Homework 3
CS229T/STATS231 (Fall 2018–2019)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. **Up to two additional points will awarded for especially well-organized and elegant solutions.**

Due date: Wednesday, Nov 28th, 2018, 11pm

1. Concentration of the landscape on generalized linear models (*20 points*) We consider the problem of learning a *generalized linear model*: suppose we observe data $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where x_i are sampled i.i.d. from some distribution P, and

$$y_i = \sigma(w_\star^\top x_i) + \epsilon_i,$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a known activation function, $\epsilon_i \in \mathbb{R}$ are i.i.d. mean-zero noise (independent with x_i), and $w_\star \in \mathbb{R}^d$ is a fixed unknown ground truth coefficent vector that we wish to recover.

Towards learning $w_\star$, we minimize the squared loss

$$\hat{L}(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sigma(w^\top x_i))^2.$$

Let $L(w) = \mathbb{E}_{x,y}[\frac{1}{2}(y - \sigma(w^\top x))^2]$ be the corresponding expected loss.

General goal of the problem: In this problem, you will be asked to work through a series of steps to demonstrate the following phenomenon: a) all the local minima of the expected loss are global minima in the squared loss objective for the generalized linear model; b) the training loss $\hat{L}$ has the same property.

Assumptions We make the following assumptions on the problem.

(1) The vectors x_i are bounded and non-degenerate: P is supported in the ball $\{x : \|x\|_2 \le B\}$, and $\mathbb{E}_{x\sim P}[xx^\top] \succeq \lambda I_{d\times d}$ for some $\lambda > 0$. Here $I_{d\times d}$ is the identity matrix of dimension $d \times d$.

(2) The ground truth coefficient vector satisfies $\|w_\star\|_2 \le R$, and $BR \ge 1$. (B is defined in (1).)

(3) The activation function σ is strictly increasing and twice differentiable. Furthermore, it satisfies the bounds

$$\sigma(t) \in [0, 1], \quad \sup_{t\in\mathbb{R}} \{|\sigma'(t)|, |\sigma''(t)|\} \le 1, \quad \text{and} \quad \inf_{t\in[-BR,BR]} \sigma'(t) \ge \gamma > 0.$$

(4) The noise ϵ_i's are mean zero and bounded: with probability 1, we have $|\epsilon_i| \le 1$.

(5) The sample size is at least the dimension: $n \ge d$.

In addition to these, you may assume without proof that the gradient operator and the expectation can always be exchanged. For example, you can directly use the equality $\nabla L(w) = \nabla \mathbb{E}[\hat{L}(w)] = \mathbb{E}[\nabla \hat{L}(w)]$.

Part I: Property of the landscape of the expected loss

We begin by demonstrating certain properties on the landscape of the expected loss $L(w)$, showing that it has a certain nice properties even though it may be non-convex.

a. (*1 point*) (non-convexity of the expected loss) Construct an example where the expected loss $L(w)$ is not a convex function of w. You can feel to work with the $d = 1$ case, and pick a particular σ, the

distribution P and the distribution of noise which satisfy the assumptions above. (Empirical demonstrations such as plotting the function and numerically computing the second order derivatives are also allowed, though the course staff suspect that it's not the simplest way to solve the problem.)

b. (*4 points*) (all stationary points of the expected loss are global minima)
Show that $w_\star$ is a global minimum of L. In addition, show that

$$\langle \nabla L(w), w - w_\star \rangle \geq \gamma^2 \lambda \cdot \|w - w_\star\|_2^2 \quad \text{for all } w \text{ such that } \|w\|_2 \leq R.$$

As a corollary of the inequality above, show that if w is a stationary point of L (i.e., $\nabla L(w) = 0$) and $\|w\|_2 \leq R$, then w is a global minimum of L.

Remark: The inequality above implies that the gradient descent on the expected loss will converge to the ground-truth $w_\star$ in polynomial time. Loosely speaking, the reason is that at any point w that is not the global minimum $w_\star$, the negative gradient direction $-\nabla L(w)$ is correlated with the ideal direction of movement $w_\star - w$. You are not required to prove this.

