Solving and Learning Nonlinear PDEs with Gaussian Processes

Yifan Chen, Caltech

Joint work with

Bamdad Hosseini, Houman Owhadi, Florian Schaefer and Andrew Stuart

SoCAMS, 2022

Roadmap

Numerical Approximation via Inference

- **Partial Differential Equations** everywhere in CSE

$$\mathcal{F}(x, t, u, \partial_t u, \nabla_x u, \nabla_x^2 u, a, \xi, ...) = 0$$

  - Spatial PDEs, dynamics, inverse problems, stochastic, UQ, ...

- **Numerical Approximation** designed by experts
  - Finite difference/element/volume
  - Spectral methods
  - Boundary integral methods (BIM)
  - Meshless methods, collocation methods, ...

- **Inference and ML** to automate computing with partial information
  - GPs for numerical integration, ODEs, and linear PDEs
  - Bayes probabilistic numerics, Bayes numerical analysis, Bayes UQ
  - Physics informed ML (Deep Ritz, PINNs, SDEs...)
  - Operator learning (Kernels, Neural Operators, DeepONets), ...

This talk

Our Goal

A general GP framework for solving and learning nonlinear PDEs

- Methodology: simple, intepretable and rigorous[1]
  - generalize RBF collocation methods and BIM
  - PDEs as a Bayes inverse problem
- Efficiency: near-linear time and space complexity[2]
  - quantitative screening effects for GPs with PDE measurements

[1] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. "Solving and learning nonlinear pdes with gaussian processes". In: *Journal of Computational Physics* (2021).

[2] Yifan Chen, Florian Schaefer, and Houman Owhadi. "Sparse Cholesky Factorization for Solving Nonlinear PDEs via Gaussian Processes". In preparation.

Roadmap

The Methodology

A nonlinear elliptic PDE Example

- Consider the stationary elliptic PDE

$$\begin{cases} -\Delta u(\mathbf{x}) + \tau(u(\mathbf{x})) = f(\mathbf{x}), & \forall \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = g(\mathbf{x}), & \forall \mathbf{x} \in \partial\Omega. \end{cases}$$

  - Domain $\Omega \subset \mathbb{R}^d$.
  - PDE data $f, g : \Omega \to \mathbb{R}$.

- PDE has a unique strong/classical solution $u^\star$.

The Methodology: A Nonlinear Elliptic PDE

1. Choose a kernel $K : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$ (Choose the prior $\mathcal{GP}(0, K)$)
   - Corresponding RKHS $\mathcal{U}$ with norm $\|\cdot\|$
2. Choose some collocation points (Choose the data/likelihood)
   - $X^{\text{int}} = \{\mathbf{x}_1^{\text{int}}, \ldots, \mathbf{x}_{M^{\text{int}}}^{\text{int}}\} \subset \Omega$
   - $X^{\text{bd}} = \{\mathbf{x}_1^{\text{bd}}, \ldots, \mathbf{x}_{M^{\text{bd}}}^{\text{bd}}\} \subset \partial\Omega$
3. Solve the optimization problem (Find the "MAP")

$$\begin{cases} \underset{u \in \mathcal{U}}{\text{minimize}} \ \|u\| \\ \text{s.t.} \quad -\Delta u(\mathbf{x}_m) + \tau(u(\mathbf{x}_m)) = f(\mathbf{x}_m), & \text{for } \mathbf{x}_m \subset X^{\text{int}} \\ \qquad\qquad u(\mathbf{x}_n) = g(\mathbf{x}_n), & \text{for } \mathbf{x}_n \subset X^{\text{bd}} \end{cases}$$

How to Solve: Introducing Slack Variables

$$
\begin{cases}
\underset{u\in\mathcal{U}}{\text{minimize}} \ \|u\| \\
\text{s.t.} \quad -\Delta u(\mathbf{x}_m) + \tau(u(\mathbf{x}_m)) = f(\mathbf{x}_m), & \text{for } \mathbf{x}_m \subset X^{\text{int}} \\
\qquad\qquad u(\mathbf{x}_n) = g(\mathbf{x}_n), & \text{for } \mathbf{x}_n \subset X^{\text{bd}}
\end{cases}
$$

$$
\Updownarrow (N = M^{\text{bd}} + 2M^{\text{int}})
$$

$$
\begin{cases}
\underset{\mathbf{z}=(\mathbf{z}^{\text{bd}},\mathbf{z}^{\text{int}},\mathbf{z}^{\text{int}}_{\Delta})\in\mathbb{R}^N}{\text{minimize}} \quad \begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} \ \|u\| \\ \text{s.t.} \quad u(X^{\text{bd}}) = \mathbf{z}^{\text{bd}} \in \mathbb{R}^{M^{\text{bd}}} \\ \qquad u(X^{\text{int}}) = \mathbf{z}^{\text{int}} \in \mathbb{R}^{M^{\text{int}}} \\ \qquad \Delta u(X^{\text{int}}) = \mathbf{z}^{\text{int}}_{\Delta} \in \mathbb{R}^{M^{\text{int}}} \end{cases} \\
\text{s.t.} \quad -\mathbf{z}^{\text{int}}_{\Delta} + \tau(\mathbf{z}^{\text{int}}) = f(X^{\text{int}}) \\
\qquad \mathbf{z}^{\text{bd}} = g(X^{\text{bd}})
\end{cases}
$$

