CS861: Theoretical Foundations of Machine Learning Lecture 20 - 10/18/2024

University of Wisconsin–Madison, Fall 2024

Lecture 20: Introduction to Online Learning

Lecturer: Kirthevasan Kandasamy *Scribed by: Guy Zamir, Lakshika Rathi*

In this lecture, we will introduce the online learning framework. In particular, we will discuss the experts problem and analyze the performance of the Hedge algorithm.

1 Online Learning and The Experts Problem

To motivate the ensuing model, we will begin with two examples.

Example 1 (Online spam detection)**.** Given a hypothesis class $\mathcal{H}$ of binary classifiers, where $\mathcal{H} \in \{h : \mathcal{X} \to \{0, 1\}\}$. Consider the following game over T rounds:

1. A learner receives an input email $n_t \in \mathcal{X}$ on round t.
2. The learner chooses some $h_t \in \mathcal{H}$ and predicts $h_t(n_t)$ (spam or not-spam).
3. Learner sees the label y_t and incurs loss $\mathbb{1}\{h_t(n_t) \neq y_t\}$.

Note that the learner knows the loss for all $h \in \mathcal{H}$.

Example 2 (Weather forecasting)**.** Given a set of models $\mathcal{H}$. Consider the following game over T rounds:

1. Learner (weather forecaster) chooses some model $h \in \mathcal{H}$ and predicts the number $\hat{y}_t$.
2. Learner observes the true weather y_t and incurs loss $\ell(y_t, \hat{y}_t)$.

We can now introduce **Expert Problem**, which proceeds over T rounds in the following fashion:

1. We are given a set of experts $\{1, \ldots, K\}$, denoted $[K]$.
2. On each round, the learner chooses an expert (a.k.a. action) $A_t \in [K]$. Simultaneously, an adversary selects a loss vector $\ell_t \in [0, 1]^K$, where $\ell_t(i)$ is the loss for expert i.
3. Learner incurs loss $\ell_t(A_t)$.
4. Learner observes ℓ_t, the loss incurred by each expert.

This type of feedback, where we observe the loss for all actions, is called full information feedback. The setting we considered in previous lectures, where we observe losses or rewards only for the action we take, is called bandit feedback. Unlike in the stochastic bandit setting, however, note that we will not assume that the loss vectors are drawn from some distribution. Then how do we define regret? Recall that in the stochastic setting, we let $a_\star = \arg\min_{i \in [K]} \mathbb{E}_{X \sim \nu_i}[X]$ be the action with the highest expected reward and defined the regret as follows:

$$R_T^{Stochastic}(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] - Ta_*.$$

We can define the regret similarly for full information feedback. Here, in the non-stochastic setting where loss vectors are arbitrary, we will compete against the best fixed action in hindsight. For a policy π, and a sequence of losses, $\ell = \ell_1, ..., \ell_t$, define

$$R'_T(\pi, \ell) = \sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a).$$

The **regret** of a stochastic policy π is defined as

$$R_T(\pi, \ell) = \mathbb{E}\left[R'_T(\,\pi, \ell)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a),$$

where $\mathbb{E}$ is with respect to the randomness of the policy.

For a given policy π, we wish to bound $R_T(\pi, \ell)$ for all loss sequences. That is $\sup_\ell R_T(\pi, \ell)$. We wish to do well even if the losses were generated by an adversary that has full knowledge of our policy π. For now, we start by considering the case where the adversary is **oblivious**, meaning that ℓ_t only depends on the action taken on the current round. We will revisit this assumption in Section 3.

2 The Hedge Algorithm

2.1 Follow The Leader

The most intuitive approach to solve this problem is to choose the action $A_t = \arg\min_{a \in [K]} \sum_{s=1}^{t-1} \ell_s(a)$ on round t. This strategy is called Follow The Leader (FTL). For instance, on the binary classification example, FTL is simply empirical risk minimization, as it selects

$$h_t = \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} 1(h(x_t) \neq y_t).$$

Unfortunately, FTL cannot guarantee sublinear regret (in fact, no deterministic policy can). Consider the following example on which FTL fails. Suppose $K = 2$, and define the loss vectors as follows:

$$\ell_t = \begin{cases} (0.5, 0) & \text{if } t = 1, \\ (1, 0) & \text{if } t \text{ is odd and } t > 1, \\ (0, 1) & \text{if } t \text{ is even.} \end{cases}$$

Then, FTL will choose

$$A_t = \begin{cases} 1 & \text{on odd rounds,} \\ 2 & \text{on even rounds.} \end{cases}$$

Observe that the total loss of FTL will be at least $T - 1$, while the best action in hindsight will have loss at most $T/2$. Hence, the regret of FTL is at least $T/2 - 1 \in \Omega(T)$.

