Sub-linear regret bounds for posterior sampling reinforcement learning with Gaussian processes

Work in progress: I might be talking moonshine

Joe Watson, Hamish Flynn, Jan Peters

02/12/2025

The question

Algorithm	Regret type	Rate
GP-UCB	worst-case	$\gamma_T \sqrt{T}$
GP-UCB	Bayesian	$\sqrt{\gamma_T T}$
GP-TS	worst-case	$\gamma_T \sqrt{T}$
GP-TS	Bayesian	$\sqrt{\gamma_T T}$

The question

Algorithm	Regret type	Rate
GP-UCB	worst-case	$\gamma_T \sqrt{T}$
GP-UCB	Bayesian	$\sqrt{\gamma_T T}$
GP-TS	worst-case	$\gamma_T \sqrt{T}$
GP-TS	Bayesian	$\sqrt{\gamma_T T}$
GP-UCRL	worst-case	$\gamma_T \sqrt{T}$
GP-UCRL	Bayesian	???
GP-PSRL	worst-case	???
GP-PSRL	Bayesian	$\gamma_T \sqrt{T}$

The question

Algorithm	Regret type	Rate
GP-UCB	worst-case	$\gamma_T \sqrt{T}$
GP-UCB	Bayesian	$\sqrt{\gamma_T T}$
GP-TS	worst-case	$\gamma_T \sqrt{T}$
GP-TS	Bayesian	$\sqrt{\gamma_T T}$
GP-UCRL	worst-case	$\gamma_T \sqrt{T}$
GP-UCRL	Bayesian	???
GP-PSRL	worst-case	???
GP-PSRL	Bayesian	$\gamma_T \sqrt{T}$

Is a Bayesian regret bound of the order $\sqrt{\gamma_T T}$ possible for GP-PSRL?

The setting

The model

Finite horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \mu, H)$.

- $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ is a set of states
- $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ is a set of actions
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the (unknown) transition kernel
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the (known) reward function
- μ is the initial state distribution
- H is the horizon

The model

Finite horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, r, \mu, H)$.

- $\mathcal{S} = \mathbb{R}^{d_s}$ is a set of states
- $\mathcal{A} \subseteq \mathbb{B}^{d_a}(R_a)$ is a set of actions
- $P(\mathbf{s}, \mathbf{a}) = \mathcal{N}(f(\mathbf{s}, \mathbf{a}), \sigma^2 \mathbf{I})$ (f is unknown, σ is known)
- $r : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is the (known) reward function
- $\mu = \mathcal{N}(0, \sigma^2 \mathbf{I})$ is the initial state distribution
- H is the horizon

The model

Finite horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, r, \mu, H)$.

- $\mathcal{S} = \mathbb{R}^{d_s}$ is a set of states
- $\mathcal{A} \subseteq \mathbb{B}^{d_a}(R_a)$ is a set of actions
- $P(\mathbf{s}, \mathbf{a}) = \mathcal{N}(f(\mathbf{s}, \mathbf{a}), \sigma^2 \mathbf{I})$ (f is unknown, σ is known)
- $r : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is the (known) reward function
- $\mu = \mathcal{N}(0, \sigma^2 \mathbf{I})$ is the initial state distribution
- H is the horizon

Write $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and $\mathbf{x} = (\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$.

Interaction protocol

The true MDP is $\mathcal{M}^\star = (\mathcal{S}, \mathcal{A}, f^\star, r, \mu, H)$. Write $f^\star = (f_1^\star, \ldots, f_{d_s}^\star)$.

Interaction protocol

The true MDP is $\mathcal{M}^\star = (\mathcal{S}, \mathcal{A}, f^\star, r, \mu, H)$. Write $f^\star = (f_1^\star, \ldots, f_{d_s}^\star)$.

At the start of the interaction, $f_i^\star \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

Interaction protocol

The true MDP is $\mathcal{M}^\star = (\mathcal{S}, \mathcal{A}, f^\star, r, \mu, H)$. Write $f^\star = (f_1^\star, \dots, f_{d_s}^\star)$.

