Homework 1[1]

CSE 599i: Interactive Learning
Instructor: Kevin Jamieson
Due 11:59 PM on January 24, 2020

Probability

Concentration inequalities are at the heart of most arguments in statistical learning theory and bandits. Refer to [1] for more details.

1.1 (Markov's Inequality) Let X be a positive random variable. Prove that $\mathbb{P}(X > \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}$.

1.2 (Jensen's Inequalty) Let X be a random vector in $\mathbb{R}^d$ and let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex. Then $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$. Show this inequality for the special case when X has discrete support. That is, for $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$, and $(x_1, \ldots, x_n) \in \mathbb{R}^n$ show that $\phi(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i \phi(x_i)$.

1.3 (Hoeffding's Lemma) Let X be a random variable with $\mathbb{E}[X] = 0$ and $X \in [a, b]$ almost surely. Show that for any $\lambda \geq 0$, $\log(\mathbb{E}[e^{\lambda X}]) \leq \frac{\lambda^2(b-a)^2}{8}$. Hint[2]

1.4 (Sub-exponential concentration) For $i = 1, \ldots, n$ let X_i be an independent, positive random variable that satisfies $\mathbb{P}(X_i > t) \leq e^{-t/a_i}$ for some $a \in \mathbb{R}^n_+$. Show that there exists a universal constant $c > 0$ such that $\mathbb{P}(\sum_{i=1}^n (X_i - a_i) \geq t) \leq \exp(-c \min\{\frac{t^2}{\|a\|_2^2}, \frac{t}{\|a\|_\infty}\})$.

The Upper Confidence Bound Algorithm.

Consider the following algorithm for the multi-armed bandit problem.

Algorithm 1: UCB

Input: Time horizon T, 1-subGaussian arm distributions $P_1, \cdots, P_n$ with unknown means $\mu_1, \cdots, \mu_n$

Initialize: At any time let $T_i(t)$ denote the number of times i has been pulled at time t and let $T_i = T_i(T)$. Pull each arm once.

for: $t = n+1, \cdots, T$

Pull arm $I_t = \arg\max_{i=1,\cdots,n} \widehat{\mu}_{i,T_{i(t-1)}} + \sqrt{\frac{2\log(2nT^2)}{T_i(t-1)}}$ and observe draw from P_i

Let $\widehat{\mu}_{i,T_i(t)}$ be the empirical mean of the first $T_i(t)$ pulls.

In the following exercises, we will compute the regret of the UCB algorithm and show it matches the regret bound from lecture. Without loss of generality, assume that the best arm is μ_1. For any $i \in [n]$, define the *sub-optimality gap* $\Delta_i = \mu_1 - \mu_i$. Define the regret at time T as $R_T = \mathbb{E}[\sum_{t=1}^T \mu^* - \mu_{I_t}] = \sum_{i=1}^n \Delta_i \mathbb{E}[T_i]$.

2.1 Consider the event

$$\mathcal{E} = \bigcap_{i\in[n]} \bigcap_{s\leq T} \left\{ |\widehat{\mu}_{i,s} - \mu_i| \leq \sqrt{\frac{2\log(2nT^2)}{s}} \right\}.$$

Show that $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{1}{T}$.

2.2 Conditioned on event $\mathcal{E}$, show that $T_i < \frac{8\log(2nT^2)}{\Delta_i^2}$ for $i \neq 1$.

2.3 Show that $\mathbb{E}[T_i] \leq \frac{8\log(2nT^2)}{\Delta_i^2} + 1$. When $n \leq T$, conclude by showing that $R_T \leq \sum_{i=1}^n \left(\frac{24\log(2T)}{\Delta_i} + \Delta_i\right)$.

---

[1] Last updated to correct for typos January 21, 2020 at 8:33 AM

[2] For any $X \in [a, b]$ we can write $X = (1-p)a + pb$ for $p = \frac{X-a}{b-a} \in [0, 1]$. Apply Jensen's inequality. This can be tricky feel free to get as far as you can.

Thompson Sampling.

