Fall 2022 *Instructor: Jacob Abernethy, Bhuvesh Kumar, Guanghui Wang*

ECE 8803: Online Decision Making in Machine Learning

Homework 3

Problem 1 (Boosting via Game Playing) 20 points There is a popular framework for constructing a hypothesis via an *ensemble*, i.e. mixing together a collection of predictors, that is known as *boosting*. What boosting does is iteratively improve the ensemble by incorporating new predictors (aka "weak learners"). The problem, as it turns out, can be formulated as solving a zero-sum game, and the iterative process can be viewed as an interaction between a no-regret learning algorithm and an "oracle" for choosing predictors. Before we begin, let's lay out some terminology.

Let $\mathcal{X}$ be a data space (e.g. $\mathbb{R}^d$), and assume we have labels $\mathcal{Y} := \{-1, +1\}$. We are given access to n examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{X}$. We also have a set of *weak learners*, $\mathcal{H} = \{h : \mathcal{X} \to \{0, 1\}\}$, which make predictions. For the remainder of this problem, let's assume $\mathcal{H}$ is finite, $|\mathcal{H}| = m$. We typically refer to these predictors as "weak" because, in practice, they will have very low complexity. A common choice of weak learner class is the set of *decision stumps*, which simply predict +1/-1 based on whether a single feature is above/below a given threshold. We don't expect a single weak learner to perform well, but perhaps we can combine them into a "stronger" predictor? Here's a classic assumption we make about the weak learning class $\mathcal{H}$.

Weak learning assumption ($\gamma > 0$)*:* Assume, for every dist. $\mathbf{p} \in \Delta_n$, there exists $h \in \mathcal{H}$

$$\Pr_{i \sim \mathbf{p}}[h(x_i) \neq y_i] \leq \frac{1}{2} - \frac{\gamma}{2}$$

This assumption is "weak" in the sense that it only guarantees that some $h \in \mathcal{H}$ is going to perform slightly better than a 50-50 guess. Of course, what we would prefer is a strong predictor, which is 100% accurate. When might we achieve this with a mixture of weak learners? Let's define this as follows.

Strong learning assumption ($\gamma > 0$)*:* Assume there is some distribution $\mathbf{q} \in \Delta_m$ for which the $\mathbf{q}$-weighted mixture of predictors $\mathcal{H}$ always has a γ-margin for every example. That is, $\forall i \in [n]$

$$\text{The weighted forecast } \sum_{j=1}^{m} q_j h_j(x_i) \text{ is } \begin{cases} \geq \gamma & \text{when } y_i = +1 \\ \leq -\gamma & \text{when } y_i = -1 \end{cases}$$

Let's now try to reformulate these ideas in the form of a min-max problem.

(a) Let's try to rewire the weak learning assumption in matrix form. Construct an $n \times m$ matrix, with values in $\{-1, +1\}$, so that the weak learning assumption is equivalent to the statement

$$\min_{\mathbf{p} \in \Delta_n} \max_{j \in [m]} \mathbf{p}^\top M \mathbf{e}_j \geq \gamma,$$

where $\mathbf{e}_j$ is the jth basis vector (all zeros, except for a 1 in the jth coordinate).

(b) Argue why the above minmax statement is equivalent to the following

$$\min_{\mathbf{p}\in\Delta_n} \max_{\mathbf{q}\in\Delta_m} \mathbf{p}^\top M\mathbf{q} \geq \gamma.$$

(c) Similarly, argue why the strong learning assumption can be reformulated as

$$\max_{\mathbf{q}\in\Delta_m} \min_{\mathbf{p}\in\Delta_n} \mathbf{p}^\top M\mathbf{q} \geq \gamma.$$

(d) Using von Neumann's minimax theorem argue that, for a fixed $\gamma > 0$, the weak learning assumption implies the strong learning assumption.

(e) (BONUS - 15 points) Formulate an algorithm, using the exponential weights method, that finds a distribution over the weak learners $\mathbf{q}^* \in \Delta_m$ which satisfies the strong learning assumption for some $\gamma > 0$. Your algorithm should require roughly $T = O\left(\frac{\log n}{\gamma^2}\right)$ updates. Your technique should iteratively find a sequence of distributions $\mathbf{p}_t$, for $t = 1, \ldots, T$, using the exponential weights method, and then the "other player" should select a weak learner $h_{j_t} \in \mathcal{H}$ which satisfies $\mathbf{p}_t^\top M\mathbf{e}_{j_t} \geq \gamma$. The latter is always guaranteed via the weak learning assumption. Your analysis should follow the standard analysis of solving a zero-sum game using no-regret algorithms, as described in class, but you'll need to think carefully how to extract the final distribution $\mathbf{q}^*$. (*Note:* this algorithm is very close to the Adaboost algorithm! It's not exactly the same only because Adaboost has an additional adaptive parameter update which doesn't require it to know γ in advance.)

