CS7545, Spring 2024: Machine Learning Theory - Solutions #3

Due: Apr. 5

1) **Doubling Trick.** Let $k = \lceil \log_2 T \rceil$. Let $T_i = 2^{i-1}$ for $i = 1, \ldots, k$, so that $T \leq \sum_{i=1}^{k} T_i$. So, the bound is (asymptotically)

$$\sum_{i=1}^{k} \sqrt{MT_i} = \sqrt{M \sum_{i=1}^{k} 2^{i-1}} = \sqrt{M} \frac{\sqrt{2}^k - 1}{\sqrt{2} - 1} \leq \sqrt{M} \frac{\sqrt{2T} - 1}{\sqrt{2} - 1} = O(\sqrt{MT})$$

2) **Dynamic Regret.** We have

$$\begin{aligned}
\text{Regret}_T &\leq \sum_{t=1}^{T} \frac{\eta}{2} G^2 + \sum_{t=1}^{T} \frac{(\|\mathbf{w}_t - \mathbf{w}_t^*\|)^2 - (\|\mathbf{w}_{t+1} - \mathbf{w}_t^*\|)^2}{2\eta} \\
&\leq \frac{TG^2\eta}{2} + \sum_{t=1}^{T} \frac{\left(\|\mathbf{w}_t\|^2 - 2\langle \mathbf{w}_t, \mathbf{w}_t^* \rangle + \|\mathbf{w}_t^*\|^2 - \|\mathbf{w}_{t+1}\|^2 + 2\langle \mathbf{w}_{t+1}, \mathbf{w}_t^* \rangle - \|\mathbf{w}_t^*\|^2\right)}{2\eta} \\
&\leq \frac{TG^2\eta}{2} + \frac{1}{2\eta}(\|\mathbf{w}_1\|^2 - \|\mathbf{w}_{T+1}\|^2) + \frac{1}{\eta} \sum_{t=1}^{T} \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \mathbf{w}_t^* \rangle \\
&\leq \frac{TG^2\eta}{2} + \frac{D^2}{2\eta} + \frac{1}{\eta}(\langle \mathbf{w}_{T+1}, \mathbf{w}_T^* \rangle - \langle \mathbf{w}_1, \mathbf{w}_1^* \rangle) + \frac{1}{\eta} \sum_{t=2}^{T} \langle \mathbf{w}_{t-1}^* - \mathbf{w}_t^*, \mathbf{w}_t \rangle \\
&\leq \frac{\eta G^2 T}{2} + \frac{7D^2}{4\eta} + \frac{D}{\eta} \sum_{t=2}^{T} \|\mathbf{w}_t^* - \mathbf{w}_{t-1}^*\| \\
&\leq \frac{\eta G^2 T}{2} + \frac{7D^2 + 4DP_T}{4\eta}.
\end{aligned}$$

where we use the following relation

$$\begin{aligned}
\|\mathbf{w}_1\|^2 = \|\mathbf{w}_1 - \mathbf{0}\|^2 &\leq D^2, \\
\mathbf{w}_{T+1}^\top \mathbf{w}_T^* \leq \|\mathbf{w}_{T+1}\| \, \|\mathbf{w}_T^*\| &\leq D^2, \\
-\mathbf{w}_1^\top \mathbf{w}_1^* \leq \tfrac{1}{4} \|\mathbf{w}_1 - \mathbf{w}_1^*\|^2 &\leq \tfrac{1}{4} D^2, \\
\langle \mathbf{w}_{t-1}^* - \mathbf{w}_t^*, \mathbf{w}_t \rangle \leq \left\|\mathbf{w}_{t-1}^* - \mathbf{w}_t^*\right\| \|\mathbf{w}_t\| &\leq D \left\|\mathbf{w}_{t-1}^* - \mathbf{w}_t^*\right\|.
\end{aligned}$$

3) **Parameter Tuning.**

(a) The minimum occurs when $T\eta = \eta^{-2}$. So, $\eta = T^{-1/3}$. So, the upper bound is $O(T^{-2/3})$.

(b) Due to the exponential term, η must have the form $\log f(T)$ for some sub-linear function f of T. Furthermore, the first term $\frac{T}{\eta}$ requires that f is an increasing function of T. For example, we can choose $\eta = \log \sqrt{T}$, and the upper bound becomes $\frac{T}{\log \sqrt{T}} + \sqrt{T}$, which is sublinear.

