Stability of learning algorithms

Maxim Raginsky

November 5, 2015

Recall our abstract formulation of the learning problem: we have a collection $Z_1, \ldots, Z_n$ of i.i.d. samples from some unknown distribution P on a set Z and a class $\mathscr{F}$ of functions $f : \mathsf{Z} \to [0,1]$. A learning algorithm is a sequence $A = \{A_n\}_{n=1}^{\infty}$ of mappings $A_n : \mathsf{Z}^n \to \mathscr{F}$ that take training data as input and generate functions in $\mathscr{F}$ as output. We say that A is consistent if

$$L(\widehat{f}_n) = \int_{\mathsf{Z}} \widehat{f}_n(z) P(\mathrm{d}z), \qquad \widehat{f}_n = A_n(Z^n)$$

converges in some sense to $L^* = \inf_{f \in \mathscr{F}} L(f)$, for any P. If a consistent algorithm exists, we say that the problem is learnable. Early on, we have identified one sufficient condition for the existence of a consistent learning algorithm: uniform convergence of empirical means (UCEM). One way of stating the UCEM property is to require that

$$\sup_P \mathbb{E}_P \| P_n - P \|_{\mathscr{F}} \xrightarrow{n \to \infty} 0, \tag{1}$$

where the expectation is with respect to an i.i.d. process $Z_1, Z_2, \ldots$ with common marginal distribution P, and P_n is the empirical distribution based on the first n samples of the process:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}.$$

We have proved that, if $\mathscr{F}$ satisfies (1), then the ERM algorithm

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i)$$

is consistent. In some cases, the UCEM property is both necessary and sufficient for learnability — for example, in the binary classification setting, where $Z = (X, Y)$ with arbitrary X, $Y \in \{0,1\}$, and $f(Z) = f(X,Y)$ taking values in $\{0,1\}$.

However, it is easy to see that, in general, one can have learnability without the UCEM property. For example, suppose that the function class $\mathscr{F}$ is such that one can find a function $\tilde{f} \notin \mathscr{F}$ with the property that $\tilde{f}(z) < \inf_{f \in \mathscr{F}} f(z)$ for every $z \in \mathsf{Z}$. Consider now a modified class $\tilde{\mathscr{F}} = \mathscr{F} \cup \{\tilde{f}\}$ obtained by adding $\tilde{f}$ to $\mathscr{F}$. Then the ERM algorithm over $\tilde{\mathscr{F}}$ will always return $\tilde{f}$, and moreover $L(\tilde{f}) \equiv L^*(\tilde{\mathscr{F}})$. Thus, not only do we have consistency, but we also have perfect generalization, and the only condition the original class $\mathscr{F}$ has to satisfy is that we can find at least one $\tilde{f}$ with the desired property. Of course, this imposes some minimal richness requirements on the ranges of all functions in $\mathscr{F}$ — for example, we could not pull this off when the functions in $\mathscr{F}$ are binary valued. And yet, the UCEM property is not required for perfect learnability!

So, what's going on here? It turns out that the main attraction of the UCEM property – namely, its algorithm-independence – is also its main disadvantage. Learnability is closely tied up with properties of learning algorithms: how well can they generalize? how good are they at rejecting obviously bad hypotheses and focusing on good ones? Thus, our goal is to connect learnability to certain properties of learning algorithms. This lecture is based primarily on a paper by Shalev-Shwartz et al. [SSSS10].

1 An in-depth view of learning algorithms

For future convenience, let us slightly modify our notation pertaining to the learning problem. As before, we will have the data space Z and a function class $\mathscr{F}$. However, now we do not require the functions in $\mathscr{F}$ to be real-valued — they will be elements of some Hilbert space. Instead, we introduce a loss function $\ell : \mathscr{F} \times \mathsf{Z} \to [0,1]$. We will still use the notation

$$L_P(f) = \mathbb{E}_P[\ell(f,Z)] \equiv \int_{\mathsf{Z}} \ell(f,z) P(\mathrm{d}z)$$

for the expected loss of f with respect to P, and will often omit the subscript P when it's clear from context. Also, given an n-tuple $Z^n = (Z_1, \ldots, Z_n)$ of i.i.d. samples from P, we have the empirical loss

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i).$$

Finally, we define

$$L^*(\mathscr{F}) \triangleq \inf_{f \in \mathscr{F}} L(f) \qquad \text{and} \qquad L_n^*(\mathscr{F}) \triangleq \inf_{f \in \mathscr{F}} L_n(f).$$

Here, $L^*(\mathscr{F})$ is a deterministic quantity that depends on ℓ, $\mathscr{F}$, and the underlying distribution P, whereas $L_n^*(\mathscr{F})$ is a random variable.

Also, let us define $\mathsf{Z}^* \triangleq \bigcup_{n=1}^{\infty} \mathsf{Z}^n$, i.e., Z^* is the collection of all tuples over Z. This definition allows us to treat a learning algorithm as a *single mapping* $A : \mathsf{Z}^* \to \mathscr{F}$ — the size of the training set is now clear from context. For example, $A(Z^n)$ is the output of A fed with an n-tuple $Z^n = (Z_1, \ldots, Z_n)$, and so, in particular,

$$L(A(Z^n)) = \int_{\mathsf{Z}} \ell(A(Z^n), z) P(\mathrm{d}z)$$

is the expected loss of the function $A(Z^n) \in \mathscr{F}$ on a fresh sample $Z \sim P$, independent of Z^n. This new notation is rather flexible: for example,

$$L_n(A(Z^n)) = \frac{1}{n} \sum_{i=1}^{n} \ell(A(Z^n), Z_i)$$

is the empirical loss of the algorithm output $A(Z^n)$ on the same sample Z^n that was supplied to A.

Our goal is to understand what makes a good learning algorithm. To keep things simple, we will focus on expected-value guarantees. We say that a learning algorithm A is *consistent* if

$$c_n(A) \triangleq \sup_P \mathbb{E}_P \left[L(A(Z^n)) - L^* \right] \xrightarrow{n \to \infty} 0. \tag{2}$$

We say that the learning problem specified by ℓ and $\mathscr{F}$ is learnable if there exists at least one consistent learning algorithm A. As we have already seen on multiple occasions, under certain conditions the ERM algorithm is consistent. A learning algorithm A is an *Asymptotic Empirical Risk Minimizer* (AERM) if

$$e_n(A) \triangleq \sup_P \mathbb{E}_P \left[L_n(A(Z^n)) - L_n^* \right] \xrightarrow{n \to \infty} 0. \tag{3}$$

Of course, is A is the exact ERM algorithm, then $e_n(A) = 0$ for all n, but there are many situations in which it is preferable to use AERM algorithms. Next, we say that A *generalizes* if

$$g_n(A) \triangleq \sup_P \mathbb{E}_P \left| L(A(Z^n)) - L_n(A(Z^n)) \right| \xrightarrow{n\to\infty} 0. \tag{4}$$

A weaker notion of generalization is as follows: A generalizes *on average* if

$$\bar{g}_n(A) \triangleq \sup_P \left| \mathbb{E}_P [L(A(Z^n)) - L_n(A(Z^n))] \right| \xrightarrow{n\to\infty} 0. \tag{5}$$

Our goal is to show that learnability is possible without requiring the UCEM property; instead, we will investigate the relationship between the above properties of learning algorithms to *stability*, i.e., weak dependence of the algorithm output on any individual training sample.

2 A primer on convex functions

In order to proceed, we first need to introduce some ideas from convex analysis. Let $\mathscr{H}$ be a Hilbert space. A subset $\mathscr{F} \subseteq \mathscr{H}$ is *convex* if

$$f_1, f_2 \in \mathscr{F} \quad \Longrightarrow \quad \lambda f_1 + (1-\lambda) f_2 \in \mathscr{F}, \; \forall \lambda \in [0,1].$$

A function $\varphi : \mathscr{F} \to \mathbb{R}$ is convex if

$$\varphi(\lambda f_1 + (1-\lambda) f_2) \le \lambda \varphi(f_1) + (1-\lambda)\varphi(f_2), \qquad \forall f_1, f_2 \in \mathscr{F}, \lambda \in [0,1].$$

A vector $g \in \mathscr{H}$ is a *subgradient* of φ at $f \in \mathscr{F}$ if

$$\varphi(f') \ge \varphi(f) + \langle g, f' - f \rangle, \qquad \forall f' \in \mathscr{F}.$$

The set of all subgradients of φ at f is denoted by $\partial\varphi(f)$ and is referred to as the *subdifferential* of φ at f. It can be shown that $\partial\varphi(f) \neq \varnothing$ for every $f \in \mathscr{F}$. We say that φ is differentiable at f if $\partial\varphi(f)$ has only one element, in which case we refer to this element as the *gradient* of φ at f and denote it by $\nabla\varphi(f)$.