How to Solve: Inner optimization

- The inner problem is linear

$$\begin{aligned} &\underset{u\in\mathcal{U}}{\text{minimize}} \ \|u\| \\ &\text{s.t.} \quad u(X^{\text{bd}}) = \mathbf{z}^{\text{bd}}, u(X^{\text{int}}) = \mathbf{z}^{\text{int}}, \Delta u(X^{\text{int}}) = \mathbf{z}_{\Delta}^{\text{int}} \end{aligned}$$

- Kernel vector and matrix

$$K(\mathbf{x}, \boldsymbol{\phi}) = \left(K(\mathbf{x}, X^{\text{bd}}), K(\mathbf{x}, X^{\text{int}}), \Delta_{\mathbf{y}} K(\mathbf{x}, X^{\text{int}})\right) \in \mathbb{R}^N$$

$$K(\boldsymbol{\phi}, \boldsymbol{\phi}) = \begin{pmatrix} K(X^{\text{bd}}, X^{\text{bd}}) & K(X^{\text{bd}}, X^{\text{int}}) & \Delta_{\mathbf{y}} K(X^{\text{bd}}, X^{\text{int}}) \\ K(X^{\text{int}}, X^{\text{bd}}) & K(X^{\text{int}}, X^{\text{int}}) & \Delta_{\mathbf{y}} K(X^{\text{int}}, X^{\text{int}}) \\ \Delta_{\mathbf{x}} K(X^{\text{int}}, X^{\text{bd}}) & \Delta_{\mathbf{x}} K(X^{\text{int}}, X^{\text{int}}) & \Delta_{\mathbf{x}}\Delta_{\mathbf{y}} K(X^{\text{int}}, X^{\text{int}}) \end{pmatrix} \in \mathbb{R}^{N\times N}$$

Minimizer $u(\mathbf{x}) = K(\mathbf{x}, \boldsymbol{\phi}) K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \mathbf{z}$

How to Solve: Finite Dimensional Representation

Combine the two level optimization:

Representer Theorem

Every minimizer $u^\dagger$ can be represented as

$$u^\dagger(\mathbf{x}) = K(\mathbf{x}, \boldsymbol{\phi}) K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \mathbf{z}^\dagger$$

where the vector $\mathbf{z}^\dagger \in \mathbb{R}^N$ is a minimizer of

$$\begin{cases} \min\limits_{\mathbf{z} \in \mathbb{R}^N} & \mathbf{z}^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \mathbf{z} \\ \text{s.t.} & F(\mathbf{z}) = \mathbf{y} \end{cases}$$

- Function $F : \mathbb{R}^N \to \mathbb{R}^M$ depends on PDE collocation constraints
- $\mathbf{y}$ contains PDE boundary and RHS data

Roadmap

Numerical Experiments: Elliptic PDEs

- Nonlinear Elliptic Equation, $\tau(u) = u^3$

$$\begin{cases} -\Delta u(\mathbf{x}) + \tau(u(\mathbf{x})) = f(\mathbf{x}), & \forall \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = g(\mathbf{x}), & \forall \mathbf{x} \in \partial\Omega. \end{cases}$$

- Truth: $d = 2$, $u^\star(\mathbf{x}) = \sin(\pi x_1)\sin(\pi x_2) + 4\sin(4\pi x_1)\sin(4\pi x_2)$
- Kernel: $K(\mathbf{x}, \mathbf{y}; \sigma) = \exp(-\frac{|\mathbf{x}-\mathbf{y}|^2}{2\sigma^2})$

