CS395T: Continuous Algorithms, Part XIX
Differential privacy

Kevin Tian

1 Basic definitions

Privacy is a major legal and ethical consideration when designing algorithms in the age of big data. When not carefully quantified, purportedly "private" algorithms can be susceptible to a variety of creative breaches (see e.g. the first lecture of [Kam20] , which gives many such examples of privacy attacks on the taxi industry, a challenge by Netflix, neural networks memorizing user data, genomic studies, a group insurance commission, etc.) Many of these examples share common themes, such as exploiting side information or low sample sizes in group summary statistics.

Perhaps surprisingly, it is possible to give a meaningful mathematical definition of privacy that provides provable guarantees against such attacks. Moreover, the source of security in this definition is information-theoretic, rather than computational as is commonly the case in cryptography (e.g. based on the conjectured hardness of solving a computational task). This definition has even recently been deployed in practice by aspects of the U.S. census data collection. The purpose of this lecture is to describe *differential privacy* and how to use it to design private algorithms.

Definition 1 (Differential privacy)**.** *Let $n \in \mathbb{N}$, $\epsilon \in \mathbb{R}_{\geq 0}$, and $\delta \in [0, 1]$. Let $\mathcal{M} : \mathcal{S}^n \to \Omega$ be a mechanism (i.e. a randomized algorithm) which acts on a dataset of n elements of $\mathcal{S}$, and returns an outcome in a sample space Ω. We say $\mathcal{M}$ satisfies* (ϵ, δ)-differential privacy *(or, $\mathcal{M}$ is (ϵ, δ)-DP) if for all neighboring datasets $D, D' \in \mathcal{S}^n$, i.e. datasets which differ only on one entry,*

$$\Pr[\mathcal{M}(D) \in A] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in A] + \delta, \text{ for all } A \subseteq \Omega. \tag{1}$$

If $\delta = 0$, we say $\mathcal{M}$ satisfies ϵ-differential privacy (or, $\mathcal{M}$ is ϵ-DP) for short.

In particular, we often think of $\mathcal{S}$ as a set of individuals from a population whose privacy we are trying to protect, so that the input to a mechanism is a set of n individuals from $\mathcal{S}$. The case of (ϵ, δ)-DP with $\delta = 0$ is sometimes called pure DP, whereas $\delta > 0$ is called approximate DP.

Remark 1. *Typically when designing (ϵ, δ)-differentially private algorithms, we wish for the additive parameter δ to be very small (e.g. polynomially small in the dataset size), whereas even constant values of ϵ provide strong privacy guarantees. Roughly speaking, this is because δ is often treated as a "failure probability" on ϵ-DP holding (akin to failure probabilities in randomized algorithms); with probability $\approx \delta$, there are no guarantees on the behavior of the algorithm. For instance, the algorithm could output some answer that is blatantly nonprivate with this probability, such as all of the features of some element in the database. Due to this asymmetry between the ϵ and δ parameters, we often aim for algorithms whose runtime or sample complexity scales polylogarithmically in $\frac{1}{\delta}$, whereas in many cases a polynomial dependence on $\frac{1}{\epsilon}$ is unavoidable. In Section 2, we develop another notion of privacy which provides smooth tradeoff curves between the ϵ and δ parameters, so that the probability of a blatantly nonprivate event is bounded "at every scale."*

One basic consequence of Definition 1 is that it generalizes straightforwardly to two datasets at arbitrary Hamming distance (i.e. the number of elements which differ between the two datasets).

Lemma 1. *Let $\mathcal{M} : \mathcal{S}^n \to \Omega$ be (ϵ, δ)-DP, and let D, D' have Hamming distance k. Then,*

$$\Pr[\mathcal{M}(D) \in A] \leq \exp(k\epsilon) \Pr[\mathcal{M}(D') \in A] + \delta \cdot \frac{\exp(k\epsilon) - 1}{\exp(\epsilon) - 1}, \text{ for all } A \subseteq \Omega.$$

Proof. Note that there is a sequence $D_0 = D, D_1, \ldots D_k = D'$ such that each pair of D_{i-1}, D_i are neighboring (for all $i \in [k]$). Iteratively applying (1) and simplifying yields the claim. □

We remark that for small values of $\epsilon \ll \frac{1}{k}$, we have $\frac{\exp(k\epsilon)-1}{\exp(\epsilon)-1} \approx \frac{k\epsilon}{\epsilon} = k$.

To gain some intuition for (1), note that it is closely-related to our definition of the total variation distance (Definition 4, Part XI). Indeed, because we can always view the outcome of a randomized algorithm as a distribution (where the particular distribution depends on the input), suppose that instead we claimed that $D_{\mathrm{TV}}(\mathcal{M}(D), \mathcal{M}(D')) \leq \epsilon$, for all neighboring datasets D, D'. This would imply that the outcomes $\mathcal{M}(D), \mathcal{M}(D')$ are stable in the following sense (by Fact 1, Part XI):

$$\Pr[\mathcal{M}(D) \in A] \leq \Pr[\mathcal{M}(D') \in A] + \epsilon, \text{ for all } A \subseteq \Omega. \tag{2}$$

Instead of the additive stability afforded by total variation distance bounds, differential privacy (1) asks for a *multiplicative* notion of stability. That is, small-probability events according to running an algorithm on one dataset should not become much larger according to a neighboring dataset.

Another motivation for the definition (1) is from the perspective of "plausible deniability." Suppose you design an experiment to learn a truth about the world, e.g. whether smoking causes cancer, by collecting statistics from n individuals. If we are to trust this experiment's conclusion, the outcome should not particularly depend on whether a particular individual was included. On the other hand, a patient's medical history is sensitive, and it is preferable that an auditor who only looks at the released statistics from the experiment (and not the dataset itself) should not be able to determine whether an individual was included. Definition 1 provides such a guarantee: indeed, (1) states no outcome is significantly more likely whether or not a patient is included. To formalize our argument that (1) guarantees plausible deniability, we provide the following intuitive claim.

Lemma 2. *Let $F : \Omega \to \Omega'$ be an arbitrary, potentially randomized, function. If $\mathcal{M} : \mathcal{S}^n \to \Omega$ is (ϵ, δ)-DP, then so is $F \circ \mathcal{M} : \mathcal{S}^n \to \Omega'$.*

Proof. We first prove this when F is a deterministic map f. Let $A' \subseteq \Omega'$ be a fixed event, and let $A := f^{-1}(A')$ denote the set of outcomes $\omega \in \Omega$ such that $f(\omega) \in A'$. Then for neighboring $D, D' \in \mathcal{S}^n$, the fact that $\mathcal{M}$ is (ϵ, δ)-DP guarantees

$$\begin{aligned}\Pr\left[f\left(\mathcal{M}\left(D\right)\right) \in A'\right] &= \Pr\left[\mathcal{M}\left(D\right) \in A\right] \\ &\leq \exp(\epsilon) \Pr\left[\mathcal{M}\left(D'\right) \in A\right] + \delta = \exp(\epsilon) \Pr\left[f(\mathcal{M}(D')) \in A'\right] + \delta.\end{aligned}$$

Finally, we can view a randomized function F as applying f drawn from a distribution $\mathcal{F}$ over deterministic functions, from which the conclusion follows from the above display and

$$\begin{aligned}\Pr\left[F(\mathcal{M}(D)) \in A\right] &= \mathbb{E}_{f \sim \mathcal{F}}\left[\Pr\left[f(\mathcal{M}(D)) \in A\right]\right] \\ &\leq \mathbb{E}_{f \sim \mathcal{F}}\left[\exp(\epsilon) \Pr\left[f(\mathcal{M}(D')) \in A'\right] + \delta\right] = \exp(\epsilon) \Pr\left[F(\mathcal{M}(D')) \in A'\right] + \delta.\end{aligned}$$

□

In other words, Lemma 2 says that we can apply any postprocessing function to the outcome of a differentially private mechanism, and the composed outcome continues to be private, as long as our postprocessing is independent of the dataset D. This is intuitive, as it means we only pay a privacy loss if we look at the dataset again before making further decisions. Lemma 2 also implies plausible deniability: if our randomized function F aims to infer membership in a dataset D by only using outcomes of an experiment run on D, it cannot do a good job distinguishing D and a neighboring dataset which excludes said member, and hence fails at its inference task.

