LEARNING GPs WITH BAYESIAN POSTERIOR OPTIMIZATION

Luiz F. O. Chamon, Santiago Paternain, and Alejandro Ribeiro

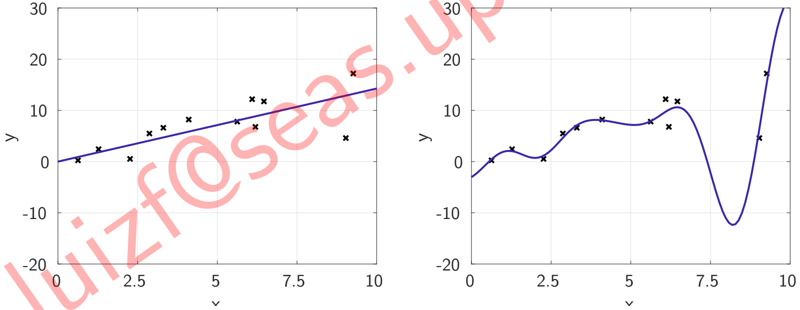
luizf@seas.upenn.edu

ASILOMAR 2019
November 4th, 2019

Dealing with complexity and uncertainty

luizf@seas.upenn.edu

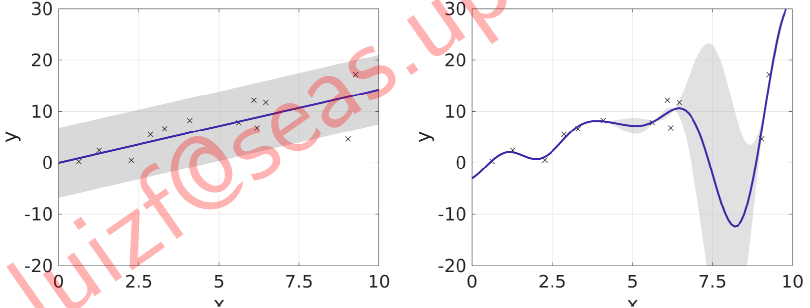
- Nonparametric methods



luizf@seas.upenn.edu

Dealing with complexity and uncertainty

- Nonparametric methods
- Bayesian methods



luizf@seas.upenn.edu

Nonparametric Bayesian methods

- All of Bayesian inference

$$\underbrace{\mathbb{P}\left(\text{model}\right)}_{\text{Prior}} + \underbrace{\mathbb{P}\left(\text{data} \mid \text{model}\right)}_{\text{Likelihood}} \to \underbrace{\mathbb{P}\left(\text{model} \mid \text{data}\right)}_{\text{Posterior}}$$

luizf@seas.upenn.edu

Nonparametric Bayesian methods

- All of Bayesian inference

$$\underbrace{\mathbb{P}\left(\text{model}\right)}_{\text{Prior}} + \underbrace{\mathbb{P}\left(\text{data} \mid \text{model}\right)}_{\text{Likelihood}} \to \underbrace{\mathbb{P}\left(\text{model} \mid \text{data}\right)}_{\text{Posterior}}$$

- Parametric models are finite dimensional

$$\mathbb{P}\left(\text{model}\right) = \mathbb{P}\left(\text{parameters}\right)$$

- Nonparametric models are infinite dimensional

luizf@seas.upenn.edu

Gaussian processes

- GPs are priors on "smooth" functions
  - ✓ easy to specify: choose a covariance function (and hyperparameters)
  - ✓ flexible: wide variety of covariance functions (degree of smoothness, periodicity...)
  - ✓ tractable

luizf@seas.upenn.edu

Gaussian processes

- GPs are priors on “smooth” functions
  - ✓ easy to specify: choose a covariance function (and hyperparameters)
  - ✓ flexible: wide variety of covariance functions (degree of smoothness, periodicity...)
  - ✓ tractable