2.2 The Hedge Algorithm

We now introduce the Hedge algorithm. Intuitively our approach is similar to FTL, but we instead use a soft version of the minimum, and we pick our action stochastically by sampling from a distribution which gives more weight to actions with small losses. We provide the Hedge algorithm below.

Algorithm 1 The Hedge Algorithm (a.k.a multiplicative weights, a.k.a exponential weights)

Given time horizon T, learning rate η
Let $L_0 \leftarrow \underline{0}_K$ (all zero vector in $\mathbb{R}^K$)
for $t = 1, ..., T$ **do**
  Set $P_t(a) \leftarrow \frac{e^{-2L_{t-1}(a)}}{\sum_{j=1}^K e^{-2L_{t-1}(j)}}, \forall a \in [K]$
  Sample $A_t \sim P_t$ (note that $P_t \in \Delta^K$)
  Incur loss $\ell_t(A_t)$, observe ℓ_t
  Update $\ell_t(a) \leftarrow L_{t-1}(a) + \ell_t(a), \forall a \in [K]$
end for

2.3 Analysis of Hedge

For any policy π, which on round t samples an action according to p_t, define the **pseudo-regret** relative to an action $k \in [K]$ as

$$\bar{R}_T(\pi, \ell, a) = \sum_{t=1}^T p_t^T \ell_t - \sum_{t=1}^T \ell_t(a).$$

Let $a^* = \arg\min_{a \in [K]} \sum_{t=1}^T \ell_t(a)$ be the best fixed action in hindsight. We have

$$\begin{aligned} R_T(\pi, \ell) &= \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t)\right] - \min_{a \in [K]} \sum_{t=1}^T \ell_t(a) \\ &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}\left[\ell_t(A_t) \mid p_t\right]\right] - \sum_{t=1}^T \ell_t(a^*) \\ &= \mathbb{E}[\bar{R}_T(\pi, \ell, a^*)]. \end{aligned}$$

Hence, if we can bound $\bar{R}_T(\pi, \ell, a)$ for any action $a \in [K]$ and any p chosen by π, then we can bound $R_T(\pi, \ell)$. This strategy is precisely how we will upper bound the regret achieved by the Hedge algorithm.

We start with a technical lemma, and then we will prove the main theorem. In the following analysis, we define define ℓ_t^2 to be the coordinate-wise square of ℓ, meaning $\ell_t^2(i) = (\ell_t(i))^2$.

Lemma 1 (Hedge Lemma)**.** *Let $p = (p_1, \ldots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. Then, for any set of loss vectors $\ell = (\ell_1, \ldots, \ell_T)$, where $\ell_t \in \mathbb{R}_+^K$, and any $a \in [K]$, if $p_t^T \ell_t \leq 1$ for all t, we have*

$$\bar{R}_T(\pi, \ell, a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^T p_t^T \ell_t^2.$$

Proof Define $\Phi_t = \frac{1}{\eta} \log\left(\sum_{a=1}^K e^{-\eta L_t(a)}\right)$. We have

$$\begin{aligned} \Phi_t - \Phi_{t-1} &= \frac{1}{\eta} \log\left(\frac{\sum_{a=1}^K e^{-\eta L_t(a)}}{\sum_{a=1}^K e^{-\eta L_{t-1}(a)}}\right) \\ &= \frac{1}{\eta} \log\left(\frac{\sum_{a=1}^K e^{-\eta L_{t-1}(a)} \cdot e^{-\eta \ell_t(a)}}{\sum_{a=1}^K e^{-\eta L_{t-1}(a)}}\right) \\ &= \frac{1}{\eta} \log\left(\sum_{a=1}^K p_t(a) e^{-\eta l_t(a)}\right) \\ &\leq \frac{1}{\eta} \log\left(\sum_{a=1}^K p_t(a)(1 - \eta \ell_t(a) + \eta^2 \ell_t^2(a))\right) \qquad (\text{as } e^{-y} \leq 1 - y + y^2 \ \forall y \geq -1) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\eta}\log(1-\eta p_t^T\ell_t+\eta^2 p_t^T\ell_t^2)\\
&\le \frac{1}{\eta}(-\eta p_t^T\ell_t+\eta^2 p_t^T\ell_t^2) \qquad\qquad (\text{since } \log(1+y)\le y\ \forall y\ge -1 \text{ and } \eta p_t^\top \ell_t \le 1)\\
&= -p_t^T\ell_t+\eta p_t^T\ell_t^2.
\end{aligned}
$$

