CS 325: Private Data Analysis **Fall 2022**

Lecture 14: Differentially Private Empirical Risk Minimization: II

Lecturer: Di Wang *Scribes: Di Wang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In the previous lecture, we discussed about the Output Perturbation Method. Generally speak, the method is based on the stability *i.e.,* the ℓ_2-norm sensitivity of some black box or white box optimization algorithm. However, there are still several issues of this method. The first one is perturbing the output may violates the requirement that the parameter will lie in the parameter space $\mathcal{C}$. The second one is that the utility of the convex case could be further improved. In this lecture, we will discuss the second type of approaches which is called Objective Perturbation Method.

14.1 Output Perturbation Method

The objective perturbation method was firstly introduced by [1]. Here we will talk about the setting in this paper. We focus on the case without constraint:

$$\arg\min_{\theta\in\mathbb{R}^d} L(\theta, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i, y_i).$$

The key idea of the objective perturbation method is that, rather than solving this optimization problem, we instead solve a related problem where the components of the parameter vector θ are penalized or rewarded by a random amount, *i.e.,* we will focus on the following problem:

$$\arg\min_{\theta\in\mathbb{R}^d} L^{+}(\theta, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i, y_i) + \frac{\lambda}{2n}\|\theta\|_2^2 + \frac{1}{n}\langle b, \theta\rangle, \tag{14.1}$$

where b is some randomized distribution. Next, we will focus on (14.3). Before that, we will provide some assumptions.

Assumption 14.1 *We assume $\ell(\cdot, x, y)$ is L-Lipschitz, twice differentiable and β-smooth.* **Moreover, we assume that the rank of the Hessian matrix for each $\ell(\theta, x, y)$ is at most 1.**

Note that Lipschitz, differentiable and smooth are commonly used assumption in machine learning or optimization community. However, the rank of the Hessian matrix to be 1 is quite wired. Later we will see how to relax this condition. But here we have to say that there is a class of loss functions that satisfies this condition, called the **Generalized Linear Loss**.

Definition 14.2 *A loss function $\ell(\theta, x, y)$ is a Generalized Linear loss if it can be written as $\ell(\theta, x, y) = f(\langle\theta, yx\rangle)$ for some convex function.*

Algorithm 1 Objective Perturbation

```
For $(\epsilon, \delta)$-DP, sample $b \sim \mathcal{N}(0, \frac{10L^2 \log \frac{1}{\delta}}{\epsilon^2} I_d)$. For $\epsilon$-DP, sample $b$ from the Gamma distribution with density $h(b) \propto \exp(-\frac{\epsilon}{2L}\|b\|_2)$
Get the exact solution of the following optimization problem
    $\theta_{priv} = \arg\min_{\theta \in \mathbb{R}^d} L^+(\theta, D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \frac{\lambda}{2n}\|\theta\|_2^2 + \frac{1}{n}\langle b, \theta \rangle,$
where $\lambda$ is some parameter.
```

The algorithm can be found in Algorithm 1.

Theorem 14.3 *Under Assumption 14.1, Algorithm 1 is* $\ln(1+\frac{2\beta}{\lambda}) + \epsilon$*-DP or* $(\frac{\epsilon}{2} + \ln(1+\frac{2\beta}{\lambda}), \delta)$*-DP.*

Proof: We assume $D \sim D'$ are neighboring datasets which differ by the i-th sample and denote θ'_{priv} as the output in the case where the input data is D'. By the definition of θ_{priv}, we have

$$\nabla L(\theta_{priv}) + \frac{\lambda}{n}\theta_{priv} + \frac{1}{n} b = 0, \tag{14.2}$$

that is

$$b = -n\nabla L(\theta_{priv}, D) - \lambda\theta_{priv}, b' = -n\nabla L(\theta_{priv}, D') - \lambda\theta_{priv} \tag{14.3}$$

Due to the ℓ_2-norm regularization, we know that $L^+(\theta, D)$ is strongly convex. And we can easily see that for strongly convex function it has a unique optimal solution. Thus, we can see that there is an one-to-one map between θ_{priv} and the value b.

Next we first consider the ϵ-DP case. Denote the probability density function of θ_{priv} as $p_{A(D)}$, the probability density function of θ'_{priv} as $p_{A(D')}$, then by the definition of ϵ-DP we want to show that for any θ in the output space we have $\frac{p_{A(D)}(\theta)}{p_{A(D')}(\theta)} \leq e^{\epsilon}$. However, there is no explicit form of the density function. What we know is that we have the explicit form of b, which is Gaussian or Gamma. And we know there is a map H such that $\theta_{priv} = H(b)$ with $H^{-1}(\theta_{priv}) = b = -n\nabla L(\theta_{priv}, D) - \lambda\theta_{priv}$. Motivated by this we can use the change of variables to calculate the density function of θ_{priv}.

