LEARNING GPs WITH
BAYESIAN POSTERIOR OPTIMIZATION

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

ASILOMAR 2019
November 4\textsuperscript{th}, 2019
Dealing with complexity and uncertainty

- Nonparametric methods

---

Chamon et al. Learning GPs with Bayesian Posterior Optimization
Dealing with complexity and uncertainty

- Nonparametric methods
- Bayesian methods
Nonparametric Bayesian methods

All of Bayesian inference

\[
P(\text{model}) + P(\text{data} \mid \text{model}) \rightarrow P(\text{model} \mid \text{data})
\]

- Prior
- Likelihood
- Posterior
Nonparametric Bayesian methods

- All of Bayesian inference

\[ P(\text{model}) + P(\text{data} \mid \text{model}) \to P(\text{model} \mid \text{data}) \]

- Parametric models are finite dimensional

\[ P(\text{model}) = P(\text{parameters}) \]

- Nonparametric models are infinite dimensional
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
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]
Maximize likelihood w.r.t. hyperparameters \[\text{[Stein'99, RW'06]}\]

- Ambiguity: multimodal likelihood, local maxima
- Indeterminacy: different parameters, same measure

\[ \theta^* = \arg\max_{\theta} \log P(y | X, \theta) \]
Which GP?

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

\[
\theta^* = \arg\max_{\theta} \log P(y \mid X, \theta)
\]

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

\[
\theta \sim P
\]
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 ($0$-loss, truncated MSE, ...)
    - ✓ Incorporate complex structures in the prior: maximum entropy, sparsity, moments, ...
    - × Non-convex, infinite dimensional optimization problem
      ⇒ simple, efficient solution using duality
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
      ⇒ simple, efficient solution using duality

Chamon et al. Learning GPs with Bayesian Posterior Optimization 6
luizf@seas.upenn.edu
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 ⇒ simple, efficient solution using duality
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...
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
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
  \[ \Rightarrow \text{simple, efficient solution using duality} \]
Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems
GP regression

- **Data:** \((x_i, y_i)\) with \(y_i \sim \mathcal{N}(f(x_i), \sigma^2)\) for an unknown \(f\)

- **Goal:** determine \(f \mid X, y\)

- **How?** Bayes’ rule and GP prior
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 $[g(x_1) \cdots g(x_n)] \sim \mathcal{N}(m, K)$ for all $n \in \mathbb{N}$

$$m = \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

- Typically, $m \equiv 0$
GP regression

- **Data:** \((x_i, y_i)\) with \(y_i \sim \mathcal{N}(f(x_i), \sigma^2)\) for an unknown \(f\)
- **Goal:** determine \((f \mid X, y)\)
- **How?** Bayes’ rule and GP prior

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

\[
\mu = \bar{k}^T K^{-1} y, \quad \Sigma = k(\bar{x}, \bar{x}) - \bar{k}^T K^{-1} \bar{k}
\]

\[
\bar{k} = \begin{bmatrix}
k(\bar{x}, x_1) \\
\vdots \\
k(\bar{x}, x_n)
\end{bmatrix}
\]

and

\[
K = \begin{bmatrix}
k(x_1, x_1) & \cdots & k(x_1, x_n) \\
\vdots & \ddots & \vdots \\
k(x_n, x_1) & \cdots & k(x_n, x_n)
\end{bmatrix}
\]
First level: unknown function $f$

$$\mathbb{P}(f \mid X, y) \propto \mathbb{P}(y \mid X, f) \mathbb{P}(f \mid X)$$
First level: unknown function $f$

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

$$\mathbb{P}(f \mid X, y) = \int \mathbb{P}(f \mid X, y, \theta) \mathbb{P}(\theta \mid X, y) d\theta$$
First level: unknown function $f$

\[
P(f \mid X, y, \theta) \propto P(y \mid X, f, \theta) P(f \mid X, \theta)
\]

\[
P(f \mid X, y) = \int P(f \mid X, y, \theta) P(\theta \mid X, y) \, d\theta
\]

\[
\theta\text{-posterior}
\]
First level: unknown function \( f \)

\[
P(f \mid X, y, \theta) \propto P(y \mid X, f, \theta) P(f \mid X, \theta)
\]

\[
P(f \mid X, y) = \int P(f \mid X, y, \theta) P(\theta \mid X, y) d\theta
\]

Second level: hyperparameters \( \theta \)

\[
P(\theta \mid X, y) \propto P(y \mid X, \theta) P(\theta \mid X)
\]
First level: unknown function $f$

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

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

Second level: hyperparameters $\theta$

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

**Issue:** choosing $\mathbb{P}(\theta \mid X)$

(interpretation, indeterminacy, informativeness)
First level: unknown function $f$

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

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

Second level: hyperparameters $\theta$

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

Issue: choosing $\mathbb{P}(\theta \mid X)$

(interpretation, indeterminacy, informativeness)
Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems
▶ Statistical learning:

\[
\phi^* = \arg\min_{\phi \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(\phi(x), y)] + R(\phi)
\]

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

\[ \phi^* = \arg\min_{\phi \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \ell(\phi(x), y) \right] + R(\phi) \]

Empirical risk minimization:

\[ \hat{\phi}^* = \arg\min_{\phi \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(\phi(x_i), y_i) + R(\phi) \]

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

▶ Statistical GP learning:

\[
\Gamma^* = \arg\min_{\Gamma \in \mathcal{GP}} \mathbb{E}_{(x,y) \sim \mathcal{D}, f \sim \Gamma} [\ell(f(x), y)] + R(\gamma)
\]