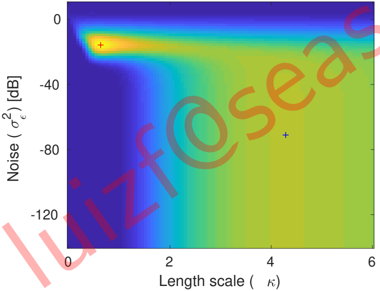
- Still... which GP?
  - × limited access to prior knowledge
  - × hard to interpret hyperparameters
  - × misspecifying GPs can be catastrophic [Bachoc'13, Beckers et al.'18, Zaytsev et al.'18]

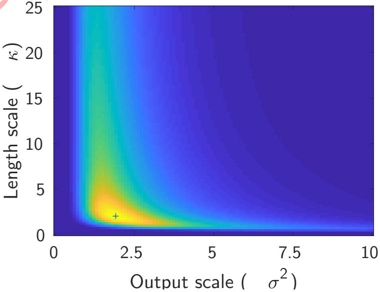
luizf@seas.upenn.edu

Which GP?

- Maximize likelihood w.r.t. hyperparameters [Stein'99, RW'06]
  - Ambiguity: multimodal likelihood, local maxima
  - Indeterminacy: different parameters, same measure

$$\theta^\star = \underset{\theta}{\operatorname{argmax}} \ \log \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$





luizf@seas.upenn.edu

Which GP?

- Maximize likelihood w.r.t. hyperparameters [Stein'99, RW'06]
  - Ambiguity: multimodal likelihood, local maxima
  - Indeterminacy: different parameters, same measure

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \log \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$

- Hierarchical models [RG'02, RW'06, Gelman et al.'13]
  - Noninformative priors $\to$ improper posteriors
  - Hard to interpret, hard to set priors
  - Indeterminacy: setting one prior affects the others

$$\boldsymbol{\theta} \sim \mathcal{P}$$

luizf@seas.upenn.edu

- Hybrid Bayesian–Optimization approach
  - **Bayesian**:
  - **Optimization**:

luizf@seas.upenn.edu

Bayesian posterior optimization

- Hybrid Bayesian–Optimization approach
  - **Bayesian**: obtain distribution over hyperparameters instead of point estimate
  - **Optimization**:

luizf@seas.upenn.edu

Bayesian posterior optimization

- Hybrid Bayesian–Optimization approach
  - **Bayesian**: obtain distribution over hyperparameters instead of point estimate
  - **Optimization**: minimize a risk measure instead of using Bayes rule

luizf@seas.upenn.edu

Bayesian posterior optimization

- Hybrid Bayesian–Optimization approach
  - **Bayesian**: obtain distribution over hyperparameters instead of point estimate
  - **Optimization**: minimize a risk measure instead of using Bayes rule

✓ Non-convex risk measures (0-1 loss, truncated MSE…)

✓ Incorporate complex structures in the prior: maximum entropy, sparsity, moments…

luizf@seas.upenn.edu

Bayesian posterior optimization

- Hybrid Bayesian–Optimization approach
  - **Bayesian**: obtain distribution over hyperparameters instead of point estimate
  - **Optimization**: minimize a risk measure instead of using Bayes rule

✓ Non-convex risk measures (0-1 loss, truncated MSE...)

✓ Incorporate complex structures in the prior: maximum entropy, sparsity, moments...

× Non-convex, infinite dimensional optimization problem

luizf@seas.upenn.edu

Bayesian posterior optimization

- Hybrid Bayesian–Optimization approach
  - **Bayesian**: obtain distribution over hyperparameters instead of point estimate
  - **Optimization**: minimize a risk measure instead of using Bayes rule

✓ Non-convex risk measures (0-1 loss, truncated MSE...)

✓ Incorporate complex structures in the prior: maximum entropy, sparsity, moments...

