Lecture 17: Differentially private estimation I: univariate estimation

November 26, 2019

In this lecture, we'll show how to use some of the techniques presented in the previous lecture (as well as some additional ones) to get estimation procedures for univariate estimation tasks which are differentially private, and which pay only a minor cost when it comes to sample complexity.

Specifically, following [1], we will consider the task of learning a Gaussian to small total variation distance error under differential privacy. As in the case of robust estimation, we'll do this by decomposing the overall problem into two sub-problems, namely, private mean estimation and private covariance estimation. Before we will do so, we'll need a couple of additional private mechanisms.

Notation As a bit of notation, in this lecture (and as is relatively standard in private ML), we will use α to denote the accuracy parameter, i.e. the error to which we learn the quantities. This is just because ε is unfortunately already taken.

1 Private histograms

Recall that a histogram over a domain $\mathcal{X}$ has the following form. We have some partition of $\mathcal{X} = B_1 \cup \ldots \cup B_k$ where the B_i are disjoint, and a function $f : \mathcal{X}^n \to \mathbb{R}^k$ where $f(X)_i = |\{j : X_j \in B_i\}|$ is simply the histogram of how many samples land in each bucket. More generically, we can ask for a potentially *infinite* partition of $\mathcal{X}$. Formally, let $\mathcal{S} \subset 2^{\mathcal{X}}$ so that for all $S, S' \in \mathcal{S}$ we have that $S \cap S' = 0$, and moreover $\cup_{S \in \mathcal{S}} S = \mathcal{X}$. Then I can define a function $f : \mathcal{X}^n \to \mathbb{R}^{\mathcal{S}}$ analogously, namely, $f(X)_S = |\{j : X_j \in S\}|$. Given such a partition, we will call f the *associated histogram* function. Clearly this generalizes the notion of histogram queries as defined before.

Recall from last class that when the partition is finite, then the Laplace mechanism gives us the following utility and privacy guarantees:

Theorem 1.1. *Let $B_1, \ldots, B_k$ be a partition of $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}^k$ be the associated histogram, and let $\varepsilon, \eta > 0$. Then $\mathcal{M}^{\mathrm{Lap}}_{f,\varepsilon}$ is $(\varepsilon, 0)$-differentially private, and moreover,*

$$\Pr\left[\left\|f(x) - \mathcal{M}^{\mathrm{Lap}}_{f.\varepsilon}(x)\right\|_\infty \geq \frac{2}{\varepsilon}\log\frac{k}{\eta}\right] \leq \eta\,.$$

However, notice that the utility of this algorithm scales to ∞ as $k \to \infty$ (albeit logarithmically). It turns out that if we are okay with approximate differential privacy, this can be avoided, which allows us to take potentially infinitely many buckets:

Theorem 1.2 ([2, 3])**.** *Let $\mathcal{S} \subset 2^{\mathcal{X}}$ be a (potentially infinite) partition of $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}^{\mathcal{S}}$ be the associated histogram function, and let $\varepsilon, \delta > 0$. Then there is an algorithm $\mathcal{A}$ which is (ε, δ) differentially private, and moreover, with probability at least 0.99, outputs $\widehat{f}$ so that $\left\|f - \widehat{f}\right\|_\infty \lesssim \frac{\log 1/\delta}{\varepsilon}$.*

The proof of this is deferred to the homework.

2 Private mean estimation

Let's first consider the case when we know the variance exactly, and all we must do is learn the mean of the Gaussian. WLOG, by scaling, we may assume that the variance is 1. Then recall that to learn a Gaussian to total variation distance α, it was equivalent to learn the mean to error α:

Fact 2.1. *Let $\mu, \mu' \in \mathbb{R}$. Then*

$$\min(|\mu - \mu'|, 1) \lesssim \mathrm{d_{TV}}(\mathcal{N}(\mu, 1), \mathcal{N}(\mu', 1)) \lesssim |\mu - \mu'| \ .$$

In this section, we'll see several algorithms for this problem, in both the pure and approximate DP setting. As we'll see, the guarantees you can achieve for these two settings are somewhat different.

