Machine Learning Theory (CS 6783)

Lecture 21: Oracle Efficient Contextual Bandits

1 ERM Oracle Efficient Contextual Bandits

Recall the contextual Bandit problem given by protocol

- For $t = 1$ to n
  - Nature produces context $x_t \in \mathcal{X}$
  - Algorithm picks arm $I_t \in [N]$ in a possibly randomized fashion while nature produces loss vector ℓ_t
  - Learner suffers loss $\ell_t[I_t]$

Goal: Minimize regret w.r.t. class of policies $\mathcal{F} \subset [N]^{\mathcal{X}}$ given by

$$\text{Reg}_n = \frac{1}{n}\sum_{t=1}^{n} \ell_t[I_t] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell_t[f(x_t)]$$

Assume $(x_t, \ell_t) \sim D$ some fixed distribution.

We would like our algorithm to make a small number of calls to the ERM oracle that, given samples $(x_1, \tilde{\ell}_1), \ldots, (x_m, \tilde{\ell}_m)$ can return ERM policy given by:

$$\widehat{f}_{\text{ERM}} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{t=1}^{m} \tilde{\ell}_t[f(x_t)]$$

We already saw that a plain ϵ-greedy algorithm would give us a regret bound of $O\left(\frac{N \log |\mathcal{F}|}{n}\right)^{1/3}$ with very few calls to an ERM oracle. Before seeing how we can get an ERM oracle efficient algorithm with optimal regret bound, we will first see an algorithm that is computationally as bad as EXP4 but enjoys optimal bound on regret like EXP4 for the stochastic case. This algorithm called policy elimination will help us build ideas for the optimal oracle efficient algorithm.

1.1 Policy Elimination

For the ϵ-greedy algorithm on every round picked the policy that optimized sum of past estimated losses with probability $1-\gamma$ and with probability γ picked the uniform distribution. In a sense we will use the same idea here, but instead of picking with probability $1-\gamma$ the ERM, we will pick with probability $1-\gamma$, a distribution $q_t(\cdot|x_t)$ and with probability γ uniformly explore as before. But the key idea we will use are, first the distribution $q_t(\cdot|x_t)$ will be a distribution over only a set $\mathcal{F}_t$ at time t that has low estimated regret so far to begin with. Further the distribution q_t we

will pick will be such that the variance of estimated losses under the distribution of our draw is bounded by N. These together will ensure optimal regret bound.

Policy Elimination Algorithm:

Initialize $\mathcal{F}_1 = \mathcal{F}$, define $\epsilon_t = \sqrt{\frac{N \log(|\mathcal{F}|n/\delta)}{t}}$ and $\gamma_t = \min\left\{1, \sqrt{\frac{N \log(|\mathcal{F}|n/\delta)}{2t}}\right\}$

For $t = 1$ to n

Pick distribution $q_t \in \Delta(\mathcal{F}_t)$ s.t.

$$\forall f \in \mathcal{F}_t, \quad \mathbb{E}_{x\sim D}\left[\frac{1}{(1-\gamma)\sum_{f'\in\mathcal{F}:f'(x)=f(x)} q_t(f') + \gamma/N}\right] \leq 2N$$

Draw policy $f_t \sim q_t$ and set $a_t = f_t(x_t)$

Observe $\ell_t[a_t]$

Build estimate $\tilde{\ell}_t$ and update $\mathcal{F}_{t+1} = \left\{f' \in \mathcal{F}_t : \frac{1}{t}\sum_{j=1}^{t} \tilde{\ell}_j[f'(x_t)] - \inf_{f\in\mathcal{F}_t} \frac{1}{t}\sum_{j=1}^{t} \tilde{\ell}_j[f(x_t)] \leq 2\epsilon_{t+1}\right\}$

End For

Theorem 1. *With probability at least $1-\delta$, for the policy elimination algorithm,*

$$\mathrm{Reg}_n \leq O\left(\sqrt{\frac{N \log(|\mathcal{F}|n/\delta)}{n}}\right)$$