We have $\Phi_t-\Phi_{t-1}\le -p_t^T\ell_t+\eta p_t^T\ell_t^2$, so $\Phi_T-\Phi_0\le -\sum_{t=1}^T p_t^T\ell_t+\eta\sum_{t=1}^T p_t^T\ell_t^2$. Moreover,

$$
\begin{aligned}
\Phi_0 &= \frac{1}{\eta}\log(\sum_{i=1}^K e^{-\eta L_0(i)}) = \frac{\log(K)}{\eta} \qquad\qquad (\text{as } L_0=\mathbf{0})\\
\Phi_T &= \frac{1}{\eta}\log(\sum_{i=1}^K e^{-\eta L_T(i)}) \ge \frac{1}{\eta}\log(e^{-\eta L_t(a)}) = -L_T(a) = -\sum_{t=1}^T \ell_t(a).
\end{aligned}
$$

Consequently,

$$
-\sum_{t=1}^T \ell_t(a)-\frac{\log(K)}{\eta}\le -\sum_{t=1}^T p_t^T\ell_t+\eta\sum_{t=1}^T p_t^T\ell_t^2,
$$

so we conclude that

$$
\bar{R}_T(\pi,\ell,a)=\sum_{t=1}^T p_t^T\ell_t-\sum_{t=1}^T \ell_t(a)\le \frac{\log(K)}{\eta}+\eta\sum_{t=1}^T p_t^T\ell_t^2.
$$

□

Theorem 3 (Regret Bound of Hedge)**.** *Suppose* $\ell_t\in[0,1]^K\ \forall t$ *and choose* $\eta=\sqrt{\frac{\log(K)}{T}}$*. Then for all* $T\ge\log(K)$*, the regret of Hedge satisfies*

$$
R_T(\pi^{Hedge},\ell)\le 2\sqrt{T\log(K)}.
$$

Proof Let us first check the conditions to satisfy the lemma:

$$
T\ge\log K\Rightarrow\eta\le 1, \qquad\qquad \ell_t\in[0,1]^K\Rightarrow p_t^T\ell_t\le 1.
$$

Then, as $\ell_t^2(a)\le 1$ for all a, we have $p_t^T\ell_t^2\le 1$. Subsequently, for any $p=(p_1,\dots,p_T)$ chosen by Hedge,

$$
\bar{R}_T(\pi,\ell,a)\le\frac{\log(K)}{\eta}+\eta T\le 2\sqrt{2\log(K)}.
$$

Thus,

$$
R_T(\pi^{\text{Hedge}},\ell)=\mathbb{E}\left[\bar{R}_T(p,\ell,a_*)\right]\le 2\sqrt{T\log(K)}.
$$

□

3 Adversarial Bandits

Thus far we have designed a policy π to minimize $\sup_\ell R_T(\pi,\ell)$ where ℓ is chosen by an oblivious adversary. As we have already stated, this means that the adversary chooses the entire loss sequence $\ell=(\ell_1,\dots,\ell_T)$ ahead of time, or in other words, ℓ_t can only depend on the action $i\in[K]$.

Now, what if ℓ was instead chosen by an **adaptive adversary**? That is, the adversary can choose a loss ℓ_t on round t depending on the history $A_1,\ell_1,\dots,A_{t-1},\ell_{t-1}$. In this setting, called adversarial bandits, we usually restrict to bandit feedback. In other words, we can view adversarial bandits as a variant of the expert problem, but where the learner observes the loss only for the action taken. It has the following components:

1. There are a set of K experts, denoted $[K]$.
2. On round t, the learner chooses an expert (a.k.a. action) $A_t \in [K]$.
3. Simultaneously, an adversary (a.k.a. the environment) picks a loss vector $\ell_t \in [0,1]^K$, where $\ell_t(i)$ is the loss for expert i.
4. The learner incurs losses $\ell_t(A_t)$.
5. The learner observes *only* $\ell_t(A_t)$ (Bandit feedback).

The regret $R_T(\pi, \ell)$ is defined exactly the same as the expert problem:

$$R'_T(\pi, \ell) = \sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

and

$$R_T(\pi, \ell) = \mathbb{E}[R'_T(\pi, \ell)].$$

As before, we are interested in minimizing $\sup_\ell R_T(\pi, \ell)$.

We summarize the settings we have seen with the following table.

	Full Information Feedback	**Bandit Feedback**
Stochastic	Trivial	Stochastic bandits
Adversarial	Experts problem	Adversarial bandits

Note that we will not analyze the stochastic, full information setting because it is easy to see that FTL is optimal. In the next lecture, we will analyze the adversarial bandits setting and present the EXP3 algorithm.

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Xinyan Wang, Zhifeng Chen, Congwei Yang, and Bo-Hsun Chen.