2.1 Pure DP algorithms

It turns out that for pure differential privacy, we will require some additional information about the unknown mean. Specifically, we will require that it lives within a known interval. WLOG let us assume that this interval is centered at 0 and has radius R, i.e. we know that $\mu \in [-R, R]$, where μ is the unknown mean. When given an interval not centered at zero, we can clearly simply shift it to be centered at zero, so this is without loss of generality.

Our first attempt will be relatively naive. Let $X_1, \ldots, X_n$ be our data set. Suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$, where μ is unknown but $\mu \in [-R, R]$. We simply observe that in this case, then by standard Gaussian concentration, we know that with extremely high probability, we have $|X_i| \leq R + O(\sqrt{\log n})$ for all $i \in [n]$. Thus, removing all points that exceed this threshold will remove no points in this case. Concretely, for some constant $C > 0$ sufficiently large, consider the estimator

$$\widetilde{\mu}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \cdot \mathbf{1}\left[|X_i| \leq R + C\sqrt{\log n}\right] \ .$$

Then we know that with high probability, if $X_i \sim \mathcal{N}(\mu, 1)$, then this will not throw away any point from out dataset with probability ≥ 0.99 (for C sufficiently large), and so this will behave like the empirical mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We already know how the empirical mean concentrates: from previous lectures, we know that with probability ≥ 0.99,

$$|\mu - \widehat{\mu}| \lesssim \sqrt{\frac{1}{n}} \ .$$

Moreover, by design, changing one point in the dataset can change $\widetilde{\mu}$ by at most $\frac{2R + 2C\sqrt{\log n}}{n}$. In particular, for any dataset, the ℓ_1 sensitivity of $\widetilde{\mu}$ is $\frac{2R + 2C\sqrt{\log n}}{n}$, and thus the randomized estimator

$$\mathcal{A}(X_1, \ldots, X_n) = \widetilde{\mu}(X_1, \ldots, X_n) + \mathrm{Lap}\left(\frac{2R + 2C\sqrt{\log n}}{n\varepsilon}\right)$$

is $(\varepsilon, 0)$-differentially private. By a standard concentration argument for Laplace random variables, we know that with probability ≥ 0.99, we have

$$|\mathcal{A}(X_1, \ldots, X_n) - \widetilde{\mu}(X_1, \ldots, X_n)| \lesssim \frac{R + C\sqrt{\log n}}{n\varepsilon} \ .$$

Combining this with the argument above, we have proven the following statement:

Lemma 2.2. *Let $\varepsilon > 0$, let $R > 0$, and let $X_1, \ldots, X_n \in \mathbb{R}$. There is an efficient $(\varepsilon, 0)$-differentially private algorithm $\mathcal{A}$ so that if X_i are i.i.d. samples from $\mathcal{N}(\mu, I)$ for some $\mu \in [-R, R]$, then with probability at least 0.97, we have*

$$|\mathcal{A}(X_1, \ldots, X_n) - \mu| \lesssim \sqrt{\frac{1}{n}} + \frac{R + \sqrt{\log n}}{n\varepsilon} \ .$$

In particular, for any $\alpha > 0$, if

$$n \gtrsim \frac{1}{\alpha^2} + \frac{R}{\alpha\varepsilon} + \frac{\sqrt{\log \frac{1}{\alpha\varepsilon}}}{\alpha\varepsilon}\ , \tag{1}$$

then $|\mathcal{A}(X_1, \ldots, X_n) - \mu| \leq \alpha$ with probability at least 0.97.

Let's parse (6) briefly. The first term is simply the rate of regular non-private mean estimation, so the second and third terms are what we pay for insisting on privacy. Notice that when $\varepsilon \approx \alpha$, then the third term disappears (up to log factors), and so we could hope to get privacy "for free" (up to logarithmic factors). However, the most salient term is arguably the second term: this algorithm pays a *linear* cost in the range. In particular, when R is large (as it may be *a priori*), we pay a lot for privacy, at least with this algorithm.

Luckily, it turns out we can dramatically improve this dependence. The main observation is that we can first get a relatively poor approximation of the mean, via a histogram, while preserving privacy. Then, we can use this constant-sized interval as a replacement for $[-R, R]$ in the above bound.

