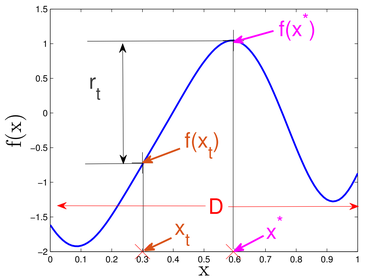
On Kernelized Multi-armed Bandits

Sayak Ray Chowdhury and Aditya Gopalan

Department of Electrical Communication Engineering, Indian Institute of Science

Problem Statement

Sequentially Maximize $f : D \to \mathbb{R}$, f unknown, $D \subset \mathbb{R}^d$



- $x^\star = \underset{x \in D}{\operatorname{argmax}}\, f(x)$
- At each round t:
  1. Learner chooses $x_t \in D$ based on past
  2. Observes noisy reward $y_t = f(x_t) + \varepsilon_t$
  3. Suffers regret $r_t = f(x^\star) - f(x_t)$

Goal: Minimize **cumulative regret** $\sum_{t=1}^{T} r_t$

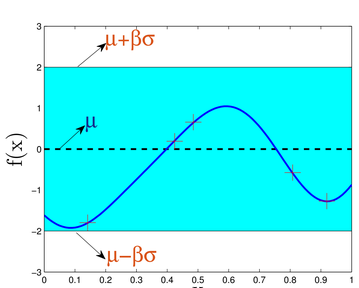
Assumptions

- Noise ε_t is R-sub-Gaussian
- f lies in **RKHS** of functions: $D \to \mathbb{R}$
- Positive semi-definite **kernel** function $k : D \times D \to \mathbb{R}$
- Reproducing property: $f(x) = \langle f, k(x, \cdot)\rangle_k$
- Induces smoothness: $|f(x) - f(y)| \le \|f\|_k \|k(x,\cdot) - k(y,\cdot)\|_k$
- D is compact, $\|f\|_k \le B$ known
- Bounded variance: $k(x,x) \le 1$, for all $x \in D$

Example Kernels

- Squared Exponential kernel: $k(x,y) = \exp\left(\frac{-\|x-y\|_2^2}{2l^2}\right)$
- Matérn kernel: $k(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\|x-y\|_2\sqrt{2\nu}}{l}\right)^{\nu} B_\nu\left(\frac{\|x-y\|_2\sqrt{2\nu}}{l}\right)$
- Stationary kernels: $k(x,y) \equiv k(x-y)$

Algorithm Design Philosophy



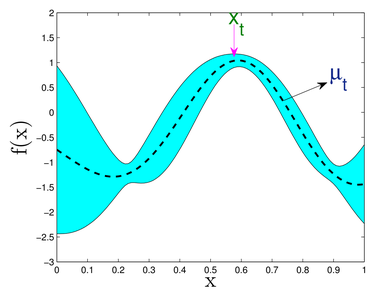
- Use **Gaussian Process (GP)** prior and **Gaussian Likelihood** model
- Prior of blue f: $GP(0, v^2 k(x,y))$
- Noise $\varepsilon_t \sim \mathcal{N}(0, \lambda v^2)$
- After t rounds, reward vector $y_{1:t} \sim \mathcal{N}(0, v^2(K_t + \lambda I))$

Posterior after t rounds is $GP(\mu_t(x), v^2 k_t(x,y))$:

$$\mu_t(x) = k_t(x)^T (K_t + \lambda I)^{-1} y_{1:t}$$
$$k_t(x,y) = k(x,y) - k_t(x)^T (K_t + \lambda I)^{-1} k_t(y)$$

Algorithm 1: Improved GP-UCB (IGP-UCB)

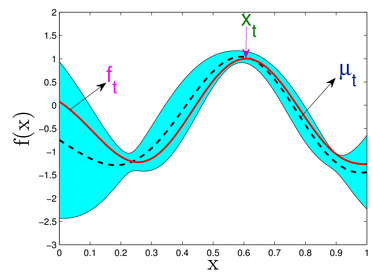
At each round t: play $x_t = \underset{x \in D}{\operatorname{argmax}}\, \mu_t(x) + \beta_t \sigma_t(x)$



- β_t trades off b/w **exploration** and **exploitation**
- **Reduced** width (β_t) of confidence interval compared to **GP-UCB** (*Srinivas et al., ICML 2010*)

Algorithm 2: Gaussian Process Thompson Sampling (GP-TS)

At each round t: sample a random function and play its maximizer



- Sample f_t from posterior of f
- Play $x_t = \underset{x \in D_t}{\operatorname{argmax}}\, f_t(x)$
- $D_t \subset D$: suitably chosen Discretization sets