At the start of the interaction, $f_i^\star \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

For a sequence of episodes $n = 1, \dots, N$, the following steps are repeated:

1. An initial state $\mathbf{s}_{n,1}$ is drawn from the initial state distribution μ
2. For steps $h = 1, \dots, H$:
   - The learner selects the action $\mathbf{a}_{n,h}$
   - The learner observes the reward $r_{n,h} = r(\mathbf{s}_{n,h}, \mathbf{a}_{n,h})$
   - The learner observes the next state $\mathbf{s}_{n,h+1} \sim \mathcal{N}(f^\star(\mathbf{s}_{n,h}, \mathbf{a}_{n,h}), \sigma^2\mathbf{I})$ (unless $h = H$)

Regret

For any MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, r, \mu, H)$, policy π and time step h, we define the value function $V^{\mathcal{M}}_{\pi,h} : \mathcal{S} \to \mathbb{R}$ as

$$V^{\mathcal{M}}_{\pi,h}(\mathbf{s}) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{j=h}^{H} r(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{s}_h = \mathbf{s}\right].$$

Regret

For any MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, r, \mu, H)$, policy π and time step h, we define the value function $V_{\pi,h}^{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$ as

$$V_{\pi,h}^{\mathcal{M}}(\mathbf{s}) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{j=h}^{H} r(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{s}_h = \mathbf{s}\right] .$$

The Bayesian regret after $T = NH$ total steps, or N episodes, is

$$\text{BayesRegret}_T = \mathbb{E}\left[\sum_{n=1}^{N} V_{\pi^\star,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1})\right] .$$

Regret

For any MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, r, \mu, H)$, policy π and time step h, we define the value function $V_{\pi,h}^{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$ as

$$V_{\pi,h}^{\mathcal{M}}(\mathbf{s}) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{j=h}^{H} r(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{s}_h = \mathbf{s}\right] .$$

The Bayesian regret after $T = NH$ total steps, or N episodes, is

$$\mathrm{BayesRegret}_T = \mathbb{E}\left[\sum_{n=1}^{N} V_{\pi^\star,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1})\right].$$

Maximum information gain: For a covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and any radius $R > 0$,

$$\gamma_T(\sigma^2, R) := \sup_{\mathbf{x}_1,\ldots,\mathbf{x}_T : \|\mathbf{x}_i\|_2 \leq R} \frac{1}{2} \log\det\left(\frac{1}{\sigma^2}\mathbf{K}_T + \mathbf{I}\right).$$

Posterior sampling reinforcement learning

In each episode, draw a random MDP from the posterior and follow the optimal policy for the sampled MDP.

Posterior sampling reinforcement learning

In each episode, draw a random MDP from the posterior and follow the optimal policy for the sampled MDP.

After episode $n-1$, the posterior $Q_n(f_i|H_{n-1})$ is the Gaussian princess with (predictive) mean and variance

$$
\begin{aligned}
\mu_{n-1,i}(\mathbf{x}) &= \mathbf{k}_{n-1}(\mathbf{x})^\top (\mathbf{K}_{n-1} + \sigma^2 \mathbf{I})^{-1} \mathbf{s}_{n-1,i} \\
\sigma^2_{n-1}(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{n-1}(\mathbf{x})^\top (\mathbf{K}_{n-1} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_{n-1}(\mathbf{x}) \,.
\end{aligned}
$$

Posterior sampling reinforcement learning

In each episode, draw a random MDP from the posterior and follow the optimal policy for the sampled MDP.

After episode $n-1$, the posterior $Q_n(f_i|H_{n-1})$ is the Gaussian princess with (predictive) mean and variance

$$\mu_{n-1,i}(\mathbf{x}) = \mathbf{k}_{n-1}(\mathbf{x})^\top(\mathbf{K}_{n-1} + \sigma^2\mathbf{I})^{-1}\mathbf{s}_{n-1,i}$$
$$\sigma^2_{n-1}(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}_{n-1}(\mathbf{x})^\top(\mathbf{K}_{n-1} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_{n-1}(\mathbf{x})\,.$$

GP-PSRL: Initialise history $H_0 = \emptyset$. For episode $n = 1, \ldots, N$:

1. Draw a random $f^{(n)}$ from the posterior Q_n
2. Find the optimal policy π_n for the MDP $\mathcal{M}_n = (\mathcal{S}, \mathcal{A}, f^{(n)}, r, \mu, H)$
3. Observe $\mathbf{s}_{n,1} \sim \mu$, and for $h = 1, \ldots, H$:
4. Update the history $H_n = H_{n-1} \cup \{\mathbf{s}_{n,1}, \mathbf{a}_{n,1}, \ldots, \mathbf{s}_{n,H}, \mathbf{a}_{n,H}\}$ and update the posterior $Q_{n+1}(f) \propto p(H_n|f)Q_1(f)$

