CS 540 DLT — Homework 4.

your NetID here.

Version 2.

Instructions.

- This homework is due **Wednesday, December 15, at 11:59pm**; no late homework accepted.
- You must work individually for this homework.
- Excluding office hours, and high-level discussions on discord, you may discuss with at most three other people; please state their NetIDs clearly on the first page of your submission.
- Homework must be typed, and submitted via gradescope. Please consider using the provided LaTeX file as a template.
- Each part of each problem is worth 3 points.
- For any problem asking you to construct something, for full credit you must always formally prove your construction works.
- General course and homework policies are on the course webpage.

Version history.

1. Initial version.

$1+\epsilon$. 3(b) now has $\ln 2$ rather than $\frac{1}{\ln 2}$.

2. 4(b,c,d) now have $1/\sqrt{n}$, instead of $1/n$ ☹; in particular, they now match 3(b).
   Additionally, $\operatorname{sgn}(\cdot)$ was added in 3(b).

1. Concentration of NTK eigenvalues.

In hw2, we established that the *infinite-width* Gram matrix from the NTK lectures has positive minimum eigenvalue. In this problem, we will prove that the finite width Gram matrix has similar (positive) eigenvalues. This settles the remaining pieces necessary to make the NTK gradient flow proof concrete.

Throughout this problem, let examples $(x_i)_{i=1}^n$ be given and fixed, and collect all examples as rows of a matrix $X \in \mathbb{R}^{n\times d}$. Suppose for simplicity the activation σ is differentiable, and $|\sigma'| \leq B$.

Define the sampled Gram matrix $\widehat{G} \in \mathbb{R}^{n\times n}$ and expected gram matrix $G \in \mathbb{R}^{n\times n}$ via

$$\widehat{G}_{ij} := \frac{1}{m}\sum_{k=1}^m x_i^\intercal x_j \sigma'(w_k^\intercal x_i)\sigma'(w_k^\intercal x_j), \qquad G_{ij} := \mathbb{E}_w x_i^\intercal x_j \sigma'(w^\intercal x_i)\sigma'(w^\intercal x_j).$$

In this problem, the random draw is over the weights $(w_j)_{j=1}^m$, not over the examples. Consequently, it is useful to define *another* family of random matrices: $(H_k)_{k=1}^m$, with

$$(H_k)_{ij} := x_i^\intercal x_j \sigma'(w_k^\intercal x_i)\sigma'(w_k^\intercal x_j), \qquad \text{whereby } \widehat{G} = \frac{1}{m}\sum_{k=1}^m H_k.$$

The approach of this problem is to apply Rademacher complexity. Let $W = (w_1, \ldots, w_m)$ denote the full random draw, analogous to the data in our usual applications of Rademacher complexity.

We will use matrix inner products:

$$\langle H_k, V\rangle = \operatorname{tr}\left(H_k^\intercal V\right), \qquad \text{and} \qquad \left|\langle H_k, V\rangle\right| \leq \|H_k\|_{\mathrm{F}} \cdot \|V\|_{\mathrm{F}}.$$

(a) Prove $\|H_k\|_{\mathrm{F}} \leq B^2\|X\|_{\mathrm{F}}^2$.

(b) Define

$$\mathcal{F} := \left\{U \mapsto \langle U, V\rangle : \|V\|_{\mathrm{F}} \leq 1\right\}, \qquad \mathcal{H} := (H_1, \ldots, H_m).$$

The relevant (unnormalized) Rademacher complexity for us is

$$\mathrm{URad}(\mathcal{F}_{|\mathcal{H}}) = \mathrm{URad}\left(\left\{(\langle H_1, V\rangle, \ldots, \langle H_m, V\rangle) : \|V\|_{\mathrm{F}} \leq 1\right\}\right).$$

Prove $\mathrm{URad}(\mathcal{F}_{|\mathcal{H}}) \leq B^2\|X\|_{\mathrm{F}}^2\sqrt{m}$.

(c) Prove $\{vv^\intercal : \|v\|_2 \leq 1\} \subseteq \{V : \|V\|_{\mathrm{F}} \leq 1\}$.