Given a convex function φ on $\mathscr{F}$, it is often of interest to minimize it, i.e., to find some $f^* \in \mathscr{F}$, such that $\varphi(f^*) \le \varphi(f)$ for all other $f \in \mathscr{F}$, in which case we say that f^* is a minimizer of φ on $\mathscr{F}$. We have the following basic result:

Lemma 1 (First-order optimality condition)**.** *Let $\varphi : \mathscr{F} \to \mathbb{R}$ be a differentiable convex function. The point $f^* \in \mathscr{F}$ is a minimizer of φ on $\mathscr{F}$ if and only if*

$$\langle \nabla\varphi(f^*), f - f^* \rangle \ge 0, \qquad \forall f \in \mathscr{F}. \tag{6}$$

Proof. To prove sufficiency, note that, by definition of the subgradient,

$$\varphi(f) \ge \varphi(f^*) + \langle g^*, f - f^* \rangle$$

for any $g^* \in \partial\varphi(f^*)$. If (6) holds, then $\varphi(f) \ge \varphi(f^*)$ for all $f \in \mathscr{F}$. (Note that here we do not require differentiability of φ.)

To prove necessity, let f^* be a minimizer of φ on $\mathscr{F}$, and suppose that (6) does not hold. That is, there exists some $f \in \mathscr{F}$, such that $\langle \nabla\varphi(f^*), f - f^* \rangle < 0$. By convexity of $\mathscr{F}$, $f^* + t(f - f^*) \in \mathscr{F}$ for all sufficiently

small $t > 0$. Consider the function $F(t) \triangleq \varphi(f^* + t(f - f^*))$. Since φ is differentiable, so is F. By the chain rule, which holds in a Hilbert space, we have

$$F'(0) = \langle \nabla\varphi(f^* + t(f - f^*)), f - f^* \rangle \Big|_{t=0} = \langle \nabla\varphi(f^*), f - f^* \rangle < 0.$$

But this means that $F(t) < F(0)$ for all small $t > 0$, which contradicts the optimality of f^*. □

Next, we say that a function $\varphi : \mathscr{F} \to \mathbb{R}$ is σ-*strongly convex*, for some $\sigma \geq 0$, if

$$\varphi(f') \geq \varphi(f) + \langle g, f' - f \rangle + \frac{\sigma}{2}\|f - f'\|^2, \tag{7}$$

for all $f, f' \in \mathscr{F}$ and all $g \in \partial\varphi(f)$. In the case $\sigma = 0$, we recover the usual definition of convexity; the real power of this condition is when $\sigma > 0$, so from now on we will only use this term when $\sigma > 0$. Fix any two $f, f' \in \mathscr{F}$ and any $\lambda \in [0, 1]$. Then

$$\varphi(f) \geq \varphi(\lambda f + (1-\lambda) f') + (1-\lambda)\langle g, f - f' \rangle + \frac{(1-\lambda)^2\sigma}{2}\|f - f'\|^2$$

and

$$\varphi(f') \geq \varphi(\lambda f + (1-\lambda) f') - \lambda\langle g, f - f' \rangle + \frac{\lambda^2\sigma}{2}\|f - f'\|^2,$$

for any $g \in \partial\varphi(\lambda f + (1-\lambda) f')$. Multiplying the first inequality by λ, the second one by $1 - \lambda$, and adding them, we get

$$\lambda\varphi(f) + (1-\lambda)\varphi(f') \geq \varphi(\lambda f + (1-\lambda) f') + \frac{\lambda(1-\lambda)\sigma}{2}\|f - f'\|^2.$$

Rearranging, we see that a σ-strongly convex function has the property that

$$\varphi(\lambda f + (1-\lambda) f') \leq \lambda\varphi(f) + (1-\lambda)\varphi(f') - \frac{\lambda(1-\lambda)\sigma}{2}\|f - f'\|^2. \tag{8}$$

In other words, the value of φ at $\lambda f + (1-\lambda) f'$ is *strictly smaller* than the weighted average $\lambda\varphi(f) + (1 - \lambda)\varphi(f')$. It can be shown that a strongly convex function φ has a unique minimizer $f^* \in \mathscr{F}$, and moreover, for any other $f \in \mathscr{F}$,

$$\varphi(f) - \varphi(f') \geq \frac{\sigma}{2}\|f - f'\|^2. \tag{9}$$

Indeed, for any $\lambda \in (0, 1)$ and any $f \in \mathscr{F}$, we have

$$\lambda\varphi(f) + (1-\lambda)\varphi(f^*) \geq \varphi(\lambda f + (1-\lambda) f^*) + \frac{\lambda(1-\lambda)\sigma}{2}\|f - f'\|^2 \geq \varphi(f^*) + \frac{\lambda(1-\lambda)\sigma}{2}\|f - f^*\|^2,$$

where the first step uses (8) and the second step uses the optimality of f^*. Thus, for any $f \in \mathscr{F}$ and any $\lambda \in (0, 1)$,

$$\varphi(f) \geq \varphi(f^*) + \frac{(1-\lambda)\sigma}{2}\|f - f'\|^2.$$

Sending $\lambda \to 0$, we obtain (9).

We say that a differentiable (not necessarily convex) function $\varphi : \mathscr{F} \to \mathbb{R}$ is β-*smooth*, for some $\beta \geq 0$, if the gradient mapping $f \mapsto \nabla\varphi(f)$ is β-Lipschitz:

$$\|\nabla\varphi(f) - \nabla\varphi(f')\| \leq \beta\|f - f'\|, \qquad \forall f, f' \in \mathscr{F}. \tag{10}$$

3 Learnability without uniform convergence

We now show that we can have learnability without assuming uniform convergence:

Theorem 1. *Suppose that $\mathscr{F}$ is a convex subset of a Hilbert space $\mathscr{H}$, and there constants $L, \sigma > 0$, such that, for every $z \in \mathsf{Z}$, the function $f \mapsto \ell(f,z)$ is σ-strongly convex and L-Lipschitz. Then the ERM algorithm*

$$\widehat{f}_n = A(Z^n) = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i)$$

is such that

$$L(\widehat{f}_n) - L^* \leq \frac{4L^2}{\delta \sigma n},$$

with probability at least $1-\delta$.

Proof. The idea is to compare the output of A on the original training data Z^n to the output of A on the modified data, with one of the training samples replaced. Specifically, let $Z'_1, \ldots, Z'_n$ be n i.i.d. samples from P, independent of Z^n. For each i, define the modified training data

$$Z^n_{(i)} \triangleq (Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n),$$

and consider the corresponding ERM solution:

$$\widehat{f}^{(i)}_n \triangleq \operatorname*{arg\,min}_{f \in \mathscr{F}} L^{(i)}_n(f),$$

where

$$L^{(i)}_n(f) \triangleq \frac{1}{n} \ell(f, Z'_i) + \frac{1}{n} \sum_{j:\, j \neq i} \ell(f, Z_j)$$

is the empirical loss of f on the modified data $Z^n_{(i)}$. Let us compare the empirical losses of $\widehat{f}^{(i)}_n$ and $\widehat{f}_n$ on the *original* training data Z^n: using the definitions, we write

$$\begin{aligned}
L_n(\widehat{f}^{(i)}_n) - L_n(\widehat{f}_n) &= \frac{1}{n} \ell(\widehat{f}^{(i)}_n, Z_i) + \frac{1}{n} \sum_{j:\, j \neq i} \ell(\widehat{f}^{(i)}_n, Z_j) - \frac{1}{n} \ell(\widehat{f}_n, Z_i) - \frac{1}{n} \sum_{j:\, j \neq i} \ell(\widehat{f}_n, Z_j) \\
&= \frac{\ell(\widehat{f}^{(i)}_n, Z_i) - \ell(\widehat{f}_n, Z_i)}{n} + \frac{\sum_{j:\, j \neq i} \left[\ell(\widehat{f}^{(i)}_n, Z_j) - \ell(\widehat{f}_n, Z_j) \right]}{n} \\
&= \underbrace{\frac{\ell(\widehat{f}^{(i)}_n, Z_i) - \ell(\widehat{f}_n, Z_i)}{n}}_{T_1} + \underbrace{L^{(i)}_n(\widehat{f}^{(i)}_n) - L^{(i)}_n(\widehat{f}_n)}_{T_2} + \underbrace{\frac{\ell(\widehat{f}_n, Z'_i) - \ell(\widehat{f}^{(i)}_n, Z'_i)}{n}}_{T_3}.
\end{aligned}$$

Now, $T_2 \leq 0$ because $\widehat{f}^{(i)}_n$ minimizes empirical loss over the modified data $Z^n_{(i)}$, whereas both T_1 and T_3 are bounded from above by $\frac{L}{n} \| \widehat{f}_n - \widehat{f}^{(i)}_n \|$, by the Lipschitz property of ℓ. Therefore,

$$L_n(\widehat{f}^{(i)}_n) - L_n(\widehat{f}_n) \leq \frac{2L}{n} \| \widehat{f}_n - \widehat{f}^{(i)}_n \|. \tag{11}$$

On the other hand, by the strong convexity assumption on ℓ, the functional $f \mapsto L_n(f)$ is σ-strongly convex, and $\widehat{f}_n$ is its minimizer of $\mathscr{F}$. Therefore, by (9),

$$L_n(\widehat{f}^{(i)}_n) - L_n(\widehat{f}_n) \geq \frac{\sigma}{2} \| \widehat{f}_n - \widehat{f}^{(i)}_n \|^2. \tag{12}$$

From Eqs. (11) and (12), it follows that

$$\|\widehat{f}_n - \widehat{f}_n^{(i)}\| \le \frac{4L}{\sigma n}. \tag{13}$$

In other words, arbitrarily replacing any one sample in Z^n by some other Z_i' has only limited effect on the ERM solution, i.e., the algorithm output $A(Z^n)$ does not depend too much on any individual sample! But, because ℓ is Lipschitz, this implies that, for any $z \in \mathsf{Z}$,

$$\left|\ell(\widehat{f}_n, z) - \ell(\widehat{f}_n^{(i)}, z)\right| \le L\|\widehat{f}_n - \widehat{f}_n^{(i)}\| \le \frac{4L^2}{\sigma n}. \tag{14}$$

