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 5, 2020

One page's overview

- **Context**: Supervised learning
- **Approach**: Gaussian process regression / kernel methods
- **Question of focus**: How to select kernels based on data
- **Algorithms in use**: Empirical Bayes and Kernel Flow
- **Achieved**: Consistency and selection bias for a Matérn model

Gaussian process regression (GPR)

- Supervised learning: recover $u^\dagger : D \subset \mathbb{R}^d \to \mathbb{R}$ from

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

- GPR solution:

$$\begin{aligned} u(\cdot, \theta, \mathcal{X}) &= \mathbb{E}\left[\xi(\cdot, \theta) \mid \xi(\mathcal{X}, \theta) = u^\dagger(\mathcal{X})\right] \\ &= K_\theta(\cdot, \mathcal{X})[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X}) \end{aligned}$$

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

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

$$\mathcal{GP} : \xi(\cdot, \theta) \sim \mathcal{N}(0, K_\theta), \text{ where } K_\theta : D \times D \to \mathbb{R}$$

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

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

Gaussian process regression (GPR)

- Supervised learning: recover $u^\dagger : D \subset \mathbb{R}^d \to \mathbb{R}$ from

$$y_i = u^\dagger(x_i), 1 \le i \le N \qquad \text{(Noiseless data)}$$

- GPR solution:

$$\begin{aligned} u(\cdot, \theta, \mathcal{X}) &= \mathbb{E}\left[\xi(\cdot, \theta) \mid \xi(\mathcal{X}, \theta) = u^\dagger(\mathcal{X})\right] \\ &= K_\theta(\cdot, \mathcal{X})[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X}) \end{aligned}$$

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

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

$$\mathcal{GP} : \xi(\cdot, \theta) \sim \mathcal{N}(0, K_\theta), \text{ where } K_\theta : D \times D \to \mathbb{R}$$

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

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

What's the problem?

- Any $\theta \in \Theta$, gets an interpolated solution on $\mathcal{X}$ (zero training loss)

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

- We need to do model selection — learn a good hierarchical parameter

Roadmap of this talk

1. Empirical Bayes' approach
2. Approximation-theoretic approach
3. Comparison of their consistency as # of data $\to \infty$, and beyond

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)

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.

Consistency

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 $\mathcal{X}$

Set-up and theorem

- 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 $q \to \infty$,

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

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

Set-up and theorem

- 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

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

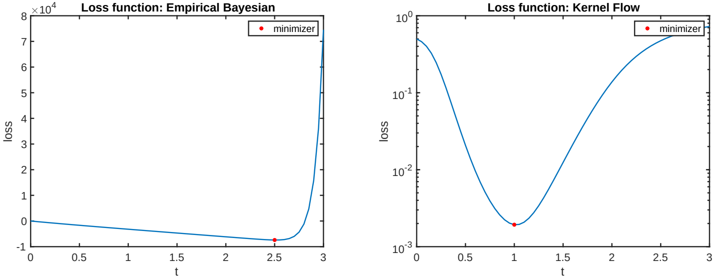

Figure: Left: EB loss; right: KF loss

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

Experiments

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

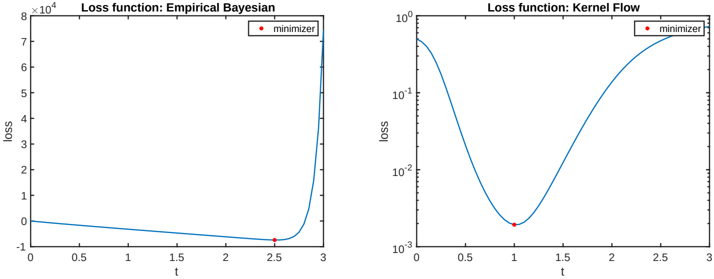

Figure: Left: EB loss; right: KF loss

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

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?
- We will look at their L^2 population errors

Experiment 1

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

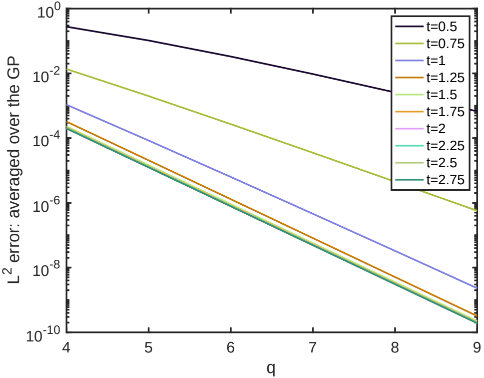

Figure: L^2 error: averaged over the GP

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

Experiment 2

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

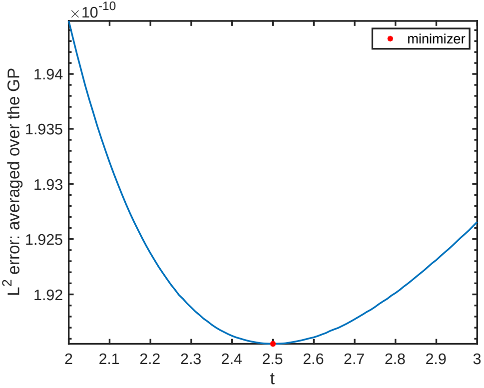

Figure: L^2 error: averaged over the GP, for $q = 9$

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

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

- 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

Hierarchical parameter learning: via Bayes or approximation-theoretic?

Thank you!