The algorithm proceeds as follows. Assume WLOG that R is an integer. Break up $[-R, R]$ into $2R - 2$ intervals, each of length 1. Denote these intervals in increasing order as $I_{-(R-1)}, \ldots, I_{-1}, I_1, \ldots, I_{R-1}$ (do something arbitrary with the boundaries). The key point is the following: if $X_1, \ldots, X_n$ are i.i.d. samples from μ, then with overwhelming probability, in the associated histogram function, the bucket that contains μ (or one of the ones around it) will contain the most points. As a result, we can run a private histogram algorithm, i.e. Theorem 1.1, on our input. This gives us a bucket I_j so that μ is close to I_j with high probability, and then we can simply run Lemma 2.2 with this a much smaller interval, and this will give us a much better dependence on R. Importantly, note that this process is an adaptive composition of two private algorithms, as the application of Lemma 2.2 uses the output of Theorem 1.1. However, this is okay, as long as we're okay with losing a constant factor in the accuracy, by composition of differential privacy.

Let's make this discussion formal. The main utility guarantee of the histogram algorithm follows from the following lemma:

Lemma 2.3. *Let f be the associated histogram function for the partition $I_{-(R-1)}, \ldots, I_{-1}, I_1, \ldots, I_{R-1}$. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$, where $\mu \in [-R, R]$. Then, with probability $\geq 1 - \exp(-\Omega(n))$, if I_j is the bucket with the most points in it, then $|\mu - j| \leq 5$. Moreover, we have that $f_j(X_1, \ldots, X_n) \geq 0.1n$, and $f_{j'}(X_1, \ldots, X_n) < 0.02n$ for all j' so that $|\mu - j| > 5$.*

Proof. Let A be the event that $|X - \mu| \leq 4$. Notice that if this occurs, then X lands within one of the buckets I_j where $|\mu - j| \leq 5$. Now we know that $\Pr_{X \sim \mathcal{N}(\mu,1)}[A] \geq 0.99$. In particular, if we let $Y_i = 1$ if $X_i \in A$ and 0 otherwise, then by a Chernoff bound, with probability $1 - \exp(-\Omega(n))$, we have that $\frac{1}{n}\sum Y_i \geq 0.98$. Condition on this event occurring for the result of the proof. If this happens, then this means that any bucket I_j so that $|j - \mu| > 5$ can have at most $0.02n$ points within it. On the other hand, this implies that $0.98n$ points fall within the 9 intervals $I = [j, j + 1]$ surrounding μ satisfying $|j - \mu| \leq 5$. This means that at least one of them must contain at least $0.98n/9 \approx 0.1n$ points, which means that the bucket containing the most points must be one of these 9 intervals, which concludes the proof. □

This yields:

Corollary 2.4. *Let $\varepsilon > 0$, let $R > 0$, and let $X_1, \ldots, X_n \in \mathbb{R}$. There is an efficient $(\varepsilon, 0)$-differentially private algorithm $\mathcal{A}$ so that if X_i are i.i.d. samples from $\mathcal{N}(\mu, I)$ for some $\mu \in [-R, R]$, then if $n \gtrsim \frac{\log R}{\varepsilon}$, then with probability at least $0.99 - \exp(-\Omega(n))$, we have that*

$$|\mathcal{A}(X_1, \ldots, X_n) - \mu| \leq 5\ .$$

Proof. The algorithm is straightforward: let f be the associated histogram function for the partition $I_{-(R-1)}, \ldots, I_{-1}, I_1, \ldots, I_{R-1}$, and let j be the maximum coordinate of $\mathcal{M}^{\mathrm{Lap}}_{f,\varepsilon}(X_1, \ldots, X_n)$. By Theorem 1.1 and our choice of n, we know that this is $(\varepsilon, 0)$-differentially private, and moreover, if c_ℓ is the ℓ-th coordinate of $\mathcal{M}^{\mathrm{Lap}}_{f,\varepsilon}(X_1, \ldots, X_n)$, then $|c_\ell - f_\ell| \leq 0.02n$ for all ℓ, with probability at least 0.99. Combining this with Lemma 2.3 immediately yields the desired result. □

