Lecture 5

Instructor: Haipeng Luo

1 Second Order Bounds and Squint

In this lecture we study one of the state-of-the-art algorithms for the expert problem, called Squint [Koolen and Van Erven, 2015], which enjoys many nice properties simultaneously.

To introduce the algorithm, recall that Hedge predicts $p_t(i) \propto \exp(-\eta L_{t-1}(i))$. Denote $r_t(i) = \langle p_t, \ell_t \rangle - \ell_t(i)$ and $R_t(i) = \sum_{\tau=1}^t r_\tau(i)$ to be the instantaneous regret and cumulative regret to expert i respectively. Hedge can then be equivalently written as $p_t(i) \propto \exp(\eta R_{t-1}(i))$. Now the first idea of Squint is to introduce a second order "correction term" in the exponent:

$$p_t(i) \propto \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)$$

where $V_t(i) = \sum_{\tau=1}^t r_\tau(i)^2$ is the cumulative square of the regret. In other words, the strategy is putting more weights on experts whose loss is closer to the algorithm's loss. This can also be seen as some kind of "self-confidence" since the algorithm is using its loss as a benchmark to evaluate the experts.

In fact, one can make the algorithm even more general by allowing some prior knowledge of the problem. This can be simply done by letting p_1 be the user's prior distribution over the experts (instead of a uniform distribution), and for $t > 1$ predicts

$$p_t(i) \propto p_1(i) \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right). \tag{1}$$

The analysis of this algorithm is also straightforward. Let $\Phi_t = \mathbb{E}_{i\sim p_1}\left[\exp\left(\eta R_t(i) - \eta^2 V_t(i)\right)\right]$ be the potential. If $\eta \le 1/2$, we have

$$\begin{aligned}
&\Phi_t - \Phi_{t-1} \\
&= \mathbb{E}_{i\sim p_1}\left[\exp\left(\eta R_t(i) - \eta^2 V_t(i)\right) - \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)\right] \\
&= \mathbb{E}_{i\sim p_1}\left[\exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)\left(\exp\left(\eta r_t(i) - \eta^2 r_t^2(i)\right) - 1\right)\right] \\
&\le \eta \mathbb{E}_{i\sim p_1}\left[\exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right) r_t(i)\right] \qquad\qquad (e^{x-x^2} \le 1 + x,\ \forall x \ge -\tfrac{1}{2}) \\
&= \eta \sum_{i=1}^N \left(p_1(i) \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)\right) r_t(i) \\
&= 0
\end{aligned}$$

where the last equality is by the fact that for any $a \in \mathbb{R}_+^N$, if $p_t(i) \propto a(i)$ then for any loss vector ℓ_t,

$$\sum_{i=1}^N a(i) r_t(i) = \sum_{i=1}^N a(i)(\langle p_t, \ell_t \rangle - \ell_t(i)) = \left(\sum_{i=1}^N a(i)\right) \sum_{i=1}^N \frac{a(i)}{\sum_{j=1}^N a(j)} \ell_t(i) - \sum_{i=1}^N a(i)\ell_t(i) = 0.$$

Therefore, the potential is non-increasing and we have

$$\Phi_T \le \Phi_{T-1} \le \cdots \Phi_0 = 1.$$

On the other hand, by the definition of Φ_T, we have for any i,

$$\Phi_T \ge p_1(i) \exp\left(\eta R_T(i) - \eta^2 V_T(i)\right).$$

Solving for $R_T(i)$ then leads to

$$R_T(i) \le \frac{\ln(1/p_1(i))}{\eta} + \eta V_T(i),$$

which is again the bound we have seen for Hedge if one sets p_1 to be uniform and upper bounds $V_T(i)$ by T. However, $V_T(i)$ can be much smaller than T and we will discuss more in the next section. For now, let's see how one can in fact obtain an even more general bound that competes with not just a single expert, but an arbitrary distribution over the experts. Indeed, for any distribution $q \in \Delta(N)$ that we want to compete with, we have

