The Algorithmic Foundations of Adaptive Data Analysis December, 2017

Final Project

Authors: Thomas Steinke, Jonathan Ullman *Scribe: Akilesh Tangella*

Subgaussian Tail Bounds via Stability Arguments

In these notes, we explain how arguments similar to those made in lecture 14 to derive high probability generalization bounds for differentially private mechanisms in adaptive data analysis can be used to derive subgaussian tail bounds.

1 Tail Bounds and Connection to Adaptive Data Analysis

The most common type of tail bounds in theoretical computer science show that sums of independent and bounded random variables have very low probability of deviating significantly from their mean (or in other words, have subgaussian tails). One such famous inequality we will discuss in these notes is a special case of Bernstein's inequality.

Theorem 1 (Tail Bound) *If $X_1, X_2, \dots, X_n$ are independent random variables in the interval $[0,1]$ and $\mu_i = \mathbb{E}[X_i]$ for all $1 \le i \le n$. Then:*

$$\forall \epsilon > 0, \Pr\left[\sum_{i=1}^{n} (X_i - \mu_i) \ge \epsilon n\right] \le e^{-\Omega(\epsilon^2 n)}$$

How does this relate to the theorems we have seen for adaptive data analysis? To see, we use an informal and convenient version of the high probability bound obtained in lecture 14.

Theorem 2 (High Probability Generalization Bound from [NS17]) *A function $f : \mathcal{X}^n \to \mathbb{R}$ has sensitivity λ if $|f(S) - f(S')| \le \lambda$ for every pair $S, S' \in \mathcal{X}^n$ differing in only one entry. Define $f(\mathcal{D}^n)$ as $\mathbb{E}_{S\sim\mathcal{D}^n}[f(S)]$. Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{F}_\lambda$ be (ϵ, δ)-differentially private where $\mathcal{F}_\lambda$ is the class of λ-sensitive functions. If $n \ge \frac{1}{\epsilon^2}\log(\frac{4\epsilon}{\delta})$, then for all distributions $\mathcal{D}$ on $\mathcal{X}$:*

$$\Pr_{\substack{S\sim\mathcal{D}^n \\ f\leftarrow\mathcal{A}(S)}}\left[|f(S) - f(\mathcal{D}^n)| \ge 18\epsilon\lambda n\right] < \frac{\delta}{\epsilon}$$

Consider the setting in which the X_i's are identically distributed according to some distribution $\mathcal{D}$ on $\mathcal{X} = [0,1]$. Then, we can view the n-tuple $(X_1, X_2, \dots, X_n)$ as drawing a dataset of n points from $\mathcal{D}^n$. Let f be the function which takes in such a dataset and outputs their sum. From now on, we refer to such a query as the SUM query. Let $\mathcal{A}$ be the function which outputs f on every input. Clearly $\mathcal{A}$ is differentially private for any value of ϵ and δ because it is a constant function. Also, f has sensitivity 1 because removing or adding an element to the dataset can change the sum of its entries by at most 1, since $\mathcal{X} = [0,1]$. Thus, we can apply theorem 2 to get:

$$\Pr_{(X_1,X_2,\dots,X_n)\sim\mathcal{D}^n}\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \ge 18\epsilon n\right] < \frac{\delta}{\epsilon} = 2^{-\Omega(\epsilon^2 n)}$$

The last equality comes from plugging in $n = \frac{1}{\epsilon^2}\log(\frac{4\epsilon}{\delta})$. This looks very similar to the tail bound in theorem 1! Given the intricate link between theorems 1 and 2, it is no surprise that the techniques used to prove theorem 2 can also be used to prove theorem 1. Thus, in the rest of these notes we detail such a proof for theorem 1 in the more general case in which the X_i's are not necessarily identically distributed.

