Gaussian Process Optimization with Adaptive Sketching: Scalable and No Regret

Daniele Calandriello, Luigi Carratino,
Alessandro Lazaric, Michal Valko, Lorenzo Rosasco

ICML, June 2019

Black-box / Bayesian / Bandit Optimization

Given A alternatives

For $t = 1, \ldots, T$

(1) Select alternative

(2) Receive noisy feedback

(3) Improve for next time

Black-box / Bayesian / Bandit Optimization

Given A alternatives

For $t = 1, \ldots, T$

(1) Select alternative

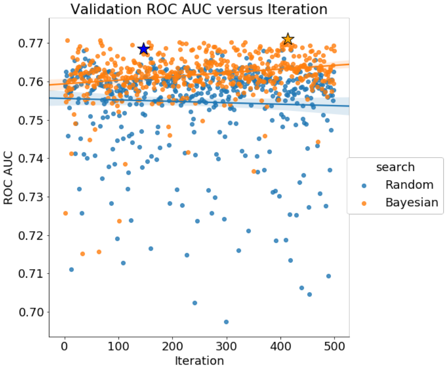
(2) Receive noisy feedback

(3) Improve for next time

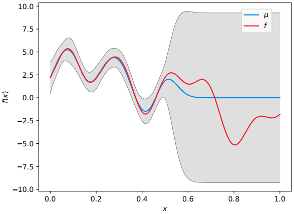
Main scientific challenges:

exploration vs exploitation

scalability

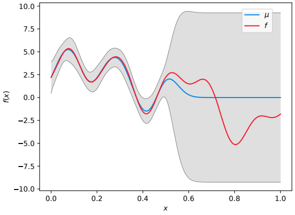


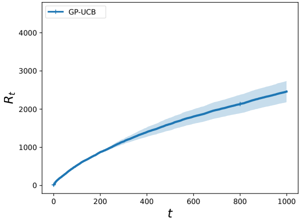
Gaussian Process Optimization:



GP-UCB

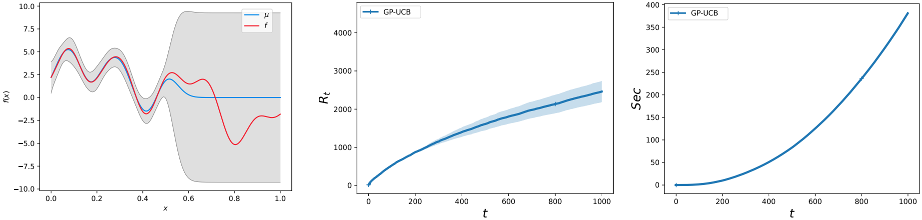
Gaussian Process Optimization: no-regret





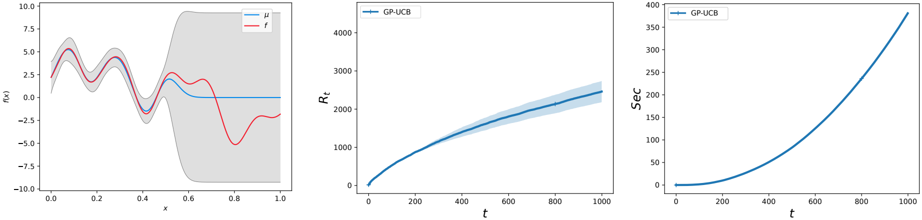
GP-UCB : no-regret [Srinivas et al., 2010]

Gaussian Process Optimization: no-regret or scalable



$\mathrm{GP\text{-}UCB}$: no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

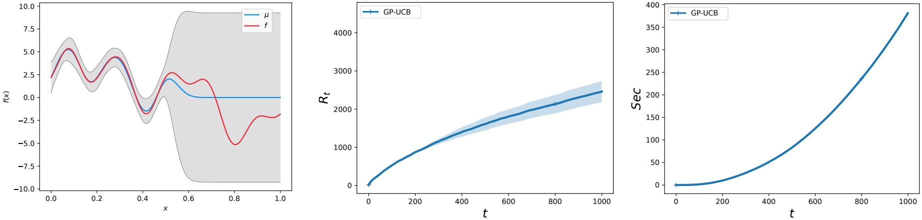
Gaussian Process Optimization: no-regret or scalable



GP-UCB : no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

Many approximations: sparse GPs, VI, RFF, Toeplitz [Huggins et al., 2019; Mutny and Krause, 2018; Quinonero-Candela et al., 2007; Wilson and Nickisch, 2015], but none or limited guarantees