Proof. The proof of the above theorem is obvious if we can show the following statement. With probability $1-\delta$, for any t and any $f \in \mathcal{F}_t$,

$$\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] - \inf_{f'\in\mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f'(x)]\right] \leq 4\epsilon_t$$

The idea then is that since we are picking $f_t \in \mathcal{F}_t$ and since every $f \in \mathcal{F}_t$ has small excess risk of $4\epsilon_t$. Hence, we can conclude that, with probability $1-\delta$,

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} \ell_t[a_t] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell_t[f(x_t)] &\leq \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f_t(x)]\right] - \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] + \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}} \\
&\leq 4\frac{1}{n}\sum_{t=1}^{n} \epsilon_t + \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}} \\
&\leq O\left(\sqrt{\frac{N \log(|\mathcal{F}|n/\delta)}{n}}\right)
\end{aligned}$$

□

Lemma 2. *With probability at least $1-\delta$, for any $t \in [n]$,*

$$\sup_{f\in\mathcal{F}_t} \left|\frac{1}{t}\sum_{j=1}^{t} \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]\right| \leq 2\epsilon_t$$

and we have that with probability $1-\delta$*, for any* t *and any* $f \in \mathcal{F}_t$*,*

$$\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] - \inf_{f^* \in \mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right] \leq 4\epsilon_t$$

Proof. First note that for any t, if we consider any $j \leq t$, for any $f \in \mathcal{F}_t$, $\tilde{\ell}_j[f(x_j)]$ is an unbiased estimator of $\mathbb{E}_{(\ell,x)}\left[\ell[f(x)]\right]$. Hence, for any $f \in \mathcal{F}_t$, $\frac{1}{t}\sum_{j=1}^t \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]$ is an average of martingale difference sequence. Just like for iid random variables we have the Bernstein concentration inequality, we have for martingale difference sequences a concentration called Freedman inequality which states the following. Let $(Y_t)_{t\in\mathbb{N}}$ be a martingale difference sequence such that Y_t is bounded by B and such that $\mathbb{E}_{t-1}\left[Y_t^2\right] \leq V_t$, then for any $\delta > 0$, with probability at least $1-\delta$, for any t

$$\left|\frac{1}{t}\sum_{j=1}^t Y_j\right| \leq \sqrt{\frac{\left(\sum_{j=1}^t V_t\right)\log(\log(t)/\delta)}{n}} + \frac{B\log(\log(t)/\delta)}{t}$$

Now note that taking $Y_j^f = \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]$ and using the above Freedman's inequality with union bound over $\mathcal{F}$ we get, that with probability $1-\delta$, for any $t \in [n]$ and any $f \in \mathcal{F}$,

$$\begin{aligned}\sup_{f\in\mathcal{F}_t}\left|\frac{1}{t}\sum_{j=1}^t \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]\right| &= \sup_{f\in\mathcal{F}}\left|\frac{1}{t}\sum_{j=1}^t Y_j^f\right| \\ &\leq \sqrt{\frac{\left(\sup_{f\in\mathcal{F}}\sum_{j=1}^t V_t^f\right)\log(n|\mathcal{F}|/\delta)}{n}} + \frac{B\log(n|\mathcal{F}|/\delta)}{t}\end{aligned}$$

However note that, $|\tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]| \leq \frac{N}{\gamma} \leq \sqrt{\frac{Nn}{\log(n|\mathcal{F}|/\delta)}} = B$ and,

$$\mathbb{E}_{t-1}\left[Y_t^2\right] \leq \mathbb{E}_{t-1}\left[\sum_{a\in[N]}\left((1-\gamma)q_t(a|x_t) + \gamma/N\right)\tilde{\ell}_t[a]^2\right] \leq \mathbb{E}_{x\sim D}\left[\frac{1}{(1-\gamma)\sum_{f'\in\mathcal{F}:f'(x)=f(x)} q_t(f') + \gamma/N}\right] \leq N$$

Hence we have that,

$$\sup_{f\in\mathcal{F}_t}\left|\frac{1}{t}\sum_{j=1}^t \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]\right| \leq O\left(\sqrt{\frac{N\log(n|\mathcal{F}|/\delta)}{t}}\right) = 2\epsilon_t$$