Regret bounds for PSRL

General recipe for PSRL regret analysis

Stochastic optimism. Since $f^\star$ and $f^{(n)}$ have the same conditional distribution,

$$\text{BayesRegret}(T) = \mathbb{E}\Bigg[\sum_{n=1}^{N} V_{\pi^\star,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}_n}(\mathbf{s}_{n,1})\Bigg] + \mathbb{E}\Bigg[\sum_{n=1}^{N} V_{\pi_n,1}^{\mathcal{M}_n}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1})\Bigg]$$

General recipe for PSRL regret analysis

Stochastic optimism. Since $f^\star$ and $f^{(n)}$ have the same conditional distribution,

$$\text{BayesRegret}(T) = \mathbb{E}\left[\sum_{n=1}^{N} V^{\mathcal{M}^\star}_{\pi^\star,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1})\right] + \mathbb{E}\left[\sum_{n=1}^{N} V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}^\star}_{\pi_n,1}(\mathbf{s}_{n,1})\right]$$

General recipe for PSRL regret analysis

Stochastic optimism. Since $f^\star$ and $f^{(n)}$ have the same conditional distribution,

$$\mathrm{BayesRegret}(T) = \mathbb{E}\Big[\sum_{n=1}^{N} V_{\pi^\star,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}_n}(\mathbf{s}_{n,1})\Big] + \mathbb{E}\Big[\sum_{n=1}^{N} V_{\pi_n,1}^{\mathcal{M}_n}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1})\Big]$$

Simulation lemma. The value estimation error is controlled by the difference between $f^{(n)}$ and $f^\star$.

$$\mathbb{E}\Big[\sum_{n=1}^{N} V_{\pi_n,1}^{\mathcal{M}_n}(\mathbf{s}_{n,1}) - V_{\pi_n,1}^{\mathcal{M}^\star}(\mathbf{s}_{n,1})\Big] \le \frac{HR_{\max}}{\sigma}\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{h=1}^{H-1} \|f^{(n)}(\mathbf{x}_{n,h}) - f^\star(\mathbf{x}_{n,h})\|_2\Big].$$

General recipe for PSRL regret analysis

Stochastic optimism. Since $f^\star$ and $f^{(n)}$ have the same conditional distribution,

$$\mathrm{BayesRegret}(T) = \mathbb{E}\Bigg[\sum_{n=1}^{N} V^{\mathcal{M}^\star}_{\pi^\star,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1})\Bigg] + \mathbb{E}\Bigg[\sum_{n=1}^{N} V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}^\star}_{\pi_n,1}(\mathbf{s}_{n,1})\Bigg]$$

Simulation lemma. The value estimation error is controlled by the difference between $f^{(n)}$ and $f^\star$.

$$\mathbb{E}\Bigg[\sum_{n=1}^{N} V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}^\star}_{\pi_n,1}(\mathbf{s}_{n,1})\Bigg] \leq \frac{HR_{\max}}{\sigma}\mathbb{E}\Bigg[\sum_{n=1}^{N}\sum_{h=1}^{H-1} \|f^{(n)}(\mathbf{x}_{n,h}) - f^\star(\mathbf{x}_{n,h})\|_2\Bigg].$$

GP concentration. Bound the differences between $f^{(n)}$ and $f^\star$ at $(\mathbf{x}_{n,h})_{n,h}$.

$$\frac{HR_{\max}}{\sigma}\mathbb{E}\Bigg[\sum_{n=1}^{N}\sum_{h=1}^{H-1} \|f^{(n)}(\mathbf{x}_{n,h}) - f^\star(\mathbf{x}_{n,h})\|_2\Bigg] \approx \mathcal{O}\big(H(d_s + d_a)\sqrt{T\gamma_T}\big).$$

General recipe for PSRL regret analysis

Stochastic optimism. Since $f^\star$ and $f^{(n)}$ have the same conditional distribution,

$$\mathrm{BayesRegret}(T) = \mathbb{E}\Big[\sum_{n=1}^{N} V^{\mathcal{M}^\star}_{\pi^\star,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1})\Big] + \mathbb{E}\Big[\sum_{n=1}^{N} V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}^\star}_{\pi_n,1}(\mathbf{s}_{n,1})\Big]$$