Answer 1.

(a) We can set the matrix M as follows, $M_{ij} := y_i h_j(x_i)$. Let us note one nice fact, that the quantity $y_i h_j(x_i)$ can be rewritten as $1 - 2\mathbb{I}[y_i \neq h_j(x_i)]$, since $y_i h_j(x_i) = +1$ when $y_i = h_j(x_i)$ and $y_i h_j(x_i) = -1$ when $y_i \neq h_j(x_i)$. With this in mind, we see that

$$\begin{aligned}
\mathbf{p}^\top M\mathbf{e}_j &= \sum_{i=1}^{n} p_i y_i h_j(x_i) = \mathbb{E}_{i\sim\mathbf{p}}[y_i h_j(x_i)] \\
&= \mathbb{E}_{i\sim\mathbf{p}}[1 - 2\mathbb{I}[y_i \neq h_j(x_i)]] \\
&= 1 - 2 \cdot \mathbb{E}_{i\sim\mathbf{p}}\mathbb{I}[y_i \neq h_j(x_i)] \\
&= 1 - 2 \cdot \Pr_{i\sim\mathbf{p}}[y_i \neq h_j(x_i)].
\end{aligned}$$

Now assume that, for some j, we have that $\Pr_{i\sim\mathbf{p}}[y_i \neq h_j(x_i)] \leq \frac{1}{2} - \frac{\gamma}{2}$. Using the above calculation, and doing some minor arithmetic, we see that this is equivalent to the statement $\mathbf{p}^\top M\mathbf{e}_j \geq \gamma$.

The final piece we need to show is that the "for all there exists" statement is equivalent to the "min max" statement. Fix $\mathbf{p}$ for the moment, and consider the statement "there exists $h_j \in \mathcal{H}$, i.e. some $j \in [m]$, such that $\mathbf{p}^\top M\mathbf{e}_j \geq \gamma$." Conveniently, if we can find a j^* such that $\mathbf{p}^\top M\mathbf{e}_{j^*} \geq \gamma$, then certainly it has to be true that $\max_{j\in[m]} \mathbf{p}^\top M\mathbf{e}_j \geq \gamma$ since j^* is one of the potential maximizers. But the converse is also true! If $\max_{j\in[m]} \mathbf{p}^\top M\mathbf{e}_j \geq \gamma$, then take the maximizing j^*, and notice that for this particular index we have $\mathbf{p}^\top M\mathbf{e}_{j^*} \geq \gamma$. We

can thus conclude that "there exists some $j \in [m]$ such that $\mathbf{p}^\top M \mathbf{e}_j \geq \gamma$" is equivalent to $\max_{j \in [m]} \mathbf{p}^\top M \mathbf{e}_j \geq \gamma$.

Let's now follow the same argument for the minimization over $\mathbf{p}$. If for every $\mathbf{p} \in \Delta_n$ we have that $\max_{j \in [m]} \mathbf{p}^\top M \mathbf{e}_j \geq \gamma$, then notice that it must hold that $\min_{\mathbf{p} \in \Delta_n} \max_{j \in [m]} \mathbf{p}^\top M \mathbf{e}_j \geq \gamma$. But if you think carefully about this, this actually shows there's an *equivalence* here. In general, for any function f, we have

$$\min_{\mathbf{p} \in \Delta_n} f(\mathbf{p}) \geq \gamma \qquad \iff \qquad \forall\, \mathbf{p} \in \Delta_n : \quad f(\mathbf{p}) \geq \gamma.$$

It is important to note that this equivalence only holds because the inequality is $f(\mathbf{p}) \geq \gamma$ – had this been a $\leq$ the equivalence would not hold!

(b) A fact about linear optimization over a polytope is that the solution always occurs at the corners of the polytope. This holds the same for the corners of the simplex Δ_m, which are the basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_m$. So, in general, we have that, for any vector v,

$$\max_{\mathbf{p} \in \Delta_m} \mathbf{p}^\top v = \max_{i \in [m]} \mathbf{e}_i^\top v.$$

Apply this reasoning to the inner maximization, and we're done.

(c) We can follow the same logic we used in part (a). Fix $\mathbf{q}$, and notice that the statement $\min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} \geq \gamma$ is equivalent to the statement $\min_{i \in [n]} \mathbf{e}_i^\top M \mathbf{q} \geq \gamma$, using the same "linear optimization over a polytope" argument from (b). Furthermore, the latter is equivalent to the statement $\forall\, i \in [n] : \ \mathbf{e}_i^\top M \mathbf{q} \geq \gamma$. Now we can use the same equivalence argument in (a) to see that $\exists \mathbf{q} \in \Delta_m : \ \min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} \geq \gamma$ is equivalent to the statement $\max_{\mathbf{q} \in \Delta_m} \min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} \geq \gamma$. Putting this all together we have that

$$\max_{\mathbf{q} \in \Delta_m} \min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} \geq \gamma \iff \exists \mathbf{q} \in \Delta_m \forall\, i \in [n] : \ \mathbf{e}_i^\top M \mathbf{q} \geq \gamma.$$

Using the same interpretation laid out in (a) we see that the RHS of the above equation is exactly the strong learning assumption.

