UniGe | MaLGa

Scalable Gaussian Process Optimization on Continuous Domain by Adaptive Discretization

Luigi Carratino
Università degli Studi di Genova

joint work with:
Daniele Calandriello, Alessandro Lazaric, Marco Rando, Lorenzo Rosasco,
Michal Valko, Silvia Villa

Problem Setup

Let (X, d) be a compact metric space (e.g. $X \subseteq \mathbb{R}^p$).
Let $f : X \to \mathbb{R}$ be an unknown continuous function.

Problem Setup

Let (X, d) be a compact metric space (e.g. $X \subseteq \mathbb{R}^p$).
Let $f : X \to \mathbb{R}$ be an unknown continuous function.

Find

$$x^* \in \operatorname*{argmax}_{x \in X} f(x),$$

with access to only sequential noisy function evaluations

$$y_t = f(x_t) + \epsilon_t, \qquad t = 1, 2, \ldots$$

Problem Setup

Given a budget $T \in \mathbb{N}$, select

$$x_1, \ldots, x_T \in X$$

with small cumulative regret

$$R_T = \sum_{t=1}^{T} (f(x^*) - f(x_t)).$$

Problem Setup

Given a budget $T \in \mathbb{N}$, select

$$x_1, \dots, x_T \in X$$

with small cumulative regret

$$R_T = \sum_{t=1}^{T} (f(x^*) - f(x_t)).$$

Assumptions:

- $f \in \mathcal{H}$, RKHS with p.d. kernel $k(\cdot, \cdot)$
- $\|f\| = F \leqslant \infty$

Bandits Optimization

For $t = 1, \ldots, T$:

(1) Select x_t
(2) Receive noisy feedback $y_t = f(x_t) + \epsilon_t$
(3) Improve for next time

Bandits Optimization

For $t = 1, \ldots, T$:

(1) Select x_t
(2) Receive noisy feedback $y_t = f(x_t) + \epsilon_t$
(3) Improve for next time

How do we model f ?

How do we improve over time?

Bandits Optimization

For $t = 1, \ldots, T$:

(1) Select x_t
(2) Receive noisy feedback $y_t = f(x_t) + \epsilon_t$
(3) Improve for next time

How do we EFFICIENTLY model f ?

How do we improve over time?

Bandits Optimization

For $t = 1, \ldots, T$:

(1) Select x_t
(2) Receive noisy feedback $y_t = f(x_t) + \epsilon_t$
(3) Improve for next time

How do we EFFICIENTLY model f ?

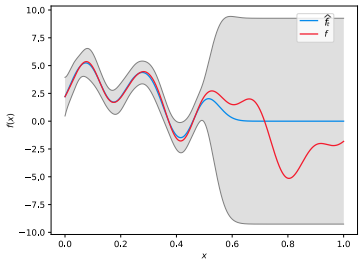
How do we EFFICIENTLY improve over time?

Kernel Ridge Regression / Gaussian Process

Given $(x_i, y_i)_{i=1}^t$, let $\widehat{K}_t \in \mathbb{R}^{t \times t}$ s.t. $(\widehat{K})_{i,j} = k(x_i, x_j)$ and $\widehat{k}_t(x) = [k(x_1, x), \dots, k(x_t, x)]^\top$, for $\lambda > 0$

$$\widehat{f}_t(x) = \widehat{k}_t(x)^\top (\widehat{K}_t + \lambda I)^{-1} \widehat{y}$$

$$\sigma_t^2(x) = k(x, x) - \widehat{k}_t(x)^\top (\widehat{K}_t + \lambda I)^{-1} \widehat{k}_t(x)$$



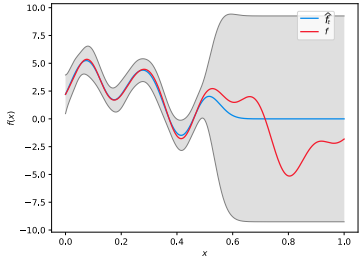
Nyström Approximation / Sparse GP

Nyström equivalent formulation:

Let $S = (\widetilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t} \rightarrow$ Nyström approximation $\widetilde{k}(x, x') = \widehat{k}_S(x)^\top \widehat{K}_S^\dagger \widehat{k}_S(x')$,
with $\widehat{K}_S \in \mathbb{R}^{M \times M}$ s.t. $(\widehat{K}_S)_{i,j} = k(\widetilde{x}_i, \widetilde{x}_j)$ and $\widehat{k}_S(x) = [k(x_1, x), \dots, k(x_M, x)]^\top$

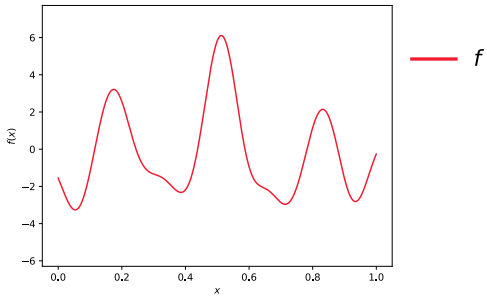
$$\widetilde{f}_t(x) = \widetilde{k}_t(x)^\top (\widetilde{K}_t + \lambda I)^{-1} \widehat{y}$$

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



BKB

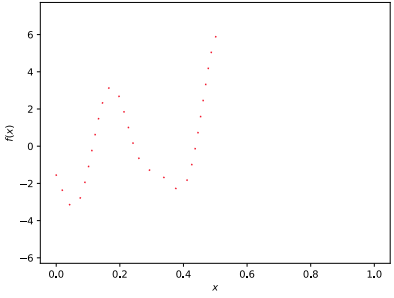
Let $f : X \to \mathbb{R}$ unknown function



(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points



f

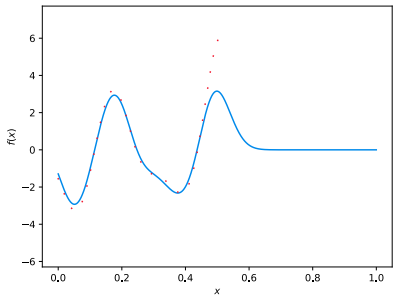
for $t = \{1, \ldots, T-1\}$ **do**

end

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points , with $S_t = (\widetilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t}$



$\widehat{f}_t$

f

for $t = \{1, \ldots, T-1\}$ **do**

end

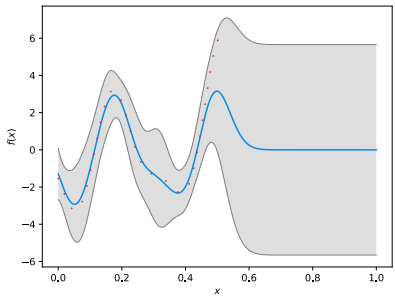
$$\widetilde{f}_t(x) = \widetilde{k}_t(x)^\top (\widetilde{K}_t + \lambda I)^{-1} \widehat{y}$$

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

UniGe | MaLGa

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points , with $S_t = (\widetilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t}$



$\widehat{f}_t$

f

for $t = \{1, \ldots, T-1\}$ **do**

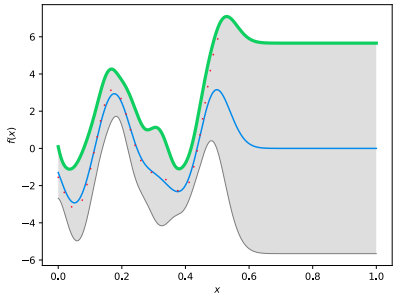
end

$$\widetilde{f}_t(x) = \widetilde{k}_t(x)^\top (\widetilde{K}_t + \lambda I)^{-1} \widehat{y} \qquad \widetilde{\sigma}_t^2(x) = \frac{1}{\lambda}\Big(k(x,x) - \widetilde{k}_t(x)^\top (\widetilde{K}_t + \lambda I)^{-1} \widetilde{k}_t(x)\Big)$$

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points , with $S_t = (\tilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t}$