$$\begin{aligned}
1 \ge \Phi_T &\ge \mathbb{E}_{i\sim q}\left[\frac{p_1(i)}{q(i)} \exp\left(\eta R_T(i) - \eta^2 V_T(i)\right)\right] \\
&= \mathbb{E}_{i\sim q}\left[\exp\left(\ln\left(\frac{p_1(i)}{q(i)}\right) + \eta R_T(i) - \eta^2 V_T(i)\right)\right] \\
&\ge \exp\left(\mathbb{E}_{i\sim q}\left[\ln\left(\frac{p_1(i)}{q(i)}\right) + \eta R_T(i) - \eta^2 V_T(i)\right]\right) && \text{(Jensen's inequality)} \\
&= \exp\left(-\mathrm{KL}(q, p_1) + \eta \mathbb{E}_{i\sim q}[R_T(i)] - \eta^2 \mathbb{E}_{i\sim q}[V_T(i)]\right) \\
&= \exp\left(-\mathrm{KL}(q, p_1) + \frac{\left(\mathbb{E}_{i\sim q}[R_T(i)]\right)^2}{4\mathbb{E}_{i\sim q}[V_T(i)]}\right)
\end{aligned}$$

where the last step is by using the optimal $\eta = \frac{\mathbb{E}_{i\sim q}[R_T(i)]}{2\mathbb{E}_{i\sim q}[V_T(i)]}$ (for simplicity assume it is in $[0, 1/2]$). Solving for $\mathbb{E}_{i\sim q}[R_T(i)]$ gives

$$\mathbb{E}_{i\sim q}[R_T(i)] \le 2\sqrt{\mathbb{E}_{i\sim q}[V_T(i)]\mathrm{KL}(q, p_1)}.$$

Note that $\mathbb{E}_{i\sim q}[R_T(i)] = \widetilde{L}_T - \langle q, L_T \rangle$ is exactly the difference between the algorithm's total loss and the total loss of a fixed strategy q. The KL divergence term implies that the closer the prior knowledge p_1 is from the competitor q, the smaller the regret becomes. Before exploring the many implications of this bound, let's first discuss how to address the learning rate tuning issue again.

Squint uses a quite different technique to deal with this issue compared to what we have seen in the last lecture. The idea is to put a prior on the learning rate $\eta \in [0, 1/2]$, which resembles a Bayesian approach. The hope is that for every possible optimal tuning of η, there is a sufficient mass around it in the prior. To state the algorithm, it is in fact easier to first look at the analysis. Specifically, let γ be a prior distribution on η to be specified later and re-define $\Phi_t = \mathbb{E}_{i\sim p_1, \eta\sim\gamma}\left[\exp\left(\eta R_t(i) - \eta^2 V_t(i)\right)\right]$. By the exact same argument as before, we have

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &\le \mathbb{E}_{i\sim p_1, \eta\sim\gamma}\left[\eta \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right) r_t(i)\right] \\
&= \sum_{i=1}^N \left(p_1(i) \mathbb{E}_{\eta\sim\gamma}\left[\eta \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)\right]\right) r_t(i)
\end{aligned}$$

and if we again want the last term to be zero, we need to set

$$p_t(i) \propto p_1(i) \mathbb{E}_{\eta\sim\gamma}\left[\eta \exp\left(\eta R_{t-1}(i) - \eta^2 V_{t-1}(i)\right)\right], \tag{2}$$

which defines the algorithm. Notice the extra η in the formula compared to Eq. (1) where η was a constant and could be dropped. Now for any distribution q that we want to compete with, let $\eta_\star = \frac{\mathbb{E}_{i\sim q}[R_T(i)]}{2\mathbb{E}_{i\sim q}[V_T(i)]}$ be the optimal tuning. For simplicity, assume again $\eta_\star$ is in $[0, 1/2]$ and has $\gamma(\eta_\star)$ mass in the prior. Then we have by a similar argument as before

$$\begin{aligned}
1 \ge \Phi_T &\ge \gamma(\eta_\star) \mathbb{E}_{i\sim p_1}\left[\exp\left(\eta_\star R_T(i) - \eta_\star^2 V_T(i)\right)\right] \\
&\ge \gamma(\eta_\star) \exp\left(-\mathrm{KL}(q, p_1) + \eta_\star \mathbb{E}_{i\sim q}[R_T(i)] - \eta_\star^2 \mathbb{E}_{i\sim q}[V_T(i)]\right) \\
&= \gamma(\eta_\star) \exp\left(-\mathrm{KL}(q, p_1) + \frac{\left(\mathbb{E}_{i\sim q}[R_T(i)]\right)^2}{4\mathbb{E}_{i\sim q}[V_T(i)]}\right)
\end{aligned}$$

