## Privacy in Statistics and Machine Learning — Spring 2023
## In-class Exercises for Lecture 12 (Gradient Descent)
## March 2, 2023

**Adam Smith (based on materials developed with Jonathan Ullman)**

*Problems with marked with an asterisk (*) are more challenging or open-ended.*

1. Show that ERM for $G$-Lipschitz losses on a set of diameter $R$ can be reduced by rescaling to ERM for 1-Lipschitz losses for a set of diameter 1. Show that the excess risk changes by a factor of $GR$ due to the rescaling. (*Reminder:* Re-scaling the input domain changes $G$ as well as $R$. )

2. The performance of gradient descent can vary a lot depending on how we choose the step size, even for convex, one-dimensional problems. Suppose we run gradient descent with the loss function $L(w) = w^2$ and no constraints (that is, $C = \mathbb{R}$).

   - If we start at $w_0 = 1$ and use step size $\eta = 2$, how will the algorithm behave? Will it converge?
   - How would your answer to the previous part change if we used $\eta = 1$? What about $\eta = \frac{1}{2}$?
   - How would you answer to the first part change if we imposed the constraint $w \in C = [-10, 10]$? What about $w \in [10, 20]$?

3. Prove the following variant of the Amplification by Subsampling Lemma.

   Suppose that given an algorithm $A$ whose input can be a data set of any size, we build a new algorithm $A'_p$ as follows: on input $\mathbf{x}$, construct a smaller data set $\mathbf{x}'$ by including each data record from $\mathbf{x}$ with probability $p$, independently of other data records. Finally, return $A(\mathbf{x}')$.

   If $A$ is $(\varepsilon, 0)$-DP under insertion/removal, then show that $A'_p$ is $(\varepsilon', 0)$-DP under insertion/removal, where $\varepsilon' = \ln(1 + p(e^{\varepsilon} - 1))$. (*NB:* As $\varepsilon$ goes to 0, we get $\varepsilon' \approx p\varepsilon$.)

4. Let's generalize the analysis of private SGD to the version where at each step, we use a uniformly random batch $B_t$ of $m$ records to estimate the gradient, so

$$\tilde{g}_t = \left( \frac{1}{m} \sum_{i \in B_t} \nabla \ell(w_{t-1}; x_i) \right) + N(0, \sigma^2) \, .$$

   Given $\delta$, we want to understand for which $\varepsilon$ this step is $(\varepsilon, \delta)$-DP. Show that, as long as

$$\sigma \geq 2G\sqrt{2\ln(1/\delta)} \cdot \frac{1}{m}$$

   the privacy cost of one step of gradient descent with subsampling is at most $e$ times higher than it would be if we had used the entire data set to estimate the gradient. In other words, subsampling has virtually no effect on privacy as long the noise level is sufficiently high.

5. We analyzed gradient descent for the setting where the diameter $R$ of $C$ is bounded. But suppose $C$ is not bounded—say $C = \mathbb{R}^d$. We could still hope to get a good bound if our initial point $w_0$ is not

too far from a true optimum $w^*$. A friend conjectures that if one has a good idea of $\|w_0 - w^*\|$, one should be able to set $\eta$ to get a bound of the form

$$L(\hat{w}) - L(w^*) \leq \frac{G \times \left(\text{some function of } \|w_0 - w^*\|\right)}{\sqrt{T}}.$$

Are they correct? What function belongs there?

6. It is common for the learning rate $\eta$ to decrease over the course of gradient descent. Suppose we set $\eta_t = \frac{1}{\sqrt{t}}$, and update the estimate as $u_t = w_{t-1} + \eta_t \nabla L(w_{t-1})$. This way of doing things has the benefit that we don't need to set the number of iterations $T$ ahead of time.

   Show that, when $G = R = 1$, we can get the same asymptotic risk bound of $O(1/\sqrt{T})$ for gradient descent.

**Exercises/facts about convex sets and functions (for reference/practice)**

7. Which of these operations, when applied to a set of convex functions, always produces a convex function? (a) Sum (b) Min (c) Max (d) Median

8. Show that the set of minima of a convex function is a convex set.

9. Show that for any closed, nonempty convex set $C \subseteq \mathbb{R}^d$, the function $f(x) = \min_{w \in C} \|x - w\|_2$ is a convex function on all of $\mathbb{R}^d$. [*Hint:* First prove this for $d = 2$. The general case $d > 2$ is a bit harder to visualize, and can be skipped if you are short on time.]

10. Show that for any closed, nonempty convex set $C \subseteq \mathbb{R}^d$, the *projection* function $\Pi_C(x) = \arg\min_{w \in C} \|x - w\|_2$ is (a) a well-defined function from $\mathbb{R}^d$ to $\mathbb{R}^d$ (that is, the minimizer is unique), (b) 1-Lipschitz, meaning that for all $x, y \in \mathbb{R}^d$, we have

$$\|\Pi_C(x) - \Pi_C(y)\|_2 \leq \|x - y\|_2$$

   [*Hint:* First prove this for $d = 2$. The general case $d > 2$ is a bit harder to visualize, and can be skipped if you are short on time.]

11. Prove Jensen's inequality (Exercise 3.4 in the notes). [*Hint:* Let $\mu = \mathbb{E}(X)$ and let $g_\mu(\cdot)$ be an affine lower bound to $f$ such that $f(\mu) = g_\mu(\mu)$. What is $\mathbb{E}\left(g_\mu(X)\right)$?]

12. The level set of a function $f$ at $a$ is the set $\{w \in C : f(w) \leq a\}$. Show that if $f$ is convex, then all of its level sets are convex. Show via a counterexample that the converse is false.

13. **(*)** Suppose $f : \mathbb{R} \to \mathbb{R}$ is convex.

    (a) Show that the subgradients of $f$ are monotone, namely: for every $x_1, x_2 \in \mathbb{R}$ such that $x_1 < x_2$, if $y_1 \in \partial f(x_1)$ and $y_2 \in \partial f(x_2)$, then $y_1 \leq y_2$. (It might be easier to first prove this for $f$ that is differentiable.)

    (b) Show that if $f$ is convex and 1-Lipschitz on a finite interval (say $[-1, 1]$), then it can be written a constant plus a convex combination of absolute value functions. Specifically, show that there is a constant $a$ and a distribution $P$ on $[-1, 1]$ such that for all $x$, $f(x) = a + \mathop{\mathbb{E}}_{Y \sim P}(|x - Y|)$.