Information-Directed Sampling

Presenter: Hao Liang

The Chinese University of Hong Kong, Shenzhen, China

December 24, 2020

Mainly based on:
Daniel Russo, Benjamin Van Roy (2018) Learning to Optimize via Information-Directed Sampling. Operations Research 66(1):230-252. https://doi.org/10.1287/opre.2017.1663

Introduction

- Classical MAB problem requires striking a balance between exploring poorly understood actions and exploiting previously acquired knowledge to attain high rewards.
- There has been significant interest in addressing problems with more complex information structures, in which sampling one action can provide information about other actions.
  - e.g. linear bandit
- UCB and TS can achieve strong performance in the linear bandit problem
- However, these approaches can perform very poorly when faced with more complex information structures.
- Information-directed sampling (IDS) is proposed to deal with online decision making with complex information structures.

Setting

- Bayesian formulation: uncertain quantities are modeled as random variables.
- The decision maker (DM) sequentially chooses actions $(A_t)_{t\in\mathbb{N}}$ from a finite action set $\mathscr{A}$ and observes the corresponding outcomes $(Y_{t,A_t})_{t\in\mathbb{N}}$.
- A random outcome $Y_{t,a} \in \mathscr{Y}$ associated with each action $a \in \mathscr{A}$ and time $t \in \mathbb{N}$.
- $Y_t \equiv (Y_{t,a})_{a\in\mathscr{A}}$ the vector of outcomes at time $t \in \mathbb{N}$.
- There is a random variable θ such that conditioned on $\theta, (Y_t)_{t\in\mathbb{N}}$ is an iid sequence.
  - MAB with independent arms: $Y_t = \theta + \eta_t$, η_t iid zero-mean noise
- Randomness in θ captures the DM's prior uncertainty about the environment, and the remaining randomness in Y_t captures intrinsic randomness in observed outcomes.

Policy

- A_t is chosen based on the history of observations $\mathscr{F}_t = \left(A_1, Y_{1,A_1}, \ldots, A_{t-1}, Y_{t-1,A_{t-1}}\right)$ up to time t.
- A randomized policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ is a sequence of deterministic functions, where $\pi_t\left(\mathscr{F}_t\right)$ specifies a probability distribution over the action set $\mathscr{A}$.
- Let $\mathscr{D}(\mathscr{A})$ denote the set of probability distributions over $\mathscr{A}$.
- $A_t \sim \pi_t\left(\mathscr{F}_t\right) \in \mathscr{D}(\mathscr{A})$
- With some abuse of notation, write $\pi_t = \pi_t(\mathscr{F}_t)$, where $\pi_t(a) = \mathbb{P}\left(A_t = a \mid \mathscr{F}_t\right)$.

Regret

- The agent associates a reward $R(y)$ with each outcome $y \in \mathscr{Y}$ via a fixed and known function $R() : \mathscr{Y} \to \mathbb{R}$.
- Let $R_{t,a} = R\left(Y_{t,a}\right)$ denote the realized reward of action a at time t.
- Uncertainty about θ induces uncertainty about $A^* \in \arg\max_{a \in \mathscr{A}} \mathbb{E}\left[R_{1,a} \mid \theta\right]$
- The expected Bayesian regret

$$\mathbb{E}[\mathrm{Regret}(T,\pi)] = \mathbb{E}\left[\sum_{t=1}^{T}\left(R_{t,A^*} - R_{t,A_t}\right)\right] = \mathbb{E}[\mathbb{E}[\sum_{t=1}^{T}(R_{t,A^*} - R_{t,A_t})|\theta]],$$

where the expectation is taken over the randomness in the actions A_t and the outcomes Y_t, and over the prior distribution over θ.

Further notations

- Set $\alpha_t(a) = \mathbb{P}\left(A^* = a \mid \mathscr{F}_t\right)$ to be the posterior distribution of A^*.
- KL divergence between P and Q is $D_{\mathrm{KL}}(P\|Q) = \int \log\left(\frac{dP}{dQ}\right) dP$
- Shannon entropy $H(P) = -\sum_{x\in\mathscr{X}} P(x)\log(P(x))$
- The mutual information under the posterior distribution between X_1 and X_2

$$I_t\left(X_1; X_2\right) := D_{\mathrm{KL}}\left[\mathbb{P}\left(\left(X_1, X_2\right) \in \cdot \mid \mathscr{F}_t\right) \| \mathbb{P}\left(X_1 \in \cdot \mid \mathscr{F}_t\right) \mathbb{P}\left(X_2 \in \cdot \mid \mathscr{F}_t\right)\right]$$

- $I_t\left(X_1; X_2\right)$ is a random variable because of its dependence on $\mathbb{P}\left(\cdot \mid \mathscr{F}_t\right)$.