Gaussian Process Optimization: no-regret and scalable

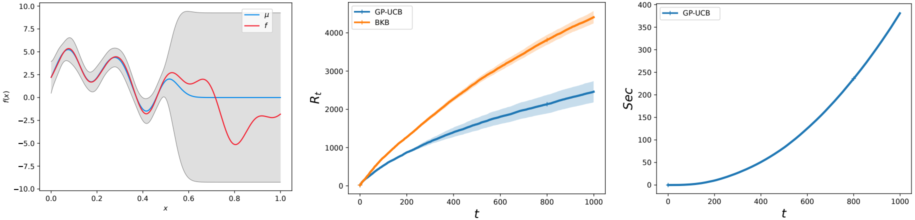


GP-UCB : no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

Many approximations: sparse GPs, VI, RFF, Toeplitz [Huggins et al., 2019; Mutny and Krause, 2018; Quinonero-Candela et al., 2007; Wilson and Nickisch, 2015], but none or limited guarantees

BKB (Budgeted Kernelized Bandits):

Gaussian Process Optimization: no-regret and scalable



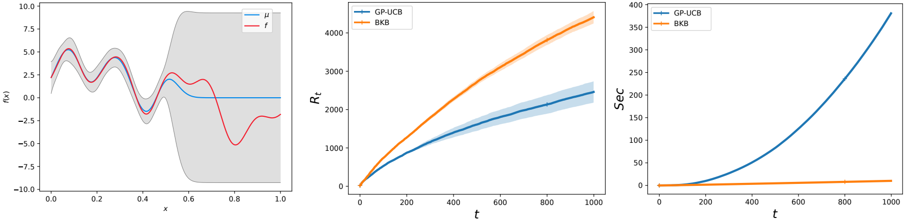
$\mathrm{GP\text{-}UCB}$: no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

Many approximations: sparse GPs, VI, RFF, Toeplitz [Huggins et al., 2019; Mutny and Krause, 2018; Quinonero-Candela et al., 2007; Wilson and Nickisch, 2015], but none or limited guarantees

BKB (Budgeted Kernelized Bandits):

no-regret: only $\mathcal{O}(\log(t))$ more than $\mathrm{GP\text{-}UCB}$

Gaussian Process Optimization: no-regret and scalable



GP-UCB : no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

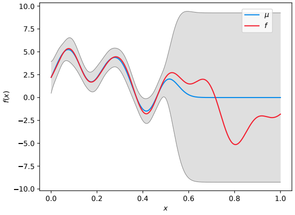
Many approximations: sparse GPs, VI, RFF, Toeplitz [Huggins et al., 2019; Mutny and Krause, 2018; Quinonero-Candela et al., 2007; Wilson and Nickisch, 2015], but none or limited guarantees

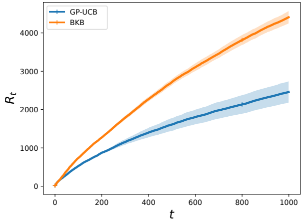
BKB (Budgeted Kernelized Bandits):

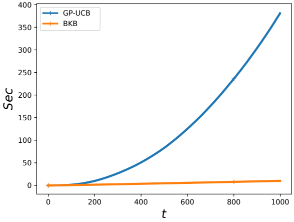
no-regret: only $\mathcal{O}(\log(t))$ more than GP-UCB

scalable: near-constant per-step complexity

Gaussian Process Optimization: no-regret and scalable







GP-UCB : no-regret [Srinivas et al., 2010] but $\mathcal{O}(t^2)$ per-step time and space

Many approximations: sparse GPs, VI, RFF, Toeplitz [Huggins et al., 2019; Mutny and Krause, 2018; Quinonero-Candela et al., 2007; Wilson and Nickisch, 2015], but none or limited guarantees

BKB (Budgeted Kernelized Bandits):

no-regret: only $\mathcal{O}(\log(t))$ more than GP-UCB

scalable: near-constant per-step complexity

no variance starvation, interpretable, extensible, ...

Black-box / Bayesian / Bandit Optimization (rigorous)

Set of arms $\mathcal{A} = \{\mathbf{x}_i\}_{i=1}^{A}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $|\mathcal{A}| = A$

Black-box / Bayesian / Bandit Optimization (rigorous)

Set of arms $\mathcal{A} = \{\mathbf{x}_i\}_{i=1}^{A}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $|\mathcal{A}| = A$

Similarity (kernel) $k(\cdot, \cdot)$ and RKHS $\mathcal{H}$

Black-box / Bayesian / Bandit Optimization (rigorous)

Set of arms $\mathcal{A} = \{\mathbf{x}_i\}_{i=1}^{A}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $|\mathcal{A}| = A$

Similarity (kernel) $k(\cdot, \cdot)$ and RKHS $\mathcal{H}$

For $t = 1, \ldots, T$

(1) Select $\mathbf{x}_t$

(2) Receive noisy feedback $y_t = f(\mathbf{x}_t) + \eta_t$

(3) Improve for next time

Black-box / Bayesian / Bandit Optimization (rigorous)

Set of arms $\mathcal{A} = \{\mathbf{x}_i\}_{i=1}^{A}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $|\mathcal{A}| = A$

