Theoretical Machine Learning
Homework 4

Instructor: Haipeng Luo

This homework is due on **12/10, 11:59pm**. See course website for more instructions on finishing and submitting your homework as well as late policy.

1. (**Stochastic MAB**) Consider MAB in the stochastic setting. In Lecture 8, we proved (in Theorem 3) that UCB strategy satisfies

$$\mathbb{E}\left[m_n(a)\right] = \mathcal{O}\left(\frac{\ln n}{\Delta_a^2}\right) \tag{1}$$

where $m_n(a)$ is the total number of times arm a is selected and $\Delta_a = \mu(a) - \mu(a^\star)$ is the optimality gap with $a^\star \in \operatorname{argmin}_a \mu(a)$ being an optimal arm. Clearly, any algorithm satisfying Equation (1) ensures the following gap-dependent pseudo regret bound:

$$\overline{\mathrm{Reg}}_n = \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu(a_t) - \mu(a^\star)\right)\right] = \sum_{a:\Delta_a>0} \Delta_a \mathbb{E}\left[m_n(a)\right] = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\ln n}{\Delta_a}\right).$$

Follow the two steps below to prove that any algorithm satisfying Equation (1) also ensures the following gap-independent bound:

$$\overline{\mathrm{Reg}}_n = \mathcal{O}\left(\sqrt{nK\ln n} + K\ln n\right). \tag{2}$$

   (a) (3pts) First show that if Equation (1) holds, then we have the following for any $\Delta \in [0,1]$:

$$\overline{\mathrm{Reg}}_n \le \Delta n + \mathcal{O}\left(\sum_{a:\Delta_a>\Delta} \frac{\ln n}{\Delta_a}\right). \tag{3}$$

   (b) (2pts) Conclude the regret bound of Equation (2) by picking an appropriate choice of Δ in Equation (3).

2. (**Multiclass Perceptron**) In this exercise you need to analyze variants of the Perceptron algorithm for multiclass classification, with either full information or bandit information. Consider a sequence of examples $x_1, \ldots, x_n \in B_2^d$ with labels $y_1, \ldots, y_n \in [K]$ where K is the number of possible classes. We assume that the following multiclass margin assumption holds: there exists a constant $\gamma > 0$ and K weight vectors $\theta_\star^1, \ldots, \theta_\star^K \in B_2^d$ such that for each $t = 1, \ldots, n$:

$$\left\langle \theta_\star^{y_t}, x_t \right\rangle \ge \left\langle \theta_\star^k, x_t \right\rangle + \gamma, \quad \forall k \ne y_t.$$

In other words, the predictor $\operatorname{argmax}_k \left\langle \theta^k, x_t \right\rangle$ makes perfect predictions for this dataset with γ margin. Now consider the following learning protocol:

For $t = 1, \dots, n$:

- receive x_t and predict $s_t \in [K]$;
- observe $\begin{cases} y_t & \text{in the full-information setting} \\ \mathbf{1}\{s_t \neq y_t\} \text{ (i.e., if the prediction is correct)} & \text{in the bandit setting} \end{cases}$

In either case, we care about the total number of mistakes $\sum_{t=1}^{n} \mathbf{1}\{s_t \neq y_t\}$.

(a) (8pts) In the full information setting, one can apply the following multiclass Perceptron algorithm, a natural generalization of its binary version studied in Lecture 7. Note that when the algorithm predicts correctly, the last update step in fact does nothing (similarly to the binary version).

Algorithm 1: Multiclass Perceptron

Initialize $\theta^1 = \cdots = \theta^K = \mathbf{0}$.
For $t = 1, \dots, n$:

- receive x_t and find $k_t \in \operatorname{argmax}_{k \in [K]} \langle \theta^k, x_t \rangle$;
- predict $s_t = k_t$;
- receive y_t and update

$$\theta^{y_t} \leftarrow \theta^{y_t} + x_t \quad \text{and} \quad \theta^{k_t} \leftarrow \theta^{k_t} - x_t.$$

Prove that under the margin assumption, this algorithm makes at most $\frac{2K}{\gamma^2}$ mistakes. Hint: similarly to the proof for Theorem 1 of Lecture 7, you need to analyze the evolution of the quantities $\sum_{k=1}^{K} \langle \theta^k, \theta_\star^k \rangle$ and $\sum_{k=1}^{K} \left\| \theta^k \right\|_2^2$. (In this process you also need the Cauchy-Schwarz inequality.)

(b) In the bandit setting, we make the following two changes to Algorithm 1: 1) first, in light of exploration vs. exploitation trade-off, it is natural to randomize the algorithm and explore every label with at least some small probability α; 2) second, the update $\theta^{y_t} \leftarrow \theta^{y_t} + x_t$ becomes invalid if the prediction is incorrect (since we do not know what y_t is), so we only do this update when we predict correctly, and we scale the update with the inverse probability of selecting the correct label, just like the idea of importance-weighted estimator in Exp3. The final algorithm is shown below.