Next, note that since $f^* \in \mathcal{F}$ is the minimizes of expected loss, we have that with probability $1-\delta$, $f^* \in \mathcal{F}_t$ for any t. Now using the above inequality, we get that with probability $1-\delta$, for any $f \in \mathcal{F}$,

$$\left|\frac{1}{t}\sum_{j=1}^t \tilde{\ell}_j[f(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]\right| \leq \epsilon_t$$

and

$$\left|\frac{1}{t}\sum_{j=1}^t \tilde{\ell}_j[f^*(x_j)] - \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right]\right| \leq \epsilon_t$$

But by definition of $\mathcal{F}_t$, we only retain those f's for which average estimated loss is close to that of ERM over $\mathcal{F}_t$ and so, all the average estimates losses within $\mathcal{F}_t$ are within ϵ_t factor and so,

$$\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] - \inf_{f^*\in\mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right] \leq 4\epsilon_t$$

This completes the proof. □

Note that the above algorithm is optimal in terms of its regret bound with high probability. However, since we need to maintain $\mathcal{F}_t$ the set of good experts on every round, the algorithm is as intractable as EXP4. But we can use the idea from this policy elimination algorithm to develop an efficient algorithm.

1.2 Oracle Efficient Algorithm

A key reason why we needed to maintain $\mathcal{F}_t$ in policy elimination was that we had to find a distribution that had low variance of N for every policy under consideration. Hence the only way we could so this and still have a distribution that had good expected regret was by shrinking $\mathcal{F}_t$ to only good policies. A soft version of policy elimination one could consider could have on every round a distribution over entire $\mathcal{F}$ but then have variance bound of N only for good policies and for bad policies allow much larger variance (of $\sqrt{t}$ on round t for instance). In fact the soft policy elimination algorithm is as follows: **Soft Policy Elimination Algorithm:**

For $t = 1$ to n

Pick distribution $q_t \in \Delta(\mathcal{F})$ s.t.

$$\mathbb{E}_{f\sim q_t}\left[\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f(x_t)]\right] - \inf_{f^*\in\mathcal{F}}\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f^*(x_t)] \leq \sqrt{\frac{N\log(|\mathcal{F}|n)}{n}}$$

and for every $f \in \mathcal{F}$,

$$\mathbb{E}_{x\sim D}\left[\frac{1}{(1-\gamma)\sum_{f'\in\mathcal{F}:f'(x)=f(x)} q_t(f') + \gamma/N}\right] \leq 2N + \frac{\sqrt{N\log(|F|n/\delta)}}{\sqrt{t}}\sum_{j=1}^{t}\tilde{\ell}_j[f(x_t)] - \inf_{f^*\in\mathcal{F}}\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f^*(x_t)]$$

Draw policy $f_t \sim q_t$ and set $a_t = f_t(x_t)$

Observe $\ell_t[a_t]$ and build estimate $\tilde{\ell}_t$ based on it.

End For

The key idea is that expected regret under the distribution is bounded by what we would like and under this distribution, the variance of loss estimated for any $f \in \mathcal{F}$ scales as order $N + \sqrt{t}\,\widehat{\mathrm{Reg}_t(f)}$ where $\widehat{\mathrm{Reg}_t(f)}$ is the estimated regret of policy f. The idea being that if a policy has large regret then variance for that policy can be quite large. For instance, policies with constant regret allow $\sqrt{t}$ additive factor on variance. IF we use the Freedman inequality with this updated bound on variance we get the following lemma.

