Efficient High Dimensional Bayesian Optimization with Additivity and Quadrature Fourier Features

Mojmír Mutný Andreas Krause

November 29, 2018

Learning & Adaptive Systems group
Institute of Machine Learning
ETH Zürich

Bayesian Optimization

- **Goal:** $\arg\max_{x \in D} g(x)$

Bayesian Optimization

- **Goal:** $\arg\max_{x \in D} g(x)$
- **Bayesian perspective:**
  $g \sim GP(0, k)$ with
  a stationary kernel k

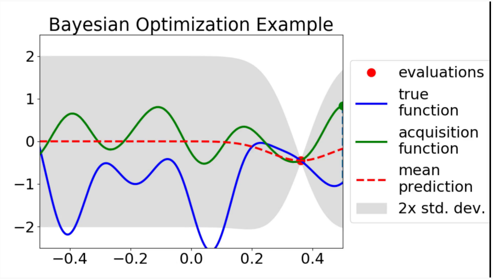
Bayesian Optimization

- **Goal:** $\arg\max_{x \in D} g(x)$
- **Bayesian perspective:**
  $g \sim GP(0, k)$ with
  a stationary kernel k

- **Metric**: cumulative regret: $R_T = \sum_{t=1}^{T} g(x_t) - g(x^*)$

Bayesian Optimization

- **Goal:** $\arg\max_{x \in D} g(x)$
- **Bayesian perspective:**
  $g \sim GP(0, k)$ with
  a stationary kernel k



- **Metric**: cumulative regret: $R_T = \sum_{t=1}^{T} g(x_t) - g(x^*)$
- **Challenge**: exploration vs. exploitation $\implies$ Bayesian Optimization.

Challenges of high dimensions

- **Statistical**

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions
  Additive functions

$$g(\boxed{x_1}\boxed{x_2}\boxed{x_3}\boxed{x_4}) = g_1(\boxed{x_1}\boxed{x_2}) + g_2(\boxed{x_3}) + g_3(\boxed{x_4})$$

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

$$g(x_1\, x_2\, x_3\, x_4) = g_1(x_1\, \overbrace{x_2) + g_2(x_2}^{\text{shared}}\, x_3) + g_3(x_4)$$

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

$$g(x_1\, x_2\, x_3\, x_4) = g_1(\underbrace{x_1\, x_2}_{\bar{d} \coloneqq \textit{size of largest group}}) + g_2(x_2\, x_3) + g_3(x_4)$$

shared

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

  shared

  $$g(x_1\,x_2\,x_3\,x_4) = g_1(\underbrace{x_1\,x_2}_{\bar{d} \coloneqq \textit{size of largest group}}) + g_2(x_2\,x_3) + g_3(x_4)$$

- **Computational**

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

  shared

  $$g(x_1\,x_2\,x_3\,x_4) = g_1(\underbrace{x_1\,x_2}_{\bar{d} \coloneqq \text{size of largest group}}) + g_2(x_2\,x_3) + g_3(x_4)$$

- **Computational**
  - Kernel inversion $\mathcal{O}(T^3)$

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

  shared

  $$g(x_1\, x_2\, x_3\, x_4) = g_1(\underbrace{x_1\, x_2}_{\bar{d} \coloneqq \textit{size of largest group}}) + g_2(x_2\, x_3) + g_3(x_4)$$

- **Computational**
  - Kernel inversion $\mathcal{O}(T^3) \rightarrow \mathcal{O}(T \log T)$

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

  shared

  $$g(x_1\, x_2\, x_3\, x_4) = g_1(\underbrace{x_1\, x_2}_{\bar{d} \,:=\, \textit{size of largest group}}) + g_2(x_2\, x_3) + g_3(x_4)$$

- **Computational**
  - Kernel inversion $\mathcal{O}(T^3) \rightarrow \mathcal{O}(T \log T)$
  - Optimization of the acquisition function

Challenges of high dimensions

- **Statistical** $\rightarrow$ needs assumptions

  Additive functions

shared

$$g(x_1\, x_2\, x_3\, x_4) = g_1(\underbrace{x_1\, x_2}_{\bar{d} \,\coloneqq\, \textit{size of largest group}}) + g_2(x_2\, x_3) + g_3(x_4)$$

- **Computational**
  - Kernel inversion $\mathcal{O}(T^3) \rightarrow \mathcal{O}(T \log T)$
  - Optimization of the acquisition function $\rightarrow$ coordinate optimization

Main tool: Quadrature Fourier Features (QFF)

$$k(x-y) \overset{\text{Bochner}}{=} \int_\Omega p(\omega) \begin{pmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{pmatrix}^\top \begin{pmatrix} \cos(\omega^\top y) \\ \sin(\omega^\top y) \end{pmatrix} d\omega$$

