal{U})} \left\{ \mathbb{E}_{\nu' \sim \mathbb{Q}} \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^\star(u') - \widehat{\mathcal{G}}_N^{(\nu)}(u') \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}_N^{(\nu)} \in \arg\min_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{G}^\star(u_n) - \mathcal{G}(u_n) \right\|_{\mathcal{Y}}^2 \right\}$$

Deriving Algorithms: An Optimize-Then-Empiricalize Approach

$(u_n \overset{\text{i.i.d.}}{\sim} \nu)$

Practice (finite validation, finite train)

$$\inf_{\nu} \left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{m=1}^{M_k} \left\| \mathcal{G}^\star(u_m^k) - \widehat{\mathcal{G}}_N^{(\nu)}(u_m^k) \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}_N^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{G}^\star(u_n) - \mathcal{G}(u_n) \right\|_{\mathcal{Y}}^2 \right\}$$

Ideal (infinite validation, finite train)

$$\inf_{\nu \in \mathscr{P}_2(\mathcal{U})} \left\{ \mathbb{E}_{\nu' \sim \mathbb{Q}} \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^\star(u') - \widehat{\mathcal{G}}_N^{(\nu)}(u') \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}_N^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{G}^\star(u_n) - \mathcal{G}(u_n) \right\|_{\mathcal{Y}}^2 \right\}$$

Continuum (infinite validation, infinite train)

$$\inf_{\nu \in \mathscr{P}_2(\mathcal{U})} \left\{ \mathbb{E}_{\nu' \sim \mathbb{Q}} \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^\star(u') - \widehat{\mathcal{G}}^{(\nu)}(u') \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \mathbb{E}_{u \sim \nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 \right\}$$

Adaptive Algorithm 1: Bilevel Minimization

Continuum: When $\mathcal{H}$ is Hilbertian, can directly minimize as a bilevel problem

$$\inf_{\nu \in \mathscr{P}_2(\mathcal{U})} \left\{ \mathbb{E}_{\nu' \sim \mathbb{Q}} \, \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^\star(u') - \widehat{\mathcal{G}}^{(\nu)}(u') \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \mathbb{E}_{u \sim \nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 \right\}$$

Bilevel minimization: Gradient descent in the space of probability measures

- Adjoint state methods in $\mathcal{H}$ (e.g., RKHS for kernel methods)
- Gradient ∇_ν, or ∇_θ if $\nu = \nu_\theta$ (parametric families)

Adaptive Algorithm 1: Bilevel Minimization

Continuum: When $\mathcal{H}$ is Hilbertian, can directly minimize as a bilevel problem

$$\inf_{\nu \in \mathscr{P}_2(\mathcal{U})} \left\{ \mathbb{E}_{\nu' \sim \mathbb{Q}} \mathbb{E}_{u' \sim \nu'} \left\| \mathcal{G}^\star(u') - \widehat{\mathcal{G}}^{(\nu)}(u') \right\|_{\mathcal{Y}}^2 \,\middle|\, \widehat{\mathcal{G}}^{(\nu)} \in \operatorname*{arg\,min}_{\mathcal{G} \in \mathcal{H}} \mathbb{E}_{u \sim \nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 \right\}$$

Bilevel minimization: Gradient descent in the space of probability measures

$$\left(\nabla_\nu \mathcal{E}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)})\right)(u) = -\underbrace{\left(\widehat{\mathcal{G}}^{(\nu)}(u) - \mathcal{G}^\star(u)\right)}_{\text{residual}} \underbrace{\left(\mathcal{K}_\nu^{-1} \mathcal{K}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)} - \mathcal{G}^\star)\right)}_{\text{reweighted residual}}(u) \qquad \text{(RKHS)}$$

Adaptive Algorithm 2: Alternating Minimization

Upper bound: Alternatively, under Lipschitz assumptions,

$$\forall \mathcal{G}\colon \mathscr{E}_{\mathbb{Q}}(\mathcal{G}) \leq \inf_{\mu} \left[\mathbb{E}_{u\sim\mu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 + \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, c^2(\mathcal{G}^\star, \mathcal{G}, \mu, \nu')} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\mu, \nu')} \right]$$

$$\Longrightarrow \inf_{\nu} \mathscr{E}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)}) \leq \inf_{(\nu,\mathcal{G})} \left[\underbrace{\mathbb{E}_{u\sim\nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2}_{\text{training error}} + \underbrace{\overbrace{\delta(\nu; \mathcal{G}^\star, \mathbb{Q})}^{\text{(strength)}} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\nu, \nu')}}_{\text{penalization on distribution misfit}} \right]$$

Adaptive Algorithm 2: Alternating Minimization

Upper bound: Alternatively, under Lipschitz assumptions,

$$\forall \mathcal{G}\colon \mathscr{E}_{\mathbb{Q}}(\mathcal{G}) \leq \inf_{\mu} \left[\mathbb{E}_{u\sim\mu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 + \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, c^2(\mathcal{G}^\star, \mathcal{G}, \mu, \nu')} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\mu, \nu')} \right]$$

$$\implies \inf_{\nu} \mathscr{E}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)}) \leq \inf_{(\nu,\mathcal{G})} \left[\underbrace{\mathbb{E}_{u\sim\nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2}_{\text{training error}} + \underbrace{\overbrace{\delta(\nu; \mathcal{G}^\star, \mathbb{Q})}^{\text{(strength)}} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\nu, \nu')}}_{\text{penalization on distribution misfit}} \right]$$

Alternating minimization: Alternating descent on upper bound:

$$\nu^{(0)} \xrightarrow{\text{train}} \widehat{\mathcal{G}}^{(0)} \xrightarrow{\text{optimize}} \nu^{(1)} \xrightarrow{\text{train}} \widehat{\mathcal{G}}^{(1)} \xrightarrow{\text{optimize}} \nu^{(2)} \cdots$$

- Objective $\mathcal{J}$ always nonincreasing as iterations progress
- Scalable: applicable to neural networks, neural operators, ...