$\hat{f}_t$
f
u

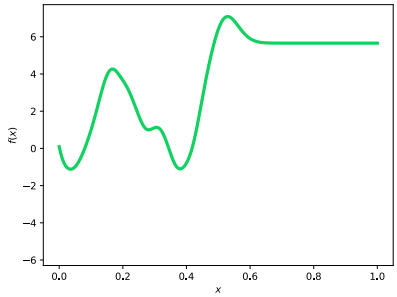
for $t = \{1, \ldots, T-1\}$ **do**
  Compute u_t
end

$$u_t(x) = \tilde{f}_t(x) + \beta_t \tilde{\sigma}_t(x)$$

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points , with $S_t = (\widetilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t}$



for $t = \{1, \ldots, T-1\}$ **do**

Compute u_t

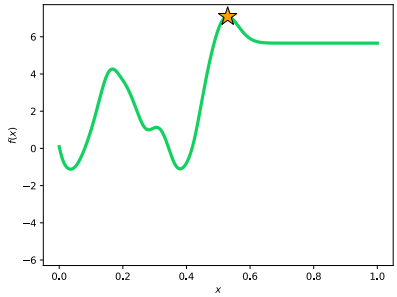
end

$$u_t(x) = \widetilde{f}_t(x) + \beta_t \widetilde{\sigma}_t(x)$$

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^{t}$ points , with $S_t = (\widetilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{t}$



u

x_{t+1}

for $t = \{1, \ldots, T-1\}$ **do**

Compute u_t

Select $x_{t+1} \leftarrow \text{argmax}_{x \in X}\, u_t(x)$;

Observe $y_{t+1} = f(x_{t+1}) + \epsilon_{t+1}$;

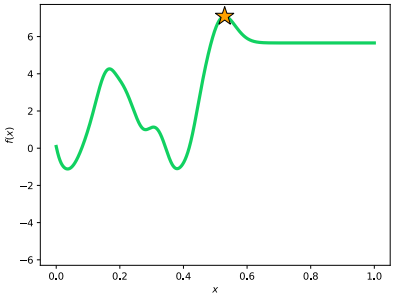
end

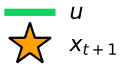
$$u_t(x) = \widetilde{f}_t(x) + \beta_t \widetilde{\sigma}_t(x) \qquad \rightarrow \qquad x_{t+1} = \underset{x \in X}{\text{argmax}}\; u_t(x)$$

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

BKB

At time t, collected $(x_i, y_i)_{i=1}^t$ points , with $S_t = (\widetilde{x}_i)_{i=1}^M \subseteq (x_i)_{i=1}^t$





for $t = \{1, \dots, T-1\}$ **do**

Compute u_t

Select $x_{t+1} \leftarrow \text{argmax}_{x \in X}\, u_t(x)$;

Observe $y_{t+1} = f(x_{t+1}) + \epsilon_{t+1}$;

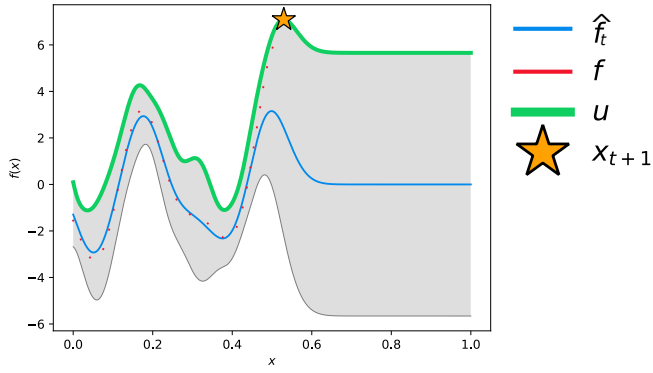
Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_{t+1})]$;

Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

end

(Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

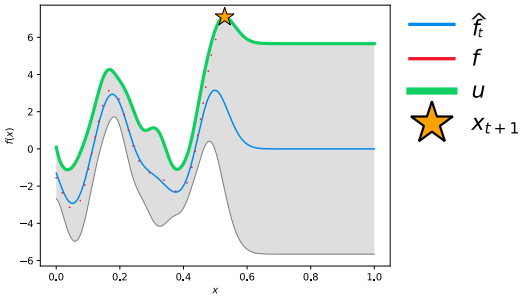
BKB: Regret



Guarantees of no-regret [Calandriello, Carratino, Lazaric, Valko, Rosasco '19]:
For the proper β_t

$$R_T \leqslant \sqrt{T}$$

BKB: Numerics



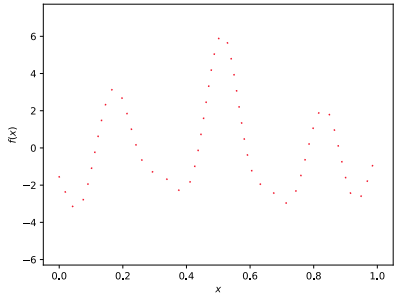
for $t = \{1, \dots, T-1\}$ **do**

- Compute u_t
- Select $x_{t+1} \leftarrow \text{argmax}_{x \in X}\, u_t(x)$;
- Observe $y_{t+1} = f(x_{t+1}) + \epsilon_{t+1}$;
- Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_{t+1})]$;
- Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

end

Computations: **?**

BKB: Numerics



f

for $t = \{1, \dots, T-1\}$ **do**

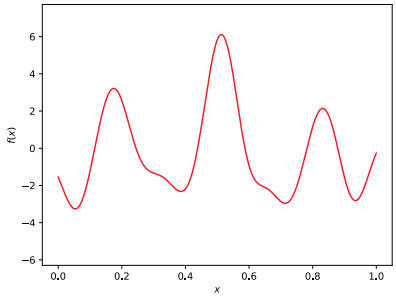
- Compute u_t
- Select $x_{t+1} \leftarrow \text{argmax}_{x \in X}\, u_t(x)$;
- Observe $y_{t+1} = f(x_{t+1}) + \epsilon_{t+1}$;
- Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_{t+1})]$;
- Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

end

- Over a discretized domain of f of cardinality A:

  Time $O(A d_{\text{eff}}^2 T)$ with $d_{\text{eff}} = |S_T| \ll T$

BKB: Numerics



for $t = \{1, \dots, T-1\}$ **do**
  Compute u_t
  Select $x_{t+1} \leftarrow \text{argmax}_{x \in X}\, u_t(x)$;
  Observe $y_{t+1} = f(x_{t+1}) + \epsilon_{t+1}$;
  Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}^2_t(x_1), \dots, \widetilde{\sigma}^2_t(x_{t+1})]$;
  Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;
end

- Over a discretized domain of f of cardinality A:
  Time $O(A d^2_{\text{eff}} T)$ with $d_{\text{eff}} = |S_T| \ll T$
- Over a continous domain of f, solving $\text{argmax}_{x \in X}\, u_t(x)$ via optimization:
  Time $O(\nu d^2_{\text{eff}} T + \nu C_\nu T)$
  with ν iterations and C_ν cost per iteration of inner optimization

Partition Trees

Definition:
Let $(X_h)_{h\in\mathbb{N}}$ be families of subsets of X.

For each depth $h \in \mathbb{N}$, the family of subsets X_h has cardinality N^h.

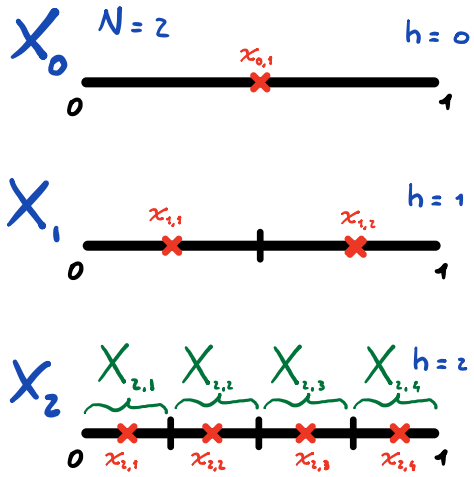
The elements of X_h are called cells $X_{h,i}$.