Solving for $\mathbb{E}_{i\sim q}[R_T(i)]$ gives

$$\mathbb{E}_{i\sim q}[R_T(i)] \leq 2\sqrt{\mathbb{E}_{i\sim q}[V_T(i)]\left(\mathrm{KL}(q,p_1)+\ln\left(\frac{1}{\gamma(\eta_\star)}\right)\right)}.$$

Therefore, if $\gamma(\eta_\star)$ is large enough, we essentially obtain the bound that we aim for with a parameter-free algorithm. The only technical difficulty now is that it is impossible to ensure sufficient mass for every η in $[0, 1/2]$. Instead, we should pick a prior so that for each q, the set of *approximately* optimal η has a sufficient mass, or in other words, every η has a sufficient mass around it.

It turns out that picking $\gamma(\eta) \approx 1/\eta$ will do the job and leads to a bound that essentially replaces $\ln(1/\gamma(\eta_\star))$ by $\ln\ln T$, which is a very small overhead. We refer the interested reader to [Koolen and Van Erven, 2015] for details on the exact prior that Squint uses. Importantly, this prior leads to a closed form for the update rule (2) and the algorithm can be efficiently implemented!

2 Implications

In this section we discuss why the Squint bound is interesting and useful. For simplicity we ignore constants and negligible terms and assume we have an algorithm that achieves for any $q \in \Delta(N)$,

$$\mathbb{E}_{i\sim q}[R_T(i)] \leq \sqrt{\mathbb{E}_{i\sim q}[V_T(i)]\mathrm{KL}(q,p_1)}. \tag{3}$$

Below we prove that this bound *simultaneously* implies three adaptive regret bounds.

2.1 Bound (3) implies "small-loss" bounds

Theorem 1. *Bound* (3) *implies*

$$\mathbb{E}_{i\sim q}[R_T(i)] \leq \sqrt{2\mathbb{E}_{i\sim q}\left[\sum_{t=1}^{T}\max\{\ell_t(i)-\langle p_t,\ell_t\rangle,0\}\right]\mathrm{KL}(q,p_1)}+\mathrm{KL}(q,p_1) \tag{4}$$

$$\leq \sqrt{2\langle q, L_T\rangle\,\mathrm{KL}(q,p_1)}+\mathrm{KL}(q,p_1) \tag{5}$$

Proof. Note that since $r_t(i) = \langle p_t, \ell_t\rangle - \ell_t(i) \in [-1, 1]$, we have

$$\begin{aligned}\mathbb{E}_{i\sim q}[V_T(i)] &\leq \mathbb{E}_{i\sim q}\left[\sum_{t=1}^{T}|r_t(i)|\right] = \mathbb{E}_{i\sim q}\left[\sum_{t=1}^{T}(r_t(i)+2\max\{-r_t(i),0\})\right]\\ &= \mathbb{E}_{i\sim q}[R_T(i)] + 2\mathbb{E}_{i\sim q}\left[\sum_{t=1}^{T}\max\{\ell_t(i)-\langle p_t,\ell_t\rangle,0\}\right].\end{aligned}$$

Plugging the above inequality into Eq. (3) and solving for $\mathbb{E}_{i\sim q}[R_T(i)]$ prove Eq. (4). Eq. (5) is simply by the fact $\max\{\ell_t(i)-\langle p_t,\ell_t\rangle,0\} \leq \ell_t(i)$. □

In particular, if one simply sets p_1 to be uniform and q to be the distribution that concentrates on the best expert $i^\star$, then Eq. (5) becomes the "small-loss" bound we saw last time:

$$R_T(i^\star) \leq \sqrt{2L_T(i^\star)\ln N} + \ln N.$$

In fact, one can obtain an even tighter bound using Eq. (4) instead.

