An Information-Theoretic Analysis of Thompson Sampling

Hao Liang

The Chinese University of Hongkong, Shenzhen

May 5, 2019

Outline

- Consider the problem of repeated decision making in the presence of model uncertainty, i.e., online optimization problem.
- *Partial feedback* leads to inherent tradeoff between *exploration and exploitation*.

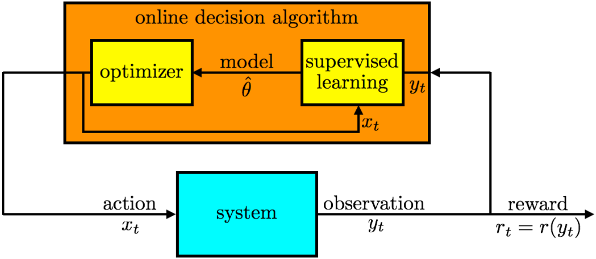

Figure: **Online decision algorithm**

Introduction

- *Thompson sampling*, *posterior sampling*, or *probability matching* is a simple algorithm to solve online optimization with partial feedback.
- We will establish performance guarantees in the form of regret bounds for TS based on an information-theoretic analysis.

Problem Formulation

- The decision–maker sequentially chooses actions $(A_t)_{t\in\mathbb{N}}$ from the action set $\mathcal{A}$ and observes the corresponding outcomes $(Y_{t,A_t})_{t\in\mathbb{N}}$.
- Let $Y_t \equiv (Y_{t,a})_{a\in\mathcal{A}}$ be the vector of all outcomes at time $t \in \mathbb{N}$ which follows the "true outcome distribution" p^*. Here p^* itself is randomly drawn from the family of distributions $\mathcal{P}$.
- We assume that, conditioned on p^*, $(Y_t)_{t\in\mathbb{N}}$ is an iid sequence distributed according to p^*.
- A fixed and known reward function maps each outcome $y \in \mathcal{Y}$ to some reward $R(y)$
- The true optimal action $A^* \in \arg\max_{a\in\mathcal{A}} \mathbb{E}\left[R(Y_{t,a})|p^*\right] = \mathbb{E}\left[R(Y_{t,a})|p_a^*\right]$ is also a random variable.

Regret and Randomized policies

- Our objective is to minimize the *Bayesian regret*

$$\mathbb{E}\left[\text{Regret}(T)\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{T}\left[R(Y_{t,A^*}) - R(Y_{t,A_t})\right]|p^*\right]\right], \qquad (1)$$

  the expectation is taken over the randomness in the actions A_t and the outcomes Y_t, and over the prior distribution over p^*.
- Actions are chosen based on the history of past observations and possibly some external source of randomness $(U_t)_{t\in\mathbb{N}}$. $(U_t)_{t\in\mathbb{N}}$ is white and independent of outcomes $\{Y_{t,a}\}_{t\in\mathbb{N},a\in\mathcal{A}}$, and p^*.
- The filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ is the sigma–algebra generated by $\left(A_1, Y_{1,A_1}, ..., A_{t-1}, Y_{t-1,A_{t-1}}\right)$. Given the history, A_t is random only through its dependence on U_t.
- *Randomized policy* π: An action is chosen at time t by randomizing according to $\pi_t(\cdot) = \mathbb{P}(A_t \in \cdot|\mathcal{F}_t)$.

Further Asumptions

Assumption 1

$$\sup_{\overline{y}\in\mathcal{Y}} R(\overline{y}) - \inf_{\underline{y}\in\mathcal{Y}} R(\underline{y}) \leq 1.$$

Assumption 2

$\mathcal{A}$ is finite.

Basic Measures and Relations in Information Theory

- Let $P(X) = \mathbb{P}(X \in \cdot)$ denote the distribution function of random variable X. Similarly, define $P(X|Y) = \mathbb{P}(X \in \cdot|Y)$ and $P(X|Y = y) = \mathbb{P}(X \in \cdot|Y = y)$.
- Suppose X is supported on a finite set $\mathcal{X}$. The *Shannon entropy* of X is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x).$$

- The first fact establishes uniform bounds on the entropy of a probability distribution.