(d) The minimax theorem states that

$$\max_{\mathbf{q} \in \Delta_m} \min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} = \min_{\mathbf{p} \in \Delta_n} \max_{\mathbf{q} \in \Delta_m} \mathbf{p}^\top M \mathbf{q}.$$

It therefore follows from the above that

$$\max_{\mathbf{q} \in \Delta_m} \min_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top M \mathbf{q} \geq \gamma \iff \min_{\mathbf{p} \in \Delta_n} \max_{\mathbf{q} \in \Delta_m} \mathbf{p}^\top M \mathbf{q} \geq \gamma.$$

Of course, we already showed in parts (b) and (c) that the RHS is the weak learning assumption and the LHS is the strong learning assumption.

(e) (BONUS - 15 points) This problem requires a complex answer. We will grade this one separately so you do not need to worry about self-grading. However, if you are interested, you can find the detailed story in Chapter 6 of this book: `https://gatech.instructure.com/courses/278984/files/folder/Materials?preview=37704671`

Problem 2 (Understanding the pseudo-regret of greedy, pure-exploration and explore-then-commit) 20 points In class, we introduced the two-armed Bernoulli bandit problem through the motivation of drug discovery. In particular, we considered a 2-armed Bernoulli bandit problem with reward parameters $p_1 = 0.2$ and $p_2 = 0.7$, corresponding to the mean efficacy of Drugs A and B respectively on a patient. We claimed that the greedy algorithm and "only-explore" algorithm were both highly suboptimal in terms of overall reward, but the explore-then-commit algorithm did better. In this problem, we will formally explore this through the metric of *pseudo-regret*.

All expectations and probabilities will be over the randomness in the reward sequence $\{G_{t,i}\}_{t\geq 1}$ for $i = 1, 2$.

(a) Consider the greedy algorithm with the default choice of picking arm 1 (drug A) on round 1, arm 2 (drug B) on round 2, and greedy thereafter. What is the probability that you will *always* pick arm 1 round 3 onwards?

(b) Use the above sub-part to *lower bound* the pseudo-regret of the greedy algorithm. What is the dependence of this lower bound on T?

Hint: Under the situation of part (a), what will the expected reward be from rounds 3 *to* T*? How does this compare to the benchmark of* $0.7(T-2)$*?*

(c) Now, consider the "pure-explore" algorithm, which picks arm 1 on odd rounds and arm 2 on even rounds. Calculate the exact pseudo-regret of this algorithm.

Hint: use the formula for pseudo-regret in terms of the expected number of times the suboptimal arm was sampled.

(d) Finally, consider the explore-then-commit (ETC) algorithm, where the arms are sampled in a round-robin fashion for $T_0 < T$ rounds. Assume that T_0 is even. Use Hoeffding's bound to derive an *upper bound* on the probability that you will always pick the suboptimal arm 1 after T_0 rounds.

(e) Use part (e) to derive an upper bound for the pseudo-regret of ETC with T_0 exploration rounds. What is the value of T_0 that minimizes the upper bound? Use this choice to state an upper bound on pseudo-regret in terms of T, the total number of rounds.

(f) (BONUS – 10 points) Repeat all of the above parts for arbitrary values of p_1, p_2. Denote $\Delta = |p_2 - p_1|$ and express your eventual bounds on pseudo-regret as a function of Δ. Suppose that you did not know Δ beforehand. Then, what is the value of T_0 that minimizes pseudo-regret over all values of Δ?

Hint: use the fact that pseudo-regret is also upper-bounded by ΔT*.*

Answer 2 (Understanding the pseudo-regret of greedy, pure-exploration and explore-then-commit) 20 points

(a) As we saw in class, the greedy algorithm will always pick arm 1 round 3 onwards if we see $G_{1,1} = 1$ *and* $G_{2,2} = 0$. The probability of this event occurring is given by $0.2 \times 0.3 = 0.06$. For completeness, we will consider the other three cases below and show that arm 1 will not always be picked round 3 onwards. (This reasoning is not required for full credit.)

i) $G_{1,1} = 1$ and $G_{2,2} = 1$: then, the greedy algorithm will randomize and pick arm 1 with probability 0.5 and arm 2 with probability 0.5.

ii) $G_{1,1} = 1$ and $G_{2,2} = 1$: then, the greedy algorithm will randomize and pick arm 1 with probability 0.5 and arm 2 with probability 0.5.

iii) $G_{1,1} = 0$ and $G_{2,2} = 1$: then, the greedy algorithm will pick arm 2 at round 3.