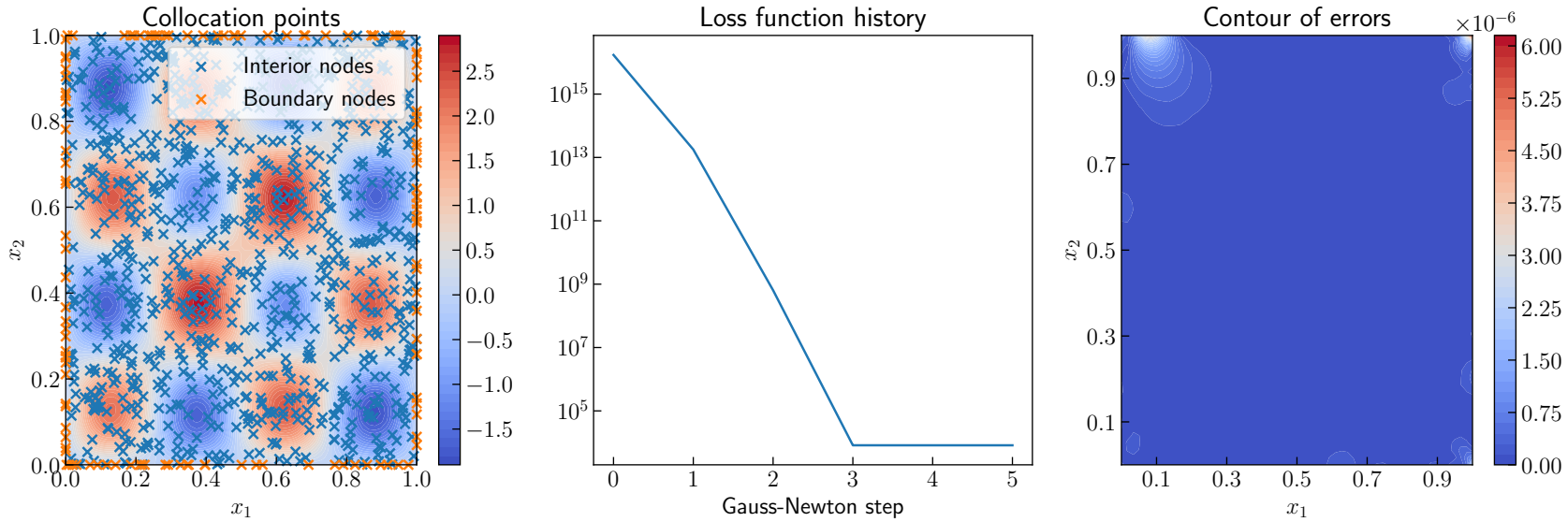


Figure: $N_{\text{domain}} = 900, N_{\text{boundary}} = 124$

Convergence Study

- For $\tau(u) = 0, u^3$, use Gaussian kernel with lengthscale σ
- L^2, L^∞ accuracy, compared with Finite Difference (FD)