Fact 1

$0 \leq H(X) \leq \log(|\mathcal{X}|)$.

Basic Measures and Relations in Information Theory

- The entropy of X conditional on a random variable $Y = y$ is

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} \mathbb{P}\left(X = x|Y = y\right) \log \mathbb{P}(X = x|Y = y)$$

- The conditional entropy of X given Y is,

$$H(X|Y) = \mathbb{E}_Y \left[-\sum_{x \in \mathcal{X}} \mathbb{P}\left(X = x|Y\right) \log \mathbb{P}(X = x|Y) \right],$$

- For two probability measures P and Q, if P is absolutely continuous with respect to Q, the *Kullback–Leibler divergence* between them is

$$D(P||Q) = \int \log \left(\frac{dP}{dQ} \right) dP \tag{2}$$

Basic Measures and Relations in Information Theory

Fact 2

(Gibbs' inequality) For any probability distributions P and Q such that P is absolutely continuous with respect to Q, $D\left(P||Q\right) \geq 0$ with equality if and only if $P = Q$ P–almost everywhere.

- The *mutual information* between X and Y

$$I(X;Y) = D\left(P(X,Y) \,||\, P\left(X\right) P\left(Y\right)\right) \tag{3}$$

the next fact states that the mutual information between X and Y is the expected reduction in the entropy due to observing Y

Fact 3

(Entropy reduction form of mutual information)

$$I\left(X;Y\right) = H(X) - H(X|Y)$$

- The mutual information between X and Y, conditional on a third random variable Z is

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z),$$

it can also be expressed as

$$I(X;Y|Z) = \mathbb{E}_Z \left[D\left(P\left((X,Y)|Z\right) || P\left(X|Z\right) P\left(X|Z\right)\right)\right].$$

Fact 4

If Z is jointly independent of X and Y, then $I(X;Y|Z) = I(X;Y)$.

Basic Measures and Relations in Information Theory

- The mutual information between a random variable X and a collection of random variables $(Z_1, ..., Z_T)$ can be expressed elegantly using the following "chain rule."

Fact 5

(Chain Rule of Mutual Information)

$$I(X; (Z_1, ...Z_T)) = I\left(X; Z_1\right) + I\left(X; Z_2 \middle| Z_1\right) + ... + I\left(X; Z_T \middle| Z_1, ..., Z_T\right).$$

Fact 6

(KL divergence form of mutual information)

$$\begin{aligned} I\left(X; Y\right) &= \mathbb{E}_X\left[D\left(P(Y|X) \,||\, P(Y)\right)\right] \\ &= \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) D\left(P(Y|X = x) \,||\, P(Y)\right) \end{aligned}$$

Notation Under Posterior Distributions

- Let
$$\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathcal{F}_t) = \mathbb{P}(\cdot|\, A_1, Y_{1,A_1}, ..., A_{t-1}, Y_{t-1,A_{t-1}})$$
and $\mathbb{E}_t\left[\cdot\right] = \mathbb{E}[\cdot|\mathcal{F}_t]$.
- Define
$$\begin{aligned} H_t(X) &= -\sum_{x\in\mathcal{X}} \mathbb{P}_t(X=x)\log\mathbb{P}_t(X=x) \\ H_t(X|Y) &= \mathbb{E}_t\left[-\sum_{x\in\mathcal{X}} \mathbb{P}_t(X=x|Y)\log\mathbb{P}_t(X=x|Y)\right] \\ I_t(X;Y) &= H_t(X) - H_t(X|Y). \end{aligned}$$
- By taking their expectation, we recover the standard definition of conditional entropy and conditional mutual information:
$$\begin{aligned} \mathbb{E}[H_t(X)] &= H(X|A_1, Y_{1,A_1}, ..., A_{t-1}, Y_{t-1,A_{t-1}}) \\ \mathbb{E}[I_t(X;Y)] &= I\left(X;Y|A_1, Y_{1,A_1}, ..., A_{t-1}, Y_{t-1,A_{t-1}}\right). \end{aligned}$$