The focus of this lecture will be privacy-utility tradeoffs: how does enforcing a certain level of differential privacy interfere with our ability to make accurate judgments, when can this be mitigated, and when is a utility loss necessary? As we will see, differential privacy also ends up being quite a fundamental idea in characterizing the *stability* of randomized algorithms, and can be used to obtain other strong properties of statistical algorithms. For instance, it can ensure *generalization* of an algorithm used to learn about a population from samples. Intuitively, this is because (1) implies we can replace one element of D with a fresh draw from the population (not used in conducting the experiment), and the outcome should not change greatly. We will formalize this notion, and additional *adaptivity* guarantees enjoyed by DP algorithms, in the next lecture.

To introduce privacy-utility tradeoffs, we begin with a simple example of mean estimation in $[0, 1]$.[1]

[1] For instance, this captures the problem of the census releasing a private estimate of the proportion of residents in a county who satisfy some property, which is an average of $\{0, 1\}$ indicator variables.

Lemma 3. *Let $D := \{X_i\}_{i\in[n]} \subseteq [0,1]$, $\alpha, \epsilon \in \mathbb{R}_{>0}$, and let $\bar{X} := \frac{1}{n}\sum_{i\in[n]} X_i$ be the empirical mean of D. If $n \geq \frac{2}{\alpha\epsilon}$, there is an ϵ-DP mechanism $\mathcal{M}$ which satisfies*

$$\mathbb{E}_{\widehat{X}\sim\mathcal{M}(D)}\left[\left|\widehat{X} - \bar{X}\right|\right] \leq \alpha. \tag{3}$$

Proof. The mechanism is simply to return $\bar{X} + \xi$ for $\xi \sim \mathrm{Lap}(\alpha)$, where $\mathrm{Lap}(b)$ is the *Laplace distribution* on $\mathbb{R}$ with density function

$$\pi(\xi) = \frac{1}{2b}\exp\left(-\frac{|\xi|}{b}\right). \tag{4}$$

The utility guarantee (3) follows immediately from the fact that $\mathbb{E}_{\xi\sim\mathrm{Lap}(b)}|\xi| = b$.[2] To prove privacy, the key observation is that for a neighboring dataset $D' = \{X'_i\}_{i\in[n]}$ to D (so that $X'_i = X_i$ for all $i \in [n]$ except for $i = j$), if we denote $\bar{X}' := \frac{1}{n}\sum_{i\in[n]} X_i$, we have

$$\left|\bar{X} - \bar{X}'\right| = \frac{1}{n}\left|X_j - X'_j\right| \leq \frac{2}{n}. \tag{5}$$

Hence, the empirical mean of a dataset in $[0,1]^n$ is a $\frac{2}{n}$-sensitive statistic; moving to a neighboring dataset can affect the true answer by at most $\frac{2}{n}$. Now (1) follows, because for any outcome $y \in \mathbb{R}$,

$$\frac{\Pr[\mathcal{M}(D) = y]}{\Pr[\mathcal{M}(D') = y]} = \frac{\exp\left(-\frac{|y-\bar{X}|}{\alpha}\right)}{\exp\left(-\frac{|y-\bar{X}'|}{\alpha}\right)} \leq \exp\left(\frac{|\bar{X} - \bar{X}'|}{\alpha}\right) \leq \exp\left(\frac{2}{\alpha n}\right) \leq \exp(\epsilon).$$

The first equality used the definition of $\mathrm{Lap}(\alpha)$ (4), the first inequality was the triangle inequality, the second used our sensitivity bound (5), and the last used our assumption $n \geq \frac{2}{\alpha\epsilon}$. □

The strategy in Lemma 3 is an instance of the *Laplace mechanism*, which applies generically when computing statistics of a dataset with bounded sensitivity. Indeed, more generally if a statistic is Δ-sensitive, we can ensure privacy by adding $\mathrm{Lap}(\frac{\Delta}{\epsilon})$ noise; Lemma 3 used this observation with $\Delta = \frac{2}{n}$. Unsurprisingly, it was crucial in Lemma 3 that the domain was bounded (which can often be ensured by clipping), and that n was sufficiently large. That is, the sensitivity of an empirical mean decays as n grows, meaning the statistic relies less on each individual (so privacy is easier).

Also, note that the problem of designing a differentially private mechanism is really just about sampling from an appropriate data-dependent distribution (e.g. a Laplace distribution). There is in fact a meta-algorithm which in some sense generalizes the Laplace mechanism, called the *exponential mechanism*, which generates a sample from an appropriate Gibbs distribution $\propto \exp(-V(x))$. This algorithm has intimate connections to our previous unit on sampling, see e.g. Section 3.4 of [DR14] or Lecture 7 of [Kam20] for a general exposition, and [GLL22] for an efficient implementation.

We now ask: how do we bound the differential privacy of an algorithm which needs to access a dataset multiple times to perform its computation, so Lemma 2 (which only applies when using a function F that is independent of D) does not hold? For instance, we could ask for multiple linear statistics of a dataset D (e.g. the mean, reweighted mean, subgroup mean, etc.), each of which depends on D and loses privacy. We can model such an algorithm as a composition of mechanisms which access the dataset once (and incur some privacy cost for doing so), and then performs some computations not using the dataset (which incurs no privacy cost, due to Lemma 2). Therefore, this question is about accounting for the total privacy loss of a mechanism $\mathcal{M} = \mathcal{M}_k \circ \ldots \circ \mathcal{M}_1$, where $\mathcal{M}_i$ models the i^{th} access of the dataset. We have the following composition theorem.

Theorem 1 (Privacy composition)**.** *Let $\mathcal{M}_i : \mathcal{S}^n \times \prod_{j\in[i-1]} \Omega_j \to \Omega_i$ for all $i \in [k]$ be (ϵ, δ)-differentially private mechanisms.*[3] *Then the composition $\mathcal{M} := \mathcal{M}_k \circ \ldots \circ \mathcal{M}_1 : \mathcal{S}^n \to \Omega_k$ satisfies $(k\epsilon, k\delta)$-differential privacy, as well as $(\epsilon', k\delta + \delta')$-differential privacy for any $\delta' > 0$ and*

$$\epsilon' := \sqrt{2k\log\left(\frac{1}{\delta'}\right)}\epsilon + k\epsilon(\exp(\epsilon) - 1). \tag{6}$$

[2] We can prove strong tail bounds on the error, e.g. $\xi \sim \mathrm{Lap}(b)$ satisfies $|\xi| \leq b\log(\frac{1}{\delta})$ with probability $\approx \delta$.

[3] We can let $\mathcal{M}_i$ depend on the outputs of all the previous mechanisms, which live in $\prod_{j\in[i-1]} \Omega_j$.

