Numerical Computation via Inference

Gaussian Processes and Multiscale Methods

Yifan Chen, Caltech

Advisors:

Thomas Hou, Houman Owhadi, and Andrew Stuart

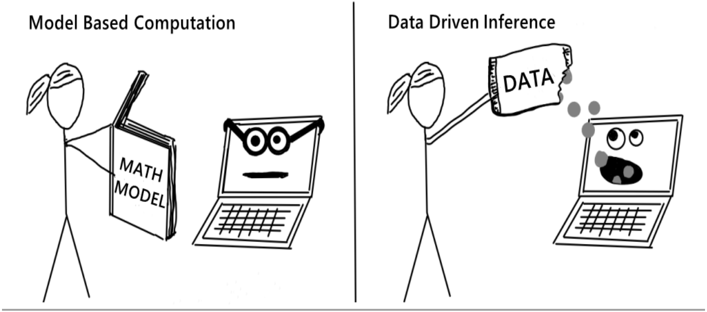
PhD Candidacy Talk, May 2022

Roadmap

When I Came to Caltech to Study

Applied and computational math research

- Model Based: ODEs/SDEs/PDEs, physics, numerical analysis, ...
- Data Driven: machine learning, optimization, statistics, ...



"Now and Future: Model + Data!"

Model Computation: Computing with Partial Information

Numerical algorithms designed to use IC/BC/RHS data wisely

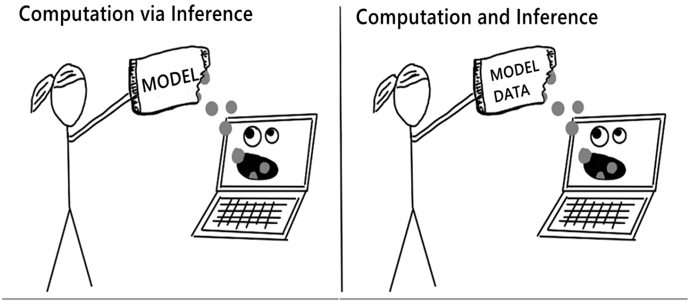
- Finite difference/element/volume
- Spectral methods
- Boundary integral methods
- Meshless methods, collocation methods
- Multiscale methods, numerical homogenization, ...

Inference and ML to automate numerical computation

- Gaussian process (GP) and kernel methods for numerical integration
- GPs and kernel methods for ODEs, linear PDEs
- Information based complexity
- Bayes probabilistic numerics, Bayes numerical analysis, UQ
- Physics informed ML (Deep Ritz methods, PINNs, SDEs...)
- Operator learning (Kernels, Neural Operators, DeepONets), ...

My PhD Wondering: Computation via Inference

Solving PDEs, surrogate models, inverse problems, via inference



The question: How to transform a given model based computational question wisely to data science inference question?

Two Stories

1. Bayes based Gaussian processes approaches
   - Interpretable, rigorous and unified framework for solving PDEs and inverse problems

Solving and learning PDEs as a Bayes inference problem

2. Approximation based multiscale methods
   - Exponentially convergent multiscale approximation for rough elliptic PDEs and Helmholtz equations

Multiscale ideas lead to exponentially efficient inference

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 $u^\star$.

A Nonlinear Elliptic PDE: The Methodology[1]

1. Choose a kernel $K : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$
   - Corresponding RKHS $\mathcal{U}$ with norm $\|\cdot\|$
2. Choose some collocation points
   - $X^{\mathsf{int}} = \{\mathbf{x}_1^{\mathsf{int}}, \ldots, \mathbf{x}_{M^{\mathsf{int}}}^{\mathsf{int}}\} \subset \Omega$
   - $X^{\mathsf{bd}} = \{\mathbf{x}_1^{\mathsf{bd}}, \ldots, \mathbf{x}_{M^{\mathsf{bd}}}^{\mathsf{bd}}\} \subset \partial\Omega$
3. Solve the optimization problem

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

Generalization of RBF collocation methods and boundary integral methods (BIM)

[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).

Bayes Inference Interpratation of the Methodology

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)
   - $X^{\mathsf{int}} = \{\mathbf{x}_1^{\mathsf{int}}, \ldots, \mathbf{x}_{M^{\mathsf{int}}}^{\mathsf{int}}\} \subset \Omega$
   - $X^{\mathsf{bd}} = \{\mathbf{x}_1^{\mathsf{bd}}, \ldots, \mathbf{x}_{M^{\mathsf{bd}}}^{\mathsf{bd}}\} \subset \partial\Omega$
3. Solve the optimization problem (Find the "MAP")

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

Generalize linear PDEs setting in Bayes probabilistic numerical methods[2]

[2] Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. "Bayesian probabilistic numerical methods". In: *SIAM Review* 61.4 (2019), pp. 756–789.