Main tool: Quadrature Fourier Features (QFF)

$$k(x-y) \overset{\text{Bochner}}{=} \int_{\Omega} p(\omega) \begin{pmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{pmatrix}^\top \begin{pmatrix} \cos(\omega^\top y) \\ \sin(\omega^\top y) \end{pmatrix} d\omega \overset{\text{Fourier F.}}{\approx} \Phi(x)^\top \underbrace{\Phi(y)}_{\mathbb{R}^m}$$

Main tool: Quadrature Fourier Features (QFF)

$$k(x-y) \overset{\text{Bochner}}{=} \int_\Omega p(\omega) \begin{pmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{pmatrix}^\top \begin{pmatrix} \cos(\omega^\top y) \\ \sin(\omega^\top y) \end{pmatrix} d\omega \overset{\text{Fourier F.}}{\approx} \Phi(x)^\top \underbrace{\Phi(y)}_{\mathbb{R}^m}$$

- Standard approach Monte Carlo estimate - sample $\omega \sim p(\omega)$, (RFF)

Main tool: Quadrature Fourier Features (QFF)

$$k(x-y) \overset{\text{Bochner}}{=} \int_\Omega p(\omega) \begin{pmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{pmatrix}^\top \begin{pmatrix} \cos(\omega^\top y) \\ \sin(\omega^\top y) \end{pmatrix} d\omega \overset{\text{Fourier F.}}{\approx} \Phi(x)^\top \underbrace{\Phi(y)}_{\mathbb{R}^m}$$

- Standard approach Monte Carlo estimate - sample $\omega \sim p(\omega)$, (RFF)
- This work: **Gaussian Quadrature**.

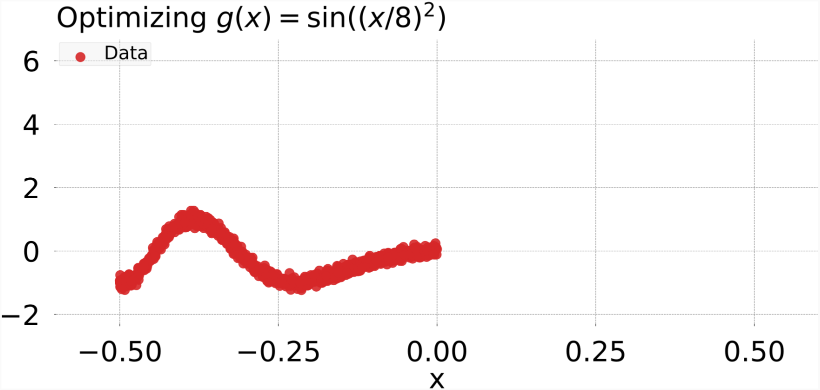
Main tool: Quadrature Fourier Features (QFF)

$$k(x-y) \overset{\text{Bochner}}{=} \int_\Omega p(\omega) \begin{pmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{pmatrix}^\top \begin{pmatrix} \cos(\omega^\top y) \\ \sin(\omega^\top y) \end{pmatrix} d\omega \overset{\text{Fourier F.}}{\approx} \Phi(x)^\top \underbrace{\Phi(y)}_{\mathbb{R}^m}$$

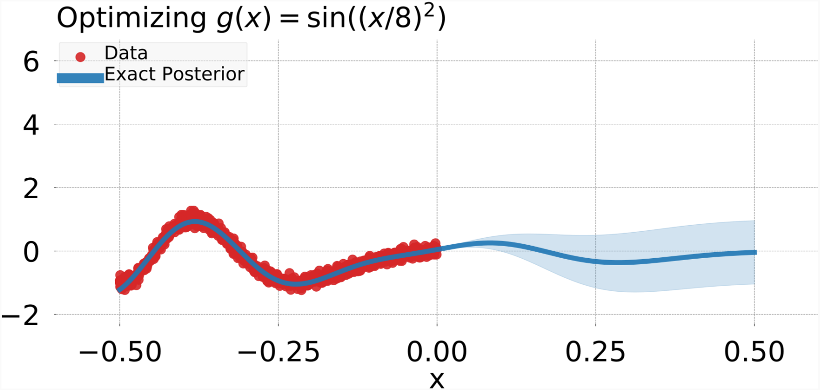
- Standard approach Monte Carlo estimate - sample $\omega \sim p(\omega)$, (RFF)
- This work: **Gaussian Quadrature**.
- (generalized) additivity $\implies$ favorable scaling with $\bar{d}$,

$$\left\| k(x,y) - \Phi(x)^\top \Phi(y) \right\|_\infty = \mathcal{O}\left(2^{\bar{d}} \rho^m\right) \ \ \rho < 1$$