We now claim that the *stability property* (14) implies

$$\mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] \le \frac{4L^2}{\sigma n}. \tag{15}$$

Indeed, since $\widehat{f}_n$ is a function of Z^n, and since $Z_1', \ldots, Z_n'$ are independent of $Z_1, \ldots, Z_n$ and are draws from the same distribution, we can write

$$\mathbb{E}L(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(\widehat{f}_n, Z_i')].$$

On the other hand, since, for every i, $\ell(\widehat{f}_n, Z_n)$ and $\ell(\widehat{f}_n^{(i)}, Z_i')$ have the same distribution, we have

$$\mathbb{E}L_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(\widehat{f}_n, Z_i)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(\widehat{f}_n^{(i)}, Z_i')]$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell(\widehat{f}_n, Z_i') - \ell(\widehat{f}_n^{(i)}, Z_i')\right] \\
&\le \frac{1}{n}\sum_{i=1}^n \sup_{z\in\mathsf{Z}}\left|\ell(\widehat{f}_n, z) - \ell(\widehat{f}_n^{(i)}, z)\right| \\
&\le \frac{4L^2}{\sigma n},
\end{aligned}$$

as claimed. Eq. (15) shows that the ERM algorithm *generalizes well on average*, i.e., the empirical loss of $\widehat{f}_n = A(Z^n)$ on the data Z^n is a good estimate of $L(\widehat{f}_n) = L(A(Z^n))$ in expectation.

Now let $f^* \in \mathscr{F}$ achieve L^*. Since f^* doesn't depend on Z^n, we have $L^* = L(f^*) = \mathbb{E}[L_n(f^*)]$, and therefore

$$\begin{aligned}
\mathbb{E}\left[L(\widehat{f}_n) - L^*\right] &= \mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n) + L_n(\widehat{f}_n) - L_n(f^*)\right] \\
&\le \mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] \\
&\le \frac{4L^2}{\sigma n}.
\end{aligned}$$

From Markov's inequality, it then follows that

$$L(\widehat{f}_n) - L^* \le \frac{4L^2}{\delta\sigma n}$$

with probability at least $1-\delta$, and we are done. □

Armed with this theorem, we can now establish the following result for a complexity-regularized ERM:

Theorem 2. *Let $\mathscr{F}$ be a convex and norm-bounded subset of a Hilbert space $\mathscr{H}$, i.e., there exists some $B < \infty$, such that $\|f\| \le B$ for all $f \in \mathscr{F}$. Suppose also that, for each $z \in \mathsf{Z}$, the function $f \mapsto \ell(f,z)$ is convex and L-Lipschitz (note: we are not assuming strong convexity). For each $\lambda > 0$, consider the complexity-regularized ERM algorithm*

$$\widehat{f}_{n,\lambda} = A_\lambda(Z^n) \triangleq \operatorname*{arg\,min}_{f \in \mathscr{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(f, Z_i) + \frac{\lambda}{2} \|f\|^2 \right\}.$$

Then $\widehat{f}_n = \widehat{f}_{n,\lambda}$ with $\lambda = \frac{L}{B\sqrt{n}}$ satisfies

$$L(\widehat{f}_n) \le L^* + \frac{LB}{2\sqrt{n}} + \frac{8LB}{\delta\sqrt{n}} + \frac{8LB}{\delta n^{3/2}}$$

with probability at least $1-\delta$.

Proof. Consider the function $\ell_\lambda(f,z) \triangleq \ell(f,z) + \frac{\lambda}{2}\|f\|^2$. This function is clearly λ-strongly convex. Moreover, for any fixed z and any $f, f' \in \mathscr{F}$, we have

$$\begin{aligned}
\left|\ell_\lambda(f,z) - \ell_\lambda(f',z)\right| &\le \left|\ell(f,z) - \ell(f',z)\right| + \frac{\lambda}{2}\left|\|f\|^2 - \|f'\|^2\right| \\
&\le L\|f - f'\| + \frac{\lambda}{2}\left|\|f\| - \|f'\|\right| \cdot \left|\|f\| + \|f'\|\right| \\
&\le L\|f - f'\| + \lambda B \|f - f'\|,
\end{aligned}$$

i.e., ℓ_λ is λ-strongly convex and $(L+\lambda B)$-Lipschitz. For each $f \in \mathscr{F}$, let $L_\lambda(f) \triangleq L(f) + \frac{\lambda}{2}\|f\|^2 \equiv \mathbb{E}[\ell_\lambda(f,Z)]$. Applying Theorem 1, we conclude that, with probability at least $1-\delta$,

$$L_\lambda(\widehat{f}_{n,\lambda}) - L^*_\lambda \le \frac{4(L+\lambda B)^2}{\delta\lambda n}, \tag{16}$$

where $L^*_\lambda \triangleq \inf_{f\in\mathscr{F}} L_\lambda(f)$. Therefore, with the same probability,

$$\begin{aligned}
L(\widehat{f}_{n,\lambda}) &\le L^*_\lambda + \frac{4(L+\lambda B)^2}{\delta\lambda n} \\
&= \inf_{f\in\mathscr{F}} \left\{ L(f) + \frac{\lambda}{2}\|f\|^2 \right\} + \frac{4(L+\lambda B)^2}{\delta\lambda n} \\
&\le L^* + \frac{\lambda}{2}\|f^*\|^2 + \frac{4(L+\lambda B)^2}{\delta\lambda n} \\
&\le L^* + \frac{\lambda B^2}{2} + \frac{4(L+\lambda B)^2}{\delta\lambda n} \\
&\le L^* + \frac{\lambda B^2}{2} + \frac{8L^2}{\delta\lambda n} + \frac{8\lambda B^2}{\delta n}.
\end{aligned}$$

□

4 Learnability and stability

Let us now see how the above ideas can be abstracted into a general set of results about learnability and stability. We say that a learning algorithm $A : \mathsf{Z}^* \to \mathscr{F}$ is *stable on average* (with respect to replace-one operation) if

$$\bar{s}_n(A) \triangleq \sup_P \frac{1}{n} \left| \sum_{i=1}^{n} \left[\ell(A(Z^n_{(i)}), Z'_i) - \ell(A(Z^n), Z'_i) \right] \right| \xrightarrow{n\to\infty} 0. \tag{17}$$

We have already proved the following:

Lemma 2. *For any learning algorithm, $\bar{g}_n(A) = \bar{s}_n(A)$. In particular, A is stable on average if and only if it generalizes on average.*

Now we can show more:

Lemma 3. *If A is an AERM that generalizes on average, then it generalizes, and moreover*

$$g_n(A) \le \bar{g}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}}. \tag{18}$$

Proof. We begin by decomposing the difference $L_n(A(Z^n)) - L(A(Z^n))$:

$$\begin{aligned} L_n(A(Z^n)) - L(A(Z^n)) &= L(A(Z^n)) - L_n^* + L_n^* - L_n(f^*) + L_n(f^*) - L(f^*) \\ &\le L(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*). \end{aligned}$$

Applying Lemma 8 in the Appendix to $U \triangleq L_n(A(Z^n)) - L(A(Z^n))$ and $V \triangleq L(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*)$, we get

$$\begin{aligned} \mathbb{E}\left|L_n(A(Z^n)) - L(A(Z^n))\right| &\le \left|\mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))]\right| + 2\mathbb{E}\left|L(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*)\right| \\ &\le \left|\mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))]\right| + 2\mathbb{E}\left|L(A(Z^n)) - L_n^*\right| + 2\mathbb{E}|L_n(f^*) - L(f^*)| \\ &\le \bar{g}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}}, \end{aligned}$$

where in the last line we have used the assumed properties of A, together with the fact that, for any f, $\mathbb{E}|L_n(f) - L(f)| = \mathbb{E}|L_n(f) - \mathbb{E}L_n(f)| \le \sqrt{\mathbb{E}(L_n(f) - \mathbb{E}L_n(f)^2)} \le \frac{1}{\sqrt{n}}$, since ℓ is bounded between 0 and 1. This completes the proof. □

Corollary 1. *If A is an AERM which is stable on average, then it generalizes, and moreover*

$$g_n(A) \le \bar{s}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}}.$$

All of this leads to the following result:

Theorem 3. *An AERM learning algorithm A is stable on average if and only if it generalizes, with*

$$g_n(A) - 2e_n(A) - \frac{2}{\sqrt{n}} \le \bar{s}_n(A) \le g_n(A). \tag{19}$$

Moreover, if A is an AERM which is stable on average, then it is consistent, with

$$c_n(A) \le \bar{s}_n(A) + e_n(A). \tag{20}$$

Proof. We have already proved that stability is equivalent to generalization on average, but for an AERM generalization on average implies generalization. Conversely, if A generalizes, then it generalizes on average, and is therefore stable on average.