Further notations (Cont')

- Information gain from an action a is

$$g_t(a) := I_t\left(A^*; Y_{t,a}\right) = \mathbb{E}\left[H\left(\alpha_t\right) - H\left(\alpha_{t+1}\right) \mid \mathscr{F}_t, A_t = a\right]$$

- Expected instantaneous regret of action a is $\Delta_t(a) := \mathbb{E}\left[R_{t,A^*} - R_{t,a} \mid \mathscr{F}_t\right]$
- $g_t(\pi) := \sum_{a \in \mathscr{A}} \pi(a) g_t(a)$ and $\Delta_t(\pi) := \sum_{a \in \mathscr{A}} \pi(a) \Delta_t(a)$
- Information ratio $\Psi_t(\pi) := \frac{\Delta_t(\pi)^2}{g_t(\pi)}$
- $\mathbb{E}_t[\cdot] = \mathbb{E}\left[\cdot \mid \mathscr{F}_t\right]$ and $\mathbb{P}_t(\cdot) = \mathbb{P}\left(\cdot \mid \mathscr{F}_t\right)$

Motivation

- In principle Bayes-optimal policy can be computed via dynamic programming.
- Computing or even storing this Bayes-optimal policy is generally infeasible.
- How to develop computationally efficient heuristics?
- IDS is motivated by accounting for kinds of information that alternatives fail to address:
  - Indirect information
  - Cumulating information
  - Irrelevant information
- Refer to IDS as a design principle rather than an algorithm.
  - Does not specify basic computational steps but only an abstract objective.
  - Need to design tractable algorithms for specific problem classes.

Information-Directed Sampling

- Information ratio (IR) $\Psi_t(\pi) = \frac{\Delta_t(\pi)^2}{g_t(\pi)}$ measures the squared regret incurred per-bit of information acquired about the optimum.
- IDS balances between exploration and exploitation via minimizing IR at each round

$$\pi_t^{\mathrm{IDS}} \in \arg\min_{\pi \in \mathscr{D}(\mathscr{A})} \Psi_t(\pi)$$

- IDS myopically minimizes this notion of cost-per-bit of information in each period.
- IDS is stationary randomized policy
  - Each action is randomly sampled
  - This action distribution is determined by the posterior distribution of θ and otherwise independent of the time period

The role of randomization in policy

- Two actions $\mathscr{A} = \{a_1, a_2\}$
- R_{a_1} is known to be distributed Ber($\frac{1}{2}$), $R_{a_2} \sim \begin{cases} \text{Ber}(\frac{3}{4}) & \text{w.p. } p_0 \\ \text{Ber}(\frac{1}{4}) & \text{w.p. } 1 - p_0 \end{cases}$
- Consider a stationary deterministic policy where each action A_t is a deterministic function of the posterior probability p_{t-1}
- Suppose that for some $p_0 > 0$, the policy selects $A_1 = a_1$
- $p_t = p_0$ and $A_t = a_1$ for all t and expected regret grows linearly with time.
- If $A_1 = a_2$ for all $p_0 > 0$ then $A_t = a_2$ for all t

The role of randomization in policy (Cont')

- For any deterministic stationary policy, there exists a prior probability p_0 such that expected regret grows linearly with time.
- A sublinear bound on (worst case) expected regret of IDS can be established.
- The expected regret of IDS does not grow linearly as does that of any stationary deterministic policy for the preceding example.
- Increasing complexity? An important property simplifies solutions.
- There exists a distribution with support of at most two actions that attains the minimum.

Alternative design principles

- UCB: $A_t \in \arg\max_{a \in \mathscr{A}} B_t(a)$ with maximal upper confidence bound
- TS: $\pi_t^{\mathrm{TS}} = \alpha_t = \mathbb{P}\left(A^* = \cdot \mid \mathscr{F}_t\right)$
  - also called probability matching: matching action distribution to posterior distribution of optimal action
- Specific UCB and TS algorithms are known to be asymptotically efficient for MAB with independent arms and satisfy strong regret bounds for problems with dependent arms.
- UCB and TS do not pursue indirect information and thus can perform very poorly relative to IDS for some natural problem classes.
- They restrict attention to sampling actions that have some chance of being optimal.