Simulation lemma. The value estimation error is controlled by the difference between $f^{(n)}$ and $f^\star$.

$$\mathbb{E}\Big[\sum_{n=1}^{N} V^{\mathcal{M}_n}_{\pi_n,1}(\mathbf{s}_{n,1}) - V^{\mathcal{M}^\star}_{\pi_n,1}(\mathbf{s}_{n,1})\Big] \leq \frac{HR_{\max}}{\sigma}\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{h=1}^{H-1} \|f^{(n)}(\mathbf{x}_{n,h}) - f^\star(\mathbf{x}_{n,h})\|_2\Big].$$

GP concentration. Bound the differences between $f^{(n)}$ and $f^\star$ at $(\mathbf{x}_{n,h})_{n,h}$.

$$\frac{HR_{\max}}{\sigma}\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{h=1}^{H-1} \|f^{(n)}(\mathbf{x}_{n,h}) - f^\star(\mathbf{x}_{n,h})\|_2\Big] \approx \mathcal{O}\big(H(d_s + d_a)\sqrt{T\gamma_T}\big).$$

Problem: With positive probability, the states $\mathbf{s}_{n,h+1} = f^\star(\mathbf{s}_{n,h}) + \boldsymbol{\varepsilon}_{n,h+1}$ exceed any finite bound.

Problems with unbounded states

Failure of uniform GP concentration. The supremum of a GP over an unbounded domain blows up.

$$\mathbb{E}\left[\sup_{\mathbf{x}\in\mathbb{B}^{d_s+d_a}(R)} \left|f_i^{(n)}(\mathbf{x}) - f_i^{\star}(\mathbf{x})\right|\right] \gtrsim (d_s + d_a)\sqrt{\log(R)}\,.$$

Problems with unbounded states

Failure of uniform GP concentration. The supremum of a GP over an unbounded domain blows up.

$$\mathbb{E}\left[\sup_{\mathbf{x}\in\mathbb{B}^{d_s+d_a}(R)} \left|f_i^{(n)}(\mathbf{x}) - f_i^{\star}(\mathbf{x})\right|\right] \gtrsim (d_s + d_a)\sqrt{\log(R)}\,.$$

Linear information gain. Take k to be a Matérn kernel or the Gaussian kernel and consider

$$\gamma_T(\sigma^2,\infty) := \sup_{\mathbf{x}_1,\ldots,\mathbf{x}_T : \mathbf{x}_i\in\mathbb{R}^{d_s+d_a}} \frac{1}{2}\log\det\left(\frac{1}{\sigma^2}\mathbf{K}_T + \mathbf{I}\right).$$

Problems with unbounded states

Failure of uniform GP concentration. The supremum of a GP over an unbounded domain blows up.

$$\mathbb{E}\left[\sup_{\mathbf{x}\in\mathbb{B}^{d_s+d_a}(R)}\left|f_i^{(n)}(\mathbf{x}) - f_i^\star(\mathbf{x})\right|\right] \gtrsim (d_s + d_a)\sqrt{\log(R)}\,.$$

Linear information gain. Take k to be a Matérn kernel or the Gaussian kernel and consider

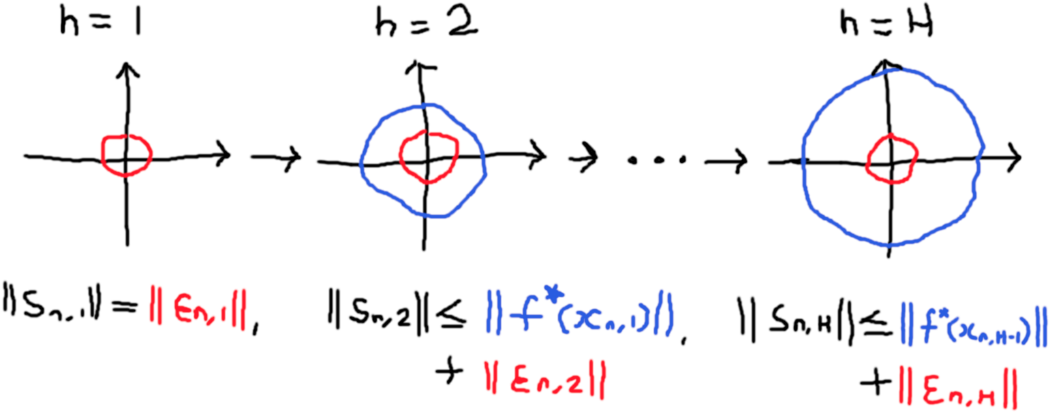
$$\gamma_T(\sigma^2, \infty) := \sup_{\mathbf{x}_1,\ldots,\mathbf{x}_T : \mathbf{x}_i \in \mathbb{R}^{d_s+d_a}} \frac{1}{2}\log\det\left(\frac{1}{\sigma^2}\mathbf{K}_T + \mathbf{I}\right).$$

