Lecture 16: Differential privacy and robust statistics

November 19, 2025

1 Differentially private learning

As mentioned in last lecture, the typical guarantee for differentially private learning is the following combination of guarantees:

- **Worst case privacy:** The algorithm should guarantee (ε, δ)-differential privacy for *any* worst-case dataset.
- **Average case utility:** If the data comes from a specific generative model as prescribed by the algorithm designer, then the output of the algorithm should be "correct" with high probability.

A canonical statistical learning task one might wish to perform privately is that of *private mean estimation*:

Definition 1.1 (Private mean estimation)**.** Let $\mathcal{D}$ be a class of distributions over $\mathbb{R}^d$. Give an (ε, δ)-differentially private algorithm which takes n data points $X_1, \dots, X_n$, with the following utility guarantee: if $X_1, \dots, X_n \sim D$ for some $D \in \mathcal{D}$ with mean $\mu \in \mathbb{R}^d$, the algorithm outputs $\widehat{\mu}$ so that $\|\widehat{\mu} - \mu\|_2$ is small with high probability.

As in the robust statistics setting (and pretty much any distribution learning setting), arguably the most natural class of distributions for which this has been studied is for the class of isotropic Gaussians, namely, $\mathcal{D} = \{\mathcal{N}(\mu, I) : \mu \in \mathbb{R}^d\}$.

2 Private algorithms for mean estimation

Recall the simplest algorithms for achieving pure and approximate DP are arguably the Laplace and Gaussian mechanisms, respectively, for which we have the following guarantees:

Theorem 2.1 (Laplace mechanism)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$, and let $\varepsilon > 0$. Let Y be a random variable whose coordinates are independent* $\mathrm{Lap}(\Delta_1(f)/\varepsilon)$*. Then $\mathcal{A}(X) = f(X) + Y$ is ε-differentially private, and moreover,*

$$\|\mathcal{A}(X) - f(X)\|_2 \lesssim \frac{\sqrt{d} \cdot \Delta_1(f) \log(1/\beta)}{\varepsilon}$$

with probability $1 - \beta$.

Theorem 2.2 (Gaussian mechanism)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$, and let $\varepsilon, \delta > 0$. Let $Y \sim \mathcal{N}(0, \sigma^2 I)$ for $\sigma = O\left(\sqrt{\log 1/\delta} \cdot \Delta_2(f)/\varepsilon\right)$. Then $\mathcal{A}(X) = f(X) + Y$ is (ε, δ)-DP, and moreover,*

$$\|\mathcal{A}(X) - f(X)\|_2 \lesssim \frac{\sqrt{d} \cdot \Delta_2(f) \sqrt{\log(1/\delta) \cdot \log(1/\beta)}}{\varepsilon}\,,$$

with probability $1 - \beta$.

Let's instantiate these mechanisms for the Gaussian mean estimation setting. Let's further assume for simplicity that $\|\mu\|_2 \le 1$. Then, given n samples $X_1, \dots, X_n \sim \mathcal{N}(\mu, I)$, the optimal estimator, ignoring other constraints, is just the empirical mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. However, it is clear that this estimator has no finite ℓ_p sensitivity, for any p, for worst-case data. One simple way to get around this is to simply use a truncated mean. Because we know that $\|\mu\|_2 \le 1$, one can show that if $X_1, \dots, X_n \sim \mathcal{N}(\mu, I)$, then

$$\|X_i\|_1 \le O(d + \sqrt{d} \cdot \log(n/\beta)) \;, \quad \text{and}$$
$$\|X_i\|_2 \le O(\sqrt{d} + \sqrt{\log(n/\beta)}) \;,$$

for all $i = 1, \dots, n$, with probability $1 - \beta$. In other words, if we let

$$f_1(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i \cdot \mathbf{1}\left[\|X_i\|_1 \le O(d + \sqrt{d} \cdot \log(n/\beta))\right] \;, \quad \text{and}$$
$$f_2(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i \cdot \mathbf{1}\left[\|X_i\|_2 \le O(\sqrt{d} + \sqrt{\log(n/\beta)})\right] \;,$$

then with high probability, if $X_1, \dots, X_n \sim \mathcal{N}(\mu, I)$, we have that

$$f_1(X) = f_2(X) = \widehat{\mu} \;,$$

and moreover, $\Delta_1(f_1) = O(d + \sqrt{d} \cdot \log(n/\beta))/n$ and $\Delta_2(f_2) = O(\sqrt{d} + \sqrt{\log(n/\beta)})/n$. Combining these bounds with the guarantees above, we obtain:

Corollary 2.3. *Let $\varepsilon, \delta > 0$, and suppose that $\|\mu\|_2 \le 1$. Let $X = (X_1, \dots, X_n)$ be a set of n independent samples from $\mathcal{N}(\mu, I)$ for some $\|\mu\|_2 \le 1$.*