Adaptive Algorithm 2: Alternating Minimization

Upper bound: Alternatively, under Lipschitz assumptions,

$$\forall \mathcal{G}\colon \mathscr{E}_{\mathbb{Q}}(\mathcal{G}) \leq \inf_{\mu} \left[\mathbb{E}_{u\sim\mu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2 + \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, c^2(\mathcal{G}^\star, \mathcal{G}, \mu, \nu')} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\mu, \nu')} \right]$$

$$\implies \inf_{\nu} \mathscr{E}_{\mathbb{Q}}(\widehat{\mathcal{G}}^{(\nu)}) \leq \inf_{(\nu,\mathcal{G})} \left[\underbrace{\mathbb{E}_{u\sim\nu} \left\| \mathcal{G}^\star(u) - \mathcal{G}(u) \right\|_{\mathcal{Y}}^2}_{\text{training error}} + \underbrace{\overbrace{\delta(\nu; \mathcal{G}^\star, \mathbb{Q})}^{\text{(strength)}} \sqrt{\mathbb{E}_{\nu'\sim\mathbb{Q}}\, \mathsf{W}_2^2(\nu, \nu')}}_{\text{penalization on distribution misfit}} \right]$$

Alternating minimization: Alternating descent on upper bound:

$$\nu^{(0)} \xrightarrow{\text{train}} \widehat{\mathcal{G}}^{(0)} \xrightarrow{\text{optimize}} \nu^{(1)} \xrightarrow{\text{train}} \widehat{\mathcal{G}}^{(1)} \xrightarrow{\text{optimize}} \nu^{(2)} \cdots$$

$$\nabla_\nu \mathcal{J}|_{(\nu,\mathcal{G})}(u) = \underbrace{\|\mathcal{G}^\star(u) - \mathcal{G}(u)\|_{\mathcal{Y}}^2}_{\text{residual}} + \underbrace{\mathcal{M}(\nu;\mathbb{Q}) \frac{\|u\|_{\mathcal{U}}^2}{2}}_{\text{moments}} + \frac{1}{\mathcal{M}(\nu;\mathbb{Q})} \underbrace{\left(\frac{\|u\|_{\mathcal{U}}^2}{2} - \mathbb{E}_{\mathbb{Q}}\, \phi_\nu(u) \right)}_{\text{OT potential}}$$

Some Takeaways

- Optimal ν depends on target $\mathcal{G}^\star$
- Information available for experimental design:
  1. Query access to $u \mapsto \mathcal{G}^\star(u)$
  2. Samples for candidate ν
  3. Fixed validation set of empirical measures $\{(\nu'_k)^{(M_k)}\}_{k=1}^K$ from $\mathbb{Q}$
- *Write down algorithms, then empiricalize*

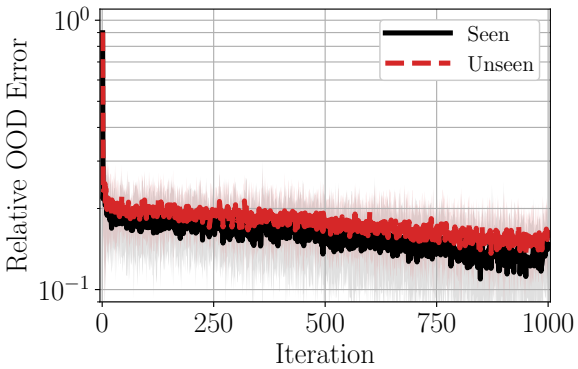
Outline

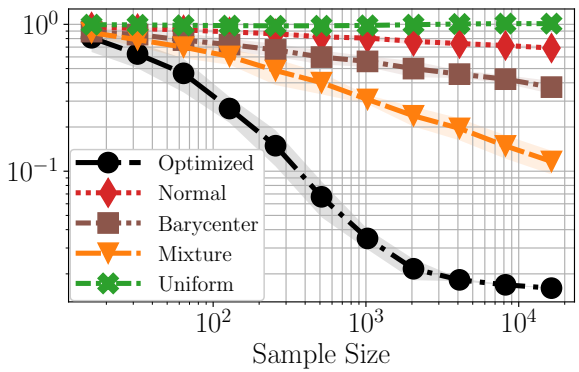
Bilevel Minimization Example: Sobol-G Function Approximation

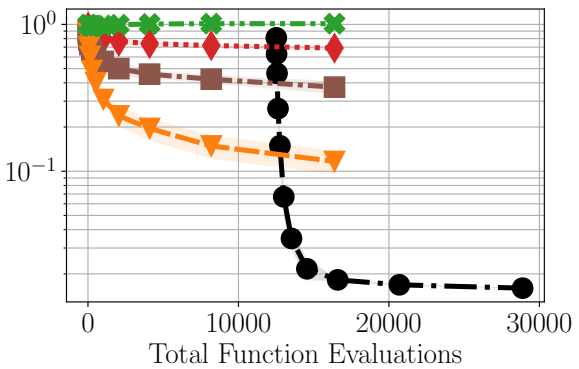
Goal: Optimize $(m, \mathcal{C})$ in $x \sim \mathcal{N}(m, \mathcal{C}) \in \mathscr{P}_2(\mathbb{R}^d)$ to learn the Sobol-G function

$$\mathcal{G}^\star(x) := \prod_{j=1}^{d} \frac{|4x_j - 2| + (j-2)/2}{1 + (j-2)/2}$$

with **kernel ridge regression** under 10 Gaussian atoms empirical measure $\mathbb{Q}$ with normal means and Wishart covariances







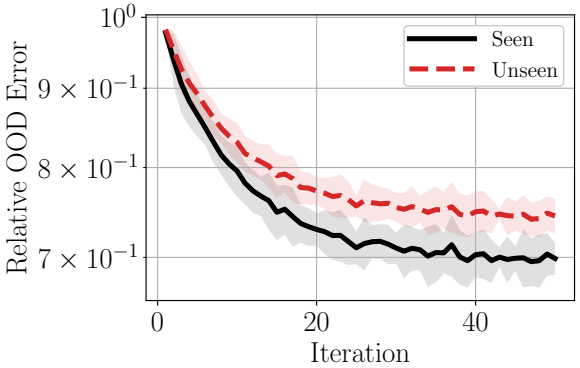
(Left) Meta validation error over 1000 iterations. (Center) After design is found, meta test error as sample size grows. (Right) Same as center, but versus number of function evaluations.