How to Solve: Introducing Slack Variables

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}}\ \|u\| \\ \text{s.t.} \quad -\Delta u(\mathbf{x}_m) + u(\mathbf{x}_m)^3 = f(\mathbf{x}_m), \quad \text{for } \mathbf{x}_m \subset X^{\mathsf{int}} \\ \qquad\qquad\qquad u(\mathbf{x}_n) = g(\mathbf{x}_n), \quad \text{for } \mathbf{x}_n \subset X^{\mathsf{bd}} \end{cases}$$

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

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

How to Solve: Inner optimization

- The inner problem is linear

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

- Measurement vector $\boldsymbol{\phi} := (\delta_{X^{\mathsf{bd}}}, \delta_{X^{\mathsf{int}}}, \delta_{X^{\mathsf{int}}}\circ\Delta) \in (\mathcal{U}^*)^{\otimes N}$
- Kernel vector and matrix

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

$$K(\boldsymbol{\phi},\boldsymbol{\phi}) =$$

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

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

How to Solve: Representation of the Minimizer

Combine the two level optimization:

Representer theorem

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

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

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

Towards A Practical Algorithm

Quadratic optimization with nonlinear constraints

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

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

"Newton's iteration for the nonlinear PDE"

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

Numerical Experiments

- 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: $K(\mathbf{x}, \mathbf{y}; \sigma) = \exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2}{2\sigma^2}\right)$

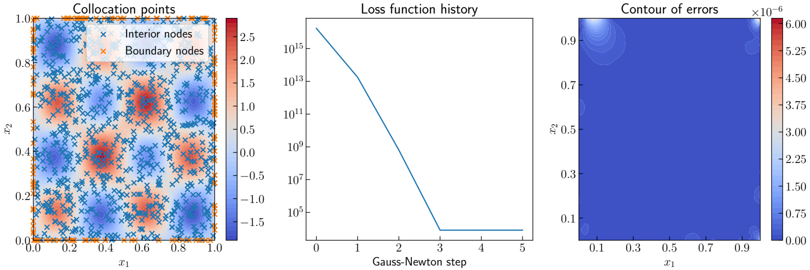


Figure: $N_{\text{domain}} = 900, N_{\text{boundary}} = 124$

Convergence Study

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

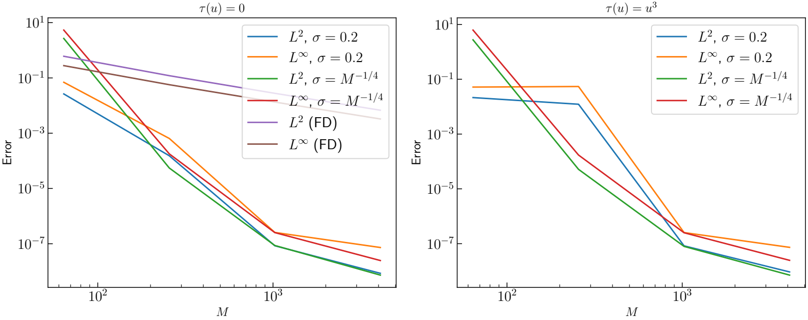


Figure: Convergence of the kernel method is fast, since the solution is smooth

Other Successful Examples

Time dependent viscous Burgers equation

- Spatio-temporal GPs approach
- Time discretization + spatial GPs: causality considered

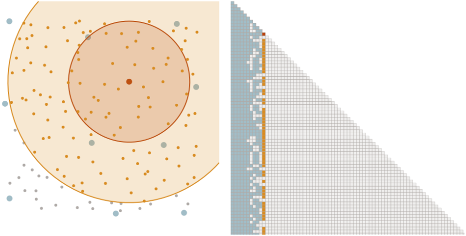
Inverse problem: Darcy flow

- PDE data + observation data treated in the same manner
- Solving PDEs and inverse problems in a unified algorithmic framework

Roadmap

Sparse Cholesky Factorization for Ordinary Kernel Matrice

Sparse Cholesky factor for kernel matrices under coarse to fine ordering[3]



Coarse to fine: max-min ordering

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

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

[3] 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[5]

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

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

[5] 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

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

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[6]

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

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

Near Linear Complexity by Sparse Cholesky

- Ignore correlation beyond $d(x, x_j) \geq \rho l_j$ (which is $O(\exp(-\rho))$)
- Once ordering and sparsity pattern determined, use KL minimization algorithm[7]: $O(N\rho^d)$ memory and $O(N\rho^{2d})$ time

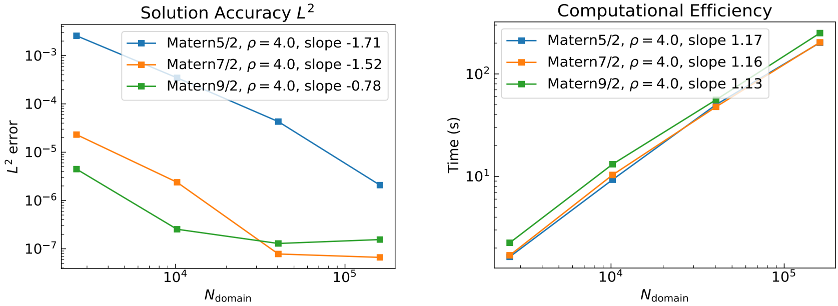


Figure: Run 3 GN iterations. Accuracy floor due to finite ρ and regularization

[7] 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.

Roadmap

Theoretical Foundation: Consistency

Consistency of the minimizer

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

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 $t \in (s^*, s)$.

