Batch Bayesian optimisation via density-ratio estimation with guarantees

Rafael Oliveira, Louis C. Tiao and Fabio Ramos
The University of Sydney, Australia
NVIDIA, USA

Global optimisation problems

Global optimisation:

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$$

- f is a "black box".

  $\hookrightarrow$ only observable via noisy and expensive evaluations

Applications:

- Hyper-parameter tuning (e.g., Optuna, HyperOpt, etc.)
- Neural architecture search
- Robotic exploration, chemical design, environmental monitoring, etc.

Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

Bayesian optimisation: the basics

- Model f as a random variable
  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)

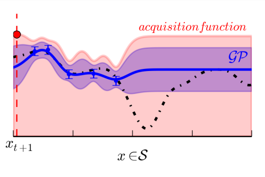
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

  $\rightarrow$ e.g., upper confidence bound (UCB):

$$a_{\text{UCB}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mu_{t-1}(\mathbf{x}) + \beta\sigma_{t-1}(\mathbf{x})$$

  $\rightarrow$ e.g., expected improvement (EI):

$$a_{\text{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mathbb{E}[\max\{0, \tau - f(\mathbf{x})\} \mid \mathcal{D}_{t-1}], \quad \tau := \min_{i<t} y_i$$



Bayesian optimisation: the basics

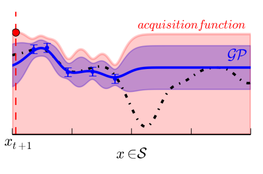
- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)

- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

  $\rightarrow$ e.g., upper confidence bound (UCB):

$$a_{\mathrm{UCB}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mu_{t-1}(\mathbf{x}) + \beta\sigma_{t-1}(\mathbf{x})$$



  $\rightarrow$ e.g., expected improvement (EI):

$$a_{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mathbb{E}[\max\{0, \tau - f(\mathbf{x})\} \mid \mathcal{D}_{t-1}], \quad \tau := \min_{i<t} y_i$$

Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)

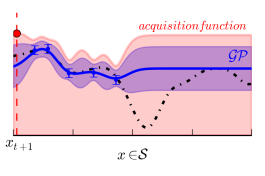
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

  $\rightarrow$ e.g., upper confidence bound (UCB):

$$a_{\mathrm{UCB}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mu_{t-1}(\mathbf{x}) + \beta\sigma_{t-1}(\mathbf{x})$$

  $\rightarrow$ e.g., expected improvement (EI):

$$a_{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mathbb{E}[\max\{0, \tau - f(\mathbf{x})\} \mid \mathcal{D}_{t-1}], \quad \tau := \min_{i<t} y_i$$



Bayesian optimisation: the basics

- Model f as a random variable

  $\hookrightarrow$ e.g., $f \sim \mathcal{GP}(0, k)$ (Gaussian process)

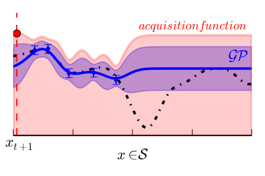
- Condition the model on past data $\mathcal{D}_{t-1} := \{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$
- Optimise an **acquisition function** $a(\mathbf{x}|\mathcal{D}_{t-1})$ to collect new data $\mathbf{x}_t, y_t$:

  $\rightarrow$ e.g., upper confidence bound (UCB):

$$a_{\text{UCB}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mu_{t-1}(\mathbf{x}) + \beta\sigma_{t-1}(\mathbf{x})$$

  $\rightarrow$ e.g., expected improvement (EI):

$$a_{\text{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) := \mathbb{E}[\max\{0, \tau - f(\mathbf{x})\} \mid \mathcal{D}_{t-1}], \quad \tau := \min_{i<t} y_i$$



Repeat for $t \in \{1, \dots, T\}$

BORE: Bayesian optimisation by density-ratio estimation (Tiao et al., 2021)

Expected improvement as a density ratio

Given $\ell(\mathbf{x}) := p(\mathbf{x}|y \leq \tau)$ and $g(\mathbf{x}) := p(\mathbf{x}|y > \tau)$, Tiao et al. (2021) showed that:

$$a_{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) \propto \frac{\ell(\mathbf{x})}{\gamma\ell(\mathbf{x}) + (1-\gamma)g(\mathbf{x})} = \gamma^{-1}\pi(\mathbf{x})$$

where $\pi(\mathbf{x}) := p(y \leq \tau|\mathbf{x}) \implies$ a **probabilistic classifier**.

- Model acquisition function a directly as $\hat{\pi}_t$ learnt from labels $z_t = \mathbb{I}[y_t \leq \tau]$

$\hookrightarrow$ Effective and scalable with flexible classifiers (e.g., deep nets, random forests, etc.)

