Thompson Sampling with Information Relaxation Penalties

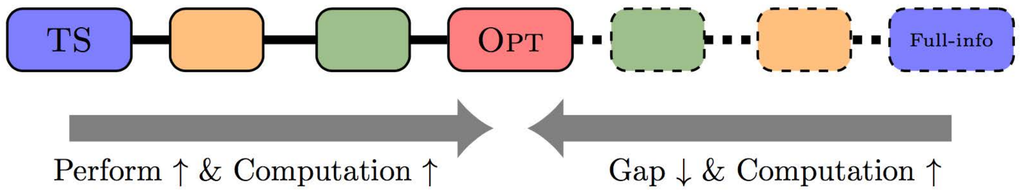
CENTER FOR FINANCIAL & BUSINESS ANALYTICS
Data Science Institute

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

Data Science Institute
COLUMBIA UNIVERSITY

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 y,

$\mathrm{IRS}(T, \mathbf{y}; z)$

Step 1: Sample the entire future

$\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y}), \quad \tilde{R}_{a,n} \overset{\text{i.i.d.}}{\sim} \mathcal{R}(\tilde{\theta}_a), \forall n \in [T]$

Step 2: Solve multi-period DP

$\max_{\mathbf{a}_{1:T}} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}; \tilde{\iota}) - z_t(\mathbf{a}_{1:t}; \tilde{\iota})$

Step 3: Make a decision

pull $\tilde{a}_1^*$

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

- **Deterministic DP.**
- **Takes $O(K), O(KT^2), O(KT^K)$ to solve depending on $z_t(\cdot)$.**
- **Obtains one sample of performance bound.**

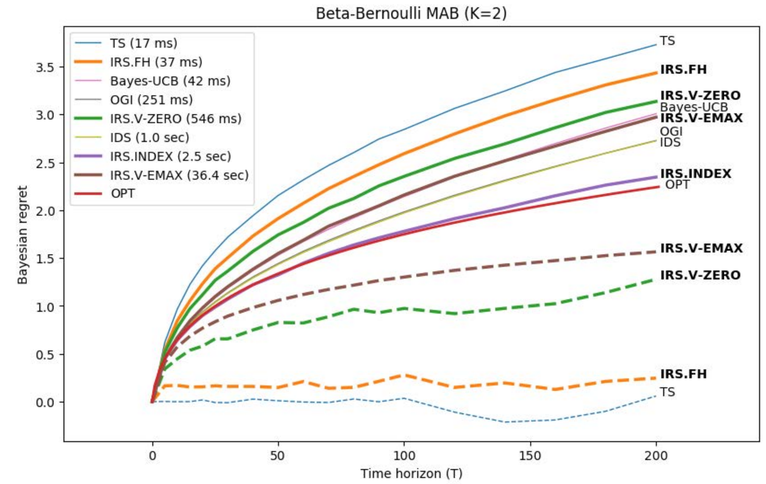
$$\mathbb{E}_{\iota \sim \mathcal{I}(T,\mathbf{y})}\left[\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}; \iota) - z_t(\mathbf{a}_{1:t}; \iota)\right]$$

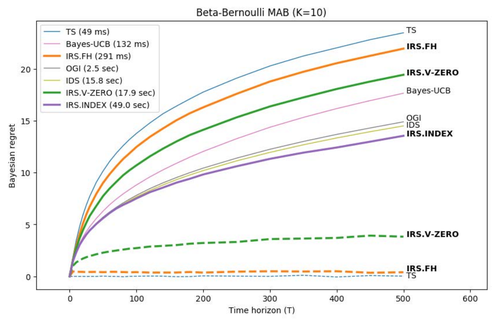
Choice of $z_t(\cdot)$ → Policy, Performance bound

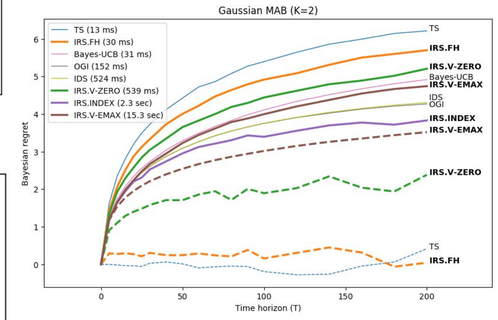
Choice of Penalty Function

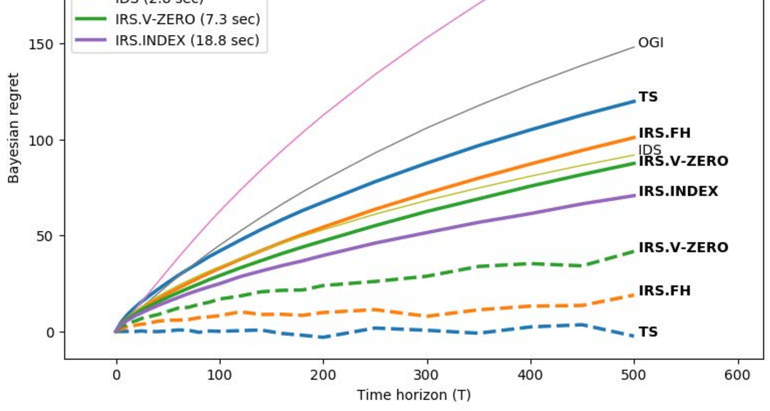
- **Thompson sampling: "find the best arm given parameters" - $O(K)$**
- **IRS.FH: "find the best arm given a finite number of observations" - $O(K)$**
- **IRS.V-Zero: "find the best allocation" - $O(KT^2)$**
- **IRS.V-EMax: "find the best action sequence" - $O(KT^K)$**

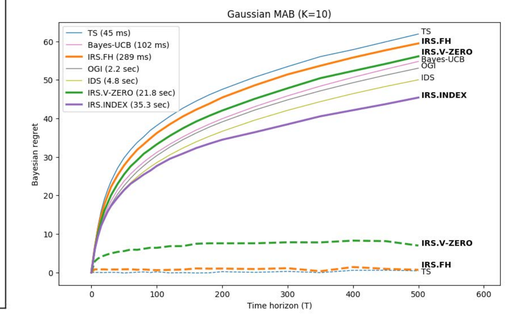
Numerical Simulations











Analysis

- **Weak duality: $V^* \leq W^z$, Strong duality: $V(\pi^{IRS.OPT}) = V^* = W^{IRS.OPT}$**
- **Monotonicity: $W^{IRS.V-Zero} \leq W^{IRS.FH} \leq W^{IRS.V-Opt}$**
- **Subopmality gaps: for Bernoulli MAB,**

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

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

$$W^{\mathrm{Irs.V\text{-}Zero}}(T,\mathbf{y}) - V(\pi^{\mathrm{Irs.V\text{-}Zero}},T,\mathbf{y}) \leq 2K + \sqrt{\log T} \times \left(2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)$$

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