**Thompson Sampling with Information Relaxation Penalties**

Seungki Min, Costis Maglaras, Ciamac C. Moallemi

Decision, Risk and Operations

---

**Motivation**

Multi-armed bandits (MAB) with $K$ indep. arms:
- Bayesian setting
- Finite-horizon $T$ (known)

Thompson sampling (TS) may explore too much:
- At $t = T$, the greedy decision is optimal.

Bayesian optimal policy (OPT) can be obtained by solving Bayesian DP (but exactly intractable).

Information relaxation provides a systematic way of computing performance bounds of a stochastic DP.

---

**Contributions**

By introducing information relaxation framework,
- Generalization of Thompson sampling & Bayesian regret benchmark.
- Choice of penalty function $z_t(\cdot)$ determines a policy $\pi^*$ and a upper bound $W^*$.

**Numerical Experiments**

![Bayesian regret comparison](image)

**Information Relaxation Sampling**

"Information relaxation": Lagrangian relaxation of the non-anticipativity constraint imposed on the policy space.

(ranomized) Policy $\pi^*$:

1. Sample the entire future
   $\theta \sim \mathcal{P}(y), R_{a,n} \sim \mathcal{R}(\theta_n), \forall n \in [T]$

2. Find a best action sequence under penalties $z_t(\cdot)$
   $\max_{a_n, y} \sum_{t=1}^{T} r_t(a_t, \omega) - z_t(a_t, \omega)$

3. Take the first action
   $\pi_1^*$

Observe outcome, update belief $y$, $T \leftarrow T - 1$

Choice of penalty functions:

- IRS(T; y, z): $W^*(T, y) \triangleq \mathbb{E}_{\omega \sim Z} \left( \sum_{t=1}^{T} r_t(a_t, \omega) - z_t(a_t, \omega) \right)$

A good penalty $\approx$ value of future information

$\Rightarrow$ prevents an action overly optimized to a particular scenario.

**Duality Theorem**

For any ex-ante zero-mean penalty function $z_t(\cdot)$,
(i.e., $\mathbb{E}[z_t(a_t, \omega) | F_{t-1}(a_{t-1}, \omega)] = 0, \forall a_{t-1} \in A'$)

(Weak duality) $W^*(T, y) \geq V(OPT, V, y)$.

There exists an ideal penalty $z_{ideal}^*$ such that

(Strong duality) $W_{ideal}(T, y) = V(OPT, V, y)$.

**Asymptotic Behavior**

When $T = 1$, all IRS policies (except TS) take the optimal action.
When $T \to \infty$, IRS.FH and IRS.V-ZERO behave like TS

$P[TS(y) = a] = \lim_{T \to \infty} P[IRS.FH(T, y) = a]$

$\approx \lim_{T \to \infty} P[IRS.V-ZERO(T, y) = a]$.

**Performance Analysis**

Monotonicity in performance bounds:

$W^{IRS.FH}(T, y) \geq W^{IRS.V-ZERO}(T, y), \quad \text{and} \quad W^{THOMPSON}(T, y) \geq W^{IRS.V-EMAX}(T, y)$.

Suboptimality gaps:

$W^{THOMPSON}(T, y) - V(TS, T, y) \leq 3K + 2\sqrt{\log T} \times 2\sqrt{K}$

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

$W^{IRS.V-EMAX}(T, y) - V^{IRS.V-EMAX}(T, y) \leq 2K + \sqrt{\log T} \times \left( 2\sqrt{K} - \frac{1}{3} \sqrt{T/K} \right)$.

---

**References**

Brown et al. (2010), Information relaxations and duality in stochastic dynamic programs.
Russo and Van Roy (2014), Learning to optimize via posterior sampling.