Consistency of Hierarchical Parameter Learning

Empirical Bayes and Kernel Flow Approaches

Yifan Chen (Caltech)

Joint work with
Andrew M. Stuart and Houman Owhadi, Caltech

September 20, 2020

One page's overview

- **Context**: Supervised learning / nonparametric regression
- **Approach**: Gaussian process regression / kernel methods
- **Question of focus**: How to select kernels based on data
  - Hierarchical parameters in the kernels
- **Algorithms in use:**
  - Bayesian: Empirical Bayes
  - Approximation theoretic: Kernel Flow
- **Contribution**:
  - Theory: Consistency and selection bias for a Matérn class model
  - Experiments: beyond Matérn model, and include model misspecification

Gaussian process regression (GPR)

- Supervised learning / nonparametric regression / interpolation

Recover $u^\dagger : D \subset \mathbb{R}^d \to \mathbb{R}$ from

$$y_i = u^\dagger(x_i), 1 \leq i \leq N \qquad \text{(Noise-free data)}$$

- GPR solution / Kernel method:

$$u(\cdot, \theta, \mathcal{X}) = K_\theta(\cdot, \mathcal{X})[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X})$$

(Depend on kernel K_θ, data set $\mathcal{X}$, and truth $u^\dagger$)

Notation: ($\theta \in \Theta$ is a *hierarchical parameter*)

$$K_\theta : D \times D \to \mathbb{R}$$

$$\mathcal{X} = \{x_1, ..., x_N\}, \text{ and } u^\dagger(\mathcal{X}) \in \mathbb{R}^N, K_\theta(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{N \times N}$$

$$K_\theta(\cdot, \mathcal{X}) : D \to \mathbb{R}^N, \text{ and } u(\cdot, \theta, \mathcal{X}) : D \to \mathbb{R}$$

Gaussian process regression (GPR)

- Supervised learning / nonparameteric regression / interpolation

Recover $u^\dagger : D \subset \mathbb{R}^d \to \mathbb{R}$ from

$$y_i = u^\dagger(x_i), 1 \leq i \leq N \qquad \text{(Noise-free data)}$$

- GPR solution / Kernel method:

$$u(\cdot, \theta, \mathcal{X}) = K_\theta(\cdot, \mathcal{X})[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X})$$

(Depend on kernel K_θ, data set $\mathcal{X}$, and truth $u^\dagger$)

Notation: ($\theta \in \Theta$ is a *hierarchical parameter*)

$$K_\theta : D \times D \to \mathbb{R}$$

$$\mathcal{X} = \{x_1, ..., x_N\}, \text{ and } u^\dagger(\mathcal{X}) \in \mathbb{R}^N, K_\theta(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{N\times N}$$

$$K_\theta(\cdot, \mathcal{X}) : D \to \mathbb{R}^N, \text{ and } u(\cdot, \theta, \mathcal{X}) : D \to \mathbb{R}$$

What's the problem?

- Any $\theta \in \Theta$, gets an interpolated solution on $\mathcal{X}$:

$$u^\dagger(x_i) = u(x_i, \theta, \mathcal{X}), 1 \leq i \leq N$$

Zero training error is not hard to get

But, for out-of-sample / generalization errors, how to pick a good θ?

- A model selection problem – learn the hierarchical parameter θ

Roadmap of this talk

1. Bayes' approach
   - Empirical Bayes estimator
2. Approximation-theoretic approach
   - Kernel Flow estimator
3. Comparison of their consistency as # of data $\to \infty$, and beyond
   - Rigorous theories for the consistency for Matérn class models
   - Experiments beyond Matérn models, and include model misspecification

Roadmap of this talk

1. Bayes' approach
   - Empirical Bayes estimator
2. Approximation-theoretic approach
   - Kernel Flow estimator
3. Comparison of their consistency as # of data $\to \infty$, and beyond
   - Rigorous theories for the consistency for Matérn class models
   - Experiments beyond Matérn models, and include model misspecification

Bayes' solution

- Put a prior on θ, and $u^\dagger|\theta \sim \mathcal{N}(0, K_\theta)$ — then calculate the posterior

- Empirical Bayes (EB) with uninformative prior:

$$\theta^{\mathrm{EB}}(\mathcal{X}, u^\dagger) = \operatorname*{argmin}_{\theta\in\Theta} \mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger)$$

$$\mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger) = u^\dagger(\mathcal{X})^\mathsf{T}[K_\theta(\mathcal{X}, \mathcal{X})]^{-1}u^\dagger(\mathcal{X}) + \log\det K_\theta(\mathcal{X}, \mathcal{X})$$

Maximum Likelihood Estimate!

