ECE 434: Final

Due on December 21, 2023

Professor Chi Jin

Evan Dogariu

I pledge my honor that I have not violated the Honor Code during this examination.

Problem 1

We consider the classification problem with input domain $X := \{x \in \mathbb{R}^d : \ \|x\|_\infty \leq D\}$ and label set $Y = \{-1, 1\}$.

(a) Consider the following linear function class on X with ℓ_1 constraints:

$$\mathcal{F}_1 := \{x \mapsto w^\top x : \ w \in \mathbb{R}^d, \|w\|_1 \leq B\}$$

Prove that the Rademacher complexity

$$R_n(\mathcal{F}_1) \leq DB\sqrt{\frac{2\log(2d)}{n}}$$

(b) Consider the following function class of 2-layer neural networks with m ReLU units:

$$\mathcal{F}_2 := \left\{x \mapsto \sum_{i \in [m]} w_i ReLU(v_i^\top x) : \ w \in \mathbb{R}^m, \|w\|_1 \leq B_2, v_i \in \mathbb{R}^d, \|v_i\|_1 \leq B_1\right\}$$

Provide an upper bound of the Rademacher complexity $R_n(\mathcal{F}_2)$.

(c) Let $\{(x_j, y_j)\}_{j=1}^n$ be the training data. Consider the setting with function class $\mathcal{F}_1$ and hinge loss $\ell(f(x), y) = \max\{0, 1 - yf(x)\}$. Write out the ERM $\hat{f}$ in this setting.

(d) Bound the excess risk of $\hat{f}$ in the setting of (c) using parameters (D, B, d, n).

(e) Consider again the linear function class $\mathcal{F}_1$ defined in (a). Prove that the sequential Rademacher complexity also has the following upper bound

$$R_n^{\text{seq}}(\mathcal{F}_1) \leq DB\sqrt{\frac{2\log(2d)}{n}}$$

(f) Prove that when choosing $\mathcal{F}_1$ as decision space, and using hinge loss to measure regret, there exists an online learning algorithm that can achieve a regret bound of $\tilde{O}(poly(D, B, \log d, \log n)\sqrt{n})$ where n is the rounds of interaction.

Solution

Proof. (a) We first prove the stronger result for (e) that

$$R_n^{\text{seq}}(\mathcal{F}_1) \leq DB\sqrt{\frac{2\log(2d)}{n}}$$

The result of (a) will then follow obviously as $R_n(\cdot) \leq R_n^{\text{seq}}(\cdot)$ always (to see this, note that for any dataset $\{(x_j, y_j)\}_j$ we may always construct a Z-valued tree $(\mathbf{x}, \mathbf{y})$ where all paths along the tree are the same sequence $(\mathbf{x}_t(\epsilon), \mathbf{y}_t(\epsilon)) = (x_t, y_t)$; then, $R_n(\cdot)$ is equal to the sequential Rademacher complexity conditioned on this tree, which is obviously upper bounded by $R_n^{\text{seq}}(\cdot)$ since the latter is the supremum over all trees). So, we proceed.

Pick any X-valued tree $\mathbf{x}$ of depth n. We will design a finite hypothesis class $\mathcal{I}_{\mathbf{x}}$ such that $\hat{R}_n^{\text{seq}}(\mathcal{F}_1; \mathbf{x}) = \hat{R}_n^{\text{seq}}(\mathcal{I}_{\mathbf{x}}; \mathbf{x})$. To this end, note that for any fixed ϵ we have that

$$\sup_{f \in \mathcal{F}_1} \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon)) = \sup_{w \in \mathbb{R}^d : \ \|w\|_1 \leq B} \sum_{t=1}^n \epsilon_t w^\top \mathbf{x}_t(\epsilon) = \sup_{w \in \mathbb{R}^d : \ \|w\|_1 \leq B} w^\top \left(\sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon)\right)$$

If we let $v_\epsilon := \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon)$, the supremum is obviously attained for $w = B\frac{v_\epsilon}{\|v_\epsilon\|_1}$, where the value is $B\frac{\|v_\epsilon\|_2^2}{\|v_\epsilon\|_1}$. Now, note that we can match this value in expectation via the hypothesis class

$$\mathcal{I}_{\mathbf{x}} := \left\{ z \mapsto \pm \frac{B}{n2^n} \sum_{\epsilon} \left(\sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon) \right)_j \; : \; j = 1, \ldots, d \right\},$$