▶ Empirical GP-risk minimization:

\[
\hat{\Gamma}^* = \arg\min_{\Gamma \in \mathcal{GP}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \Gamma} [\ell(f(x_i), y_i)] + R(\gamma)
\]

- \(\mathcal{D}\) is an unknown probability distribution over pairs \((x, y)\)
- \(\ell: \mathbb{R}^2 \rightarrow \mathbb{R}_+\) is a loss function
- \(R\) is a regularizer
- \(\mathcal{GP}\) is the “space of Gaussian processes”
- Data: \((x_i, y_i) \sim \mathcal{D}\)
Empirical GP-risk minimization:

\[
\hat{\Gamma}^* = \arg\min_{\Gamma \in \mathcal{GP}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \Gamma} [\ell(f(x_i), y_i)] + R(\gamma)
\]

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

\[
\hat{\Gamma}^* = \arg\min_{\Gamma \in \mathcal{GP}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \Gamma} [\ell(f(x_i), y_i)] + R(\gamma)
\]

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

Leverage the first level of the hierarchical model

\[
\mathbb{P}(f \mid X, y) = \int \mathbb{P}(f \mid X, y, \theta) \mathbb{P}(\theta \mid X, y) \, d\theta
\]

“Parameterize” (a subset of) \( \mathcal{GP} \) using \( \mathbb{P}(\theta \mid X, y) \)
Bayesian posterior optimization:

\[ p^* = \arg\min_{p \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \text{GP}(0, k_{\theta})} [\ell(f(x_i), y_i)] p(\theta) d\theta + R(p) \]

\[ \hat{\Gamma}_p^* = \int \mathbb{P}(f \mid X, y, \theta) p^*(\theta) d\theta \]

- Optimization variable: \( p(\theta) = \mathbb{P}(\theta \mid X, y) \)
- Alternative interpretation: mixture of GPs with weights \( p(\theta) \)
- Non-convex, infinite dimensional optimization problem
Roadmap

The Bayesian part

The optimization part

Solving Bayesian posterior optimization problems
Assumptions

Bayesian posterior optimization:

\[
\minimize_{p \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \mathcal{GP}(0, k_\theta)} [\ell(f(x_i), y_i)] p(\theta) d\theta + R(p)
\]
Assumptions

- **Bayesian posterior optimization:**

  $$\minimize_{p \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \mathcal{GP}(0, k_\theta)} [\ell(f(x_i), y_i)] p(\theta) d\theta + R(p)$$

- **Measure** $p$ is non-atomic and absolutely continuous.
Assumptions

- **Bayesian posterior optimization**:

  $$\min_{p \in L_1^+} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f \sim \text{GP}(0, k_{\theta})} \left[ \ell(f(x_i), y_i) \right] p(\theta) d\theta + R(p)$$

  subject to $$\int p(\theta) d\theta = 1$$

- Measure $p$ is non-atomic and absolutely continuous
Bayesian posterior optimization:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \text{GP}(0, k_{\theta})} [\ell(f(x_i), y_i)] p(\theta) d\theta + R(p) \\
\text{subject to} & \quad \int p(\theta) d\theta = 1
\end{align*}
\]

- Measure \( p \) is non-atomic and absolutely continuous
- \( R \) is a separable functional
Assumptions

Bayesian posterior optimization:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \text{GP}(0,k_{\theta})} [\ell(f(x_i), y_i)] p(\theta) d\theta \\
& + \sum_{j=1}^{m} \lambda_j \int h_j [p(\theta), \theta] d\theta \\
\text{subject to} & \quad \int p(\theta) d\theta = 1
\end{align*}
\]

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

- **Bayesian posterior optimization:**

\[
\minimize_{p \in L_1^+} \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{E}_{f \sim \text{GP}(0, k_\theta)} [\ell(f(x_i), y_i)] p(\theta) \, d\theta \\
+ \sum_{j=1}^{m} \lambda_j \int h_j [p(\theta), \theta] \, d\theta
\]

subject to \( \int p(\theta) \, d\theta = 1 \)

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

- Strong duality
  (BPO problem is an SFP [Chamon’19])
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])
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)
Optimizing posteriors

1) \( \bar{\ell}(\theta) = \int \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right] df \)
\[ \mathcal{P}(f(\bar{x}) | \theta) \]
\[ \mathcal{N}(f(\bar{x}) | \mu_\theta, \Sigma_\theta) \]

[Gauss-Hermite quadrature]

2) \( p_d(\theta, \mu) = \arg\min_{p \geq 0} (\bar{\ell}(\theta) + \mu) p + \sum_{j=1}^{m} \lambda_j h_j(p, \theta) \)

[(often) closed-form]

3) \( \mu^* = \arg\max_{\mu \in \mathbb{R}} \mathcal{L}(p_d(\theta, \mu), \mu) \)

[SGD or PBA]

4) \( p^*(\theta) = p_d(\mu^*, \theta) \)
Numerical examples

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

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

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

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

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

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

- Loss: $\ell_2$-norm
- Regularization: negative entropy
Numerical examples

- Loss: $\ell_2$-norm
- Regularization: negative entropy + $L_0$
Numerical examples

- Loss: $\ell_2$-norm
- Regularization: negative entropy + $L_0 + \mathbb{E} [\kappa]$
Numerical examples

- Loss: Leave-one-out $\ell_2$-norm
- Regularization: negative entropy + $L_0$
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
LEARNING GPs WITH BAYESIAN POSTERIOR OPTIMIZATION

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

ASILOMAR 2019
November 4th, 2019