CS 6840 Algorithmic Game Theory October 21st, 2024

Lecture 23: Learning with Partial Feedback

Instructor: Eva Tardos *Scribe: Shaleen Baral*

Throughout these notes we require that our utility be bounded within the unit interval, i.e. $u^\tau(s) \in [0,1]$. Furthermore, the number of actions are assumed to be a finite value $K < \infty$.

1 Introduction

In the previous lecture, we considered the Follow the Perturbed Leader algorithm that chooses a strategy at time t by using the following quantities,

$$\sum_{\tau=1}^{t-1} u^\tau(s), \qquad \forall s \in S.$$

Even the Hedge algorithm needs knowledge of the utility vector at all prior time steps to update its internal probability distribution over strategies. In this sense, both these algorithms assume a setting where we have full information about $u^\tau(\cdot)$ at all time steps, $\tau \in [t-1]$.

It is also natural to consider a situation where at every time step τ, we are provided the bare minimum—when we choose a strategy s^τ, we are only revealed $u^\tau(s^\tau)$ as opposed to the entire utility vector, u^τ. Is it still possible to get a good bound on regret in this context of partial feedback?

2 Propensity Scoring

At every time step τ, only $u^\tau(s^\tau)$ is revealed where s^τ denotes the strategy chosen. A modest guess for what the utility vector looks like at time step τ could be,

$$\tilde{u}^\tau(s) = \begin{cases} u^t(s^t) & \text{if } s = s^\tau, \\ 0 & \text{otherwise.} \end{cases}$$

If ΔS denotes the distribution from which we draw our strategy at time step τ, we have for any $s \in S$,

$$\mathbb{E}_{\Delta S}\left[\tilde{u}^\tau(s)\right] = \mathbb{P}_{\Delta S}\left(s^\tau = s\right) \cdot u^t\left(s\right).$$

For our purposes, it will be more convenient to work with a rescaling of $\tilde{u}^\tau$ as follows,

Definition 1 (Propensity Scoring)**.**

$$\overline{u}^\tau(s) = \begin{cases} \frac{u^t(s^t)}{\mathbb{P}_{\Delta S}(s^t)} & \text{if } s = s^\tau, \\ 0 & \text{otherwise.} \end{cases}$$

Specifically, this definition implies that, at the very least, the vector $\overline{u}^\tau$ is an *unbiased estimator* of u^τ, in the sense that, for any $s \in S$,

$$\mathbb{E}_{\Delta S}\left[\overline{u}^\tau(s)\right] = \frac{\mathbb{E}_{\Delta S}\left[\tilde{u}^\tau\left(s\right)\right]}{\mathbb{P}_{\Delta S}\left(s^\tau\right)} = u^\tau(s^\tau).$$

What happens when we try to use a learning algorithm with this estimator? Note that computing $\overline{u}^\tau$ itself may be challenging if we don't know $\mathbb{P}(s^\tau)$. In fact, a major aspect of Follow the Perturbed Leader was that we avoided having to explicitly calculate these probabilities. This isn't a big problem for us as the first learning algorithm we encountered, Hedge, internally maintains a probability distribution to make its decisions!

3 Multiplicative Weights

We now consider running Hedge using our estimator $\overline{u}^\tau$. We have a simple recipe for performing regret analysis:

a. Imagine that $\overline{u}^\tau$ corresponds to the true utility and analyze the regret,

$$\frac{1}{T}\left[\sum_{\tau=1}^{T}\overline{u}^\tau(s^\tau) - \min_{s\in S}\sum_{\tau=1}^{T}\overline{u}^\tau(s)\right]$$

where $\{s^\tau\}$ denotes the strategy chosen by Hedge.

b. Combine part (a) with the fact that $\mathbb{E}_{\Delta S}\left[\overline{u}^\tau\right] = u^\tau$ to compute the true regret.