Thompson Sampling

- The Thompson sampling algorithm simply samples actions according to the posterior probability they are optimal.
- Actions are chosen randomly at time t according to the sampling distribution $\pi_t^{\mathrm{TS}} = \mathbb{P}(A^* = \cdot|\mathcal{F}_t)$.
- Consider the case where $\mathcal{P} = \{p_\theta\}_{\theta\in\Theta}$ is some parametric family of distributions, and p^* corresponds to a random index $\theta^* \in \Theta$ in the sense that $p^* = p_{\theta^*}$ almost surely.
- Practical implementations of TS use two simple steps:
  - An index $\hat{\theta}_t \sim \mathbb{P}\left(\theta^* \in \cdot|\mathcal{F}_t\right)$ is sampled from the posterior distribution of the true index θ^*.
  - Selects the action $A_t \in \underset{a\in\mathcal{A}}{\arg\max}\, \mathbb{E}\left[R(Y_{t,a})|\theta^* = \hat{\theta}_t\right]$ that would be optimal if the sampled parameter were actually the true parameter.

Example of TS: Beta-Bernouli Bandit

- Action $a \in \mathcal{A}$ is yields either a a success ($Y_a = 1$) or a failure ($Y_a = 0$), and the outcomes are rewards, i.e., $R(y) = y$.
- Suppose action a produces a success with probability θ_a^*, therefore for each $a \in \mathcal{A}$, $\underset{y \sim p_a^*}{\mathbb{E}}[R(y)] = \theta_a^*$ and $A^* \in \underset{a \in \mathcal{A}}{\arg\max}\, \theta_a^*$.
- Since beta distribution is the *conjugate prior* of Bernouli distribution, we take independent priors over each θ_a^* to be beta-distributed with $\alpha = (\alpha_1, ..., \alpha_{|\mathcal{A}|})$ and $\beta = (\beta_1, ..., \beta_{|\mathcal{A}|})$.
- For each action a, the prior probability density function of θ_a^* is

$$p(\theta_a^*) = \frac{\Gamma(\alpha_a + \beta_a)}{\Gamma(\alpha_a)\Gamma(\beta_a)} (\theta_a^*)^{\alpha_a - 1} (1 - \theta_a^*)^{\beta_a - 1},$$

Example of TS: Beta-Bernouli Bandit

- Due to conjugacy properties, each action's posterior distribution is also beta with parameters that can be updated according to a simple rule:

$$(\alpha_a, \beta_a) \leftarrow \begin{cases} (\alpha_a, \beta_a) & \text{if } A_t \neq a \\ (\alpha_a, \beta_a) + (R_t, 1 - R_t) & \text{if } A_t = a. \end{cases}$$

Algorithm 1 Beta-Bernouli Thompson Sampling

1: **Sample Model**:
$\hat{\theta}_t \sim \text{Beta}(\alpha_t, \beta_t)$
2: **Select Action**:
$A_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_{t,a}$
Apply A_t and observe R_t
3: **Update Statistics**:
$(\alpha_{A_t}, \beta_{A_t}) \leftarrow (\alpha_{A_t}, \beta_{A_t}) + (R_t, 1 - R_t)$
4: **Increment t and Goto Step 1**

The Information Ratio

- The expected information gain is defined as the expected reduction in the entropy of the posterior distribution of A^*, i.e., $I_t\left(A^*; (A_t, Y_{t,A_t})\right)$
- We relate the expected regret of Thompson sampling to its expected information gain by *information ratio*,

$$\Gamma_t := \frac{\mathbb{E}_t\left[R(Y_{t,A^*}) - R(Y_{t,A_t})\right]^2}{I_t\left(A^*; (A_t, Y_{t,A_t})\right)}$$

- The information ratio provides a natural measure of each problem's information structure, i.e., the relations between actions and rewards.
- The expected regret is bounded in terms of the inforation ratio and information gain.