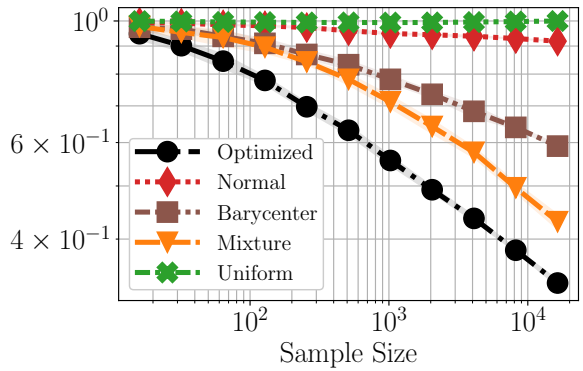
Bilevel Minimization Example: Friedmann Function Approximation

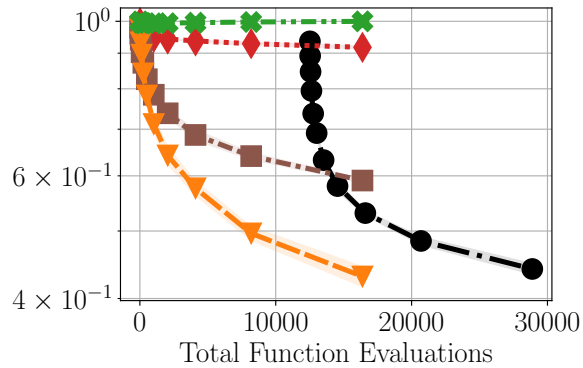
Goal: Optimize $(m, \mathcal{C})$ in $x \sim \mathcal{N}(m, \mathcal{C}) \in \mathscr{P}_2(\mathbb{R}^d)$ to learn the function

$$\mathcal{G}^\star(x) := 10 \sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5$$

with **kernel ridge regression** under 10 Gaussian atoms empirical measure $\mathbb{Q}$ with normal means and Wishart covariances



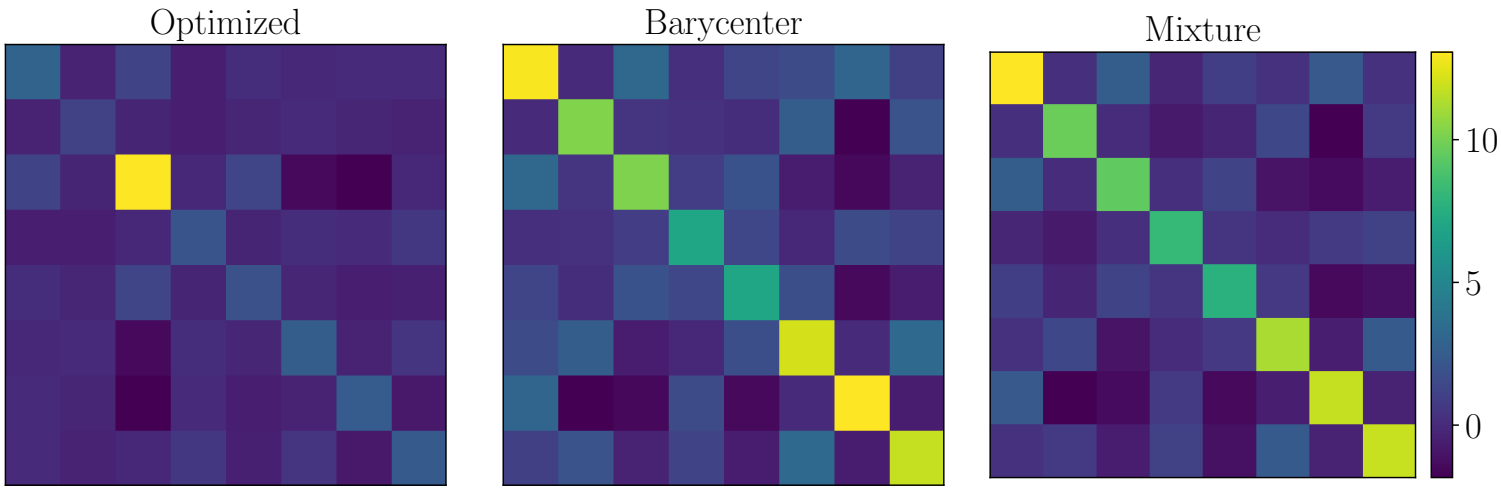




(Left) Meta validation error over 50 iterations. (Center) After design is found, meta test error as sample size grows. (Right) Same as center, but versus number of function evaluations.

Optimal Distributions are Interpretable

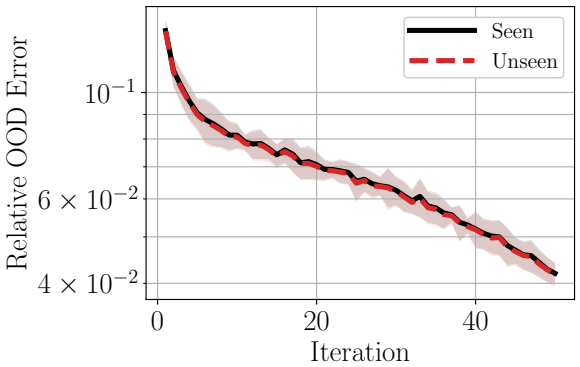
$$\mathcal{G}^\star(x) := 10\sin(\pi x_1 x_2) + 20\underbrace{(x_3 - 1/2)^2}_{\text{largest term}} + 10x_4 + 5x_5$$

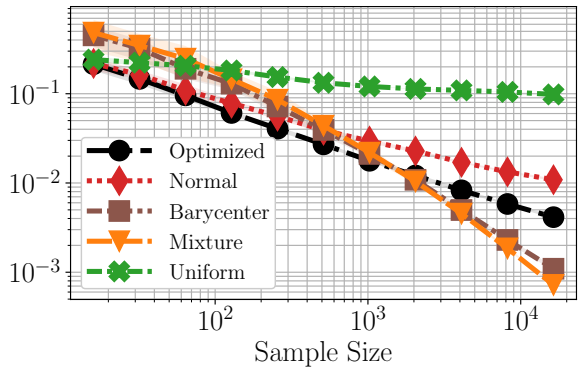


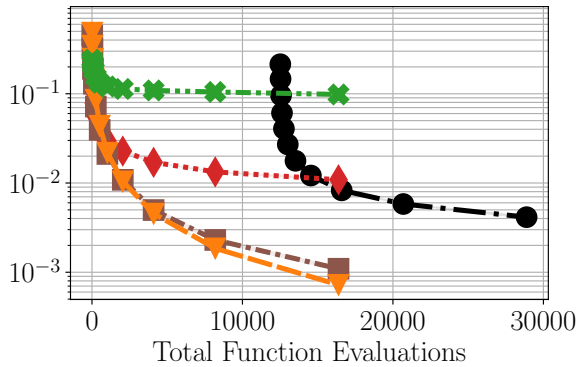
Learned (left) vs. empirical $\mathbb{Q}$-barycenter (center) vs. empirical $\mathbb{Q}$-mixture covariance (right)