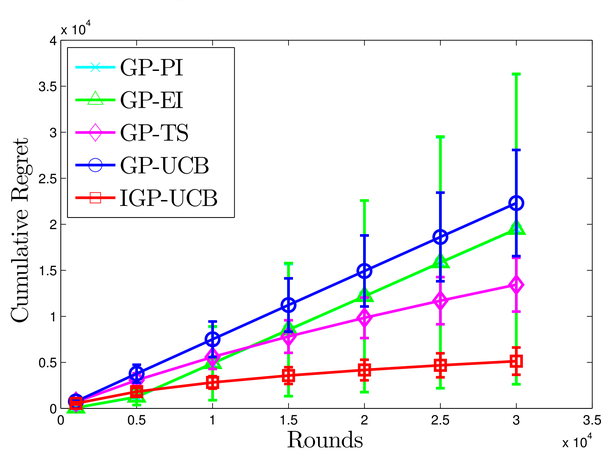
Key Tool: New Self-Normalized Concentration Inequality for RKHS-valued Martingales

For all t: $\|S_t\|^2_{V_t^{-1}} \le 2R^2 \ln\left(\frac{\sqrt{\det(K_t+I)}}{\delta}\right)$ with probability at least $1-\delta$

- Feature map $\varphi : D \to$ **RKHS**
- $S_t = \sum_{s=1}^{t} \varepsilon_s \varphi(x_s)$ ← RKHS-valued Martingale
- $V_t = I + \sum_{s=1}^{t} \varphi(x_s)\varphi(x_s)^T$ ← possibly of infinite dimension
- **Generalizes** finite-dimensional result for vector-valued Martingales (*Abbasi-Yadkori et al., NIPS 2011*)
- Uses method of mixtures technique
- Curse of Dimensionality → Mixing over Gaussian Processes

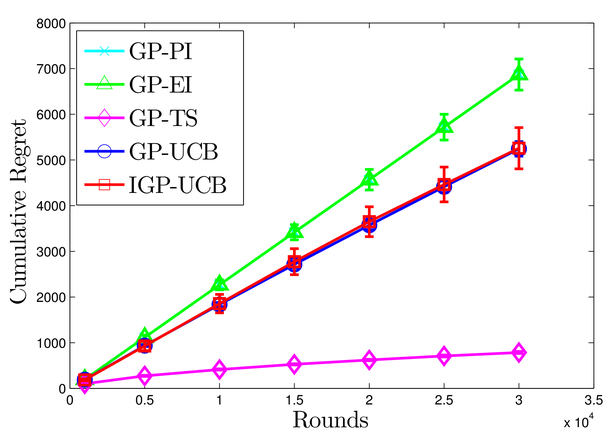
Numerical Results

f sampled from RKHS (SE kernel)



IGP-UCB improves over GP-UCB, GP-TS fares pretty well

Temperature Sensor Data (Intel Research)



IGP-UCB performs same as GP-UCB, GP-TS fares the best

http://proceedings.mlr.press/v70/chowdhury17a.html

Regret Bound for IGP-UCB

Regret of **IGP-UCB**: $O\left(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T)\right)$ whp with the choice of confidence width $\beta_t \approx B + \sqrt{\gamma_t}$ for all t

- γ_T is **Maximum Information Gain** after T rounds: $\gamma_T = \max_{A \subset D : |A| = T} I(y_A; f_A)$
  - Mutual Information b/w function values and rewards at A
  - **Reduction in uncertainty** about f after observing rewards
  - SE kernel: $\gamma_T = O((\ln T)^{d+1})$ → **sublinear** regret
- Regret of **GP-UCB**: $O\left(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T \ln^{3/2} T)\right)$ whp and so we improve by $O(\ln^{3/2} T)$!

Regret Bound for GP-TS

Regret of **GP-TS**: $O\left(\sqrt{Td\ln(BdT)}(B\sqrt{\gamma_T} + \gamma_T)\right)$ whp

- First **frequentist** regret guarantee of **TS** in the **non-parametric** setting of infinite action spaces
- $\sqrt{d\ln(BdT)}$ ← Consequence of **Discretization**
- Open Question: Can the logarithmic dependency be removed?

Recovering Linear Bandits

- Linear kernel: $k(x,y) = x^T y$
- $f(x) = \theta^T x$, $\theta \in \mathbb{R}^d$ unknown parameter
- **Maximum Information Gain**: $\gamma_T = O(d \ln T)$
- Regret of **IGP-UCB**: $\tilde{O}(d\sqrt{T})$ and **GP-TS**: $\tilde{O}(d^{3/2}\sqrt{T})$
- Exactly recovers regrets of **OFUL** (*Abbasi-Yadkori et al., 2013*) and **Linear TS** (*Agrawal and Goyal, ICML 2013*)
- **Lower Bound**: $\Omega(d\sqrt{T})$ (*Dani et al., COLT 2008*)

Conclusion

For Non-parametric Bandits, we have **improved** the existing UCB based algorithm, **introduced** a new Thompson Sampling based algorithm and **developed** a novel self-normalized concentration inequality for RKHS-valued martingales.

Selected References

[1] Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. **Improved algorithms for linear stochastic bandits.** *In Advances in Neural Information Processing Systems, 2011.*

[2] Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, and Seeger, Matthias. **Gaussian process optimization in the bandit setting: No regret and experimental design.** *In Proceedings of the 27th International Conference on Machine Learning, 2010.*