Sampling from Gaussian Process Posteriors using Stochastic Gradient Descent

Jihao Andreas Lin*, Javier Antorán*, Shreyas Padhy*, David Janz, José Miguel Hernández-Lobato, Alexander Terenin

1. Gaussian Process Regression

$\mathbf{x} \in X^N, \mathbf{y} \in \mathbb{R}^N$

$$\mu_{f\mid\mathbf{y}}(\cdot) = \mathbf{K}_{(\cdot)\mathbf{x}}(\mathbf{K}_{\mathbf{xx}} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \qquad k_{f\mid\mathbf{y}}(\cdot,\cdot') = \mathbf{K}_{(\cdot,\cdot')} - \mathbf{K}_{(\cdot)\mathbf{x}}(\mathbf{K}_{\mathbf{xx}} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{\mathbf{x}(\cdot')}$$

$$(f\mid\mathbf{y})(\cdot) = f(\cdot) + \mathbf{K}_{(\cdot)\mathbf{x}}\underbrace{(\mathbf{K}_{\mathbf{xx}} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - f(\mathbf{x}) - \epsilon)}_{=\alpha^*} \qquad f \sim \mathrm{GP}(\cdot\mid 0, k) \quad \epsilon \sim \mathcal{N}(\cdot\mid 0, \sigma^2 I)$$

2. Optimization-based Learning

$\mathcal{O}(N)$ per iteration

$$\alpha^* = \underset{\alpha\in\mathbb{R}^N}{\arg\min}\ \frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{K}_{\mathbf{xx}}\alpha\|_2^2 + \|\alpha\|^2_{\mathbf{K}_{\mathbf{xx}}}$$

$$= \sum_{i=1}^{N}(y_i - \mathbf{K}_{x_i\mathbf{x}}\alpha)^2 \approx \frac{N}{D}\sum_{i=1}^{D}(y_i - \mathbf{K}_{x_i\mathbf{x}}\alpha)^2$$

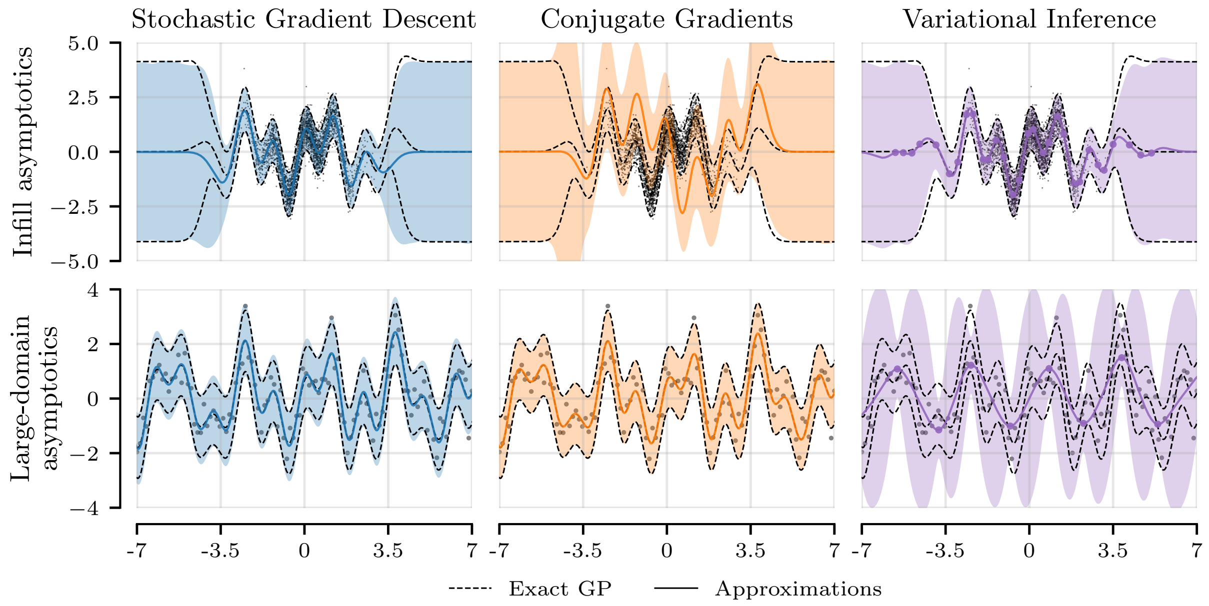
Estimate with mini-batches

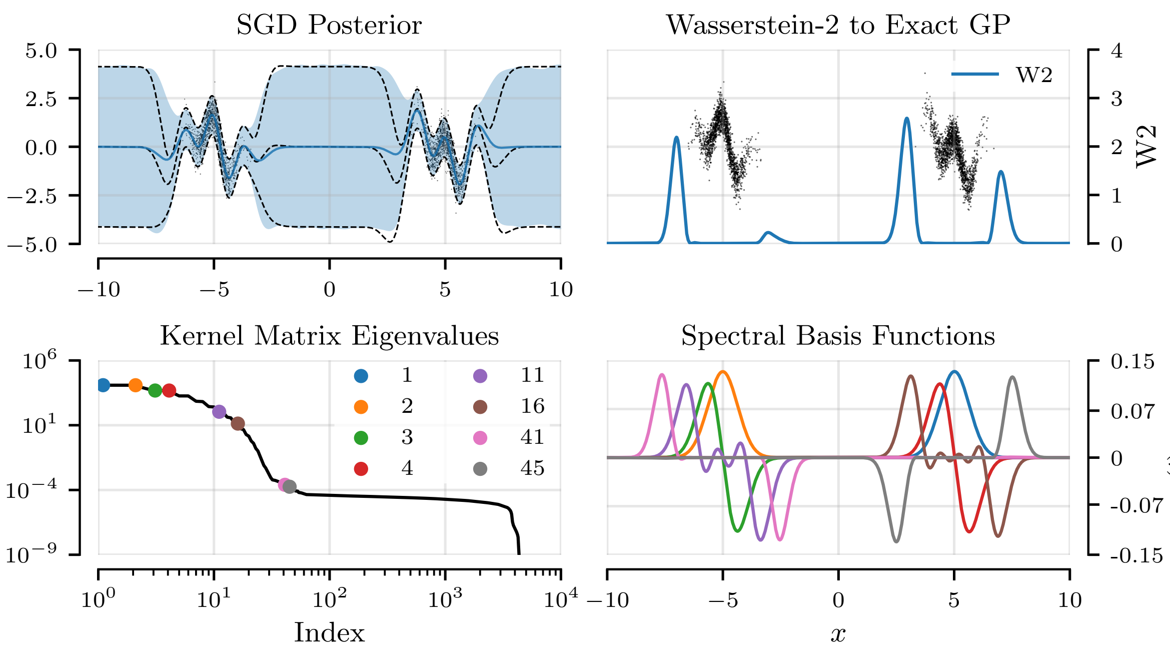
Estimate with random features

$$= \alpha^\top\mathbf{K}_{\mathbf{xx}}\alpha \approx \alpha^\top\Phi(\mathbf{x})^\top\Phi(\mathbf{x})\alpha$$

$$= \|\Phi(\mathbf{x})\alpha\|_2^2 = \sum_{l=1}^{L}(\phi_l(\mathbf{x})^\top\alpha)^2$$

3. Illustrative Visualizations





4. Large-scale Experiments