Example 2: A Revealing Action

- $\mathscr{A} = \{0, 1, \ldots, K\}$ and θ is drawn uniformly at random from a finite set $\Theta = \{1, \ldots, K\}$
- $Y_{t,a} = R_{t,a}$. Under θ, the reward of action a is $R_{t,a} = \begin{cases} 1 & \theta = a, \\ 1/2\theta & a = 0, \\ 0 & otherwise. \end{cases}$
- Action 0 never yields the maximal reward, and is therefore never selected by TS or UCB.
- They will select among actions $\{1, \ldots, K\}$, ruling out only a single action at a time until a reward 1 is earned and the optimal action is identified.
- Their expected regret therefore grows linearly in K.

Regret bounds

- Establishes regret bounds for IDS for several classes of online optimization problems
  - Uncorrelated arms
  - Linear bandit
  - Full information
- These regret bounds follow from the information theoretic analysis of TS (Russo and Van Roy 2016), where regret bound for any policy is bounded in terms of its IR.
- Because the IR of IDS is always smaller than that of TS, the bounds on regret of TS immediately yield regret bounds for IDS.

General bound

Proposition 1.

For any policy $\pi = (\pi_1, \pi_2, \pi_3, \ldots)$ *and time* $T \in \mathbb{N}$,

$$\mathbb{E}[\mathrm{Regret}(T, \pi)] \leqslant \sqrt{\bar{\Psi}_T(\pi) H\left(\alpha_1\right) T},$$

where $\bar{\Psi}_T(\pi) \equiv \frac{1}{T} \sum_{t=1}^{I} \mathbb{E}_\pi\left[\Psi_t\left(\pi_t\right)\right]$ *is the average expected information ratio under* π.

Corollary 0.1.

For any $\pi = (\pi_1, \pi_2, \ldots)$ *such that* $\Psi_t\left(\pi_t\right) \leqslant \lambda$ *almost surely for each* $t \in [T]$. *Then,*

$$\mathbb{E}[\mathrm{Regret}(T, \pi)] \leqslant \sqrt{\lambda H\left(\alpha_1\right) T}$$

- $H(\alpha_1)$ captures the magnitude of the decision-maker's prior uncertainty about which action is optimal.

Specialized Bounds on the Minimal Information Ratio

- The bounds on the IR roughly captures the extent to which sampling some actions allows the DM to make inferences about other actions.
  - Worst case/independent arms: $\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leq |\mathscr{A}|/2$
  - Best case/full information: $\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leq 1/2$
  - Intermediate case/linear bandit: $\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leq d/2$
- The proofs of these bounds follow from the analysis of TS and the fact that $\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leqslant \Psi_t\left(\pi_t^{\mathrm{TS}}\right)$
- Some work by Bubeck et al. (2015) and Bubeck and Eldan (2016) bounds the IR when the reward function is convex.
- Assumption 1: $\sup_{\bar{y}\in\mathscr{Y}} R(\bar{y}) - \inf_{\underline{y}\in\mathscr{Y}} R(\underline{y}) \leqslant 1$

Worst case

Proposition 2.

For any $t \in \mathbb{N}, \Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leqslant |\mathscr{A}|/2$ *almost surely.*

- Combining Proposition 2 with Corollary 0.1 shows that $\mathbb{E}\left[\operatorname{Regret}\left(T, \pi^{\mathrm{IDS}}\right)\right] \leqslant \sqrt{\frac{1}{2}|\mathscr{A}| H\left(\alpha_1\right) T}$.
- This bound holds for general MAB problems with arbitrary information structure. Can be much smaller under specific information structures.

Full information

- The outcome $Y_{t,a}$ is perfectly revealed by observing $Y_{t,\tilde{a}}$ for some $\tilde{a} \neq a$.

Proposition 3.

Suppose for each $t \in \mathbb{N}$ there is a random variable $\mathrm{Z}_t : \Omega \rightarrow \mathscr{Z}$ such that for each $a \in \mathscr{A}, Y_{t,a} = (a, \mathrm{Z}_t)$. Then for all $t \in \mathbb{N}, \Psi_t \left(\pi_t^{\mathrm{IDS}}\right) \leqslant \frac{1}{2}$ almost surely.

