Lecture 4

Instructor: Haipeng Luo

In the following lectures, we focus on the expert problem and study more *adaptive* algorithms. Although Hedge is proven to be worst-case optimal, one may wonder how well it would actually perform when dealing with a practical problem that is probably not the worst case or even relatively easy. Indeed, the regret bound we proved for Hedge only says that for all problem instances, Hedge's regret is uniformly bounded by $\mathcal{O}(\sqrt{T \ln N})$. However, ideally we want to have an algorithm that enjoys a much smaller regret in many easy situations, but in the worst case still guarantees the minimax regret $\mathcal{O}(\sqrt{T \ln N})$. Deriving adaptive algorithms and adaptive regret bounds is exactly one way to achieve this goal.

1 "Small-loss" Bounds

We start with the arguably simplest adaptive bound, sometimes called "small-loss" bound or first order bound. Recall that we proved the following intermediate bound for Hedge:

$$\mathcal{R}_T = \widetilde{L}_T - L_T(i^\star) \le \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{N} p_t(i) \ell_t^2(i),$$

where L_T is the cumulative loss vector, $i^\star$ is the best expert and we define $\widetilde{L}_T = \sum_{t=1}^{T} \langle p_t, \ell_t \rangle$ to be the cumulative loss of the algorithm. By boundedness of losses the last term above can be bounded by $\eta \widetilde{L}_T$. If $\eta < 1$, then rearranging gives

$$\mathcal{R}_T \le \frac{1}{1-\eta} \left(\frac{\ln N}{\eta} + \eta L_T(i^\star) \right).$$

Therefore, if for a moment we assume we knew the quantity $L_T(i^\star)$ ahead of time and was able to set $\eta = \min\left\{ \frac{1}{2}, \sqrt{\frac{\ln N}{L_T(i^\star)}} \right\}$, then we arrive at

$$\mathcal{R}_T \le 2 \left(\max\left\{ 2\ln N, \frac{\ln N}{\sqrt{(\ln N)/L_T(i^\star)}} \right\} + \sqrt{\frac{\ln N}{L_T(i^\star)}} L_T(i^\star) \right) = \mathcal{O}\left(\sqrt{L_T(i^\star) \ln N} + \ln N \right).$$

The final bound above is the so-called "small-loss" bound, which essentially replaces the dependence on T in the minimax bound $\sqrt{T \ln N}$ by the loss of the best expert $L_T(i^\star)$. Note that $L_T(i^\star)$ is bounded by T, therefore the "small-loss" bound is not worse than the minimax bound. More importantly, it can be much smaller than T when the best expert indeed suffers very small loss. In particular, if the best expert makes no mistakes at all and have $L_T(i^\star) = 0$, then the "small-loss" bound is only $\mathcal{O}(\ln N)$, independent of T. This is one typical example of adaptive bounds that we are aiming for.

Of course, one obvious issue in the above derivation is that the learning rate has to be set in terms of the unknown quantity $L_T(i^\star)$. In fact, this becomes an even more severe problem in a non-oblivious environment since $L_T(i^\star)$ can depend on the algorithm's actions and thus η, making the definition of η circular.

Fortunately, there are many different ways to address this issue, and we explore one of them here. The idea is to use a more adaptive and time-varying learning rate schedule. Specifically, the algorithm predicts $p_t(i) \propto \exp(-\eta_t L_{t-1}(i))$ where

$$\eta_t = \sqrt{\frac{\ln N}{\widetilde{L}_{t-1} + 1}}. \tag{1}$$

Note that $\widetilde{L}_{t-1} = \sum_{\tau=1}^{t-1} \langle p_\tau, \ell_\tau \rangle$ is the cumulative loss of the algorithm up to round $t-1$ and is thus available at the beginning of round t. This is sometimes called a "self-confident" learning rate since the algorithm is confident that its loss is close to the loss of the best expert and thus uses it as a proxy for the loss of the best expert to tune the learning rate. We next prove that this algorithm indeed provides a "small-loss" bound.