2 Tail Bounds via Proxies

Recall how we proved the Chernoff-like bound in lecture 1. Instead of directly dealing with a probability statement, we dealt with the expectation of a proxy random variable (the moment generating function). We will do the same thing here, except with a different proxy. Let $Y = \sum_{i=1}^{n}(X_i - \mu_i)$. Let $Y^1, Y^2, \ldots, Y^m$ be m independent copies of Y. Then, the proxy we will use is $\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}$. The following lemma suggests why such an approach should work.

Lemma 3 *Let Y be a random variable and let $Y^1, Y^2, \ldots, Y^m$ be independent copies of Y. Then:*

$$\Pr\left[Y \geq 2\,\mathbb{E}\left[max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right]\right] \leq \frac{\ln(2)}{m}$$

Proof Let $y = 2\,\mathbb{E}\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right]$ and let $\psi = \Pr[Y \geq y]$. We want to show $\psi \leq \frac{\ln(2)}{m}$. Suppose for the sake of contradiction that this is not the case. On one hand, by Markov's inequality:

$$\Pr\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\} \geq y\right] \leq \frac{\mathbb{E}\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right]}{y} = \frac{1}{2}$$

On the other hand, we have:

$$\begin{aligned}
\Pr\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\} \geq y\right] &= 1 - \Pr\left[\forall j \in [m],\ Y^j < y\right] \\
&= 1 - \Pr[Y < y]^m \\
&= 1 - (1-\psi)^m \\
&> 1 - e^{-\psi m} \\
&> 1 - e^{-m\cdot\frac{\ln(2)}{m}} = 1 - e^{-\ln(2)} = \frac{1}{2}
\end{aligned}$$

The first inequality arises from the fact that $(1 - x) < e^{-x}$ for positive x and the second inequality comes from our assumption that $\psi > \frac{\ln(2)}{m}$. Overall, we have $\frac{1}{2} \geq \Pr\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\} \geq y\right] > \frac{1}{2}$, a contradiction! So $\psi \leq \frac{\ln(2)}{m}$, as desired. ■

If we can upper bound $\mathbb{E}\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right]$ and tactically set m, then it seems possible to recover theorem 1 from lemma 3. This is our remaining mission.

3 Bounding the Proxy via a Stability Argument

Before delving into the proof, let us draw another parallel to lecture 14. Recall lemma 3 in lecture 14 states that if a differentially private algorithm is given multiple datasets and allowed to choose one of these datasets and a query (with bounded sensitivity) to evaluate this dataset on, it cannot significantly overfit in expectation. We can view Y as the generalization error of the SUM query, where datasets are drawn from $(X_1, X_2, \ldots, X_n)$ (we think of this as a distribution on n-tuples). Then evaluating $\text{argmax}_{j\in[m+1]}\left\{Y^j\right\}$ (let $Y^{m+1} = 0$ and the dataset corresponding to Y^{m+1} be the dataset of the true means) is equivalent to the following process: an algorithm is given multiple datasets each drawn from the aforementioned distribution on n-tuples and has to choose one of them to maximize the generalization error of the SUM query. We do not have tools to bound the expected generalization of the algorithm's output in such a setting, but if our adversary is differentially private then this is similar to lemma 3 from lecture 14. Thus, the overall idea is to use a differentially private algorithm with useful accuracy guarantees to choose a dataset, such that in expectation, the resulting generalization error closely approximates the maximum generalization error. Via techniques similar to those in lecture 14, we can bound the expected generalization error of a differentially private algorithm. Combining these two steps allows us to bound the expected maximum generalization error, as desired.

Note a few discrepancies in our analogy. Firstly, our differentially private algorithm will not output queries

like the setting in lemma 3. But this is because our query is not adaptively chosen based on the datasets, but rather fixed beforehand as the SUM query. In other words, we could just output the SUM query on every input (along with the selected dataset) and keep the format of our output similar to lemma 3's, but this is useless since a constant output will not compromise stability. Secondly, in our setting, the n points in each dataset do not necessarily come from the same distribution. This difference also arose in section 1 and is the main reason why the proofs from lecture 14 cannot be repeated verbatim.