We pause to interpret Theorem 1. The first result, i.e. $(k\epsilon, k\delta)$-DP of the composition, behaves exactly as one would expect, and we will prove it in the special case $\delta = 0$ shortly. For this reason it is called *basic composition* in the literature. The second result, (6) is more sophisticated and shows that if we are willing to pay a small additive error in the δ parameter, the growth in the ϵ parameter is actually more like $\sqrt{k}$ instead of the k factor given by basic composition. Intuitively, this is because the *privacy loss* parameter, the logarithm of the density ratio between neighboring datasets, behaves like a $\pm\epsilon$ random variable. Hence, summing k times scales as $\sqrt{k}\epsilon$ with high probability via a Chernoff bound; formalizing this argument loses an additive term in (6). If $\sqrt{k}\epsilon \lesssim 1$ (so the privacy of Theorem 1 is at most a constant), then this $\approx k\epsilon^2$ term is low-order.

On the other hand, the bound (6) is somewhat unwieldy, and to our knowledge it is not straightforward to state a generalization to uneven privacy parameters (e.g. if $\mathcal{M}_i$ is (ϵ_i, δ_i)-DP). For this reason we defer a full proof to Theorem B.1, [DR14] (for basic composition) and Theorem 3.20, [DR14] (for advanced composition). In the following Section 2, we present an alternative privacy accounting scheme, Rényi differential privacy (RDP), which we find somewhat more flexible. As a side effect, we will see that RDP qualitatively recovers Theorem 1 in a simple way.

As a first example, we now prove Theorem 1 in the special case $k = 2$, $\delta = 0$.

Lemma 4. *Let $\mathcal{M}_1 : \mathcal{S}^n \to \Omega_1$ be an ϵ_1-DP mechanism, and let $\mathcal{M}_2 : \mathcal{S}^n \times \Omega_1 \to \Omega_2$ be an ϵ_2-DP mechanism. Then the composition $\mathcal{M}_2 \circ \mathcal{M}_1 : \mathcal{S}^n \to \Omega_2$ is $(\epsilon_1 + \epsilon_2)$-DP.*

Proof. Let $D, D' \in \mathcal{S}^n$ be neighboring datasets, and let $\omega_2 \in \Omega_2$ be an arbitrary event. Then,

$$\begin{aligned}\frac{\Pr\left[\mathcal{M}_2(D, \mathcal{M}_1(D)) = \omega_2\right]}{\Pr\left[\mathcal{M}_2(D', \mathcal{M}_1(D')) = \omega_2\right]} &= \frac{\mathbb{E}_{\omega_1 \sim \mathcal{M}_1(D)}\left[\Pr\left[\mathcal{M}_2(D, \omega_1) = \omega_2\right]\right]}{\mathbb{E}_{\omega_1 \sim \mathcal{M}_1(D')}\left[\Pr\left[\mathcal{M}_2(D', \omega_1) = \omega_2\right]\right]} \\ &\le \exp(\epsilon_2) \cdot \frac{\mathbb{E}_{\omega_1 \sim \mathcal{M}_1(D)}\left[\Pr\left[\mathcal{M}_2(D, \omega_1) = \omega_2\right]\right]}{\mathbb{E}_{\omega_1 \sim \mathcal{M}_1(D')}\left[\Pr\left[\mathcal{M}_2(D, \omega_1) = \omega_2\right]\right]} \le \exp(\epsilon_1 + \epsilon_2),\end{aligned}$$

where the first inequality used that $\mathcal{M}_2(\cdot, \omega_1)$ is DP for all $\omega_1 \in \Omega$, and the second used that because $\mathcal{M}_1$ is DP, $\Pr[\mathcal{M}_1(D) = \omega_1] \le \exp(\epsilon_1)\Pr[\mathcal{M}_1(D') = \omega_1]$ for all $\omega_1 \in \Omega$. □

By recursively applying Lemma 4, we give a simple proof of basic composition when $\delta = 0$.

Corollary 1. *Let $\mathcal{M}_i : \mathcal{S}^n \times \prod_{j \in [i-1]} \Omega_j \to \Omega_i$ be an ϵ_i-DP mechanism, for all $i \in [k]$. Then the composition $\mathcal{M} := \mathcal{M}_k \circ \ldots \circ \mathcal{M}_1 : \mathcal{S}^n \to \Omega_k$ is $\sum_{i \in [k]} \epsilon_i$-DP.*

Proof. We induct on k; the base case $k = 2$ is Lemma 4. Now if the statement of the corollary is true for $k \leftarrow k - 1$, we can apply Lemma 4 with $\mathcal{M}_1 \leftarrow \mathcal{M}_{k-1} \circ \ldots \circ \mathcal{M}_1$, but extending its output space to $\Omega_1 \leftarrow \prod_{j \in [k-1]} \Omega_j$ (i.e. we concatenate the outputs of the first $k - 1$ mechanisms), as well as $\mathcal{M}_2 \leftarrow \mathcal{M}_k$. Induction then follows from Lemma 4 and the inductive hypothesis. □

Indeed, Corollary 1 extends (a special case of) the basic composition result in Theorem 1, to hold for a family of ϵ_i-DP mechanisms $\mathcal{M}_i$ with unequal values of ϵ_i. We now demonstrate the power of composition by privately computing multiple linear functions of a dataset. Specifically, we give a private mechanism for releasing *one-way marginals* of a dataset $D \in [0,1]^{n \times d} \equiv ([0,1]^d)^n$. Here, we view the samples as elements of $\mathcal{S} := [0,1]^d$ (e.g. n individuals each with d features), and our goal is to privately compute empirical averages of each feature, which we call the one-way marginals.

Proposition 1 (One-way marginals)**.** *Let $D := \{X_i\}_{i \in [n]} \in ([0,1]^d)^n$, $\alpha, \epsilon \in (0,1)$, and let $\mu(D) \in [0,1]^d$ be the one-way marginals of D, i.e. $\mu(D)]_j := \frac{1}{n}\sum_{i \in [n]} X_{ij}$ for all $j \in [d]$. If $n \ge \frac{2d}{\alpha\epsilon}$, there is an ϵ-DP mechanism $\mathcal{M}$ which satisfies*

$$\mathbb{E}_{\hat{\mu} \sim \mathcal{M}(D)}\left[|\hat{\mu}_j - [\mu(D)]_j|\right] \le \alpha \text{ for all } j \in [d]. \tag{7}$$

Moreover, if $\delta \in (0,1)$ and $n \ge \frac{6\sqrt{d \log \frac{1}{\delta}}}{\alpha\epsilon}$, there is an (ϵ, δ)-DP mechanism $\mathcal{M}$ which satisfies (7).

