Mathematics of Machine Learning: Bandits **Summer 2019**

Problem Set

Lecturer: Kevin Jamieson *TA: Lalit Jain*

Disclaimer: *These exercises have not been subjected to the usual scrutiny reserved for formal publications. They may contain mistakes.*

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: Pull each arm once. At any time let $T_i(t)$ denote the
number of times $i$ has been pulled at time $t$ and let $T_i = T_i(T)$.
for: $t = 1, \cdots, T$
    Choose $I_t = \max_{i=1,\cdots,k} \widehat{\mu}_{i,t} + \sqrt{\frac{2\log(nT^2)}{T_i(t)}}$
    Observe $X_{I_t,t} \sim P_{I_t}$. $T_{I_t}(t+1) \leftarrow T_{I_t}(t) + 1$, update $\widehat{\mu}_{i,t+1}$
```

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 $i \in [k]$, define the *sub-optimality gap* $\Delta_i = \mu_1 - \mu_i$.
1. (Warm-up). Show that $R_T = \sum_{i=1}^n \Delta_i \mathbb{E}[T_i]$.

2. Consider the event

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

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

3. Conditioned on event $\mathcal{E}$, show that $T_i < \frac{8\alpha\log(nT^2)}{\Delta_i^2}$.
4. Show that $\mathbb{E}[T_i] \leq \frac{8\log(nT^2)}{\Delta_i^2} + 1$. When $n \leq T$, conclude by showing that $R_T \leq \sum_{i=1}^k \left(\frac{8\alpha\log(T)}{\Delta_i} + 2\Delta_i\right)$.

5. Implement UCB, Action Elimination, and your personal algorithm for two arms, $\mu_1 = \Delta$ and $\mu_2 = 0$ and run it for a time horizon of around 2000 samples. What do you notice? What if you have three arms?

6. (Challenge Problem.) The above algorithm relies on knowing the horizon T. Can you remove this dependency providing an algorithm with a similar *anytime* regret guarantee?
Remark: The $\log(T)$ term can in fact be removed. See [1, 2] for more details.

Explore-then-Commit.

```
Algorithm 1: Explore-then-Commit

    Input: Time horizon $T$, $\alpha > 2$, 1-subGaussian 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, $I_t = \arg\max_i \widehat{\mu}_{i,mn}$
```

1. (Warm-up). Show that $\mathbb{E}[T_i] \leq mn + (T - mn)\mathbb{P}(\widehat{\mu}_{i,mn} \geq \max_{j \neq i} \widehat{\mu}_{i,mn})$, where $\widehat{\mu}_{i,mn}$ is the empirical mean of arm i at time mn.

2. Show that the regret of Explore-then-Commit is bounded by

$$R_T \leq m \sum_{i=1}^{n} \Delta_i + (n - mn) \sum_{i=1}^{n} \exp(-m\Delta_i^2/4).$$

3. Assume you have only two arms. Minimize the above expression for m. How much regret would you incur if you know the gap, $\Delta_2 = \mu_1 - \mu_2$?

4. Implement Explore then commit and UCB in the case of two arms (assume that $\mu_1 = 0$ and $\mu_2 = \Delta$. Experiment with different values of m for $T \approx 1000$. What do you observe? How does ETC at the optimal value of m compare to UCB?

5. (Challenge) The choice of m in part 3 depended on knowing the smallest gap. Show that there is a choice of the smallest m independent of the gap so that the regret is $O(T^{2/3})$.

Lower Bounds on Hypothesis Testing

Consider n samples $X_1, \cdots, X_n \sim P$ where $P \in \{P_0, P_1\}$. 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 $dP_i = p_i(x)dx$ In this problem, we will lower bound the number of samples needed by *any* hypothesis test on a fixed number of samples.

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: bound the max below by the average, explicitly compute the optimal ϕ, and show the result).

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

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

4. The final quantity is known as the KL-Divergence between distributions. Now assume that $P_0 = N(\mu_0 I_n, I_n)$ and $P_1 = N(\mu_1 I_n, I_n)$ where I_n is the $n \times n$ identity matrix. Show (or look up) $KL(P_0||P_1)$.

5. Conclude that to acheive a test with a probability of error less than δ, then we necessarily have $n \geq 2\Delta^{-2}\log(1/4\delta)$.
Remark: The art of lower bounds is well established and extensive in statistics. See [?] for more details in the hypothesis testing setting. In the bandit setting, see [3].

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

2. (Hoeffding's Lemma) Let $X \in [a, b]$. Show that for any $\lambda \geq 0$, $\mathbb{E}[e^{\lambda X}] \leq \frac{\lambda^2(b-a)^2}{8}$.

The Kiefer-Wolfowitz Theorem Let $x_1, \cdots, x_n \subset \mathbb{R}^{\ltimes}$. In this exercise, we will prove the celebrated Kiefer-Wolfowitz theorem. Let $A(\lambda) = \sum_x \lambda_x x x^\top$ and $\rho = \min_{\lambda \in \Delta_n} \max_{i \in [n]} \|x_i\|^2_{A(\lambda)^{-1}}$.
1. Show $\min_{\lambda \in \Delta_n} \max_{i \in [n]} \|x_i\|_{A(\lambda)^{-1}} = \min_{\lambda \in \Delta_n} \max_{\rho \in \Delta_n} \|x_i\|_{A(\lambda)^{-1}}$.

2. Next, show that for any $\lambda \in \Delta_n$, $\sum_x \lambda_x \|x\|^2_{A(\lambda)^{-1}} \geq d$.

3. Now, consider the function $f(\lambda) = \log\det(A(\lambda))$. Show that $\nabla_\lambda f(\lambda) = \|x\|^2_{A(\lambda)^{-1}}$. (Hint: use a Frechet derivative. Even if you black-box this part, convince yourself it is true).

4. Show that $f(\lambda)$ is concave in λ .

5. Let $\lambda^* = \arg\max_{\lambda \in \Delta_n} f(\lambda)$. Apply the first order optimality conditions to show that $0 \geq \langle \nabla f(\lambda^*), \lambda - \lambda^* \rangle$ for any $\lambda \in \Delta_n$.

6. Use the result of the previous exercise along with to show part 2 to conclude. You actually got a bonus result. What is it?

Thompson Sampling

We consider the following setting. Assume that we have access to n arm distributions $\nu_1(\theta), \cdots, \nu_n(\theta)$ each supported on $[0, 1]$, where $\theta \in \mathbb{R}$ is a real parameter, and the mean of the i-th distribution is $\mu_i(\theta)$. We also assume access to a prior $p_0(\theta)$.

