**The Algorithmic Foundations of Adaptive Data Analysis** November 6, 2017

# Lecture 14: Strong Generalization Bounds

*Lecturer: Adam Smith* *Scribe: Adam Smith*

So far, we have only provided bounds on the *expected* generalization error of a interactive process. Today we will see how, for $\epsilon, \delta$-differential privacy, those bounds can be converted to high probability bounds. These allows us to give confidence intervals around the values estimated by an algorithm.

## 1 The Monitor Technique and High-Probability Bounds

**Theorem 1** *Let $\epsilon \in [\sqrt{\frac{12}{n}}, \frac{1}{5}]$, and $\delta \leq \frac{\epsilon}{16}$. Let $M : \mathcal{X}^n \to \mathcal{Q}_{1/n}$ be $(\epsilon, \delta)$-max-KL stable where $\mathcal{Q}_{1/n}$ is the class of $\frac{1}{n}$-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathcal{D}$ be a distribution on $\mathcal{X}$.*

$$\Pr_{\substack{\mathbf{S}\sim\mathcal{D}^n \\ q\leftarrow M(\mathbf{S})}} (|q(\mathcal{D}) - q(\mathbf{S})| \geq 6\epsilon) = \max\left(\tfrac{4\delta}{\epsilon}, e^{-\epsilon^2 n/8}\right).$$

A consequence of this theorem is that high-probability bounds on *empirical* accuracy (such as we already proved for all the differentially private mechanisms we discussed) translate to high-probability bounds on population accuracy.

**Corollary 2 (Informal)** *For every distribution $\mathcal{D}$, and $\epsilon, \delta$ as in Theorem 1: If $M$ is a mechanism for answering adaptive lienar queries that is $(\epsilon, \epsilon\cdot\delta)$-differentially private and $(\epsilon, \delta)$-*empirically *accurate (for every particular data set), then it is $(O(\epsilon), O(\delta))$-*population *accurate for data sets drawn i.i.d. from $\mathcal{D}$.*

To get some intuition for the proof, suppose that we have a differentially private algorithm which does not satisfy any reasonable high-probability bound—perhaps with probability $p$ it manages to overfit significantly. The proof is to uses a monitor argument (like that from Lectures 7–10) to come up with a new differentially private algorithm that overfits with constant probability. Roughly: If we were to run $T \approx 1/p$ independent copies of the mechanism (on independently selected data sets), then the probability that at least one of them overfits becomes constant. The technical work lies in showing that the monitor can use the exponential mechanism to find which copy overfit differentially privately, and then deriving a contradiction.

One important technical tool is a lemma bounding the expected generalization of an $(\epsilon, \delta)$ algorithm in a setting where the algorithm gets several independently selected data sets as input, and can choose to overfit to any of them.

**Lemma 3** *For every $\epsilon > 0, \delta > 0, T \in \mathbb{N}$, and distribution $\mathcal{D}$ on $\mathcal{X}$: If $\mathcal{W} : (\mathcal{X}^n)^T \to \{1, ..., T\} \times \mathcal{Q}_{1/n}$ is $(\epsilon, \delta)$-differentially private, and we select $\overrightarrow{\mathbf{S}} = (\mathbf{S}^{(1)}, ..., \mathbf{S}^{(T)}) \sim (\mathcal{D}^n)^T$, then*

$$\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T \\ (t,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(q(\mathbf{S}^{(t)}) - q(\mathcal{D})\right) \leq (e^\epsilon - 1) + T\delta\,.$$

We first use the lemma to prove our main theorem.

**Proof** [of Theorem 1] Given and algorithm $M$ and distrbution $\mathcal{D}$ and an integer $T$ (to be fixed later), consider a monitor algorithm $\mathcal{W}$ that takes $T$ data sets, each of size $n$, and runs $T$ independent copies of $M$. See Algorithm 1

First, observe that if $M$ is $(\epsilon, \delta)$-differentially private, then $M$ is $(\epsilon + \epsilon', \delta)$-differentially private—regardless of how large $T$ is—because each data point affects only one instance of $M$.