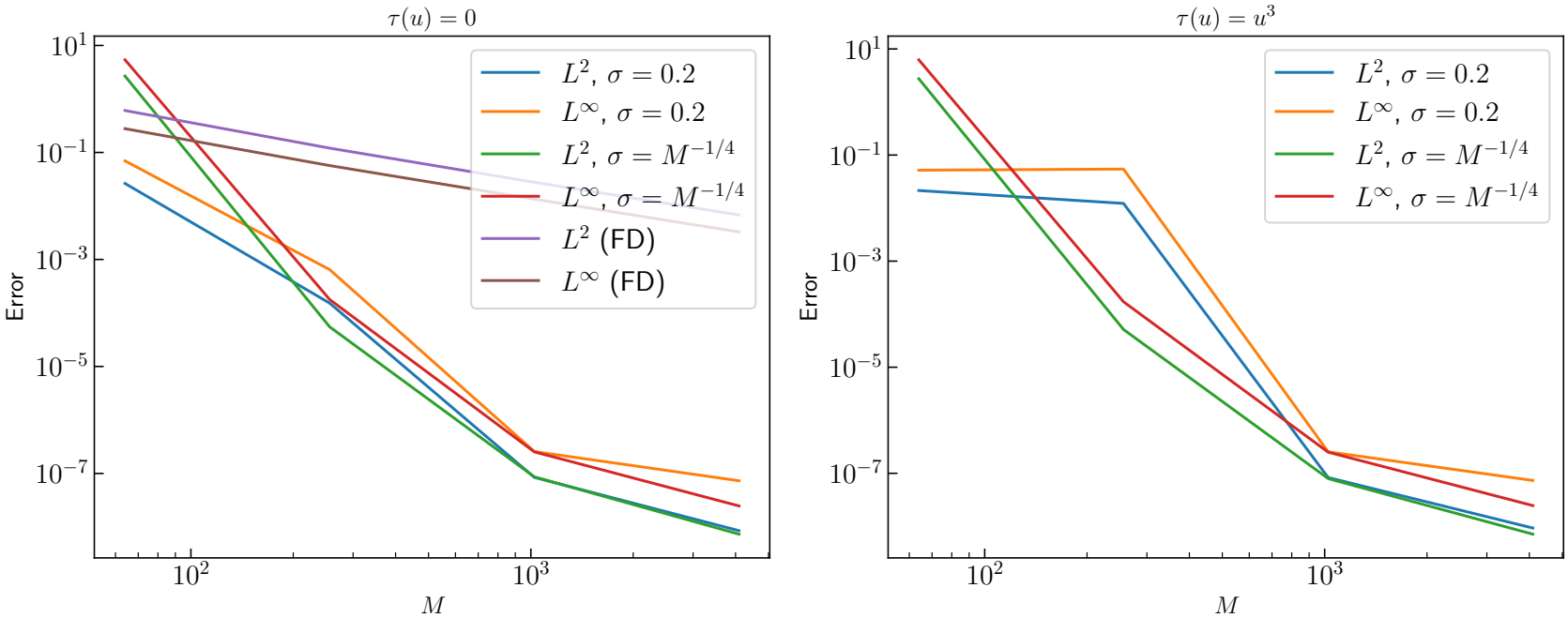


Figure: Convergence of the kernel method is fast, since the solution is smooth

Roadmap

Numerical Experiments: Viscous Burgers' Equation

- Kernel: $K((s,t),(s',t')) = \exp\left(-20^2|s-s'|^2 - 3^2|t-t'|^2\right)$

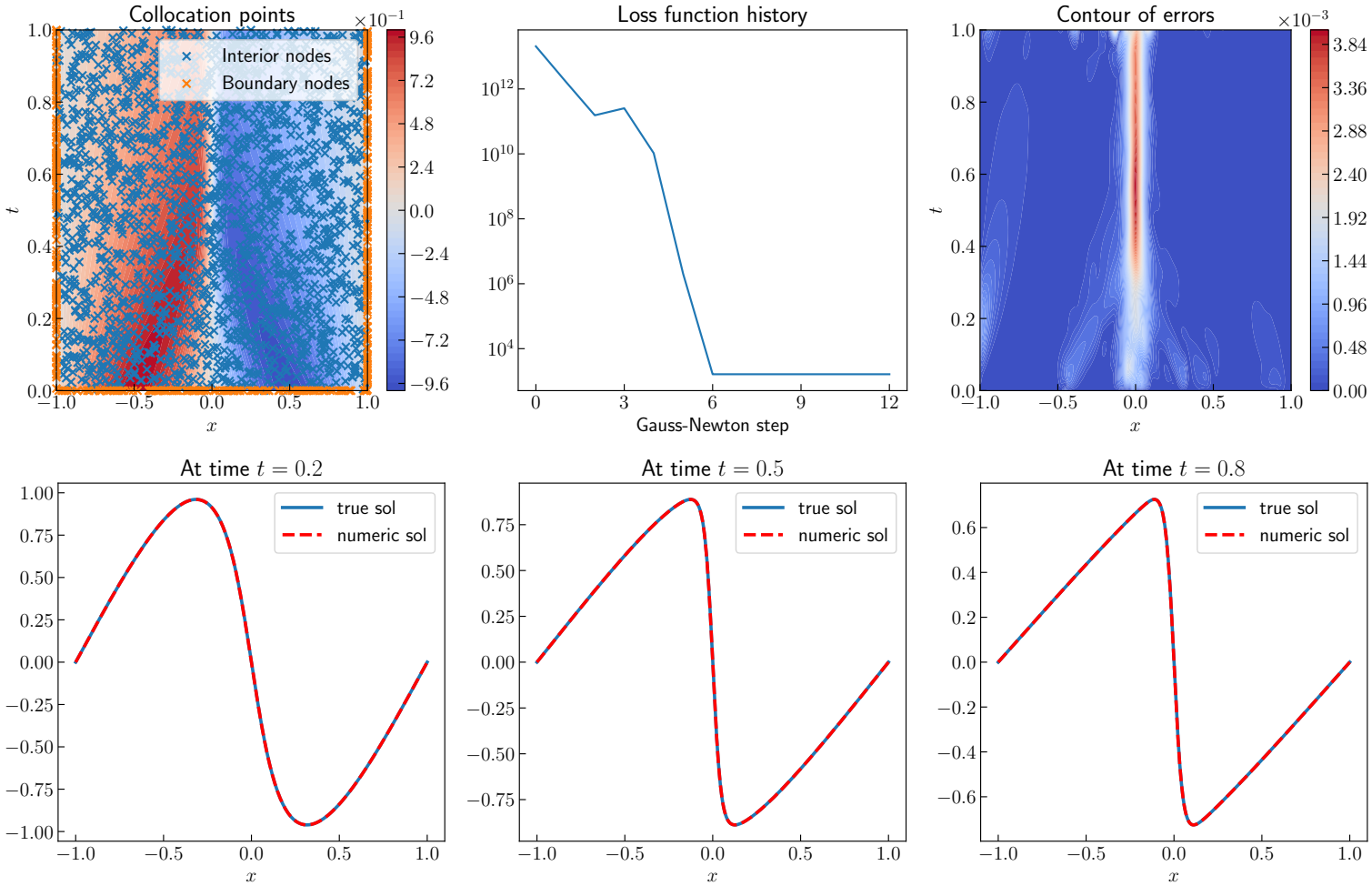


Figure: $N_{\text{domain}} = 2000, N_{\text{boundary}} = 400$

Push to Small Viscosity

Discretize in time first, then apply the methodology to the resulting spatial PDE: dimension of kernel matrices is reduced