(b) First, recall that the formula for the pseudo-regret is given by

$$\overline{R}_T = \sum_{a \neq a^*} \Delta_a \cdot \mathbb{E}[N_a(T)] = 0.5 \cdot \mathbb{E}[N_1(T)] \tag{1}$$

in our example. From the reasoning in part (a), when $G_{1,1} = 1$ and $G_{2,2} = 0$ we will have $N_1(T) = T-2+1 = T-1$. In all other cases, we have $N_1(T) \geq 0$. The law of total probability then gives us

$$\mathbb{E}[N_a(T)] \geq 1 + \mathbb{P}[G_{1,1} = 1, G_{2,2} = 0] \cdot (T-2) = 0.06(T-2) + 1$$

and so the pseudo-regret is $\overline{R}_T \geq 0.5 \cdot (0.06(T-2)+1) = 0.03(T-2) + 0.5$. This is clearly linear in T.

NOTE: an alternative solution would have involved using the formula for pseudoregret

$$\overline{R}_T := T\mu^* - \mathbb{E}[G_T] = 0.7T - \mathbb{E}[G_T]$$

in our example. Then, similar reasoning to the above gives us the same answer: now, we just need to *upper-bound* $\mathbb{E}[G_T]$ in order to lower bound $\overline{R}_T$. From the reasoning in part (a), when $G_{1,1} = 1$ and $G_{2,2} = 0$ we will have $G_T = 0.2+0.7+0.2(T-2) = 0.2(T-1)+0.7$. In all other cases we will have $G_T \leq 0.7T$. Consequently, we get $\mathbb{E}[G_T] \leq 0.06 \cdot (0.2(T-1)+0.7)+0.94 \cdot 0.7T$, and subtracting this upper bound from $T\mu^* = 0.7T$ gives us the same answer.

These calculations clearly demonstrate that the alternative solution is more complicated to reason about, and show the convenience of working with the formula for pseudo-regret instead of the expected number of suboptimal arm pulls.

(c) We will again use the provided formula for pseudo-regret in terms of the expected number of suboptimal arm pulls, i.e. $\overline{R}_T = 0.5 \cdot \mathbb{E}[N_1(T)]$. Now, the "pure explore" algorithm picks arm 1 on all odd rounds $1, 3, 5, \ldots$, and arm 2 on all even rounds $2, 4, 6, \ldots$ *regardless* of what the realized rewards are. Consider the impact that this has on the pseudo-regret for T rounds in two cases:

i) if T is even, then $N_1(T) = T/2$ and so the pseudo-regret is given by $\overline{R}_T = 0.25T$.

ii) if T is odd, then $N_1(T) = (T+1)/2$ (verify this for yourself with a simple example, e.g. $T = 5$ yields $N_1(T) = 3 = 6/2$), and so the pseudo-regret is given by $\overline{R}_T = 0.5 \cdot (T+1)/2 = 0.25(T+1)$.

In both cases, we see that the pseudo-regret is linear in T.

(d) First, we will use Hoeffding's bound to upper bound the probability that the suboptimal arm 1 is picked after T_0 rounds. Our upper bound will be a function of T_0. We note that

$$\begin{aligned} \mathbb{P}[\text{ arm 1 picked on all rounds } t > T_0] &= \mathbb{P}\left[\widehat{\mu}_{1,T_0} > \widehat{\mu}_{2,T_0}\right] \\ &= \mathbb{P}\left[\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0} > 0\right]. \end{aligned}$$

Next, we note that $\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0}$ is an unbiased estimator of $\mu_1 - \mu_2 = 0.2 - 0.7 = -0.5$, constructed from $T_0/2$ samples of the random variable $G_1 - G_2$, where G_1 and G_2 denote the reward random variables for arms 1 and 2 respectively. By definition, we have $G_1 \sim$ Bernoulli(0.2) and $G_2 \sim$ Bernoulli(0.7). Because G_1 and G_2 both take values in $\{0, 1\}$, we have $G_1 - G_2 \in [-1, 1]$. Therefore, we substitute Hoeffding's lemma with the values $a = -1, b = 1$ (Theorem 1 in the review lecture on probability and statistics) to get

$$\begin{aligned}
\mathbb{P}\left[\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0} > 0\right] &= \mathbb{P}\left[\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0} - (-0.5) > 0.5\right] \\
&\leq \exp\left(-\frac{2(T_0/2)(0.5)^2}{(1-(-1))^2}\right) \\
&= \exp\left(-\frac{0.25T_0}{4}\right).
\end{aligned}$$