The first step appears to be simple since we have already analyzed the regret for Hedge. However, our analysis requires that the utility lie within $[0,1]$. As $\mathbb{P}(s^\tau) > 0$ can be arbitrarily small, $\overline{u}^\tau(s) \in [0,\infty)$. We fix this problem by modifying our algorithm as follows, for some fixed $\delta > 0$:

Algorithm 1 $(1-\delta)$-Hedge

```
for every time step $\tau$ do
    choose, with probability $\delta$, a strategy $s \in S$, uniformly at random,
    or, choose, with probability $(1-\delta)$, the strategy $s^\tau$ recommended by Hedge.
end for
```

This guarantees that $\mathbb{P}(s^\tau) \geq \delta \cdot 1/K$, where $K = |S|$ denotes the total number of strategies. Since $u^\tau \in [0,1]^K$,

$$\overline{u}^\tau(s) \in \left[0, \frac{1}{\mathbb{P}(s^\tau)}\right] \subseteq \left[0, \frac{K}{\delta}\right].$$

We can interpret δ as being a parameter tuning how much our algorithm 'explores' the strategy space or 'exploit' strategies that we already know about. No matter what we have learned, there is a constant probability that we try out an arbitrary strategy. This is in start contrast to what Hedge would normally do wherein the probability of choosing an action that yields low utility may diminish arbitrarily.

Now that we have bounded $\overline{u}$ appropriately, we may use our results from previous lectures. Particularly, we can just scale our utility vector to be

$$\frac{\delta}{K} \cdot \overline{u}^\tau \in [0,1]^K.$$

We then run the above algorithm and then afterwards rescale our obtained utilities by K/δ to recover values in the original, intended range. The upshot is that we can borrow our analysis of Hedge, only paying the cost of a K/δ-factor on the regret bound we had derived.

Let $\{\overline{s}^\tau\}$ denote the sequence of strategies played by $(1-\delta)$-Hedge and $\{s^\tau\}$ denote the sequence of strategies recommended by the Hedge subroutine. Explicitly,

$$\mathop{\mathbb{E}}_{\substack{\delta,\\ \text{Hedge}}}\left[\sum_{\tau=1}^{T}\overline{u}^\tau(\overline{s}^\tau)\right] \geq (1-\delta)\mathop{\mathbb{E}}_{\text{Hedge}}\left[\sum_{\tau=1}^{T}\overline{u}^\tau(s^\tau)\right]$$

For any $s \in S$, we then have

$$\geq (1-\delta)\left[\sum_{\tau=1}^{T}\overline{u}^\tau(s)\right] - (1-\delta)\frac{K}{\delta}O(\sqrt{T\ln K}).$$

Finally, we use the fact that $\mathbb{E}_{\Delta S}[\overline{u}] = u$ as folows,

$$\mathbb{E}_{\Delta S}\left[\mathop{\mathbb{E}}_{\substack{\delta,\\ \text{Hedge}}}\left[\sum_{\tau=1}^{T}\overline{u}^\tau(\overline{s}^\tau)\right]\right] \geq (1-\delta)\mathbb{E}_{\Delta S}\left[\sum_{\tau=1}^{T}\overline{u}^\tau(s)\right] - (1-\delta)\frac{K}{\delta}O(\sqrt{T\ln K})$$
$$\implies \mathop{\mathbb{E}}_{\substack{\delta,\\ \text{Hedge}}}\left[\sum_{\tau=1}^{T}u^\tau(\overline{s}^\tau)\right] \geq (1-\delta)\left[\sum_{\tau=1}^{T}u^\tau(s)\right] - (1-\delta)\frac{K}{\delta}O(\sqrt{T\ln K})$$

From here, it is merely a manner of rearrangement and choosing the appropriate δ to obtain a nice regret bound as we had done before. For completion, we include the next section with attempts at computing this more precisely.

Determining the Regret Bound

Note that $O(\sqrt{T\ln K})$ factor above is quite explicitly $2\sqrt{T\ln K}$. Rearranging what we obtained prior and bounding the utility of the best-fixed response by T gives us,

$$\mathcal{R}^{(1-\delta)\text{-Hedge}}(T) \leq 2(1-\delta)\frac{K}{\delta}\sqrt{T\ln K} + \delta T \leq 2\frac{K}{\delta}\sqrt{T\ln K} + \delta T.$$

For $T > 4K^2\ln K$, this gives us a bound of $\mathcal{R} \leq 4\sqrt{2}K^{\frac{1}{2}}(\ln K)^{\frac{1}{4}}T^{\frac{3}{4}}$ when we take $\delta = \sqrt{2}K^{\frac{1}{2}}(\ln K)^{\frac{1}{4}}T^{-\frac{1}{4}} < 1$.