Part II: Property of the landscape of the training loss

We now analyze the training loss $\hat{L}(w)$. The goal of this part is to prove that the training loss has no bad local minimum, as stated in the following theorem:

Theorem 1 (Training loss has no bad local minima)**.** *Under the problem assumptions, we have the "*concentration of directional gradients*" in the sense that: with probability at least $1 - \delta$, for all $w \in \{w : \|w\|_2 \leq R\}$, we have*

$$\langle \nabla \hat{L}(w) - \nabla L(w), w - w_\star \rangle \leq C_1 \cdot B \sqrt{\frac{d(C_2 + \log(nBR)) + \log \frac{1}{\delta}}{n}} \cdot \|w - w_\star\|_2,$$

where $C_1, C_2 > 0$ are universal constants that do not depend on (B, R, d, n, δ). Conditioned on the event above, the training loss has no local minima outside a small neighborhood of $w_\star$: for any w such that $\|w\|_2 \leq R$, if $\nabla \hat{L}(w) = 0$, then

$$\|w - w_\star\|_2 \leq \frac{C_1 B}{\gamma^2 \lambda} \sqrt{\frac{d(C_2 + \log(nBR)) + \log \frac{1}{\delta}}{n}}.$$

Remark: Theorem 1 shows that all stationary points of $\hat{L}(w)$ has to be within a small neighborhood of $w_\star$, but there is still the question of whether there are multiple local minima near $w_\star$. While we won't go into detail on that, it can be resolved by proving a similar concentration of the Hessian $\nabla^2 \hat{L}(w)$ under additional assumptions. The concentration of the Hessian can imply that with high probability, $\hat{L}(w)$ is strongly convex in a neighborhood of $w_\star$, and thus the stationary point (or local minimum) has to be unique. (You are not required to prove this.)

Towards proving Theorem 1, you are allowed to choose one of the following two options. In option 1, you are asked to directly prove Theorem 1. Full credit will be given for a correct proof. Partial credits for intermediate steps will be weighted by their relative importance in a similar fashion as in option 2, which we describe below. In option 2, you are asked to prove a sequence of parts involving several tools and intermediate steps, and use them to establish Theorem 1 in the last part. We allow two options for you because we are aware of multiple potential routes to proving Theorem 1.

Option 1 (*15 points*): Prove Theorem 1.

Option 2 (*15 points*): Do the following parts.

c. (*2 points*) **(upper bounds on the Hessians)** We start by bounding the operator norm of the Hessian as a preparation. Show that

$$\|\nabla^2 L(w)\|_{\mathrm{op}} \le 2B^2 \quad \text{and} \quad \|\nabla^2 \hat{L}(w)\|_{\mathrm{op}} \le 3B^2.$$

Full credit will be given for upper bounds of the form CB^2 for a universal constant $C > 0$. Recall that B is defined in the problem assumptions. As a consequence (which you don't need to argue), for any $w_1, w_2 \in \mathbb{R}^d$, we have the Lipschitzness of gradients

$$\|\nabla L(w_1) - \nabla L(w_2)\|_2 \le 2B^2\|w_1 - w_2\|_2 \quad \text{and} \quad \|\nabla \hat{L}(w_1) - \nabla \hat{L}(w_2)\|_2 \le 3B^2\|w_1 - w_2\|_2.$$

Hint: you may use the fact that for any symmetric matrix A, $\|A\|_{\mathrm{op}} = \sup_{\|v\|_2 \le 1} v^\top A v$.