✓ Non-convex, infinite dimensional optimization problem
$\Rightarrow$ *simple, efficient solution using duality*

luizf@seas.upenn.edu

Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems

luizf@seas.upenn.edu

GP regression

- **Data:** $(\boldsymbol{x}_i, y_i)$ with $y_i \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma_\epsilon^2)$ for an unknown f
- **Goal:** determine $(f \mid \boldsymbol{X}, \boldsymbol{y})$
- **How?** Bayes' rule and GP prior

luizf@seas.upenn.edu

Gaussian processes

- A GP is a stochastic process whose finite dimensional marginals are jointly Gaussian

- Formally, $\mathbb{GP}(m, k)$ is a distribution over functions g such that $[g(\boldsymbol{x}_1) \ \cdots \ g(\boldsymbol{x}_n)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$ for all $n \in \mathbb{N}$

$$\boldsymbol{m} = \begin{bmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{bmatrix} \text{ and } \boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}$$

- Typically, $m \equiv 0$

luizf@seas.upenn.edu

GP regression

- **Data:** $(\boldsymbol{x}_i, y_i)$ with $y_i \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma_\epsilon^2)$ for an unknown f
- **Goal:** determine $(f \mid \boldsymbol{X}, \boldsymbol{y})$
- **How?** Bayes' rule and GP prior

$$(f(\bar{\boldsymbol{x}}) \mid \boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \bar{\boldsymbol{k}}^T \boldsymbol{K}^{-1} \boldsymbol{y}, \qquad \boldsymbol{\Sigma} = k(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - \bar{\boldsymbol{k}}^T \boldsymbol{K}^{-1} \bar{\boldsymbol{k}}$$

$$\bar{\boldsymbol{k}} = \begin{bmatrix} k(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) \\ \vdots \\ k(\bar{\boldsymbol{x}}, \boldsymbol{x}_n) \end{bmatrix} \text{ and } \boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}$$

luizf@seas.upenn.edu

- First level: unknown function f

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, f\right) \mathbb{P}\left(f \mid \boldsymbol{X}\right)$$

luizf@seas.upenn.edu

- First level: unknown function f

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, f, \boldsymbol{\theta}\right) \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right) d\boldsymbol{\theta}$$

luizf@seas.upenn.edu

- First level: unknown function f

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, f, \boldsymbol{\theta}\right) \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \underbrace{\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right)}_{\boldsymbol{\theta}\text{-posterior}} d\boldsymbol{\theta}$$

luizf@seas.upenn.edu

Back to Ba(ye)sics

- First level: unknown function f

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, f, \boldsymbol{\theta}\right) \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \underbrace{\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right)}_{\boldsymbol{\theta}\text{-posterior}} d\boldsymbol{\theta}$$

- Second level: hyperparameters $\boldsymbol{\theta}$

$$\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}\right) \mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right)$$

luizf@seas.upenn.edu

- First level: unknown function f

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, f, \boldsymbol{\theta}\right) \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{\theta}\right)$$

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \underbrace{\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right)}_{\boldsymbol{\theta}\text{-posterior}} d\boldsymbol{\theta}$$

- Second level: hyperparameters $\boldsymbol{\theta}$

$$\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right) \propto \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}\right) \mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right)$$

- **Issue:** choosing $\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right)$
(interpretation, indeterminacy, informativeness)

luizf@seas.upenn.edu

Back to Ba(ye)sics

- First level: unknown function f

$$\mathbb{P}(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}) \propto \mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, f, \boldsymbol{\theta})\, \mathbb{P}(f \mid \boldsymbol{X}, \boldsymbol{\theta})$$

$$\mathbb{P}(f \mid \boldsymbol{X}, \boldsymbol{y}) = \int \mathbb{P}(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta})\, \underbrace{\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y})}_{\boldsymbol{\theta}\text{-posterior}}\, d\boldsymbol{\theta}$$

- Second level: hyperparameters $\boldsymbol{\theta}$

$$\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}) \propto \cancel{\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{X})}$$

- **Issue:** choosing $\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{X})$
(interpretation, indeterminacy, informativeness)

luizf@seas.upenn.edu

Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems

luizf@seas.upenn.edu

- **Statistical learning:**

$$\phi^\star = \underset{\phi \in \mathcal{F}}{\operatorname{argmin}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell(\phi(\boldsymbol{x}), y)\right] + R(\phi)$$