Lemma 3. *With probability $1-\delta$, for any t and any $f \in \mathcal{F}$,*

$$\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f(x_j)] - \inf_{f' \in \mathcal{F}} \frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f'(x_j)] \leq 2\left(\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] - \inf_{f^* \in \mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right]\right) + \sqrt{\frac{N\log(|\mathcal{F}|n/\delta)}{t}}$$

and

$$\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right] - \inf_{f^* \in \mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right] \leq 2\left(\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f(x_j)] - \inf_{f' \in \mathcal{F}} \frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f'(x_j)]\right) + \sqrt{\frac{N\log(|\mathcal{F}|n/\delta)}{t}}$$

The proof uses the same steps as the proof of lemma 2 except that for every $f \in \mathcal{F}$ we use the variance of $f \in \mathcal{F}$ that is bounded as $N + \sqrt{t}\,\widehat{\mathrm{Reg}_t(f)}$ and then simply apply the fact that $\sqrt{ab} \leq a/2 + b/2$ to get the factor 2 on regret and estimated regrets. However, since q_t the distribution over $\mathcal{F}$ that we get is such that

$$\mathbb{E}_{f\sim q_t}\left[\frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f(x_t)]\right] - \inf_{f^* \in \mathcal{F}} \frac{1}{t}\sum_{j=1}^{t}\tilde{\ell}_j[f^*(x_t)] \leq \sqrt{\frac{N\log(|\mathcal{F}|n/\delta)}{t}}$$

, combined with the above it implies that

$$\mathbb{E}_{f\sim q_t}\left[\mathbb{E}_{(x,\ell)\sim D}\left[\ell[f(x)]\right]\right] - \inf_{f^* \in \mathcal{F}} \mathbb{E}_{(x,\ell)\sim D}\left[\ell[f^*(x)]\right] \leq 2\sqrt{\frac{N\log(|\mathcal{F}|n/\delta)}{t}}$$

Using this we can conclude with simple concentration that with probability $1-\delta$,

$$\mathrm{Reg}_n \leq O\left(\sqrt{\frac{N\log(|\mathcal{F}|n/\delta)}{n}}\right)$$

which is the optimal bound.

But have we done anything useful at all? note that q_t is still a distribution over $\mathcal{F}$ just like in EXP4 case. So why can we hope to implement this method oracle efficiently. Well while I shall skip the proof for this, the idea is that q_t that we need to get will be a distribution that is sparse and has only a small support in $\mathcal{F}$. Further, this sparse distribution can be computed by performing coordinate descent and each coordinate can be computed using the ERM oracle. Hence overall we can compute this distribution q_t which is over a large set in an efficient manner.

2 Online Square Loss Regression Oracle

While the above approach does give an algorithm that is ERM oracle efficient, the above requires finding the ERM. However note that even if we had only two actions, the ERM optimization can be as bad as optimizing w.r.t. binary classification loss which in most cases is again computationally hard. The typical way out of this for classification is to replace the zero-one loss by some nicer losses like square loss or logistic loss etc. In this section, under the so called realizability assumption we can show that one can get contextual bandit algorithms that are online squared loss regression oracle efficient. To be able to achieve this, the algorithms require the following realizability assumption.

Assumption 4. *Assume that there is a class $\mathcal{L} \subset \mathbb{R}^{\mathcal{X}\times[N]}$ such that for some $g^* \in \mathcal{L}$, we have that for any $x \in \mathcal{X}$ and $a \in [N]$,*

$$\mathbb{E}\left[\ell_t(a)|x_t = x\right] = g^*(x, a)$$

The assumption tells us that the conditional expected loss of any action given a context is modeled will by a member g^* of some class $\mathcal{L}$. The rough idea then is to learn this model in some sense and take actions that are (close to) optimal w.r.t. this model given a context. More specifically, on every round we make a prediction of the loss given context for every action based on our ability to solve online squared loss regression w.r.t. class $\mathcal{L}$. Then we take a distribution that is skewed towards making the best decision based on this model for losses. Specifically we use the following algorithm.