- Combining this result with Corollary 0.1 shows that $\mathbb{E}\left[\mathrm{Regret}\left(T, \pi^{\mathrm{IDS}}\right)\right] \leqslant \sqrt{\frac{1}{2} H\left(\alpha_1\right) T}$.
- A worst-case bound on $H(\alpha_1)$ yields $\mathbb{E}\left[\mathrm{Regret}\left(T, \pi^{\mathrm{IDS}}\right)\right] \leqslant \sqrt{\frac{1}{2} \log(|\mathscr{A}|) T}$.
- Dani et al. (2007) show this bound is order optimal: $\inf_\pi \mathbb{E}[\mathrm{Regret}(T, \pi)] \geqslant c_0 \sqrt{\log(|\mathscr{A}|) T}$

Linear bandit

- Observations from taking one action allow the DM to make inferences about other actions.

Proposition 4.

If $\mathscr{A} \subset \mathbb{R}^d, \Theta \subset \mathbb{R}^d$, *and* $\mathbb{E}\left[R_{t,a} \mid \theta\right] = a^T\theta$ *for each action* $a \in \mathscr{A}$, *then* $\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leqslant d/2$ *almost surely for all* $t \in \mathbb{N}$.

- This result shows the inequalities $\mathbb{E}\left[\mathrm{Regret}\left(T, \pi^{\mathrm{IDS}}\right)\right] \leqslant \sqrt{\frac{1}{2} H\left(\alpha_1\right) dT} \leqslant \sqrt{\frac{1}{2}\log(|\mathscr{A}|)dT}$ for linear bandit problems.
- Dani et al. (2007) again show this bound is order optimal in the sense that, when the action set is $\mathscr{A} = \{0,1\}^d$ such that $\inf_\pi \mathbb{E}[\mathrm{Regret}(T,\pi)] \geqslant c_0\sqrt{\log(|\mathscr{A}|)dT}$.

Computational methods

- Provide guidance and examples of designing efficient computational methods that implement IDS for specific problem classes.
- Assume posterior distributions can be efficiently computed and stored, e.g., tractable finite uncertainty sets or conjugate priors.
- Focus in the problem of generating an action A_t given the posterior distribution over θ.
- Two of the algorithms approximate IDS using samples from the posterior distribution.

Evaluating the Information Ratio

- Given a finite action set $\mathscr{A} = \{1, \ldots, K\}$, view action distribution π as a K-dimensional vector of problem probabilities.
- No general efficient procedure for computing $\vec{\Delta}$ and $\vec{g}$ given a posterior distribution, require computing integrals over possibly high-dimensional spaces.
- Such computation can often be carried out efficiently by leveraging the functional form of the specific posterior distribution and often requires numerical integration.

Finite Sets

- $\Theta = \{1, \ldots, L\}$, $\mathscr{A} = \{1, \ldots, K\}$, $\mathscr{Y} = \{1, \ldots, N\}$.
- The reward function $R : y \mapsto \mathbb{R}$ is arbitrary.
- Let p_1 be the prior probability mass function of θ and let $q_{\theta,a}(y)$ be the probability, conditioned on θ, of observing y when action a is selected.
- p_t can be computed recursively via Bayes' rule:

$$p_{t+1}(\theta) \leftarrow \frac{p_t(\theta) q_{\theta, A_t}\left(Y_{t, A_t}\right)}{\sum_{\theta' \in \Theta} p_t\left(\theta'\right) q_{\theta', A_t}\left(Y_{t, A_t}\right)}$$

Finite Sets (Cont')

Algorithm 1 (finiteIR (L, K, N, R, p, q))

$1 : \Theta_a \leftarrow \{\theta \mid a = \arg\max_{a'} \Sigma_y q_{\theta,a'}(y) R(y)\}, \quad \forall\theta$

2: $p(a^*) \leftarrow \sum_{\theta\in\Theta_{a^*}} p(\theta), \quad \forall a^*$

3: $p_a(y) \leftarrow \Sigma_\theta p(\theta) q_{\theta,a}(y), \quad \forall a, y, \theta$

$4 : p_a(a^*, y) \leftarrow \frac{1}{p(a^*)} \sum_{\theta\in\Theta_{a^*}} q_{\theta,a}(y), \quad \forall a, y, a^*$

$5 : R^* \leftarrow \sum_a \sum_{\theta\in\Theta_a} \Sigma_y p(\theta) q_{\theta,a}(y) R(y)$

$6 : \vec{g}_a \leftarrow \Sigma_{a^*,y} p_a(a^*, y) \log \frac{p_a(a^*,y)}{p(a^*) p_a(y)}, \quad \forall a$