where the sum is over all the 2^n possibilities for ϵ and $(\cdot)_j$ denotes the j^{th} coordinate. In expectation, the best choice from this hypothesis class will match the best choice from $\mathcal{F}_1$. However, we note that $|\mathcal{I}_{\mathbf{x}}| = 2d$ since for each j there is a choice between $\pm$. Furthermore, we note that for each $f \in \mathcal{I}_{\mathbf{x}}$ we have that $\|f\|_\infty \leq DB$. So, by Theorem 2 in Lecture 10, we see that

$$\hat{R}_n^{\text{seq}}(\mathcal{I}_{\mathbf{x}}; \mathbf{x}) \leq DB\sqrt{\frac{2\log(2d)}{n}},$$

and so

$$\hat{R}_n^{\text{seq}}(\mathcal{F}_1; \mathbf{x}) \leq DB\sqrt{\frac{2\log(2d)}{n}}$$

as well. Since this holds for all trees $\mathbf{x}$, taking the supremum yields

$$R_n^{\text{seq}}(\mathcal{F}_1) \leq DB\sqrt{\frac{2\log(2d)}{n}}$$

as desired.

(b) We begin by noting that

$$\mathcal{F}_2 = \left\{ x \mapsto \sum_{i \in [m]} w_i ReLU(f(x)) : \; w \in \mathbb{R}^m, \|w\|_1 \leq B_2, f \in \mathcal{F}_1 \right\},$$

where $\mathcal{F}_1$ is with norm bound B_1. Write $\mathcal{G}_{w_i} := \{x \mapsto w_i ReLU(f(x)) : \; f \in \mathcal{F}_1\}$. Noting that $ReLU(x) = \max\{0, x\}$ is 1-Lipschitz and sends 0 to 0, we see that by Theorem 1(3, 4) of Lecture 4,

$$R_n(\mathcal{G}_{w_i}) \leq 2|w_i| R_n(\mathcal{F}_1)$$

From here, since $\mathcal{F}_2 = \{x \mapsto \sum_{i=1}^m f_i(x) : \; f_i \in \mathcal{G}_{w_i}, \|w\|_1 \leq B_2\}$, we may apply Theorem 1(6) from Lecture 4 to see that

$$R_n(\mathcal{F}_2) \leq 2B_2 R_n(\mathcal{F}_1),$$

where we noted that $\sum_{i=1}^m |w_i| \leq B_2$ for all allowable w. Combining this with the result from (a), we have that

$$R_n(\mathcal{F}_2) \leq 2DB_1B_2\sqrt{\frac{2\log(2d)}{n}}$$

(c) We see that the empirically risk-minimizing selection of w is given by

$$\hat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d : \; \|w\|_1 \leq B} \left\{ \sum_{j=1}^n \max\{0, 1 - y_j w^\top x_j\} \right\}$$

For this selection of $\hat{w}$, we get the empirical risk minimizer

$$\hat{f}(x) = \hat{w}^\top x = \left(\operatorname*{arg\,min}_{w \in \mathbb{R}^d : \; \|w\|_1 \leq B} \left\{ \sum_{j=1}^n \max\{0, 1 - y_j w^\top x_j\} \right\} \right)^\top x$$

(d) Write the function class

$$\mathcal{G} := \{(x, y) \mapsto \ell(f(x), y) : \ f \in \mathcal{F}_1\}$$

We know that for any (x, y) in our dataset and any $f \in \mathcal{F}_1$, we have that

$$|yf(x)| = |w^\top x| = \left|\sum_{i=1}^d w_i x_i\right| \le \|x\|_\infty \sum_{i=1}^d |w_i| = \|x\|_\infty \|w\|_1 \le DB$$

So, we get that $\ell(f(x), y) = \max\{0, 1 - yf(x)\} \in [0, 1 + DB]$ always. We know from Lecture 3 that with probability $\ge 1 - \delta$,

$$\text{excess risk} \le 4R_n(\mathcal{G}) + 2(1 + DB)\sqrt{\frac{\log(1/\delta)}{2n}}$$

To bound the Rademacher complexity of $\mathcal{G}$, we first define

$$\mathcal{H} := \{(x, y) \mapsto yf(x) : \ f \in \mathcal{F}_1\}$$

Since $y_i \in \{-1, 1\}$, for any fixed y_i we know that the distribution of ϵ_i and $y_i\epsilon_i$ are the same. Thus, $R_n(\mathcal{H}) = R_n(\mathcal{F}_1)$. Next, if we define $\mathcal{J} := \{(x, y) \mapsto 1 - h(x, y) : \ h \in \mathcal{H}\}$, Theorem 1(5) from Lecture 4 tells us that $R_n(\mathcal{J}) \le \frac{1}{\sqrt{n}} + R_n(\mathcal{H}) = \frac{1}{\sqrt{n}} + R_n(\mathcal{F}_1)$. Lastly, since $\ell(f(x), y) = \max\{0, 1 - yf(x)\}$, we may use that $\max\{0, \cdot\}$ is 1-Lipschitz along with Theorem 1(4) from Lecture 4 to see that

$$R_n(\mathcal{G}) \le 2R_n(\mathcal{F}_1) + \frac{2}{\sqrt{n}} \le 2DB\sqrt{\frac{2\log(2d)}{n}} + \frac{2}{\sqrt{n}}$$