To prove the second part, since $\bar{s}_n(A) = \bar{g}_n(A)$, we have

$$\begin{aligned}\mathbb{E}[L(A(Z^n)) - L^*] &= \mathbb{E}[L(A(Z^n)) - L_n(A(Z^n)) + L_n(A(Z^n)) - L_n(f^*)] \\ &\le \mathbb{E}[L(A(Z^n)) - L_n(A(Z^n))] + \mathbb{E}[L_n(A(Z^n)) - L_n^*] \\ &\le \bar{g}_n(A) + e_n(A) \\ &= \bar{s}_n(A) + e_n(A).\end{aligned}$$

□

5 Stability of stochastic gradient descent

WARNING: This section is still very rough, needs more editing.

One of the most popular algorithms for learning over complicated hypothesis classes (such as deep neural networks) is the Stochastic Gradient Descent (SGD) algorithm. The basic idea behind SGD is as follows. For a fixed training set $Z^n = (Z_1, \dots, Z_n)$, the usual ERM approach requires minimizing the function

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) \tag{21}$$

over the hypothesis class $\mathscr{F}$. One way to go about this is to use gradient descent: assuming that the function $f \mapsto \ell(f, z)$ is differentiable for each $z \in \mathsf{Z}$, we can set the initial condition $f_0 \in \mathscr{F}$ and iteratively compute

$$f_t = f_{t-1} - \alpha_t \nabla L_n(f_{t-1}), \qquad t = 1, 2, \dots \tag{22}$$

where

$$\nabla L_n(f_{t-1}) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(f_{t-1}, Z_i)$$

is the gradient of L_n at f_{t-1}, and $\{\alpha_t\}_{t=1}^{\infty}$ is a monotonically decreasing sequence of nonnegative reals typically referred to as step sizes. (For simplicity, we are assuming here that the updates defined in (22) stay in $\mathscr{F}$; in general, one needs an additional step of projecting onto $\mathscr{F}$.) Under certain mild conditions on ℓ and $\mathscr{F}$, and with appropriately tuned step sizes, one can guarantee that

$$L_n(f_t) \to \inf_{f \in \mathscr{F}} L_n(f) \equiv L_n^* \qquad \text{as } t \to \infty.$$

In other words, for each n, we can find a large enough T_n, such that $A(Z^n) = f_{T_n}$ is an AERM algorithm.

However, one disadvantage of gradient descent is that each update (22) requires a sweep through the entire sample Z^n in order to compute the gradient ∇L_n. Thus, the complexity of each step of the gradient descent method scales as $O(n)$. SGD offers a way around this limitation and allows to reduce

the complexity of each iteration to $O(1)$. If we look at (21), we see that the empirical loss $L_n(f)$ can be written as an average of n functions of f:

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f), \qquad \text{where } \ell_i(f) \triangleq \ell(f, Z_i).$$

In each iteration of SGD, we pick a random index $I_t \in \{1,\dots,n\}$ and update

$$f_t = f_{t-1} - \alpha_t \nabla \ell_{I_t}(f_{t-1}) \equiv f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, Z_{I_t}). \tag{23}$$

Thus, SGD is a *randomized* algorithm. Two popular choices for selecting the indices I_t are:

- **Random selection:** At each time step t, I_t is drawn uniformly at random from $\{1,\dots,n\}$, independently of all past realizations $I_1,\dots,I_{t-1}$.
- **Random permutation:** At time $t = 0$, we draw a permutation σ of the set $\{1,\dots,n\}$ uniformly at random and then cycle through the samples $Z_{\sigma(1)}, Z_{\sigma(2)},\dots,Z_{\sigma(n)}$:

$$I_t = \sigma(1 + [(t-1) \bmod n]).$$

In a recent paper, Hardt et al. [HRS15] have shown that SGD with suitably tuned step sizes and number of updates gives a stable learning algorithm. Under different assumptions on the loss function ℓ, we end up with different conditions for stability. In order to proceed, let us first examine the evolution of SGD updates for a fixed training set Z^n. Fix a differentiable function $\varphi : \mathscr{F} \to \mathbb{R}$ and a step size $\alpha \geq 0$, and define an operator $G_{\varphi,f} : \mathscr{F} \to \mathscr{F}$ by

$$G(f) \triangleq f - \alpha \nabla \varphi(f). \tag{24}$$

Again, to keep things simple, we assume that the image of $\mathscr{F}$ under G is contained in $\mathscr{F}$. Then we can write the tth update of SGD as

$$f_t = G_t(f_{t-1}), \qquad \text{where } G_t = G_{\ell(\cdot,Z_{I_t}),\alpha_t}. \tag{25}$$

Now let us fix some $i \in \{1,\dots,n\}$ and consider running SGD with the same realization of the random indices $\{I_t\}$ on another training set $Z'^n = (Z'_1,\dots,Z'_n)$ that differs from Z^n only in one sample. Denoting by $\{f'_t\}$ the corresponding updates with $f'_0 = f_0$, we can write

$$f'_t = G'_t(f'_{t-1}), \qquad \text{where } G'_t = \begin{cases} G_t, & \text{if } Z_{I_t} = Z'_{I_t} \\ G_{\ell(\cdot,Z'_{I_t}),\alpha_t}, & \text{otherwise.} \end{cases} \tag{26}$$

For each $t = 0, 1, \dots$, let $\delta_t \triangleq \|f_t - f'_t\|$, with the initial condition $\delta_0 = 0$. We can now track the evolution of δ_t as follows:

- If $G_t = G'_t$, then

$$\begin{aligned} \delta_t &= \|f_t - f'_t\| && (27)\\ &= \|G_t(f_{t-1}) - G_t(f'_{t-1})\| && (28)\\ &\leq \eta_t \|f_{t-1} - f'_{t-1}\| && (29)\\ &\equiv \eta_t \delta_t, && (30) \end{aligned}$$

  where we have defined

$$\eta_t \triangleq \sup_{f,f' \in \mathscr{F}} \frac{\|G_t(f) - G_t(f')\|}{\|f - f'\|}. \tag{31}$$

- If $G_t \neq G'_t$, then, on the one hand,

$$\delta_t = \|G_t(f_{t-1}) - G'_t(f_{t-1})\| \quad \leq \|G_t(f_{t-1}) - f_{t-1}\| + \|f_{t-1} - f'_{t-1}\| + \|G'_{t-1}(f_{t-1}) - f'_{t-1}\| \leq 2c_t + \delta_{t-1}, \tag{32}$$

where we have defined

$$c_t \triangleq \max\left\{\sup_{f\in\mathscr{F}} \|G_t(f) - f\|, \sup_{f\in\mathscr{F}} \|G'_t(f) - f\|\right\}, \tag{33}$$

and on the other hand,

$$\begin{aligned}
\delta_t &= \|G_t(f_{t-1}) - G'_t(f'_{t-1})\| && (34)\\
&\leq \|G_t(f_{t-1}) - G_t(f'_{t-1})\| + \|G_t(f'_{t-1}) - f'_{t-1}\| + \|G'_t(f'_{t-1}) - f'_{t-1}\| && (35)\\
&\leq \eta_t \delta_{t-1} + 2c_t. && (36)
\end{aligned}$$

This gives us the bound

$$\delta_t \leq (1 \wedge \eta_t)\delta_{t-1} + 2c_t. \tag{37}$$

Summarizing:

$$\delta_t \leq \begin{cases} \eta_t \delta_{t-1}, & \text{if } G_t = G'_t \\ (1\wedge\eta_t)\delta_{t-1} + 2c_t, & \text{otherwise} \end{cases}. \tag{38}$$

This will be our main tool for analyzing the stability of SGD. Another tool is the following estimate:

Lemma 4. *Suppose that, for each $z \in \mathsf{Z}$, the function $f \mapsto \ell(f,z)$ is L-Lipschitz, and takes values in $[0,1]$. Let $\{f_t\}_{t=0}^T$ and $\{f'_t\}_{t=0}^T$ be the updates of SGD (either with random selection or with random permutation) run on two datasets Z^n and Z'^n that differ only in one sample, with $f_0 = f'_0$. Then, for any $t_0 \in \{0,1,\ldots,n\}$ and for any $z \in \mathsf{Z}$,*

$$\mathbb{E}\left[|\ell(f_T,z) - \ell(f'_T,z)|\right] \leq L\mathbb{E}[\delta_T \mathbf{1}\{\delta_{t_0} = 0\}] + \frac{t_0}{n}. \tag{39}$$

Proof. We start by writing

$$\begin{aligned}
|\ell(f_T,z) - \ell(f'_T,z)| &= |\ell(f_T,z) - \ell(f'_T,z)|\mathbf{1}\{\delta_{t_0} = 0\} + |\ell(f_T,z) - \ell(f'_T,z)|\mathbf{1}\{\delta_{t_0} \neq 0\} && (40)\\
&\leq L\|f_T - f'_T\|\mathbf{1}\{\delta_{t_0} = 0\} + \mathbf{1}\{\delta_{t_0} \neq 0\}, && (41)
\end{aligned}$$

where in the second step we have used the fact that ℓ is L-Lipschitz and takes values in $[0,1]$. Taking expectations, we get

$$\mathbb{E}\left[|\ell(f_T,z) - \ell(f'_T,z)|\right] \leq L\mathbb{E}\left[\delta_T \mathbf{1}\{\delta_{t_0} = 0\}\right] + \mathbb{P}\left[\delta_{t_0} \neq 0\right]. \tag{42}$$