Theoretical Foundation: Kernel Learning

- Bayes approach built in GPs: e.g. Empirical Bayes (EB)

$$\theta^{\mathrm{EB}} = \operatorname*{argmin}_{\theta} \|u^\dagger(\cdot, X, \theta)\|_{K_\theta}^2 + \log\det K_\theta(X, X)$$

  where, $u^\dagger(\cdot, X, \theta)$ is the solution using collocation points X and kernel K_θ, and $\|\cdot\|_{K_\theta}$ is the RKHS norm for the kernel K_θ

- Kernel Flow (KF)[8]: a variant of cross-validation

$$\theta^{\mathrm{KF}} = \operatorname*{argmin}_{\theta} \ \mathbb{E}_\pi \frac{\|u^\dagger(\cdot, X, \theta) - u^\dagger(\cdot, \pi X, \theta)\|_{K_\theta}^2}{\|u^\dagger(\cdot, X, \theta)\|_{K_\theta}^2}$$

  where, πX is a subsampling of X

Consistency and robustness of EB and KF for learning Matérn-like kernels: both has large data limit, EB optimal while KF robust[9]

[8] Houman Owhadi and Gene Ryan Yoo. "Kernel flows: From learning kernels from data into the abyss". In: *Journal of Computational Physics* 389 (2019), pp. 22–47.

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

Roadmap

Recap

What so far

PDEs treated as nonlinear combination of linear differential measurement data of a GP, then solved via inference of the GP (MAP estimator)

- Framework: choose the GP prior, choose the data, then inference
- MAP estimator: generalization of RBF collocation methods and BIM
- Efficient algorithm, theoretical consistency, parameter learning

Potential issue in the prior choice

- Kernel selection unrelated to the specific PDE

Potential issue in the data choice

- Collocation methods, require strong solution

A Linear Rough Elliptic PDE Example

For $a \in L^\infty(\Omega), f \in L^2(\Omega)$:

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

Choose kernel K, apply the methodology:

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} \ \|u\| \\ \text{s.t.} \quad -\nabla \cdot (a\nabla u)(\mathbf{x}_m) = f(\mathbf{x}_m), & \text{for } \mathbf{x}_m \subset X^{\text{int}} \\ \qquad\qquad u(\mathbf{x}_n) = 0, & \text{for } \mathbf{x}_n \subset X^{\text{bd}} \end{cases}$$

Not work, since $u \in H_0^1(\Omega)$ only

The collocation data we formulate from the PDE is not appropriate!

Recall the Framework

1. Choose the prior $\mathcal{GP}(0, K)$
2. Choose the data from the computational problem
3. Find the "MAP" / optimal recovery

$$\begin{cases} \underset{u \in \mathcal{U}}{\text{minimize}} \; \|u\| \\ \text{s.t. } \text{Data of } u \end{cases}$$

Switch to Choose Weak Data

Choose kernel K that satisfies BC, and choose $\psi_i \in H_0^1(\Omega), 1 \leq i \leq N$

$$\begin{cases} \underset{u \in \mathcal{U}}{\text{minimize}} \ \|u\| \\ \text{s.t.} \quad \langle \nabla \psi_i, a \nabla u \rangle = \langle \psi_i, f \rangle \text{ for } 1 \leq i \leq N \end{cases}$$

If K is the Green function[10] of $-\nabla \cdot (a \nabla \cdot)$, then apply Lagrangian dual:

$$-\underset{v \in \mathsf{span}\{\psi_i, 1 \leq i \leq N\}}{\text{minimize}} \left(\frac{1}{2} \langle \nabla v, a \nabla v \rangle - \langle v, f \rangle \right)$$

Recover Galerkin methods using basis functions $\psi_i, 1 \leq i \leq N$

[10] If $d > 1$, $\mathcal{U}$ is the more general Cameron-Martin space rather than RKHS

Choose Weak Data Dependent on the Green Function

If choosing

$$\mathrm{span}\{\psi_i, 1 \leq i \leq N\} = (-\nabla \cdot (a\nabla \cdot))^{-1}\,\mathrm{span}\{\phi_i, 1 \leq i \leq N\}$$

Then the equivalent inference problem becomes a simple one

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}}\ \|u\| \\ \text{s.t.} \quad \langle \phi_i, u\rangle \text{ known, for } 1 \leq i \leq N \end{cases}$$

Some incomplete literature:

- ϕ_i finite element function of local support $O(H)$[11]
- ϕ_i piecewise constant function of local support $O(H)$[12]

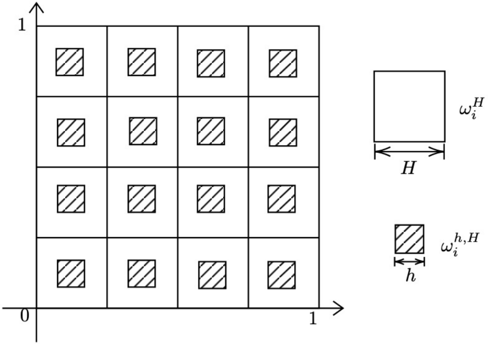
Accuracy: $O(H)$ in $H^1_a(\Omega)$ norm
Localization: ψ_i can be localized of size $O(H\log(1/H))$

[11] Axel Målqvist and Daniel Peterseim. "Localization of elliptic multiscale problems". In: *Mathematics of Computation* 83.290 (2014), pp. 2583–2603.

[12] Houman Owhadi. "Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games". In: *SIAM Review* 59.1 (2017), pp. 99–149.

Possibility: Subsampled Measurement Functions?

