UC Davis, STA 250
Homework 1
Instructor: Spencer Frei

Version 1.0, released Tuesday, January 16, 2024.

Problem 1

In this problem, we consider a two-layer leaky ReLU network trained by gradient descent on the first-layer weights. Let $m \in \mathbb{N}$, $\phi(t) = \max(t, \gamma t)$ for $\gamma \in (0, 1]$, let $W \in \mathbb{R}^{m \times d}$ have rows $w_j^\top$, and let $a_j \in \{\pm 1/\sqrt{m}\}$ (the a_j can take arbitrary values in this set). Consider

$$f(x; W) := \sum_{j=1}^{m} a_j \phi(\langle w_j, x\rangle).$$

Let us assume that $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ are such that $\|x_i\| \leq 1$ for each i, and there exists $v \in \mathbb{R}^d$ such that $y_i \langle v, x_i\rangle \geq 1$ for all i. Let

$$\widehat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i; W)).$$

Let $\alpha > 0$ be a step size, and consider gradient descent on the logistic loss $\ell(t) = \log(1 + \exp(-t))$,

$$W^{(t+1)} = W^{(t)} - \alpha \nabla \widehat{L}(W^{(t)}).$$

In this problem, we will show that although $\widehat{L}(W)$ is not smooth, we can still show convergence of gradient descent using what is known as a "Perceptron-style" proof. This is so-named because of its similarity to the proof of convergence of the Perceptron algorithm for learning halfspaces with linear classifiers (see, e.g., Theorem 9.1 of Shalev-Shwartz and Ben-David's book.)

1. Show that $\widehat{L}(W)$ is not necessarily β-smooth.

2. Show that there exists $V \in \mathbb{R}^{m \times d}$ satisfying $\|V\|_F = 1$ and $c > 0$ such that for any training point (x_i, y_i) and for any $W \in \mathbb{R}^{m \times d}$, we have

$$y_i \langle \nabla f(x_i; W), V\rangle \geq c.$$

Hint: it suffices to take a matrix V where every row is a multiple of a single vector.

3. Let $H_t := \langle W^{(t)}, V \rangle$ be the correlation between the weights found by G.D. and the matrix V from the previous part of the problem, and let

$$\widehat{G}(W) := \frac{1}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; W)).$$

Show that there exists $c' > 0$, independent of α, such that for any $t \geq 0$,

$$H_{t+1} - H_t \geq c' \alpha \widehat{G}(W^{(t)}).$$

Hint: use that ℓ is Lipschitz and decreasing.

4. Let $F_t := \|W^{(t)}\|_F$. Show that $F_{t+1}^2 \leq F_t^2 + 2\alpha + \alpha^2$ for any $t \geq 0$.

Hint: use that ϕ is 1-homogeneous.

5. Use the above to conclude that for any $\varepsilon > 0$, there exists a finite $T = T(\varepsilon, m, \gamma, \alpha)$ for which $\widehat{G}(W^{(T)}) \leq \varepsilon$.

Hint: Consider how quickly the quantity $H_t^2 := \langle W^{(t)}, V \rangle^2$ grows as t increases, and use Cauchy–Schwarz.

6. Use this to conclude that for any $\varepsilon > 0$, there exists a finite $T = T(\varepsilon, m, \gamma, \alpha)$ for which $\widehat{L}(W^{(T)}) \leq \varepsilon$. What are the conditions on α under which this result holds?

Problem 2

Let $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ for $i = 1, \ldots, n$; call $S = \{(x_i, y_i)\}_{i=1}^n$. Let $R_{\mathsf{min}}^2 := \min_i \|x_i\|^2$ and $R_{\mathsf{max}}^2 := \max_i \|x_i\|^2$ and $R^2 := R_{\mathsf{max}}^2 / R_{\mathsf{min}}^2$, and assume $R_{\mathsf{min}} > 0$. Let us call the training dataset *p-orthogonal* if,

$$R_{\mathsf{min}}^2 \geq p R^2 n \max_{i \neq j} |\langle x_i, x_j \rangle|.$$

In particular, if the examples x_i are exactly orthogonal, then S is p-orthogonal for every $p > 0$.

Recall the definition of the ℓ_2-max margin solution (MM) and the ℓ_2-minimum norm interpolator (MNI)

$$\begin{aligned} w_{\mathsf{MM}} &:= \operatorname{argmin}\{\|w\|_2^2 : w \in \mathbb{R}^d,\ y_i \langle w, x_i \rangle \geq 1 \text{ for all } i = 1, \ldots, n\}, \\ w_{\mathsf{MNI}} &:= \operatorname{argmin}\{\|w\|_2^2 : w \in \mathbb{R}^d,\ \langle w, x_i \rangle = y_i \text{ for all } i = 1, \ldots, n\}. \end{aligned}$$

1. Suppose that $x_i \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$. For $\delta \in (0, 1/2)$, state sufficient conditions under which we can guarantee that the training dataset S is p-orthogonal with probability at least $1 - \delta$.

2. Show that if S is p-orthogonal for some $p \geq 3$, then w_{MM} exists and $w_{\mathsf{MM}} = w_{\mathsf{MNI}}$. What does this imply about training on the logistic loss vs. training on the squared loss when the training data is p-orthogonal?