It remains to bound the probability on the right-hand side. To that end, let $i^* \in \{1,\ldots,n\}$ be the position where the two training sets differ. Define an integer-valued random variable I as the first time that the SGD algorithm uses the sample Z_{i^*}. Now observe that if $I > t_0$, then $\delta_{t_0} = 0$, because the updates f_t and f'_t are the same for all $t < I$. Consequently, $\mathbb{P}[\delta_{t_0} \neq 0] \leq \mathbb{P}[I \leq t_0]$. To upper-bound the latter probability, we consider the two index selection rules separately:

- In the case of random permutation, the samples are reshuffled as

$$(Z_1,\dots,Z_{i^*},\dots,Z_n)\mapsto(Z_{\sigma(1)},\dots,Z_{\sigma(i^*)},\dots,Z_{\sigma(n)})$$
$$(Z'_1,\dots,Z'_{i^*},\dots,Z'_n)\mapsto(Z'_{\sigma(1)},\dots,Z'_{\sigma(i^*)},\dots,Z'_{\sigma(n)}).$$

  While the two original datasets Z^n and Z'^n differed in the i^*th position, the permuted datasets now differ in position $J^*=\sigma^{-1}(i^*)$, where σ is the permutation of $\{1,\dots,n\}$ drawn uniformly at random at time $t=0$. Therefore, $I=J^*$. It is not hard to see that the event $\{J^*=j\}$ has probability $1/n$ for all $j=1,\dots,n$ — there are $n!$ permutations, and among these there are $(n-1)!$ permutations with $\sigma(j)=i^*$. Therefore, in the random permutation case,

$$\mathbb{P}[I\le t_0]=\mathbb{P}[\sigma^{-1}(i^*)\le t_0]=\frac{t_0}{n}. \tag{43}$$

- In the case of random selection, the index I_t is drawn uniformly at random from $\{1,\dots,n\}$, independently of the past realizations $I_1,\dots,I_{t-1}$. Therefore,

$$\mathbb{P}[I\le t_0]=\mathbb{P}\left[\bigcup_{t=1}^{t_0}\{I_t=i^*\}\right]\le\sum_{t=1}^{t_0}\mathbb{P}[I_t=i^*]=\frac{t_0}{n}. \tag{44}$$

Thus, in both cases, the probability $\mathbb{P}[\delta_{t_0}\neq 0]\le t_0/n$. □

Now we can analyze the stability of SGD under several assumptions on the loss ℓ:

Theorem 4. *Suppose that, for each $z\in\mathsf{Z}$, the loss function $f\mapsto\ell(f,z)$ is convex, β-smooth, and L-Lipschitz. Suppose that we run SGD with step sizes $\alpha_t\le 2/\beta$ for T time steps. Then, for any two datasets Z^n and Z'^n that differ in only one sample,*

$$\sup_{z\in\mathsf{Z}}\mathbb{E}|\ell(f_T,z)-\ell(f'_T,z)|\le\frac{2L^2}{n}\sum_{t=1}^{T}\alpha_t, \tag{45}$$

where the expectation is only with respect to the internal randomness of SGD (i.e., index selection).

Proof. It can be shown that, if φ is a convex, β-smooth, L-Lipschitz function, then the mapping $G=G_{\varphi,\alpha}$ with $\alpha\le 2/\beta$ satisfies

$$\sup_{f,f'\in\mathscr{F}}\frac{\|G(f)-G(f')\|}{\|f-f'\|}\le 1\qquad\text{and}\qquad\sup_{f\in\mathscr{F}}\|G(f)-f\|\le\alpha L. \tag{46}$$

Applying Lemma 4 with $t_0=0$, we have

$$\mathbb{E}|\ell(f_T,z)-\ell(f'_T,z)|\le L\mathbb{E}[\delta_T]. \tag{47}$$

In order to bound $\mathbb{E}[\delta_T]$, we will use (38). Let's consider what happens at time t. Let F_t denote the event that, at time t, the SGD algorithm does not use different samples in Z^n and Z'^n. If F_t occurs, then $G_t=G'_t$, otherwise $G_t\neq G'_t$. In that case,

$$\mathbb{E}[\delta_t]=\mathbb{E}[\delta_t\mathbf{1}_{F_t}]+\mathbb{E}[\delta_t\mathbf{1}_{F_t^c}] \tag{48}$$
$$\le\mathbb{E}[\eta_t\delta_{t-1}\mathbf{1}_{F_t}]+\mathbb{E}\left[\left((1\wedge\eta_t)\delta_{t-1}+2c_t\right)\mathbf{1}_{F_t^c}\right]. \tag{49}$$

By (46), we can take $\eta_t = 1$ and $c_t = \alpha_t L$. Thus,

$$\mathbb{E}[\delta_t] \le \mathbb{E}[\delta_{t-1}] + 2\alpha_t L \mathbb{P}[F_t^c]. \tag{50}$$

In both the random permutation and the random selection case, $\mathbb{P}[F_t^c] = 1/n$. Therefore, we end up with the recursion

$$\mathbb{E}[\delta_t] \le \mathbb{E}[\delta_{t-1}] + \frac{2\alpha_t L}{n}, \qquad t = 1, 2, \ldots, T \tag{51}$$

with the initial condition $\delta_0 = 0$. Unwinding the recursion, we get

$$\mathbb{E}[\delta_T] \le \frac{2L}{n} \sum_{t=1}^{T} \alpha_t. \tag{52}$$

Substituting this into Eq. (47), we are done. □

For example, if we set $T = n$ and $\alpha_t = 2/\beta\sqrt{n}$ for all t, then

$$\sum_{t=1}^{T} \alpha_t = \frac{2\sqrt{n}}{\beta}, \tag{53}$$

and then the algorithm $A(Z^n)$ obtained by running SGD for $\sqrt{n}$ steps with constant step size $\alpha = 2/\beta\sqrt{n}$ is stable with $\bar{s}_n(A) \le 4L^2/\beta\sqrt{n}$.

If we now assume that ℓ is also strongly convex, we get a bound that does not depend on the number of iterations T:

Theorem 5. *Suppose that ℓ satisfies the conditions of Theorem 4, and also that the function $f \mapsto \ell(f, z)$ is γ-strongly convex for each $z \in \mathsf{Z}$. Suppose that we run SGD with a constant step size $\frac{1}{\beta} \le \alpha \le \frac{2}{\beta+\gamma}$ for T time steps. Then, for any two datasets Z^n and Z'^n that differ in only one sample,*

$$\sup_{z \in \mathsf{Z}} \mathbb{E}|\ell(f_T, z) - \ell(f'_T, z)| \le \frac{2L^2}{\gamma n}, \tag{54}$$

where the expectation is only with respect to the internal randomness of SGD (i.e., index selection).

Proof. The proof is similar to the proof of Theorem 4. First of all, it can be shown that, under our assumptions on ℓ and on α, we have

$$\eta_t \le 1 - \alpha\gamma \qquad \text{and} \qquad c_t \le \alpha L. \tag{55}$$

Then the same steps that led to (51) give

$$\mathbb{E}[\delta_t] \le (1 - \alpha\gamma)\mathbb{E}[\delta_{t-1}] + \frac{2\alpha L}{n}, \tag{56}$$

with the initial condition $\delta_0 = 0$. Unwinding the recursion, we get

$$\mathbb{E}[\delta_T] \le \frac{2\alpha L}{n} \sum_{t=1}^{T} (1 - \alpha\gamma)^{t-1} \le \frac{2\alpha L}{n} \cdot \frac{1}{\alpha\gamma} = \frac{2L}{\gamma n}. \tag{57}$$

The result follows. □

Finally, we derive a stability estimate for SGD without requiring convexity, but still assuming Lipschitz-continuity and smoothness:

Theorem 6. *Suppose that, for each $z \in \mathsf{Z}$, the loss function $f \mapsto \ell(f,z)$ β-smooth and L-Lipschitz. Suppose that we run SGD with step sizes $\alpha_t \le c/t$ for T time steps, where $c > 0$ is some constant. Then, for any two datasets Z^n and Z'^n that differ in only one sample,*

$$\sup_{z \in \mathsf{Z}} \mathbb{E}|\ell(f_T,z) - \ell(f'_T,z)| \le \frac{1+1/\beta c}{n} (2cL^2)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}, \tag{58}$$

where the expectation is only with respect to the internal randomness of SGD (i.e., index selection).