(c) At the minimum, all three terms are equal. So, we want $\frac{T\epsilon}{\eta} = T\eta$, and $\frac{T\epsilon}{\eta} = \frac{N}{\epsilon}$

The first condition implies $\epsilon = \eta^2$. The second condition implies $T\epsilon^2 = N\eta$, which then implies $\eta = \left(\frac{N}{T}\right)^{1/3}$ after the substitution $\epsilon = \eta^2$,

So, we get an upper bound $3T\eta = O(T\eta) = O(T^{\frac{2}{3}}N^{\frac{1}{3}})$.

(d) For simplicity, we let $f(\eta, \epsilon) = \frac{\log N}{\eta} + \frac{\eta T}{\epsilon^2} + 2\epsilon T$, which is an upper bound of the original objective. Taking the derivative and setting to zero, we have

$$\frac{\partial f}{\partial \epsilon} = -2(\eta T)\epsilon^{-3} + 2T = 0,$$

which implies $\epsilon^* = \eta^{\frac{1}{3}}$. It is obvious that this is the global minimum. Plugging into f, we have

$$f(\eta; \epsilon^*) = \frac{\log N}{\eta} + 3\eta^{\frac{1}{3}}T.$$

We can again take the derivative w.r.t. η, then we have

$$\frac{\partial f}{\partial \eta} = -\frac{\log N}{\eta^2} + \eta^{-\frac{2}{3}}T = 0$$

and we get

$$\eta^* = \left(\frac{\log N}{T}\right)^{\frac{3}{4}}.$$

We have

$$f(\eta^*, \epsilon^*) = 4(\log N)^{\frac{1}{4}}T^{\frac{3}{4}} = O\left((\log N)^{\frac{1}{4}}T^{\frac{3}{4}}\right),$$

which is an upper bound of our original objective.

(e) Write the bound as

$$\frac{\log N}{1 - \exp(-\eta)} + \frac{\eta T}{1 - \exp(-\eta)}.$$

The first term is a decreasing function of η, and the second term is an increasing function of η (To verify, take the derivative $\frac{e^n(e^n - n - 1)}{(e^n - 1)^2} > 0, \forall n > 0$). Since $T >> \log N$, the optimal value for η must be small, say less than 1.

Note that for $\eta \in (0, 1)$,

$$\frac{\eta}{1 - e^-\eta} \leq (\eta + 1)$$

and

$$(1 - \exp(-\eta))^{-1} \leq 2/\eta$$

Using the above inequalities, the bound becomes

$$\frac{\log N}{1 - \exp(-\eta)} + \frac{\eta T}{1 - \exp(-\eta)} \leq \frac{2\log N}{\eta} + (1 + \eta)T = \left(\frac{2\log N}{\eta} + \eta T\right) + T$$

Take $\eta \leftarrow \sqrt{\frac{2\log N}{T}}$ and we have an upper bound $O(\sqrt{2T\log N} + T)$.

4) **Online Non-Convex Optimization.** We partition X into 2 -norm ϵ-balls. Each ϵ ball has size $O\left(\epsilon^n\right)$ and we need $N := O\left(1/\epsilon^n\right)$ of those to cover X. We treat each ball as an expert and run Hedge.

Hedge suffers $O(\sqrt{T \log N})$ regret with respect to the best expert. Now we need to analyze the best expert's regret with respect to the best fixed-point prediction. Let $x^* = \arg\min_{x \in X} \sum_t f_t(x)$. Then, one of the experts must satisfy $\|x - x^*\|_2 \leq \epsilon$, which by Lipschitz assumption, implies $f_t(x) - f_t\left(x^*\right) \leq \epsilon$ for all t. So, the best expert suffers at most $T\epsilon$ regret with respect to the best fixed-point prediction. The total regret of the algorithm is therefore upper-bounded by

$$O(\sqrt{T \log N} + T\epsilon) = O\left(\sqrt{Tn \log \frac{1}{\epsilon}} + T\epsilon\right).$$

Set $\epsilon = 1/T$, and the regret now becomes $O(\sqrt{nT \log T})$.