  - $\mathcal{D}$ is an *unknown* probability distribution over pairs $(\boldsymbol{x}, y)$
  - $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function
  - R is a regularizer
  - $\mathcal{F}$ is a space of functions $\phi : \mathbb{R}^d \to \mathbb{R}$

luizf@seas.upenn.edu

Statistical learning

- **Statistical learning:**

$$\phi^\star = \operatorname*{argmin}_{\phi \in \mathcal{F}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell(\phi(\boldsymbol{x}), y)\right] + R(\phi)$$

- **Empirical risk minimization:**

$$\hat{\phi}^\star = \operatorname*{argmin}_{\phi \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(\phi(\boldsymbol{x}_i), y_i) + R(\phi)$$

  - $\mathcal{D}$ is an *unknown* probability distribution over pairs $(\boldsymbol{x}, y)$
  - $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function
  - R is a regularizer
  - $\mathcal{F}$ is a space of functions $\phi : \mathbb{R}^d \to \mathbb{R}$
  - Data: $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}$

luizf@seas.upenn.edu

Statistical GP learning

- **Statistical GP learning:**

$$\Gamma^\star = \operatorname*{argmin}_{\Gamma \in \mathcal{GP}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D},\, f\sim\Gamma}\left[\ell(f(\boldsymbol{x}), y)\right] + R(\gamma)$$

- **Empirical GP-risk minimization:**

$$\hat{\Gamma}^\star = \operatorname*{argmin}_{\Gamma \in \mathcal{GP}} \; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f\sim\Gamma}\left[\ell(f(\boldsymbol{x}_i), y_i)\right] + R(\gamma)$$

  - $\mathcal{D}$ is an *unknown* probability distribution over pairs $(\boldsymbol{x}, y)$
  - $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function
  - R is a regularizer
  - $\mathcal{GP}$ is the "space of Gaussian processes"
  - Data: $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}$

luizf@seas.upenn.edu

Back to the first level

- **Empirical GP-risk minimization:**

$$\hat{\Gamma}^\star = \underset{\Gamma \in \mathcal{GP}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \Gamma} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] + R(\gamma)$$

- Challenge: optimizing over $\mathcal{GP}$
  (isomorphic to the space of positive semi-definite functions)

luizf@seas.upenn.edu

Back to the first level

- **Empirical GP-risk minimization:**

$$\hat{\Gamma}^\star = \underset{\Gamma \in \mathcal{GP}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \Gamma} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] + R(\gamma)$$

- Challenge: optimizing over $\mathcal{GP}$
  (isomorphic to the space of positive semi-definite functions)

- Leverage the first level of the hierarchical model

$$\mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}\right) = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) \mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right) d\boldsymbol{\theta}$$

- "Parameterize" (a subset of) $\mathcal{GP}$ using $\mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right)$

luizf@seas.upenn.edu

- **Bayesian posterior optimization:**

$$p^\star = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathbb{GP}(0, k_{\boldsymbol{\theta}})} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + R(p)$$

$$\hat{\Gamma}_p^\star = \int \mathbb{P}\left(f \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}\right) p^\star(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

- Optimization variable: $p(\boldsymbol{\theta}) = \mathbb{P}\left(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y}\right)$
- Alternative interpretation: mixture of GPs with weights $p(\boldsymbol{\theta})$

× Non-convex, infinite dimensional optimization problem

luizf@seas.upenn.edu

Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems

luizf@seas.upenn.edu

- **Bayesian posterior optimization:**

$$\underset{p \in \mathcal{P}}{\text{minimize}} \;\; \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathbb{GP}(0, k_{\boldsymbol{\theta}})} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + R(p)$$

luizf@seas.upenn.edu

Assumptions

- **Bayesian posterior optimization:**

$$\operatorname*{minimize}_{p \in \mathcal{P}} \ \frac{1}{n}\sum_{i=1}^{n}\int \mathbb{E}_{f \sim \mathbb{GP}(0,k_{\boldsymbol{\theta}})}\left[\ell(f(\boldsymbol{x}_i), y_i)\right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + R(p)$$