Let $p_\alpha$ be the probability that an execution of $M$ on a data set $\mathbf{S} \sim \mathcal{D}^n$ outputs a query with score at least $\alpha$. (Recall that our goal is to get an upper bound on $p_\alpha$ when $\alpha = 6\epsilon$.) Then the probability that

**Algorithm 1:** $\mathcal{W}$

**Input:** $\overrightarrow{\mathbf{s}} = (\mathbf{s}^{(1)}, ..., \mathbf{s}^{(T)}) \in (\mathcal{X}^n)^T$

Fixed parameters: $\epsilon, \epsilon' > 0$, $T \in \mathbb{N}$, mechanism $M : \mathcal{X}^n \to \mathcal{Q}_{1/n}$, distribution $\mathcal{D}$ on $\mathcal{X}$

1 **for** $t = 1$ *to* $T$ **do**

2 $\quad q_t \leftarrow M(\mathbf{s}^{(t)})$;

3 $F \leftarrow \bigcup_{t=1}^{T} \{(t, q_t), (t, 1 - q_t)\}$, define $score : F \to \mathbb{R}$ via $\mathsf{score}(t, q) = q(\mathbf{s}^{(t)}) - q(\mathcal{D})$ ;

4 Sample $(t^*, q^*)$ from $F$ using the exponential mechanism with parameter $\epsilon'$

5 (that is, sample $(q, t) \in F$ with probability $\propto \exp(\frac{n\epsilon'}{2}(q(\mathbf{s}^{(t)}) - q(\mathcal{D})))$;

6 **return** $(t^*, q^*)$.

at least one of the $T$ executions run by $\mathcal{W}$ outputs such a query is $1 - (1 - p_\alpha)^T$. Furthermore, there is always an item with nonnegative score (since we include every query with both positive and negative signs). Thus, if we let $S_{max}$ denote the maximum score among the pairs in $F$ (at Line 3), we can lower bound its expectation

$$\mathbb{E}(S_{max}) \geq \alpha(1 - (1 - p_\alpha)^T) + 0 \cdot (1 - p_\alpha)^T .$$

Next, observe that if any of the copies of $M$ outputs a query that significantly overfits to its data set, then $\mathcal{W}$ is likely to produce a query that significantly overfits. Specifically, conditioned on any particular value $S_{max}$ of the maximum score among items in $F$ at line 3, the exponential mechanism will select a pair $(t^*, q^*)$ with expected score at least $S_{max} - \frac{2\ln(T)}{\epsilon' n}$. Thus, the expected score of the final output of $\mathcal{W}$ satisfies

$$\mathbb{E}\left(q^*(\mathbf{S}^{(t^*)}) - q^*(\mathcal{D})\right) \geq \mathbb{E}\left(S_{max} - \frac{2\ln(T)}{\epsilon' n}\right) \geq \alpha(1 - (1 - p_\alpha)^T) - \frac{2\ln(T)}{\epsilon' n} .$$

Finally, we can apply Lemma 3—the generalization of the basic stability lemma to a setting where $T$ data sets are available—to bound the expected score of the $\mathcal{W}$'s output, and conclude that

$$\alpha(1 - (1 - p_\alpha)^T) - \frac{2\ln(T)}{\epsilon' n} \leq (e^{\epsilon+\epsilon'} - 1) + T\delta . \tag{1}$$

We could now optimize over the choice over $T$ and $\epsilon'$ to get the best possible bound, but the result would be unwieldly. Here, we extract a special case. Let $T = \lfloor 1/p_\alpha \rfloor$ (so that $1 \geq Tp_\alpha \geq 1 - p_\alpha$). In that case, $(1 - p_\alpha)^T \leq e^{-p_\alpha T} \leq e^{-1+p_\alpha} \leq 1/2$ for $p_\alpha \leq 1/4$. This allows us to write

$$\alpha - 2(e^{2\epsilon} - 1) \leq \frac{4\ln(T)}{\epsilon n} + 2T\delta \leq \frac{4\ln(1/p_\alpha)}{\epsilon n} + \frac{2\delta}{p_\alpha} .$$

Now take $\alpha = 6\epsilon$, and $\epsilon \leq \frac{1}{5}$. In that case, $\alpha - 2(2^\epsilon - 1) \geq \epsilon$. One of the two terms on the right-hand side must therefore be at least $\epsilon/2$, from which we conclude $p_\alpha \leq \max(\frac{4\delta}{\epsilon}, e^{-\epsilon^2 n/8})$. This bound is indeed below $1/4$ when $\delta \leq \epsilon/16$ and $\epsilon \geq \sqrt{12/n}$. ■

We can now turn to the proof of the main technical lemma.