Bilevel Minimization Example: RBF Function Approximation

Goal: Optimize $(m, \mathcal{C})$ in $x \sim \mathcal{N}(m, \mathcal{C}) \in \mathscr{P}_2(\mathbb{R}^d)$ to learn a sum of 1000 RBFs with random coefficients with **kernel ridge regression** under 10 Gaussian atoms empirical measure $\mathbb{Q}$ with normal means and Wishart covariances







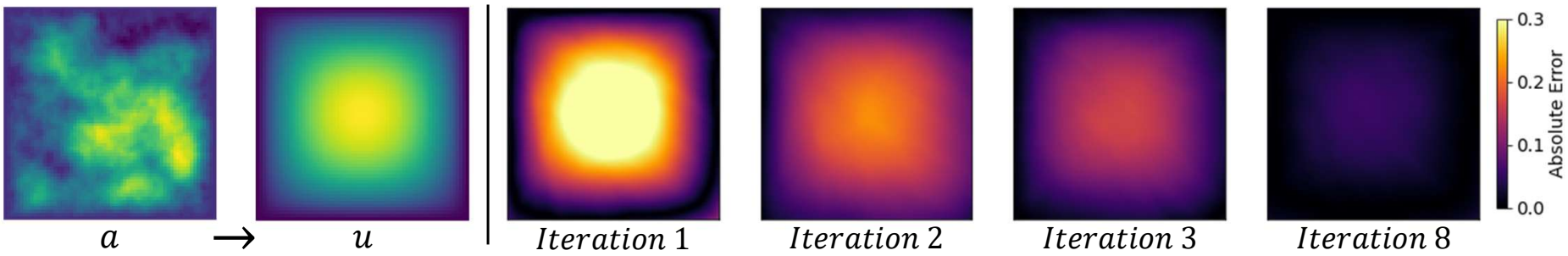
(Left) Meta validation error over 50 iterations. (Center) After design is found, meta test error as sample size grows. (Right) Same as center, but versus number of function evaluations.

Alternating Minimization Example: Darcy Flow Operator Learning

Goal: Optimize the mean function m in random field $a \sim \mathcal{N}(m, \mathcal{C}_0)$, where $\mathcal{C}_0$ is fixed, to learn the map $\mathcal{G}^\star : a \mapsto u$ implicit in

$$\begin{cases} -\nabla \cdot (a \nabla u) = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

with **DeepONet** under 20 GP atoms empirical measure $\mathbb{Q}$ with Matérn covariances



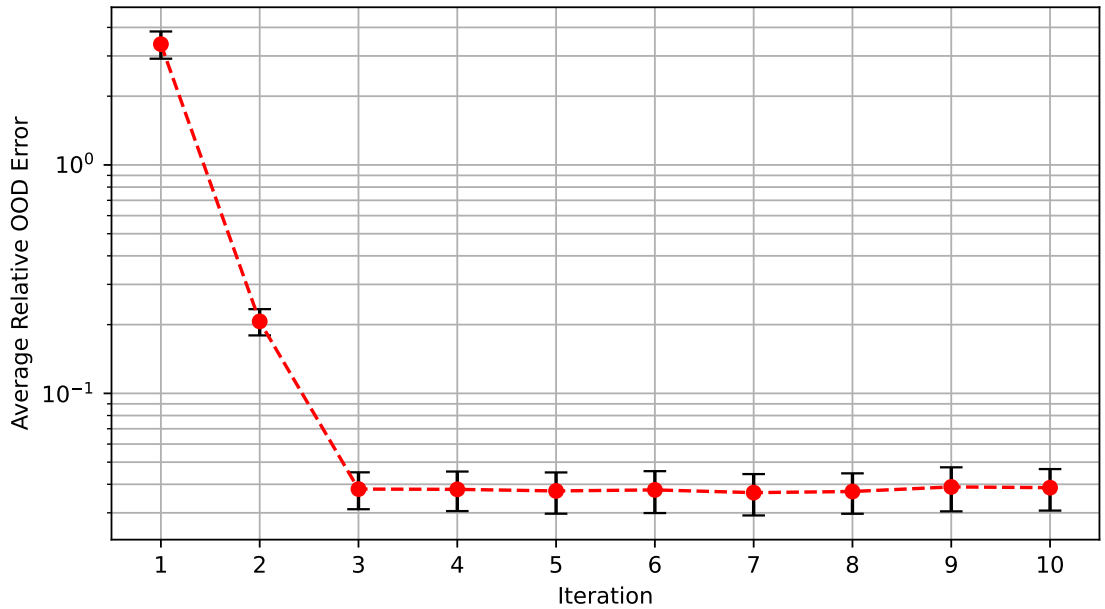
Evolution of solution errors as alternating minimization progresses

Alternating Minimization Example: Darcy Flow Operator Learning

Goal: Optimize the mean function m in random field $a \sim \mathcal{N}(m, \mathcal{C}_0)$, where $\mathcal{C}_0$ is fixed, to learn the map $\mathcal{G}^\star : a \mapsto u$ implicit in