We can always choose $\mathbf{x}_1, \ldots, \mathbf{x}_T$ to be arbitrarily far apart, which means

$$\gamma_T(\sigma^2, \infty) = \frac{1}{2}\log\det\left(\frac{1}{\sigma^2}\mathbf{I} + \mathbf{I}\right) = \frac{T}{2}\log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

The fix

General idea



Tool 1: GP concentration

SHARPER BOUNDS FOR GAUSSIAN AND EMPIRICAL PROCESSES[1]

BY M. TALAGRAND

Tool 1: GP concentration

SHARPER BOUNDS FOR GAUSSIAN AND EMPIRICAL PROCESSES[1]

BY M. TALAGRAND

Let $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ be a random draw from a zero mean Gaussian process with a covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies $C := \sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$ and

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{B}^{d_s + d_a}(R)\,, \quad |k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{x}')| \leq L \|\mathbf{x} - \mathbf{x}'\|_2\,.$$

Tool 1: GP concentration

SHARPER BOUNDS FOR GAUSSIAN AND EMPIRICAL PROCESSES[1]

BY M. TALAGRAND

Let $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ be a random draw from a zero mean Gaussian process with a covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies $C := \sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$ and

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{B}^{d_s + d_a}(R) \,, \quad |k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{x}')| \leq L \|\mathbf{x} - \mathbf{x}'\|_2 \,.$$

Then there exists a universal constant D, such that for all $z \geq \sqrt{d_s} + \sqrt{2d_s(d_s + d_a)}$,

$$\mathbb{P}\left(\sup_{\mathbf{x} \in \mathbb{B}^{d_s + d_a}(R)} \left\| f^\star(\mathbf{x}) \right\|_2 \geq z \right) \leq 2d_s \left(\frac{D\sqrt{C^2 + 4LR}}{C\sqrt{2d_s(d_s + d_a)}} z \right)^{2(d_s + d_a)} \exp\left(- \frac{z^2}{2d_s C^2} \right) .$$

Importantly, if $z \geq D(d_s + d_a)\sqrt{\log(1/\delta)}$, then $\mathbb{P}(\sup_{\mathbf{x} \in \mathbb{B}^{d_s + d_a}(R)} \|f^\star(\mathbf{x})\|_2 \geq z) \leq \delta$.

Tool 2: indicator trick

For any finite collection of events $(A_h)_{h=1}^H$,

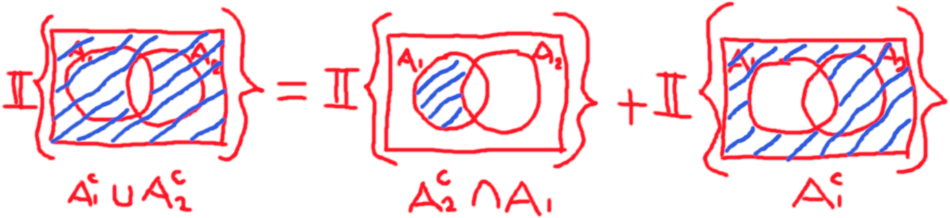
$$\mathbb{I}\{\cup_{h=1}^H A_h^{\mathsf{c}}\} = \sum_{h=1}^H \mathbb{I}\{A_h^{\mathsf{c}}\}\mathbb{I}\{\cap_{j=1}^{h-1} A_j\}\,.$$

Tool 2: indicator trick

For any finite collection of events $(A_h)_{h=1}^H$,

$$\mathbb{I}\{\cup_{h=1}^H A_h^c\} = \sum_{h=1}^H \mathbb{I}\{A_h^c\}\mathbb{I}\{\cap_{j=1}^{h-1} A_j\}\,.$$

Easy proof for $H = 2$.