Subsampled measurements: $\phi_i^{h,H}$ supported in $\omega_i^{h,H}$



The middle between Diracs ($h = 0$) and $h = H$

Accuracy and Localization for Subsampled Data

Approximation accuracy[13]: $O(H\rho_d(\frac{H}{h}))$ in the $H^1_a(\Omega)$ norm

$$\rho_d(t) = \begin{cases} 1, & d < 2 \\ \sqrt{\log(1+t)}, & d = 2 \\ t^{\frac{d-2}{2}}, & d > 2\,. \end{cases}$$

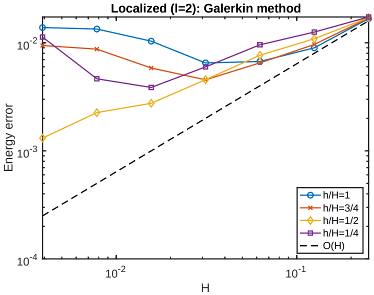
Localization[14]: exponential decay rate of $\psi_i^{h,H}$ exhibits non-monotone behavior regarding h

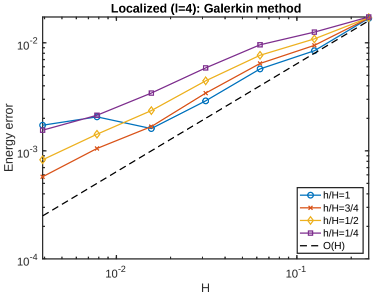
A trade-off between approximation and localization: ratio h/H matters

[13] Yifan Chen and Thomas Y Hou. "Function approximation via the subsampled Poincaré inequality". In: *Discrete & Continuous Dynamical Systems* 41.1 (2021), p. 169.

[14] Yifan Chen and Thomas Y Hou. "Multiscale elliptic PDE upscaling and function approximation via subsampled data". In: *Multiscale Modeling & Simulation* 20.1 (2022), pp. 188–219.

Numerical Examples





Summary for now

Solving PDEs from GP inference perspectives

Choose prior:

- Parametric kernel + kernel learning
- Green function as the kernel

Choose data:

- Collocation data
- Weak form data

Question: convergence rates, i.e. inference efficiency?

- Depend on the smoothness of the solution
- Usually algebraic, unless the solution is smooth

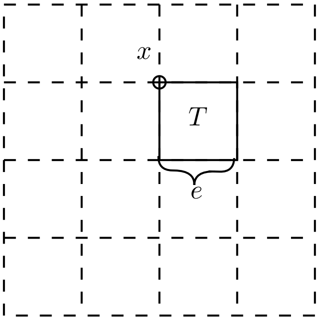
Can we choose the data more thoughtfully to get exponential convergence, even for nonsmooth solution?

Roadmap

- Consider Hemholtz equation
  $-\nabla\cdot(a\nabla u)-k^2u=f$
- Local decomposition:
  mesh size $H=O(1/k)$,
  in each T, $u=u_T^{\mathsf{h}}+u_T^{\mathsf{b}}$

$$\begin{cases}-\nabla\cdot(a\nabla u_T^{\mathsf{h}})-k^2u_T^{\mathsf{h}}=0, & \text{in } T\\ u_T^{\mathsf{h}}=u, & \text{on } \partial T\end{cases}$$

$$\begin{cases}-\nabla\cdot(a\nabla u_T^{\mathsf{b}})-k^2u_T^{\mathsf{b}}=f, & \text{in } T\\ u_T^{\mathsf{b}}=0, & \text{on } \partial T\end{cases}$$



$x\in\mathcal{N}_H, e\in\mathcal{E}_H, T\in\mathcal{T}_H$

- Global function:
  $u^{\mathsf{h}}(x)=u_T^{\mathsf{h}}(x), u^{\mathsf{b}}(x)=u_T^{\mathsf{b}}(x)$
  when $x\in T$ for each T
- Coarse-fine decomposition:
  $u=u^{\mathsf{h}}+u^{\mathsf{b}}$
  u^{h} coarse part, u^{b} fine part

Stick to Case $k = 0$ and Dirichlet BC for Simplicity

Coarse and fine scale space

- $u = u^{\mathsf{h}} + u^{\mathsf{b}} \in V^{\mathsf{h}} \oplus_a V^{\mathsf{b}}$

$$V^{\mathsf{h}} = \{v \in H_0^1(\Omega) : -\nabla \cdot (a\nabla v) = 0 \text{ in every } T \in \mathcal{T}_H\}$$
$$V^{\mathsf{b}} = \{v \in H_0^1(\Omega) : v = 0 \text{ on } \partial T, \text{ for every } T \in \mathcal{T}_H\}$$

$$\boxed{H_0^1(\Omega) = V^{\mathsf{h}} \oplus_a V^{\mathsf{b}}}$$

- Fine scale part u^{b} solved locally
- Coarse scale part u^{h} depends on edge values of u

Recall the inference framework: How to get data of u^{h}?
Choose test function $\psi \in V^{\mathsf{h}}$, then

$$\langle \psi, f \rangle = \langle \nabla\psi, a\nabla u \rangle = \langle \nabla\psi, a\nabla u^{\mathsf{h}} \rangle$$

This is a measurement of u^{h}

Roadmap

How to approximate u^{h} using basis functions?