BORE: Bayesian optimisation by density-ratio estimation (Tiao et al., 2021)

Expected improvement as a density ratio

Given $\ell(\mathbf{x}) := p(\mathbf{x}|y \leq \tau)$ and $g(\mathbf{x}) := p(\mathbf{x}|y > \tau)$, Tiao et al. (2021) showed that:

$$a_{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) \propto \frac{\ell(\mathbf{x})}{\gamma\ell(\mathbf{x}) + (1-\gamma)g(\mathbf{x})} = \gamma^{-1}\pi(\mathbf{x})$$

where $\pi(\mathbf{x}) := p(y \leq \tau|\mathbf{x}) \implies$ a **probabilistic classifier**.

- Model acquisition function a directly as $\hat{\pi}_t$ learnt from labels $z_t = \mathbb{I}[y_t \leq \tau]$

$\hookrightarrow$ Effective and scalable with flexible classifiers (e.g., deep nets, random forests, etc.)

BORE: Bayesian optimisation by density-ratio estimation (Tiao et al., 2021)

Expected improvement as a density ratio

Given $\ell(\mathbf{x}) := p(\mathbf{x}|y \leq \tau)$ and $g(\mathbf{x}) := p(\mathbf{x}|y > \tau)$, Tiao et al. (2021) showed that:

$$a_{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_{t-1}) \propto \frac{\ell(\mathbf{x})}{\gamma\ell(\mathbf{x}) + (1-\gamma)g(\mathbf{x})} = \gamma^{-1}\pi(\mathbf{x})$$

where $\pi(\mathbf{x}) := p(y \leq \tau|\mathbf{x}) \implies$ a **probabilistic classifier**.

- Model acquisition function a directly as $\hat{\pi}_t$ learnt from labels $z_t = \mathbb{I}[y_t \leq \tau]$

$\hookrightarrow$ Effective and scalable with flexible classifiers (e.g., deep nets, random forests, etc.)

BORE^{++}: BORE with improved uncertainty estimates (our contributions)

Can BORE be equipped with theoretical guarantees?

BORE^{++}: Optimising an upper confidence bound $\pi_{t,\delta}$ instead of the best-fit $\hat{\pi}_t$ leads to bounded regret:

$$\min_{t \leq T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \in \tilde{\mathcal{O}}\left(T^{-1/2}\right)$$

BORE^{++}: BORE with improved uncertainty estimates (our contributions)

Can BORE be equipped with theoretical guarantees?

BORE^{++}: Optimising an upper confidence bound $\pi_{t,\delta}$ instead of the best-fit $\hat{\pi}_t$ leads to bounded regret:

$$\min_{t \leq T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \in \tilde{\mathcal{O}}\left(T^{-1/2}\right)$$

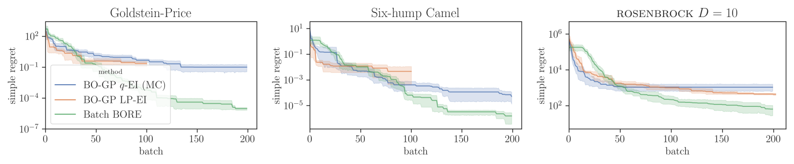
Can we collect observations in batches instead of single points?

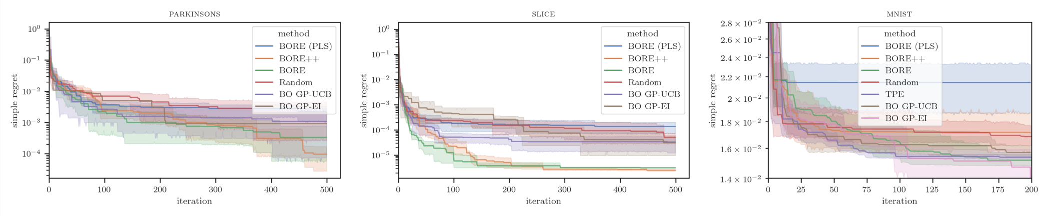
Batch BORE$^{(++)}$: Solve the batch selection problem via approximate inference:

$$\{\mathbf{x}_{t,i}\}_{i=1}^{M} \sim q_t \in \operatorname*{argmin}_{q \in \mathcal{Q}} D_{\mathrm{KL}}(q || \hat{p}_t)$$

where $\hat{p}_t \propto \pi_{t,\delta}$ or $\hat{\pi}_t$. We solve it via Stein variational gradient descent (SVGD).

Experiments on global optimisation benchmarks





Experimental results on synthetic (top) and real-data (bottom) benchmarks

Conclusion

Contributions

- Theoretical guarantees for BORE algorithms
- Batch BORE extension and its guarantees
- Experimental results on global optimisation benchmarks

Please, come to our poster session for Q&A