- The EB solution: just pick $\theta^{\mathrm{EB}}(\mathcal{X}, u^\dagger)$
  - depend on data set $\mathcal{X}$, truth $u^\dagger$ (and the prior)

Bayes' solution

- Put a prior on θ, and $u^\dagger|\theta \sim \mathcal{N}(0, K_\theta)$ — then calculate the posterior
- Empirical Bayes (EB) with uninformative prior:

$$\theta^{\mathrm{EB}}(\mathcal{X}, u^\dagger) = \operatorname*{argmin}_{\theta\in\Theta} \mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger)$$

$$\mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger) = u^\dagger(\mathcal{X})^{\mathsf{T}}[K_\theta(\mathcal{X}, \mathcal{X})]^{-1}u^\dagger(\mathcal{X}) + \log\det K_\theta(\mathcal{X}, \mathcal{X})$$

Maximum Likelihood Estimate!

- The EB solution: just pick $\theta^{\mathrm{EB}}(\mathcal{X}, u^\dagger)$
  - depend on data set $\mathcal{X}$, truth $u^\dagger$ (and the prior)

Roadmap of this talk

1. Bayes' approach
   - Empirical Bayes estimator
2. Approximation-theoretic approach
   - Kernel Flow estimator
3. Comparison of their consistency as # of data $\to \infty$, and beyond
   - Rigorous theories for the consistency for Matérn class models
   - Experiments beyond Matérn models, and include model misspecification

Approximation-theoretic approach

- Why $\theta, u^\dagger$ have a prior distribution? — may be brittle to misspecification
- Go straightforward: set a target cost d, and optimize$_\theta$ $\mathsf{d}(u^\dagger, u(\cdot, \theta, \mathcal{X}))$
- Problem: $u^\dagger$ not available — solution: approximation

$$\min_\theta \mathsf{d}(u(\cdot, \theta, \mathcal{X}), u(\cdot, \theta, \pi\mathcal{X})) \qquad \text{(One example)}$$

π: subsampling operator (similar to cross-validation)

Approximation-theoretic approach

- Why $\theta, u^\dagger$ have a prior distribution? — may be brittle to misspecification
- Go straightforward: set a target cost d, and optimize$_\theta$ $\mathsf{d}(u^\dagger, u(\cdot, \theta, \mathcal{X}))$
- Problem: $u^\dagger$ not available — solution: approximation

$$\min_\theta \mathsf{d}(u(\cdot, \theta, \mathcal{X}), u(\cdot, \theta, \pi\mathcal{X})) \qquad \text{(One example)}$$

π: subsampling operator (similar to cross-validation)

Approximation-theoretic approach

- Why $\theta, u^\dagger$ have a prior distribution? — may be brittle to misspecification
- Go straightforward: set a target cost d, and optimize$_\theta$ $\mathsf{d}(u^\dagger, u(\cdot, \theta, \mathcal{X}))$
- Problem: $u^\dagger$ not available — solution: approximation

$$\min_\theta \mathsf{d}(u(\cdot, \theta, \mathcal{X}), u(\cdot, \theta, \pi\mathcal{X})) \qquad \text{(One example)}$$

π: subsampling operator (similar to cross-validation)

Kernel Flow

A specific choice of d: [Owhadi, Yoo 2018 & 2020], [Hamzi, Owhadi 2020]

$$\theta^{\mathrm{KF}}(\mathcal{X}, \pi\mathcal{X}, u^\dagger) = \operatorname*{argmin}_{\theta\in\Theta} \mathsf{L}^{\mathrm{KF}}(\theta, \mathcal{X}, \pi\mathcal{X}, u^\dagger)$$

$$\mathsf{L}^{\mathrm{KF}}(\theta, \mathcal{X}, \pi\mathcal{X}, u^\dagger) = \frac{\|u(\cdot, \theta, \mathcal{X}) - u(\cdot, \theta, \pi\mathcal{X})\|_{K_\theta}^2}{\|u(\cdot, \theta, \mathcal{X})\|_{K_\theta}^2}$$

where

- π: a subsampling operator, so $\pi\mathcal{X} \subset \mathcal{X}$
- $\|\cdot\|_{K_\theta}$: RKHS norm determined by K_θ