Theorem $(d=2)$[15] [16]

On a mesh of size $H = O(1/k)$, there exist c_i, d_i such that

$$u^{\mathsf{h}} = \sum_{i \in I_1} c_i \psi_i^{\mathrm{MsFEM}} + \sum_{i \in I_2} d_i \psi_i^{\mathrm{Edge}} + \; O\left(\exp\left(-m^{\frac{1}{d+1}-\epsilon}\right)\right)$$

where the approximation is in the energy norm, and

- ψ_i^{MsFEM} is the MsFEM basis with linear BC $\qquad \#I_1 = O(1/H^2)$
- ψ_i^{Edge} computed by solving local equation and spectral problems $\qquad \#I_2 = O(2m/H^2)$

[16] Yifan Chen, Thomas Y Hou, and Yixuan Wang. "Exponential convergence for multiscale linear elliptic PDEs via adaptive edge basis functions". In: *Multiscale Modeling & Simulation* 19.2 (2021), pp. 980–1010.

[16] Yifan Chen, Thomas Y Hou, and Yixuan Wang. "Exponentially convergent multiscale methods for high frequency heterogeneous Helmholtz equations". In: *arXiv preprint arXiv:2105.04080* (2021).

The Detailed Approximation (For Elliptic Case)

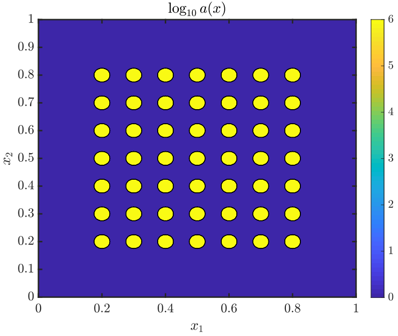
1. Interpolation: $u^{\mathsf{h}} - I_H u^{\mathsf{h}}$ vanishes on edge nodes
   where: I_H: piecewise linear interpolation on the edge (MsFEM)
   Put those interpolation functions into basis functions
2. Oversampling: $e \subset \omega_e$, then on e,

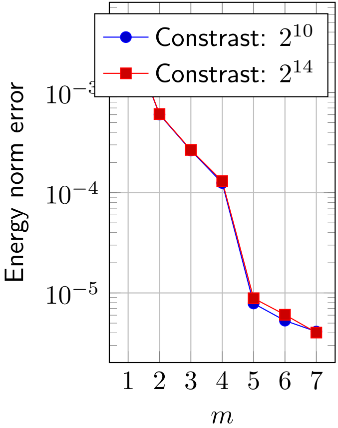
$$(u^{\mathsf{h}} - I_H u^{\mathsf{h}})|_e = (u - I_H u)|_e = \underbrace{(u^{\mathsf{h}}_{\omega_e} - I_H u^{\mathsf{h}}_{\omega_e})|_e}_{a\text{-harmonic function in } \omega_e} + \underbrace{(u^{\mathsf{b}}_{\omega_e} - I_H u^{\mathsf{b}}_{\omega_e})|_e}_{\text{locally computable}}$$

   where, $u^{\mathsf{h}}_{\omega_e}$ is the a-harmonic part of u decomposed in domain ω_e
3. There exists basis functions v_e^j on each e which solve local spectral problems such that
   $(u^{\mathsf{h}}_{\omega_e} - I_H u^{\mathsf{h}}_{\omega_e})|_e = \sum_{j=1}^{m-1} c_j v_e^j + O\left(\exp\left(-m^{\frac{1}{d+1}-\epsilon}\right)\|u^{\mathsf{h}}\|_{H_a^1(\omega_e)}\right)$
   where the approximation is in the $\mathcal{H}^{1/2}(e)$ norm: the $H_a^1(\Omega)$ norm of the a-harmonic extension of function on e
   Key: the restriction of a-harmonic functions is of low complexity

Numerical Examples

The coefficient a has high contrast, $H = 1/32$.





Connection to Multiscale Methods in the Literature

Compared to Generalized FEM, MsFEM, GMsFEM ...

- Our method uses a noval edge coupling[17]
- Nearly exponential convergence results for rough elliptic equations were achieved via partition of unity (PUM)[18]
- Orthogonality of u^{h} and u^{b} preserved
- Noval results for Helmholtz equation

Compared to Variational Multiscale Methods, LOD, Gamblets ...

- We use coarse-fine decomposition as well
- Exponential convergence is achieved

---

[17] Thomas Y Hou and Pengfei Liu. "Optimal Local Multi-scale Basis Functions for Linear Elliptic Equations with Rough Coefficient". In: *Discrete and Continuous Dynamical Systems* 36.8 (2016), pp. 4451–4476.

[18] Ivo Babuska and Robert Lipton. "Optimal local approximation spaces for generalized finite element methods with application to multiscale problems". In: *Multiscale Modeling & Simulation* 9.1 (2011), pp. 373–406.

