Gaussian Process Optimization with Adaptive Sketching: Scalable and No Regret

Daniele Calandriello*, Luigi Carratino†, Alessandro Lazaric‡, Michal Valko§, Lorenzo Rosasco*,†,¶
*LCSL - Istituto Italiano di Tecnologia, †Università degli Studi di Genova, ‡Facebook AI Research Paris,
§INRIA Lille - Nord Europe, ¶Massachusetts Institute of Technology

In a nutshell

Gaussian process-UCB (GP-UCB) is a popular Bayesian/bandit optimization alternative to grad student descent. However, it requires $\mathcal{O}(T^3)$ time and $\mathcal{O}(T^2)$ space and does not scale.

We introduce the first general GP optimization algorithm (BKB) that is no regret and provably scalable, with near-linear runtime $\mathcal{O}(Td_{\text{eff}}^2)$. It also maintains valid posterior variance estimates at all steps, while previous approaches could under/over-estimate the confidence intervals of the GP. BKB main ingredient is a novel adaptive selection of inducing point using approximate posterior variance sampling.

Gaussian process optimization

Arms $\mathcal{A} = \{\mathbf{x}_i\}_{i=1}^{A}$ with $\mathbf{x}_i \in \mathbb{R}^d$, similarity (kernel) $\mathrm{k}(\cdot,\cdot)$ and RKHS $\mathcal{H}$

For $t \in [1,\dots,T]$: (1) select $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}_i} u_t(\mathbf{x}_i)$
(2) Receive noisy feedback $y_{t+1} = f(\mathbf{x}_{t+1}) + \eta_{t+1}$
(3) Improve u_{t+1} for next time

Goal: minimize regret $R_T = \sum_{t=1}^{T} f(\mathbf{x}_*) - f(\mathbf{x}_t)$ vs. $\mathbf{x}_* = \arg\max_{\mathbf{x}_i} f(\mathbf{x}_i)$
Assumption: $f \in \mathcal{H}$ arbitrary but $\|f\| \le F$ (frequentist/bandit regret)

Measuring the complexity of a GP

Maximum information gain: $\gamma_T \triangleq \max_{\mathcal{D}\subset\mathcal{A}:|\mathcal{D}|=T} \log\det(\mathbf{K}_{\mathcal{D}}/\lambda + \mathbf{I})$
Effective dimension/rank: $d_{\text{eff}} \triangleq \sum_{i=1}^{T} \sigma_T^2(\widetilde{\mathbf{x}}_i)$
From γ_T to d_{eff}: $\log\det(\mathbf{K}_T + \mathbf{I}) \le d_{\text{eff}} \log(T) \ll \gamma_T \log(T)$

GP-UCB and sparse GPs

GP-UCB: $u_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta_t\sigma_t(\mathbf{x})$

$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^\intercal(\mathbf{K}_t + \lambda\mathbf{I})^{-1}\mathbf{y}_t$

$\sigma_t^2(\mathbf{x}) = \frac{1}{\lambda}\left(\mathrm{k}(\mathbf{x},\mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\intercal(\mathbf{K}_t + \lambda\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x})\right)$

No regret $R_T^{\text{GP-UCB}} \le \widetilde{\mathcal{O}}(\sqrt{T}(\gamma_T + \sqrt{\gamma_T F}))$ but too slow $\mathcal{O}(At^2)$ per step

Sparse GPs: given m inducing points $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^{m}$ (a.k.a. dictionary) replace $\mathrm{k}(\mathbf{x}_i,\mathbf{x}_j)$ with $\widetilde{\mathrm{k}}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{k}_{\mathcal{S}}(\mathbf{x}_i)^\intercal\mathbf{K}_{\mathcal{S}}^{+}\mathbf{k}_{\mathcal{S}}(\mathbf{x}_j)$

GP-UCB + DTC: $\widetilde{u}_t(\mathbf{x}) = \widetilde{\mu}_t(\mathbf{x}) + \widetilde{\beta}_t\widetilde{\sigma}_t(\mathbf{x})$

$\widetilde{\mu}_t(\mathbf{x}) = \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\mathbf{y}_t$

$\widetilde{\sigma}_t^2(\mathbf{x}) = \frac{1}{\lambda}\left(\mathrm{k}(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})\right)$

Deterministic training conditional (DTC) a.k.a. projected GP
↳ $\mathcal{O}(Am^2 + m^3)$ per step since $\mathrm{Rank}(\widetilde{\mathbf{K}}_t) = m$ but is it no regret?

Problem: how to choose $\mathcal{S}$ for good accuracy/regret?