- *There is an ε-DP algorithm $\mathcal{A}(X)$ so that with high probability,*

$$\|\mathcal{A}(X) - \mu\|_2 \le \widetilde{O}\left(\frac{d^{3/2}}{\varepsilon n}\right) + \|\widehat{\mu} - \mu\|_2 \;.$$

- *There is an (ε, δ)-DP algorithm $\mathcal{A}(X)$ so that with high probability,*

$$\|\mathcal{A}(X) - \mu\|_2 \le \widetilde{O}\left(\frac{d\sqrt{\log 1/\delta}}{\varepsilon n}\right) + \|\widehat{\mu} - \mu\|_2 \;.$$

Here, $\widetilde{O}$ suppresses polylogarithmic factors in the input.

How do we interpret these results? A classic result from high-dimensional concentration bounds states that $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{d/n}$ with high probability. In other words, to obtain error α, one needs $n \ge O(d/\alpha^2)$ samples. If we attempt to achieve the same error using the guarantees of Corollary 2.3, we obtain that the two algorithms require

$$n \ge \widetilde{O}\left(\frac{d^{3/2}}{\varepsilon\alpha}\right) + O\left(\frac{d}{\alpha^2}\right) \;, \quad \text{and} \quad n \ge \widetilde{O}\left(\frac{d\sqrt{\log 1/\delta}}{\varepsilon\alpha}\right) + O\left(\frac{d}{\alpha^2}\right) \;,$$

respectively. In particular, the overhead we pay for pure DP is *polynomially worse* in d, even when ε is a large constant. On the other hand, the overhead we pay for approximate DP is only some polylogarithmic factors. However, this begs a natural question: is such a overhead for pure DP necessary?

3 Differentially private mean estimation via robust estimation

It turns out there is a very powerful, almost black-box way, to turn a robust estimator into a private one. This reduction as stated was first presented in two concurrent works of [1, 2], although qualitatively similar (but not as clean) connections have been made in earlier works, see e.g. [3, 4].

Specifically, let us consider a deterministic (although this assumption can be removed) estimator $f(X)$ which takes a dataset $X = (X_1, \dots, X_n)$ with the following guarantee: for some non-decreasing $\alpha : [0, 1] \to \mathbb{R}$, and for any $\eta \in [0, 1]$, if X is an η-corrupted set of samples from a distribution D with associated parameter θ, the algorithm outputs $\|f(X) - \theta\|_2 \leq \alpha(\eta)$ with probability at least $1 - \beta$. One can state these results for more general norms as well, but we will not do so in this lecture for simplicity.

The key idea will be the following procedure. Let α be some target error, and let η_0 be so that $\alpha(\eta_0) = \alpha$. Define the score function

$$s(\theta) = \min \{\|X' - X\|_0 : \|f(X') - \theta\|_2 \leq \alpha(\eta_0)\} \ . \tag{1}$$

That is, $s(\theta)$ is how many data points we would need to change from our dataset so that our robust estimator outputs something close to θ. If we cannot change any number of data points to obtain this guarantee, we declare that $s(\theta) = n$. Clearly, it is not hard to see that s is a 1-stable function, and hence the exponential mechanism instantiated with this score function is ε-DP. We also have the following utility guarantee:

Theorem 3.1. *For any η_0, let $s(\theta)$ be defined as in (1). Let $X_1, \dots, X_n$ be drawn from distribution D with parameter $\theta^* \in \mathbb{R}^d$. Let θ be drawn from the distribution with pdf p given by*

$$p(\theta) \propto \exp(-\varepsilon s(\theta)) \ .$$

Then, $\|\theta - \theta^*\|_2 \leq 2\alpha(\eta_0)$ with probability at least $1 - \beta$ as long as

$$n \geq \max_{\eta \in [\eta_0, 1]} \frac{d \cdot \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log 1/\beta + O(\log \eta n)}{\eta \varepsilon} \ .$$

Before we prove this theorem, let us see how it is useful. Consider again the setting of mean estimation for Gaussians with bounded mean. Here, by using the Tukey median, we know that as long as $n \gtrsim O(d/\alpha^2)$ and $\eta \leq 1/2$, we can obtain $\alpha(\eta) = O(\alpha + \eta)$, and for $\eta > 1/2$, since the mean has norm at most 1, we declare $\alpha(\eta) = 1$. Plugging in $\eta_0 = \alpha$ into the theorem statement, we see that this is maximized by $\eta = \eta_0$, in which case we obtain that

$$n \gtrsim \frac{d + \log 1/\beta}{\eta \alpha} + \frac{d}{\alpha^2}$$

samples suffices to guarantee α-closeness.