Thus, we see an exponentially decaying relationship in the probability of picking the suboptimal arm 1 after round T_0 as a function of T_0: the more we explore, the less likely we are to pick the suboptimal arm. This further gives us $N_1(T) \leq \frac{T_0}{2}$ with probability at most 1, and $N_1(T) \leq T$ with probability at most $\exp\left(-\frac{0.25T_0}{4}\right)$. Consequently, we get

$$\begin{aligned}
\mathbb{E}[N_1(T)] &= \mathbb{P}[\text{arm 1 picked after round } T_0] \cdot T + \mathbb{P}[\text{arm 1 eliminated after round } T_0] \cdot \frac{T_0}{2} \\
&\leq \exp\left(-\frac{0.25T_0}{4}\right) \cdot T + 1 \cdot \frac{T_0}{2}
\end{aligned}$$

and substituting this into Equation (1) gives us

$$\overline{R}_T \leq 0.5 \cdot \frac{T_0}{2} + \exp\left(-\frac{0.25T_0}{4}\right) \cdot 0.5T. \tag{2}$$

NOTE: we will allow for flexibility in the eventual expression that was derived here, particularly in the dependence on the constants. As long as you got the essence of these steps correct, you will get full credit for your answer. There are also possibly sharper expressions that you could obtain in the dependences on T and T_0; if you obtained those expressions instead, you get full credit.

(e) Equation (2) in the solution to part (d) gives us an upper bound on the pseudo-regret for different values of T_0, which is the number of exploration rounds. We now explore the choice of T_0 that would minimize the upper bound

$$f(T_0) := 0.5 \cdot \frac{T_0}{2} + \exp\left(-\frac{0.25T_0}{4}\right) \cdot 0.5T.$$

Since $f(T_0)$ is a convex function in T_0, we can differentiate it with respect to T_0 and set the derivative to 0 to find the value of T_0 that minimizes it. The derivative is given by

$$f'(T_0) = \frac{0.5}{2} - \frac{0.25}{4} \cdot 0.5T \cdot \exp\left(-\frac{0.25T_0}{4}\right).$$

Now, setting this to 0 yields

$$\frac{0.25}{4} \cdot 0.5T \cdot \exp\left(-\frac{0.25T_0}{4}\right) = \frac{0.5}{2}$$
$$\implies \exp\left(\frac{0.25T_0}{4}\right) = \frac{T}{8}$$
$$\implies T_0 = \frac{4}{0.25}\log\left(\frac{T}{8}\right) = 16\log\left(\frac{T}{8}\right).$$

This choice of T_0 then gives us

$$\overline{R}_T \leq 4\log\left(\frac{T}{8}\right) + 4,$$

which is logarithmic in T.

NOTE: your constants may differ depending on the answer that you derived to part (d). This suffices for full credit.

(f) (BONUS) We provide brief solutions for each of the parts (a) – (e), and then provide the special solution for the second half of the question that is specific to this subpart. You can assume that $p_1 < p_2$ to get full credit on this question (but you are welcome to evaluate the case where $p_1 > p_2$ as well). The solutions that we present below will assume that $p_1 < p_2$.

Parts (a) and (b): The probability that $G_{1,1} = 1$ and $G_{2,2} = 0$ is equal to $p_1(1-p_2)$, and the suboptimality gap is equal to $\Delta = p_2 - p_1$. Therefore, the pseudo-regret of the greedy algorithm is lower-bounded by $\overline{R}_T \geq p_1(1-p_2) \cdot \Delta T$.

Part (c): The "pure-explore" continues to pick the suboptimal arm 1 $\frac{T}{2}$ times when T is even, and $\frac{T+1}{2}$ times when T is odd. Therefore, the pseudo-regret of the pure-explore algorithm is equal to $\frac{\Delta T}{2}$ when T is odd, and $\frac{\Delta(T+1)}{2}$ when T is even.

Part (d): We note that $\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0}$ is an unbiased estimator of $\mu_1 - \mu_2 = p_1 - p_2 = -\Delta$. This gives us

$$\begin{aligned}\mathbb{P}\left[\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0} > 0\right] &= \mathbb{P}\left[\widehat{\mu}_{1,T_0} - \widehat{\mu}_{2,T_0} - (-\Delta) > \Delta\right] \\ &\leq \exp\left(-\frac{2(T_0/2)(\Delta)^2}{(1-(-1))^2}\right) \\ &= \exp\left(-\frac{T_0\Delta^2}{4}\right).\end{aligned}$$

As before, we see an exponentially decaying relationship in the probability of picking the suboptimal arm 1 after round T_0 as a function of T_0: the more we explore, the less likely we are to pick the suboptimal arm. This further gives us $N_1(T) \leq \frac{T_0}{2}$ with probability at most 1, and $N_1(T) \leq T$ with probability at most $\exp\left(-\frac{T_0\Delta^2}{4}\right)$. Consequently, we get

$$\mathbb{E}[N_1(T)] \leq \frac{T_0}{2} + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot T,$$

and substituting this into Equation (1) gives us

$$\overline{R}_T \leq \Delta \cdot \frac{T_0}{2} + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot \Delta T. \tag{3}$$