- Measure p is non-atomic and absolutely continuous

luizf@seas.upenn.edu

Assumptions

- **Bayesian posterior optimization:**

$$\begin{aligned} \underset{p \in L_1^+}{\text{minimize}} \quad & \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathbb{GP}(0, k_{\boldsymbol{\theta}})} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + R(p) \\ \text{subject to} \quad & \int p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 \end{aligned}$$

- Measure p is non-atomic and absolutely continuous

luizf@seas.upenn.edu

Assumptions

- **Bayesian posterior optimization:**

$$\begin{aligned} \underset{p \in L_1^+}{\text{minimize}} \quad & \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathbb{GP}(0, k_{\boldsymbol{\theta}})} \left[\ell(f(\boldsymbol{x}_i), y_i) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + R(p) \\ \text{subject to} \quad & \int p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 \end{aligned}$$

- Measure p is non-atomic and absolutely continuous
- R is a separable functional

luizf@seas.upenn.edu

Assumptions

- **Bayesian posterior optimization:**

$$\begin{aligned}\underset{p \in L_1^+}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathbb{GP}(0,k_{\boldsymbol{\theta}})} \left[\ell(f(\boldsymbol{x}_i), y_i)\right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ & + \sum_{j=1}^{m} \lambda_j \int h_j \left[p(\boldsymbol{\theta}), \boldsymbol{\theta}\right] d\boldsymbol{\theta} \\ \text{subject to} \quad & \int p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1\end{aligned}$$

- Measure p is non-atomic and absolutely continuous
- R is a separable functional

luizf@seas.upenn.edu

- **Bayesian posterior optimization:**

$$\begin{aligned}\underset{p \in L_1^+}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n}\int \mathbb{E}_{f\sim\mathbb{GP}(0,k_{\boldsymbol{\theta}})}\left[\ell(f(\boldsymbol{x}_i),y_i)\right]p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ & + \sum_{j=1}^{m}\lambda_j \int h_j\left[p(\boldsymbol{\theta}),\boldsymbol{\theta}\right]d\boldsymbol{\theta} \\ \text{subject to} \quad & \int p(\boldsymbol{\theta})d\boldsymbol{\theta} = 1\end{aligned}$$

- Measure p is non-atomic and absolutely continuous
- R is a separable functional
- ℓ and h_j are (possibly non-convex) normal integrands

luizf@seas.upenn.edu

Optimizing posteriors

- Strong duality
  (BPO problem is an SFP [Chamon'19])

luizf@seas.upenn.edu

Optimizing posteriors

- Strong duality
  (BPO problem is an SFP [Chamon'19])

- Exchangeability of the infimum and integral operators
  (normal integrand + separability of L_1^+ [Rockafellar'76])

luizf@seas.upenn.edu

Optimizing posteriors

- Strong duality
  (BPO problem is an SFP [Chamon'19])

- Exchangeability of the infimum and integral operators
  (normal integrand + separability of L_1^+ [Rockafellar'76])

- Lagrangian can be minimized efficiently, often in closed-form
  (separability + Gaussian integrals)

luizf@seas.upenn.edu

Optimizing posteriors

1) $$\bar{\ell}(\boldsymbol{\theta}) = \int \left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)\right] \underbrace{\mathbb{P}\left(f(\bar{\boldsymbol{x}}) \mid \boldsymbol{\theta}\right)}_{\mathcal{N}(f(\bar{\boldsymbol{x}})|\boldsymbol{\mu_\theta},\boldsymbol{\Sigma_\theta})} df$$

[Gauss-Hermite quadrature]