Theorem 1. *Hedge with adaptive learning rate schedule* (1) *ensures*

$$\mathcal{R}_T \le 3\sqrt{(L_T(i^\star) + 1)\ln N} + 9 \ln N.$$

Proof. Let $\Phi_t(\eta) = \frac{1}{\eta} \ln\left(\frac{1}{N}\sum_{i=1}^N \exp(-\eta L_t(i))\right)$. In Lecture 2 we already proved

$$\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t) \le -\langle p_t, \ell_t \rangle + \eta_t \sum_{i=1}^N p_t(i)\ell_t^2(i).$$

Summing over t and rearranging give

$$\begin{aligned}
\widetilde{L}_T &\le \Phi_0(\eta_1) - \Phi_T(\eta_T) + \sum_{t=1}^T \eta_t \sum_{i=1}^N p_t(i)\ell_t^2(i) + \sum_{t=1}^{T-1} \left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right) \\
&\le \frac{\ln N}{\eta_T} - \frac{1}{\eta_T} \ln\left(\exp(-\eta_T L_T(i^\star))\right) + \sum_{t=1}^T \eta_t \sum_{i=1}^N p_t(i)\ell_t(i) + \sum_{t=1}^{T-1} \left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right) \\
&= \sqrt{(\widetilde{L}_{T-1} + 1)\ln N} + L_T(i^\star) + \sum_{t=1}^T \eta_t \langle p_t, \ell_t \rangle + \sum_{t=1}^{T-1} \left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right).
\end{aligned}$$

To bound the term $\sum_{t=1}^T \eta_t \langle p_t, \ell_t \rangle$, note that

$$\begin{aligned}
\sum_{t=1}^T \frac{\langle p_t, \ell_t \rangle}{\sqrt{\widetilde{L}_{t-1} + 1}} &= \sum_{t=1}^T \frac{\widetilde{L}_t - \widetilde{L}_{t-1}}{\sqrt{\widetilde{L}_{t-1} + 1}} \\
&= \sum_{t=1}^T \frac{\widetilde{L}_t - \widetilde{L}_{t-1}}{\sqrt{\widetilde{L}_t + 1}} + \sum_{t=1}^T (\widetilde{L}_t - \widetilde{L}_{t-1}) \left(\frac{1}{\sqrt{\widetilde{L}_{t-1} + 1}} - \frac{1}{\sqrt{\widetilde{L}_t + 1}}\right) \\
&\le \sum_{t=1}^T \frac{\widetilde{L}_t - \widetilde{L}_{t-1}}{\sqrt{\widetilde{L}_t + 1}} + \sum_{t=1}^T \left(\frac{1}{\sqrt{\widetilde{L}_{t-1} + 1}} - \frac{1}{\sqrt{\widetilde{L}_t + 1}}\right) \qquad (\widetilde{L}_t - \widetilde{L}_{t-1} \le 1) \\
&\le 1 + \sum_{t=1}^T \frac{\widetilde{L}_t - \widetilde{L}_{t-1}}{\sqrt{\widetilde{L}_t + 1}} \\
&\le 1 + \int_{\widetilde{L}_0}^{\widetilde{L}_T} \frac{dx}{\sqrt{x+1}} \\
&\le 2\sqrt{\widetilde{L}_T + 1},
\end{aligned}$$

and thus $\sum_{t=1}^T \eta_t \langle p_t, \ell_t \rangle \le 2\sqrt{(\widetilde{L}_T + 1)\ln N}$.

To bound $\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)$, we prove that $\Phi_t(\eta)$ in increasing in η and thus $\Phi_t(\eta_{t+1}) \le \Phi_t(\eta_t)$. It suffices to prove that the derivative is non-negative. Indeed, direct calculation shows that with

$p^\eta_{t+1}(i) \propto \exp(-\eta L_t(i))$,

$$\begin{aligned}
\eta^2 \Phi_t'(\eta) &= \eta^2 \left(-\frac{1}{\eta^2} \ln \left(\frac{1}{N} \sum_{i=1}^N \exp(-\eta L_t(i)) \right) - \frac{1}{\eta} \frac{\sum_{i=1}^N L_t(i) \exp(-\eta L_t(i))}{\sum_{i=1}^N \exp(-\eta L_t(i))} \right) \\
&= \ln N - \sum_{i=1}^N p^\eta_{t+1}(i) \left(\ln \left(\sum_{j=1}^N \exp(-\eta L_t(j)) \right) + \eta L_t(i) \right) \\
&= \ln N - \sum_{i=1}^N p^\eta_{t+1}(i) \ln \left(\frac{\sum_{j=1}^N \exp(-\eta L_t(j))}{\exp(-\eta L_t(i))} \right) \\
&= \ln N - \sum_{i=1}^N p^\eta_{t+1}(i) \ln \frac{1}{p^\eta_{t+1}(i)} \geq 0,
\end{aligned}$$

where the last step is by the fact that entropy is maximized by the uniform distribution. To sum up, we have proven that

$$\mathcal{R}_T = \widetilde{L}_T - L_T(i^\star) \leq 3\sqrt{(\widetilde{L}_T + 1)\ln N}.$$