We consider the following Bayesian setting. Consider n arms and let p_0 be an n-dimensional prior distribution over $[-1,1]^n$ such that $\theta^* \sim p_0$ is drawn before the start of the game (e.g., p_0 is uniform over $[-1,1]^n$). At any time t, when we pull arm $i \in [n]$ we observe a random variable $X_{i,t} \in [-1,1]$ where $\mathbb{E}[X_{i,t}] = \theta_i^*$.

Algorithm 1: Thompson Sampling

> **Input:** Time horizon T, arm distributions $\nu_1, \cdots, \nu_n$
> Assume the prior p_0 is known and that $\theta^* \sim p_0$. Let $p_t(\cdot|I_1, X_{I_1,1}, \cdots, I_t, X_{I_t,t})$ be the posterior distribution on θ^* at time t.
> **for:** $t = 1, \cdots, T$
> Sample $\theta^{(t)} \sim p_{t-1}$
> Pull arm $I_t = \arg\max_{i \le n} \theta_i^{(t)}$ to observe $X_{I_t,t}$
> Update $T_{I_t}(t+1) \leftarrow T_{I_t}(t) + 1$
> Compute exact posterior update p_t

Denote the σ-algebra generated by the observations at time t by $\mathcal{F}_t = \sigma(I_1, X_{I_1,1}, \cdots, I_t, X_{I_t,t})$ (if you are unfamiliar with σ-algebras, don't worry too much - conditioning on the σ-algebra just means conditioning on the choices of arms and the rewards observed). The *Bayesian Regret* of an algorithm is

$$\begin{aligned} BR_T &= \mathbb{E}_{\theta^* \sim p_0}\left[\sum_{t=1}^{T} \max_{i=1,\dots,n} \theta_i^* - \theta_{I_t}^*\right] \\ &= \mathbb{E}_{\theta^* \sim p_0}\left[\mathbb{E}\left[\sum_{t=1}^{T} \max_{i=1,\dots,n} \theta_i^* - \theta_{I_t}^* \middle| \theta_*\right]\right] \end{aligned}$$

Assume that expectations, if not explicitly specified, are with respect to all randomness including $\theta_* \sim p_0$, $I_1, \dots, I_T$, and observations.

3.1 On a given run of the algorithm, let $\widehat{\theta}_{i,s}$ denote the empirical mean of the first s pulls from arm i, note that $\mathbb{E}[\widehat{\theta}_{i,s}] = \theta_i^*$. Let the good event be

$$\mathcal{E} = \bigcap_{i \in [n]} \bigcap_{t \le T} \left\{ |\widehat{\theta}_{i,t} - \theta_i^*| \le \sqrt{\frac{2\log(2/\delta)}{t}} \right\}.$$

Show that $\mathbb{P}(\mathcal{E}^c) \le nT\delta$.

3.2 (Key idea.) Argue that $\mathbb{P}(i^* = \cdot|\mathcal{F}_{t-1}) = \mathbb{P}(I_t = \cdot|\mathcal{F}_{t-1})$.

3.3 Define $U_t(i) = \min\{1, \widehat{\theta}_{i,T_i(t)} + \sqrt{\frac{2\log(2/\delta)}{T_i(t)}}\}$. If $i^* = \arg\max_i \theta_i^*$, show that $\mathbb{E}_{\theta^* \sim p_0}[\mathbb{E}_{I_t}[\theta_{i^*}^* - \theta_{I_t}^*|\mathcal{F}_{t-1}]] = \mathbb{E}_{\theta^* \sim p_0}[\theta_{i^*}^* - U_t(i^*)] + \mathbb{E}_{\theta^* \sim p_0}[\mathbb{E}_{I_t}[U_t(I_t) - \theta_{I_t}^*|\mathcal{F}_{t-1}]]$. Conclude that $BR_T = \mathbb{E}_{\theta^* \sim p_0}[\sum_{t=1}^T \theta_{i^*}^* - U_t(i^*) + \sum_{t=1}^T \mathbb{E}_{I_t}[U_t(I_t) - \theta_{I_t}^*|\mathcal{F}_{t-1}]]$. Hint[3].