Roadmap

Summary

Solving computational PDEs from an inference perspective

Gaussian processes for nonlinear PDEs

- Generalize collocation methods and BIM
- Automatic and unified framework for solving and learning PDEs
- Near linear complexity sparse Cholesky factorization
- Kernel learning (theory for linear problems)
- Weak form data, Galerkin methods and subsampled measurements

Multiscale methods for rough elliptic and Helmholtz equations

- Coarse-fine scale decomposition
- Edge coupling extending MsFEM
- Coarse scale solution is of low complexity: exponential convergence

References

- Chen, Yifan, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. "Solving and learning nonlinear pdes with gaussian processes". In: *Journal of Computational Physics* (2021).
- Cockayne, Jon, Chris J Oates, Timothy John Sullivan, and Mark Girolami. "Bayesian probabilistic numerical methods". In: *SIAM Review* 61.4 (2019), pp. 756–789.
- Schäfer, F, 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.
- Stein, Michael L. "The screening effect in kriging". In: *Annals of statistics* 30.1 (2002), pp. 298–323.
- Chen, Yifan, Florian Schaefer, and Houman Owhadi. "Sparse Cholesky Factorization for Solving Nonlinear PDEs via Gaussian Processes". In preparation.
- Schäfer, Florian, Matthias Katzfuss, and Houman Owhadi. "Sparse Cholesky Factorization by Kullback–Leibler Minimization". In: *SIAM Journal on Scientific Computing* 43.3 (2021), A2019–A2046.
- Owhadi, Houman and Gene Ryan Yoo. "Kernel flows: From learning kernels from data into the abyss". In: *Journal of Computational Physics* 389 (2019), pp. 22–47.
- Chen, Yifan, Houman Owhadi, and Andrew Stuart. "Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation". In: *Mathematics of Computation* (2021).
- Målqvist, Axel and Daniel Peterseim. "Localization of elliptic multiscale problems". In: *Mathematics of Computation* 83.290 (2014), pp. 2583–2603.
- Owhadi, Houman. "Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games". In: *SIAM Review* 59.1 (2017), pp. 99–149.
- Chen, Yifan and Thomas Y Hou. "Function approximation via the subsampled Poincaré inequality". In: *Discrete & Continuous Dynamical Systems* 41.1 (2021), p. 169.
- – ."Multiscale elliptic PDE upscaling and function approximation via subsampled data". In: *Multiscale Modeling & Simulation* 20.1 (2022), pp. 188–219.
- Chen, Yifan, Thomas Y Hou, and Yixuan Wang. "Exponential convergence for multiscale linear elliptic PDEs via adaptive edge basis functions". In: *Multiscale Modeling & Simulation* 19.2 (2021), pp. 980–1010.
- – ."Exponentially convergent multiscale methods for high frequency heterogeneous Helmholtz equations". In: *arXiv preprint arXiv:2105.04080* (2021).
- Hou, Thomas Y and Pengfei Liu. "Optimal Local Multi-scale Basis Functions for Linear Elliptic Equations with Rough Coefficient". In: *Discrete and Continuous Dynamical Systems* 36.8 (2016), pp. 4451–4476.
- Babuska, Ivo and Robert Lipton. "Optimal local approximation spaces for generalized finite element methods with application to multiscale problems". In: *Multiscale Modeling & Simulation* 9.1 (2011), pp. 373–406.

Backup Slides

Numerical Experiments: Time Dependent Problems

Viscous Burgers' Equation

- Viscosity $\nu = 0.02$

$$\begin{cases} \partial_t u + u\partial_s u - \nu\partial_s^2 u = 0, \qquad \forall (s,t) \in (-1,1)\times(0,1]. \\ u(s,0) = -\sin(\pi s), \\ u(-1,t) = u(1,t) = 0. \end{cases}$$

- Shock when $\nu = 0$. Problem harder for smaller ν
- Choose an anisotropic spatio-temperal GP

Numerical Experiments: Viscous Burgers' Equation

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

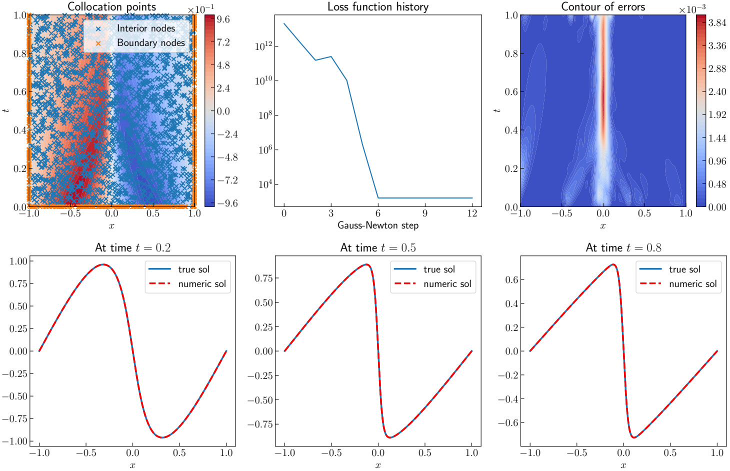


Figure: $N_{\text{domain}} = 2000, N_{\text{boundary}} = 400$

Push to Small Viscosity

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

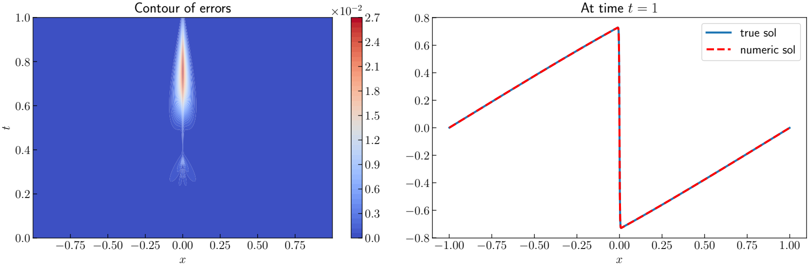


Figure: $\nu = 10^{-3}$; number of spatial points 2000; time step size 0.01; Matern7/2 kernel with lengthscale 0.02; use 2 GN iterations