A General Regret Bound

- We provide a general upper bound on the expected regret of Thompson sampling that depends on the time horizon T, $H(A^*)$, and any worst–case upper bound on the information ratio Γ_t.

Proposition 1

For any $T \in \mathbb{N}$, if $\Gamma_t \leq \overline{\Gamma}$ almost surely for each $t \in \{1, .., T\}$,

$$\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] \leq \sqrt{\overline{\Gamma} H(A^*) T}.$$

- We will provide bounds on Γ_t for some classes of online optimization problems.

A General Regret Bound

Proof.

Recall that $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$ and we use I_t to denote mutual information evaluated under the base measure $\mathbb{P}_t$. Then,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] &\overset{(a)}{=} \mathbb{E}\sum_{t=1}^{T} \mathbb{E}_t\left[R(Y_{t,A^*}) - R(Y_{t,A_t})\right] \\
&= \mathbb{E}\sum_{t=1}^{T} \sqrt{\Gamma_t I_t\left(A^*; (A_t, Y_{t,A_t})\right)} \\
&\leq \sqrt{\overline{\Gamma}} \left(\mathbb{E}\sum_{t=1}^{T} \sqrt{I_t\left(A^*; (A_t, Y_{t,A_t})\right)}\right) \\
&\overset{(b)}{\leq} \sqrt{\overline{\Gamma} T \mathbb{E}\sum_{t=1}^{T} I_t\left(A^*; (A_t, Y_{t,A_t})\right)},
\end{aligned}$$

A General Regret Bound

Proof.

For the remainder of this proof, let $Z_t = (A_t, Y_{t,A_t})$. Then,

$$\mathbb{E}\left[I_t\left(A^*; Z_t\right)\right] = I\left(A^*; Z_t | Z_1, ..., Z_{t-1}\right),$$

and therefore

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{T} I_t\left(A^*; Z_t\right) &= \sum_{t=1}^{T} I\left(A^*; Z_t | Z_1, ..., Z_{t-1}\right) \overset{(c)}{=} I\left(A^*; Z_1, ...Z_T\right) \\
&= H(A^*) - H(A^* | Z_1, ...Z_T) \\
&\overset{(d)}{\leq} H(A^*).
\end{aligned}$$

$\square$

Bounding the Information Ratio

- By Proposition 1, we can get explicit regret bounds by establishing bounds on the information ratio.
- The information ratio captures the influence of sampling some actions on making inferences about *different* actions, which depends on the class of problems.
  - Worst case: bounded by the number of actions; actions could provide no information about others.
  - Best case: bounded by a numerical constant; full information, sampling one action perfectly reveals the rewards for any other action.
  - Linear bandit case: bounded by the dimension of action space; sampling actions could provide some information about others.

An Alternative Representation of the Information Ratio

- To simplify notation, from now on we will omit the subscript t from $\mathbb{E}_t, \mathbb{P}_t, P_t, A_t, Y_t, H_t$, and I_t.
- The following proposition expresses the information ratio of Thompson sampling in a form that facilitates further analysis.

Proposition 2

$$\begin{aligned} I\left(A^*;(A,Y_A)\right) &= \sum_{a\in\mathcal{A}} \mathbb{P}(A=a) I(A^*;Y_a) \\ &= \sum_{a,a^*\in\mathcal{A}} \mathbb{P}(A^*=a)\mathbb{P}(A^*=a^*)\left[D\left(P(Y_a|A^*=a^*)\,||\,P(Y_a)\right)\right]. \end{aligned}$$

and

$$\mathbb{E}\left[R(Y_{A^*})-R(Y_A)\right] = \sum_{a\in\mathcal{A}} \mathbb{P}(A^*=a)\left(\mathbb{E}\left[R(Y_a)|A^*=a\right]-\mathbb{E}[R(Y_a)]\right.$$

An Alternative Representation of the Information Ratio

- The numerator captures how much knowing that the *selected action is optimal* influences the expected reward observed.
- The denominator measures how much, on average, knowing *which action is optimal* changes the observations at the selected action.