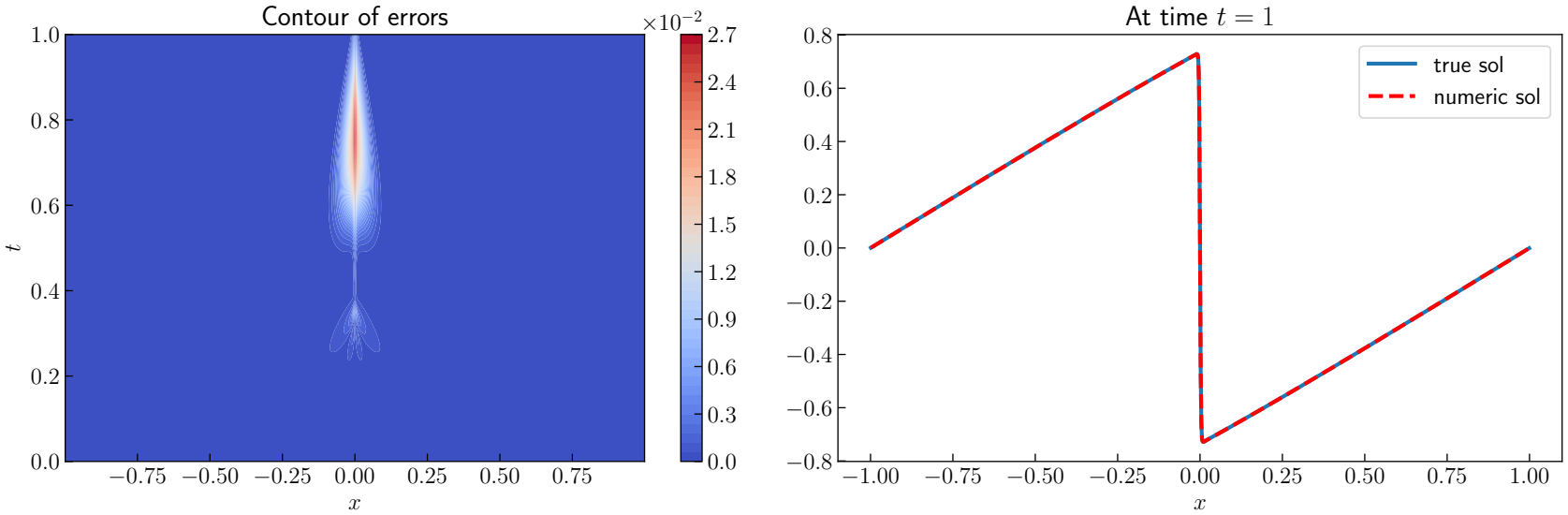


Figure: $\nu = 10^{-3}$; number of spatial points 2000; time step size 0.01; Matern7/2 kernel with lengthscale 0.02; use 2 GN iterations

At time $t = 1$, L^2 accuracy: 10^{-4}

Roadmap

Numerical Experiments: Inverse Problems

Darcy Flow inverse problems

$$\begin{cases} \min\limits_{u,a} \|u\|_K^2 + \|a\|_\Gamma^2 + \dfrac{1}{\gamma^2}\displaystyle\sum_{j=1}^{I} |u(\mathbf{x}_j) - o_j|^2, \\ \text{s.t.} \qquad -\mathsf{div}(\exp(a)\nabla u)(\mathbf{x}_m) = 1, \qquad \forall \mathbf{x}_m \in (0,1)^2 \\ \qquad\qquad\qquad\quad u(\mathbf{x}_m) = 0, \qquad\quad \forall \mathbf{x}_m \in \partial(0,1)^2. \end{cases}$$

- Recover a from pointwise measurements of u
- Model (u, a) as independent GPs
- Impose PDE constraints and formulate Bayesian inverse problem

Numerical Experiments: Darcy Flow

- Kernel $K(\mathbf{x}, \mathbf{x}'; \sigma) = \exp\left(-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2\sigma^2}\right)$ for both u and a