```
Algorithm 1: Thompson Sampling

    Input: Time horizon $T$, arm distributions $\nu_1, \cdots, \nu_n$
    Let $p_t(|I_1, R_1, \cdots, I_{t-1}, R_{t-1})$ be the posterior distribution on $\theta$ at
    time $t$. for: $t = 1, \cdots, T$
        Sample $\theta_t \sim p_t$
        Choose $I_t = \arg\max_{i \leq n} \mu_t(\theta_t)$
```

Denote the σ-algebra generated by the observations at time t by $\mathcal{F}_t = \sigma(I_1, R_1, \cdots, I_{t-1}, R_{t-1})$ (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

$$BR_T = \mathbb{E}_{\theta \sim p_0}\left[\sum_{t=1}^{T} \mu^* - \mu_{I_t}(\theta)\right]$$

where $i^* = \arg\max_{i \in [n]} \mu_i(\theta)$ (it's a random variable depending on θ).
1. Let the good event be

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

Show that $\mathbb{P}(\mathcal{E}) \leq nT\delta$.

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

3. Define $U_t(I_t) = \sqrt{\frac{2\log(2/\delta)}{T_i(t)}}$. Using the above, show that $\mathbb{E}[\mu_{i^*} - \mu_{I_t}|\mathcal{F}_{t-1}] = \mathbb{E}[\mu_{i^*} - U_t(i^*)|\mathcal{F}_{t-1}] + \mathbb{E}[U_t(I_t) - \mu_{I_t}|\mathcal{F}_{t-1}]$.

4. Conclude that $BR_T = \mathbb{E}[\sum_{t=1}^T \mu_{i^*} - \mu_{I_t} + \sum_{t=1}^T \mu_{I_t} - \mu_{i^*}]$ Hint: Tower rule of expectation.

5. On the event $\mathcal{E}$, show that $\mathbf{1}(\mathcal{E})BR_T \leq \sum U_t(I_t) - \mu_{I_t} \leq O(\sqrt{Tk\log(1/\delta)})$.

6. 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 [4] and the tutorial [5] for more details.

The Doubling Trick

Assume we have access to a fixed-horizon algorithm $\mathcal{A}$ that depends on knowing the time horizon T and the regret $\mathcal{A}$ incurs is given by $f(T)$. In this algorithm we develop a strategy to turn $\mathcal{A}$ into an anytime algorithm.

1. Now assume that $f(T) \leq \sqrt{T}$. Come up with a strategy so that at anytime t, the cumulative regret is bounded $C\sqrt{t}$ for some fixed constant C.

2. Now assume that $f(T) \leq \log T$. Come up with a strategy so that at anytime t, the cumulative regret is bounded $C\log t$ for some fixed constant C.

Notes: The presentation of Explore-then-Commit and Thompson sampling are motivated by those in [6].

References

[1] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. 2009.

[2] Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

[3] 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.

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

[5] 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.

[6] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.