SquareCB:

Set $\gamma = \sqrt{\frac{4N}{\text{RegSQ}_n(\mathcal{L})}}$

For $t = 1$ to n

Receive context $x_t \in \mathcal{X}$

For each action $a \in [N]$, compute $\hat{y}_t[a] = \hat{y}_t(x_t, a)$ by feeding x_t, a as input to the square loss regression algorithm for round t

Set $b_t = \underset{b\in[N]}{\text{argmin}}\, \hat{y}_t[b]$

$\forall a \neq b_t$, set $p_t(a) = \frac{1}{N+\gamma(\hat{y}_t[a]-\hat{y}_t[b_t])}$, set $p_t(b_t) = 1 - \sum_{a\neq b_t} p_t(a)$

Draw action $a_t \sim p_t$ and observe $\ell_t[a_t]$

Use online regression algorithm by feeding it input instance (x_t, a_t) and output $\ell_t[a_t]$

End For

Theorem 5. *Assume we have access to an online regression oracle that guarantees that for any sequence of context action pairs $(x_1, a_1), \ldots, (x_n, a_n)$ as input and any labels $y_1, \ldots, y_n$ produced possibly by an adversary, we have an online learning algorithm that guarantees that:*

$$\frac{1}{n}\sum_{t=1}^{n}((\hat{y}_t - y_t)^2 - \inf_{g\in\mathcal{L}} \frac{1}{n}\sum_{t=1}^{n}\left(g^*(a_t, x_t) - y_t\right)^2 \leq \text{RegSQ}_n(\mathcal{L})$$

where $\text{RegSQ}_n(\mathcal{L})$ *is the bound guaranteed for online squared loss regression against* $\mathcal{L}$*. Then, the Square CB algorithm enjoys the regret bound,*

$$\mathbb{E}\left[\text{Reg}_n\right] \leq \sqrt{3N\ \text{RegSQ}_n(\mathcal{L})}$$

where

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_n\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t(a_t) - \frac{1}{n}\sum_{t=1}^{n}\ell_t(f^*(x_t))\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\ell_t(a_t)|x=x_t\right] - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\ell_t(f^*(x_t))|x=x_t\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}g^*(a_t,x_t) - \frac{1}{n}\sum_{t=1}^{n}g^*(f^*(x_t),x_t)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\left(g^*(a_t,x_t) - g^*(f^*(x_t),x_t) - \frac{\gamma}{2}\left(\hat{y}_t(x_t,a_t) - g^*(a_t,x_t)\right)^2\right)\right] \\
&\qquad + \frac{\gamma}{2}\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t(x_t,a_t) - g^*(a_t,x_t)\right)^2\right]
\end{aligned}$$

But due to realizability, $\left(\hat{y}_t(x_t,a_t) - g^*(a_t,x_t)\right)^2 = \mathbb{E}\left[\left(\hat{y}_t(x_t,a_t) - \ell_t[a_t]\right)^2 - \left(g^*(a_t,x_t) - \ell_t[a_t]\right)^2 | x_t = x\right]$

$$\begin{aligned}
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\left(g^*(a_t,x_t) - g^*(f^*(x_t),x_t) - \frac{\gamma}{2}\left(\hat{y}_t(x_t,a_t) - g^*(a_t,x_t)\right)^2\right)\right] \\
&\qquad + \frac{\gamma}{2}\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\left(\left(\hat{y}_t(x_t,a_t) - \ell_t[a_t]\right)^2 - \left(g^*(a_t,x_t) - \ell_t[a_t]\right)^2\right]
\end{aligned}$$

replacing $g^*(\cdot,x_t)$ by taking supremum over vector $g^* \in [0,1]^N$ for each round, and replacing $f^*(x_t)$ by macimum $a^* \in [N]$ we move to upper bound,

$$\begin{aligned}
&\leq \frac{1}{n}\sum_{t=1}^{n}\sup_{g^*\in[0,1]^N}\max_{a^*\in[N]}\mathbb{E}_{a_t\sim p_t}\left[g^*[a_t] - g^*[a^*] - \frac{\gamma}{2}\left(\hat{y}_t(x_t,a_t) - g^*[a_t]\right)^2\right] \\
&\qquad + \frac{\gamma}{2}\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\left(\left(\hat{y}_t(x_t,a_t) - \ell_t[a_t]\right)^2 - \left(g^*(a_t,x_t) - \ell_t[a_t]\right)^2\right] \\
&\leq \frac{1}{n}\sum_{t=1}^{n}\sup_{g^*\in[0,1]^N}\max_{a^*\in[N]}\mathbb{E}_{a_t\sim p_t}\left[g^*[a_t] - g^*[a^*] - \frac{\gamma}{2}\left(\hat{y}_t(x_t,a_t) - g^*[a_t]\right)^2\right] \\
&\qquad + \frac{\gamma}{2}\mathrm{RegSQ}_n(\mathcal{L})
\end{aligned}$$