For convenience, we will think of each of our datasets as column vectors, and the conglomeration of these column vectors as forming a matrix. We begin with some notation and prove a general lemma about random matrices. Let $\mathbf{X}$ be a random $n \times m$ matrix with entries on the interval $[0,1]$. Let X_i^j be the random variable corresponding to the entry in the i^{th} row and j^{th} column of $\mathbf{X}$. Let $\mathbf{x}$, $\mathbf{x}_i$, and x_i^j be the realizations of $\mathbf{X}$, its i^{th} row, and its $(i,j)^{th}$ entry, respectively.

Lemma 4 *Let $\boldsymbol{X}$ be a random $n \times m$ matrix with entries in the interval $[0,1]$ and independent rows. Then:*

$$\forall \eta > 0, \mathbb{E}\left[\max_{j\in[m]} \sum_{i=1}^{n} X_i^j\right] \leq e^{\eta} \max_{j\in[m]} \mathbb{E}\left[\sum_{i=1}^{n} X_i^j\right] + \frac{2\ln(m)}{\eta}$$

Proof We consider the following procedure, $S_\eta : [0,1]^{n\times m} \longrightarrow [m]$, for selecting a column j:

$$\Pr\left[S_\eta(\mathbf{x}) = j\right] = \frac{\exp\left(\frac{\eta}{2} \cdot \sum_{i=1}^{n} x_i^j\right)}{\sum_{k=1}^{m} \exp\left(\frac{\eta}{2} \cdot \sum_{i=1}^{n} x_i^k\right)}$$

Using the lingo from lecture 11, this is just an instantiation of the exponential mechanism with privacy parameter η and score sensitivity 1. But why is the score sensitivity 1? To see this, we must first define our notion of neighboring databases (which in our case are really matrices).

Definition 5 *Two matrices are neighboring if they differ in exactly one row. We write $(\boldsymbol{x}_{-i}, \tilde{\boldsymbol{x}}) \in [0,1]^{n\times m}$ to be the matrix $\boldsymbol{x} \in [0,1]^{n\times m}$ with its i^{th} row replaced by $\tilde{\boldsymbol{x}} \in [0,1]^m$.*

Note that the outputs of our instantiation of the exponential mechanism are columns and the scores are the sums of the elements of the columns. When we replace a row of the matrix, at most one entry can change in any given column. This means the sum of the elements in any column changes by the difference of the new entry and old entry, but this difference is at most 1 since the elements of the matrix are in the interval $[0,1]$. Thus, the familiar theorems for the stability (differential privacy) and accuracy of the exponential mechanism must hold.

Claim 6 (Stability of Exponential Mechanism) *For every $\boldsymbol{x} \in [0,1]^{n\times m}$, $\tilde{\boldsymbol{x}} \in [0,1]^m$, $i \in [n]$, and $j \in [m]$:*

$$e^{-\eta} \Pr\left[S_\eta(\boldsymbol{x}) = j\right] \leq \Pr\left[S_\eta(\boldsymbol{x}_{-i}, \tilde{\boldsymbol{x}}) = j\right] \leq e^{\eta} \Pr\left[S_\eta(\boldsymbol{x}) = j\right]$$

We state a accuracy theorem for the exponential mechanism similar to theorem 6 from lecture 11. It is exactly the same bound as what we used in lecture 14.

Claim 7 (Accuracy of Exponential Mechanism) *For every $\boldsymbol{x} \in [0,1]^{n\times m}$:*

$$\mathbb{E}_{S_\eta}\left[\sum_{i=1}^{n} x_i^{S_\eta(\boldsymbol{x})}\right] \geq \max_{j\in[m]} \left(\sum_{i=1}^{n} x_i^j\right) - \frac{2\ln(m)}{\eta}$$

Note that the accuracy guarantee of the exponential mechanism is for any fixed matrix, but our matrix itself is randomized. The following claim shows that in expectation over the randomization of the matrix and the exponential mechanism, the sum of the entries of the column that the exponential mechanism outputs is not too much smaller than the sum of the entries of any fixed column.