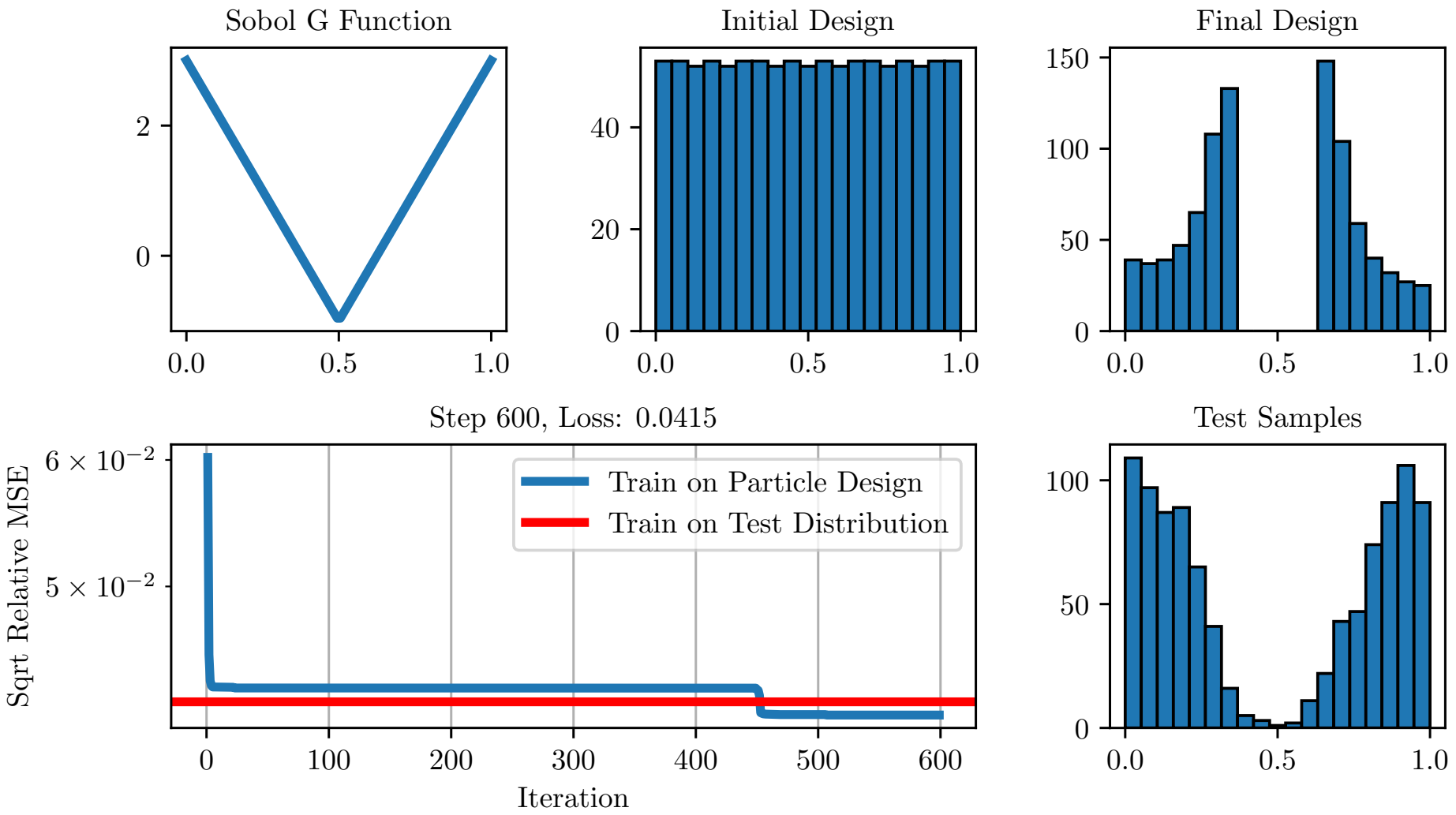
$$\begin{cases} -\nabla \cdot (a \nabla u) = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

with **DeepONet** under 20 GP atoms empirical measure $\mathbb{Q}$ with Matérn covariances



Evolution of average OOD test errors as alternating minimization progresses

Nonparametric Particle Design (1D illustration)



Gradient-based distribution design with Wasserstein gradient flows

Outline

Conclusion

Summary

- Bilevel and alternating optimization over training distributions
- Optimal ν depends on target map $\mathcal{G}^\star$