Proof.

The action A is selected based on past observations and independent random noise. Therefore, conditioned on the history, A is jointly independent of A^* and the outcome vector $Y \equiv (Y_a)_{a \in \mathcal{A}}$.

$$
\begin{aligned}
& \mathbb{E}\left[R(Y_{A^*}) - R(Y_A)\right] \\
= & \sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a)\mathbb{E}\left[R(Y_a) | A^* = a\right] - \sum_{a \in \mathcal{A}} \mathbb{P}(A = a)\mathbb{E}[R(Y_a) | A = a] \\
= & \sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a)\left(\mathbb{E}\left[R(Y_a) | A^* = a\right] - \mathbb{E}[R(Y_a)]\right),
\end{aligned}
$$

An Alternative Representation of the Information Ratio

Proof.

$$
\begin{aligned}
& I(A^*;(A,Y_A)) \\
\overset{(a)}{=} \quad & I(A^*;A) + I(A^*;Y_A|A) \\
\overset{(b)}{=} \quad & I(A^*;Y_A|A) \\
= \quad & \sum_{a\in\mathcal{A}} \mathbb{P}(A=a) I(A^*;Y_A|A=a) \\
\overset{(c)}{=} \quad & \sum_{a\in\mathcal{A}} \mathbb{P}(A=a) I(A^*;Y_a) \\
\overset{(d)}{=} \quad & \sum_{a\in\mathcal{A}} \mathbb{P}(A=a) \left(\sum_{a^*\in\mathcal{A}} \mathbb{P}(A^*=a^*) D\left(P(Y_a|A^*=a^*) \,||\, P(Y_a)\right) \right) \\
= \quad & \sum_{a,a^*\in\mathcal{A}} \mathbb{P}(A^*=a)\mathbb{P}(A^*=a^*) \left[D\left(P(Y_a|A^*=a^*) \,||\, P(Y_a)\right)\right].
\end{aligned}
$$

□

Preliminaries

- Here we state two basic facts that are used in bounding the information ratio.
- The first fact lower bounds the Kullback–Leibler divergence between two bounded random variables in terms of the difference between their means.

Fact 7

For any distributions P and Q such that that P is absolutely continuous with respect to Q, any random variable $X : \Omega \to \mathcal{X}$ and any $g : \mathcal{X} \to \mathbb{R}$ such that $\sup g - \inf g \le 1$,

$$\mathbb{E}_P\left[g(X)\right] - \mathbb{E}_Q\left[g(X)\right] \le \sqrt{\frac{1}{2} D\left(P||Q\right)},$$

where $\mathbb{E}_P$ and $\mathbb{E}_Q$ denote the expectation operators under P and Q.

Preliminaries

- Because of Assumption 1, this fact shows

$$\mathbb{E}\left[R(Y_a)|A^* = a^*\right] - \mathbb{E}\left[R(Y_a)\right] \leq \sqrt{\frac{1}{2} D\left(P(Y_a|A^* = a^*) \mid\mid P(Y_a)\right)}.$$

- For any rank r matrix $M \in \mathbb{R}^{n \times n}$ with singular values $\sigma_1, ..., \sigma_r$, let

$$\|M\|_* := \sum_{i=1}^{r} \sigma_i, \qquad \|M\|_F := \sqrt{\sum_{k=1}^{m} \sum_{j=1}^{n} M_{i,j}^2} = \sqrt{\sum_{i=1}^{r} \sigma_i^2},$$

denote respectively the Nuclear norm and Frobenius norm of M.