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

- Observation: accuracy not monotone regarding time t
- Implication: further improvement through time-adaptive kernels

Numerical Experiments: Inverse Problems

Darcy Flow inverse problems

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

- 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

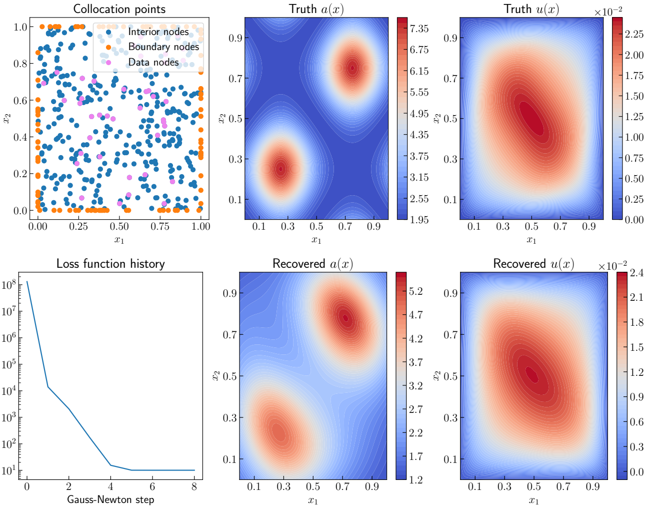


Figure: $N_{\text{domain}} = 400, N_{\text{boundary}} = 100, N_{\text{observation}} = 50$

Consistency?

Question: How do θ^{EB} and θ^{KF} behave, as $\#$ of data $\to \infty$?

- We answer the question for some specific model of $u^\dagger, \theta$ and $\mathcal{X}$

Theory: set-up and theorem

A specific Matérn-like regularity model:

- Domain: $D = \mathbb{T}^d = [0,1]^d_{\text{per}}$
- Lattice data $\mathcal{X}_q = \{j \cdot 2^{-q}, j \in J_q\}$
  where $J_q = \{0, 1, ..., 2^q - 1\}^d$, # of data: 2^{qd}
- Kernel $K_\theta = (-\Delta)^{-t}$, and $\theta = t$
- Subsampling operator in KF: $\pi \mathcal{X}_q = \mathcal{X}_{q-1}$

Theorem (Y. Chen, H. Owhadi, A.M. Stuart, 2020)

Informal: if $u^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$ for some s, then as $q \to \infty$,

$$\theta^{\text{EB}} \to s \quad \text{and} \quad \theta^{\text{KF}} \to \frac{s - d/2}{2} \quad \text{in probability}$$

- Equivalently, $u^\dagger$ is the solution to $(-\Delta)^{s/2} u^\dagger = f$ for white noise f
  Thus, can learn the *fractional physical laws* underlying the data
- Analysis based on multiresolution decomposition and uniform convergence of random series

Theory: set-up and theorem

A specific Matérn-like regularity model:

- Domain: $D = \mathbb{T}^d = [0,1]^d_{\text{per}}$
- Lattice data $\mathcal{X}_q = \{j \cdot 2^{-q}, j \in J_q\}$
  where $J_q = \{0, 1, ..., 2^q - 1\}^d$, # of data: 2^{qd}
- Kernel $K_\theta = (-\Delta)^{-t}$, and $\theta = t$
- Subsampling operator in KF: $\pi \mathcal{X}_q = \mathcal{X}_{q-1}$

Theorem (Y. Chen, H. Owhadi, A.M. Stuart, 2020)

Informal: if $u^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$ for some s, then as $q \to \infty$,

$$\theta^{\mathsf{EB}} \to s \quad \text{and} \quad \theta^{\mathsf{KF}} \to \frac{s - d/2}{2} \quad \text{in probability}$$

- Equivalently, $u^\dagger$ is the solution to $(-\Delta)^{s/2} u^\dagger = f$ for white noise f
  Thus, can learn the *fractional physical laws* underlying the data
- Analysis based on multiresolution decomposition and uniform convergence of random series

Experiments justifying the theory

How it works in practice?

- $d = 1, s = 2.5$, # of data $N = 2^9$, mesh size 2^{-10}

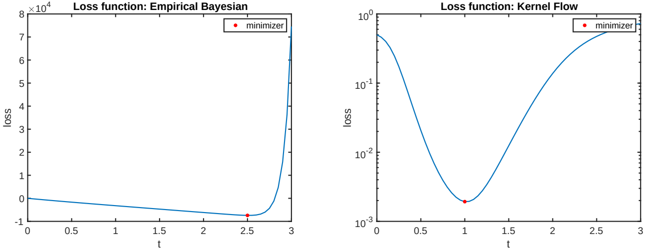


Figure: Left: EB loss; right: KF loss

- Patterns in the loss function (our theory can predict!)
  - EB: first linear, then blow up quickly
  - KF: more symmetric

Experiments justifying the theory

How it works in practice?