So, we get that with probability $\ge 1 - \delta$,

$$\text{excess risk} \le 8DB\sqrt{\frac{2\log(2d)}{n}} + \frac{8}{\sqrt{n}} + 2(1 + DB)\sqrt{\frac{\log(1/\delta)}{2n}}$$

(e) We already proved this result in part (a).

(f) Let

$$\mathcal{G} := \{(x, y) \mapsto \ell(f(x), y) : \ f \in \mathcal{F}_1\}$$

and

$$\mathcal{H} := \{(x, y) \mapsto 1 - yf(x) : \ f \in \mathcal{F}_1\}$$

as we did in (d). We claim that $R_n^{\text{seq}}(\mathcal{H}) \le R_n^{\text{seq}}(\mathcal{F}_1)$. To see this, note that for all Z-valued trees $\mathbf{x}, \mathbf{y}$, we have

$$\begin{aligned}\hat{R}_n^{\text{seq}}(\mathcal{H}; (\mathbf{x}, \mathbf{y})) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}_1} \sum_{t=1}^n \epsilon_t(1 - \mathbf{y}_t(\epsilon) f(\mathbf{x}_t(\epsilon)))\right] \\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}_1} \sum_{t=1}^n -\epsilon_t \mathbf{y}_t(\epsilon) f(\mathbf{x}_t(\epsilon))\right],\end{aligned}$$

where to get to the second line we note that ϵ_t is 0 in expectation. Consider the mapping from $\{-1, 1\}^n \to \{-1, 1\}^n$ sending

$$\epsilon \mapsto \mathbf{s} := (-\epsilon_1 \mathbf{y}_1(\epsilon), \ldots, -\epsilon_n \mathbf{y}_n(\epsilon))$$

for our fixed value of $\mathbf{y}$. We claim that this is a bijection, which follows since we may iteratively construct the inverse mapping via $\epsilon_t = -s_t \mathbf{y}_t(\epsilon_{1:t-1})$, just as we did in the proof of Lemma 1 in Lecture 10. As such, we may construct a tree $\mathbf{x}'$ such that $\mathbf{x}'_t(\mathbf{s}) = \mathbf{x}_t(\epsilon)$ for all t. This, since the distribution over $\mathbf{s}$'s is the same as the distribution over ϵ's via the bijection,

$$\hat{R}_n^{\text{seq}}(\mathcal{H}; (\mathbf{x}, \mathbf{y})) = \frac{1}{n}\mathbb{E}_{\mathbf{s}}\left[\sup_{f \in \mathcal{F}_1} \sum_{t=1}^{n} s_t f(\mathbf{x}'_t(s))\right] = \hat{R}_n^{\text{seq}}(\mathcal{F}_1; \mathbf{x}')$$

Taking suprema over both sides, we get that

$$R_n^{\text{seq}}(\mathcal{H}) \leq R_n^{\text{seq}}(\mathcal{F}_1)$$

We note that $R_n^{\text{seq}}(\mathcal{G}) \leq R_n^{\text{seq}}(\mathcal{H}) \cdot \mathcal{O}(\log^{3/2}(n))$ by a bound akin to that of Lemma 2 in Lecture 10, where we are using that $\max\{0, \cdot\}$ is 1-Lipschitz. From Lecture 10, we therefore find that

$$\mathcal{V}^{\text{seq}}(\mathcal{F}_1, n) \leq 2R_n^{\text{seq}}(\mathcal{G}) \leq 2R_n^{\text{seq}}(\mathcal{F}_1) \cdot \mathcal{O}(\log^{3/2}(n)) \leq 2DB\sqrt{\frac{2\log(2d)}{n}} \cdot \mathcal{O}(\log^{3/2}(n))$$