Lemma 14.4 *Consider two random variables x, y with $y = H(x)$ for some surjective map H. Suppose the density function of x is $f(\cdot)$ then the density function of y, g has the following form:*

$$g(y) = f(H^{-1}(y))|det(\frac{dH^{-1}(z)}{dz}|_{z=y})|,$$

with the differential regarded as the Jacobian of the inverse of H, evaluated at y.

Thus we have

$$\frac{p_{A(D)}(\theta)}{p_{A(D')}(\theta)} = \frac{p_D(b)}{p_{D'}(b')} \cdot \frac{|\det(\frac{dH_D^{-1}(z)}{dz}|_{z=y})|}{|\det(\frac{dH_{D'}^{-1}(z)}{dz}|_{z=y})|},$$

where $p_D(b)$ is the density function of b in the case where the input data is D and H_D^{-1} is the map (14.3) in the case where the input data is D.

We first consider the term $\frac{|\det(\frac{dH_D^{-1}(z)}{dz}|_{z=y})|}{|\det(\frac{dH_{D'}^{-1}(z)}{dz}|_{z=y})|}$. For $|\det(\frac{dH_D^{-1}(z)}{dz}|_{z=y})|$ since we know $H_{D'}-1(\theta_{priv}) = b = -n\nabla L(\theta_{priv}, D') - \lambda\theta_{priv}$, by some simple calculation we have

$$A = \det(\frac{dH_{D'}^{-1}(z)}{dz}) = -n\nabla^2 L(z, D') - \lambda I_{d\times d},$$

where $I_{d\times d}$ is the identity matrix. Thus we have

$$\det(\frac{dH_D^{-1}(z)}{dz}) = -n\nabla^2 L(z, D) - \lambda I_{d\times d} = -n\nabla^2 L(z, D') + \nabla^2\ell(z, x_i', y_i') - \nabla^2\ell(z, x_i, y_i) - \lambda I_{d\times d} = A + E,$$

where $E = \nabla^2\ell(z, x_i', y_i') - \nabla^2\ell(z, x_i, y_i)$ is a matrix whose rank is at most 2 by Assumption 1. Thus we want to bound $\frac{\det(A+E)}{\det(A)}$, we will use the following lemma

Lemma 14.5 ([1]) *If A is a full rank matrix and if E is a matrix with rank at most 2, then*

$$\frac{det(A+E)}{det(A)} = 1 + \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E)\lambda_2(A^{-1}E).$$

We know that the smallest eigenvalue of A is λ, and the largest eigenvalue of E is 2β by Assumption. Thus

$$\frac{\det(A+E)}{\det(A)} \leq (1 + \frac{2\beta}{\lambda})^2.$$

We now focus on the first term

$$\frac{p_D(b)}{p_{D'}(b')} \leq \exp(\frac{\epsilon}{2L}\|\nabla\ell(\theta, x_i, y_i) - \nabla\ell(\theta, x_i', y_i')\|_2) \leq \exp(\epsilon). \tag{14.4}$$

Thus, the algorithm is $\ln(1 + \frac{2\beta}{\lambda}) + \epsilon$-DP. Thus if we have a budget ϵ' then we can choose $\epsilon = \frac{\epsilon'}{2}$ and $\lambda = \frac{2\beta}{e^{\frac{\epsilon'}{2}} - 1}$.

Next we consider the (ϵ, δ)-DP case. The same as the ϵ-DP case, we just consider the term of $\frac{p_D(b)}{p_{D'}(b')}$, denote the term $g = b' - b$ then we have

$$\frac{p_D(b)}{p_{D'}(b')} = \frac{\exp(-\frac{\|b\|_2^2}{2\sigma^2})}{\exp(-\frac{\|b'\|_2^2}{2\sigma^2})} = \exp(\frac{\|b+g\|_2^2 - \|b\|_2^2}{2\sigma^2}) = \exp(\frac{\|g\|_2^2 + 2\langle b, g\rangle}{2\sigma^2}). \tag{14.5}$$

For the term $\|g\|_2^2 = \|\nabla\ell(\theta, x_i, y_i) - \nabla\ell(\theta, x_i', y_i')\|_2^2 \leq 4L^2$. For the term $\langle b, g\rangle$ since $b \sim \mathcal{N}(0, \sigma^2 I_d)$ we have

$$\langle b, g\rangle \sim \mathcal{N}(0, \|g\|_2^2\sigma^2).$$

Thus we have for any t,

$$\mathbb{P}(|\langle b, g\rangle| \geq 2L\sigma t) \leq \exp(-\frac{t^2}{2}). \tag{14.6}$$