where in the above, $\mathrm{RegSQ}_n(\mathcal{L})$ is the regret bound for online square loss regression w.r.t. loss class $\mathcal{L}$. It can be shown that for the choice of distribution p_t (shown in next lemma), we have that for any t:

$$\sup_{g^*\in[0,1]^N}\max_{a^*\in[N]}\mathbb{E}_{a_t\sim p_t}\left[g^*[a_t] - g^*[a^*] - \frac{\gamma}{2}\left(\hat{y}_t(x_t,a_t) - g^*[a_t]\right)^2\right] \leq \frac{3N}{2\gamma}$$

Using this we can conclude that:

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \frac{3N}{2\gamma} + \frac{\gamma}{2}\mathrm{RegSQ}_n(\mathcal{L})$$

using $\gamma = \sqrt{\frac{3N}{\text{RegSQ}_n(\mathcal{L})}}$ we obtain that:

$$\mathbb{E}\left[\text{Reg}_n\right] \leq \sqrt{3N \ \text{RegSQ}_n(\mathcal{L})}$$

□

The point to note is that for a finite class $\mathcal{L}$, it turns out the exponential weights algorithm can actually ensure that $\text{RegSQ}_n(\mathcal{L}) \leq \frac{\log|\mathcal{L}|}{n}$ and so for finite $\mathcal{L}$ class one has,

$$\mathbb{E}\left[\text{Reg}_n\right] \leq \sqrt{\frac{2N\log|\mathcal{L}|}{n}}$$

Lemma 6. *For any vector $\hat{y} \in [0,1]^N$, let $b^* = \operatorname{argmin}_{a\in[N]} \hat{y}[a]$. Let distribution $p \in \Delta_N$ be given by, $\forall a \neq b^*, p(a) = \frac{1}{N+\gamma(\hat{y}[a]-\hat{y}[b^*])}$ and $p(b^*) = 1 - \sum_{a\neq b^*} p(a)$, then,*

$$\sup_{g\in[0,1]^N} \max_{a^*\in[N]} \mathbb{E}_{a\sim p}\left[g[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right] \leq \frac{3N}{2\gamma}$$

Proof. Now consider any $a^* \in [N]$ and any $g \in [0,1]^N$. Note that,

$$\begin{aligned}
&\mathbb{E}_{a\sim p}\left[g[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right] \\
&\quad= \sum_{a\in[A]} p(a)\left(g[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right) \\
&\quad= \sum_{a\neq a^*} p(a)\left(g[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right) - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad= \sum_{a\neq a^*} p(a)\left(g[a] - \hat{y}[a] + \hat{y}[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right) - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2
\end{aligned}$$

But now note that $g[a] - \hat{y}[a] \leq \frac{\gamma}{2}(\hat{y}[a]-g[a])^2 + \frac{1}{2\gamma}$ and so we have,

$$\begin{aligned}
&\mathbb{E}_{a\sim p}\left[g[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right] \\
&\quad= \sum_{a\neq a^*} p(a)\left(g[a] - \hat{y}[a] + \hat{y}[a] - g[a^*] - \frac{\gamma}{2}(\hat{y}[a]-g[a])^2\right) - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad\leq \sum_{a\neq a^*} p(a)\left(\hat{y}[a] - g[a^*] + \frac{1}{2\gamma}\right) - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad= \sum_{a\neq a^*} p(a)\left(\hat{y}[a] - g[a^*]\right) + \frac{1-p(a^*)}{2\gamma} - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad= \sum_{a\neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[a^*] + \hat{y}[a^*] - g[a^*]\right) + \frac{1-p(a^*)}{2\gamma} - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad= \sum_{a\neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[a^*]\right) + (1-p[a^*])\left(\hat{y}[a^*] - g[a^*]\right) + \frac{1-p(a^*)}{2\gamma} - \frac{p(a^*)\gamma}{2}(\hat{y}[a^*]-g[a^*])^2 \\
&\quad= \sum_{a\neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[a^*]\right) + (1-p[a^*])\left(\hat{y}[a^*] - g[a^*] - \frac{p(a^*)\gamma}{2(1-p[a^*])}(\hat{y}[a^*]-g[a^*])^2\right) + \frac{1-p(a^*)}{2\gamma}
\end{aligned}$$