Combining the output of this algorithm with the Laplace noise-based approach with the ε for both algorithms internally set to $\varepsilon/2$ and invoking composition of differential privacy yields:

Theorem 2.5. *Let $\varepsilon > 0$, let $R > 0$, and let $X_1, \dots, X_n \in \mathbb{R}$. There is an efficient $(\varepsilon, 0)$-differentially private algorithm $\mathcal{A}$ so that if X_i are i.i.d. samples from $\mathcal{N}(\mu, I)$ for some $\mu \in [-R, R]$, then with probability at least $0.97 - \exp(-\Omega(n))$, if $n \gtrsim \frac{\log R}{\varepsilon}$, then we have*

$$|\mathcal{A}(X_1, \dots, X_n) - \mu| \lesssim \sqrt{\frac{1}{n}} + \frac{\sqrt{\log n}}{n\varepsilon} .$$

In particular, for any $\alpha > 0$, if

$$n \gtrsim \frac{1}{\alpha^2} + \frac{\log R}{\varepsilon} + \frac{\sqrt{\log \frac{1}{\alpha\varepsilon}}}{\alpha\varepsilon} , \tag{2}$$

then $|\mathcal{A}(X_1, \dots, X_n) - \mu| \leq \alpha$ with probability at least 0.96.

This is much better! In particular, we get privacy "for free" in a much larger range of parameters. However, one could still ask if this could be improved. Unfortunately, it turns out that for pure differential privacy, $\log R$ is unavoidable:

Theorem 2.6. *Suppose $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}$ is an $(\varepsilon, 0)$-differentially private algorithm, which, given $(X_1, \dots, X_n) \sim \mathcal{N}(\mu, I)$ for some $\mu \in [-R, R]$, outputs μ' so that with probability $\geq 1/2$, we have that $|\mu' - \mu| < 0.1$. Then $n \gtrsim \frac{\log R}{\varepsilon}$.*

Proof. Let $Z_1, \dots, Z_n \sim \mathcal{N}(0, 1)$. By correctness, we must have that

$$\Pr\left[|\mathcal{A}(Z_1, \dots, Z_n)| < 0.1\right] \geq 1/2 . \tag{3}$$

On the other hand, suppose that $X_1, \dots, X_n \sim \mathcal{N}(j, 1)$ for some $j \neq 0$. Then by correctness, we must have that

$$\Pr\left[|\mathcal{A}(X_1, \dots, X_n) - j| < 0.1\right] \geq 1/2 . \tag{4}$$

However, by group privacy (just by replacing all of the X_i with Z_i), this implies that

$$\Pr\left[|\mathcal{A}(Z_1, \dots, Z_n) - j| < 0.1\right] \geq e^{-\varepsilon n}/2 . \tag{5}$$

Consider the events $E_j = \{|\mathcal{A}(Z_1, \dots, Z_n) - j| < 0.1\}$. These events are obviously disjoint, and moreover $\Pr[E_0] \geq 1/2$, which in particular implies that $\sum_{j \neq 0} \Pr[E_j] < 1/2$. Thus, we have that

$$(2R - 3)\frac{\varepsilon^{-\varepsilon n}}{2} \leq \sum_{j \neq 0} \Pr[E_0] < \frac{1}{2} ,$$

which by solving for n immediately implies the desired result. □

This turns out to be a special case of a fairly widespread phenomena. In particular, as long as there is a "packing" of distinct problem instances of size S, the same technique will give you that $(\varepsilon, 0)$-differential privacy will require $\Omega(\log S)$ samples. One can also show that the third term in this expression is unavoidable, up to the log factor:

Lemma 2.7 ([4])**.** *Let P, Q be two distributions over $\mathcal{X}$ so that $\mathrm{d_{TV}}(P, Q) = \alpha$. Then, suppose there is an $(\varepsilon, 0)$-differentially private algorithm $\mathcal{A}$, which, given n samples from either P or Q, correctly classifies if the samples are from P or from Q, with probability at least 0.99. Then $n = \Omega(\frac{1}{\alpha\varepsilon})$.*

