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

Lecture 17: Differentially Private Empirical Risk Minimization: V

Lecturer: Di Wang *Scribes: Di Wang*

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

So far we have discussed several methods for DP-ERM. However, besides the ERM problem which is about the training, we also want to make our model has good generalization property on unseen or testing data. This problem is called the Stochastic Optimization. Mathematically, for the dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, we assume each data record is sampled from some unknown distribution $\mathcal{P}$. And instead of the empirical risk, our goal is to minimize the population risk function based on the dataset D

$$\theta^* = \arg\min_{\theta \in \mathcal{C}} L_{\mathcal{P}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(\theta, x, y)]. \tag{17.1}$$

For any private estimator θ_{priv}, we will use the (expected) excess population risk to measure the performance of θ_{priv}:

$$\mathbb{E}L_{\mathcal{P}}(\theta_{priv}) - L_{\mathcal{P}}(\theta^*).$$

In the following we will consider the case where the loss function is convex, and we call the problem as Differentially Private Stochastic Convex Optimization (DP-SCO). And we will mainly focus on the (ϵ, δ)-DP model. Note that compared with DP-ERM, DP-SCO is much hard since we do not have the access to $L_{\mathcal{P}}(\theta)$. In this lecture, we will introduce two approaches to get upper bound of the excess population risk.

17.1 From Empirical to Population Risk

The most direct approach is transferring from the bound of excess empirical risk to population risk. [3] first proposed this approach, which is based on the following lemma:

Lemma 17.1 *If the loss function is L-Lipschitz, and α strongly convex. Then with probability at least $1-\gamma$ over the randomness of sampling the dataset D we have*

$$\mathbb{E}L_{\mathcal{P}}(\theta_{priv}) - L_{\mathcal{P}}(\theta^*) \leq O(\frac{L}{\sqrt{\alpha}}\sqrt{L(\theta_{priv}, D) - \min_{\theta\in\mathcal{C}} L(\theta, D)} + \frac{L^2}{\gamma\alpha n}). \tag{17.2}$$

Since for the empirical risk, the optimal bound is $O(\frac{d\log\frac{1}{\delta}}{\alpha n^2\epsilon^2})$. Thus, if we omit other terms, the excess population risk will be

$$\mathbb{E}L_{\mathcal{P}}(\theta_{priv}) - L_{\mathcal{P}}(\theta^*) \leq O(\frac{\sqrt{d\log\frac{1}{\delta}}}{\alpha n\epsilon}).$$

For the general convex loss case, we can do the same thing. That is we first add a regularization term to the loss to make it to be Δ-strongly convex, where Δ is a parameter,

$$\tilde{\ell}(\theta, x) = \ell(\theta, x) + \frac{\Delta}{2}\|\theta\|_2^2. \tag{17.3}$$

Thus, we have

$$\mathbb{E}L_{\mathcal{P}}(\theta_{priv}) - L_{\mathcal{P}}(\theta^*) = \mathbb{E}\tilde{L}_{\mathcal{P}}(\theta_{priv}) - \tilde{L}_{\mathcal{P}}(\theta^*) + \frac{\Delta}{2}\|\theta^*\|_2^2 - \frac{\Delta}{2}\|\theta_{priv}\|_2^2$$
$$\leq \mathbb{E}\tilde{L}_{\mathcal{P}}(\theta_{priv}) - \min_{\theta\in\mathcal{C}}\tilde{L}_{\mathcal{P}}(\theta) + \frac{\Delta}{2}\|\mathcal{C}\|_2^2 = O(\frac{\sqrt{d\log\frac{1}{\delta}}}{\Delta n\epsilon} + \frac{\Delta}{2}\|\mathcal{C}\|_2^2)$$

Balancing Δ we have

$$\mathbb{E}L_{\mathcal{P}}(\theta_{priv}) - L_{\mathcal{P}}(\theta^*) \leq O(\frac{\sqrt[4]{d\log\frac{1}{\delta}}}{\sqrt{n\epsilon}}).$$

We can see now the excess population risk degrades to $O(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon})$ and $O(\frac{\sqrt[4]{d\log\frac{1}{\delta}}}{\sqrt{n\epsilon}})$ from $O(\frac{d\log 1/\delta}{n^2\epsilon^2})$ and $O(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon})$ for strongly and general convex, respectively. Our question is, can we further improve the upper bound? We will show the second approach which is based on the algorithmic stability.