Fact 8

For any matrix $M \in \mathbb{R}^{k \times k}$,

$$\mathrm{Trace}\left(M\right) \leq \sqrt{\mathrm{Rank}(M)} \|M\|_{\mathrm{F}}.$$

Worst Case Bound

- The next proposition provides a bound on the information ratio that holds whenever rewards are bounded, and this scaling cannot be improved in general.

Proposition 3

For any $t \in \mathbb{N}$, $\Gamma_t \leq |\mathcal{A}|/2$ *almost surely.*

- Combining Proposition 3 with Proposition 1 shows that $\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] \leq \sqrt{\frac{1}{2}|\mathcal{A}|H(A^*)T}$.

Worst Case Bound

Proof.

$$
\begin{aligned}
& \mathbb{E}\left[R(Y_{A^*}) - R(Y_A)\right]^2 \\
\overset{(a)}{=} \; & \left(\sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a) \left(\mathbb{E}\left[R(Y_a) | A^* = a\right] - \mathbb{E}[R(Y_a)] \right) \right)^2 \\
\overset{(b)}{\leq} \; & |\mathcal{A}| \sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a)^2 \left(\mathbb{E}\left[R(Y_a) | A^* = a\right] - \mathbb{E}[R(Y_a)] \right)^2 \\
\leq \; & |\mathcal{A}| \sum_{a, a^* \in \mathcal{A}} \mathbb{P}(A^* = a) \mathbb{P}(A^* = a^*) \left(\mathbb{E}\left[R(Y_a) | A^* = a^*\right] - \mathbb{E}[R(Y_a)] \right)^2 \\
\overset{(c)}{\leq} \; & \frac{|\mathcal{A}|}{2} \sum_{a, a^* \in \mathcal{A}} \mathbb{P}(A^* = a) \mathbb{P}(A^* = a^*) D\left(P(Y_a | A^* = a^*) \;||\; P(Y_a) \right) \\
\overset{(d)}{=} \; & \frac{|\mathcal{A}| I(A^*; (A, Y))}{2}.
\end{aligned}
$$

□

Full Information

- Problems with full information is an extreme case of our formulation. The outcome $Y_{t,a}$ is perfectly revealed by observing $Y_{t,\tilde{a}}$ for any $\tilde{a} \neq a$, in other words, what is learned does not depend on the selected action.

Proposition 4

Suppose for each $t \in \mathbb{N}$ there is a random variable $Z_t : \Omega \to \mathcal{Z}$ such that for each $a \in \mathcal{A}$, $Y_{t,a} = (a, Z_t)$. Then for all $t \in \mathbb{N}$, $\Gamma_t \leq 1/2$ almost surely.

- Combining this result with Proposition 1 shows $\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] \leq \sqrt{\frac{1}{2}H(A^*)T}$.

Full Information

Proof.

$$
\begin{aligned}
& \mathbb{E}\left[R(Y_{A^*}) - R(Y_A)\right] \\
\overset{(a)}{=} \quad & \sum_{a\in\mathcal{A}} \mathbb{P}(A^* = a)\left(\mathbb{E}\left[R(Y_a)|A^* = a\right] - \mathbb{E}[R(Y_a)]\right) \\
\overset{(b)}{\leq} \quad & \sum_{a\in\mathcal{A}} \mathbb{P}(A^* = a)\sqrt{\frac{1}{2}D\left(P(Y_a|A^* = a)\,||\,P(Y_a)\right)} \\
\overset{(c)}{\leq} \quad & \sqrt{\frac{1}{2}\sum_{a\in\mathcal{A}} \mathbb{P}(A^* = a)D\left(P(Y_a|A^* = a)\,||\,P(Y_a)\right)} \\
\overset{(d)}{=} \quad & \sqrt{\frac{1}{2}\sum_{a,a^*\in\mathcal{A}} \mathbb{P}(A^* = a)\mathbb{P}(A^* = a^*)D\left(P(Y_a|A^* = a^*)\,||\,P(Y_a)\right)} \\
\overset{(e)}{=} \quad & \sqrt{\frac{I(A^*;(A,Y))}{2}}.
\end{aligned}
$$

□

Linear Optimization Under Bandit Feedback

- In this setting, each action is associated with a finite dimensional feature vector, and the mean reward generated by an action is the inner product between its known feature vector and some unknown parameter vector.