3. Show that there exist training datasets S for which $w_{\mathsf{MNI}} \neq w_{\mathsf{MM}}$.

4. Show that if S is p-orthogonal for some $p \geq 3$, then there exist $s_i > 0$ such that $w_{\mathsf{MM}} = \sum_{i=1}^n s_i y_i x_i$ and the s_i satisfy $\max_{i,j} s_i / s_j \leq R^2 \left(1 + \frac{1}{\Omega(pR^2)}\right)$. In particular, if p is large and the norms of the examples are close to each other, the max-margin classifier is approximately proportional to the uniform average of the training data, $\sum_{i=1}^n y_i x_i$.

Problem 3

Let us again consider the training of a two-layer leaky ReLU network $f(x; W)$ by gradient descent on the logistic loss training only the first-layer weights (the setting of Problem 1). We shall show a partial result concerning the implicit bias of gradient descent towards rank minimization in neural networks when the training data is p-orthogonal. Towards this end, for a matrix $M \in \mathbb{R}^{m\times d}$, let us recall the definition of the Frobenius norm and spectral norm:

$$\|M\|_F^2 := \sum_{i,j} ([M]_{i,j})^2, \quad \|M\|_2 := \sup_{\|v\|_2=1} \|Mv\|_2.$$

We define the *stable rank* of M as

$$\mathsf{StableRank}(M) := \frac{\|M\|_F^2}{\|M\|_2^2}.$$

The stable rank is a continuous version of the rank of a matrix. Consider, e.g., $M \in \mathbb{R}^{d\times d}$ with $M = \mathrm{diag}(1, \ldots, 1, \varepsilon)$ for $\varepsilon \in [0, 1]$. For any $\varepsilon > 0$, the rank of M is d, while for $\varepsilon = 0$ the rank abruptly changes to $d - 1$. On the other hand, $\mathsf{StableRank}(M)$ smoothly changes from $d - 1$ to d as ε goes from 0 to 1. Similarly, if $M = \mathrm{diag}(1, \exp(-d), \ldots, \exp(-d))$, then the rank of M is equal to d for all d, while $\mathsf{StableRank}(M) = 1 + (d - 1)\exp(-2d) = 1 + o_d(1)$.

1. Suppose that $[W^{(0)}]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$ for some $\sigma > 0$. A classical result in random matrix theory states the following.[1] For some $c > 0$ and for any $t \geq 0$,

   $$\mathbb{P}(\sigma^{-1}\|W^{(0)}\|_2 \geq \sqrt{m} + \sqrt{d} + t) \leq 2\exp(-ct^2).$$

   Use this to show that with probability at least $1 - o_d(1)$, $\mathsf{StableRank}(W^{(0)}) \geq \Omega(\min(m, d))$.

2. Suppose that the training data is p-orthogonal, and consider $W^{(1)} = W^{(0)} - \alpha\nabla\widehat{L}(W^{(0)})$ as in Problem 1, where $[W^{(0)}]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$. Show that if p is sufficiently large, then there exists some $\underline{\alpha}, \bar{\alpha} > 0$, $\bar{\sigma} > 0$, such that for $\underline{\alpha} \leq \alpha \leq \bar{\alpha}$ and $0 < \sigma \leq \bar{\sigma}$, it holds that $\mathsf{StableRank}(W^{(1)}) \leq C$ for some universal constant C which is independent of m and d. In particular, gradient descent reduces the stable rank of the weight matrix from order $\Omega(\min(m, d))$ to constant order in one step.

   Hint 1: You need to prove an upper bound on $\|W^{(1)}\|_F^2$ and a lower bound on $\|W^{(1)}\|_2^2$, and show they are within a constant of one another. The proof of both bounds should explicitly use the fact that the training data is p-orthogonal; you may find some of the proof ideas from Problem 1 helpful.

   Hint 2: By taking σ sufficiently small, the approximation $W^{(1)} \approx -\alpha\nabla\widehat{L}(W^{(0)})$ holds; see what happens if you treat this as an equality.

3. Consider training a two-layer leaky ReLU network, with biases, on the cross-entropy loss with $\gamma = 0.05$ and $m = 150$ neurons for the MNIST classification task. (Unlike in Problem 1 and the above subproblem, we are now considering training on both layers and with bias terms.) Initialize the network with i.i.d. mean zero Gaussians with standard deviation $\sigma = 0.02$. Find a suitable learning rate such that you can produce a network which achieves less than 5% training error within 20 minutes of training on your laptop/Google Colab; call $W^{(T)}$ the weights found at the end. Now examine what

[1] See, e.g., Corollary 7.3.3 of Vershynin's *High-Dimensional Probability*.

happens when you train with the same learning rate and for the same number of steps T as you vary σ so that $\sigma \in \{0.0002, 0.002, 0.02, 0.2, 2\}$.

Produce a plot with the following characteristics:

- σ on the x-axis,
- For each $t \in \{1, T/10, T/5, T/2, T\}$, have a curve with values $\frac{\mathsf{StableRank}(W^{(t)})}{\mathsf{StableRank}(W^{(0)})}$ as a function of σ, i.e. the relative rank of the weights at time t vs. at time 0. In particular, there should be 5 separate curves, with different colors and line styles, for each of the times $t \in \{1, T/10, T/5, T/2, T\}$, so each curve corresponds to the relative rank decrease as a function of the number of gradient descent steps. Are there any noteworthy findings?