Example

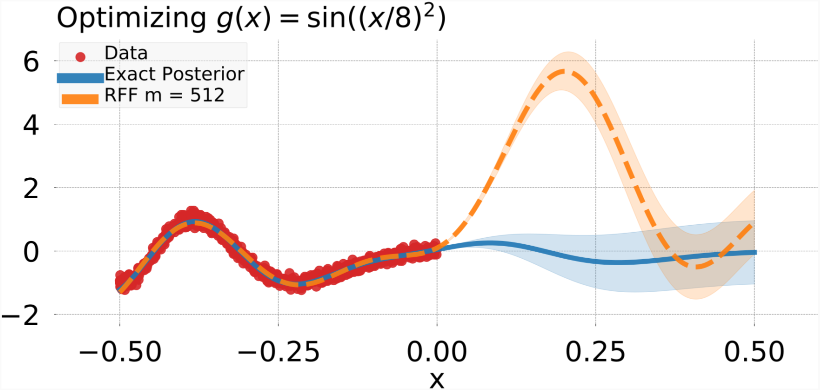


Example

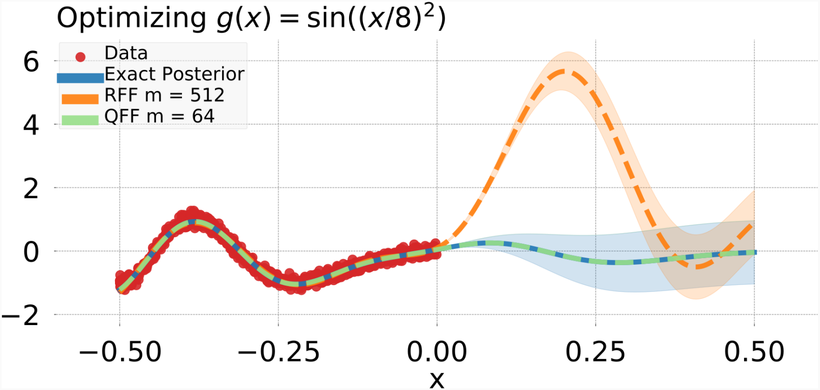


Optimizing $g(x) = \sin((x/8)^2)$
Data
Exact Posterior
6
4
2
0
−2
−0.50
−0.25
0.00
0.25
0.50
x

Example


Optimizing $g(x) = \sin((x/8)^2)$
Data
Exact Posterior
RFF m = 512
6
4
2
0
−2
−0.50
−0.25
0.00
0.25
0.50
x

Example


Optimizing $g(x) = \sin((x/8)^2)$
Data
Exact Posterior
RFF m = 512
QFF m = 64
6
4
2
0
−2
−0.50
−0.25
0.00
0.25
0.50
x

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

- QFF are provably better than RFF for additive kernels

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

- QFF are provably better than RFF for additive kernels
- High quality approximation and analytical form of posterior samples from a GP.

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

- QFF are provably better than RFF for additive kernels
- High quality approximation and analytical form of posterior samples from a GP.
- With additive model, the Thompson sampling acquisiton function decomposes over the variable groups - one pass coordinate ascent.

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

- QFF are provably better than RFF for additive kernels
- High quality approximation and analytical form of posterior samples from a GP.
- With additive model, the Thompson sampling acquisiton function decomposes over the variable groups - one pass coordinate ascent.
- Thompson sampling with QFF for squared exponential kernel is no-regret.

Algorithm

Main Contribution

First efficient and provably accurate high dimensional Bayesian optimization using additive models.

- QFF are provably better than RFF for additive kernels
- High quality approximation and analytical form of posterior samples from a GP.
- With additive model, the Thompson sampling acquisiton function decomposes over the variable groups - one pass coordinate ascent.
- Thompson sampling with QFF for squared exponential kernel is no-regret.

Please come to the poster #23.

Room 210 & 230 AB

Kandasamy, K. and Yu, Y. (2016).
Additive approximations in high dimensional nonparametric regression via the SALSA.
In *International Conference on Machine Learning*, ICML'16, pages 69–78. JMLR.org.

Rahimi, A., Recht, B., et al. (2007).
Random features for large-scale kernel machines.
In *Advances in Neural Information Processing Systems*, volume 3, page 5.

Rolland, P., Scarlett, J., Bogunovic, I., and Cevher, V. (2018).
High-dimensional Bayesian optimization via additive models with overlapping groups.
International Conference on Artificial Intelligence and Statistics, 84.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Feitas, N. (2016).
Bayesian optimization in a billion dimensions via random embeddings.
Journal of Artificial Intelligence Research, 55:361–387.

Wang, Z. and Jegelka, S. (2017).
Max-value entropy search for efficient Bayesian optimization.
International Conference on Machine Learning.