3.4 Show that $BR_T \le 4\delta nT^2 + \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\left(\sum_{t=1}^T U_t(I_t) - \theta_{I_t}^*\right)\middle|\theta^*\right]\right] \le O(\delta nT^2 + \sqrt{Tn\log(1/\delta)})$. Hint[4]

3.5 Make an appropriate choice of δ and state a final regret bound.

In general, giving frequentist bounds on the regret is significantly more difficult. We refer the interested reader to [2] and the tutorial [3] for more details. This exercise is motivated by [4]

[3]Tower rule of expectation.
[4]Apply Jensen's to $\sum_{i=1}^n \sqrt{T_i}$.

Algorithm 1: Explore-then-Commit

Input: Time horizon T, $m \in \mathbb{N}$, 1-sub-Gaussian arm distributions $P_1, \cdots, P_n$ with unknown means $\mu_1, \cdots, \mu_n$
for: $t = 1, \cdots, T$
  If $t \leq mn$, choose $I_t = (t \mod n) + 1$
  Else, $I_t = \arg\max_i \widehat{\mu}_{i,m}$

Empirical Experiments
Implement UCB, Thompson Sampling (TS), and Explore-then-Commit (ETC). Let $P_i = \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, n$. For Thompson sampling, define the prior for the ith arm as $\mathcal{N}(0, 1)$.

4.1 Let $n = 10$ and $\mu_1 = 0.1$ and $\mu_i = 0$ for $i > 1$. On a single plot, for an appropriately large T to see expected effects, plot the regret for the UCB, TS, and ETC for several values of m.

4.2 Let $n = 40$ and $\mu_1 = 1$ and $\mu_i = 1 - 1/\sqrt{i-1}$ for $i > 1$. On a single plot, for an appropriately large T to see expected effects, plot the regret for the UCB, TS, and ETC for several values of m.

Lower Bounds on Hypothesis Testing
Consider n samples $X_1, \cdots, X_n \sim P$ where $P \in \{P_0, P_1\}$ (assume for simplicity that these are probability distributions on $\mathbb{R}$). A *hypothesis test* for $H_0 : P = P_0, H_1 : P = P_1$ is a function $\phi(x_1, \cdots, x_n) : \mathbb{R}^n \to \{0, 1\}$ that takes the data as input and returns the null or the alternative. Assume that the $dP_i = p_i(x)dx$ so that the probability density function exists (think: $p_i(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_i)^2/2}$). If $x \in \mathbb{R}^n$ is the vector of n observations, define $p_i(x) := \prod_{j=1}^n p_i(x_j)$. In this problem, we will lower bound the number of samples needed by *any* hypothesis test on a fixed number of samples. Convince yourself, at least intuitively, that any best-arm identification algorithm for two arms will take at least as many samples as this hypothesis test takes.

5.1 Show $\inf_\phi \max\{\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0)\} \geq \frac{1}{2} \int_{\mathbb{R}^n} \min(p_0(x), p_1(x))dx$. Hint[5].

5.2 Let's continue on. Show $\frac{1}{2} \int_{x \in \mathbb{R}^n} \min(p_0(x), p_1(x))dx \geq \frac{1}{4} \left(\int_{x \in \mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx \right)^2$. Hint[6][7].

5.3 One more step. Show $\left(\int_{x \in \mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx \right)^2 \geq \exp\left(- \int_{x \in \mathbb{R}^n} \log\left(\frac{p_1(x)}{p_0(x)}\right) p_1(x)dx \right)$. Hint[8].

5.4 The final quantity is known as the KL-Divergence between distributions. Now assume that $P_0 = N(\mu_0 \mathbf{1}_n, I_n)$ and $P_1 = N(\mu_1 \mathbf{1}_n, I_n)$ where I_n is the $n \times n$ identity matrix and $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector. Show (or look up) $KL(P_0||P_1)$.

5.5 Conclude that to acheive a test that accurately determines whether the sample of size n came from P_0 or P_1 with a probability of error less than δ, we necessarily have $n \geq 2\Delta^{-2} \log(1/4\delta)$ where $\Delta = \mu_1 - \mu_0$.

Remark: The art of lower bounds is well established and extensive in statistics. See [5] for more details in the hypothesis testing setting. In the bandit setting, see [6].

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[5] Bound the max below by the average
[6] Note that for $a, b > 0$ we have $ab = \min\{a, b\} \max\{a, b\}$.
[7] Apply Cauchy Schwartz.
[8] Jensen's inequality.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[3] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[4] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[5] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009.

[6] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.