Solving for $\sqrt{\widetilde{L}_T + 1}$ leads to

$$\sqrt{\widetilde{L}_T + 1} \leq \frac{3}{2}\sqrt{\ln N} + \sqrt{L_T(i^\star) + 1 + \frac{9}{4}\ln N}.$$

Finally squaring both sides and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ give

$$\widetilde{L}_T \leq 9 \ln N + L_T(i^\star) + 3\sqrt{(L_T(i^\star) + 1)\ln N},$$

which completes the proof. □

Besides enjoying a better theoretical regret bound, this algorithm is also intuitively more reasonable since it tunes the learning rate adaptively based on observed data. In general, learning rate tuning is an important topic in machine learning and could be of great practical importance.

2 Quantile Bounds

"Small-loss" bounds improve the dependence on T in the minimax regret bound to $L_T(i^\star)$. Is it possible to improve the other term $\ln N$ in the minimax bound to something better? To answer this question, consider again Hedge with a fixed learning rate for simplicity, and note that we proved in Lecture 2,

$$\widetilde{L}_T \leq \frac{\ln N}{\eta} - \frac{1}{\eta} \ln \left(\sum_{i=1}^N \exp(-\eta L_T(i)) \right) + T\eta.$$

Without loss of generality, assume $L_T(1) \leq \cdots \leq L_T(N)$ so that expert i is the i-th best expert. Previously we obtained the final regret bound by lower bounding $\sum_{i=1}^N \exp(-\eta L_T(i))$ by $\max_i \exp(-\eta L_T(i)) = \exp(-\eta L_T(1))$. In general, however, for each i we have

$$\sum_{j=1}^N \exp(-\eta L_T(j)) \geq \sum_{j=1}^i \exp(-\eta L_T(j)) \geq i \exp(-\eta L_T(i)),$$

and we therefore have the following regret bound against the i-th best expert:

$$\widetilde{L}_T - L_T(i) \leq \frac{\ln\left(\frac{N}{i}\right)}{\eta} + T\eta. \tag{2}$$

With η optimally tuned to $\sqrt{\ln\left(\frac{N}{i}\right)/T}$, the bound becomes $2\sqrt{T \ln\left(\frac{N}{i}\right)}$. This is called the quantile bound and it states that the algorithm suffers at most this amount of regret for all but i/N fraction of the experts. Of course, at the end of the day what we care about is actually the loss of the algorithm. So assuming we had the knowledge of L_T for a moment, then we could pick the optimal η to achieve

$$\widetilde{L}_T \leq \min_{i\in[N]} \left(L_T(i) + 2\sqrt{T \ln\left(\frac{N}{i}\right)} \right), \tag{3}$$

which is a strictly better bound compared to $L_T(1) + 2\sqrt{T\ln N}$. To understand the improvement, consider the case when N is huge but there are many similar experts so that for example the top 1% of them all have the same cumulative loss. Then bound (3) is at most

$$L_T(1\% \times N) + 2\sqrt{T\ln\left(\frac{N}{1\%\times N}\right)} = L_T(1) + 2\sqrt{T\ln(100)},$$

which is independent of N.

Just as in the previous discussion, one obvious issue in the derivation of bound (3) above is again that the learning rate η needs to be tuned based on unknown knowledge. To address the issue, here we explore a quite different approach. The idea is to have different instances of Hedge running with different learning rates, and have a "master" Hedge to combine the predictions of these "meta-experts". To this end, we use Hedge(η) to denote an instance of Hedge running with learning rate η. The algorithm is shown below.

Algorithm 1: Hedge with Quantile Bounds

Input: master learning rate $\eta > 0$, base learning rates $\eta_1, \ldots, \eta_M$
Initialize: M Hedge algorithms Hedge(η_1), ..., Hedge(η_M), $C_0(j) = 0$ for all $j \in [M]$
for $t = 1, \ldots, T$ **do**
- let p_t^j be the prediction of Hedge(η_j) on round t
- compute $p_t = \sum_{j=1}^M q_t(j) p_t^j$ where $q_t(j) \propto \exp(-\eta C_{t-1}(j))$
- play p_t and observe loss vector $\ell_t \in [0,1]^N$
- update $C_t(j) = C_{t-1}(j) + \left\langle p_t^j, \ell_t \right\rangle$ for all $j \in [M]$
- pass ℓ_t to Hedge(η_1), ..., Hedge(η_M).