17.2 Algorithmic Stability Approach

To show the main idea of this approach, which is proposed by [2, 5], we need to first decompose the excess population risk into the the sum of the excess empirical risk and the generalization error:

Lemma 17.2 *Consider a randomized algorithm $\mathcal{A}$, then the excess population risk of its output $\mathcal{A}(D)$ can be decomposed as*

$$\mathbb{E}L_{\mathcal{P}}(\mathcal{A}(D)) - L_{\mathcal{P}}(\theta^*) \leq \mathbb{E}_{A,D}[L(\mathcal{A}(D), D)] - \min_{\theta\in\mathcal{C}} L(\theta, D) + |\mathbb{E}[L(\mathcal{A}(D), D)] - L_{\mathcal{P}}(\mathcal{A}(D))|. \tag{17.4}$$

Now we first consider the case where the loss function is strongly convex, then the first term in (17.4) is just the empirical risk of the output $\mathcal{A}(D)$. The second term, which is the difference of empirical risk and population risk at the the output $\mathcal{A}(D)$, is called the **generalization error**. It could be bounded by the algorithmic stability of the algorithm $\mathcal{A}$.

Lemma 17.3 *Suppose the algorithm A is γ-stable, that is for any neighboring data $D \sim D'$ and any data x we have*

$$|\ell(A(D), x) - \ell(A(D'), x)| \leq \gamma. \tag{17.5}$$

Then the generalization error

$$|\mathbb{E}[L(\mathcal{A}(D), D)] - L_{\mathcal{P}}(\mathcal{A}(D))| \leq 2\gamma. \tag{17.6}$$

Thus, based on previous lemma, our goal is to design some private algorithm that has low excess empirical risk and has low stability. Actually, we can use the output perturabtion method:

Theorem 17.4 ([6]) *If the loss function is L-Lipschitz, β-smooth and α-strongly convex. If we run gradient descent(GD) algorithm with constant step size $\eta \leq \frac{1}{\alpha+\beta}$ for T steps, then the ℓ_2-norm-sensitivity of GD for any bounded convex set $\mathcal{C}$,*

$$\Delta_T \leq \frac{5L(\alpha+\beta)}{n\alpha\beta}.$$

Thus, the stability of the algorithm is bounded by $\frac{5L^2(\alpha+\beta)}{n\alpha\beta}$.

Algorithm 1 Output Perturbation base on GD

1: Initialize a start parameter θ_0
2: **for** $t = 0, \cdots T-1$ **do**
3: $\quad \theta_t = \Pi_{\mathcal{C}}(\theta_{t-1} - \eta_{t-1}\nabla L(\theta_{t-1}, D))$.
4: **end for**
5: Return $\theta_{priv} = \theta_T + Z$, where $Z \sim \mathcal{N}(0, O(\frac{\Delta_T^2 \log \frac{1}{\delta}}{\epsilon^2})I_d)$ and Δ_T is the ℓ_2-norm sensitivity of θ_T.

Theorem 17.5 *Under the same assumption as in Theorem 17.4*

$$\mathbb{E}L(\theta_{priv}, D) - L(\theta^*, D) \leq O(\frac{d \log \frac{1}{\delta}}{n^2\epsilon^2} + \exp(-\frac{2\alpha\beta T}{(\alpha+\beta)^2})).$$

Moreover, if we take $T = O(\frac{\beta}{\alpha} \log n)$ *then we have*

$$\mathbb{E}L(\theta_{priv}, D) - L(\theta^*, D) \leq O(\frac{d \log \frac{1}{\delta}}{n^2\epsilon^2}).$$

In total we have the output of Algorithm 1 has the excess population risk of $O(\frac{d \log 1/\delta}{n^2\epsilon^2} + \frac{1}{n})$. Moreover, this is bound is optimal and cannot be improved [1].

For the general convex loss function, as we mentioned in Lecture 11, even for the empirical risk, it is not optimal. Thus, we cannot use the same method here for the general convex case. [4] first provide the optimal algorithm and show it is possible to achieve an error of $O(\frac{\sqrt{d \log \frac{1}{\delta}}}{n\epsilon} + \frac{1}{\sqrt{n}})$.

References

[1] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

[2] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

[3] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[4] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[5] Zhiyu Xue, Shaoyang Yang, Mengdi Huai, and Di Wang. Differentially private pairwise learning revisited.

[6] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.