Proof. Recall one definition of total variation distance: that there exists a coupling (X, Y) of P and Q so that $\Pr[X \neq Y] \leq \alpha$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n i.i.d. copies of this coupled random variable. Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. Since $X_1, \ldots, X_n$ are n i.i.d. samples from P, then $\Pr_{\mathcal{A},X}[\mathcal{A}(X) \text{ outputs } P] > 0.99$, and similarly $\Pr_{\mathcal{A},Y}[\mathcal{A}(Y) \text{ outputs } Q] > 0.99$. For any $z \in \mathcal{X}^n$, and any $R \in \{P, Q\}$, let

$$f_R(z) = \Pr_{\mathcal{A}}[\mathcal{A}(z) \text{ outputs } R] \, .$$

Then in particular we have $\mathbb{E}_X[F_P(X)] \geq 0.99$, so since f_R is a $[0,1]$-valued function, we have that $\Pr_X[F_P(X) > 0.9] \geq 0.9$, and similarly $\Pr_Y[F_Q(Y) > 0.7] \geq 0.9$. Moreover, we have $\mathbb{E}\left[\|X - Y\|_0\right] \leq \alpha n$, and hence by Markov's inequality, $\Pr\left[\|X - Y\|_0 > 10\alpha n\right] < 0.1$. Thus, by a union bound, we may assume that the following three events occur simultaneously with probability at least 0.7:

$$\begin{aligned} F_P(X) &> 0.9 \, , \\ F_Q(Y) &> 0.9 \, , \\ \|X - Y\|_0 &< 10\alpha n \, . \end{aligned}$$

Condition on these three events occurring simultaneously. By group privacy, we have

$$0.9 < F_P(X) = \Pr_{\mathcal{A}}[\mathcal{A}(X) \text{ outputs } P] \leq \varepsilon^{10\alpha n} \cdot \Pr_{\mathcal{A}}\left[\mathcal{A}(Y) \text{ outputs } P\right] \leq e^{10\varepsilon\alpha n} \cdot 0.1 \, .$$

Solving this expression for n yields the desired result. □

2.2 Approximate DP algorithms

It turns out that in general, a dependence of the form $\frac{1}{\alpha\varepsilon}$ is unavoidable for essentially notion of privacy. However, this term is generally pretty minor: when $\varepsilon = \Theta(1)$, this is often dominated by the non-private rate. On the other hand, if we're willing to compromise a little, and ask for only approximate differential privacy, then we can avoid any dependence on R. In particular, the only part of the above algorithm which requires a dependence on R is the first histogramming step. Specifically, if we're okay with approximate differential privacy, we can invoke Theorem 1.2 with an infinite decomposition of $\mathbb{R}$ into unit length intervals instead of Theorem 1.1, and then follow the same recipe as before. Straightforward computation yields:

Theorem 2.8. *Let $\varepsilon, \delta > 0$, and let $X_1, \ldots, X_n \in \mathbb{R}$. There is an efficient (ε, δ)-differentially private algorithm $\mathcal{A}$ so that if X_i are i.i.d. samples from $\mathcal{N}(\mu, I)$ for some $\mu \in \mathbb{R}$, then with probability at least $0.97 - \exp(-\Omega(n))$, if $n \gtrsim \frac{\log 1/\delta}{\varepsilon}$, then we have*

$$|\mathcal{A}(X_1, \ldots, X_n) - \mu| \lesssim \sqrt{\frac{1}{n}} + \frac{\sqrt{\log n}}{n\varepsilon} \, .$$

In particular, for any $\alpha > 0$, if

$$n \gtrsim \frac{1}{\alpha^2} + \frac{\log 1/\delta}{\varepsilon} + \frac{\sqrt{\log \frac{1}{\alpha\varepsilon}}}{\alpha\varepsilon} \, , \tag{6}$$

then $|\mathcal{A}(X_1, \ldots, X_n) - \mu| \leq \alpha$ with probability at least 0.96.

References

[1] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[2] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *Journal of Machine Learning Research*, 20(94):1–34, 2019.

[3] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

[4] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6879–6891. Curran Associates Inc., 2018.