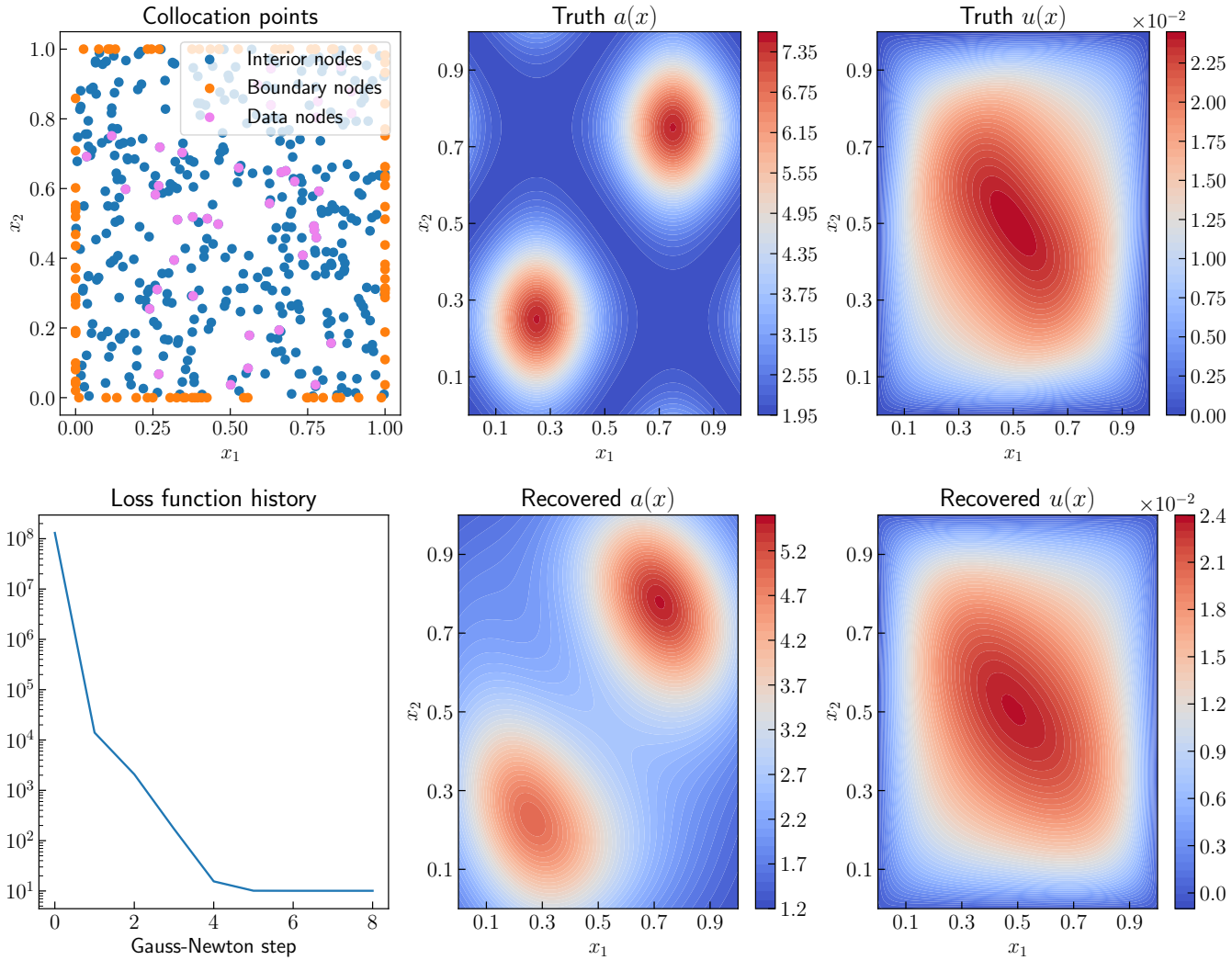


Figure: $N_{\text{domain}} = 400, N_{\text{boundary}} = 100, N_{\text{observation}} = 50$

Roadmap

Scalability: Sparse Cholesky Factorization

- Sparse Cholesky for kernel matrices under coarse to fine ordering[3]
- $O(N\rho^d)$ memory and $O(N\rho^{2d})$ time (ρ sparsity parameter)