Algorithm 2: Multiclass Perceptron with Bandit Feedback

Input: exploration parameter $\alpha \in (0, \frac{1}{2K}]$.
Initialize $\theta^1 = \cdots = \theta^K = \mathbf{0}$.
For $t = 1, \dots, n$:

- receive x_t and find $k_t \in \operatorname{argmax}_{k \in [K]} \langle \theta^k, x_t \rangle$;
- predict s_t drawn from p_t, where $p_t(k) = (1 - \alpha K)\mathbf{1}\{k = k_t\} + \alpha, \forall k$;
- receive $\mathbf{1}\{s_t \neq y_t\}$ and update

$$\theta^{y_t} \leftarrow \theta^{y_t} + \frac{x_t \mathbf{1}\{s_t = y_t\}}{p_t(y_t)} \quad \text{and} \quad \theta^{k_t} \leftarrow \theta^{k_t} - x_t.$$

Follow the steps below to prove that this algorithm makes at most $\mathcal{O}\left(\frac{K\sqrt{n}}{\gamma^2}\right)$ mistakes in expectation. We will use the notation $\theta_t^1, \dots, \theta_t^K$ to denote the value of the weight vectors $\theta^1, \dots, \theta^K$ at the beginning of round t.

i. (5pts) Prove for any $t = 1, \dots, n$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \langle \theta_{t+1}^k, \theta_\star^k \rangle\right] \geq \mathbb{E}\left[\sum_{k=1}^{K} \langle \theta_t^k, \theta_\star^k \rangle\right] + \gamma \mathbb{E}\left[\mathbf{1}\{k_t \neq y_t\}\right].$$

ii. (8pts) Prove for any $t = 1, \dots, n$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \left\|\theta_{t+1}^{k}\right\|_2^2\right] \leq \mathbb{E}\left[\sum_{k=1}^{K} \left\|\theta_{t}^{k}\right\|_2^2\right] + 2\mathbb{E}\left[\frac{\mathbf{1}\{k_t \neq y_t\}}{\alpha} + K\mathbf{1}\{k_t = y_t\}\right].$$

Hint: consider the two cases $k_t \neq y_t$ and $k_t = y_t$ separately.

iii. (5pts) Combine the results from the last two questions to conclude

$$M\gamma \leq \sqrt{2K\left(\frac{M}{\alpha} + Kn\right)}$$

where $M = \mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{k_t \neq y_t\}\right]$ (you need the Cauchy-Schwarz inequality again and the fact $\mathbb{E}[\sqrt{z}] \leq \sqrt{\mathbb{E}[z]}$ for any $z \geq 0$). Solving for M gives $M = \mathcal{O}\left(\frac{K}{\alpha\gamma^2} + \frac{K\sqrt{n}}{\gamma}\right)$ (you do not need to prove this step). Use this fact to prove that the expected total number of mistakes is bounded as

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{s_t \neq y_t\}\right] = \mathcal{O}\left(\frac{K\sqrt{n}}{\gamma} + \frac{K^2}{\gamma^2}\right)$$

with an appropriate choice of α.

(c) (8pts) Recall that in Lecture 7, we also show a mistake bound of order $\mathcal{O}\left(d \ln \frac{1}{\gamma}\right)$ for the binary case, by discretizing the weight vector space and applying Halving. Can you generalize this argument to show a mistake bound of order $\mathcal{O}\left(dK^2 \ln\left(\frac{1}{\gamma}\right)\right)$ for the multiclass case with bandit information? Describe the algorithm and show the analysis. (Hint: for the analysis you might find the inequality $1 + z \leq e^z, \forall z$ useful)

3. (**MAB with graph feedback**) Consider the following variant of MAB, where each of the K arms can be seen as a node of some fixed graph $G = ([K], E)$. Here, $E \subset [K] \times [K]$ is the set of directed edges of this graph, and we denote by $S(a) = \{b \in [K] \mid (a, b) \in E\}$ the set of arms to which arm a is connected. The learning protocol is as follows.

> The environment decides the loss vectors $\ell_1, \dots, \ell_n \in \{0, 1\}^K$ ahead of time.
> For $t = 1, \dots, n$:
> - learner selects an arm $a_t \in [K]$, and observes $\ell_t(b)$ for all $b \in S(a)$.

MAB is clearly a special case of this problem where E is the set of all self-loops.

(a) (3pts) Show that this problem is an instance of partial monitoring by specifying a loss matrix ℓ and a feedback matrix Φ.

(b) Note that because the loss structure is symmetric among the actions, every pair of actions in this partial monitoring problem is a neighboring pair.

i. (4pts) Show that if the graph is such that for every node $b \in [K]$, there exists $a \in [K]$ with $(a, b) \in E$, then the corresponding partial monitoring problem is globally observable.

ii. (4pts) Show that if the graph is such that for every node $b \in [K]$, either $(b, b) \in E$ or $(a, b) \in E$ for all $a \neq b$, then the corresponding partial monitoring problem is locally observable.