Bounded states

Let $A_{n,h} := \{\|\mathbf{s}_{n,h}\|_2 \leq R_h\}$ and define the good event $A := \cap_{n=1}^{N} \cap_{h=1}^{H} A_{n,h}$.

Bounded states

Let $A_{n,h} := \{\|\mathbf{s}_{n,h}\|_2 \leq R_h\}$ and define the good event $A := \cap_{n=1}^{N} \cap_{h=1}^{H} A_{n,h}$. Using the indicator trick,

$$\mathbb{P}(A^{\mathsf{c}}) = \mathbb{E}[\mathbb{I}\{\cup_{n=1}^{N} \cup_{h=1}^{H} A_{n,h}^{\mathsf{c}}\}] = \sum_{n=1}^{N} \sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{A_{n,h}^{\mathsf{c}}\}\mathbb{I}\{(\cap_{i=1}^{n-1} \cap_{j=1}^{H} A_{i,j}) \cap (\cap_{j=1}^{h-1} A_{n,j})\}]$$

$$\leq \sum_{n=1}^{N} \sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \geq R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \leq R_{h-1}\}] \,.$$

Bounded states

Let $A_{n,h} := \{\|\mathbf{s}_{n,h}\|_2 \le R_h\}$ and define the good event $A := \cap_{n=1}^{N} \cap_{h=1}^{H} A_{n,h}$. Using the indicator trick,

$$\begin{aligned}\mathbb{P}(A^{\mathsf{c}}) = \mathbb{E}[\mathbb{I}\{\cup_{n=1}^{N} \cup_{h=1}^{H} A_{n,h}^{\mathsf{c}}\}] &= \sum_{n=1}^{N}\sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{A_{n,h}^{\mathsf{c}}\}\mathbb{I}\{(\cap_{i=1}^{n-1} \cap_{j=1}^{H} A_{i,j}) \cap (\cap_{j=1}^{h-1} A_{n,j})\}] \\ &\le \sum_{n=1}^{N}\sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \ge R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \le R_{h-1}\}]\,.\end{aligned}$$

If $\|\mathbf{s}_{n,h}\|_2 \le R_h$, then $\|\mathbf{x}_{n,h}\|_2 \le \sqrt{R_h^2 + R_a^2} =: \widetilde{R}_h$. Therefore,

$$\begin{aligned}\mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \ge R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \le R_{h-1}\}] &\le \mathbb{P}\big(\sup_{\mathbf{x}\in\mathbb{B}^{d_s+d_a}(\widetilde{R}_{h-1})}\|f^\star(\mathbf{x})\|_2 > R_h/2\big) + \mathbb{P}(\|\boldsymbol{\varepsilon}_{n,h}\|_2 > R_h/2) \\ &\le 2d_s\left(\frac{D\sqrt{C^2+4L\widetilde{R}_{h-1}}}{C\sqrt{2d_s(d_s+d_a)}}R_h\right)^{2(d_s+d_a)} \exp\left(-\frac{R_h^2}{8d_sC^2}\right) \\ &\quad + 2^{d_s/2}\exp\left(-\frac{R_h^2}{16\sigma^2}\right).\end{aligned}$$

Bounded states

Let $A_{n,h} := \{\|\mathbf{s}_{n,h}\|_2 \leq R_h\}$ and define the good event $A := \cap_{n=1}^{N} \cap_{h=1}^{H} A_{n,h}$. Using the indicator trick,

$$\mathbb{P}(A^{\mathsf{c}}) = \mathbb{E}[\mathbb{I}\{\cup_{n=1}^{N} \cup_{h=1}^{H} A_{n,h}^{\mathsf{c}}\}] = \sum_{n=1}^{N}\sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{A_{n,h}^{\mathsf{c}}\}\mathbb{I}\{(\cap_{i=1}^{n-1} \cap_{j=1}^{H} A_{i,j}) \cap (\cap_{j=1}^{h-1} A_{n,j})\}]$$
$$\leq \sum_{n=1}^{N}\sum_{h=1}^{H} \mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \geq R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \leq R_{h-1}\}]\,.$$