Similarity (kernel) $k(\cdot, \cdot)$ and RKHS $\mathcal{H}$

For $t = 1, \ldots, T$

(1) Select $\mathbf{x}_t$

(2) Receive noisy feedback $y_t = f(\mathbf{x}_t) + \eta_t$

(3) Improve for next time

Assumptions: $f \in \mathcal{H}$ arbitrary but $\|f\| \leq F$ (frequentist/bandit regret)
Goal: minimize regret w.r.t. $\mathbf{x}_* = \arg\max_{\mathbf{x}_i \in A} f(\mathbf{x}_i)$

$$R_T = \sum_{t=1}^{T} f(\mathbf{x}_*) - f(\mathbf{x}_t)$$

GP-UCB [Srinivas et al., 2010]

Select $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathbf{X}_A} u_t(\mathbf{x})$

$$u_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta_t \sigma_t(\mathbf{x}),$$

GP-UCB [Srinivas et al., 2010]

Select $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathbf{X}_A} u_t(\mathbf{x})$

$$u_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta_t \sigma_t(\mathbf{x}),$$
$$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^\mathsf{T}(\mathbf{K}_t + \lambda \mathbf{I})^{-1}\mathbf{y}_t$$

GP-UCB [Srinivas et al., 2010]

Select $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathbf{X}_A} u_t(\mathbf{x})$

$$
\begin{aligned}
u_t(\mathbf{x}) &= \mu_t(\mathbf{x}) + \beta_t \sigma_t(\mathbf{x}), \\
\mu_t(\mathbf{x}) &= \mathbf{k}_t(\mathbf{x})^\intercal (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{y}_t \\
\sigma_t^2(\mathbf{x}) &= \frac{1}{\lambda}\Big(k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\intercal (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{k}_t(\mathbf{x}) \Big)
\end{aligned}
$$

GP-UCB [Srinivas et al., 2010]

Select $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathbf{X}_A} u_t(\mathbf{x})$

$$u_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta_t \sigma_t(\mathbf{x}),$$

$$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{y}_t \qquad \mathcal{O}(t)$$

$$\sigma_t^2(\mathbf{x}) = \frac{1}{\lambda}\Big(k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{k}_t(\mathbf{x})\Big) \qquad \mathcal{O}(t^2)$$

Too slow: $\mathcal{O}(At^2)$ per step

Sparse GP

Choose subset of m inducing points $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^m$ (a.k.a. dictionary)

Replace $k(\mathbf{x}_i, \mathbf{x}_j)$ with approximate $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j)$

$$\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{k}_{\mathcal{S}}(\mathbf{x}_i)^\mathsf{T} \mathbf{K}_{\mathcal{S}}^{+} \mathbf{k}_{\mathcal{S}}(\mathbf{x}_j)$$

,

Sparse GP

Choose subset of m inducing points $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^{m}$ (a.k.a. dictionary)

Replace $k(\mathbf{x}_i, \mathbf{x}_j)$ with approximate $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j)$

$$\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{k}_{\mathcal{S}}(\mathbf{x}_i)^{\mathsf{T}} \mathbf{K}_{\mathcal{S}}^{+} \mathbf{k}_{\mathcal{S}}(\mathbf{x}_j) = \mathbf{z}(\mathbf{x}_i)^{\mathsf{T}} \mathbf{z}(\mathbf{x}_j),$$

$$\mathbf{z}(\cdot) \triangleq \left(\mathbf{K}_{\mathcal{S}}^{1/2}\right)^{+} \mathbf{k}_{\mathcal{S}}(\cdot) : \mathbb{R}^{d} \to \mathbb{R}^{m},$$

Sparse GP

Choose subset of m inducing points $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^m$ (a.k.a. dictionary)

Replace $k(\mathbf{x}_i, \mathbf{x}_j)$ with approximate $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j)$

$$\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{k}_{\mathcal{S}}(\mathbf{x}_i)^\mathsf{T} \mathbf{K}_{\mathcal{S}}^{+} \mathbf{k}_{\mathcal{S}}(\mathbf{x}_j) = \mathbf{z}(\mathbf{x}_i)^\mathsf{T} \mathbf{z}(\mathbf{x}_j),$$

$$\mathbf{z}(\cdot) \triangleq \left(\mathbf{K}_{\mathcal{S}}^{1/2}\right)^{+} \mathbf{k}_{\mathcal{S}}(\cdot) : \mathbb{R}^d \to \mathbb{R}^m,$$

$$\mathbf{Z}_t \triangleq [\mathbf{z}(\mathbf{x}_1), \ldots, \mathbf{z}(\mathbf{x}_t)]^\mathsf{T} \in \mathbb{R}^{t \times m}$$

Deterministic Training Conditional (DTC) sparse GP-UCB

[Seeger et al., 2003]

Select $\widetilde{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathbf{X}_A} \widetilde{u}_t(\mathbf{x})$

$$\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t\widetilde{\sigma}_t(\mathbf{x}),$$

Deterministic Training Conditional (DTC) sparse GP-UCB

[Seeger et al., 2003]