By Eq. (2), we have for each Hedge(η_j) and each expert i

$$\sum_{t=1}^T \left\langle p_t^j, \ell_t \right\rangle - L_T(i) \leq \frac{\ln\left(\frac{N}{i}\right)}{\eta_j} + T\eta_j.$$

On the other hand, for the master Hedge, we have for each meta-expert j,

$$\sum_{t=1}^T \sum_{j=1}^M q_t(j) \left\langle p_t^j, \ell_t \right\rangle - C_T(j) \leq \frac{\ln M}{\eta} + T\eta.$$

Note that by construction, we have $\sum_{j=1}^M q_t(j)\left\langle p_t^j, \ell_t\right\rangle = \langle p_t, \ell_t\rangle$ and $C_T(j) = \sum_{t=1}^T \left\langle p_t^j, \ell_t \right\rangle$. Therefore summing up the above two inequalities lead to

$$\sum_{t=1}^T \langle p_t, \ell_t\rangle - L_T(i) \leq \frac{\ln\left(\frac{N}{i}\right)}{\eta_j} + T\eta_j + \frac{\ln M}{\eta} + T\eta = \frac{\ln\left(\frac{N}{i}\right)}{\eta_j} + T\eta_j + 2\sqrt{T\ln M},$$

where the last step is by picking the optimal $\eta = \sqrt{\ln M/T}$. Note that the above holds for all j and all i. Therefore, suppose we have (a) for each i, there is an η_j such that $\frac{1}{\eta_j}\ln\left(\frac{N}{i}\right) + T\eta_j = \mathcal{O}\left(\sqrt{T\ln\left(\frac{N}{i}\right)}\right)$, and (b) M is much smaller than N, then we obtain bound (3).

Setting $M = N$ and $\eta_j = \sqrt{\ln(\frac{N}{j})/T}$ would clearly satisfy (a), but not (b). Fortunately, it turns out that one only needs to create $M \approx \ln N$ meta-experts and still satisfy (a). Specifically, let

$$\eta_j = \frac{1}{2^{j-1}}\sqrt{\frac{\ln N}{T}} \quad \text{and} \quad M = \left\lceil \log_2 \sqrt{\frac{\ln N}{\ln(\frac{N}{N-1})}} \right\rceil + 1.$$

Now clearly for each $i \neq N$, there exist a j such that $\eta_j \leq \sqrt{\ln\left(\frac{N}{i}\right)/T} \leq 2\eta_j$ and therefore

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t, \ell_t \rangle - L_T(i) &\leq \frac{\ln\left(\frac{N}{i}\right)}{\eta_j} + T\eta_j + 2\sqrt{T \ln M} \\
&\leq \frac{\ln\left(\frac{N}{i}\right)}{\frac{1}{2}\sqrt{\ln\left(\frac{N}{i}\right)/T}} + T\sqrt{\ln\left(\frac{N}{i}\right)/T} + 2\sqrt{T \ln M} \\
&= 3\sqrt{T \ln\left(\frac{N}{i}\right)} + 2\sqrt{T \ln M}.
\end{aligned}$$

It remains to show that M is small enough. Indeed, since $\ln(1+x) \geq x/2,\ \forall x \leq 1$, we have

$$\ln\left(\frac{N}{N-1}\right) = \ln\left(1 + \frac{1}{N-1}\right) \geq \frac{1}{2(N-1)},$$

and therefore $M = \mathcal{O}(\ln(N \ln N))$. So as least for the case when N/i is larger than $\mathcal{O}(\ln(N \ln N))$, the term $\sqrt{T \ln M}$ is dominated by $\sqrt{T \ln\left(\frac{N}{i}\right)}$ in the regret bound. We summarize the result in the following theorem.

Theorem 2. *Algorithm 1 with* $\eta = \sqrt{\frac{\ln N}{T}}$, $\eta_j = \frac{1}{2^{j-1}}\sqrt{\frac{\ln N}{T}}$ *and* $M = \left\lceil \log_2 \sqrt{\frac{\ln N}{\ln(\frac{N}{N-1})}} \right\rceil + 1$ *ensures*

$$\widetilde{L}_T \leq \min_{i \neq N} \left(L_T(i) + 3\sqrt{T \ln\left(\frac{N}{i}\right)} \right) + \mathcal{O}(\sqrt{T \ln(\ln(N \ln N))}).$$

This idea of combining algorithms using Hedge is useful for many other problems. It is usually a quick and easy way to verify whether some regret bound is possible or not in theory. However, the resulting algorithm might not be so elegant and practical. In the next lecture, we will study a different algorithm that not only guarantees a quantile bound (in fact even better than the one proven here), but also enjoys several more useful properties.