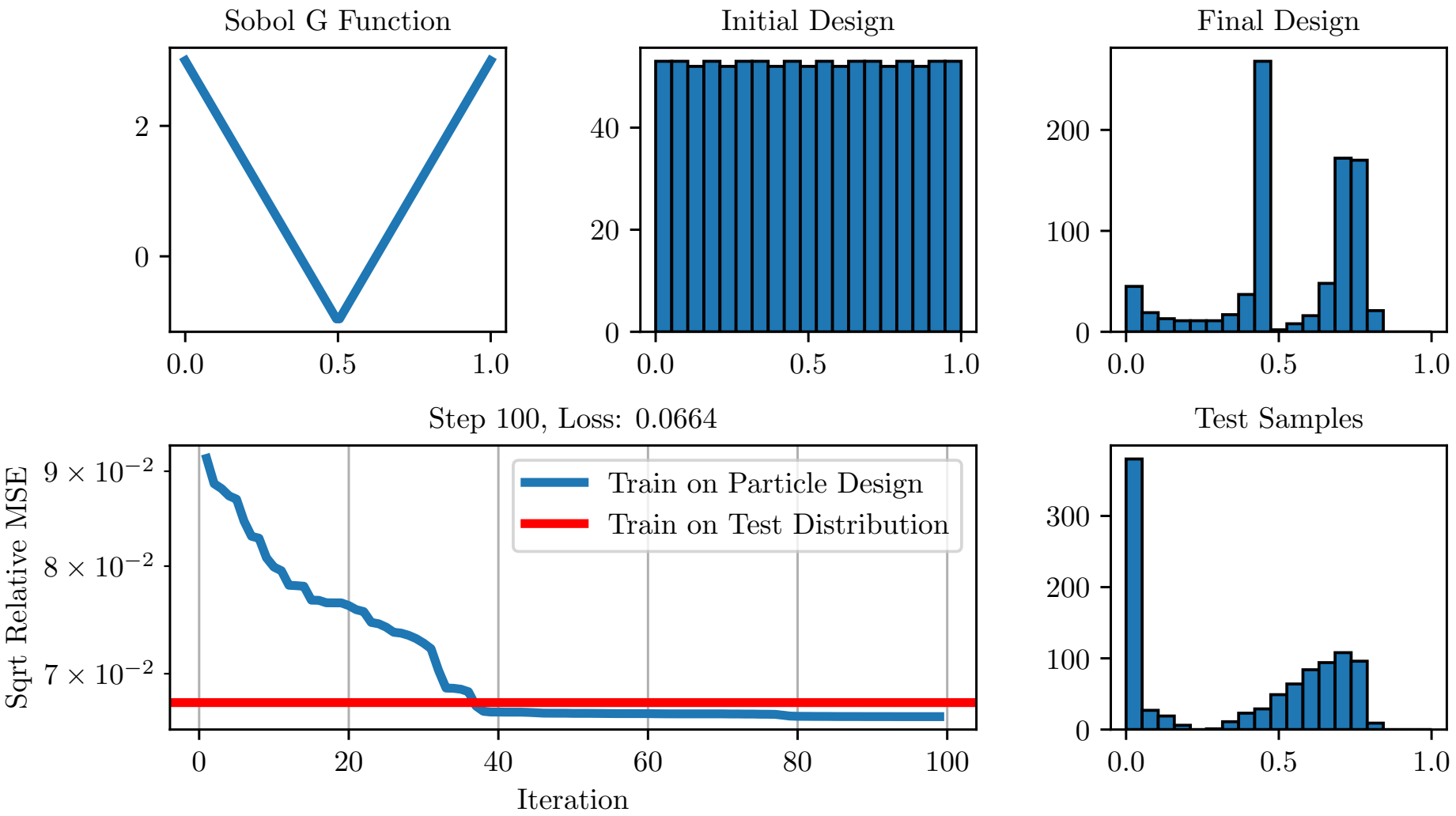
Future work

- Scalable nonparametric distribution families
- Cost vs. accuracy tradeoffs

Paper: arXiv:2505.21626 cs.LG
Code: github.com/nicolas-guerra/learning-where-to-learn

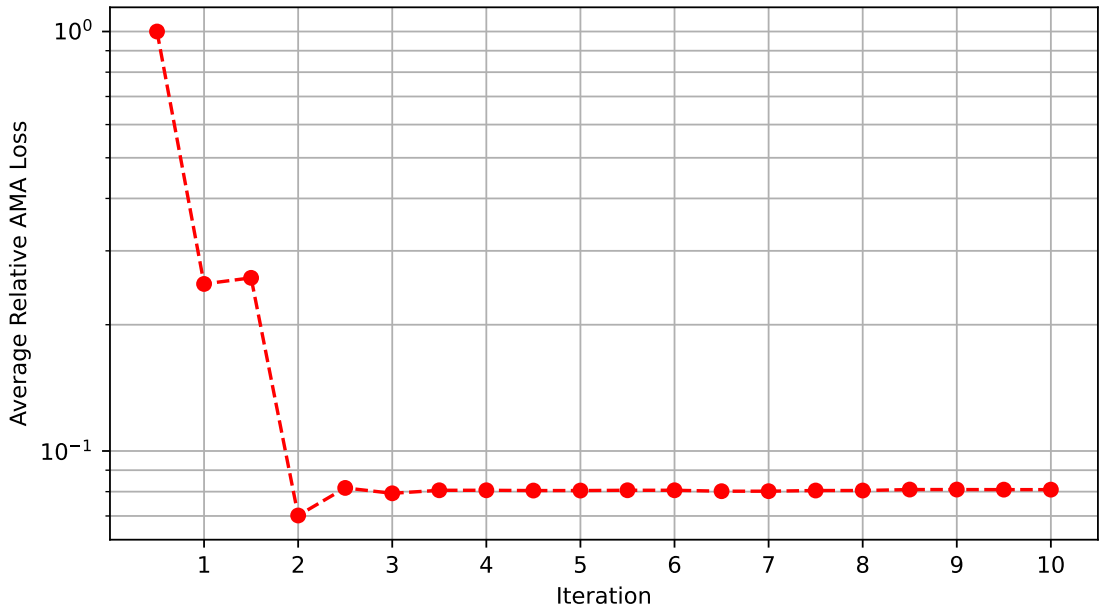
BACKUP

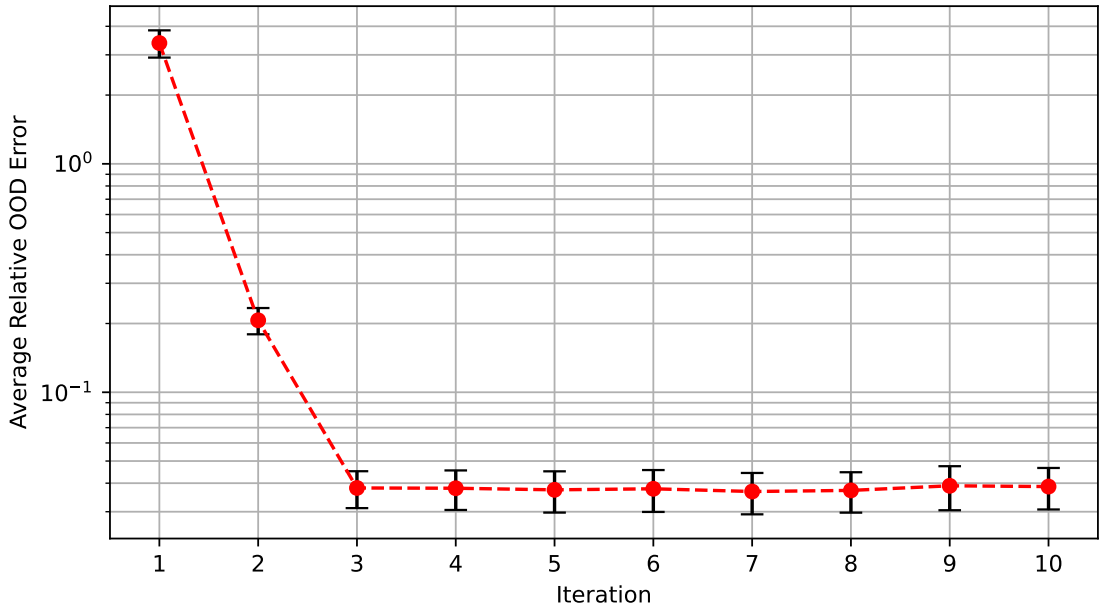
Nonparametric Particle Design (1D illustration)