Select $\widetilde{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x} \in \mathbf{X}_A} \widetilde{u}_t(\mathbf{x})$

$$\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t \widetilde{\sigma}_t(\mathbf{x}),$$

$$\begin{aligned}\widetilde{\mu}_t(\mathbf{x}) &\triangleq \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal (\widetilde{\mathbf{K}}_t + \lambda \mathbf{I})^{-1} \mathbf{y}_t \\ &= \mathbf{z}(\mathbf{x})^\intercal (\mathbf{Z}_t^\intercal \mathbf{Z}_t + \lambda \mathbf{I})^{-1} \mathbf{Z}_t^\intercal \mathbf{y}_t,\end{aligned}$$

Deterministic Training Conditional (DTC) sparse GP-UCB

[Seeger et al., 2003]

Select $\widetilde{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathbf{X}_A} \widetilde{u}_t(\mathbf{x})$

$$\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t\widetilde{\sigma}_t(\mathbf{x}),$$

$$\begin{aligned}\widetilde{\mu}_t(\mathbf{x}) &\triangleq \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\mathbf{y}_t \\ &= \mathbf{z}(\mathbf{x})^\mathsf{T}(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}\mathbf{Z}_t^\mathsf{T}\mathbf{y}_t,\end{aligned}$$

$$\begin{aligned}\widetilde{\sigma}_t^2(\mathbf{x}) &\triangleq \frac{1}{\lambda}\left(k(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})\right) \\ &= \frac{1}{\lambda}\left(k(\mathbf{x},\mathbf{x}) - \mathbf{z}(\mathbf{x})^\mathsf{T}\underbrace{\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}}_{m\times m\text{ matrix}}\mathbf{z}(\mathbf{x})\right),\end{aligned}$$

Deterministic Training Conditional (DTC) sparse GP-UCB

[Seeger et al., 2003]

Select $\widetilde{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathbf{X}_A} \widetilde{u}_t(\mathbf{x})$

$$\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t\widetilde{\sigma}_t(\mathbf{x}),$$

$$\begin{aligned}\widetilde{\mu}_t(\mathbf{x}) &\triangleq \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\mathbf{y}_t \\ &= \mathbf{z}(\mathbf{x})^\mathsf{T}(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}\mathbf{Z}_t^\mathsf{T}\mathbf{y}_t, \qquad \mathcal{O}(m)\end{aligned}$$

$$\begin{aligned}\widetilde{\sigma}_t^2(\mathbf{x}) &\triangleq \tfrac{1}{\lambda}\Big(k(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})\Big) \\ &= \tfrac{1}{\lambda}\Big(k(\mathbf{x},\mathbf{x}) - \mathbf{z}(\mathbf{x})^\mathsf{T}\underbrace{\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}}_{m \times m \text{ matrix}}\mathbf{z}(\mathbf{x})\Big), \qquad \mathcal{O}(m^2)\end{aligned}$$

Efficient: $\mathcal{O}(Am^2 + m^3)$ per step

Deterministic Training Conditional (DTC) sparse GP-UCB

[Seeger et al., 2003]

Select $\widetilde{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathbf{X}_A} \widetilde{u}_t(\mathbf{x})$

$$\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t \widetilde{\sigma}_t(\mathbf{x}),$$

$$\begin{aligned}\widetilde{\mu}_t(\mathbf{x}) &\triangleq \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\mathbf{y}_t \\ &= \mathbf{z}(\mathbf{x})^\mathsf{T}(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}\mathbf{Z}_t^\mathsf{T}\mathbf{y}_t, \qquad \mathcal{O}(m)\end{aligned}$$

$$\begin{aligned}\widetilde{\sigma}_t^2(\mathbf{x}) &\triangleq \frac{1}{\lambda}\Big(k(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\mathsf{T}(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})\Big) \\ &= \frac{1}{\lambda}\Big(k(\mathbf{x},\mathbf{x}) - \mathbf{z}(\mathbf{x})^\mathsf{T}\underbrace{\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t(\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t + \lambda\mathbf{I})^{-1}}_{m \times m \text{ matrix}}\mathbf{z}(\mathbf{x})\Big), \qquad \mathcal{O}(m^2)\end{aligned}$$

Efficient: $\mathcal{O}(Am^2 + m^3)$ per step

How to choose $\mathcal{S}$ for good accuracy?