Each cell is identified by the centroid $x_{h,i} \in X_{h,i}$ s.t.

$$X_{h,i} = \{x \in X : d(x, x_{h,i}) \leqslant d(x, x_{h,j}) \quad \forall j \neq i\}.$$

For all $h \in \mathbb{N}$ and $i = 1, \dots, N^h$,

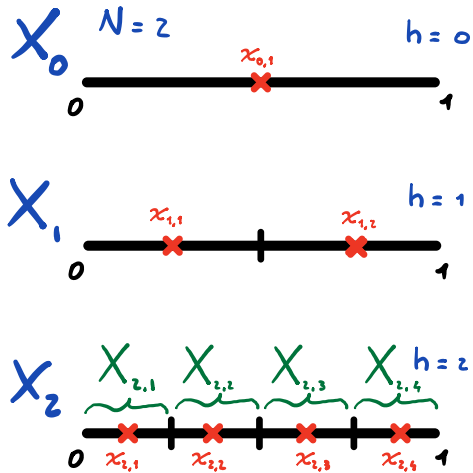
$$X_{h,i} = \cup_{j=N(i-1)+1}^{Ni} X_{h+1,j}.$$



Partition Trees

Assumption (1):
There exist $\rho \in (0,1)$ and $0 < c_2 \leqslant 1 \leqslant c_1$ s.t. for $h \geqslant 0$ and all $i = 1, \dots, N^h$

$$B(x_{h,i}, c_2\rho^h) \subset X_{h,i} \subset B(x_{h,i}, c_1\rho^h)$$



Partition Trees

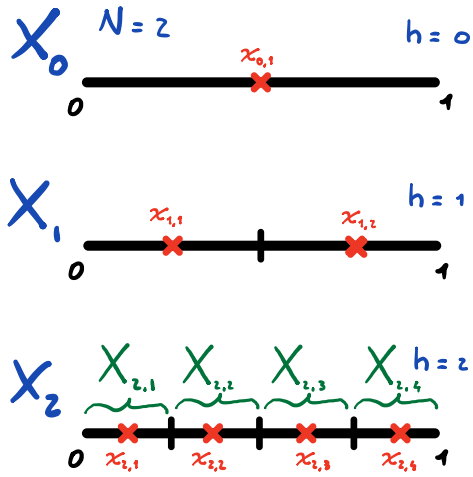
Assumption (1):
There exist $\rho \in (0,1)$ and $0 < c_2 \leqslant 1 \leqslant c_1$ s.t. for $h \geqslant 0$ and all $i = 1, \dots, N^h$

$$B(x_{h,i}, c_2\rho^h) \subset X_{h,i} \subset B(x_{h,i}, c_1\rho^h)$$

Lemma [Rando, C., Villa, Rosasco '21]:
Under (mild) and Ass. 1,
for $h \geqslant 0$ and for all $1 \leqslant i \leqslant N^h$,

$$\sup_{x,x' \in X_{h,i}} |f(x) - f(x')| \leqslant V_h,$$

with $\quad V_h \propto F\rho^h$



Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if **then**

  else

    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end

$I_t(x_{h,i}) \approx u_t(x_{h,i}) + V_h$ h.p. upper bound of f in cell $X_{h,i}$

Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \widetilde{\sigma}(x_t) \leqslant V_h$ **then**

  else

    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end

Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \tilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

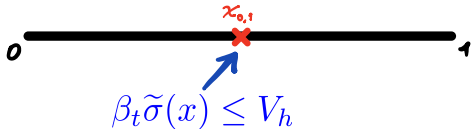
    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\tilde{p}_{t+1} \propto [\tilde{\sigma}_t^2(x_1), \dots, \tilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \tilde{p}_{t+1}$;

  end

end



$$\text{N children of } x_{h,i} \quad \leftarrow \quad \{x_{h+1,j} | N(i-1)+1 \leqslant j \leqslant Ni\}$$

Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \tilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\tilde{p}_{t+1} \propto [\tilde{\sigma}_t^2(x_1), \dots, \tilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \tilde{p}_{t+1}$;

  end

end

Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \widetilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

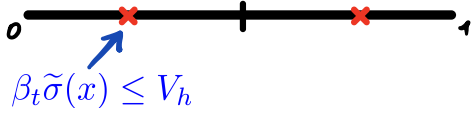
    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end



Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \widetilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end

Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \widetilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