Now, we define

$$f_\Delta(T_0) = \Delta \cdot \frac{T_0}{2} + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot \Delta T,$$

and examine the value of T_0 that will minimize the upper bound above. In particular, taking the first derivative yields

$$f'_\Delta(T_0) = \frac{\Delta}{2} - \frac{\Delta^3 T}{4} \cdot \exp\left(-\frac{T_0\Delta^2}{4}\right) = 0 \text{ when}$$
$$T_0 = \frac{4}{\Delta^2} \cdot \log\left(\frac{T\Delta^2}{2}\right).$$

and substituting this back in gives us regret

$$\overline{R}_T \leq \frac{2}{\Delta} \cdot \log\left(\frac{T\Delta^2}{2}\right) + \frac{2}{\Delta},$$

which is logarithmic in T for a fixed Δ. So in this case, explore-then-commit can indeed achieve logarithmic regret.

However, there is an important catch. Note that the above choice of T_0 depends on Δ. The last part of the bonus question asked for the value of T_0 that minimizes pseudo-regret for all values of $\Delta \in [0, 1]$. For this, recall that we had

$$f_\Delta(T_0) = \Delta \cdot \frac{T_0}{2} + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot \Delta T$$

as the upper bound on the regret of explore-then-commit for an instance with gap Δ and exploration amount set to T_0. Essentially, we wish to find T_0 to *minimize* an upper bound on $f^*(T_0) := \max_{\Delta\in[0,1]} f_\Delta(T_0)$. We begin by upper-bounding $f^*(T_0)$. We do this in a number of steps:

- First, for $\Delta \leq \frac{2}{\sqrt{T_0}}$, we have $f_\Delta(T_0) \leq \Delta T \leq \frac{2T}{\sqrt{T_0}}$, where the equality is achieved at $\Delta = \frac{2}{\sqrt{T_0}}$.
- Second, for $\Delta \geq \frac{2}{\sqrt{T_0}}$, we have $f_\Delta(T_0) = \Delta \cdot \frac{T_0}{2} + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot \Delta T \leq T_0 + \exp\left(-\frac{T_0\Delta^2}{4}\right) \cdot \Delta T$. Then, we can show that the function $g(\Delta) := \Delta \cdot \exp\left(-\frac{T_0\Delta^2}{4}\right)$ is *decreasing* in Δ when $\Delta \geq \frac{2}{\sqrt{T_0}}$. Therefore, we have $\Delta \cdot \exp\left(-\frac{T_0\Delta^2}{4}\right) \leq \frac{2}{\sqrt{T_0}} \cdot e^{-1} \leq \frac{2}{\sqrt{T_0}}$, and altogether we get

$$f_\Delta(T_0) \leq T_0 + \frac{2T}{\sqrt{T_0}} \text{ for all } \Delta \geq \frac{2}{\sqrt{T_0}}.$$

Together, these give us

$$f^*(T_0) \leq \max\{\frac{2T}{\sqrt{T_0}}, T_0 + \frac{2T}{\sqrt{T_0}}\} \leq T_0 + \frac{2T}{\sqrt{T_0}}$$

Thus, we need to find T_0 that minimizes $T_0 + \frac{2T}{\sqrt{T_0}}$. This value of T_0 will be the one that satisfies

$$\frac{2T}{\sqrt{T_0}} = T_0$$
$$\implies T_0 = (2T)^{2/3}.$$

Plugging this back in gives us the regret bound $\overline{R}_T = \mathcal{O}(T^{2/3})$ for the explore-then-commit algorithm.

Please be liberal in grading yourselves for the last part of this bonus question. As you can see above, solving for the exactly optimal value of T_0 is a little tedious. Any heuristic reasoning that has the right idea, and gets the right answer of $R_T = \mathcal{O}(T^{2/3})$ suffices to get full credit on this part.

Problem 3 (The successive arm elimination algorithm) 20 points In this problem, we examine the psudo-regret of the successive arm elimination (SAE) algorithm for the 2-armed Bernoulli bandit problem with reward parameters μ_1 and μ_2 both bounded between $[0, 1]$. Without loss of generality, assume $\mu_1 < \mu_2$.

This algorithm is very similar in some ways to the UCB algorithm, but a little different in its day-to-day behavior. Instead of constantly keeping all arms in play like UCB, it plays a current set of "active" arms in a round-robin fashion, and eliminates arms that seem to be performing suboptimally at the end of each round-robin turn. In more detail, the basic procedure of SAE at round t, when both arms are in play, is as follows:

- Define the upper confidence bound of arm $a \in \{1, 2\}$ as $\mathsf{UCB}(a, t) = \hat{\mu}_{a,t-1} + \sqrt{\frac{\log(1/\delta)}{2N_{t-1}(a)}}$, and the lower confidence bound $\mathsf{LCB}(a, t) = \hat{\mu}_{a,t-1} - \sqrt{\frac{\log(1/\delta)}{2N_{t-1}(a)}}$.
- If t is odd, play arm 1.
- If t is even:
  - Play arm 2.
  - If $\mathsf{UCB}(1, t+1) < \mathsf{LCB}(2, t+1)$, eliminate arm 1 and *play arm* 2 *on all rounds there-after.*
  - Else, if $\mathsf{UCB}(2, t+1) < \mathsf{LCB}(1, t+1)$, then eliminate arm 2 and *play arm* 1 *on all rounds there-after.*
  - If neither elimination criterion is met, then keep both arms in play and proceed to the next round.