Proposition 5

If $\mathcal{A} \subset \mathbb{R}^d$ and for each $p \in \mathcal{P}$ there exists $\theta_p \in \mathbb{R}^d$ such that for all $a \in \mathcal{A}$

$$\mathop{\mathbb{E}}_{y \sim p_a} [R(y)] = a^T \theta_p,$$

then for all $t \in \mathbb{N}$, $\Gamma_t \leq d/2$ almost surely.

- This result shows that $\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] \leq \sqrt{\frac{1}{2} H(A^*) dT} \leq \sqrt{\frac{1}{2} \log(|\mathcal{A}|) dT}$ for linear bandit problems.

Linear Optimization Under Bandit Feedback

Proof.

Write $\mathcal{A} = \{a_1, ..., a_K\}$ and let $\alpha_i = \mathbb{P}(A^* = a_i)$. Define $M \in \mathbb{R}^{K \times K}$ by

$$M_{i,j} = \sqrt{\alpha_i \alpha_j} \left(\mathbb{E}[R(Y_{a_i}) | A^* = a_j] - \mathbb{E}[R(Y_{a_i})] \right),$$

for all $i, j \in \{1, .., K\}$. Then, by Proposition 2,

$$\mathbb{E}\left[R(Y_{A^*}) - R(Y_A)\right] = \sum_{i=1}^{K} \alpha_i \left(\mathbb{E}[R(Y_{a_i}) | A^* = a_i] - \mathbb{E}[R(Y_{a_i})] \right) = \mathrm{Trace(M)}.$$

Similarly, by Proposition 2,

$$\begin{aligned} I(A^*; (A, Y_A)) &= \sum_{i,j} \alpha_i \alpha_j D\left(P(Y_{a_i} | A^* = a_j) \,||\, P(Y_{a_i}) \right) \\ &\overset{(a)}{\geq} 2 \sum_{i,j} \alpha_i \alpha_j \left(\mathbb{E}[R(Y_{a_i}) | A^* = a_j] - \mathbb{E}[R(Y_{a_i})] \right)^2 \\ &= 2 \|M\|_{\mathrm{F}}^2, \end{aligned}$$

Proof.

This shows, by Fact 8, that

$$\frac{\mathbb{E}\left[R(Y_{A^*}) - R(Y_A)\right]^2}{I(A^*; (A, Y_A))} \leq \frac{\mathrm{Trace(M)}^2}{2\|M\|_{\mathrm{F}}^2} \leq \frac{\mathrm{Rank}(M)}{2}.$$

We now show $\mathrm{Rank}(M) \leq d$. Define

$$\mu = \mathbb{E}\left[\theta_{p^*}\right] \qquad \mu^j = \mathbb{E}\left[\theta_{p^*} | A^* = a_j\right].$$

We have $M_{i,j} = \sqrt{\alpha_i \alpha_j}((\mu^j - \mu)^T a_i)$ and therefore

$$M = \begin{bmatrix} \sqrt{\alpha_1} a_1^T \\ \vdots \\ \vdots \\ \sqrt{\alpha_K} a_K^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha_1}(\mu^1 - \mu) & \cdots & \cdots & \sqrt{\alpha_K}(\mu^K - \mu) \end{bmatrix}.$$

□

References

- Russo, Daniel, and Benjamin Van Roy. "An information-theoretic analysis of thompson sampling." The Journal of Machine Learning Research 17.1 (2016): 2442-2471.
- Russo, Daniel J., et al. "A tutorial on thompson sampling." Foundations and Trends in Machine Learning 11.1 (2018): 1-96.