A kernel is good, if subsampling data does not influence solution much

Roadmap of this talk

1. Bayes' approach
   - Empirical Bayes estimator
2. Approximation-theoretic approach
   - Kernel Flow estimator
3. Comparison of their consistency as # of data $\to \infty$, and beyond
   - Rigorous theories for the consistency for Matérn class models
   - Experiments beyond Matérn models, and include model misspecification

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 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 (Chen, Owhadi, Stuart, 2020)

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

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

- Analysis based on multiresolution decomposition and uniform convergence of random series

Theory: Set-up and theorem

A specific Matérn 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 (Chen, Owhadi, Stuart, 2020)

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

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

- Analysis based on multiresolution decomposition and uniform convergence of random series

Experiments

How it works in practice?

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

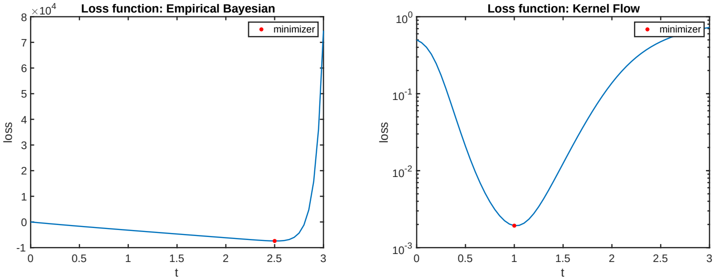

Figure: Left: EB loss; right: KF loss

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

Experiments

How it works in practice?

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

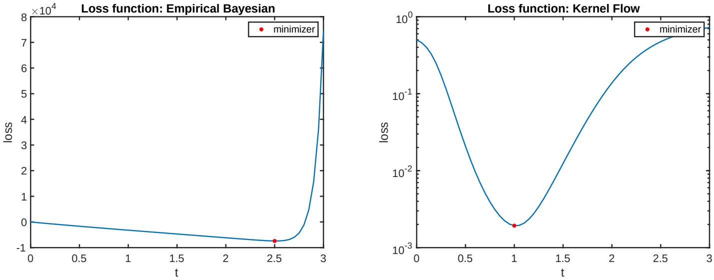

Figure: Left: EB loss; right: KF loss

- 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 1

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

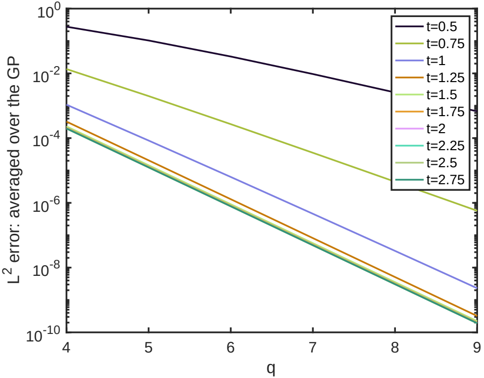

Figure: L^2 error: averaged over the GP

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

Experiment 2

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

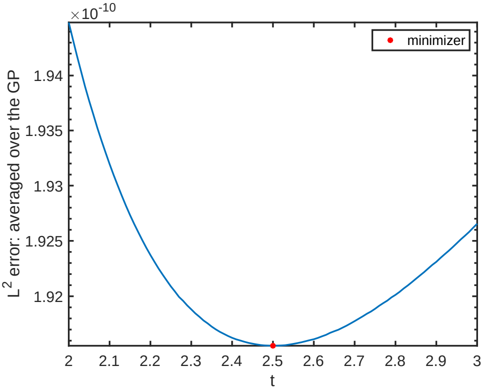

Figure: L^2 error: averaged over the GP, for $q = 9$

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

Summary of Our Theory

For Matérn-like model, EB and KF have different selection bias

- EB selects the t that achieves the minimal L^2 error in expectation
- KF selects the minimal t that suffices for the fastest rate of L^2 error

Beyond Matérn class model?

Summary of Our Theory

For Matérn-like model, EB and KF have different selection bias

- EB selects the t that achieves the minimal L^2 error in expectation
- KF selects the minimal t that suffices for the fastest rate of L^2 error

Beyond Matérn class model?