$7 : \vec{\Delta}_a \leftarrow R^* - \sum_\theta p(\theta) \Sigma_y q_{\theta,a}(y) R(y), \quad \forall a$

8: return $\vec{\Delta}, \vec{g}$

Optimizing the Information Ratio

- IDS selects an action by solving

$$\min_{\pi \in \mathscr{S}_K} \frac{\left(\pi^\top \vec{\Delta}\right)^2}{\pi^\top \vec{g}} \tag{1}$$

where $\mathscr{S}_K = \left\{\pi \in \mathbb{R}_+^K : \Sigma_k \pi_k = 1\right\}$ is the K-dimensional unit simplex.

Proposition 5.

For all $\vec{\Delta}, \vec{g} \in \mathbb{R}_+^K$ such that $\vec{g} \neq 0$, the function $\pi \mapsto \left(\pi^\top \vec{\Delta}\right)^2 / \pi^\top \vec{g}$ is convex on $\left\{\pi \in \mathbb{R}^K : \pi^\top \vec{g} > 0\right\}$. Moreover, this function is minimized over $\mathscr{S}_K$ by some π^* for which $|\{k : \pi_k^* > 0\}| \leqslant 2$

Optimizing the Information Ratio (Cont')

- While IDS is a randomized policy, it suffices to randomize over two actions.
- q can be computed by solving for the first-order necessary condition or approximated by a bisection method.
- The compute time of this algorithm scales with K^2.

Algorithm $3(\mathrm{IDSAction}(K, \vec{\Delta}, \vec{g}))$

$1: q_{a,a'} \leftarrow \arg\min_{q' \in [0,1]} \left[q' \vec{\Delta}_a + (1-q')\, \vec{\Delta}_{a'}\right]^2 / \left[q' \vec{g}_a + (1-q')\, \vec{g}_{a'}\right], \quad \forall a < K, a' > a$

$2: (a^*, a^{**}) \leftarrow \arg\min_{a<K, a'>a} \left[q_{a,a'} \vec{\Delta}_a + (1-q_{a,a'})\, \vec{\Delta}_{a'}\right]^2 / \left[q_{a,a'} \vec{g}_a + (1-q_{a,a'})\, \vec{g}_{a'}\right]$

3: Sample $b \sim$ Bernoulli (q_{a^*,a^*})

4: return $ba^* + (1-b)a^{**}$

Approximating the Information Ratio

- The dominant source of complexity in computing $\vec{\Delta}$ and $\vec{g}$ is in the calculation of requisite integrals, which can require integration over high-dimensional spaces.
- Replace integrals with sample-based estimates.
- Takes as input M representative samples of θ

Algorithm $2\,($ Sample$\,\mathrm{IR}\left(K, q, R, M, \theta^1, \ldots, \theta^M\right))$

$1: \hat{\Theta}_a \leftarrow \left\{m \mid a = \arg\max_{a'} \sum_y q_{\theta^m, a'}(y) R(y)\right\}$

2: $\hat{p}\left(a^*\right) \leftarrow \left|\hat{\Theta}_{a^*}\right| / M, \quad \forall a^*$

3: $\hat{p}_a(y) \leftarrow \Sigma_m q_{a, \theta^m}(y) / M, \quad \forall y$

4: $\hat{p}_a\left(a^*, y\right) \leftarrow \Sigma_{m \in \Theta_a} q_{a, \theta^m}(y) / M, \quad \forall a^*, y$

$5: \hat{R}^* \leftarrow \sum_{a, y} \hat{p}_a(a, y) R(y)$

$6: \vec{g}_a \leftarrow \Sigma_{a^*, y} \hat{p}_a\left(a^*, y\right) \log \frac{\hat{p}_a\left(a^*, y\right)}{\hat{p}\left(a^*\right) \hat{p}_a(y)}, \quad \forall a$

7: $\vec{\Delta}_a \leftarrow R^* - M^{-1} \Sigma_m \Sigma_y q_{\theta^m, a}(y) R(y), \quad \forall a$

