Thompson Sampling with Information Relaxation Penalties

Seungki Min, Costis Maglaras, Ciamac C. Moallemi
Decision, Risk and Operations
Graduate School of Business

Thompson Sampling + Information Relaxation Framework

We propose a general framework that produces a series of policies and performance bounds for finite-time horizon multi-armed bandits problems. We introduce the information relaxation penalties to obtain tight performance bounds, each of which has a corresponding online decision making policy.

- Thompson sampling/full-information benchmark is a special case.
- The ideal penalty produces Bayesian optimal policy/true optimal value.
- Several penalty function designs in the middle.

Information Relaxation Sampling

Given a penalty function $z_t(\cdot)$, when remaining time is $T$ and current belief is $\psi$, observe outcome, update belief $\psi$, $T \leftarrow T - 1$

**Choice of Penalty Function**
- Thompson sampling: “find the best arm given parameters” - $O(K)$
- IRS.FH: “find the best arm given a finite number of observations” - $O(KT)$
- IRS.V-Zero: “find the best allocation” - $O(KT^2)$
- IRS.V-EMax: “find the best action sequence” - $O(KT^K)$

**Numerical Simulations**

**Analysis**
- Weak duality: $V^* \leq W^2$, Strong duality: $V\left(\pi^{IRS.OPT}\right) = V^* = W^{IRS.OPT}$
- Monotonicity: $W^{IRS.V-\text{Zero}} \leq W^{IRS.FH} \leq W^{IRS.V-\text{Opt}}$
- Suboptimality gaps: for Bernoulli MAB,

$$W^{TS}(T, \psi) - V(\pi^{TS}, T, \psi) \leq 3K + 2\sqrt{\log T} \times 2\sqrt{KT}$$
$$W^{IRS.FH}(T, \psi) - V(\pi^{IRS.FH}, T, \psi) \leq 3K + 2\sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)$$
$$W^{IRS.V-Zero}(T, \psi) - V(\pi^{IRS.V-Zero}, T, \psi) \leq 2K + \sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)$$

**References**

David Brown et al. (2010), Information Relaxations and Duality in Stochastic Dynamic Programs
Daniel Russo and Benjamin Van Roy (2014), Learning to Optimize via Posterior Sampling