Proof. To see the first claim, we simply let $\mathcal{M}_j$ apply the mechanism from Lemma 3 to the j^{th} column, with a privacy parameter $\epsilon \leftarrow \frac{\epsilon}{d}$. The composition of these mechanisms is then ϵ-DP by

Theorem 1 with $k \leftarrow d$.[4] To see the second claim, we instead apply the mechanism from Lemma 3 to the j^{th} column, with a privacy parameter $\epsilon \leftarrow \frac{\epsilon}{3\sqrt{d \log \frac{1}{\delta}}}$. Then, Theorem 1 yields (ϵ, δ)-DP, as

$$\sqrt{2d \log\left(\frac{1}{\delta}\right)} \cdot \frac{\epsilon}{3\sqrt{d \log\left(\frac{1}{\delta}\right)}} + 2d \cdot \left(\frac{\epsilon}{3\sqrt{d \log\left(\frac{1}{\delta}\right)}}\right)^2 \leq \epsilon,$$

where we used that $\exp(c) - 1 \leq 2c$ for $c \leq 1$ to bound the second term in (6). □

We can see that Proposition 1 already illustrates the advantage of advanced composition over basic composition; if we are willing to tolerate a small failure probability, we can improve the sample complexity of privately computing d one-way marginals from $\approx d$ to $\approx \sqrt{d}$. Both of the bounds in Proposition 1 turn out to be tight [HT10, SU16, BUV18]. Interestingly, the expected per-coordinate guarantee in (7) can be boosted to an expected all-coordinates guarantee, i.e. of the form $\mathbb{E}_{\hat{\mu} \sim \mathcal{M}(D)}[\|\hat{\mu} - \mu(D)\|_\infty] \leq \alpha$, with the same sample complexities up to constant factors as in Proposition 1 [GKM21, DK22, CLN+23], by using a careful recursive filtering scheme.[5]

2 Rényi differential privacy

In this section, we introduce an alternative framework for parameterizing privacy guarantees, called Rényi differential privacy (RDP). RDP allows for somewhat simpler and tighter calculations in many cases when working with approximate differential privacy (i.e. (ϵ, δ)-DP with $\delta > 0$). To introduce this framework, we first state the central objects used in its definition: the *Rényi divergences*, a family of distances between probability distributions.

Definition 2 (Rényi divergence)**.** *For $\alpha > 1$ and two distributions P, Q on the same sample space Ω, we define their* Rényi divergence of order α *by*[6]

$$D_\alpha(P\|Q) := \frac{1}{\alpha - 1} \log\left(\int \left(\frac{P(\omega)}{Q(\omega)}\right)^{\alpha-1} P(\omega) \mathrm{d}\omega\right) = \frac{1}{\alpha - 1} \log\left(\mathbb{E}_{\omega \sim P}\left[\left(\frac{P(\omega)}{Q(\omega)}\right)^{\alpha-1}\right]\right). \quad (8)$$

Intuitively, the definition (8) asks for $\approx (\alpha - 1)$ moments of the density ratio $\frac{P}{Q}$ to be bounded. The Rényi divergence is nonincreasing in α, which reflects this intuition; the more moments we want to bound, the worse the bound (by Jensen's inequality). Additionally, as $\alpha \to 1$ the Rényi divergence approaches the KL divergence $\mathbb{E}_Q[\log \frac{P}{Q}]$, so D_1 is often used synonymously with D_{KL}. Another notable Rényi divergence is $D_2(P\|Q) = \log(1 + \chi^2(P\|Q))$, where χ^2 is the chi-squared divergence. We are now ready to define the Rényi differential privacy of a mechanism.

Definition 3 (Rényi differential privacy)**.** *Let $n \in \mathbb{N}$, $\alpha > 1$, and $\rho \in \mathbb{R}_{>0}$. Following notation in Definition 1, we say that a mechanism $\mathcal{M} : \mathcal{S}^n \to \Omega$ satisfies* (α, ρ)-Rényi differential privacy *(or, $\mathcal{M}$ is (α, ρ)-RDP) if for all neighboring datasets $D, D' \in \mathcal{S}^n$,*

$$D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \alpha\rho. \quad (9)$$

If (9) holds for all values of α simultaneously for all neighboring datasets $D, D' \in \mathcal{S}^n$, we say that $\mathcal{M}$ satisfies ρ-concentrated differential privacy *(or, $\mathcal{M}$ is ρ-CDP).*[7]

The CDP definition is appealing, because it only has one parameter. Conveniently, the ubiquitous *Gaussian mechanism* in any dimension (i.e. adding Gaussian noise to a statistic) satisfies CDP. We begin by computing the Rényi divergence between two spherical Gaussians with unequal means.

Lemma 5. *Let $\mu \in \mathbb{R}^d$ and $\sigma \geq 0$. For all $\alpha \geq 1$,*

$$D_\alpha\left(\mathcal{N}\left(\mathbb{0}_d, \sigma^2 \mathbf{I}_d\right) \| \mathcal{N}\left(\mu, \sigma^2 \mathbf{I}_d\right)\right) = \frac{\alpha \|\mu\|_2^2}{2\sigma^2}.$$

[4] In this application of DP composition, each mechanism did not need to depend on the outputs of the previous mechanisms. However, in Section 3 we give an example where the ability to handle dependences is important.

[5] This is surprising because the expected maximum of d Laplace random variables is $\approx \log d$, via a union bound.

[6] This definition extends straightforwardly to the discrete setting, where $\mathrm{d}\omega$ should be thought of as the counting measure, so integration is replaced by summation over a finite set (similarly to Definition 4, Part XI).

[7] This naming convention was introduced by [DR16] and CDP was further developed by [BS16, BDRS18]. Our definition of Rényi DP is slightly different than the definition in [Mir17], for consistency with the CDP definition.

Proof. By direct computation, for $P := \mathcal{N}(\mathbb{0}_d, \sigma^2 \mathbf{I}_d)$ and $Q := \mathcal{N}(\mu, \sigma^2 \mathbf{I}_d)$,

$$\begin{aligned}\int \left(\frac{P(x)}{Q(x)}\right)^{\alpha-1} P(x)\mathrm{d}x &= \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int \exp\left(-\frac{\alpha \|x\|_2^2}{2\sigma^2} + \frac{(\alpha-1)\|x-\mu\|_2^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int \exp\left(-\frac{\|x\|_2^2 - 2(\alpha-1)\langle x, \mu\rangle - (\alpha-1)\|\mu\|_2^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \exp\left(\frac{\alpha(\alpha-1)\|\mu\|_2^2}{2\sigma^2}\right) \cdot \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int \exp\left(-\frac{\|x-(\alpha-1)\mu\|_2^2}{2\sigma^2}\right)\mathrm{d}x}_{=1}.\end{aligned}$$

The claim follows by taking logarithms and dividing by $\alpha - 1$. □

Moreover, Rényi divergences satisfy monotonicity under postprocessing, analogously to Lemma 2.

Lemma 6. *Let P, Q be distributions on Ω, and let $F : \Omega \to \Omega'$ be an arbitrary, potentially randomized, function. Then for all $\alpha \geq 1$,*

$$D_\alpha\left(F(P)\|F(Q)\right) \leq D_\alpha\left(P\|Q\right),$$

where $F(P)$ denotes the distribution of $F(\omega)$ for $\omega \sim P$ (analogously defining $F(Q)$).

Proof. As in Lemma 2, first we consider the case where f is a deterministic map. Then, letting $f^{-1}(\omega') := \{\omega \in \Omega \mid f(\omega) = \omega'\}$ for all $\omega' \in \Omega'$,

$$\begin{aligned}(\alpha-1)\exp\left(D_\alpha\left(f(P)\|f(Q)\right)\right) &= \int \left(\frac{\int_{f^{-1}(\omega')} P(\omega)\mathrm{d}\omega}{\int_{f^{-1}(\omega')} Q(\omega)\mathrm{d}\omega}\right)^\alpha \left(\int_{f^{-1}(\omega')} Q(\omega)\mathrm{d}\omega\right)\mathrm{d}\omega' \\ &\leq \int \left(\frac{P(\omega)}{Q(\omega)}\right)^\alpha Q(\omega)\mathrm{d}\omega = (\alpha-1)\exp\left(D_\alpha\left(P\|Q\right)\right),\end{aligned}$$

from which the conclusion follows by monotonicity of log and division by $\alpha - 1$. The only inequality was Jensen's: letting $Q\,|_{\omega'}$ be the distribution which draws $\omega \in f^{-1}(\omega')$ proportionally to Q,