- $d = 1, s = 2.5$, # of data $N = 2^9$, mesh size 2^{-10}

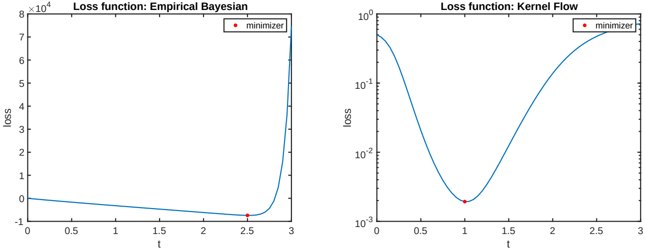


Figure: Left: EB loss; right: KF loss

- Patterns in the loss function (our theory can predict!)
  - EB: first linear, then blow up quickly
  - KF: more symmetric

Selection Bias

Next Question: How are the limits $s\ (= 2.5)$ and $\frac{s-d/2}{2}\ (= 1)$ special?

- What is the *implicit bias* of EB and KF algorithms?
- Our strategy: look at their L^2 population errors

Experiment I

- # of data: 2^q; compute $\mathbb{E}_{u^\dagger}\|u^\dagger(\cdot) - u(\cdot, t, \mathcal{X}_q)\|_{L^2}^2$

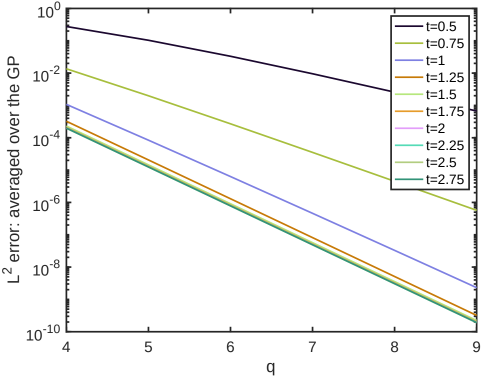


Figure: L^2 error: averaged over the GP

- $\frac{s-d/2}{2}$ $(= 1)$ is the minimal t that suffices for the fastest rate of L^2 error

Experiment II

- # of data: $2^q, q = 9$; compute $\mathbb{E}_{u^\dagger} \|u^\dagger(\cdot) - u(\cdot, t, \mathcal{X}_q)\|_{L^2}^2$

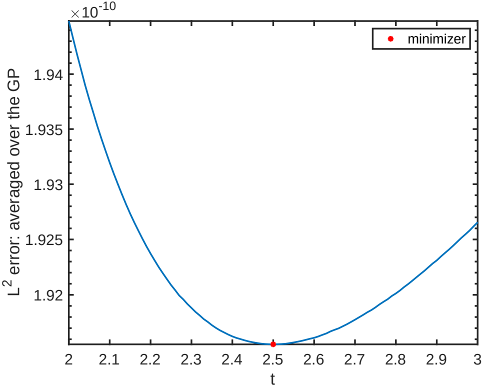


Figure: L^2 error: averaged over the GP, for $q = 9$

- $s\ (= 2.5)$ is the t that achieves the minimal L^2 error in expectation

Take-aways

- For Matérn-like kernel model, EB and KF have different selection bias
  - EB selects the θ that achieves the minimal L^2 error in expectation
  - KF selects the minimal θ that suffices for the fastest rate of L^2 error
- More comparisons between EB and KF in our paper
  - Estimate amplitude and lengthscale in $\mathcal{N}(0, \sigma^2(-\Delta + \tau^2 I)^{-s})$
  - Variance of estimators
  - Robustness to model misspecification (important!)
  - Computational cost

Parameter learning: via Bayes or approximation-theoretic?

Localization of ψ_i

Representation of ψ_i (Lagrangian dual)

$$\begin{aligned}\psi_i = \mathsf{argmin}_{\psi \in H_0^1(\Omega)} \quad & \|\psi\|^2_{H_a^1(\Omega)} \\ \text{s.t.} \quad & \langle \psi, \phi_j \rangle = \delta_{i,j} \ \text{ for } 1 \le j \le N\,. \end{aligned}$$

Local spectral approximation

- The $\mathcal{H}^{1/2}(e)$ norm:

$$\|\tilde{\psi}\|^2_{\mathcal{H}^{1/2}(e)} := \int_\Omega a|\nabla\psi|^2$$

where ψ is the a-harmonic extension of $\tilde{\psi}$ on e

- $R_e : (V_{\omega_e}, \|\cdot\|_{H^1_a(\Omega)}) \to (\mathcal{H}^{1/2}(e), \|\cdot\|_{\mathcal{H}^{1/2}(e)})$ such that $R_e v = (v - I_H v)|_e$ where, V_{ω_e} is the space of a-harmonic functions in ω_e

For any a-harmonic functions v in ω_e and any $\epsilon > 0$, there exists an $N_\epsilon > 0$, such that for all $m > N_\epsilon$, we can find an $(m-1)$ dimensional space $W_e^m = \mathrm{span}\ \{\tilde{v}_e^k\}_{k=1}^{m-1}$ so that

$$\min_{\tilde{v}_e \in W_e^m} \|R_e v - \tilde{v}_e\|_{\mathcal{H}^{1/2}(e)} \leq C \exp\left(-m^{(\frac{1}{d+1}-\epsilon)}\right) \|v\|_{H^1_a(\omega_e)}$$

- Proof technique combines [Babuska, Lipton 2011] and C^α estimates