d. (*1 point*) **(norm bound via covering)** We now establish a useful tool for showing the concentration of gradients in sequel. Let $N(\mathcal{B}(1), 1/2)$ be a $\frac{1}{2}$-covering of the unit ball $\mathcal{B}(1) = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ with respect to ℓ_2 norm. Recall that we have the bound $|N(\mathcal{B}(1), 1/2)| \le 5^d$ (by hw1 2.b). Show that for any vector $x \in \mathbb{R}^d$,

$$\|x\|_2 \le 2 \sup_{v \in N(\mathcal{B}(1), 1/2)} \langle x, v \rangle.$$

e. (*5 points*) **(point-wise concentration of gradients)** For any $w \in \mathbb{R}^d$ and $t > 0$, show that

$$\mathbb{P}(\|\nabla \hat{L}(w) - \nabla L(w)\|_2 \ge t) \le \exp\left(-\frac{nt^2}{32B^2} + d \log 5\right).$$

Here the probability is over the randomness of the training examples (that are used to define $\hat{L}$.)

Full credit will be given for bounds of the form $\exp(-c_1 nt^2/B^2 + c_2 d)$ for universal constants $c_1, c_2 > 0$. *Hint*: You may use part (d) to upper bound the 2-norm of any vector by its max inner product with $v \in N(1/2)$.

f. (*1 point*) **(union bounds)** Now, consider the ball $\mathcal{B}(R) = \{w : \|w\|_2 \le R\}$, which we know contains the ground truth $w_\star$. Show that for any $0 < \epsilon < R$, we can take $N(\mathcal{B}(R), \epsilon)$ to be an ϵ-covering of $\mathcal{B}(R)$ and get

$$\mathbb{P}\left(\sup_{w \in N(\mathcal{B}(R), \epsilon)} \|\nabla \hat{L}(w) - \nabla L(w)\|_2 \ge t\right) \le \exp\left(-\frac{nt^2}{32B^2} + d \log \frac{15R}{\epsilon}\right).$$

Full credit will be given for bounds of the form $\exp(-c_1 nt^2/B^2 + (c_2 + \log(R/\epsilon))d)$ for universal constants $c_1, c_2 > 0$.

g. (*5 points*) **(uniform convergence of gradients)** Combining part (c) and (f), by performing a discretization argument with a suitable ϵ, show that with probability at least $1-\delta$ (over the randomness of (x_i, y_i)'s),

$$\sup_{w \in \mathcal{B}(R)} \|\nabla \hat{L}(w) - \nabla L(w)\|_2 \leq C_1 \cdot B \sqrt{\frac{d(C_2 + \log(nBR)) + \log \frac{1}{\delta}}{n}},$$

where $C_1, C_2 > 0$ are universal constants. Conditioned on the event above, show that for all $w \in \mathcal{B}(R)$, we have

$$\langle \nabla \hat{L}(w) - \nabla L(w), w - w_\star \rangle \leq C_1 \cdot B \sqrt{\frac{d(C_2 + \log(nBR)) + \log \frac{1}{\delta}}{n}} \cdot \|w - w_\star\|_2.$$

h. (*1 point*) **(the training loss has no bad local minima)** The uniform convergence of projected gradients in (g) combined with the property of the population gradient in (b) gives a concrete result that the training loss $\hat{L}$, though non-convex, has no stationary points outside of a certain neighborhood of $w_\star$. (As a consequence, many local search algorithms such as gradient descent on $\hat{L}$ will converge into the neighborhood. You are not required to show this.)

Building on part g, show that with probability at least $1-\delta$ (over the randomness of (x_i, y_i)'s), the training loss has no local minima outside a small neighborhood of $w_\star$: for any w such that $\|w\|_2 \leq R$, if $\nabla \hat{L}(w) = 0$, then

$$\|w - w_\star\|_2 \leq \frac{C_1 B}{\gamma^2 \lambda} \sqrt{\frac{d(C_2 + \log(nBR)) + \log \frac{1}{\delta}}{n}}.$$

This part completes the proof of Theorem 1.

2. Varying step sizes (*20 points*)

(Note that this problem requires knowledge that will be covered in class in the week 7 and 8. Therefore we suggest you solve it after Wed of Week 8.)

