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:
- e.g., at $t=1$, the greedy decision is optimal.

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

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

---

**Contributions**

By introducing information relaxation framework, we can:
- Generalization of Thompson sampling & Bayesian regret benchmark (up to OPT).
- A family of intuitive policies $\pi^z$ and upper bounds $W^z$.

**Choice of Penalty Functions**

<table>
<thead>
<tr>
<th>Penalty function</th>
<th>Policy $\pi^z$</th>
<th>Performance bound $W^z$</th>
<th>Inner problem</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{TS}$</td>
<td>TS</td>
<td>$W_{TS}$</td>
<td>Find a best arm given parameters.</td>
<td>$O(K)$</td>
</tr>
<tr>
<td>$z_{IRS}$</td>
<td>IRS-FH</td>
<td>$W_{IRS}$</td>
<td>Find a best arm given finite observations.</td>
<td>$O(KT^2)$</td>
</tr>
<tr>
<td>$z_{IRS}$</td>
<td>IRS-V-Zero</td>
<td>$W_{IRS}$</td>
<td>Find an optimal allocation of $T$ pulls.</td>
<td>$O(KT^3)$</td>
</tr>
<tr>
<td>$z_{IRS}$</td>
<td>IRS-V-EMAX</td>
<td>$W_{IRS}$</td>
<td>Solve Bellman equations.</td>
<td>–</td>
</tr>
</tbody>
</table>

---

**Information Relaxation**

“Lagrangian relaxation of information structure”

$\Rightarrow$ Allow the DM to use future information, but penalize for the additional benefit.

- Weak duality: if penalties have a zero mean ex-ante, $\sup_{\pi} E[\sum_{t=1}^T r(\pi^z_t, \omega)] \leq \sup_{\pi} E[\sum_{t=1}^T r(\pi^z_t, \omega) - z(\pi^z_t, \omega)]$ where $G$ is a relaxation of $\mathcal{F}$ (i.e., $\mathcal{F}_t \subseteq G_t$).

- Strong duality: equality holds under ideal penalty $z_{ideal}(\pi^z_t, \omega) \equiv E[Q(\pi^z_t, \omega)|G_{t-1}] - E[Q(\pi^z_t, \omega)|F_{t-1}]$ Approximate $Q_1^*$ to obtain a tighter upper bound.

---

**Numerical Experiments**

**Beta-Bernoulli MAB with two arms:**

$\theta_1 \sim \text{Beta}(1, 1), \theta_2 \sim \text{Beta}(0.5, 0.5)$

**Gaussian MAB with heteroscedastic noise:**

$R_{n \sim N(0, \sigma_n^2), \theta_n \sim N(0, 1)}$

---

**Asymptotic Behavior**

When $T=1$, all IRS policies take the optimal action. When $T \rightarrow \infty$, IRS-FH and IRS-V-ZERO behave like TS. Naturally incorporates the horizon constraint.

---

**PERFORMANCE ANALYSIS**

- Monotonicity in performance bounds: for any MAB,
  $W_{TS}(T, y) \geq W_{IRS-FH}(T, y) \geq W_{IRS-V-ZERO}(T, y)$, and $W_{TS}(T, y) \geq W_{IRS-V-EMAX}(T, y)$.

- Improvement in suboptimality gaps: in Beta-Bernoulli MAB,
  $W_{TS}(T, y) - V(T, y) \leq 3K + 2\sqrt{\log T} \times 2\sqrt{KT}$,
  $W_{IRS-FH}(T, y) - V(T, y) \leq 3K + 2\sqrt{\log T} \times 2\sqrt{KT}$,
  $W_{IRS-V-ZERO}(T, y) - V(T, y) \leq 2K + 2\sqrt{\log T} \times (2\sqrt{KT} - \frac{1}{2}\sqrt{KT})$.

---

**REFERENCES**

Brown, Smith, and Sun (2010), Information relaxations and duality in stochastic dynamic programs.

Russo and Van Roy (2014), Learning to optimize via posterior sampling.