2) $$p_d(\boldsymbol{\theta}, \mu) = \underset{p \geq 0}{\operatorname{argmin}} \ (\bar{\ell}(\boldsymbol{\theta}) + \mu)p + \sum_{j=1}^{m} \lambda_j h_j\left(p, \boldsymbol{\theta}\right)$$

[(often) closed-form]

3) $$\mu^\star = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \ \mathcal{L}(p_d(\boldsymbol{\theta}, \mu), \mu)$$

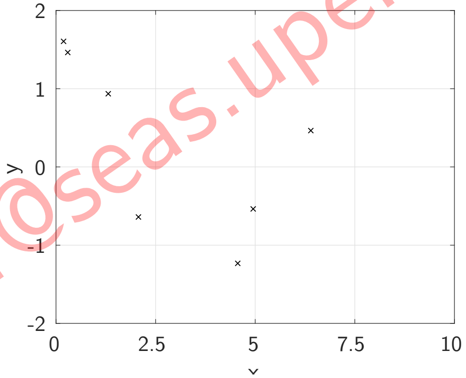
[SGD or PBA]

4) $$p^\star(\boldsymbol{\theta}) = p_d(\mu^\star, \boldsymbol{\theta})$$

luizf@seas.upenn.edu

- GP (RBF): $\sigma^2 = 1$, $\kappa = 1$, and $\sigma_\epsilon^2 = 10^{-1}$

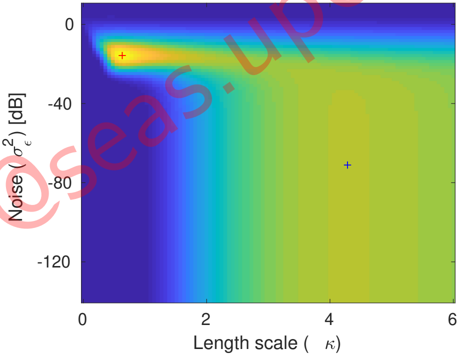
$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left[-\kappa \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\right] + \sigma_\epsilon^2 \boldsymbol{I}$$



luizf@seas.upenn.edu

- GP (RBF): $\sigma^2 = 1$, $\kappa = 1$, and $\sigma_\epsilon^2 = 10^{-1}$

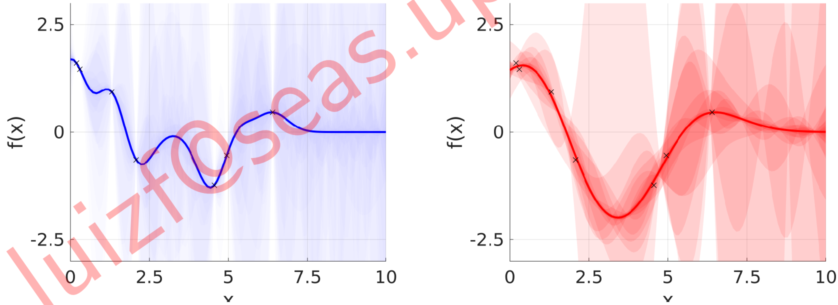
$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left[-\kappa \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\right] + \sigma_\epsilon^2 \boldsymbol{I}$$



luizf@seas.upenn.edu

- GP (RBF): $\sigma^2 = 1$, $\kappa = 1$, and $\sigma_\epsilon^2 = 10^{-1}$

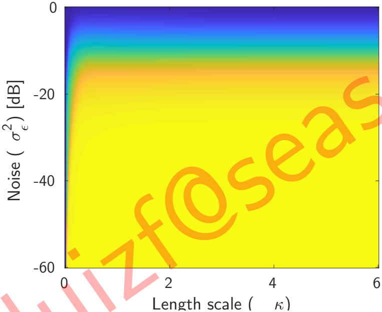
$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left[-\kappa \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\right] + \sigma_\epsilon^2 \boldsymbol{I}$$