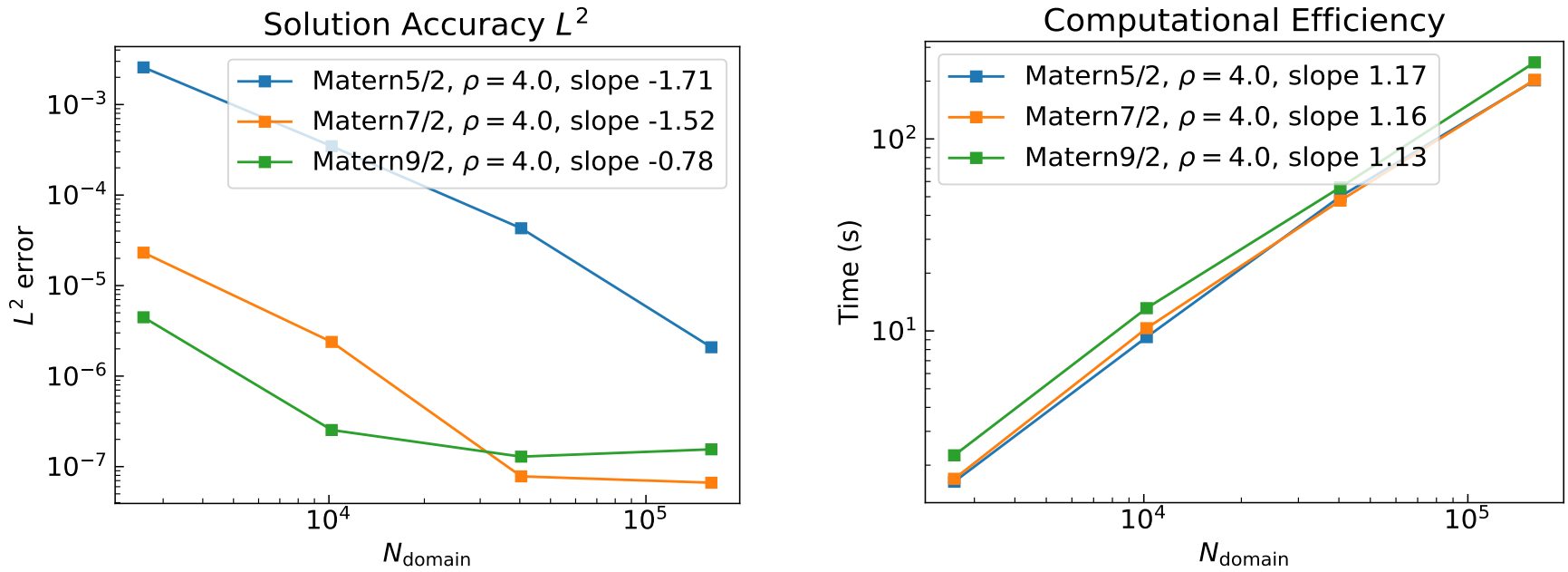


Figure: Run 3 GN iterations. Accuracy floor due to finite ρ and regularization

[3] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. "Sparse Cholesky Factorization by Kullback–Leibler Minimization". In: *SIAM Journal on Scientific Computing* 43.3 (2021), A2019–A2046.

Theoretical Foundation: Consistency

Consistency of the minimizer

$$\begin{cases} \min\limits_{u \in \mathcal{U}} & \|u\| \\ \text{s.t.} & \text{PDE constraints at } \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \in \overline{\Omega}. \end{cases}$$

Convergence theory

- K is chosen so that
  - $\mathcal{U} \subseteq H^s(\Omega)$ for some $s > s^*$ where $s^* = d/2 +$ order of PDE.
  - $u^\star \in \mathcal{U}$.
- Fill distance of $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \to 0$ as $M \to \infty$.

Then as $M \to \infty$, $u^\dagger \to u^\star$ pointwise in Ω and in $H^t(\Omega)$ for $t \in (s^*, s)$.

Roadmap

Take-aways

Solving and Learning Nonlinear PDEs with Gaussian Processes

Algorithm

- A simple framework for solving and learning nonlinear PDEs
- Near-linear complexity treatment of the dense kernel matrices
- Experiments: stationary PDEs, time dependent, inverse problems

Theory

- Consistency as fill-in distance goes to 0 (asymptotic only)
- Not covered: learning the kernel[4]

Thank you!

[4] Yifan Chen, Houman Owhadi, and Andrew Stuart. "Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation". In: *Mathematics of Computation* (2021).

Backup Slides

Towards A Practical Algorithm

Quadratic optimization with nonlinear constraints

- A simple linearization algorithm $\mathbf{z}^k \to \mathbf{z}^{k+1}$

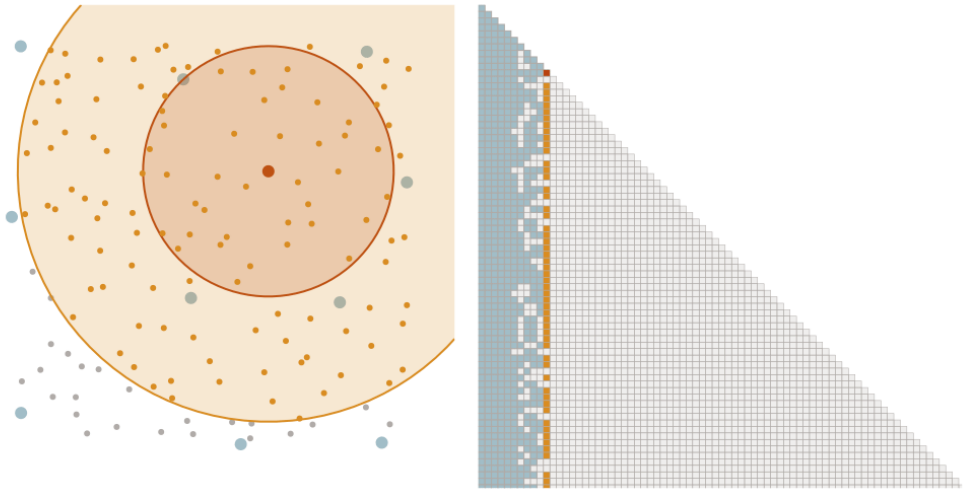
$$\begin{cases} \min\limits_{\mathbf{z}\in\mathbb{R}^N} & \mathbf{z}^T K(\boldsymbol{\phi},\boldsymbol{\phi})^{-1}\mathbf{z} \\ \text{s.t.} & F(\mathbf{z}^k)+F'(\mathbf{z}^k)(\mathbf{z}-\mathbf{z}^k)=\mathbf{y}. \end{cases}$$