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
$\hookrightarrow$ $\mathcal{S}_t$ must change with t

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
↳ $\mathcal{S}_t$ must change with t

Accuracy-efficiency tradeoff of m
↳ adaptively resize $\mathcal{S}_t$

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
↳ $\mathcal{S}_t$ must change with t

Accuracy-efficiency tradeoff of m
↳ adaptively resize $\mathcal{S}_t$

$\sigma_t^2(\cdot)$ captures informative arms
↳ include $\mathbf{x}_i$ with large $\sigma_t^2(\mathbf{x}_i)$

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
↳ $\mathcal{S}_t$ must change with t

Accuracy-efficiency tradeoff of m
↳ adaptively resize $\mathcal{S}_t$

$\sigma_t^2(\cdot)$ captures informative arms
↳ include $\mathbf{x}_i$ with large $\sigma_t^2(\mathbf{x}_i)$

Greedy inclusion hard to analyze
↳ random inclusion $p_{t,i} \propto \sigma_t^2(\cdot)$

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
↳ $\mathcal{S}_t$ must change with t

Accuracy-efficiency tradeoff of m
↳ adaptively resize $\mathcal{S}_t$

$\sigma_t^2(\cdot)$ captures informative arms
↳ include $\mathbf{x}_i$ with large $\sigma_t^2(\mathbf{x}_i)$

Greedy inclusion hard to analyze
↳ random inclusion $p_{t,i} \propto \sigma_t^2(\cdot)$

$\sigma_t^2(\cdot)$ expensive to compute
↳ approximate $\sigma_t^2(\cdot) \approx \widetilde{\sigma}_t^2(\cdot)$

Budgeted Kernelized Bandits

$\mathbf{X}_t$ changes over time
↳ $\mathcal{S}_t$ must change with t

Accuracy-efficiency tradeoff of m
↳ adaptively resize $\mathcal{S}_t$

$\sigma_t^2(\cdot)$ captures informative arms
↳ include $\mathbf{x}_i$ with large $\sigma_t^2(\mathbf{x}_i)$

Greedy inclusion hard to analyze
↳ random inclusion $p_{t,i} \propto \sigma_t^2(\cdot)$

$\sigma_t^2(\cdot)$ expensive to compute
↳ approximate $\sigma_t^2(\cdot) \approx \widetilde{\sigma}_t^2(\cdot)$

Algorithm 6: BKB

Data: Arm set $\mathcal{A}$, q, $\{\beta_t\}_{t=1}^T$
Result: Arm choices $\mathcal{D}_T \leftarrow \{(\widetilde{\mathbf{x}}_t, y_t)\}$
Select uniformly at random $\mathbf{x}_1$;
Observe y_1;
Initialize $\mathcal{S}_1 \leftarrow \{\mathbf{x}_1\}$;
for $t = \{1, \dots, T-1\}$ **do**
    Compute $\widetilde{\mu}_t(\mathbf{x}_i)$ and $\widetilde{\sigma}_t^2(\mathbf{x}_i)$ for all $\mathbf{x}_i \in \mathcal{A}$;
    Select $\widetilde{\mathbf{x}}_{t+1} \leftarrow \arg\max_{\mathbf{x}_i \in \mathcal{A}} \widetilde{u}_t(\mathbf{x}_i)$;
    for $i = \{1, \dots, t+1\}$ **do**
        Set $\widetilde{p}_{t+1,i} \leftarrow \overline{q} \cdot \widetilde{\sigma}_t^2(\widetilde{\mathbf{x}}_i)$;
        Draw $q_{t+1,i} \sim$ *Bernoulli* $(\widetilde{p}_{t+1,i})$;
        If $q_{t+1} = 1$ **then** include $\widetilde{\mathbf{x}}_i$ in $\mathcal{S}_{t+1}$;
    end
end

Measuring the complexity of GP optimization

Maximum information gain [Srinivas et al., 2010]

$$\gamma_T \triangleq \max_{\mathcal{D} \subset \mathcal{A} : |\mathcal{D}| = T} \frac{1}{2} \log\det(\mathbf{K}_{\mathcal{D}}/\lambda + \mathbf{I}).$$

Measuring the complexity of GP optimization

Maximum information gain [Srinivas et al., 2010]

$$\gamma_T \triangleq \max_{\mathcal{D}\subset\mathcal{A}:|\mathcal{D}|=T} \frac{1}{2}\log\det(\mathbf{K}_{\mathcal{D}}/\lambda + \mathbf{I}).$$

Effective dimension (a.k.a effective rank) [Alaoui and Mahoney, 2015]

$$d_{\text{eff}}(\lambda, \widetilde{\mathbf{X}}_T) \triangleq \sum_{i=1}^{T} \sigma_T^2(\widetilde{\mathbf{x}}_i) = \text{Tr}(\mathbf{K}_T(\mathbf{K}_T + \lambda\mathbf{I})^{-1})$$

Measuring the complexity of GP optimization

Maximum information gain [Srinivas et al., 2010]

$$\gamma_T \triangleq \max_{\mathcal{D}\subset\mathcal{A}:|\mathcal{D}|=T} \frac{1}{2}\log\det(\mathbf{K}_{\mathcal{D}}/\lambda + \mathbf{I}).$$

Effective dimension (a.k.a effective rank) [Alaoui and Mahoney, 2015]

$$d_{\text{eff}}(\lambda, \widetilde{\mathbf{X}}_T) \triangleq \sum_{i=1}^{T} \sigma_T^2(\widetilde{\mathbf{x}}_i) = \operatorname{Tr}(\mathbf{K}_T(\mathbf{K}_T + \lambda\mathbf{I})^{-1})$$