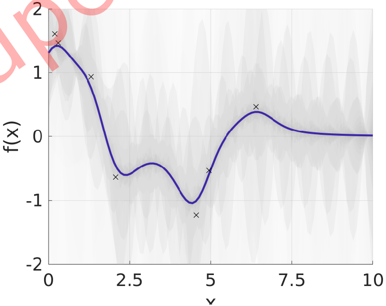


luizf@seas.upenn.edu

Numerical examples

- Loss: ℓ_2-norm
- Regularization: negative entropy

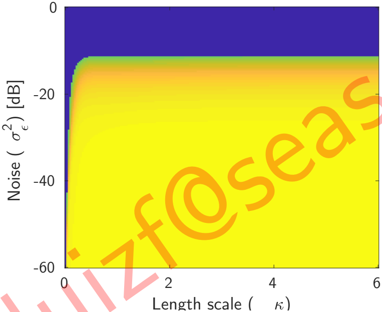


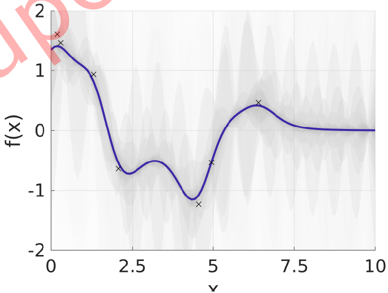


luizf@seas.upenn.edu

Numerical examples

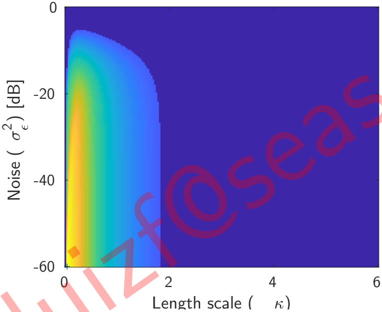
- Loss: ℓ_2-norm
- Regularization: negative entropy $+ L_0$

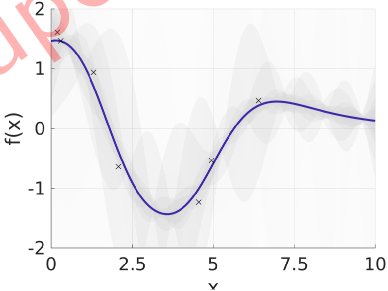




luizf@seas.upenn.edu

- Loss: ℓ_2-norm
- Regularization: negative entropy $+ L_0 + \mathbb{E}[\kappa]$

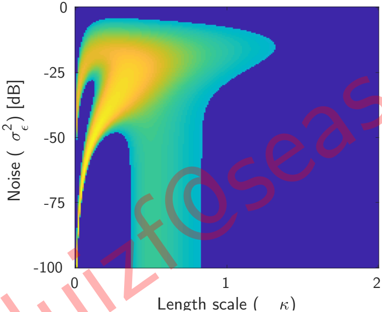


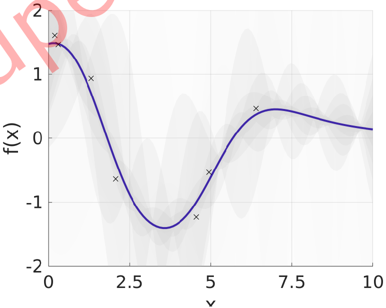


luiz@seas.upenn.edu

Numerical examples

- Loss: Leave-one-out ℓ_2-norm
- Regularization: negative entropy $+ L_0$





luiz@seas.upenn.edu

Conclusion

- Priors for nonparametric Bayesian methods are hard to specify and learning them from data is challenging
- *Bayesian posterior optimization*: replace the prior by a statistical optimization problem
- Despite the non-convexity and infinite dimensionality, posterior optimization problems can be solved efficiently

luizf@seas.upenn.edu

LEARNING GPs WITH BAYESIAN POSTERIOR OPTIMIZATION

Luiz F. O. Chamon, Santiago Paternain, and Alejandro Ribeiro

luizf@seas.upenn.edu

ASILOMAR 2019
November 4$^{\text{th}}$, 2019