**Proof** [of Lemma 3] For simplicity, we will prove this lemma only for the case of statistical queries (given by a function $q : \mathcal{X} \to [0, 1]$). The case of general $1/n$-sensitive queries is similar but involves an additional step.

The idea is to write the expectation that we want to bound as the sum of $T$ different expectations, one for each choice of $T^*$. Mathematically, we can do this by multiplying by the indicator function $\mathbf{1}_{\{t^*=t\}}$.

$$\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}} \sim (\mathcal{D}^n)^T \\ (t^*, q) \leftarrow \mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(q(\mathbf{S}^{(t)})\right) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}} \sim (\mathcal{D}^n)^T \\ (t^*, q) \leftarrow \mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(\mathbf{1}_{\{t^*=t\}} \cdot q(\mathbf{S}^{(t)})\right)$$

The straightforward approach to bounding this sum (separately for each term) would give a bound of about $T(\epsilon + \delta)$ on the overall generalization error. However, the bound we are aiming for has the form $\epsilon + T\delta$. To obtain it, we need to take advantage of the fact that most of the terms in this sum are small—they are nonnegative and add to at most 1.

We'll need the following simple claim:

**Claim 4** *Let $X, Y$ be distributions on a set $\mathcal{Y}$ such that $X \approx_{\epsilon,\delta} Y$, and let $f : \mathcal{Y} \to [0,1]$ be a bounded real-valued function. Then*

$$\mathbb{E}(f(X)) \leq e^{\epsilon}\,\mathbb{E}(f(Y)) + \delta.$$

**Proof** Since $f$ is nonnegative, $\mathbb{E}(f(X)) = \int_{z=0}^{1} \Pr(f(X) \leq z)dz \leq \int_{z=0}^{1} \big(e^{\epsilon}\Pr(f(Y) \leq z) + \delta\big)dz = e^{\epsilon}\,\mathbb{E}(f(Y)) + \delta$. ■

Using the argument from Lecture 8, we can imagine replacing one sample from $\mathbf{S}^{(t)}$ (where $t$ depends which term in the sum we are bounding) with a fresh sample from $\mathcal{D}$. Given this fresh sample $x'$, we use $\overrightarrow{\mathbf{S}}_{t,i\to x'}$ to denote the new data set. For each $t$, we have

$$\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T \\ (t^*,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(\mathbf{1}_{\{t^*=t\}}\cdot q(\mathbf{S}^{(t)})\right) \leq e^{\epsilon}\left(\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T, X'\sim\mathcal{D} \\ (t^*,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}_{t,i\to X'}\right)}} \left(\mathbf{1}_{\{t^*=t\}}\cdot q(\mathbf{S}^{(t)})\right)\right) + \delta$$

As before, we can observe that this is the same *in expectation* as evaluating $q$ on a random point $X_i^{(t)}$ in $\mathbf{S}^{(t)}$, and—this was the slick part—we can swap $X'$ and $X_i^{(t)}$. We get

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T \\ (t^*,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(\mathbf{1}_{\{t^*=t\}}\cdot q(\mathbf{S}^{(t)})\right) &\leq e^{\epsilon}\left(\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T, X'\sim\mathcal{D} \\ (t^*,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(\mathbf{1}_{\{t^*=t\}}\cdot q(X')\right)\right) + \delta\,. \\ &= e^{\epsilon}\left(\mathop{\mathbb{E}}_{\substack{\overrightarrow{\mathbf{S}}\sim(\mathcal{D}^n)^T \\ (t^*,q)\leftarrow\mathcal{W}\left(\overrightarrow{\mathbf{S}}\right)}} \left(\mathbf{1}_{\{t^*=t\}}\cdot q(\mathcal{D})\right)\right) + \delta\,.\end{aligned}$$

Summing over all the values of $t$, we get $e^{\epsilon}\,\mathbb{E}(q(\mathcal{D})) + T\delta$, as desired. ■

## 2 Notes

The generalization bound given here appears in [Nis15, BNS+16]. This type of argument can also be used to prove concentration bounds with no apparent connection to data privacy [NS17].

## References


[BNS+16] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1046–1059. ACM, 2016.

[Nis15] On the generalization properties of differential privacy. *CoRR*, abs/1504.05800, 2015. Withdrawn.

[NS17] Kobbi Nissim and Uri Stemmer. Concentration bounds for high sensitivity functions through differential privacy. *CoRR*, abs/1703.01970, 2017.