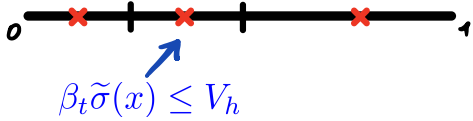
    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end



Ada-BKB

$\bar{X}_0 = \{x_{0,1}\}$

for $t = \{1, \dots, T\}$ **do**

  Compute I_t

  Select $x_t \leftarrow \text{argmax}_{x \in \bar{X}} I_t(x)$;

  if $\beta_t \widetilde{\sigma}(x_t) \leqslant V_h$ **then**

    $\bar{X}_{t+1} = (\bar{X}_t \setminus \{x_t\}) \cup \{\text{N children of } x_t\}$;

  else

    Observe $y_t = f(x_t) + \epsilon_t$;

    Set $\widetilde{p}_{t+1} \propto [\widetilde{\sigma}_t^2(x_1), \dots, \widetilde{\sigma}_t^2(x_t)]$;

    Sample $S_{t+1} \sim \widetilde{p}_{t+1}$;

  end

end

Ada-BKB: Regret and Numerics

Guarantees of no-regret [Rando, Carratino, Villa, Rosasco '21]:
For the proper β_t

$$R_T \leqslant \sqrt{T}$$

Computations:

Time $O(d_{\text{eff}}^2 T^2)$ with $d_{\text{eff}} \ll T$

Computation comparison

Fixed discretization of f:

- GP-UCB: **Time** $O(AT^3)$
  (Srinivas, Krause, Kakade, Seeger '10)

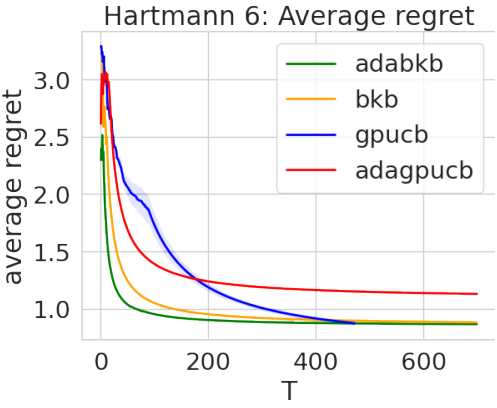
- BKB: **Time** $O(Ad_{\text{eff}}^2 T)$, with $d_{\text{eff}} \ll T$
  (Calandriello, Carratino, Lazaric, Valko, Rosasco '19)

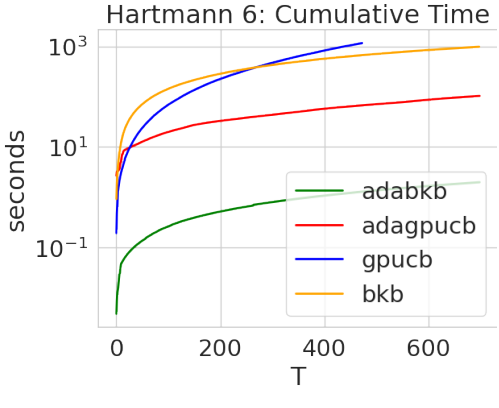
Adaptive discretization of f:

- AdaGP-UCB: **Time** $O(T^4)$
  (Shekhar, Javidi '18)

- Ada-BKB: **Time** $O(d_{\text{eff}}^2 T^2)$, with $d_{\text{eff}} \ll T$
  (Rando, Carratino, Villa, Rosasco '21)

In practice





Sublinear regret in a fraction of the time

Contribution and Open Questions

Contributions:

Blending sketching, optimization and adaptive discretization techniques to derive provably efficient algorithms in bandits optimization

Open questions:

- Batching?
- Random features BKB?
- ...

Questions?

Assumptions:

- There exists a non-decresing function $g : [0, \infty) \to [0, \infty)$ such that $g(0) = 0$ and such that for all $x, x' \in X$
$$d_k(x, x') \leqslant g(d(x, x')).$$
- There exist $\delta_k > 0$, $\alpha \in (0, 1]$, and $C'_k, C_k > 0$ such that
$$(\forall r \leqslant \delta_k) \qquad C_k r^\alpha \leqslant g(r) \leqslant C'_k r^\alpha$$