$$\begin{aligned}\left(\frac{\int_{f^{-1}(\omega')} P(\omega)\mathrm{d}\omega}{\int_{f^{-1}(\omega')} Q(\omega)\mathrm{d}\omega}\right)^\alpha &= \left(\mathbb{E}_{\omega\sim Q|_{\omega'}}\left[\frac{P(\omega)}{Q(\omega)}\right]\right)^\alpha \\ &\leq \mathbb{E}_{\omega\sim Q|_{\omega'}}\left[\left(\frac{P(\omega)}{Q(\omega)}\right)^\alpha\right] = \int_{f^{-1}(\omega')} \left(\frac{P(\omega)}{Q(\omega)}\right)^\alpha \frac{Q(\omega)}{\int_{f^{-1}(\omega)} Q(\omega)\mathrm{d}\omega}\mathrm{d}\omega,\end{aligned} \tag{10}$$

for all $\omega' \in \Omega'$. For randomized functions F which draw $f \sim \mathcal{F}$, the conclusion follows from *quasi-convexity* of Rényi divergences,[8] which states that if $\{P_f, Q_f\}_{f\in S}$ are densities and $\mathcal{F}$ is a distribution over S, and $P_F := \mathbb{E}_{f\sim\mathcal{F}}[P_f]$, $Q_F := \mathbb{E}_{f\sim\mathcal{F}}[Q_f]$,

$$D_\alpha\left(P_F\|Q_F\right) \leq \max_{f\in\mathcal{F}} \left(P_f\|Q_f\right).$$

This fact can be seen as follows: because $x \to \frac{1}{\alpha-1}\log x$ is monotone, it is enough to show that $(P, Q) \to \int P(\omega)^\alpha Q(\omega)^{1-\alpha}\mathrm{d}\omega$ is quasiconvex in its arguments. Indeed, the scalar function on $\mathbb{R}^2_{\geq 0}$, $(p, q) \to p^\alpha q^{1-\alpha}$, is jointly convex (and hence quasiconvex) in its arguments, so applying this pointwise to $p \leftarrow P(\omega)$, $q \leftarrow Q(\omega)$ yields that $(P, Q) \to \int P(\omega)^\alpha Q(\omega)^{1-\alpha}\mathrm{d}\omega$ is quasiconvex. □

The strategy of applying Jensen's inequality to a coarsening of a probability space in (10) is quite general, and more broadly is an example of the *data processing inequality* for f-divergences. Informally, it states that convex information measures, such as Rènyi divergences, decrease when applying a postprocessing function (i.e. "information is lost" under such a coarsening). Now, combining Lemmas 5 and 6 bounds the CDP of the Gaussian mechanism.

[8] As a word of warning, Rényi divergences do not in general satisfy the stronger notion of joint convexity in its arguments, except in the special case of $\alpha = 1$ (i.e. the KL divergence is jointly convex).

Proposition 2 (Gaussian mechanism). *Let $f : \mathcal{S}^n \to \mathbb{R}^d$ be a d-dimensional statistic of a dataset with ℓ_2-sensitivity bounded by Δ, i.e. for all neighboring datasets $D, D' \in \mathcal{S}^n$, $\|f(D) - f(D')\|_2 \leq \Delta$. Then the mechanism which outputs a sample from $\mathcal{N}(f(D), \sigma^2 \mathbf{I}_d)$ satisfies $\frac{\Delta^2}{2\sigma^2}$-CDP.*

Proof. It suffices to show that for all neighboring datasets $D, D' \in \mathcal{S}^n$, and all $\alpha \geq 1$,

$$D_\alpha \left(\mathcal{N}\left(f(D), \sigma^2 \mathbf{I}_d\right) \| \mathcal{N}\left(f(D'), \sigma^2 \mathbf{I}_d\right) \right) \leq \frac{\alpha \Delta^2}{2\sigma^2}.$$

Indeed, for $\mu := f(D') - f(D)$ and $v := f(D)$, this follows from $\|\mu\|_2 \leq \Delta$ and Lemmas 5 and 6 (with $F(x) \leftarrow x + v$), which imply

$$D_\alpha \left(\mathcal{N}(v, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(\mu + v, \sigma^2 \mathbf{I}_d) \right) \leq D_\alpha \left(\mathcal{N}(\mathbb{0}_d, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(\mu, \sigma^2 \mathbf{I}_d) \right) \leq \frac{\alpha \Delta^2}{2\sigma^2}.$$

□

Additionally, as mentioned in Remark 1, RDP naturally offers privacy tradeoff curves; for every failure probability δ in approximate DP, RDP yields an $(\epsilon(\delta), \delta)$-DP guarantee for some value $\epsilon(\delta)$. This can be viewed as giving a DP guarantee at every scale of the probability of "catastrophic failure" (i.e. with probability δ, we may violate the DP guarantee by a potentially arbitrary amount).

Lemma 7. *If $\mathcal{M} : \mathcal{S}^n \to \Omega$ is (α, ρ)-RDP, it is also $(\epsilon(\delta), \delta)$-DP for any $\delta \in (0, 1)$, where*

$$\epsilon(\delta) := \alpha\rho + \frac{\log \frac{1}{\delta}}{\alpha - 1}.$$

Proof. Throughout, for brevity let $P := \mathcal{M}(D)$ and $Q := \mathcal{M}(D')$ for neighboring databases $D, D' \in \mathcal{S}^n$. By Hölder's inequality with dual exponents $(\alpha, \frac{\alpha}{\alpha-1})$, we have for any event $A \subseteq \Omega$ that

$$\begin{aligned}
\Pr_{\omega \sim P}[\omega \in A] &= \int_A P(\omega) \mathrm{d}\omega \leq \left(\int_A P(\omega)^\alpha Q(\omega)^{1-\alpha} \mathrm{d}\omega \right)^{\frac{1}{\alpha}} \left(\int_A Q(\omega) \mathrm{d}\omega \right)^{\frac{\alpha-1}{\alpha}} \\
&= \exp\left(\frac{\alpha - 1}{\alpha} D_\alpha \left(P \| Q \right) \right) \left(\int_A Q(\omega) \mathrm{d}\omega \right)^{\frac{\alpha-1}{\alpha}} = \left(\exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \right)^{\frac{\alpha-1}{\alpha}}.
\end{aligned}$$

Now if $\exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \leq \delta^{\frac{\alpha}{\alpha-1}}$, it follows that $\Pr_{\omega \sim P}[\omega \in A] \leq \exp(c) \Pr_{\omega \sim Q}[\omega \in A] + \delta$ for any $c \in \mathbb{R}$. Otherwise, we have the conclusion from

$$\begin{aligned}
\left(\exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \right)^{\frac{\alpha-1}{\alpha}} &= \exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \cdot \left(\exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \right)^{-\frac{1}{\alpha}} \\
&\leq \exp(\alpha\rho) \Pr_{\omega \sim Q}[\omega \in A] \cdot \delta^{-\frac{1}{\alpha-1}} = \exp\left(\alpha\rho + \frac{\log \frac{1}{\delta}}{\alpha - 1} \right) \Pr_{\omega \sim Q}[\omega \in A].
\end{aligned}$$

□

We now provide some intuition for Lemma 7. First, observe that ϵ-DP implies that the likelihood ratio of neighboring mechanisms is pointwise bounded by $\exp(\epsilon)$, so plugging this into the second formula in (8), we have that ϵ-DP implies $(\alpha, \frac{\epsilon}{\alpha})$-RDP. It would be ideal if this implication went both ways, but pure DP (i.e. with $\delta = 0$) is stronger than RDP; the latter requires bounds on $\alpha - 1$ moments of likelihood ratios, whereas the former requires bounds on all moments. To convert from (α, ρ)-RDP to DP, we hence need to ask; what is the probability that the likelihood ratio is larger than ϵ, if the expected $(\alpha - 1)^{\text{th}}$ moment is $\approx \alpha\rho$? If we set this failure probability to δ, Lemma 7 says we can obtain the desired approximate DP by increasing $\alpha\rho$ to $\alpha\rho + \frac{\log \frac{1}{\delta}}{\alpha - 1}$. Unsurprisingly, the larger α is (the more moments we have control over), the smaller this increase is.