8: return $\vec{\Delta}, \vec{g}$

Variance-based information ratio

$$
\begin{aligned}
g_t(a) =& I_t\left(A^*; Y_{t,a}\right) \\
=& \sum_{a^* \in \mathbb{A}} \mathbb{P}_t\left(A^* = a^*\right) \cdot D_{\mathrm{KL}}\left(\mathbb{P}_t\left(Y_{t,a} = \cdot \mid A^* = a^*\right) \| \mathbb{P}_t\left(Y_{t,a} = \cdot\right)\right) \\
\geqslant& \sum_{a^* \in A} \mathbb{P}_t\left(A^* = a^*\right) \cdot D_{\mathrm{KL}}\left(\mathbb{P}_t\left(R_{t,a} = \cdot \mid A^* = a^*\right) \| \mathbb{P}_t\left(R_{t,a} = \cdot\right)\right) \\
\geqslant& \, 2 \sum_{a^* \in \mathscr{A}} \mathbb{P}_t\left(A^* = a^*\right)\left(\mathbb{E}_t\left[R_{t,a} \mid A^* = a^*\right] - \mathbb{E}_t\left[R_{t,a}\right]\right)^2 \\
=& 2\mathbb{E}_t\left[\left(\mathbb{E}_t\left[R_{t,a} \mid A^*\right] - \mathbb{E}_t\left[R_{t,a}\right]\right)^2\right] \\
=& 2 \operatorname{Var}_t\left(\mathbb{E}_t\left[R_{t,a} \mid A^*\right]\right)
\end{aligned}
$$

- Let $v_t(a) := \operatorname{Var}_t\left(\mathbb{E}_t\left[R_{t,a} \mid A^*\right]\right)$
- Implication: Actions with high variance $v_t(a)$ must yield substantial information about which action is optimal.

Variance-based information ratio

- Variance-based IDS

$$\min_{\pi \in \mathscr{S}_K} \frac{\left(\pi^\top \vec{\Delta}\right)^2}{\pi^\top \vec{v}}$$

Proposition 6.

Suppose $\sup_y R(y) - \inf_y R(y) \leqslant 1$ *and*

$$\pi_t \in \underset{\pi \in \mathscr{S}_K}{\arg\min} \frac{\Delta_t(\pi)^2}{v_t(\pi)}$$

Then $\Psi_t(\pi_t) \leqslant |\mathscr{A}|/2$. *Moreover, if* $\mathscr{A} \subset \mathbb{R}^d, \Theta \subset \mathbb{R}^d$, *and* $\mathbb{E}\left[R_{t,a} \mid \theta\right] = a^T \theta$ *for each action* $a \in \mathscr{A}$, *then* $\Psi_t(\pi_t) \leqslant d/2$.

Variance-based IDS: linear bandit

- $\mathscr{A} = \{1, \ldots, K\}, y = \mathbb{R}$, and $R(y) = y$. $\theta \in \mathbb{R}^d \sim \mathcal{N}(\mu_1, \Sigma_1)$.
- A known feature matrix $\Phi = [\Phi_1, \cdots, \Phi_K] \in \mathbb{R}^{d \times K}$, $Y_{t,A_t} | \theta, A_t \sim \mathcal{N}\left(\Phi_{A_t}^\top \theta, \eta^2\right)$.
- $\Sigma_{t+1} = \left(\Sigma_t^{-1} + \Phi_{A_t} \Phi_{A_t}^\top / \eta^2\right)^{-1}$, $\mu_{t+1} = \Sigma_{t+1}\left(\Sigma_t^{-1} \mu_t + Y_{t,A_t} \Phi_{A_t} / \eta^2\right)$,
- Let $\mu_t^a = \mathbb{E}_t\left[\theta \mid A^* = a\right]$ and $L_t = \mathbb{E}_t\left[\left(\mu_t^{A^*} - \mu_t\right)\left(\mu_t^{A^*} - \mu_t\right)^\top\right]$
- $$\begin{aligned} v_t(a) &= \mathrm{Var}_t\left(\mathbb{E}_t\left[R_{t,a} \mid A^*\right]\right) \\ &= \mathrm{Var}_t\left(\mathbb{E}_t\left[\Phi_a^\top \theta \mid A^*\right]\right) \\ &= \mathrm{Var}_t\left(\Phi_a^\top \mathbb{E}_t\left[\theta \mid A^*\right]\right) \\ &= \Phi_a^\top \mathbb{E}_t\left[\left(\mu_t^{A^*} - \mu_t\right)\left(\mu_t^{A^*} - \mu_t\right)^\top\right] \Phi_a \\ &= \Phi_a^\top L_t \Phi_a \end{aligned}$$

Variance-based IDS: linear bandit (Cont')