Choosing step sizes properly is crucial in optimization and can have a dramatic impact on performance (as you probably saw in the previous problem). The version of online gradient descent (OGD) that we studied in class uses the same step size η on each iteration. Could we do better if we allowed the learning algorithm to change the step size? In this problem, we will devlop an analysis conducive to answering this question, and then use it to obtain bounds that inform us on how to choose a good step size schedule.

In this problem, we will develop a single analysis that lets us simultaneously explore two advantages of varying step sizes: (i) taking advantage of strong convexity to obtain better regret bounds, and (ii) setting the step size without having to know T in advance.

Let $S \subseteq \mathbb{R}^d$ be a convex set of experts, and let $f_1, \ldots, f_T$ be a sequence of loss functions. We assume each f_t is strongly convex w.r.t to ℓ_2 norm; that is, there exists a $K_t \geq 0$ (called the strong convexity parameter) such that for all $w \in S$ and any subgradient $z_t \in \partial f_t(w)$, the following holds for all $u \in S$:

$$f_t(u) \geq f_t(w) + z_t \cdot (u - w) + \frac{K_t}{2} \|u - w\|_2^2. \tag{1}$$

Intuitively, f_t is lower bounded by not just the usual linearization $z_t \cdot (u - w)$, but an additional quadratic.

We will consider the following OGD update based on eager projection (as opposed to lazy projection that we discussed in class; make sure that you see the difference between the two):

$$w_{t+1} = \Pi_S \left(w_t - \eta_t z_t \right), \quad z_t \in \partial f_t(w_t), \tag{2}$$

where Π_S is the projection onto S and $w_1 = 0$. Note that the step size η_t depends on t. Let's figure out how to set it properly to obtain low regret. In the following, let $u \in S$ be any expert.

Remark (Subgradients): Though the problem considers the general setting of subgradients, it's okay if you only work with differentiable functions, where the subgradient is a singleton consisting of the gradient: $\partial f_t(w) = \{\nabla f_t(w)\}$.

Remark/Hint: To some extent, this question will guide you to go through a route of analyzing online learning algorithms that is different from (but related to) what was presented in the class. Therefore, the high-level tools taught in the class (such as BTL lemma) may not be directly useful for solving the question (whereas the low-level tools such as properties of strongly convex functions are certainly necessary.)

a. (*4 points*) Let $\text{Regret}(u)$ denote the excess loss of the algorithm with respect to the expert u:

$$\text{Regret}(u) \stackrel{\text{def}}{=} \sum_{t=1}^{T} [f_t(w_t) - f_t(u)].$$

Prove the following regret bound:

$$\text{Regret}(u) \leq \sum_{t=1}^{T} \left[\frac{1}{2\eta_t} \left(\|w_t - u\|_2^2 - \|w_{t+1} - u\|_2^2 \right) + \frac{1}{2}\eta_t \|z_t\|_2^2 - \frac{K_t}{2}\|w_t - u\|_2^2 \right]. \tag{3}$$

Remark: The bound has a form similar to the one presented in class. The first term in the summand measures the size of u (though the varying step size makes the expression more complex), and the second term measures the size of z_t. Finally, notice how strong convexity helps: the larger K_t is, the lower the regret.

b. (*4 points*) Now assume that we have bounds on the norms of our experts and of our gradients: that $\|u\|_2 \leq B$ for all $u \in S$, and that $\|z_t\|_2 \leq L$ for all t.

Furthermore, assume that, when we pick varying step sizes, we will always pick them to satisfy $\frac{1}{\eta_t} \geq \frac{1}{\eta_{t-1}} + K_t$. Intuitively, the step size decreases at least as fast as what strong convexity dictates. (By convention, let $\eta_0 = \infty$.)

Note: if we had no strong convexity, i.e. if $K_t = 0$ for all t, then this assumption simply means that we require non-increasing step sizes.