To get a feel for how to use Lemma 7 in applications, let us revisit the one-way marginals problem in Proposition 1. For fixed $\epsilon, \delta \in (0, 1)$, Lemma 7 says that we can obtain an (ϵ, δ)-DP mechanism by instead designing a (α, ρ)-RDP mechanism, for

$$\alpha = \frac{3 \log \frac{1}{\delta}}{\epsilon}, \ \rho = \frac{\epsilon}{2\alpha} = \frac{\epsilon^2}{6 \log \frac{1}{\delta}}. \tag{11}$$

Moreover, it is straightforward to see that $\mu(D) \in \mathbb{R}^d$ is a statistic of the dataset D with ℓ_2-sensitivity at most $\Delta = \frac{2}{n} \cdot \sqrt{d}$ (following notation in Proposition 2), since each of the d one-way marginals can only change by $\frac{2}{n}$ in magnitude when moving to a neighboring dataset. Hence, the Gaussian mechanism (Proposition 2) yields ρ-CDP (and hence (α, ρ)-RDP) by taking

$$\sigma = \frac{2\Delta\sqrt{\log\frac{1}{\delta}}}{\epsilon} = \frac{4\sqrt{d\log\frac{1}{\delta}}}{n\epsilon} \implies \frac{\Delta^2}{2\sigma^2} \leq \frac{\epsilon^2}{6\log\frac{1}{\delta}}. \tag{12}$$

Now because the Gaussian mechanism simply adds $\mathcal{N}(0, \sigma^2)$ noise to each coordinate of the output, the utility guarantee (7) holds with $\alpha \approx \sigma$. Hence, for a target accuracy level α,

$$\frac{2\Delta\sqrt{\log\frac{1}{\delta}}}{\epsilon} = \frac{4\sqrt{d\log\frac{1}{\delta}}}{n\epsilon} \leq \alpha \iff n \geq \frac{4\sqrt{d\log\frac{1}{\delta}}}{\alpha\epsilon}.$$

We have thus recovered Proposition 2 (with a slightly better constant) without appealing to advanced composition from Theorem 1. Indeed, this recovery of advanced composition via RDP is generic, as derived in Corollary 1, [Mir17]. We conclude the section with a simple composition theorem for RDP mechanisms, analogously to Lemma 4 and Corollary 1.

Lemma 8. *Let $\alpha \geq 1$, and let $\mathcal{M}_i : \mathcal{S}^n \times \prod_{j\in[i-1]} \Omega_j \to \Omega_i$ be an (α, ρ_i)-RDP mechanism, for all $i \in [k]$. Then the composition $\mathcal{M} := \mathcal{M}_k \circ \ldots \circ \mathcal{M}_1$ is $(\alpha, \sum_{i\in[k]} \rho_i)$-RDP.*

Proof. We prove the claim when $k = 2$, and then the conclusion follows by performing the same induction as in Corollary 1. For brevity, let $P_1 := \mathcal{M}_1(D)$ and $Q_1 := \mathcal{M}_1(D')$ for neighboring datasets $D, D' \in \mathcal{S}^n$, and for each $\omega \in \Omega$, let $P_2^\omega := \mathcal{M}_2(D, \omega)$ and $Q_2^\omega := \mathcal{M}_2(D', \omega)$. Then,

$$\begin{aligned}
\exp\left((\alpha-1)D_\alpha\left(\mathcal{M}(D)\|\mathcal{M}(D')\right)\right) &\leq \int\int \left(P_1(\omega)P_2^\omega(\omega')\right)^\alpha \left(Q_1(\omega)Q_2^\omega(\omega')\right)^{1-\alpha} \mathrm{d}\omega'\mathrm{d}\omega \\
&= \int \left(P_1(\omega)^\alpha Q_1(\omega)^{1-\alpha}\right) \left(\int (P_2^\omega(\omega')^\alpha Q_2^\omega(\omega')^{1-\alpha}\mathrm{d}\omega'\right) \mathrm{d}\omega \\
&\leq \exp\left((\alpha-1)\alpha\rho_2\right) \int P_1(\omega)^\alpha Q_1(\omega)^{1-\alpha}\mathrm{d}\omega \\
&\leq \exp\left((\alpha-1)\alpha(\rho_1+\rho_2)\right).
\end{aligned}$$

The first inequality used that the Rényi divergence between the distributions of (ω, ω') is larger than that between the distributions of only ω' by Lemma 6, as marginalization is a postprocessing. The other two inequalities used the RDP assumptions on $\mathcal{M}_1$ and $\mathcal{M}_2(\cdot, \omega)$ for any $\omega \in \Omega$. □

As we saw in (11), the parameter ρ in RDP can be thought of as having "units" ϵ^2, where ϵ is the privacy loss in (1).[9] Because ρ grows linearly upon composition by Lemma 8, it is now unsurprising that the privacy loss grows as $\approx \sqrt{k}$ (with some failure probability), yielding advanced composition.

3 Differentially private optimization

We now show how to use our developments thus far to give an algorithm for a differentially private variant of the empirical risk minimization problem introduced in Section 5, Part III, for convex sample functions. We first formally define the optimization problem we study.

Definition 4 (DP-ERM and DP-SCO)**.** *Let $\epsilon, \delta \in (0, 1)$, let $\mathcal{P}$ be a distribution over $\mathcal{S}$, and suppose that there is a family of L-Lipschitz, differentiable[10] convex functions $\{f(\cdot; s)\}_{s\in\mathcal{S}}$ so each $f(\cdot; s) : \mathcal{X} \to \mathbb{R}$ is indexed by an element $s \in \mathcal{S}$, for $\mathcal{X} \subseteq \mathbb{R}^d$, with $\mathrm{diam}(\mathcal{X}) \leq R$. For $n \in \mathbb{N}$, we receive a dataset $D := \{s_i\}_{i\in[n]} \subseteq \mathcal{S}$ sampled i.i.d. $\sim \mathcal{P}$, and denote $f_i(\cdot) := f(\cdot; s_i)$ for all $i \in [n]$.*

1. *In the* differentially private empirical risk minimization *(DP-ERM) problem, our goal is to design an (ϵ, δ)-DP mechanism computing an approximate minimizer of the empirical risk*

$$F_{\mathrm{erm}}(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x). \tag{13}$$

[9] This is also suggested by Proposition 2, since ρ scales as the square of the sensitivity Δ.

[10] Differentiability is assumed only for simplicity; using subgradients appropriately removes this requirement.

2. *In the* differentially private stochastic convex optimization *(DP-SCO) problem, our goal is to design an* (ϵ, δ)*-DP mechanism computing an approximate minimizer of the population risk*

$$F_{\text{pop}}(x) := \mathbb{E}_{s\sim\mathcal{P}}\left[f(x;s)\right]. \tag{14}$$

The problem formulation in Definition 4 is quite general, and is a common model for statistical learning induced by a convex loss function. Indeed, as discussed in Section 5, Part III, it encompasses private variants of learning many popular statistical models such as linear and logistic regression, support vector machines, as well as mean and median estimation, etc.