Algorithm 3 (linearSampleVIR $\left(K, d, M, \theta^1, \ldots, \theta^M\right)$)

$1: \hat{\mu} \leftarrow \Sigma_m \theta^m / M$

$2: \hat{\Theta}_a \leftarrow \left\{m : \left(\Phi^\top \theta^m\right)_a = \max_{a'} \left(\Phi \theta^m\right)_{a'}\right\}, \quad \forall a$

3: $\hat{p}^*(a) \leftarrow \left|\hat{\Theta}_a\right| / M, \quad \forall a$

$4: \hat{\mu}^a \leftarrow \Sigma_{\theta \in \hat{\Theta}_a} \theta / \left|\hat{\Theta}_a\right|, \quad \forall a$

$5: \hat{L} \leftarrow \Sigma_a \hat{p}^*(a) \left(\hat{\mu}^a - \hat{\mu}\right) \left(\hat{\mu}^a - \hat{\mu}\right)^\top$

$6: \rho^* \leftarrow \Sigma_a \hat{p}^*(a) \Phi_a^\top \hat{\mu}^a$

$7: \vec{v}_a \leftarrow \Phi_a^\top \hat{L} \Phi_a^\top, \quad \forall a$

$8: \vec{\Delta}_a \leftarrow \rho^* - \Phi_a^\top \hat{\mu}, \quad \forall a$

9: return $\vec{\Delta}, \vec{v}$

Beta-Bernoulli Bandit

- The mean reward of each arm is drawn from Beta(1,1)/$\mathcal{U}(0,1)$

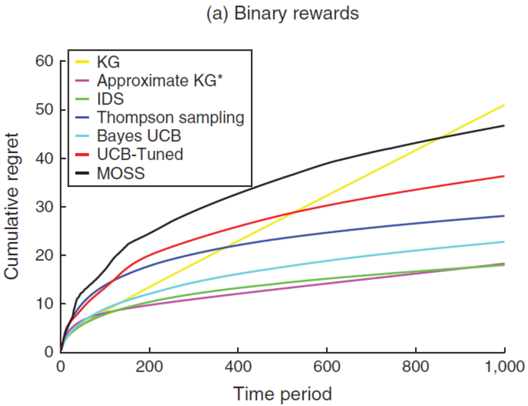

Figure: 1,000 independent trials of an experiment with 10 arms and a time horizon of 1,000

Beta-Bernoulli Bandit (Cont')

	Time horizon agnostic						Optimized for time horizon		
Algorithm	IDS	V-IDS	TS	Bayes UCB	UCB1	UCB-Tuned	MOSS	KG	KG*
Mean regret	18.0	18.1	28.1	22.8	130.7	36.3	46.7	51.0	18.4
Standard error	0.4	0.4	0.3	0.3	0.4	0.3	0.2	1.5	0.6
Quantile 0.10	3.6	5.2	13.6	8.5	104.2	24.0	36.2	0.7	2.9
Quantile 0.25	7.4	8.1	18.0	12.5	117.6	29.2	40.0	2.9	5.4
Quantile 0.50	13.3	13.5	25.3	20.1	131.6	35.2	45.2	11.9	8.7
Quantile 0.75	22.5	22.3	35.0	30.6	144.8	41.9	51.0	82.3	16.3
Quantile 0.90	35.6	36.5	46.4	40.5	154.9	49.5	57.9	159.0	46.9
Quantile 0.95	51.9	48.8	53.9	47.0	160.4	54.9	64.3	204.2	76.6

Figure: Realized Regret Over 2,000 Trials in Bernoulli Experiment

Independent Gaussian Bandit

Table 2. Realized Regret Over 2,000 Trials in Independent Gaussian Experiment

	Time horizon agnostic				Optimized for time horizon		
Algorithm	V-IDS	TS	Bayes UCB	GPUCB	Tuned GPUCB	KG	KG*
Mean regret	58.4	69.1	63.8	157.6	53.8	65.5	50.3
Standard error	1.7	0.8	0.7	0.9	1.4	2.9	1.9
Quantile 0.10	24.0	39.2	34.7	108.2	24.2	16.7	19.4
Quantile 0.25	30.3	47.6	43.2	130.0	30.1	20.8	24.0
Quantile 0.50	39.2	61.8	57.5	156.5	41.0	25.9	29.9
Quantile 0.75	56.3	80.6	76.5	184.2	58.9	36.4	40.3
Quantile 0.90	104.6	104.5	97.5	207.2	86.1	155.3	74.7
Quantile 0.95	158.1	126.5	116.7	222.7	112.2	283.9	155.6