If $\|\mathbf{s}_{n,h}\|_2 \leq R_h$, then $\|\mathbf{x}_{n,h}\|_2 \leq \sqrt{R_h^2 + R_a^2} =: \widetilde{R}_h$. Therefore,

$$\mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \geq R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \leq R_{h-1}\}] \leq \mathbb{P}\big(\sup_{\mathbf{x}\in\mathbb{B}^{d_s+d_a}(\widetilde{R}_{h-1})} \|f^\star(\mathbf{x})\|_2 > R_h/2\big) + \mathbb{P}(\|\boldsymbol{\varepsilon}_{n,h}\|_2 > R_h/2)$$
$$\leq 2d_s\left(\frac{D\sqrt{C^2+4L\widetilde{R}_{h-1}}}{C\sqrt{2d_s(d_s+d_a)}}R_h\right)^{2(d_s+d_a)} \exp\left(-\frac{R_h^2}{8d_sC^2}\right)$$
$$+\, 2^{d_s/2}\exp\left(-\frac{R_h^2}{16\sigma^2}\right).$$

One can set each R_h such that $\widetilde{R}_h \leq \widetilde{C}_h(d_s+d_a)\sqrt{\log(T)}$,

$$\mathbb{E}[\mathbb{I}\{\|\mathbf{s}_{n,h}\|_2 \geq R_h\}\mathbb{I}\{\|\mathbf{s}_{n,h-1}\|_2 \leq R_{h-1}\}] \leq \tfrac{1}{T^2}\,, \qquad \text{and} \qquad \mathbb{P}(A^{\mathsf{c}}) \leq \frac{1}{T}\,.$$

Main result

For a bounded and Lipschitz kernel function k, the Bayesian regret of PSRL (with $f_1^\star, \ldots, f_{d_s}^\star \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')))$ satisfies

$$\text{BayesRegret}_T = \mathcal{O}\big(H(d_s + d_a)\sqrt{\gamma_T(\sigma^2, (d_s + d_a)\sqrt{\log(T)})T\log(T)}\big)\,.$$

Main result

For a bounded and Lipschitz kernel function k, the Bayesian regret of PSRL (with $f_1^\star, \ldots, f_{d_s}^\star \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')))$ satisfies

$$\text{BayesRegret}_T = \mathcal{O}\big(H(d_s + d_a)\sqrt{\gamma_T(\sigma^2, (d_s + d_a)\sqrt{\log(T)})T\log(T)}\big)\,.$$

"Proof." Follow the recipe from before, but under the good event A.

Main result

For a bounded and Lipschitz kernel function k, the Bayesian regret of PSRL (with $f_1^\star, \dots, f_{d_s}^\star \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')))$ satisfies

$$\mathrm{BayesRegret}_T = \mathcal{O}\big(H(d_s + d_a)\sqrt{\gamma_T(\sigma^2, (d_s + d_a)\sqrt{\log(T)})T \log(T)}\big)\,.$$

"Proof." Follow the recipe from before, but under the good event A.

If k is the Matérn kernel (with smoothness parameter ν), then

$$\gamma_T(\sigma^2, (d_s + d_a)\sqrt{\log(T)}) = \mathcal{O}\big((d_s + d_a)^{2\nu} T^{\frac{d_s + d_a}{2\nu + d_s + d_a}} \log^{\max(\nu, \frac{2\nu}{\nu+1})}(T)\big)\,,$$

and

$$\mathrm{BayesRegret}_T = \mathcal{O}\big(H(d_s + d_a)^{1+\nu} T^{\frac{\nu + d_s + d_a}{2\nu + d_s + d_a}} \log^{1+\max(\frac{\nu}{2}, \frac{\nu}{\nu+1})}(T)\big)\,.$$

Conclusion

In the Bayesian regret bound for GP-PSRL, you can get the information gain underneath the square root.

Conclusion

In the Bayesian regret bound for GP-PSRL, you can get the information gain underneath the square root.

What's left?

- What if the kernel is not uniformly (on $\mathcal{S} \times \mathcal{A}$) bounded/Lipschitz? (done)

Conclusion

In the Bayesian regret bound for GP-PSRL, you can get the information gain underneath the square root.

What's left?

- What if the kernel is not uniformly (on $\mathcal{S} \times \mathcal{A}$) bounded/Lipschitz? (done)
- What if an approximate posterior is used? (somewhat done)

Conclusion

In the Bayesian regret bound for GP-PSRL, you can get the information gain underneath the square root.

What's left?

- What if the kernel is not uniformly (on $\mathcal{S} \times \mathcal{A}$) bounded/Lipschitz? (done)
- What if an approximate posterior is used? (somewhat done)
- What about worst-case regret? (not done)

The end. Thank you!