The optimal attainable error for DP-ERM and DP-SCO (in expectation over the randomness of the algorithm and sampled dataset) respectively scale as

$$\Theta\left(LR\cdot\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right),\ \Theta\left(LR\cdot\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)\right). \tag{15}$$

Indeed, observe that this rate for DP-ERM closely resembles the optimal error attainable for the one-way marginals problem, as discussed after Proposition 1.[11] This is no coincidence; [BST14] attained their DP-ERM error lower bound by designing a simple construction of a hard linear function, based on one-way marginals. The second error rate in (15) is simply the sum of the DP-ERM lower bound and the optimal non-private SCO loss, which scales as $\approx \frac{1}{\sqrt{n}}$ (witnessed by classical hard instances such as learning the bias of a coin, i.e. 1-d mean estimation). One interpretation of the bound (15) is that for DP-SCO, there is no asymptotic "cost of (ϵ, δ)-privacy" once we have enough samples $n = \Omega(\frac{d\log\frac{1}{\delta}}{\epsilon^2})$, as the non-private SCO loss term $\frac{1}{\sqrt{n}}$ then dominates.

We focus on the DP-ERM problem in this lecture, but mention that there are generic techniques for extending DP-ERM algorithms with some error rate to DP-SCO algorithms achieving the same error rate, up to a $LR\cdot\frac{1}{\sqrt{n}}$ additive overhead (see e.g. Theorem 5.1, [KLL21]). These techniques are based on a reduction called *iterative localization* building upon [SSSS09, FKT20]; the former work noted that *strongly convex* variants of ERM automatically yield generalization guarantees, and the latter gave a reduction from strongly convex ERM to its weakly convex counterpart.

Remark 2 (One-pass algorithms)**.** *In light of the discussion in Section 5, Part III, the reader may be surprised to find out that the reduction from SCO to ERM discussed above is not immediate. Indeed, if there was a* one-pass *DP-ERM algorithm (i.e. one that only queries each sample function* $f(\cdot; s_i)$ *through a first-order oracle once), then generalization guarantees to* F_{pop} *are typically straightforward. This is because we can treat sample gradients as unbiased for the population gradient* ∇F_{pop}*, since there are no dependencies between iterations (as each sample is a fresh draw from the population). Indeed, standard one-pass stochastic gradient descent (as developed in Section 5, Part III in the non-private setting) satisfies this property. Unfortunately, there currently does not exist such an algorithm; existing DP-ERM algorithms which attain the optimal error in* (15) *take multiple passes over the dataset, except in certain situations discussed at the end of the section. This has been quoted as a major open problem in the area of private optimization [Tal22].*

We now give a simple analysis of an optimal (multi-pass) DP-ERM algorithm, with quadratic sample gradient complexity. Our DP-ERM algorithm is simply an instantiation of the Gaussian mechanism in Proposition 2, applied to privatize our computation of gradients of F_{erm}.

Theorem 2 (DP-ERM with optimal error rates)**.** *In the setting of Definition 4, let* $x_0 \in \mathcal{X}$ *be arbitrary, and consider iterating*

$$\begin{aligned} &x_{t+1} \leftarrow \text{argmin}_{x\in\mathcal{X}}\left\{\left\langle\eta\left(\nabla F_{\text{erm}}(x_t)+\xi_t\right), x\right\rangle + \frac{1}{2}\left\|x-x_t\right\|_2^2\right\},\ \textit{for } 0\le t<T,\\ &\textit{where } \xi_t\sim\mathcal{N}(0,\sigma^2\mathbf{I}_d),\ \eta\leftarrow\frac{R}{L}\left(2T+\frac{32dT^2\log\frac{1}{\delta}}{\epsilon^2n^2}\right)^{-\frac{1}{2}},\ T\ge\frac{n^2\epsilon^2}{32d\log\frac{1}{\delta}},\ \sigma:=\frac{4L\sqrt{T\log\frac{1}{\delta}}}{\epsilon n}. \end{aligned} \tag{16}$$

[11] The multiplicative overhead of LR is to preserve scale invariance of the optimal error.

The mechanism which returns the average iterate $\bar{x} := \frac{1}{T}\sum_{0\le t<T} x_t$ *satisfies* (ϵ, δ)*-differential privacy, and letting* $x^\star \in \operatorname{argmin}_{x\in\mathcal{X}} F_{\mathrm{erm}}(x)$,

$$\mathbb{E}\left[F_{\mathrm{erm}}(\bar{x}) - F_{\mathrm{erm}}(x^\star)\right] \le 8LR \cdot \frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}.$$

Proof. We begin with the privacy analysis. Note that the gradient of the empirical risk at any point $x \in \mathcal{X}$, $\nabla F_{\mathrm{erm}}(x)$, is a statistic of the dataset D with ℓ_2-sensitivity $\Delta := \frac{2L}{n}$. Indeed, for neighboring datasets $D = \{s_i\}_{i\in[n]}, D' = \{s_i'\}_{i\in[n]}$ where $s_i = s_i'$ except when $i = j$,

$$\left\|\frac{1}{n}\sum_{i\in[n]}\nabla f(x; s_i) - \frac{1}{n}\sum_{i\in[n]}\nabla f(x; s_i')\right\|_2 = \frac{1}{n}\left\|\nabla f(x; s_j) - \nabla f(x; s_j')\right\| \le \frac{2L}{n}, \tag{17}$$

where we used the triangle inequality and L-Lipschitzness of each sample function $f(\cdot; s)$. Therefore, by Proposition 2, each iteration (16) which takes as input all iterates produced by the algorithm thus far, and outputs the next iterate, is a $\frac{\Delta^2}{2\sigma^2}$-CDP mechanism. By Lemma 8, the overall mechanism which iterates (16) T times thus satisfies (α, ρ)-RDP with α set in (11), and

$$\rho := \frac{T\Delta^2}{2\sigma^2} = \frac{2TL^2}{\sigma^2 n^2} \le \frac{\epsilon^2}{6\log\frac{1}{\delta}},$$

by using our choice of σ. Since outputting the average iterate is a postprocessing of the outputs of each iteration (16) (i.e. no further accesses to the dataset are necessary), Lemma 6 and the calculation in (11) guarantee (ϵ, δ)-differential privacy of the overall mechanism.

We now prove the utility claim. Note that each stochastic gradient $\nabla F_{\mathrm{erm}}(x_t) + \xi_t$ is unbiased for $\nabla F_{\mathrm{erm}}(x_t)$, and has a second moment bound (using Lipschitzness of F_{erm})

$$\mathbb{E}\left\|\nabla F_{\mathrm{erm}}(x_t) + \xi_t\right\|_2^2 \le 2\left\|\nabla F_{\mathrm{erm}}(x_t)\right\|_2^2 + 2\mathbb{E}\left\|\xi_t\right\|_2^2 \le 2L^2 + 2\sigma^2 d = 2L^2\left(1 + \frac{16dT\log\frac{1}{\delta}}{\epsilon^2 n^2}\right). \tag{18}$$

Applying Corollary 4, Part III with $\varphi(x) := \frac{1}{2}\|x\|_2^2$ then gives

$$\begin{aligned}\mathbb{E}\left[F_{\mathrm{erm}}(\bar{x}) - F_{\mathrm{erm}}(x^\star)\right] &\le \frac{R^2}{2\eta T} + \eta L^2\left(1 + \frac{16dT\log\frac{1}{\delta}}{\epsilon^2 n^2}\right) \\ &\le LR\cdot\sqrt{\frac{2}{T}\left(1 + \frac{16dT\log\frac{1}{\delta}}{\epsilon^2 n^2}\right)} \le 8LR\cdot\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n},\end{aligned} \tag{19}$$

where we plugged in our choices of η and T respectively in the last two inequalities. □

As promised, Theorem 2 yields the optimal rate in (15). Moreover, following the discussion after (15), if we take just enough samples $n \approx \frac{d\log\frac{1}{\delta}}{\epsilon^2}$ so there is no asymptotic cost of privacy, the iteration count T in (16) scales linearly in n. Finally, note that computing $\nabla F_{\mathrm{erm}}(x_t)$ in each iteration requires querying each of the n sample gradients $\nabla f(x_t; s_i)$, so the total number of sample gradients used is $\approx n^2$. Hence, Theorem 2 falls short of being a one-pass algorithm (Remark 2).