"Newton's iteration for the nonlinear PDE"

- Poor conditioning of $K(\boldsymbol{\phi},\boldsymbol{\phi})$, and scale imbalance between blocks
  Adding scale-aware regularization $K(\boldsymbol{\phi},\boldsymbol{\phi})+\lambda\text{diag}(K(\boldsymbol{\phi},\boldsymbol{\phi}))$

Sparse Cholesky Factorization for Ordinary Kernel Matrice

Sparse Cholesky factor for kernel matrices under coarse to fine ordering[5]



Coarse to fine: max-min ordering

$$x_k = \text{argmax}_{x_i} d(x_i, \{x_j, 1 \leq j < k\})$$

with lengthscale $l_k = d(x_k, \{x_j, 1 \leq j < k\})$

[5] F Schäfer, TJ Sullivan, and H Owhadi. "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity". In: *Multiscale Modeling & Simulation* 19.2 (2021), pp. 688–730.

Why Sparse? Cholesky Factors and Screening Effects

Let $\Theta \in \mathbb{R}^{d\times d}$, $\Theta_{ij} = k(x_i, x_j)$, and $X \sim \mathcal{N}(0, \Theta)$

- Cholesky factor of the covariance matrix $\Theta = LL^T$

$$\frac{L_{ij}}{L_{jj}} = \frac{\mathrm{Cov}[X_i, X_j | X_{1:j-1}]}{\mathrm{Var}[X_j | X_{1:j-1}]} \qquad (i \geq j)$$

- Cholesky factor of the precision matrix $\Theta^{-1} = UU^T$

$$\frac{U_{ij}}{U_{jj}} = (-1)^{i\neq j} \frac{\mathrm{Cov}[X_i, X_j | X_{1:j-1\setminus\{i\}}]}{\mathrm{Var}[X_j | X_{1:j-1\setminus\{i\}}]} \qquad (i \leq j)$$

Screening effects: $x_{1:j}$ ordered from coarse to fine; scale of x_j is l_j, then for certain kernel arsing from PDEs[7]

$$\mathrm{Cov}[X_i, X_j | X_{1:j-1}] \lesssim \exp\left(-\frac{d(x_i, x_j)}{l_j}\right)$$

[6] Michael L Stein. "The screening effect in kriging". In: *Annals of statistics* 30.1 (2002), pp. 298–323.

[7] Schäfer, Sullivan, and Owhadi, "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity".

Screening Effects with PDE measurements

Recall the kernel matrices

$$\begin{pmatrix} K(X^{\text{bd}}, X^{\text{bd}}) & K(X^{\text{bd}}, X^{\text{int}}) & \Delta_{\mathbf{y}} K(X^{\text{bd}}, X^{\text{int}}) \\ K(X^{\text{int}}, X^{\text{bd}}) & K(X^{\text{int}}, X^{\text{int}}) & \Delta_{\mathbf{y}} K(X^{\text{int}}, X^{\text{int}}) \\ \Delta_{\mathbf{x}} K(X^{\text{int}}, X^{\text{bd}}) & \Delta_{\mathbf{x}} K(X^{\text{int}}, X^{\text{int}}) & \Delta_{\mathbf{x}}\Delta_{\mathbf{y}} K(X^{\text{int}}, X^{\text{int}}) \end{pmatrix}$$

How to order when there are derivative measurements?

- Order pointwise measurements from coarse to fine
- PDE measurements follow behind (with the same ordering)

Theorem: screening effects hold for such ordering[8]

Theory: need technical assumptions

- The kernel is the Green function of some differential operator $\mathcal{L} : H_0^s(\Omega) \to H^{-s}(\Omega)$

Practice: works more generally

[8] Chen, Schaefer, and Owhadi, "Sparse Cholesky Factorization for Solving Nonlinear PDEs via Gaussian Processes".