So, we see that by definition of the value of a sequential game, letting $z_{1:n}$ denote the adversarial environment (chosen either obliviously or adversarially),

$$\inf_{\text{Alg}} \sup_{z_{1:n}} \mathbb{E}[\text{Reg}(\mathcal{F}_1, n)] \leq n\mathcal{V}^{\text{seq}}(\mathcal{F}_1, n) \leq \sqrt{n} \cdot \mathcal{O}\left(DB \log^{1/2}(2d) \log^{3/2}(n)\right)$$

There exists an algorithm that comes arbitrarily close to the infimum; in particular, for any $\delta > 0$ there must be an algorithm that achieves a regret bound of $\delta + \sqrt{n} \cdot \mathcal{O}\left(DB \log^{1/2}(2d) \log^{3/2}(n)\right)$. ∎

Problem 2

In this question, we consider a convex differentiable function f which is *not necessarily* smooth. Consider the Proximal Point Algorithm (PPA) with parameter ℓ, which has the following update equation:

$$x_{t+1} = \arg\min_x \left\{ f(x) + \frac{\ell}{2}\|x - x_t\|^2 \right\}$$

(a) Show that $x_{t+1} = x_t - \frac{1}{\ell}\nabla f(x_{t+1})$ and $f(x_{t+1}) \leq f(x_t)$.

(b) Prove that for any $t \in \mathbb{N}$, we have $f(x_{t+1}) - f(x^*) \leq \frac{\ell}{2}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$.

(c) Use the above results to prove the following theorem:

Theorem 1. *For any $\ell \geq 0$ and any convex function f, PPA with parameter ℓ satisfies:*

$$f(x_t) - f(x^*) \leq \mathcal{O}\left(\frac{\ell\|x_0 - x^*\|^2}{t}\right) \quad (\forall t > 0)$$

(d) Suppose now that f is α-strongly convex. Then, show that for any $t \in \mathbb{N}$, we have $\|x_{t+1} - x^*\|^2 \leq e^{-\alpha/(\ell+\alpha)}\|x_t - x^*\|^2$.

(e) Use the above results to prove the following theorem:

Theorem 2. *For any $\ell \geq 0$ and any α-strongly convex function f, PPA with parameter ℓ satisfies:*

$$f(x_t) - f(x^*) \leq \mathcal{O}\left(\ell\|x_0 - x^*\|^2 e^{-\alpha t/(\ell+\alpha)}\right) \quad (\forall t > 0)$$

Solution

Proof. We note that if $\ell = 0$ then $x_t = x^*$ for all t, and every desired result follows trivially. So, we suppose for everything below that $\ell > 0$. For this entire problem, we denote by f_t the function mapping $x \mapsto f(x) + \frac{\ell}{2}\|x - x_t\|^2$.

(a) Note that the function f_t is convex and differentiable, and so it has a unique global minimum precisely at the point x where $\nabla f_t(x) = 0$. We compute

$$\nabla f_t(x) = \nabla f(x) + \ell(x - x_t)$$

By definition of PPA and the above discussion, x_{t+1} will be the unique value of x for which this expression equals 0. So,

$$0 = \nabla f(x_{t+1}) + \ell(x_{t+1} - x_t) \implies x_{t+1} = x_t - \frac{1}{\ell}\nabla f(x_{t+1})$$

By convexity of f,

$$\begin{aligned} f(x_t) &\geq f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1} \rangle \\ &= f(x_{t+1}) + \left\langle \nabla f(x_{t+1}), \frac{1}{\ell}\nabla f(x_{t+1}) \right\rangle \\ &= f(x_{t+1}) + \frac{1}{\ell}\|\nabla f(x_{t+1})\|^2 \\ &\geq f(x_{t+1}) \end{aligned}$$

(b) Since f is convex, we see that f_t is ℓ-strongly convex. So, we get

$$f_t(x^*) - f_t(x_{t+1}) \geq \langle \nabla f_t(x_{t+1}), x^* - x_{t+1} \rangle + \frac{\ell}{2}\|x^* - x_{t+1}\|^2$$

Problem 2 continued on next page...

However, since x_{t+1} is the minimum of f_t by construction, we know that $\nabla f_t(x_{t+1}) = 0$. Thus,

$$f_t(x^*) - f_t(x_{t+1}) \geq \frac{\ell}{2}\|x^* - x_{t+1}\|^2$$

From this, we can compute

$$\begin{aligned}
f_t(x^*) - f_{t+1}(x^*) &= f_t(x^*) - f_t(x_{t+1}) + f_t(x_{t+1}) - f_{t+1}(x^*) \\
&\geq \frac{\ell}{2}\|x^* - x_{t+1}\|^2 + f(x_{t+1}) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 - f(x^*) - \frac{\ell}{2}\|x^* - x_{t+1}\|^2 \\
&= f(x_{t+1}) - f(x^*) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\
&\geq f(x_{t+1}) - f(x^*),
\end{aligned}$$

where we applied our earlier observation and plugged in the definitions of f_t and f_{t+1} to get the second line. From here, we plug in the definitions of f_t and f_{t+1} to see

$$f_t(x^*) - f_{t+1}(x^*) = f(x^*) + \frac{\ell}{2}\|x^* - x_t\|^2 - \left(f(x^*) + \frac{\ell}{2}\|x^* - x_{t+1}\|^2\right) = \frac{\ell}{2}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$$