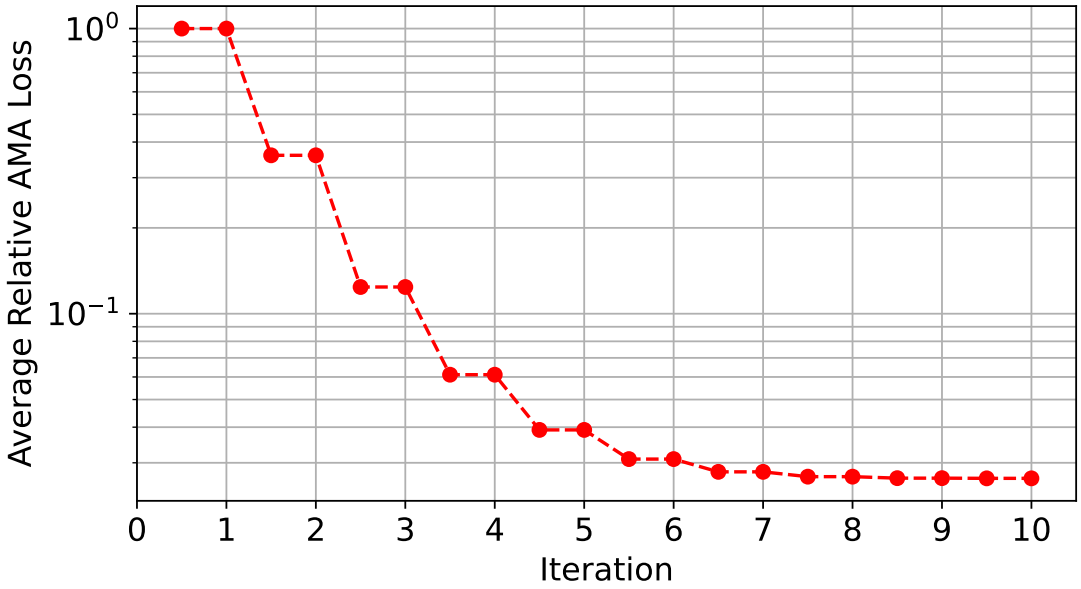
Gradient-based distribution design with Wasserstein gradient flows