Budgeted Kernelized Bandits (BKB)

$\mathbf{X}_t$ 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 $p_{t,i} \propto \sigma_t^2(\cdot)$

for $t = \{1,\dots,T-1\}$ **do**
  Compute $\widetilde{\mu}_t(\mathbf{x}_i)$ and $\widetilde{\sigma}_t^2(\mathbf{x}_i)$ for all $\mathbf{x}_i$;
  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

Main result: BKB is scalable and no regret

Theorem: Let $\widetilde{\beta}_t \triangleq 2\sqrt{\left(\sum_{s=1}^{t}\widetilde{\sigma}_t^2(\widetilde{\mathbf{x}}_s)\right)\log(t) + \log(1/\delta)} + 3\sqrt{\lambda}F$. Then w.p. $1-\delta$, for all $t \in [T]$ and all $\mathbf{x} \in \mathcal{A}$, we have

$$\sigma_t^2(\mathbf{x})/2 \le \widetilde{\sigma}_t^2(\mathbf{x}) \le 2\sigma_t^2(\mathbf{x}) \quad \text{and} \quad |\mathcal{S}_t| \le \mathcal{O}(d_{\text{eff}}\log(t/\delta)),$$

and BKB suffers at most regret

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

$R_T^{\text{BKB}} \le 16 R_T^{\text{GP-UCB}}\log(T)$: no regret but only $\widetilde{\mathcal{O}}(TAd_{\text{eff}}^2)$ time!

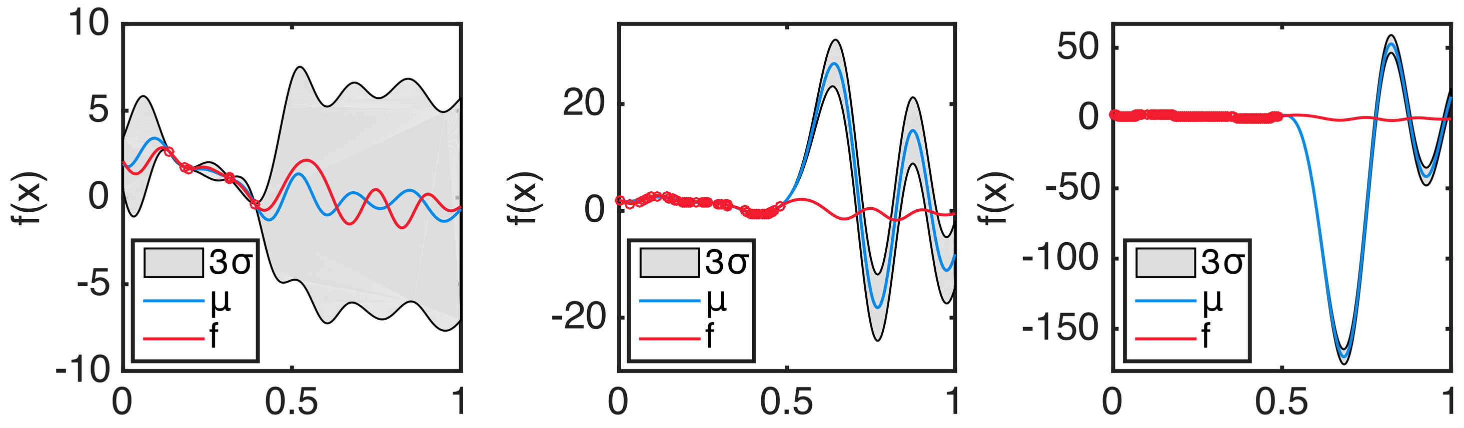
- $\widetilde{\beta}_t$ computable in $\widetilde{\mathcal{O}}(Ad_{\text{eff}}^2)$ time replacing worst-case bounds on γ_T
- No assumptions on k (e.g., not only stationary k)
- No free lunch: worst-case falls back to GP-UCB
- Not incremental: have to recompute $\mathcal{S}_t$ at each step

Alg.	$\mathrm{k}(\cdot,\cdot)$	m	$R_T/R_T^{\text{GP-UCB}}$
TS-QFF [2]	stationary	$2^d\gamma_T$	16
SOFUL [3]	linear	k	$1 + \sum_{i=k+1}^{T}\lambda_i(\mathbf{K}_T)$
BKB	any	d_{eff}	$16\log(T)$