From γ_T to $d_{\text{eff}}(\lambda, \widetilde{\mathbf{X}}_T)$ [Calandriello et al., 2017]

$$\log\det\left(\mathbf{K}_T/\lambda + \mathbf{I}\right) \leq 2d_{\text{eff}}(\lambda, \widetilde{\mathbf{X}}_T)\log\left(T/\lambda\right) \ll 2\gamma_T\log(T/\lambda).$$

Accuracy and computational guarantees

Theorem

With probability $1-\delta$*, for all* $t \in [T]$ *and all* $\mathbf{x} \in \mathcal{A}$*, we have*

$$\sigma_t^2(\mathbf{x})/2 \leq \widetilde{\sigma}_t^2(\mathbf{x}) \leq 2\sigma_t^2(\mathbf{x}) \quad \textit{and} \quad |\mathcal{S}_t| \leq \mathcal{O}(d_{\mathit{eff}}(\lambda, \widetilde{\mathbf{X}}_t)\log(t/\delta)).$$

Accuracy and computational guarantees

Theorem

With probability $1-\delta$, for all $t \in [T]$ and all $\mathbf{x} \in \mathcal{A}$, we have

$$\sigma_t^2(\mathbf{x})/2 \leq \widetilde{\sigma}_t^2(\mathbf{x}) \leq 2\sigma_t^2(\mathbf{x}) \quad \textit{and} \quad |\mathcal{S}_t| \leq \mathcal{O}(d_{\mathit{eff}}(\lambda, \widetilde{\mathbf{X}}_t)\log(t/\delta)).$$

Note that $d_{\text{eff}} \leq \gamma_T$, when $\gamma_T \ll T$

Time: near-constant $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2 + d_{\text{eff}}^3) \leq \widetilde{\mathcal{O}}(A\gamma_T^2 + \gamma_T^3)$ per-step

Space: near-constant $\widetilde{\mathcal{O}}(d_{\text{eff}}^2) \leq \widetilde{\mathcal{O}}(\gamma_T^2)$

Accuracy and computational guarantees

Theorem

With probability $1-\delta$*, for all* $t \in [T]$ *and all* $\mathbf{x} \in \mathcal{A}$*, we have*

$$\sigma_t^2(\mathbf{x})/2 \leq \widetilde{\sigma}_t^2(\mathbf{x}) \leq 2\sigma_t^2(\mathbf{x}) \quad \textit{and} \quad |\mathcal{S}_t| \leq \mathcal{O}(d_{\mathit{eff}}(\lambda, \widetilde{\mathbf{X}}_t)\log(t/\delta)).$$

Note that $d_{\text{eff}} \leq \gamma_T$, when $\gamma_T \ll T$

Time: near-constant $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2 + d_{\text{eff}}^3) \leq \widetilde{\mathcal{O}}(A\gamma_T^2 + \gamma_T^3)$ per-step

Space: near-constant $\widetilde{\mathcal{O}}(d_{\text{eff}}^2) \leq \widetilde{\mathcal{O}}(\gamma_T^2)$

$\widetilde{\sigma}_t(\cdot)$ always close to $\sigma_t(\cdot)$: no variance starvation
↳ previously only k stationary and $|\mathcal{S}_t| \approx \mathcal{O}(\log(t)^d) \gg \mathcal{O}(d_{\text{eff}}\log(t))$

Accuracy and computational guarantees

Theorem

With probability $1-\delta$, for all $t \in [T]$ and all $\mathbf{x} \in \mathcal{A}$, we have

$$\sigma_t^2(\mathbf{x})/2 \leq \widetilde{\sigma}_t^2(\mathbf{x}) \leq 2\sigma_t^2(\mathbf{x}) \quad \textit{and} \quad |\mathcal{S}_t| \leq \mathcal{O}(d_{\mathit{eff}}(\lambda, \widetilde{\mathbf{X}}_t)\log(t/\delta)).$$

Note that $d_{\text{eff}} \leq \gamma_T$, when $\gamma_T \ll T$

Time: near-constant $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2 + d_{\text{eff}}^3) \leq \widetilde{\mathcal{O}}(A\gamma_T^2 + \gamma_T^3)$ per-step

Space: near-constant $\widetilde{\mathcal{O}}(d_{\text{eff}}^2) \leq \widetilde{\mathcal{O}}(\gamma_T^2)$

$\widetilde{\sigma}_t(\cdot)$ always close to $\sigma_t(\cdot)$: no variance starvation