Proof. Here the idea is to apply Lemma 4 with an arbitrary $t_0 \in \{0,1,\ldots,n\}$, and then optimize over t_0. Under the assumptions on ℓ, we can apply (38) with

$$\eta_t \le 1 + \alpha_t \beta \qquad \text{and} \qquad c_t \le \alpha_t L. \tag{59}$$

By Lemma 4, for a fixed $t_0 \in \{0,1,\ldots,n\}$, we have

$$\mathbb{E}\left[|\ell(f_T,z) - \ell(f'_T,z)|\right] \le L\mathbb{E}[\delta_T \mathbf{1}\{\delta_{t_0} = 0\}] + \frac{t_0}{n}. \tag{60}$$

Let us denote $\Delta_t \triangleq \mathbb{E}[\delta_t | \delta_{t_0} = 0]$. For any time $t > t_0$, we have

$$\begin{aligned}
\Delta_t &= \mathbb{E}[\delta_t \mathbf{1}_{F_t} | \delta_{t_0} = 0] + \mathbb{E}[\delta_t \mathbf{1}_{F_t^c} | \delta_{t_0} = 0] && (61)\\
&\le (1+\alpha_t\beta)\mathbb{E}[\delta_{t-1}\mathbf{1}_{F_t}|\delta_{t_0}=0] + \mathbb{E}[\delta_{t-1}\mathbf{1}_{F_t^c}|\delta_{t_0=0}] + 2\alpha_t L \mathbb{P}[F_t^c|\delta_{t_0}=0] && (62)\\
&\le \left(1 + \frac{\beta c}{t}\right)\Delta_{t-1} + \frac{2cL}{tn} && (63)\\
&\le \exp\left(\beta c/t\right)\Delta_{t-1} + \frac{2cL}{tn}, && (64)
\end{aligned}$$

where in the last line we have used the inequality $1 + u \le \exp(u)$. Unwinding the recursion down to $t = t_0 + 1$ and using the initial condition $\Delta_{t_0} = 0$, we have

$$\begin{aligned}
\Delta_T &\le \sum_{t=t_0+1}^{T} \prod_{k=t+1}^{T} \exp(\beta c/k) \frac{2cL}{tn} && (65)\\
&= \sum_{t=t_0+1}^{T} \exp\left(\beta c \sum_{k=t+1}^{T} \frac{1}{k}\right) \frac{2cL}{tn} && (66)\\
&\le \sum_{t=t_0+1}^{T} \exp\left(\beta c \log \frac{T}{t}\right) \frac{2cL}{tn} && (67)\\
&\le \frac{2cL}{n} T^{\beta c} \sum_{t=t_0+1}^{T} t^{-(1+\beta c)} && (68)\\
&\le \frac{2cL}{n} T^{\beta c} \frac{1}{\beta c} \left(t_0^{-\beta c} - T^{-\beta c}\right) && (69)\\
&\le \frac{2L}{n\beta} \left(\frac{T}{t_0}\right)^{\beta c}. && (70)
\end{aligned}$$

Plugging this estimate into (60), we get

$$\mathbb{E}\left[|\ell(f_T,z)-\ell(f'_T,z)|\right] \le \frac{t_0}{n}+\frac{2L^2}{n\beta}\left(\frac{T}{t_0}\right)^{\beta c}. \tag{71}$$

The right-hand side is (approximately) minimized by setting

$$t_0 = \left(2cL^2\right)^{\frac{1}{q+1}} T^{\frac{q}{q+1}}, \qquad q=\beta c$$

which gives

$$\mathbb{E}\left[|\ell(f_T,z)-\ell(f'_T,z)|\right] \le \frac{1+1/\beta c}{n}(2cL^2)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}. \tag{72}$$

□

In this case, we can set $T = n^{\varepsilon(1+1/\beta c)}$ for any $\varepsilon \in (0,1)$, and obtain the stability bound

$$s_n(A) \le \frac{1+1/\beta c}{n}(2cL^2)^{\frac{1}{\beta c+1}} n^{\varepsilon-1} \tag{73}$$

for $A(Z^n) = f_T$.

6 Differentially private algorithms and generalization

Recall that we have defined a randomized learning algorithm A to be stable if the outputs $A(Z^n)$ and $A(Z'^n)$ of A on two training sets Z^n and Z'^n that differ in only one example are close in terms of their losses: for example, A is ε-uniformly stable if

$$\sup_{z\in\mathsf{Z}}\left[\mathbb{E}\ell(A(Z^n),z)-\mathbb{E}\ell(A(Z'^n),z)\right] \le \varepsilon \tag{74}$$

for all Z^n and Z'^n that differ in only one example.

In this section, we will examine a much stronger stability property that pertains to the sensitivity of the conditional distribution of the output of A given $Z^n = z^n$ to individual training examples comprising Z^n. For this purpose, it is convenient to think of $F = A(Z^n)$ as a random object taking values in the hypothesis class $\mathscr{F}$. Then the operation of A is fully described by the conditional distribution $P_{F|Z^n}$. Moreover, we can rewrite the stability condition (74) in the following equivalent form:

$$\sup_{z\in\mathsf{Z}}\left[\mathbb{E}[\ell(F,z)|Z^n=z^n]-\mathbb{E}[\ell(F,z)|Z^n=z'^n]\right] \le \varepsilon \tag{75}$$

for any two training sets z^n, z'^n that differ in at most one example. Let us now consider a stronger property that compares the conditional *distribution* of F given $Z^n = z^n$ against the one given $Z^n = z'^n$:

Definition 1. *A randomized algorithm A specified by the conditional distribution $P_{F|Z^n}$ is* (ε,δ)-differentially private *if, for any measurable subset B of $\mathscr{F}$ and for any two trainig sets z^n, z'^n that differ in at most one example, we have*

$$P\left[F\in B|Z^n=z^n\right] \le e^{\varepsilon}P\left[F\in B|Z^n=z'n\right]+\delta. \tag{76}$$

Equivalently, $P_{F|Z^n}$ is (ε,δ)-differentially private if, for any function $g:\mathscr{F}\to[0,1]$,

$$\mathbb{E}[g(F)|Z^n=z^n] \le e^{\varepsilon}\mathbb{E}[g(F)|Z^n=z'^n]+\delta. \tag{77}$$

This definition was proposed by Cynthia Dwork in the context of protecting individual information in statistical databases [Dwo06]. Of course, it is useful only for $\delta \in [0,1)$ and for suitably small values of ε.

We start with the following simple observation:

Lemma 5. *If a learning algorithm $P_{F|Z^n}$ is (ε,δ)-differentially private, then it is $(e^\varepsilon - 1 + \delta)$-uniformly stable in the sense of (74). If $\varepsilon \in [0,1]$, then the algorithm is $(2\varepsilon+\delta)$-uniformly stable.*

Proof. A direct consequence of the definition: let z^n, z'^n be two training sets differing in only one example. Then, for any $z \in \mathsf{Z}$,

$$\begin{aligned}\mathbb{E}[\ell(F,z)|Z^n = z^n] - \mathbb{E}[\ell(F,z)|Z^n = z'^n] &\leq (e^\varepsilon - 1)\mathbb{E}[\ell(F,z)|Z^n = z'^n] + \delta \\ &\leq e^\varepsilon - 1 + \delta.\end{aligned}$$

Since $e^u - 1 \leq 2u$ for $u \in [0,1]$, we also obtain the second part of the lemma. □

This stability estimate immediately implies that a differentially private algorithm should generalize. However, the resulting bounds are rather loose. We will now present a tighter bound, due to Nissim and Stemmer [NS15].

First, we need to collect some preliminaries on the properties of differentially private algorithms. Fix a randomized algorithm $A = P_{F|Z^n}$ and consider a new algorithm obtained by running M copies of A in parallel on m training sets $(Z_{j,1},\ldots,Z_{j,n})$, $1 \leq j \leq m$. In other words, we form the matrix

$$Z^{m\times n} = \begin{pmatrix} Z_{1,1} & Z_{1,2} & \ldots & Z_{1,n} \\ Z_{2,1} & Z_{2,2} & \ldots & Z_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m,1} & Z_{m,2} & \ldots & Z_{m,n} \end{pmatrix}$$

and let F_j be the output of A on the jth row of $Z^{m\times n}$. This defines a new algorithm, which we denote by A^m and which is described by the following conditional distribution $P_{F^m|Z^{m\times n}}$: For any m measurable sets $B_1,\ldots,B_m \subset \mathscr{F}$,

$$P\left[F_1 \in B_1,\ldots,F_m \in B_n | Z^{m\times n} = z^{m\times n}\right] = \prod_{j=1}^m P\left[F \in B_m | Z^n = (z_{j,1},\ldots,z_{j,n})\right].$$

If A is (ε,δ)-differentially private, then the algorithm A^m constructed in this way is also (ε,δ)-differentially private. This follows almost immediately from the fact that the jth component of the output of the new algorithm depends only on the jth row of the matrix $Z^{m\times n}$.

Another way of combining algorithms is by *adaptive composition*. Consider two randomized algorithms, $A_1 = P_{F_1|Z^n}$ and $A_2 = P_{F_2|Z^n,F_1}$. Here, the first algorithm takes a dataset Z^n and produces an output $F_1 \in \mathscr{F}_1$; the second algorithm takes a dataset Z^n and an additional $\mathscr{F}_1$-valued input F_1 and produces an output $F_2 \in \mathscr{F}_2$. The *adaptive composition* of A_1 and A_2 takes Z^n as input and produces an output $F_2 \in \mathscr{F}_2$ using a two-stage procedure:

- Generate F_1 by running A_1 on Z^n.
- Generate F_2 by running A_2 on Z^n and on F_1 generated by A_1.