Thus, in total we have with probability at least $1 - \delta$ we have

$$\frac{p_D(b)}{p_{D'}(b')} \leq \exp(\frac{2L^2 + \sqrt{2}\sigma L\sqrt{\log\frac{1}{\delta}}}{\sigma^2}) \leq \exp(\frac{\epsilon}{2}).$$

Thus, the algorithm will be $(\frac{\epsilon}{2} + \ln(1 + \frac{2\beta}{\lambda}), \delta)$-DP. ■

Next we will focus on the utility, here we only consider the general convex case. Before that we recall some definitions

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} L(\theta, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i, y_i)$$
$$\theta_{priv} = \arg\min_{\theta\in\mathbb{R}^d} L^+(\theta, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i, y_i) + \frac{\lambda}{2n}\|\theta\|_2^2 + \frac{1}{n}\langle b, \theta\rangle$$
$$\bar{\theta} = \arg\min_{\theta\in\mathbb{R}^d} \bar{L}(\theta, D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i, y_i) + \frac{\lambda}{2n}\|\theta\|_2^2$$

To show the utility, we have

$$\begin{aligned}L(\theta_{priv}, D) - L(\theta^*, D) &\leq \bar{L}(\theta_{priv}, D) - \bar{L}(\bar{\theta}, D) + \bar{L}(\bar{\theta}, D) - \bar{L}(\theta^*, D) + \frac{\lambda}{2n}\|\theta^*\|_2^2 - \frac{\lambda}{2n}\|\theta_{priv}\|_2^2\\ &\leq \bar{L}(\theta_{priv}, D) - \bar{L}(\bar{\theta}, D) + \frac{\lambda}{2n}\|\theta^*\|_2^2 - \frac{\lambda}{2n}\|\theta_{priv}\|_2^2\end{aligned}$$

For the term of $\bar{L}(\theta_{priv}, D) - \bar{L}(\bar{\theta}, D)$ we have since $L^+(\theta, D)$ is $\frac{\lambda}{n}$-strongly convex

$$\begin{aligned}\bar{L}(\bar{\theta}, D) + \frac{1}{n}\langle b, \bar{\theta}\rangle &\geq \bar{L}(\theta_{priv}, D) + \frac{1}{n}\langle b, \theta_{priv}\rangle + \frac{\lambda}{2n}\|\bar{\theta} - \theta_{priv}\|_2^2\\ \iff \bar{L}(\theta_{priv}, D) - \bar{L}(\bar{\theta}, D) &\leq \frac{\|b\|_2\|\theta_{priv} - \bar{\theta}\|_2}{2n} \leq \frac{\|b\|_2^2}{n\lambda}.\end{aligned}$$

Moreover we have

$$\begin{aligned}\bar{L}(\bar{\theta}, D) + \frac{1}{n}\langle b, \bar{\theta}\rangle &\geq \bar{L}(\theta_{priv}, D) + \frac{1}{n}\langle b, \theta_{priv}\rangle + \frac{\lambda}{2n}\|\bar{\theta} - \theta_{priv}\|_2^2\\ \iff \frac{1}{n}\|b\|_2\|\theta_{priv} - \bar{\theta}\|_2 &\geq \frac{\lambda}{2n}\|\bar{\theta} - \theta_{priv}\|_2^2\\ \iff \|\bar{\theta} - \theta_{priv}\|_2 &\leq \frac{2\|b\|_2}{\lambda}.\end{aligned}$$

In total we have

$$L(\theta_{priv}, D) - L(\theta^*, D) \leq O(\frac{\|b\|_2^2}{n\lambda} + \frac{\lambda}{2n}\|\theta^*\|_2^2). \tag{14.7}$$

Based on different noise b we have the following theorem.

Theorem 14.6 *Under Assumption 14.1, for ϵ'-DP, set $\lambda = O(\frac{p\log p}{\|\theta^*\|_2})$ we have*

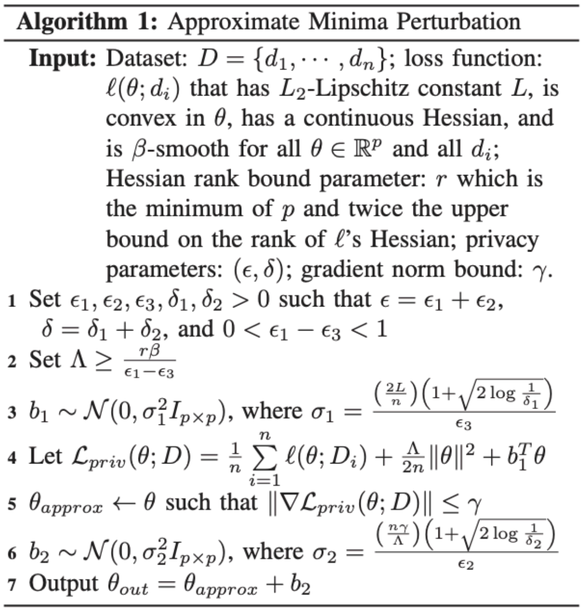
$$\mathbb{E}L(\theta_{priv}, D) - L(\theta^*, D) \leq O(\frac{\|\theta^*\|_2 p\log p}{n\epsilon'}). \tag{14.8}$$