Roadmap of this talk

1. Bayes' approach
   - Empirical Bayes estimator
2. Approximation-theoretic approach
   - Kernel Flow estimator
3. Comparison of their consistency as # of data $\to \infty$, and beyond
   - Rigorous theories for the consistency for Matérn class models
   - Experiments beyond Matérn models, and include model misspecification

Recovery of other parameters in Matérn-like model

- Matérn-like model: $u^\dagger \sim \mathcal{N}(0, \sigma^2(-\Delta + \tau^2 I)^{-s})$
  - σ: amplitude; τ: lengthscale; s: regularity

- Experiments: $D = \mathbb{T}^d, d = 1$
  - EB can recover s and σ (respectively & simultaneously), not τ
  - KF can only recover $\frac{s-d/2}{2}$, not σ and τ

Variance of regularity estimation

- Earlier model: $u^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$, $s = 2.5, d = 1$
- Variance (# of data 2^9):

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 7.8 \times 10^{-5} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s-d/2)/2)^2} \approx 4 \times 10^{-3}$$

For well-specification model: variance of EB better than KF

Other well-specified models: 1st

- Model: $u^\dagger \sim \mathcal{N}(0, (-\nabla \cdot (a\nabla \cdot))^{-s})$ on one-dim torus
  $K_\theta = (-\nabla \cdot (a\nabla \cdot))^{-\theta})$, $\mathcal{X}$ uniform lattice (# of data: 2^9)

$$a(x) = \begin{cases} 1 & x \in [0, 1/2] \\ 2 & x \in (1/2, 1] \end{cases}$$

- Variance:

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 7.8 \times 10^{-5} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s - d/2)/2)^2} \approx 4 \times 10^{-3}$$

Other well-specified models: 2nd

- Model: $-\nabla \cdot (a_{1/2} \nabla u^{\dagger}) = \xi \sim \mathcal{N}(0, (-\Delta)^{-1})$

$$a_{\theta}(x) = \begin{cases} 1 & x \in [0, \theta] \\ 2 & x \in (\theta, 1] \end{cases}.$$

  $K_{\theta} = (-\nabla \cdot (a_{\theta} \nabla \cdot))^{-1} (-\Delta)^{-s} (-\nabla \cdot (a_{\theta} \nabla \cdot))^{-1}$ for $s = 1$
  $\mathcal{X}$ uniform lattice (# of data: 2^9)

- Experimental Result: Both EB and KF recover $\theta = 1/2$

Model Misspecification: 1st

- Model: $u^\dagger \sim \mathcal{N}(0, (-\nabla \cdot (a\nabla \cdot))^{-s})$

$$a(x) = \begin{cases} 1 & x \in [0, 1/2] \\ 2 & x \in (1/2, 1] \end{cases}$$

  $K_\theta = (-\Delta)^{-\theta}$, $\mathcal{X}$ uniform lattice (# of data: 2^9)

- Variance:

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 5.9 \times 10^{-4} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s-d/2)/2)^2} \approx 6.8 \times 10^{-4}$$

Model Misspecification: 2nd

- Model: $-\nabla \cdot (a_{1/2} \nabla u^\dagger) = \xi \sim \mathcal{N}(0, (-\Delta)^{-1})$

$$a_\theta(x) = \begin{cases} 1 & x \in [0, \theta] \\ 2 & x \in (\theta, 1] \end{cases}.$$

  $K_\theta = (-\nabla \cdot (a_\theta \nabla \cdot))^{-1} (-\Delta)^{-s} (-\nabla \cdot (a_\theta \nabla \cdot))^{-1}$ for $s = 5$
  $\mathcal{X}$ uniform lattice (# of data: 2^9)

- Experimental Result: KF recovers $\theta = 1/2$, EB fails

Model Misspecification: 3nd

- Model: $(-\Delta)^s u^\dagger(\cdot) = \delta(\cdot - 1/2)$ deterministic
  $K_\theta = (-\Delta)^{-\theta}$, $\mathcal{X}$ uniform lattice (# of data: 2^9)

- Experimental Result: EB recovers $2s$, while KF recovers s

Takeaway messages

- For Matérn-like kernel model, EB and KF have different selection bias
  - EB selects the t that achieves the minimal L^2 error in expectation
  - KF selects the minimal t that suffices for the fastest rate of L^2 error
- Comparisons between EB and KF
  - 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

Hierarchical parameter learning: via Bayes or approximation-theoretic?

Thank you!