(d) Prove that with probability at least $1-\delta$ over the draw of $(w_1, \ldots, w_m)$, simultaneously for every $\|u\|_2 \leq 1$,

$$\left|u^\intercal G u - u^\intercal \widehat{G} u\right| \leq \frac{2B^2\|X\|_{\mathrm{F}}^2}{\sqrt{m}} + 6B^2\|X\|_{\mathrm{F}}^2\sqrt{\frac{\ln(4/\delta)}{2m}} =: \star.$$

(e) Prove that, with probability at least $1-\delta$, simultaneously

$$\lambda_{\min}(\widehat{G}) \geq \lambda_{\min}(G) - \star \qquad \text{and} \qquad \lambda_{\max}(\widehat{G}) \leq \lambda_{\max}(G) + \star,$$

where $\star$ is the right hand side of the bound in the previous part.

Remark. The point is that we can make $\star$ as small as we want just by increasing m.

Solution. *(If using this template, please write your solution here.)*

2. Shallow Rademacher bound near initialization.

Throughout this problem, consider our usual shallow NTK setup:

$$F(x;W) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma_{\mathrm{r}}(w_j^{\intercal} x),$$

where $\sigma_{\mathrm{r}}(z) = \max\{0, z\}$ is the ReLU, $s_j \in \{-1, +1\}$ is fixed and independent of the data, $W \in \mathbb{R}^{m \times d}$ with jth row $w_j^{\intercal}$, and W will always be close to a fixed (and independent of the data) $W_0 \in \mathbb{R}^{m \times d}$, whose jth row is $w_{0,j}^{\intercal}$.

Let $S = ((x_i, y_i))_{i=1}^n$ denote a data sample, and let $X \in \mathbb{R}^{n \times d}$ be a matrix whose ith row is $x_i^{\intercal}$.

(a) Show that for any $j \in \{1, \dots, m\}$ and any $r \geq 0$,

$$\mathrm{URad}\left(\left\{ x \mapsto s_j \sigma_{\mathrm{r}}(w_j^{\intercal} x) : \|w_j - w_{0,j}\|_2 \leq r \right\}_{|S} \right) \leq r \|X\|_{\mathrm{F}}.$$

(b) Show that for any $r \geq 0$,

$$\mathrm{URad}\left(\left\{ x \mapsto F(x;W) : \|(W - W_0)^{\intercal}\|_{2,\infty} \leq \frac{r}{\sqrt{m}} \right\}_{|S} \right) \leq r \|X\|_{\mathrm{F}},$$

where $\|(W - W_0)^{\intercal}\|_{2,\infty} = \max_j \|w_j - w_{0,j}\|_2$.

(c) Let $\ell(z) = \ln(1 + \exp(-z))$ denote the logistic loss, define a truncation $\tilde{\ell}(z) = \min\{\ell(z), 1\}$, and suppose $\|x\| \leq 1$. Show that with probability at least $1 - \delta$ over the draw of S, every W with $\|(W - W_0)^{\intercal}\|_{2,\infty} \leq r/\sqrt{m}$ satisfies

$$\mathbb{E}_{x,y} \tilde{\ell}(yF(x;W)) \leq \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(y_i F(x_i;W)) + \frac{2r}{\sqrt{n}} + 3\sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}.$$

Remark. What happens if ℓ is used in place of $\tilde{\ell}$?

Solution. *(If using this template, please write your solution here.)*

3. "Rethinking" generalization.

The "rethinking generalization" paper contains many claims, but one is that Rademacher complexity can not be applied to neural networks, since they are universal approximators. This problem will study this claim.

(a) Let a sequence of function classes $(\mathcal{F}_i)_{i\geq 1}$ be given, and define $\mathcal{F} := \cup_{i\geq 1}\mathcal{F}_i$. Suppose that for each $\mathcal{F}_i$, we have a standard generalization bound: there exist constants a_i and b_i so that with probability at least $1-\delta$, for every $f \in \mathcal{F}_i$ simultaneously,

$$\mathcal{R}(f) \leq \widehat{\mathcal{R}}(f) + \frac{a_i}{\sqrt{n}} + b_i\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}.$$

Show that with probability at least $1-\delta$, for every $f \in \mathcal{F}$ simultaneously,

$$\mathcal{R}(f) \leq \widehat{\mathcal{R}}(f) + \inf_{\substack{i\geq 1\\ f\in\mathcal{F}_i}}\left[\frac{a_i}{\sqrt{n}} + b_i\sqrt{\frac{\ln(i+1)}{n}} + b_i\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right].$$