Suppose that A_1 is (ε_1,δ_1)-differentially private, and that, for each $f_1 \in \mathscr{F}_1$, $P_{F_2|Z^n,F_1=f_1}$ is (ε_2,δ_2)-differentially private. Then their adaptive composition is $(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2)$-differentially private. We will now prove this: Fix an arbitrary function $g:\mathscr{F}_2 \to [0,1]$ and two datasets z^n, z'^n that differ in only one sample, and write

$$\begin{aligned}\int_{\mathscr{F}_2} g(f_2) P_{F_1,F_2|Z^n=z^n}(\mathrm{d}f_1,\mathrm{d}f_2) &= \int_{\mathscr{F}_1}\left(\int_{\mathscr{F}_2} g(f_2) P_{F_2|Z^n=z^n,F_1=f_1}(\mathrm{d}f_2)\right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1)\\ &\le \int_{\mathscr{F}_1} \min\left(1, e^{\varepsilon_2}\int_{\mathscr{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2) + \delta_2\right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1)\\ &= \int_{\mathscr{F}_1} \min\left(1, e^{\varepsilon_1}\int_{\mathscr{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2)\right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) + \delta_2, \end{aligned} \tag{78}$$

where we have used the differential privacy assumption on A_2. Now, for a fixed realization z'^n, we can define the function

$$g'(f_1) \triangleq \min\left(1, e^{\varepsilon_2}\int_{\mathscr{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2)\right)$$

that takes values in $[0,1]$. Therefore, by the differential privacy assumption on A_1,

$$\begin{aligned}&\int_{\mathscr{F}_1} \min\left(1, e^{\varepsilon_2}\int_{\mathscr{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2)\right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1)\\ &\quad= \int_{\mathscr{F}_1} g'(f_1) P_{F_1|Z^n=z^n}(\mathrm{d}f_1)\\ &\quad\le e^{\varepsilon_1}\int_{\mathscr{F}_1} g'(f_1) P_{F_1|Z^n=z'^n}(\mathrm{d}f_1) + \delta_1\\ &\quad\le e^{\varepsilon_1+\varepsilon_2}\int_{\mathscr{F}_1}\int_{\mathscr{F}_2} g(f_2) P_{F_1,F_2|Z^n=z'^n}(\mathrm{d}f_1,\mathrm{d}f_2) + \delta_1. \end{aligned} \tag{79}$$

Using the bound (79) in (78), we obtain

$$\mathbb{E}[g(F_2)|Z^n=z^n] \le e^{\varepsilon_1+\varepsilon_2}\mathbb{E}[g(F_2)|Z^n=z'^n] + (\delta_1+\delta_2).$$

Since g was arbitrary, we have established the desired differential privacy property.

Finally, we will need a particular differentially private algorithm, the so-called *exponential mechanism* of McSherry and Talwar [MT07]. Suppose that we are given a function $U:\mathscr{S}\times \mathsf{Z}^n \to \mathbb{R}$, where $\mathscr{S}$ is a finite set, such that

$$\max_{s\in\mathscr{S}} |U(s,z^n) - U(s,z'^n)| \le 1$$

for all z^n, z'^n that differ in only one sample. Consider a randomized algorithm that takes input Z^n and generates an output S taking values in $\mathscr{S}$ according to the following distribution:

$$P_{S|Z^n=z^n}(s) = \frac{e^{\varepsilon U(s,z^n)/2}}{\sum_{s'\in\mathscr{S}} e^{\varepsilon U(s',z^n)/2}}. \tag{80}$$

We have the following:

Lemma 6. *The exponential algorithm (80) has the following properties:*

1. *It is ε-differentially private.*
2. *Let $U^*(z^n) \triangleq \max_{s\in\mathscr{S}} U(s,z^n)$. Then, for any t,*

$$P\left[U(S,Z^n) < U^*(Z^n) - t | Z^n = z^n\right] \le |\mathscr{S}| e^{-\varepsilon t/4}. \tag{81}$$

Proof. For part 1, fix z^n, z'^n differing in only one sample. Then we have

$$\begin{aligned}
\frac{P[S=s|Z^n=z^n]}{P[S=s|Z^n=z'^n]} &= \frac{e^{\varepsilon U(s,z^n)/2}}{e^{\varepsilon U(s,z'^n)/2}} \cdot \frac{\sum_{s'\in\mathscr{S}} e^{\varepsilon U(s',z'^n)/2}}{\sum_{s'\in\mathscr{S}} e^{\varepsilon U(s',z^n)/2}} \\
&= \exp\left(\frac{\varepsilon(U(s,z^n)-U(s,z'^n))}{2}\right) \cdot \frac{\sum_{s'\in\mathscr{S}} e^{\varepsilon U(s',z'^n)/2}}{\sum_{s'\in\mathscr{S}} e^{\varepsilon U(s',z^n)/2}} \\
&\le e^{\varepsilon/2} \cdot \frac{|\mathscr{S}| e^{(\varepsilon/2)\max_{s\in\mathscr{S}} U(s,z^n)}}{|\mathscr{S}| e^{(\varepsilon/2)\min_{c\in\mathscr{S}} U(s,z'^n)}} \\
&\le e^{\varepsilon/2} \cdot \exp\left(\frac{\varepsilon}{2} \cdot \max_{s\in\mathscr{S}} |U(s,z^n)-U(s,z'^n)|\right) \\
&\le e^{\varepsilon}.
\end{aligned}$$

For part 2: for each t, define the set

$$\mathscr{S}_t \triangleq \left\{s \in \mathscr{S} : U(s,z^n) \ge U^*(z^n) - t\right\}.$$

Then

$$\begin{aligned}
P\left[S \in \mathscr{S}_t | Z^n = z^n\right] &= \frac{\sum_{s\in\mathscr{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathscr{S}} e^{\varepsilon U(s,z^n)/2}} \\
&= \frac{\sum_{s\in\mathscr{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathscr{S}_{t/2}} e^{\varepsilon U(s,z^n)/2} + \sum_{s\in\mathscr{S}_{t/2}^c} e^{\varepsilon U(s,z^n)/2}} \\
&\le \frac{\sum_{s\in\mathscr{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathscr{S}_{t/2}^c} e^{\varepsilon U(s,z^n)/2}} \\
&\le e^{(\varepsilon/2)(U^*(z^n)-t)} e^{-(\varepsilon/2)(U^*(z^n)-t/2)} |\mathscr{S}_t| \\
&\le |\mathscr{S}| e^{-\varepsilon t/4}.
\end{aligned}$$

□

Finally, we need the following result, due to Nissim and Stemmer [NS15]:

Lemma 7. *Let the parameters ε, δ be such that $0 < \delta \le \varepsilon \le \frac{1}{5}$ and $m = \frac{\varepsilon}{\delta}$ is an integer. Consider an algorithm B that takes an input $Z^{m\times n}$ and outputs a pair $(F_J, J) \in \mathscr{F} \times \{1,\ldots,m\}$. If B is (ε,δ)-differentially private, then*

$$\mathbb{P}\left[L_n^{(J)}(F_J) \le L(F_J) + 5\varepsilon\right] \ge \varepsilon. \tag{82}$$

Here, for each $j \in \{1,\ldots,m\}$, $L_n^{(j)}(f)$ denotes the empirical loss of $f \in \mathscr{F}$ on the jth row of the matrix $Z^{m\times n}$.

Proof. We first derive a version of this result that holds in expectation. Let $Z'^{m\times n}$ be an independent copy of $Z^{m\times n}$, and let $Z_{(ji)}^{m\times n}$ be obtained from $Z^{m\times n}$ by replacing the sample Z_{ji} in the jth row and the ith column with Z'_{ji}.

Now, we write

$$\begin{aligned}
\mathbb{E}\left[L_n^{(J)}(F_J)\right] &= \sum_{j=1}^{m} \mathbb{E}\left[L_n^{(J)}(F_J)\mathbf{1}\{J=j\}\right] \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \mathbb{E}[\ell(F_J, Z_{ji})\mathbf{1}\{J=j\}] \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z_{ji}) \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}),
\end{aligned} \tag{83}$$

where in the last line we have used the assumption that the entries of $Z^{m\times n}$ and $Z'^{m\times n}$ are i.i.d. draws from the same distribution. Since $P_{(F_J,J)|Z^{m\times n}}$ is (ε,δ)-differentially private, we have

$$\int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) \le e^{\varepsilon}\int P_{(F_J,J)|Z^{m\times n}=z^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) + \delta.$$

Averaging with respect to $Z^{m\times n}$ and $Z'^{m\times n}$ and exploiting independence, we obtain

$$\begin{aligned}
&\int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) \\
&\le e^{\varepsilon}\int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) + \delta \\
&= e^{\varepsilon}\mathbb{E}[\ell(F_J, Z'_{ji})\mathbf{1}\{J=j\}] + \delta \\
&= e^{\varepsilon}\mathbb{E}[L(F_J)\mathbf{1}\{J=j\}] + \delta.
\end{aligned}$$

Substituting this into (83), we obtain

$$\mathbb{E}\left[L_n^{(J)}(F_J)\right] \le e^{\varepsilon}\mathbb{E}[L(F_J)] + m\delta \le \mathbb{E}[L(F_J)] + 2\varepsilon + m\delta \le \mathbb{E}[L(F_J)] + 3\varepsilon, \tag{84}$$

where the second step follows from the inequality $ae^x \le 2x + a$ for $a, x \in [0,1]$. The probability bound (82) follows from Lemma 9 in the Appendix. □

Now we are ready to state and prove the main result of this section:

Theorem 7 (Nissim–Stemmer)**.** *Let the parameters ε, δ be such that $0 < \delta \le \varepsilon \le \frac{1}{10}$ and $m = \frac{\varepsilon}{\delta}$ is an integer. Let $A = P_{F|Z^n}$ be an (ε,δ)-differentially private randomized learning algorithm operaitng on a sample of size $n \ge \frac{4}{\varepsilon^2}\log\frac{8}{\delta}$. Assume that $\varepsilon \ge \delta$. Then, for any loss function $\ell : \mathscr{F}\times \mathsf{Z} \to [0,1]$,*

$$\mathbb{P}\left[|L_n(F) - L(F)| > 13\varepsilon\right] \le \frac{2\delta}{\varepsilon}\log\frac{2}{\varepsilon}. \tag{85}$$