Claim 8 *Let $\nu = \max_{j\in[m]} \mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{n} X_i^j\right]$ (for notational convenience). Then:*

$$\mathbb{E}_{\boldsymbol{X}, S_\eta}\left[\sum_{i=1}^{n} X_i^{S_\eta(\boldsymbol{X})}\right] \leq e^{\eta}\nu$$

Proof Let $\tilde{\mathbf{X}}$ be an independent and identical copy of $\mathbf{X}$. Since the rows in each matrix are independent and $\mathbf{X}_i$ and $\tilde{\mathbf{X}}_i$ are identically distributed, we have that $\left(\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i\right)$ is distributed identically to $\mathbf{X}$. Since $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are identically distributed, X_i^j and $\tilde{X}_i^j$ are identically distributed. Also, X_i^j is independent of $\left(\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i\right)$ since $\mathbf{X}_i$ is independent of $\mathbf{X}_{-i}$ and $\tilde{\mathbf{X}}_i$. Since $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are independent, $\tilde{X}_i^j$ is independent of $\mathbf{X}$. Putting this altogether, we have that the pair $\left(\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i\right)$ and X_i^j is distributed the same as the pair $\mathbf{X}$ and $\tilde{X}_i^j$. This implies the pair $\mathcal{S}_\eta\left(\left(\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i\right)\right)$ and X_i^j is distributed identically to the pair $\mathcal{S}_\eta(\mathbf{X})$ and $\tilde{X}_i^j$. We have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X},\mathcal{S}_\eta}\left[\sum_{i=1}^{n} X_i^{\mathcal{S}_\eta}\right] &= \mathbb{E}_{\mathbf{X}}\left[\sum_{j=1}^{m}\sum_{i=1}^{n} \Pr\left[\mathcal{S}_\eta(\mathbf{X}) = j\right] X_i^j\right] \\
&\leq \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[\sum_{j=1}^{m}\sum_{i=1}^{n} e^\eta \Pr\left[\mathcal{S}_\eta\left(\left(\mathbf{X}_{-i}, \tilde{\mathbf{X}}_i\right)\right) = j\right] X_i^j\right] \\
&= \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[\sum_{j=1}^{m}\sum_{i=1}^{n} e^\eta \Pr\left[\mathcal{S}_\eta(\mathbf{X}) = j\right] \tilde{X}_i^j\right] \\
= \sum_{j=1}^{m} \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[e^\eta \Pr\left[\mathcal{S}_\eta(\mathbf{X}) = j\right] \sum_{i=1}^{n} \tilde{X}_i^j\right] &= \sum_{j=1}^{m} \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[e^\eta \Pr\left[\mathcal{S}_\eta(\mathbf{X}) = j\right]\right] \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[\sum_{i=1}^{n} \tilde{X}_i^j\right] \\
&\leq \mathbb{E}_{\mathbf{X},\tilde{\mathbf{X}}}\left[e^\eta \sum_{j=1}^{m} \Pr\left[\mathcal{S}_\eta(\mathbf{X}) = j\right]\right] \nu = e^\eta \nu
\end{aligned}$$

The first line follows from the definition of expected value, the second from the stability of the exponential mechanism, the third from the independence of the pairs of random variables discussed above, the fourth from the linearity of expectation and independence of $\mathbf{X}$ and $\tilde{\mathbf{X}}$, and the fifth from the definition of ν and linearity of expectation. ■

Taking the expected value with respect to $\mathbf{X}$ on both sides of claim 7, and combining with claim 8, we obtain:

$$\mathbb{E}_{\mathbf{X}}\left[\max_{j\in[m]} X_i^j - \frac{2\ln(m)}{\eta}\right] \leq \mathbb{E}_{\mathbf{X},\mathcal{S}_\eta}\left[X_i^{\mathcal{S}_\eta(\mathbf{X})}\right] \leq e^\eta \nu = e^\eta \max_{j\in[m]} \mathbb{E}_{\mathbf{X}}\left[\sum_{i=1}^{n} X_i^j\right]$$