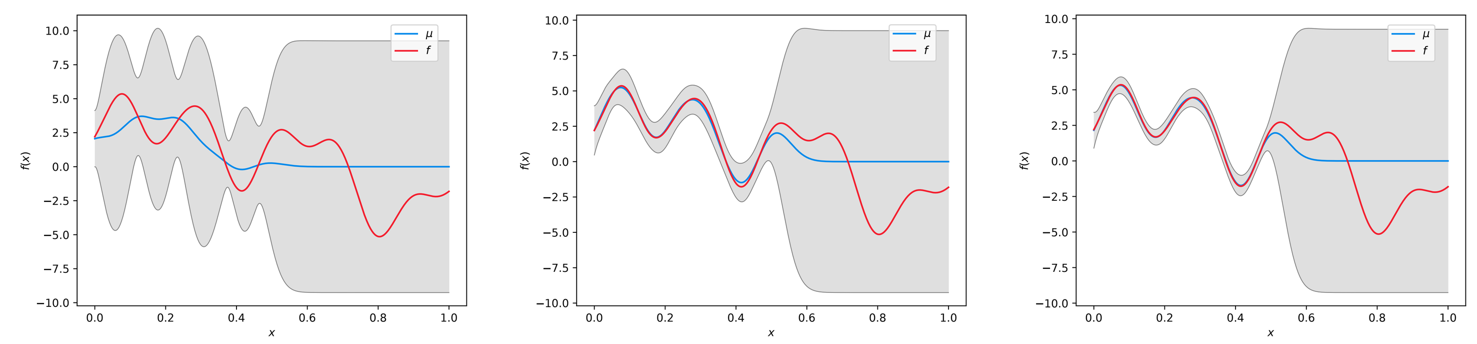
DTC is not a GP (not consistent), but now a justified heuristic
Easy extension to infinite $\mathcal{A}$, but how to optimize posterior?

No variance starvation

Sparse GPs become over/underconfident when $d_{\text{eff}} \gg m$ [5]



BKB adaptively matches GP rank m and d_{eff}



Alg.	$\widetilde{\sigma}_t(\cdot)$	$\widetilde{k}(\cdot,\cdot)$	worst case
SoR	$\widetilde{\mathrm{k}}(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})$	fixed	0
DTC	$\mathrm{k}(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})$	fixed	$\mathrm{k}(\mathbf{x},\mathbf{x})$
BKB	$\mathrm{k}(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_t(\mathbf{x})^\intercal(\widetilde{\mathbf{K}}_t + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{k}}_t(\mathbf{x})$	adaptive	$(1 \pm 1/2)\sigma_t(\cdot)$

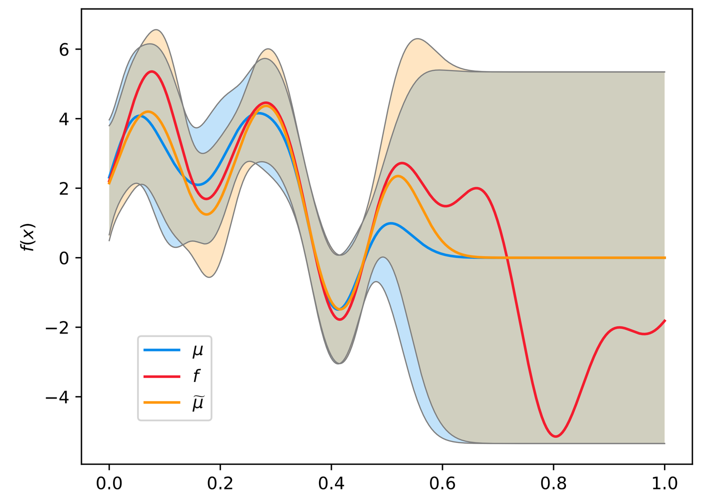
Not simply an approximate GP-UCB

Confidence intervals

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

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

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



Analysis: OFUL [4] + regularization with projection $\mathbf{P}_t$ on $\mathrm{Im}(\mathcal{S}_t)$

$$\widetilde{\beta}_t \le \underbrace{\lambda\|f\|}_{\text{regular. bias}} + \underbrace{\eta_t^\intercal\mathbf{P}_t\mathbf{K}_t\mathbf{P}_t(\mathbf{P}_t\mathbf{K}_t\mathbf{P}_t + \lambda\mathbf{I})^{-1}\eta_t}_{\text{reduced variance}} + \underbrace{\|(\mathbf{I}-\mathbf{P}_t)\mathbf{K}_t^{1/2}f\|}_{\text{extra approximation bias}}$$

Lemma: $\mathbf{I} - \mathbf{P}_t \preceq (1+\varepsilon)\lambda(\mathbf{K}_t + \lambda\mathbf{I})^{-1}$ when $\mathcal{S}_t$ sampled $\sigma_t(\cdot)$

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

Non-uniform error: self-normalized bias focuses on essential parts of $\mathcal{A}$
no need for uniform bounds, ε-grids, and $\exp\{d\}$ dependencies

[1] Srinivas et al. Gaussian process optimization in the bandit setting: No regret and experimental design. ICML 2010 [2] Mutný et al. Efficient high-dimensional Bayesian optimization with additivity and quadrature Fourier features. NeurIPS 2018
[3] Kuzborskij et al. Efficient linear bandits through matrix sketching. AISTATS 2019 [4] Abbasi-Yadkori et al. Improved algorithms for linear stochastic bandits. NeurIPS 2011 [5] Wang et al. Batched Large-scale Bayesian Optimization in High-dimensional Spaces. AISTATS 2018