Combining the above, we have shown that

$$f(x_{t+1}) - f(x^*) \leq \frac{\ell}{2}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$$

(c) Now, we may use the above results to prove the first theorem. Define $\delta_s := f(x_s) - f(x^*)$ for notation. From part (a), we saw that $\delta_t \leq \delta_s$ for all $s \leq t$. Summing over all $s \leq t$,

$$t\delta_t \leq \sum_{s=1}^{t} \delta_s$$

From part (b), we saw that $\delta_s \leq \frac{\ell}{2}(\|x_{s-1} - x^*\|^2 - \|x_s - x^*\|^2)$. Plugging this in,

$$t\delta_t \leq \frac{\ell}{2}\sum_{s=1}^{t}(\|x_{s-1} - x^*\|^2 - \|x_s - x^*\|^2) = \frac{\ell}{2}(\|x_0 - x^*\|^2 - \|x_t - x^*\|^2),$$

where we used the fact that this sum telescopes. Since $\|x_t - x^*\|^2 \geq 0$, we find that $t\delta_t \leq \frac{\ell\|x_0 - x^*\|^2}{2}$. Plugging in δ_t and dividing by t yields

$$f(x_t) - f(x^*) \leq \frac{\ell\|x_0 - x^*\|^2}{2t}$$

So, the result of the theorem holds for all $t \geq 1$.

(d) Suppose now that f is α-strongly convex. Therefore, f_t is $(\alpha + \ell)$-strongly convex. By strong convexity of f_t,

$$f_t(x^*) - f_t(x_{t+1}) \geq \langle \nabla f_t(x_{t+1}), x^* - x_{t+1}\rangle + \frac{\ell + \alpha}{2}\|x^* - x_{t+1}\|^2$$

Since x_{t+1} is the minimum of f_t by design, $\nabla f_t(x_{t+1}) = 0$, and so

$$f_t(x_{t+1}) - f_t(x^*) \leq -\frac{\ell + \alpha}{2}\|x^* - x_{t+1}\|^2$$

We know that $f(x_{t+1}) \leq f_t(x_{t+1})$, and so

$$f(x_{t+1}) - f(x^*) - \frac{\ell}{2}\|x^* - x_t\|^2 \leq f_t(x_{t+1}) - f_t(x^*) \leq -\frac{\ell + \alpha}{2}\|x^* - x_{t+1}\|^2$$

Also, we know that by optimality of x^*, it holds that $f(x_{t+1}) - f(x^*) \geq 0$. Thus,

$$-\frac{\ell}{2}\|x^* - x_t\|^2 \leq -\frac{\ell + \alpha}{2}\|x^* - x_{t+1}\|^2 \implies \|x^* - x_{t+1}\|^2 \leq \frac{\ell}{\ell + \alpha}\|x^* - x_t\|^2$$

Applying the fact that $1 - z \leq e^{-z} \quad \forall z \in \mathbb{R}$ with the value $z = \frac{\alpha}{\ell+\alpha}$ and noting that $1 - z = \frac{\ell}{\ell+\alpha}$, we find that

$$\|x_{t+1} - x^*\|^2 \leq e^{-\alpha/(\ell+\alpha)}\|x_t - x^*\|^2$$

as desired.

(e) From (b), we know that $f(x_{t+1}) - f(x^*) \leq \frac{\ell}{2}\|x_t - x^*\|^2$ since $\|x_{t+1} - x^*\|^2 \geq 0$. However, from repeated application of (d) we know that

$$\|x_t - x^*\|^2 \leq e^{-\alpha/(\ell+\alpha)}\|x_{t-1} - x^*\|^2 \leq \ldots \leq e^{-t\alpha/(\ell+\alpha)}\|x_0 - x^*\|^2$$

Taken together, these results show that

$$f(x_{t+1}) - f(x^*) \leq \frac{\ell\|x_0 - x^*\|^2}{2}e^{-t\alpha/(\ell+\alpha)}$$

The result of the theorem follows. ■