(a) Set $\delta = 1/T^2$ (as we did for UCB), and define the "good event" $\mathcal{A}$ as $\mathcal{A} = \{\mu_i \in [\mathsf{LCB}(i, t), \mathsf{UCB}(i, t)]$ for all $i \in \{1, 2\}, t \in \{1, \ldots, T\}.\}$. Prove that $\mathcal{A}$ occurs with probability at least $1 - 2/T$.

(b) Let $\hat{t}$ denote the round on which we eliminate an arm. Conditioned on the "good event" $\mathcal{A}$, prove that the pseudo-regret incurred on rounds $t = \hat{t}+1, \ldots, T$ is equal to 0.

Hint: Use the elimination criterion to show that when the "good event" $\mathcal{A}$ holds, the eliminated arm must be the suboptimal arm 1. Drawing a picture of the confidence intervals $[\mathsf{LCB}, \mathsf{UCB}]$ for each arm might help.

(c) Next, let's consider the pseudo-regret for $t < \hat{t}$. According to the elimination rule, we know that round $\hat{t}-2$ is the last round that we do *not* drop any arm. Based on this fact, prove that $\Delta = |\mu_1 - \mu_2| \leq 4\sqrt{\frac{2\log(T)}{(\hat{t}-2)}}$ under the "good event" $\mathcal{A}$.

Hint: use the solution to part (a) together with the elimination criterion to make a conclusion about Δ.

(d) Based on part (c), provide an upper-bound on the pseudo-regret of SAE from round 1 to round $\hat{t}$ conditioned on the "good event" $\mathcal{A}$ that depends only on T and Δ.

(e) Combine the above results and provide an upper-bound on the overall pseudo-regret for SAE from round 1 to round T (considering both the "good event" $\mathcal{A}$ and the "bad event" $\mathcal{A}^C$). Like part (d), your upper bound should depend only on T and Δ.

Answers to Problem 3 (The successive arm elimination algorithm) 20 points

(a) Let $\hat{\mu}_a(n)$ be the mean rewards of pulling arm a for n times. Then, based on Hoeffding's inequality, we have

$$\mathbb{P}\left[\hat{\mu}_a(n) \leq \mu_a - \sqrt{\frac{\log T}{n}}\right] \leq \exp\left(-\frac{2n\log T}{n}\right) = \frac{1}{T^2}$$

and

$$\mathbb{P}\left[\hat{\mu}_a(n) \geq \mu_a + \sqrt{\frac{\log T}{n}}\right] \leq \exp\left(-\frac{2n\log T}{n}\right) = \frac{1}{T^2}$$

Thus, taking a union bound for all values $n = 1, \ldots, T$ and all arms $a \in \{1, 2\}$, we get that the bad event happens with probability at most $4 \cdot T \cdot \frac{1}{T^2} = \frac{4}{T}$. Therefore, the good event happens with probability at least $1 - \frac{4}{T}$.

(b) We present an algebraic solution below, but you may find it helpful to draw a picture to get intuition about the solution. Let $[\mathsf{LCB}(i,t), \mathsf{UCB}(i,t)]$ be the "confidence region" of arm i. Let j denote the identity of the eliminated arm, and j' the identity of the arm that remains in play—and let $\hat{t}$ denote the round on which j is eliminated. At round $\hat{t}$, the elimination criterion uses the UCB and LCB indexes that we would obtain* at round $\hat{t}+1$. Then, based on the elimination rule of SAE, we have $\mathsf{UCB}(j, \hat{t}+1) < \mathsf{LCB}(j', \hat{t}+1)$. Moreover, conditioned on the "good event" $\mathcal{A}$, we know that the true means are within the confidence regions; in other words, we have

$$\begin{aligned}\mu_j &\leq \mathsf{UCB}(j, \hat{t}+1) \text{ and} \\ \mu_{j'} &\geq \mathsf{LCB}(j', \hat{t}+1).\end{aligned}$$

Putting all of this together gives us $\mu_j \leq \mathsf{UCB}(j, \hat{t}+1) < \mathsf{LCB}(j', \hat{t}+1) \leq \mu_{j'}$ and so in summary we have $\mu_j < \mu_{j'}$. Since $\mu_1 < \mu_2$, this means that the eliminated arm must be $j = 1$ (and the arm that remains must be $j' = 2$).