2.2 Bound (3) implies quantile bounds

Theorem 2. *Assume $L_T(1) \leq \cdots \leq L_T(N)$ without loss of generality. With p_1 being the uniform distribution Bound* (3) *implies*

$$\widetilde{L}_T \leq \min_{i\in[N]}\left(\ell_T(i) + \sqrt{2\ell_T(i)\ln\left(\frac{N}{i}\right)} + \ln\left(\frac{N}{i}\right)\right).$$

Proof. For any $i \in [N]$, setting $q(j) = 1/i$ for all $j \le i$ and $q(j) = 0$ for all $j > i$ and using Eq. (5) give

$$\widetilde{L}_T \le \frac{1}{i}\sum_{j=1}^{i} \ell_T(j) + \sqrt{2\left(\frac{1}{i}\sum_{j=1}^{i} \ell_T(j)\right)\ln\left(\frac{N}{i}\right)} + \ln\left(\frac{N}{i}\right).$$

Noting that $\frac{1}{i}\sum_{j=1}^{i} \ell_T(j) \le \ell_T(i)$ and the above holds for all i simultaneously finishes the proof. $\square$

Note that this is an improved version of the quantile bound that we proved last time, and it combines both the quantile bound and the "small-loss" bound.

2.3 Bound (3) implies constant regret in a stochastic setting

Finally, we consider a specific stochastic setting where there is a good expert that is distinguishable from others in expectation.

Theorem 3. *Suppose there exists a good expert $i^\star$ and a constant gap $\Delta \in (0, 1]$ such that $\mathbb{E}_t[\ell_t(i) - \ell_t(i^\star)] \ge \Delta$ for all t and $i \ne i^\star$, where $\mathbb{E}_t$ denotes the conditional expectation given all the randomness up to round t. Then Bound* (3) *implies*

$$\mathbb{E}[R_T(i^\star)] \le \frac{\ln\left(\frac{1}{p_1(i^\star)}\right)}{\Delta}.$$

Proof. By the condition we have

$$\mathbb{E}_t[r_t(i^\star)] = \mathbb{E}_t\left[\sum_{i \ne i^\star} p_t(i)(\ell_t(i) - \ell_t(i^\star))\right] \ge \Delta(1 - p_t(i^\star)),$$

and therefore $\mathbb{E}[R_T(i^\star)] \ge \Delta B$ where we define $B = \mathbb{E}\left[\sum_{t=1}^{T}(1 - p_t(i^\star))\right]$. On the other hand,

$$\begin{aligned}
\mathbb{E}[V_T(i^\star)] &\le \mathbb{E}\left[\sum_{t=1}^{T} |\langle p_t, \ell_t\rangle - \ell_t(i^\star)|\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)|\ell_t(i) - \ell_t(i^\star)|\right] && \text{(Jensen's inequality)} \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i \ne i^\star} p_t(i)\right] = B.
\end{aligned}$$

Therefore, by setting q to be the distribution that concentrates on $i^\star$, Eq. (3) implies

$$\begin{aligned}
\Delta B \le \mathbb{E}[R_T(i^\star)] &\le \mathbb{E}\sqrt{V_T(i^\star)\ln\left(\frac{1}{p_1(i^\star)}\right)} \\
&\le \sqrt{\mathbb{E}[V_T(i^\star)]\ln\left(\frac{1}{p_1(i^\star)}\right)} && \text{(Jensen's inequality)} \\
&\le \sqrt{B\ln\left(\frac{1}{p_1(i^\star)}\right)}. && (6)
\end{aligned}$$

Solving for B gives $B \le \ln\left(\frac{1}{p_1(i^\star)}\right)/\Delta^2$. Plugging this back to Eq. (6) finishes the proof. $\square$

For a uniform prior p_1, the bound simply becomes $(\ln N)/\Delta$, which is independent of T!

References

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *28th Annual Conference on Learning Theory*, 2015.