6.1 The proof of Theorem 7

Suppose, to the contrary, that A does not generalize, i.e., that

$$\mathbb{P}\left[L_n(F) - L(F) > 13\varepsilon\right] > \frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}. \tag{86}$$

Draw $m+1$ independent datasets $(Z_{j,1},\dots,Z_{j,n})$, $j\in\{1,\dots,m+1\}$, from P_Z, and form the $(m+1)\times n$ matrix

$$Z^{(m+1)\times n} == \begin{pmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,n} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m,1} & Z_{m,2} & \dots & Z_{m,n} \\ Z_{m+1,1} & Z_{m+1,2} & \dots & Z_{m+1,n} \end{pmatrix}.$$

Think of the first m rows of this matrix as m independent training sets, and of the last row as a separate validation set. For $j\in\{1,\dots,m\}$, let $F_j\in\mathscr{F}$ be the output of an independent copy of A on the jth row of this matrix. Next, for each $f\in\mathscr{F}$, define the empirical losses

$$L_n^{(j)}(f)\triangleq\frac{1}{n}\sum_{i=1}^{n}\ell(f,Z_{j,i}),\qquad j\in\{1,\dots,m+1\}$$

of f on the jth training set and on the validation set. For the random set $\mathscr{S}=\{(F_j,j)\}_{j=1}^m$, define a function $U:\mathscr{S}\times\mathsf{Z}^{(m+1)\times n}\to\mathbb{R}$ by

$$U((F_j,j),z^{(m+1)\times n})\triangleq n\left(L_n^{(j)}(F_j)-L_n^{(m+1)}(F_j)\right) \tag{87}$$

and generate a random pair $(F_I,I)\in\mathscr{S}$ by running the McSherry–Talwar exponential algorithm (80) with this function U on $Z^{(m+1)\times n}$.

For $j\in\{1,\dots,m\}$, denote by E_j the event $\{L_n^{(j)}(F_j)-L(F_j)>13\varepsilon\}$, and let $E=\bigcup_{j=1}^m E_j$. By hypothesis, cf. Eq. (86), $\mathbb{P}[E_j]>\frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}$ for each j. Since the events $E_1,\dots,E_m$ are independent,

$$\mathbb{P}[E^c]=\mathbb{P}\left[\bigcap_{j=1}^m E_j^c\right]=\prod_{j=1}^m\mathbb{P}[E_j^c]\le\left(1-\frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}\right)^m=\left(1-\frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}\right)^{\varepsilon/\delta}\le\frac{\varepsilon}{2}.$$

Next, for each $j\in\{1,\dots,m\}$, let G_j denote the event that $|L_n^{(m+1)}(F_j)-L(F_j)|\le\varepsilon$, and let $G=\bigcap_{j=1}^m G_j$. On the other hand, since $F_1,\dots,F_m$ are independent of the last row of the matrix $Z^{(m+1)\times n}$, Hoeffding's lemma and the union bound guarantee that

$$\mathbb{P}[G]\ge 1-2me^{-2n\varepsilon^2}=1-\frac{2\varepsilon}{\delta}e^{-2n\varepsilon^2}.$$

By the union bound,

$$\begin{aligned}\mathbb{P}[E\cap G]&=1-\mathbb{P}[E^c\cup G^c]\ge 1-(\mathbb{P}[E^c]+\mathbb{P}[G^c])\\&\ge 1-\frac{2\varepsilon}{\delta}e^{-2n\varepsilon^2}-\frac{\varepsilon}{2}.\end{aligned}$$

Consequently, if we choose $n\ge\frac{1}{2\varepsilon^2}\log\frac{8}{\delta}$, then we will have $\mathbb{P}[E\cap G]\ge 1-\frac{3\varepsilon}{4}$.

Now, on the event $E\cap G$, the function U defined in (87) will satisfy

$$\begin{aligned}U^*(Z^{(m+1)\times n})&=\max_{j\in\{1,\dots,m\}}U(j,Z^{(m+1)\times n})\\&=n\max_{j\in\{1,\dots,m\}}\left[L_n^{(j)}(F_j)-L_n^{(m+1)}(F_j)\right]\\&=n\max_{j\in\{1,\dots,m\}}\left[\left(L_n^{(j)}(F_j)-L(F_j)\right)+\left(L(F_j)-L_n^{(m+1)}(F_j)\right)\right]\\&\ge 12n\varepsilon.\end{aligned}$$

Therefore, on the event $E \cap G$, with probability at least $1 - me^{-n\varepsilon^2/4}$, the output (F_I, I) the exponential mechanism with (87) will be such that

$$\begin{aligned} L_n^{(I)}(F_I) - L(F_I) &= L_n^{(I)}(F_I) - L_n^{(m+1)}(F_I) + L_n^{(m+1)}(F_I) - L(F_I) \\ &= \frac{U(I, Z^{(m+1)\times n})}{n} + L_n^{(m+1)}(F_I) - L(F_I) \\ &\geq \frac{U^*(Z^{(m+1)\times n})}{n} - 2\varepsilon > 10\varepsilon. \end{aligned}$$

Thus, if n is also chosen to be larger then $\frac{4}{\varepsilon^2}\log\frac{8}{\delta}$, then the output (F_I, I) will satisfy

$$\mathbb{P}\left[L_n^{(I)}(F_I) \leq L(F_I) + 10\varepsilon\right] \leq \varepsilon. \tag{88}$$

By Lemma 7, this is impossible if we can show that the algorithm $B = P_{(F_J,J)|Z^{(m+1)\times n}}$ is $(2\varepsilon, \delta)$-differentially private.

To see this, we observe that the algorithm $B = P_{(F_J,J)|Z^{(m+1)\times n}}$ is an adaptive composition of A^m and the McSherry–Talwar algorithm. Since A is (ε, δ)-differentially private, so is A^m, and the McSherry–Talwar algorithm is $(\varepsilon, 0)$-differentially private. Therefore, B is $(2\varepsilon, \delta)$-differentially private. Therefore, by Lemma 7, its output must satisfy

$$\mathbb{P}\left[L_n^{(J)}(F_J) \leq L(F_J) + 10\varepsilon\right] \geq 2\varepsilon.$$

which contradicts (88).

A Technical lemmas

Lemma 8. *Let U and V be two random variables, such that $U \leq V$ almost surely. Then*

$$\mathbb{E}|U| \leq |\mathbb{E}U| + 2\mathbb{E}|V|. \tag{89}$$

Proof. We have

$$\mathbb{E}|U| = \mathbb{E}|(V-U) - V| \leq \mathbb{E}|V-U| + \mathbb{E}|V| = \mathbb{E}[V-U] + \mathbb{E}|V| \leq |\mathbb{E}U| + 2\mathbb{E}|V|.$$

□

Lemma 9. *Let U and V be two random variables, such that $0 \leq U, V \leq 1$ almost surely, and*

$$\mathbb{E}[U] \leq \mathbb{E}[V] + 3\varepsilon$$

for some $0 \leq \varepsilon \leq \frac{1}{5}$. Then

$$\mathbb{P}[U \leq V + 5\varepsilon] \geq \varepsilon. \tag{90}$$

Proof. Suppose, by way of contradiction, that $\mathbb{P}[U \leq V + 5\varepsilon] < \varepsilon$. Then

$$\begin{aligned} \mathbb{E}[U] &= \mathbb{E}[U\mathbf{1}\{U - V \leq 5\varepsilon\}] + \mathbb{E}[U\mathbf{1}\{U - V > 5\varepsilon\}] \\ &> \mathbb{E}[(5\varepsilon + V)\mathbf{1}\{U - V > 5\varepsilon\}] \\ &= (5\varepsilon)\,\mathbb{P}[U - V > 5\varepsilon] + \mathbb{E}[V\mathbf{1}\{U - V > 5\varepsilon\}]. \end{aligned}$$

On the other hand, since $0 \le V \le 1$

$$\mathbb{E}[V\mathbf{1}\{U - V \le 5\varepsilon\}] \le \mathbb{E}[\mathbf{1}\{U - V \le 5\varepsilon\}] = \mathbb{P}[U - V \le 5\varepsilon] < \varepsilon.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[U] &> (5\varepsilon)\,\mathbb{P}[U - V > 5\varepsilon] + \mathbb{E}[V] - \varepsilon \\
&> (5\varepsilon)(1 - \varepsilon) + \mathbb{E}[V] - \varepsilon \\
&= \mathbb{E}[V] + 4\varepsilon - 5\varepsilon^2 \\
&\ge \mathbb{E}[V] + 3\varepsilon,
\end{aligned}$$

which contradicts the assumption that $\mathbb{E}[U] \le \mathbb{E}[V] + 3\varepsilon$. □

References

[Dwo06] C. Dwork. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1–12, 2006.

[HRS15] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: stability of stochastic gradient descent. arXiv preprint 1509.01240, September 2015.

[MT07] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103, 2007.

[NS15] K. Nissim and U. Stemmer. On the generalization properties of differential privacy. arXiv preprint 1504.05800, April 2015.

[SSSS10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability, and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.