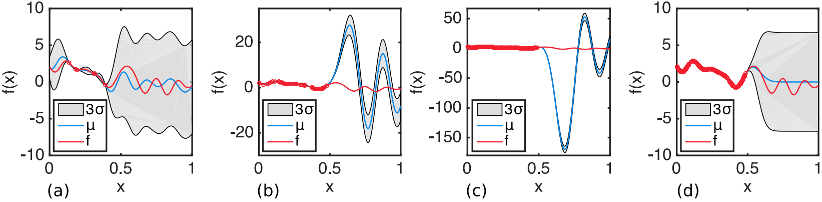
↳ previously only k stationary and $|\mathcal{S}_t| \approx \mathcal{O}(\log(t)^d) \gg \mathcal{O}(d_{\text{eff}}\log(t))$

Proof: $\sigma_t(\mathbf{x}_i)$ is the λ-ridge leverage score of $\mathbf{x}_i$ w.r.t. $k(\cdot,\cdot)$ and $\mathbf{X}_t$

↳ we can leverage literature on leverage score sampling

Variance starvation

Problem: hard to judge negative correlation far from $\mathcal{S}$ [Wang et al., 2018]



Fixed-rank sparse GPs become overconfident when $n \gg m$

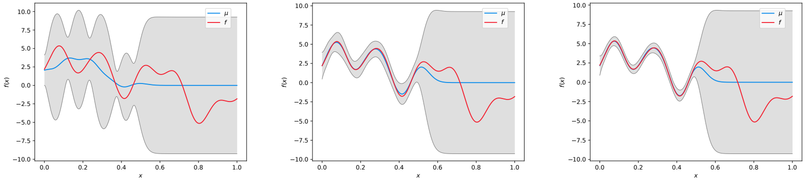
Prior approaches to avoid variance starvation:
[Huggins et al., 2019; Mutny and Krause, 2018]

Require stationary k and/or additive kernel

Build ε-grid of the space, $\exp\{d\}$ dependencies

Variance starvation

Solution: BKB adaptively matches sparse GP rank and d_{eff}



DTC approximation also crucial to be accurate RLS estimator

No need for ε-grid, focus on essential parts of $\mathbf{X}_t$

Regret guarantees

Theorem

If we run BKB *with* $\widetilde{\beta}_t \triangleq 2\xi\sqrt{\left(\sum_{s=1}^{t} \widetilde{\sigma}_t^2(\widetilde{\mathbf{x}}_s)\right)\log(t) + \log(1/\delta)} + 3\sqrt{\lambda}F$, *then, with probability of at least* $1-\delta$,

$$R_T^{\mathrm{BKB}} \leq 32\sqrt{T}\left(\xi d_{\mathit{eff}}\log(T) + \sqrt{\lambda F^2 d_{\mathit{eff}}\log(T)} + \xi\log(1/\delta)\right)$$

$R_T^{\mathrm{BKB}} \leq 16 R_T^{\mathrm{GP\text{-}UCB}} \log(T)$: no-regret

$\widetilde{\beta}_t$ computable in $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2)$ time

No assumptions on k, $\mathcal{A}$

DTC is not a GP (not consistent), but now a justified heuristic

No free lunch: learning complexity is computational complexity

Related results

Same regret as $\mathrm{GP\text{-}UCB}$, but improve from $\widetilde{\mathcal{O}}(At^2)$ time to $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2)$

Related results

Same regret as $\mathrm{GP\text{-}UCB}$, but improve from $\widetilde{\mathcal{O}}(At^2)$ time to $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2)$

Vs. methods without regret guarantees:

[Huggins et al., 2019; Wilson and Nickisch, 2015]

↳ same sparsity level $\widetilde{\mathcal{O}}(d_{\text{eff}}) \approx \widetilde{\mathcal{O}}(\gamma_T)$ for generic k

Related results

Same regret as $\mathrm{GP\text{-}UCB}$, but improve from $\widetilde{\mathcal{O}}(At^2)$ time to $\widetilde{\mathcal{O}}(Ad_{\mathrm{eff}}^2)$

Vs. methods without regret guarantees:

[Huggins et al., 2019; Wilson and Nickisch, 2015]

↳ same sparsity level $\widetilde{\mathcal{O}}(d_{\mathrm{eff}}) \approx \widetilde{\mathcal{O}}(\gamma_T)$ for generic k

Vs. scalable methods with regret guarantees:

Thompson sampling with quadrature RFF (GP-Opt) [Mutny and Krause, 2018]

↳ small d: same sparsity level and regret, generic k

large d: no need for ε-grid, no $\exp\{d\}$ dependency

OFUL with Frequent Direction sketch (Linear Bandit) [Kuzborskij et al., 2019]

↳ same sparsity level and lower regret

Related results

Same regret as $\mathrm{GP\text{-}UCB}$, but improve from $\widetilde{\mathcal{O}}(At^2)$ time to $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2)$

Vs. methods without regret guarantees:

[Huggins et al., 2019; Wilson and Nickisch, 2015]

↳ same sparsity level $\widetilde{\mathcal{O}}(d_{\text{eff}}) \approx \widetilde{\mathcal{O}}(\gamma_T)$ for generic k

Vs. scalable methods with regret guarantees:

Thompson sampling with quadrature RFF (GP-Opt) [Mutny and Krause, 2018]

↳ small d: same sparsity level and regret, generic k

large d: no need for ε-grid, no $\exp\{d\}$ dependency

OFUL with Frequent Direction sketch (Linear Bandit) [Kuzborskij et al., 2019]

↳ same sparsity level and lower regret

QFF/VI based methods can exploit kernel additivity:[Huggins et al., 2019]

↳ TS-QFF can optimize exactly posterior for small d

can BKB for small m do the same?