For (ϵ', δ)-DP set $\lambda = O(\frac{\sqrt{p\log 1/\delta}}{\|\theta^*\|_2})$ we have

$$\mathbb{E}L(\theta_{priv}, D) - L(\theta^*, D) \leq O(\frac{\|\theta^*\|_2\sqrt{p\log 1/\delta}}{n\epsilon'}). \tag{14.9}$$

14.1.1 Discussion

So far we have discussed the Objective Perturbation method, we can see that the strength of this method comparing with the output perturbation method is that it is quite flexible and we have improve upper bound for general convex loss functions. However, there are still several issues:



```
Algorithm 1: Approximate Minima Perturbation
Input: Dataset: $D = \{d_1, \cdots, d_n\}$; loss function:
    $\ell(\theta; d_i)$ that has $L_2$-Lipschitz constant $L$, is
    convex in $\theta$, has a continuous Hessian, and
    is $\beta$-smooth for all $\theta \in \mathbb{R}^p$ and all $d_i$;
    Hessian rank bound parameter: $r$ which is
    the minimum of $p$ and twice the upper
    bound on the rank of $\ell$'s Hessian; privacy
    parameters: $(\epsilon, \delta)$; gradient norm bound: $\gamma$.
Set $\epsilon_1, \epsilon_2, \epsilon_3, \delta_1, \delta_2 > 0$ such that $\epsilon = \epsilon_1 + \epsilon_2$,
    $\delta = \delta_1 + \delta_2$, and $0 < \epsilon_1 - \epsilon_3 < 1$
Set $\Lambda \geq \frac{r\beta}{\epsilon_1 - \epsilon_3}$
$b_1 \sim \mathcal{N}(0, \sigma_1^2 I_{p\times p})$, where $\sigma_1 = \frac{\left(\frac{2L}{n}\right)\left(1+\sqrt{2\log\frac{1}{\delta_1}}\right)}{\epsilon_3}$
Let $\mathcal{L}_{priv}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; D_i) + \frac{\Lambda}{2n}\|\theta\|^2 + b_1^T\theta$
$\theta_{approx} \leftarrow \theta$ such that $\|\nabla \mathcal{L}_{priv}(\theta; D)\| \leq \gamma$
$b_2 \sim \mathcal{N}(0, \sigma_2^2 I_{p\times p})$, where $\sigma_2 = \frac{\left(\frac{n\gamma}{\Lambda}\right)\left(1+\sqrt{2\log\frac{1}{\delta_2}}\right)}{\epsilon_2}$
Output $\theta_{out} = \theta_{approx} + b_2$
```

1. We can see that although we do not need the strongly convex condition in Assumption 14.1, here we must assume the rank of the Hessian matrix is at most 1, which is quite strict.

2. We need to exactly solve the perturbed objective function, which is impossible in practice.

3. Due to the equation (14.3) we need to assume $\mathcal{C}$ is the whole space.

We will take about the recent development of this method. In [2], the authors aimed to address the issue 1 and 2. See Figure 14.1.1 for details.

[3] addressed the issue 3. They extend the original Algorithm 1 to any bounded convex set $\mathcal{C}$ without changing any parameter. Thus Algorithm 1 is DP for any bounded convex set $\mathcal{C}$. Moreover, they relax the differentiable condition of the loss function to the following regularized ERM problem:

$$\arg\min_{\theta\in\mathbb{R}^d} L(\theta, D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i, y_i) + r(\theta),$$

where $r(\theta)$ is some regularization that could be non-smooth and is independent on the dataset, such that lasso where $r(\theta) = \lambda\|\theta\|_1$.

However, there is no algorithm that can handle all the three issues together. This is an open problem.

On the other side, [4] recently extend to the case where the loss function can be non-convex. However, the constraint set $\mathcal{C}$ must be discrete and thus cannot be used to most machine learning problems.

References

[1] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[2] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

[3] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[4] Seth Neel, Aaron Roth, Giuseppe Vietri, and Steven Wu. Oracle efficient private non-convex optimization. In *International Conference on Machine Learning*, pages 7243–7252. PMLR, 2020.