Again using AM-GM to note that $\left(\hat{y}[a^*] - g[a^*] - \frac{p(a^*)\gamma}{2(1-p[a^*])}(\hat{y}[a^*] - g[a^*])^2\right) \leq \frac{1-p(a^*)}{2\gamma p(a^*)}$ we conclude that,

$$\leq \sum_{a \neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[a^*]\right) + \frac{(1-p(a^*))^2}{2\gamma p(a^*)} + \frac{1-p(a^*)}{2\gamma}$$
$$= \sum_{a \neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[a^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$

Recall that $b^* = \underset{a \in [N]}{\operatorname{argmin}}\, \hat{y}[a]$,

$$= \sum_{a \neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[b^*] + \hat{y}[b^*] - \hat{y}[a^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$
$$= \sum_{a \neq a^*} p(a)\left(\hat{y}[a] - \hat{y}[b^*]\right) + (1-p(a^*))\left(\hat{y}[b^*] - \hat{y}[a^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$
$$= \sum_{a=1}^{N} p(a)\left(\hat{y}[a] - \hat{y}[b^*]\right) - \left(\hat{y}[a^*] - \hat{y}[b^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$
$$= \sum_{a: a \neq b^*} \frac{(\hat{y}[a] - \hat{y}[b^*])}{N + \gamma((\hat{y}[a] - \hat{y}[b^*]))} - \left(\hat{y}[a^*] - \hat{y}[b^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$
$$= \sum_{a: a \neq b^*} \frac{1}{\frac{N}{(\hat{y}[a] - \hat{y}[b^*])} + \gamma} - \left(\hat{y}[a^*] - \hat{y}[b^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$

since each $(\hat{y}[a] - \hat{y}[b^*]) \leq 1$,

$$\leq \frac{N-1}{N+\gamma} - \left(\hat{y}[a^*] - \hat{y}[b^*]\right) + \frac{(1-p(a^*))}{2p(a^*)\gamma}$$
$$\leq \frac{N-1}{N+\gamma} + \max\left\{\frac{(1-p(b^*))}{2p(b^*)\gamma}, \max_{a \neq b^*}\left\{\frac{(1-p(a))}{2p(a)\gamma} - \left(\hat{y}[a] - \hat{y}[b^*]\right)\right\}\right\}$$
$$= \frac{N-1}{N+\gamma} + \max\left\{\frac{(1-p(b^*))}{2p(b^*)\gamma}, \max_{a \neq b^*}\left\{\frac{N + \gamma\left(\hat{y}[a] - \hat{y}[b^*]\right)}{2\gamma} - \frac{1}{2\gamma} - \left(\hat{y}[a] - \hat{y}[b^*]\right)\right\}\right\}$$
$$= \frac{N-1}{N+\gamma} + \max\left\{\frac{(1-p(b^*))}{2p(b^*)\gamma}, \max_{a \neq b^*}\left\{\frac{N-1}{2\gamma} - \frac{1}{2}\left(\hat{y}[a] - \hat{y}[b^*]\right)\right\}\right\}$$

Note that $p(b^*) \geq 1/N$ because we are picking b^* with highest probability, hence

$$\leq \frac{N-1}{N+\gamma} + \max\left\{\frac{N-1}{2\gamma}, \max_{a \neq b^*}\left\{\frac{N-1}{2\gamma} - \frac{1}{2}\left(\hat{y}[a] - \hat{y}[b^*]\right)\right\}\right\}$$
$$= \frac{N-1}{N+\gamma} + \frac{N-1}{2\gamma} \leq \frac{\frac{3}{2}(N-1)}{\gamma}$$

□