Table 3. Competitive Performance Without Knowing the Time Horizon

Time horizon T	10	25	50	75	100	250	500	750	1,000	2,000
Regret of V-IDS	9.8	16.1	21.1	24.5	27.3	36.7	48.2	52.8	58.3	68.4
Regret of KG(T)	9.2	15.3	20.5	22.9	25.4	35.2	45.3	52.3	62.9	80.0

Figure: $R_a \sim \mathcal{N}(\theta_a, 1)$, $\theta_a \sim \mathcal{N}(0, 1)$

Asymptotic Optimality

- The seminal work of Lai and Robbins (1985) provides the asymptotic lower bound $\liminf_{T\to\infty} \frac{\mathbb{E}[\text{Regret}(T,\pi)|\theta]}{\log T} \geqslant \sum_{a\neq A^*} \frac{\theta_{A^*}-\theta_a}{D_{\text{KL}}(\theta_{A^*}\|\theta_a)} := c(\theta).$
- when applied with an independent uniform prior, both Bayes UCB and TS are known to attain this lower bound (Kaufmann et al. 2012a, b).

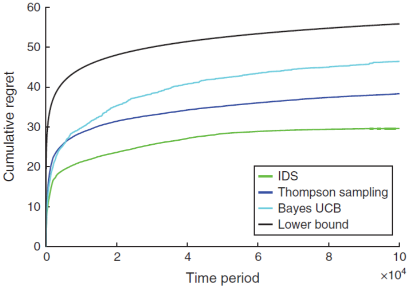

Figure: $\theta = (0.3, 0.2, 0.1)$. 10,000 time periods. 200 independent trials. Uniform prior.

Linear Bandit

- $a \in \mathbb{R}^5$, $R_a = a^T\theta + \epsilon_t$ where $\theta \sim \mathcal{N}(0, 10I)$ and $\epsilon_t \sim \mathcal{N}(0, 1)$
- $\mathscr{A}$ contains 30 actions, each with features $\sim \mathcal{U}([-1/\sqrt{5}, 1/\sqrt{5}])$

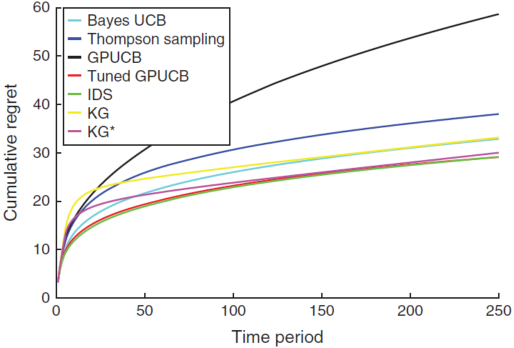

Figure: Regret in Linear-Gaussian Model

Runtime Comparison

Table 6. Bernoulli Experiment: Compute Time per Decision in Seconds

Arms	IDS	V-IDS	TS	Bayes UCB	UCB1	KG	Approx KG*
10	0.011013	0.01059	0.000025	0.000126	0.000008	0.000036	0.074618
30	0.047021	0.047529	0.000023	0.000147	0.000005	0.000017	0.215145
50	0.104328	0.10203	0.000024	0.000176	0.000005	0.000017	0.358505
70	0.18556	0.178689	0.000028	0.000167	0.000005	0.000017	0.494455

Table 7. Independent Gaussian Experiment: Compute Time per Decision in Seconds

Arms	V-IDS	TS	Bayes UCB	GPUCB	KG	KG*
10	0.00298	0.000008	0.00002	0.00001	0.000146	0.001188
30	0.012597	0.000005	0.000009	0.000005	0.000097	0.003157
50	0.023084	0.000006	0.000009	0.000005	0.000094	0.005146
70	0.03913	0.000006	0.000009	0.000005	0.000098	0.006364

Table 8. Linear Gaussian Experiment: Compute Time per Decision in Seconds

Arms	Dimension	V-IDS	TS	Bayes UCB	GPUCB	KG	KG*
15	3	0.004305	0.000178	0.000139	0.000048	0.002709	0.311935
30	5	0.008635	0.000064	0.000048	0.000038	0.004789	0.589998
50	20	0.026222	0.000077	0.000083	0.000068	0.008356	1.051552
100	30	0.079659	0.000115	0.000148	0.00013	0.017034	2.067123

Figure