More generally, if we only have the guarantee that $\|\mu\|_2 \leq R$, one obtains the guarantee that

$$n \gtrsim \frac{d + \log 1/\beta}{\eta \alpha} + \frac{d \log R}{\alpha} + \frac{d}{\alpha^2}$$

samples suffice, and this is known to be tight [5]. Note that this improves upon the rate of the Laplace mechanism by polynomial factors.

One can also show that by plugging in the appropriate robust estimators, this mechanism recovers optimal statistical rates

Proof of Theorem 3.1. Condition on the event that the robustness guarantee holds; we know that this holds with probability at least $1 - \beta$ by assumption. For all $r > 0$, let B_r denote the ball of radius r around $f(X)$, and let $V_r \propto r^d$ denote its volume. Let θ have score ηn. By definition, we know that $\|f(X') - \theta\|_2 \leq \alpha(\eta_0)$ for some $\|X' - X\|_0 \leq \eta n$. By robustness, this implies that $\|\theta - \theta^*\|_2 \leq \alpha(\eta) + \alpha(\eta_0) \leq 2\alpha(\eta)$, so long as $\eta \geq \eta_0$, so in other words, $\theta \in B_{2\alpha(\eta)}$. Note also that if $\theta \in B_{\alpha(\eta_0)}$, then $s(\theta) = 0$, by definition.

The above argument shows that it suffices to show that with probability $1-\beta$, we sample a point θ with score at most $\eta_0 n$, so it suffices to find an upper bound on the probability that we sample a point with score t for $t > \eta_0 n$. Recall that the form of the probability distribution is

$$p(\theta) = \frac{1}{Z}\exp(-\varepsilon s(\theta)) \;, \quad \text{where} \quad Z = \int \exp(-\varepsilon s(\theta))d\theta \;.$$

We note that we can lower bound the normalization constant by $Z \geq V_{\alpha(\eta_0)}$, since every point in that set has score 0.

In particular, this implies that:

$$\begin{aligned}
\Pr\left[s(\theta) = t\right] &= \frac{\int_{\theta : s(\theta)=t} \exp(-\varepsilon t)d\theta}{\int \exp(-\varepsilon s(\theta))d\theta} \\
&\leq \frac{\exp(-\varepsilon t)\cdot V_{2\alpha(t/n)})}{Z} \\
&\leq \frac{\exp(-\varepsilon \eta n)\cdot V_{2\alpha(t/n)}}{V_{\alpha(\eta_0)}} \\
&\leq \exp(-\varepsilon t)\left(\frac{2\alpha(t/n)}{\alpha(\eta_0)}\right)^d \;.
\end{aligned}$$

Summing over all $t \geq \eta n$, we obtain:

$$\begin{aligned}
\Pr\left[s(\theta) \geq \eta_0 n\right] &\leq \sum_{t=\eta_0 n}^{n} \exp(-\varepsilon t)\left(\frac{2\alpha(t/n)}{\alpha(\eta_0)}\right)^d \\
&\leq \sum_{t=\eta_0 n}^{n} \exp(-\varepsilon t)\left(\frac{2\alpha(t/n)}{\alpha(\eta_0)}\right)^d \cdot t^2 \cdot (1/t^2) \\
&\leq O(1)\cdot \max_{\eta_0 \leq \eta \leq 1}\left((\eta n)^2 \cdot \exp(-\varepsilon \eta n)\cdot \left(\frac{2\alpha(t/n)}{\alpha(\eta_0)}\right)^d\right) \;.
\end{aligned}$$

Solving for n then yields the desired expression. □

Computational efficiency There is a major remaining issue with this reduction: namely, this gives a computationally inefficient algorithm. In particular, the score function itself seems difficult to compute, and even given the score function, implementing the exponential mechanism with this score function seems quite challenging. We will not have time to go over this in this lecture, but it turns out that both of these issues can be dealt with. At an extremely high level:

- While we cannot sample from $p \propto \exp(-\varepsilon s(\theta))$ for general s, we can efficiently sample from these distributions if s is *convex*; in this case, p is a log-concave distribution, and we have efficient samplers for such distributions, assuming appropriate access to s.
- It turns out that while s is hard to compute, for mean estimation, it turns out that there exists a computationally efficient proxy that can be computed using the Sum-of-Squares hierarchy that also simultaneously guarantees that s is convex; this is quite non-trivial, and we defer the reader to [1] for the details.

Putting these together yields an efficient private estimator for mean estimation that obtains pure DP guarantees with the optimal statistical rates.

References

[1] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023.

[2] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *International Conference on Machine Learning*, pages 1121–1146. PMLR, 2023.

[3] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[4] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.

[5] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *arXiv preprint arXiv:1905.13229*, 2019.