One natural direction towards improvement is to subsample each gradient $\nabla F_{\mathrm{erm}}(x_t)$ by randomly outputting $\nabla f_i(x_t)$, for $i \sim_{\mathrm{unif.}} [n]$. In the case of non-private ERM, this immediately improves standard stochastic gradient descent to be a one-pass algorithm, since the second moment bound on subsampled gradients has the same scaling of $O(L^2)$ as the average dataset gradient (see (18)). Unfortunately, this strategy increases the sensitivity of each statistic by a factor of n (corresponding to (17)), so naïvely we must take a larger noise level σ which cancels out any gains from subsampling. On the other hand, there is hope: if a mechanism first uses a subsampled dataset, then the probability we need to pay a privacy loss should also decrease. This idea can be formalized through a framework called *privacy amplification by subsampling*. One such amplification result which applies to the Gaussian mechanism can be found in Theorem 14, [BDRS18].

Unfortunately, this amplification alone does not improve upon the query complexity in Theorem 2. To see why, suppose that a mechanism $\mathcal{M}$ is ϵ-DP on datasets of size $m = pn$, but acts on an input

dataset $D \in \mathcal{S}^n$ by first subsampling $E \subseteq D$ by randomly including m elements of D (say, without replacement). Then if $D, D' \in \mathcal{S}^n$ are neighboring, the subsampled datasets E and E' are exactly the same except with probability $p = \frac{m}{n}$ (the probability that the differing sample is included). We should thus expect the overall privacy loss ϵ' in (1) to behave like

$$\exp(\epsilon') \approx p \exp(\epsilon) + (1-p)\exp(0) = 1 + p(\exp(\epsilon) - 1) \approx \exp(p\epsilon).$$

If all of the above calculations held, then the new privacy parameter $\epsilon' = p\epsilon$ would be a factor of p smaller, as desired. However, note that the last approximation above only holds if ϵ is already sufficiently small, as otherwise $\exp(\epsilon) \gg 1 + \epsilon$. The takeaway from this discussion is that privacy amplification by subsampling (with probability p) only boosts the privacy loss by a factor $\approx p$ if the mechanism is already $O(1)$-private, *before* applying privacy amplification. This is problematic in the context of Theorem 1's proof, where the amplification factor of $p = \frac{1}{n}$ turns out to exactly cancel out the sensitivity loss factor of n in (17), leading to no overall improvement.[12]

Recently, the community has been able to improve on the basic analysis in Theorem 2, giving DP-ERM and DP-SCO algorithms with substantially improved sample gradient query complexities. For instance, it was observed in [FKT20] that under a smoothness assumption on the sample functions $f(\cdot; s)$ (in addition to Lipschitzness), a one-pass algorithm in the sense of Remark 2 actually is attainable. This result was proven in two different ways (one of which uses an alternative privacy amplification framework called *privacy amplification by iteration* [FMTT18]). The core idea of both strategies is that smooth gradient steps are contractive (see Proposition 2.10, [FKT20]), so that we can benefit from a single noise addition for many consecutive iterations, where iterates taken with respect to neighboring datasets' sample gradients do not drift further apart.

Using a careful combination of privacy amplification strategies and smoothening techniques, a line of work [AFKT21, KLL21, CJJ+23] has yielded a DP-SCO algorithm which queries at most

$$\approx n + \frac{(nd)^{\frac{2}{3}}}{\epsilon}$$

sample gradients. Notably, if $n \gg \frac{d^2}{\epsilon^3}$ is sufficiently large, the above query complexity of [CJJ+23] is linear in n. Hence, this result shows that in addition to there being no asymptotic cost of privacy in terms of utility (15), there is also no asymptotic cost in terms of computational complexity. Nevertheless, the more general problem of designing a one-pass DP-SCO algorithm without further assumptions (or showing that such an algorithm cannot exist) remains an exciting open direction.[13]

[12] More concretely, one can check that to apply amplification by subsampling with $p \approx \frac{1}{n}$, i.e. a single sample gradient per iteration, we must take the noise level $\sigma \approx L\epsilon^{-\frac{1}{2}}$. Then to balance the two terms in (19) to obtain the optimal error, we require $T \gtrsim \frac{n^2}{\epsilon}$, so we use quadratically many queries. By using a larger batch size of $\approx \sqrt{n}$, one can improve upon this bound slightly [AFKT21], but this still does not result in a one-pass algorithm.

[13] We also remark that there are several extraneous logarithmic factors lost by the [CJJ+23] algorithm, which are also interesting to improve (potentially via a simpler framework).

Source material

Portions of this lecture are based on reference material in [DR14, Kam20], as well as the author's own experience working in the field.

References

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in L1 geometry. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 393–403. PMLR, 2021.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 74–86. ACM, 2018.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Proceedings, Part I*, volume 9985 of *Lecture Notes in Computer Science*, pages 635–658, 2016.

[BST14] Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014*, pages 464–473. IEEE Computer Society, 2014.

[BUV18] Mark Bun, Jonathan R. Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM J. Comput.*, 47(5):1888–1938, 2018.

[CJJ+23] Yair Carmon, Arun Jambulapati, Yujia Jin, Yin Tat Lee, Daogao Liu, Aaron Sidford, and Kevin Tian. Resqueing parallel and private stochastic convex optimization. In *64th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2023*, pages 2031–2058. IEEE, 2023.

[CLN+23] Edith Cohen, Xin Lyu, Jelani Nelson, Tamás Sarlós, and Uri Stemmer. Generalized private selection and testing with high confidence. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, volume 251 of *LIPIcs*, pages 39:1–39:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[DK22] Yuval Dagan and Gil Kur. A bounded-noise mechanism for differential privacy. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 625–661. PMLR, 2022.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[DR16] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[FKT20] 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, STOC 2020*, pages 439–449. ACM, 2020.

[FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018*, pages 521–532. IEEE Computer Society, 2018.

[GKM21] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. On avoiding the union bound when answering multiple differentially private queries. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2133–2146. PMLR, 2021.

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory, 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 1948–1989. PMLR, 2022.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010*, pages 705–714. ACM, 2010.

[Kam20] Gautam Kamath. Algorithms for private data analysis. http://www.gautamkamath.com/CS860-fa2020.html, 2020. Accessed: 2024-04-16.

[KLL21] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth ERM and SCO in subquadratic steps. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 4053–4064, 2021.

[Mir17] Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017*, pages 263–275. IEEE Computer Society, 2017.

[SSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT 2009 - The 22nd Conference on Learning Theory, 2009*, 2009.

[SU16] Thomas Steinke and Jonathan R. Ullman. Between pure and approximate differential privacy. *J. Priv. Confidentiality*, 7(2), 2016.

[Tal22] Kunal Talwar. Ppml workshop talk: Open questions in differentially private machine learning. https://machinelearning.apple.com/video/open-questions, 2022. Accessed: 2022-11-06.