(b) Suppose the notation (and bounds) of the previous problem: specifically, a shallow ReLU network prediction mapping $x \mapsto F(x;W)$, where $W \in \mathbb{R}^{m\times d}$ is a parameter matrix, and there is some fixed $W_0 \in \mathbb{R}^{m\times d}$ which is independent of the data, and lastly ℓ and $\tilde{\ell}$ are respectively the regular and truncated logistic losses. Show that, with probability at least $1-\delta$, every $W \in \mathbb{R}^{m\times d}$ simultaneously satisfies

$$\begin{aligned}\ln(2)\cdot\Pr\left[\operatorname{sgn}(F(x;W)) \neq y\right] &\leq \mathbb{E}\tilde{\ell}(yF(x;W))\\ &\leq \frac{1}{n}\sum_{i=1}^n \ell(y_iF(x_i;W)) + \frac{2\sqrt{m}\left(1+\|(W-W_0)^\intercal\|_{2,\infty}\right)}{\sqrt{n}}\\ &\quad + 3\sqrt{\frac{\ln\left(2+\|(W-W_0)^\intercal\|_{2,\infty}\right)}{n}} + 3\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}.\end{aligned}$$

Remark. The intuitive difference between this bound and the one in the previous problem is that we could magically guess the appropriate value of r.

(c) Returning to the "rethinking generalization" paper, since $\operatorname{URad}(\{\pm 1\}^n) = n$ and since neural networks are universal approximators, it is suggested that Rademacher bounds for neural networks are always at least 1, and thus vacuous.

Does the bound in the previous part provide a resolution to this? Why, or why not? Use 1-5 sentences for your answer.

Solution. *(If using this template, please write your solution here.)*

4. Generalization experiments near initialization.

This problem will check some basic generalization properties near initialization. Starter code is provided in `hw4.py`: it is nearly the same as `hw3.py`, but additionally it fixes a test/train split. This problem will once again run gradient descent; please refer to the relevant discussion in `hw3.py`.

(a) Using the provided shallow ReLU network class, plot the *empirical logistic risk over the test set minus the empirical logistic risk over the training set* for 4096 gradient descent iterations (on the training set) with step size 1.0 and widths $m \in \{64, 256, 1024\}$. (Only difference with last time is that $m = 4$ has been dropped.)

For full points, include the plot in your hand-in, and describe it qualitatively in 1-3 sentences.

Remark. If the terminology is unclear, see `hw4.py` for the relevant treatment of the (linear) logistic regression case.

(b) For the same setup as the previous problem, plot the main term of the previous generalization bound: namely, plot

$$\frac{\sqrt{m}\|(W_t - W_0)^\intercal\|_{2,\infty}}{\sqrt{n}}$$

along the vertical axis for all three widths, with iteration counter t running along the horizontal axis.

For full points, include the plot in your hand-in, and describe it qualitatively in 1-3 sentences, including a comparison to the observed excess risk from the previous part.

(c) For the same setup as the previous problem, let's drop the "$-W_0$" term: plot

$$\frac{\sqrt{m}\|W_t^\intercal\|_{2,\infty}}{\sqrt{n}}$$

along the vertical axis for all three widths, with iteration counter t running along the horizontal axis.

For full points, include the plot in your hand-in, and describe it qualitatively in 1-3 sentences, including a comparison to the previous bound and to the observed excess risk.

(d) Lastly, let's plot the dominant term from the Golowich-Rakhlin-Shamir bound given in lecture: specifically, since the outer layer signs satisfy $\|\vec{s}\|_2 = \sqrt{m}$, plot

$$\frac{\sqrt{m}\|W_t\|_{\mathrm{F}}}{\sqrt{n}}$$

along the vertical axis for all three widths, with iteration counter t running along the horizontal axis.

For full points, include the plot in your hand-in, and describe it qualitatively in 1-5 sentences, including a comparison to the previous bounds and to the observed excess risk.

Solution. *(If using this template, please write your solution here.)*

5. Course feedback.

You receive full credit for this question so long as you write at least one sentence for each answer. Please be honest and feel free to be critical.

(a) What are some topics you wish had been covered?

(b) What are some things the instructor can improve for next time?

Solution. *(If using this template, please write your solution here.)*