Proof sketch

DTC is also known as projected process approximation
↳ show equivalence to projected OFUL [Abbasi-Yadkori et al., 2011]

Proof sketch

DTC is also known as projected process approximation
↳ show equivalence to projected OFUL [Abbasi-Yadkori et al., 2011]

Orthogonal projection $\mathbf{P}_t$ on $\mathrm{Span}(\mathcal{S}_t)$ regularizes
↳ reduces variance but introduces extra bias $\|(\mathbf{I} - \mathbf{P}_t)\mathbf{K}_t^{1/2} f\|^2$

Proof sketch

DTC is also known as projected process approximation
↳ show equivalence to projected OFUL [Abbasi-Yadkori et al., 2011]

Orthogonal projection $\mathbf{P}_t$ on Span($\mathcal{S}_t$) regularizes
↳ reduces variance but introduces extra bias $\|(\mathbf{I} - \mathbf{P}_t)\mathbf{K}_t^{1/2} f\|^2$

Lemma

When $\mathcal{S}_t$ sampled according to RLS $\mathbf{I} - \mathbf{P}_t \preceq (1+\varepsilon)\lambda(\mathbf{K}_t + \lambda\mathbf{I})^{-1}$

Proof sketch

DTC is also known as projected process approximation
↳ show equivalence to projected OFUL [Abbasi-Yadkori et al., 2011]

Orthogonal projection $\mathbf{P}_t$ on $\mathrm{Span}(\mathcal{S}_t)$ regularizes
↳ reduces variance but introduces extra bias $\|(\mathbf{I} - \mathbf{P}_t)\mathbf{K}_t^{1/2} f\|^2$

Lemma

When $\mathcal{S}_t$ sampled according to RLS $\mathbf{I} - \mathbf{P}_t \preceq (1+\varepsilon)\lambda(\mathbf{K}_t + \lambda\mathbf{I})^{-1}$

self-normalized bias

$$\|(\mathbf{I} - \mathbf{P}_t)\mathbf{K}_t^{1/2} f\|^2 \leq (1+\varepsilon)\lambda\|(\mathbf{K}_t + \lambda\mathbf{I})^{-1/2}\mathbf{K}_t^{1/2} f\| \leq (1+\varepsilon)\lambda\|f\|$$

Proof sketch

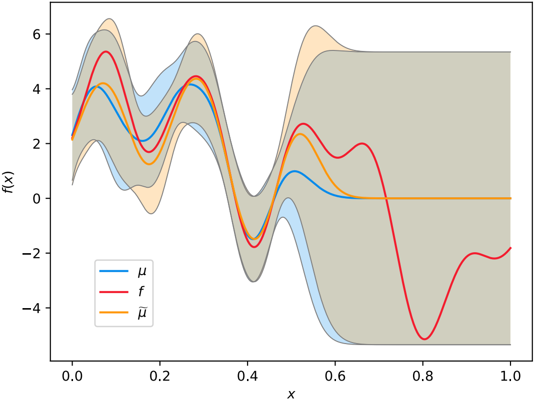
BKB is not simply a $\mathrm{GP\text{-}UCB}$ approximation

Confidence intervals

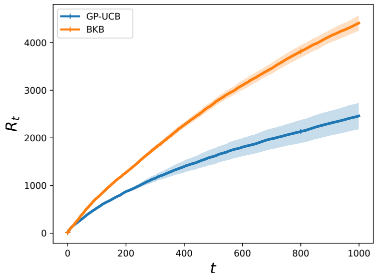
$$C_t = [\mu_t(\mathbf{x}) \pm \beta_t \sigma_t(\mathbf{x})],$$

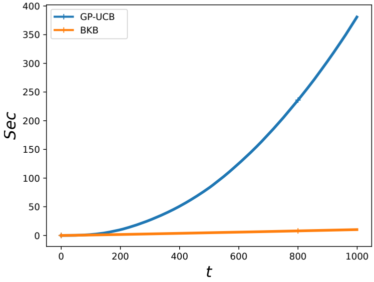
$$\widetilde{C}_t = [\widetilde{\mu}_t(\mathbf{x}) \pm \widetilde{\beta}_t \widetilde{\sigma}_t(\mathbf{x})]$$

$$C_t \not\subset \widetilde{C}_t, \quad \widetilde{C}_t \not\subset C_t$$



Experiments





Dataset: Cadata ($A \approx 10^4$), Kernel: RBF with $\sigma^2 = 5$