Rearranging yields lemma 4. ■

Lemma 9 (Proxy Bound) *Let $X_1, X_2, \ldots, X_n$ be independent random variables in the interval $[0,1]$ and let $\mu_i = \mathbb{E}[X_i]$. Let $Y = \sum_{i=1}^{n}(X_i - \mu_i)$. Fix $m \in \mathbb{N}$ and let $Y^1, Y^2, \ldots, Y^m$ be m independent copies of Y. Then:*

$$\mathbb{E}\left[max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right] \leq 4\sqrt{n \cdot \ln(m+1)}$$

Proof If $m \geq e^n - 1$, then $\ln(m+1) \geq n$, so $4\sqrt{n \cdot \ln(m+1)} \geq 4n$, but the maximum of Y is n since each $X_i \in [0,1]$ so the lemma is trivial in this case. Thus, we assume $m < e^n - 1$ from now on. Let $\mu = \sum_{i=1}^{n} \mu_i$. For every $i \in [n]$ let $X_i^1, X_i^2, \ldots, X_i^m$ be m independent copies of X_i, such that $Y^j = \sum_{i=1}^{n}\left(X_i^j - \mu_i\right)$. Let $X_i^{m+1} = \mu_i$ for each $i \in [n]$. Consider the $n \times (m+1)$ random matrix $\mathbf{X}$ with (i,j) entry X_i^j. Since the X_i's are independent the rows of $\mathbf{X}$ are independent, so we may apply lemma 4 to get:

$$\forall \eta > 0, \mathbb{E}\left[\max_{j\in[m+1]} \sum_{i=1}^{n} X_i^j\right] \leq e^\eta \max_{j\in[m+1]} \mathbb{E}\left[\sum_{i=1}^{n} X_i^j\right] + \frac{2\ln(m+1)}{\eta}$$

$\mathbb{E}\left[\sum_{i=1}^{n} X_i^j\right]$ is just $\sum_{i=1}^{n} \mu_i = \mu$ for any $j \in [m+1]$. Also, $Y^j = \sum_{i=1}^{n}\left(X_i^j - \mu_i\right) = \sum_{i=1}^{n} X_i^j - \mu \implies Y^j + \mu = \sum_{i=1}^{n} X_i^j$ for each $j \in [m]$. Since, $X_i^{m+1} = \mu_i$, we have $\sum_{i=1}^{n} X_i^{m+1} = 0 + \mu$. Substituting this in, we get:

$$\forall \eta > 0, \mathbb{E}\left[\max\{Y^1 + \mu, Y^2 + \mu, \ldots, 0 + \mu\}\right] \leq e^\eta \mu + \frac{2\ln(m+1)}{\eta}$$

Since μ is a constant we can bring it out and subtract it from both sides to get:

$$\forall \eta > 0, \mathbb{E}\left[\max\{Y^1, Y^2, \ldots, 0\}\right] \leq (e^\eta - 1)\mu + \frac{2\ln(m+1)}{\eta}$$

Since for $1 \leq i \leq n$ each $\mu_i \leq 1$, $\mu \leq n$. Consider $e^\eta - 1$. It is convex. At $\eta = 0$ it is 0 and at $\eta = 1$ it is $e - 1$. The linear interpolation of these two points is $(e-1)\eta$ which must lie above the curve for $\eta \in (0,1)$. Thus, we have $2\eta \geq (e-1)\eta \geq e^\eta - 1$ for all $\eta \in (0,1)$. Applying these bounds yields:

$$\forall \eta \in (0,1), \mathbb{E}\left[\max\{Y^1, Y^2, \ldots, 0\}\right] \leq 2\eta n + \frac{2\ln(m+1)}{\eta}$$