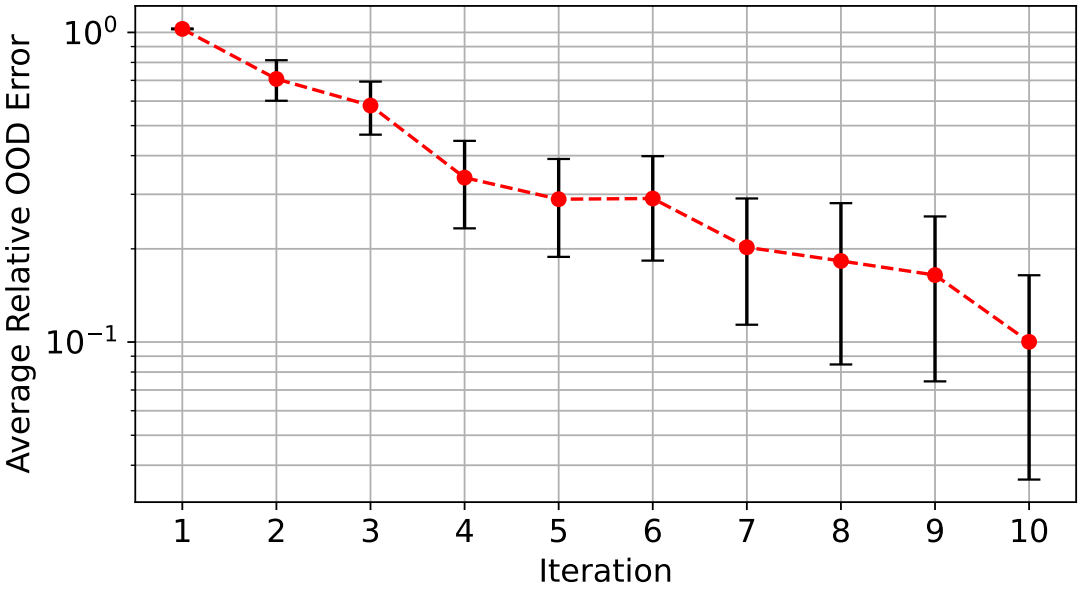
Forward Darcy



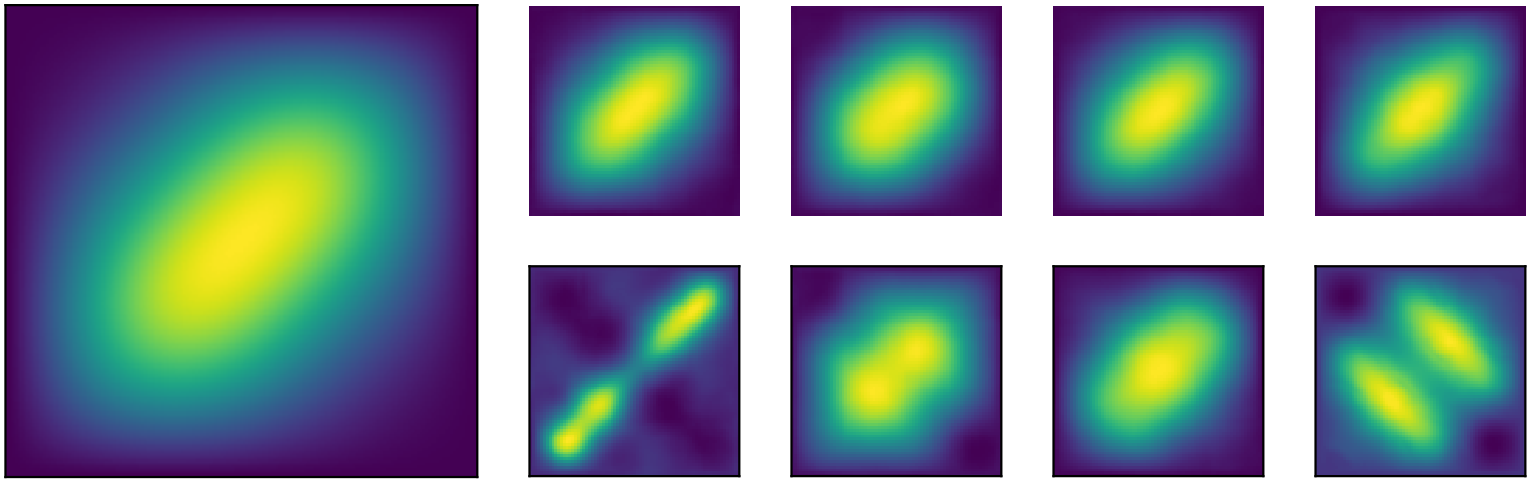


Forward NtD Map from EIT



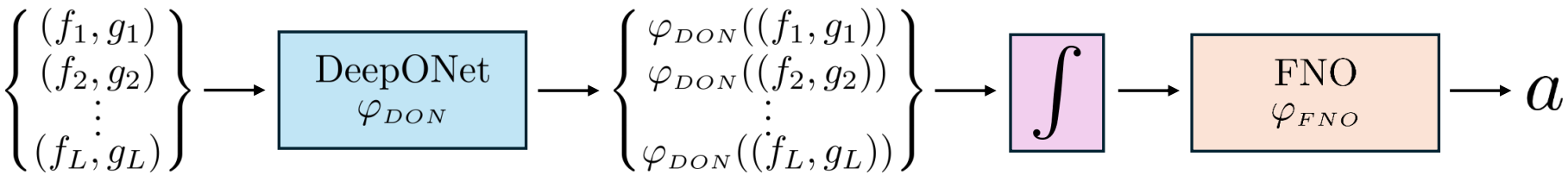


EIT Inverse Problem: Likelihood Shift in Forward Data



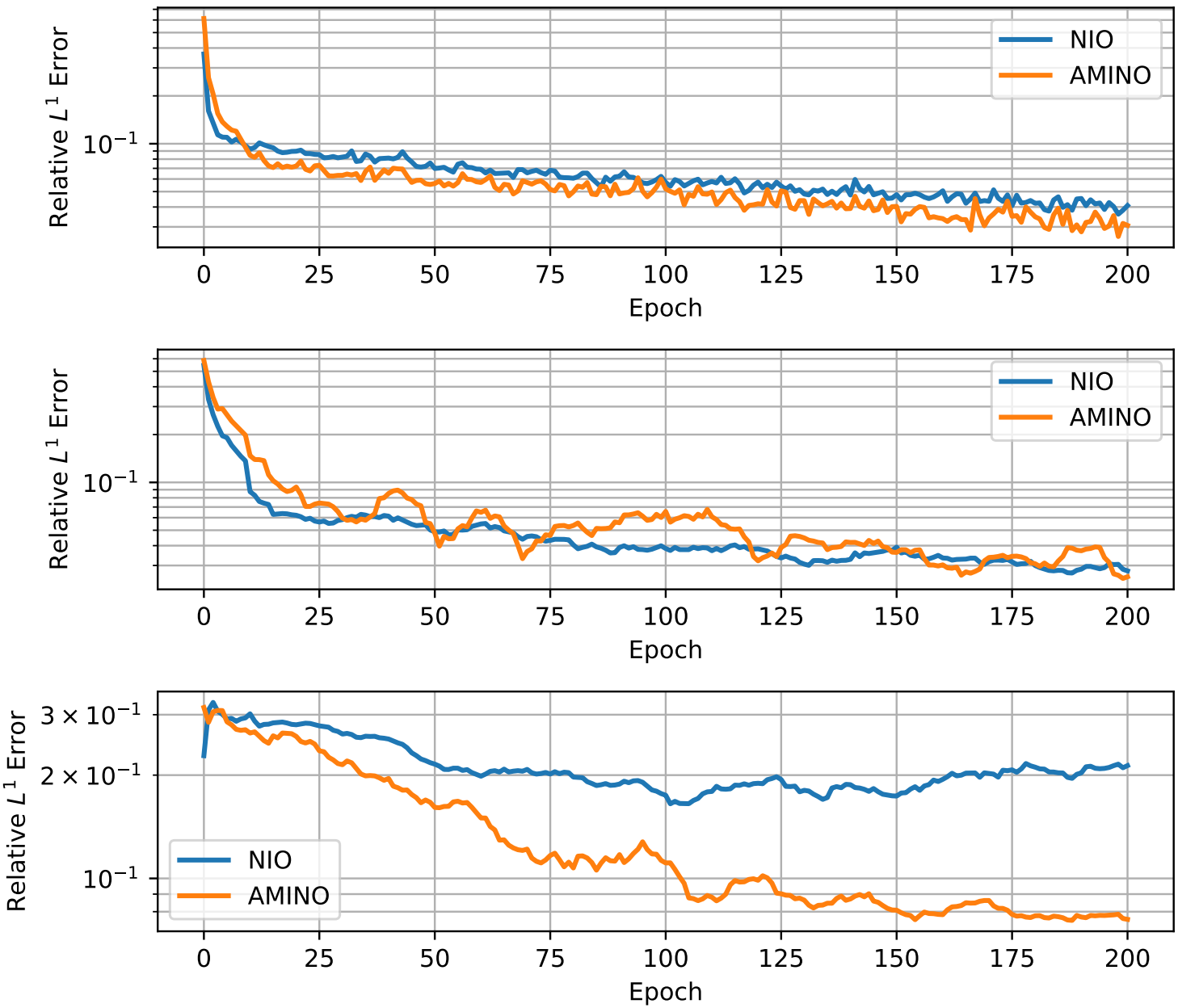
(top row) in-distribution; (bottom row) out-of-distribution

AMINO



$$\Psi_{\text{AMINO}}\left(\frac{1}{M}\sum_{m=1}^{M}\delta_{(g_m,y_m)}\right) = \varphi_{\text{FNO}}\left(\frac{1}{M}\sum_{m=1}^{M}\varphi_{\text{DON}}\big((g_m,y_m)\big)\right).$$

AMINO vs. NIO


Relative L^1 Error
10^{-1}
NIO
AMINO
0
25
50
75
100
125
150
175
200
Epoch
Relative L^1 Error
10^{-1}
NIO
AMINO
0
25
50
75
100
125
150
175
200
Epoch
Relative L^1 Error
3×10^{-1}
2×10^{-1}
10^{-1}
NIO
AMINO
0
25
50
75
100
125
150
175
200
Epoch