**The indexing of UCB and LCB can be confusing. If you instead wrote the indices in terms of $\hat{t}$ for the UCB and LCB, we will give you full credit as long as you used the correct reasoning provided above.*

(c) Since neither of the arms has been eliminated before round $\hat{t}$, we must have $\mathsf{UCB}(1, \hat{t}-1) \geq \mathsf{LCB}(2, \hat{t}-1)$ (using the fact that arm 1 in particular is not eliminated yet). Recall that we denote $N_t(1), N_t(2)$ as the number of times arms 1 and 2 have been pulled by round t. Since neither arm has been eliminated, we have $N_{\hat{t}-2}(1) = N_{\hat{t}-2}(2) = \frac{\hat{t}-2}{2}$. We denote this value as

$N_{\hat{t}-2}$ as shorthand. This gives us

$$\begin{aligned}
\mathsf{UCB}(1,\hat{t}-1) \geq \mathsf{LCB}(2,\hat{t}-1) &\implies \hat{\mu}_{1,\hat{t}-2} + \sqrt{\frac{\log T}{N_{\hat{t}-2}}} \geq \hat{\mu}_{2,\hat{t}-2} - \sqrt{\frac{\log T}{N_{\hat{t}-2}}} \\
&\implies \hat{\mu}_{1,\hat{t}-2} - \sqrt{\frac{\log T}{N_{\hat{t}-2}}} \geq \hat{\mu}_{2,\hat{t}-2} - 3\sqrt{\frac{\log T}{N_{\hat{t}-2}}} \\
&\implies \hat{\mu}_{1,\hat{t}-2} - \sqrt{\frac{\log T}{N_{\hat{t}-2}}} \geq \hat{\mu}_{2,\hat{t}-2} + \sqrt{\frac{\log T}{N_{\hat{t}-2}}} - 4\sqrt{\frac{\log T}{N_{\hat{t}-2}}} \qquad (4) \\
&\implies \mathsf{LCB}(1,\hat{t}-1) \geq \mathsf{UCB}(2,\hat{t}-1) - 4\sqrt{\frac{\log T}{N_{\hat{t}-2}}} \\
&\implies \mathsf{UCB}(2,\hat{t}-1) - \mathsf{LCB}(1,\hat{t}-1) \leq 4\sqrt{\frac{\log T}{N_{\hat{t}-2}}}.
\end{aligned}$$

On the other hand, under the good event $\mathcal{A}$, we have $\mu_1 \geq \mathsf{LCB}(1,\hat{t}-1)$ and $\mu_2 \leq \mathsf{UCB}(2,\hat{t}-1)$. This gives us

$$\begin{aligned}
\Delta := \mu_2 - \mu_1 &\leq \mathsf{UCB}(2,\hat{t}-1) - \mathsf{LCB}(1,\hat{t}-1) \\
&\leq 4\sqrt{\frac{\log T}{N_{\hat{t}-2}}} = 4\sqrt{\frac{2\log T}{(\hat{t}-2)}},
\end{aligned}$$

which is exactly what we want.

Note: As in part (b), the indexing of UCB and LCB can be confusing. If you instead wrote the indices in terms of $\hat{t}-2$ for the UCB and LCB, we will give you full credit as long as you used the correct reasoning provided above (and, importantly, used the statistics $N_{\hat{t}-2}(1), N_{\hat{t}-2}(2)$).

(d) Part (c) gave us $\Delta \leq 4\sqrt{\frac{2\log T}{(\hat{t}-2)}}$. Squaring both sides and rearranging terms, we get

$$\hat{t} \leq \frac{32\log T}{\Delta^2} + 2.$$

From round 1 to $\hat{t}$, we pulled the sub-optimal arm $\frac{\hat{t}}{2}$ times. Thus, the regret under the good event $\mathcal{A}$ is upper bounded by

$$N_a(\hat{t}) \cdot \Delta = \frac{\hat{t}}{2} \cdot \Delta \leq \frac{16\log T}{\Delta} + \Delta.$$

(e) As we did in the lecture notes (for the UCB algorithm), we can write the pseudo-regret as

$$\begin{aligned}
\overline{R}_T &\leq \mathbb{P}[\mathcal{A} \text{ happens}] \left(\frac{16\log T}{\Delta} + \Delta\right) + (1 - \mathbb{P}[\mathcal{A} \text{ happens}]) \cdot \Delta T \\
&\leq \frac{16\log T}{\Delta} + \Delta + \frac{4}{T} \cdot \Delta T \\
&= \frac{16\log T}{\Delta} + \Delta + 4\Delta = \frac{16\log T}{\Delta} + 5\Delta.
\end{aligned}$$

The first inequality substitutes the upper bound on regret on the good event $\mathcal{A}$ from part (d), and under the bad event notes that the regret will be at most ΔT. The second inequality uses the crude upper bound $\mathbb{P}[\mathcal{A} \text{ happens}] \leq 1$, and substitutes the upper bound from part (a).