Since we are only dealing with the case where $m < e^n - 1$, we can set $\eta = \sqrt{\frac{\ln(m+1)}{n}}$, which exactly gives the proxy bound. ■

Theorem 10 (Theorem 1 Revisited with Constants) *If $X_1, X_2, \ldots, X_n$ are independent random variables in the interval $[0,1]$ and $\mu_i = \mathbb{E}[X_i]$ for all $1 \leq i \leq n$. Then:*

$$\Pr\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq \epsilon n\right] \leq e^{1-\frac{\epsilon^2 n}{64}}$$

Proof Set $m = \left\lfloor e^{\frac{\epsilon^2 n}{64}} - 1 \right\rfloor$. Then $8\sqrt{n\ln(m+1)} \leq \epsilon n$. Combining this fact with the proxy bound (lemma 9) and lemma 3, we get:

$$\Pr[Y \geq \epsilon n] \leq \Pr\left[Y \geq 8\sqrt{n\ln(m+1)}\right] \leq \Pr\left[Y \geq 2\,\mathbb{E}\left[\max\left\{0, Y^1, Y^2, \ldots, Y^m\right\}\right]\right] \leq \frac{\ln(2)}{m} \leq \frac{\ln(2)}{e^{\frac{\epsilon^2 n}{64}} - 2}$$

Thus, we have:

$$\Pr[Y \geq \epsilon n] \leq \min\left\{1, \frac{\ln(2)}{e^{\frac{\epsilon^2 n}{64}} - 2}\right\}$$

If the right hand side of the above equation is 1, then:

$$1 \leq \frac{\ln(2)}{e^{\frac{\epsilon^2 n}{64}} - 2} \implies e^{\frac{\epsilon^2 n}{64}} - 2 \leq \ln(2) \implies 1 \leq (2 + \ln(2))e^{-\frac{\epsilon^2 n}{64}}$$

So we get

$$\Pr[Y \geq \epsilon n] \leq 1 \leq (2 + \ln(2))e^{-\frac{\epsilon^2 n}{64}}$$

A similar argument shows that when the right hand side is $\frac{\ln(2)}{e^{\frac{\epsilon^2 n}{64}} - 2}$, we get the same bound, so overall, we have:

$$\Pr[Y \geq \epsilon n] \leq (2 + \ln(2))e^{-\frac{\epsilon^2 n}{64}} \leq e^{1-\frac{\epsilon^2 n}{64}}$$

■

The proof of the proxy bound, particularly lemma 4, may seem rather magical. So, based on the analogies we drew to lecture 14 at the beginning of this section, let us recap what happened. To prove the proxy bound, we realized two things had to be done:

- We had to use a differentially private algorithm to select a dataset which closely approximates the expected maximum generalization error. This was the purpose of claim 7.
- We had to exploit the stability of our selection procedure to derive a bound on its expected generalization error. This was the purpose of claim 8.

Combining the two steps above gave lemma 4, which with a bit of additional work resulted in the proxy bound. At this point, one last thing might be irksome. Our instantiation of the exponential mechanism selected columns from a matrix whose entry in each row i was an independent copy of X_i. The score for a column was the sum of its entries, or equivalently, the sum of the X_i's for $1 \leq i \leq n$. This is not the generalization error (Y), however, so the exponential mechanism is not giving accuracy guarantees with respect to the maximum generalization error. But realizing that $Y = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i$ and that $\mu = \sum_{i=1}^{n} \mu_i$ is a constant implies that the column with the maximum sum of the X_i's is the column with the maximum generalization error, and so the behavior of the exponential mechanism is exactly what we want. This transformation between sums of X_i's and Y via addition/subtraction of μ is seen in lemma 9.

References

[NS17] Kobi Nissim and Uri Stemmer. Concentration bounds for high sensitivity functions through differential privacy. 2017.

[RS17] Aaron Roth and Adam Smith. Lecture notes in adaptive data analysis. 2017.

[SU17] Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. 2017.