Use part (a) to obtain the regret bound

$$\text{Regret} \leq 2B^2 \left(\frac{1}{\eta_T} - \sum_{t=1}^{T} K_t \right) + \frac{1}{2} L^2 \sum_{t=1}^{T} \eta_t. \tag{4}$$

c. (*4 points*) Let's use this result to bound regret in a setting where we have a simple strong convexity assumption and a simple schedule for our step sizes. Assume that all the loss functions are at least K-strongly convex, that is, $K_t \geq K > 0$ for all t. Show that if we set the step size according to $\eta_t = \frac{1}{\sum_{i=1}^{t} K_i}$, we obtain the following bound:

$$\text{Regret} \leq \frac{L^2(\log(T) + 1)}{2K}. \tag{5}$$

Remark: This shows that we get *logarithmic* regret for *all* strongly convex functions (not just quadratic functions), provided we set the step size to decay as $O(1/t)$. Note, however, that this bound does become increasingly bad as we lose strong convexity ($K \to 0$).

d. (*4 points*) Actually, not all hope is lost for our analysis when we lose strong convexity. Suppose that $K_t = 0$ (meaning that we make no assumptions on the f_t other than convexity). Show that if we set η_t to $\frac{B}{L\sqrt{t}}$ then we obtain the bound

$$\text{Regret} \leq 3BL\sqrt{T}. \tag{6}$$

Remark: This final result is significant because it shows that we can obtain a regret of $O(\sqrt{T})$ even without knowing T in advance.

e. (*4 points*) Let's see how our analysis guides our choice of step sizes when faced with a concrete problem. Assume that at each time step t, we get a feature vector $x_t \in \mathbb{R}^d$ for some $d \geq 2$. Suppose that $\|u\|_2 \leq B$ for all $u \in S$, $\|x_t\|_2 \leq C$, and $|y_t| \leq a$. We'll look at two fairly common loss functions:

1. Linear least-squares regression: $f_t(w) = (y_t - w \cdot x_t)^2$
2. Regularized SVM: $f_t(w) = \max\{0, 1 - y_t(w \cdot x_t)\} + \lambda\|w\|_2^2$.

How would you pick the step sizes for these two scenarios? Write your selections for the step size η_t (as a function of a, B, C, t) and give upper bounds on the resulting regret based on your calculations above. Please justify your answer.

3. Bonus Problem: Proof of Bobkov-Gotze Theorem (*0 points*)

Let μ be a probability measure over on a metric space $(\mathcal{X}, d)$.[1] Here, for simplicity, we assume $\mathcal{X}$ is a finite set. (Although the result holds for general metric space.) Let μ be a probability distribution over $\mathcal{X}$. Recall that a function f is 1-Lipschitz with respect to the metric d if

$$\forall x, y \in \mathcal{X}, |f(x) - f(y)| \leq d(x, y) \tag{7}$$

We will show that for any $\sigma > 0$ the following two statements (which are properties of μ) are equivalent:

1. Let $X \sim \mu$. For every 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, $f(X)$ is subgaussian with varriance proxy σ^2.
2. For any distribution ν over $\mathcal{X}$, $W_1(\nu, \mu) \leq \sqrt{2\sigma^2 \mathrm{KL}\,(\nu\|\mu)}$.

This result says that as long as the distribution μ is suitably regular (satisfying condition 1), then $\mathrm{KL}\,(\nu\|\mu)$ is lower bounded by $W_1(\nu, \mu)^2$ up to some constant, which will be a useful tool for analyzing distance between distributions such as in Wasserstein GANs.

a. (*3 bonus points*) Prove the following equality:

$$\log \mathbb{E}_\mu[e^f] = \sup_\nu \{\mathbb{E}_\nu f - \mathrm{KL}\,(\nu\|\mu)\} \tag{8}$$

Here $\mathbb{E}_\mu[e^f]$ is a shorthand of $\mathbb{E}_{X\sim\mu}[e^{f(X)}]$ and $\mathbb{E}_\mu f$ is a shorthand of $\mathbb{E}_{X\sim\mu} f(X)$, and the supremum is over all probability measures on $\mathcal{X}$.

b. (*4 bonus points*) Prove the equivalence between 1 and 2.

[1] Please see `https://en.wikipedia.org/wiki/Metric_space` for the definition of a metric space.