EN.601.774 Theory of Replicable ML Spring 2025

Lecture 16

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

Last time, we showed how to extend our "in expectation" generalization guarantees for a single TV-stable or (ϵ, δ)-private mechanism to "in expectation" generalization guarantees for the adaptive sequential composition of k such mechanisms. We showed that answering queries using the Laplace mechanism with noise rate $\varepsilon \in \tilde{O}\left(\frac{\tau}{\sqrt{km}}\right)$ would the guarantee us expected generalization error $O(\tau)$.

Over the next few lectures, we'll build on and strengthen these results in two ways:

1. We'll show high probability generalization bounds for (ε, δ)-DP algorithms
2. We'll show that the noise rate required by stable mechanisms to obtain those high probability generalization bounds are not so large that the generalization guarantees are vacuous

We'll start with the high probability generalization bounds. Recall that we tried to get high probability guarantees for TV-stability the same way we got them for UCO stability. We ran into challenges with applying McDiarmid's inequality, however, because we only had stability guarantees for the expected generalization error $\mathbb{E}_r[q_{S;r}(S) - q_{S;r}(D)]$, not $q_{S;r}(S) - q_{S;r}(D)$.

To get high probability guarantees, we're going to follow a different proof approach: proof by contradiction. Suppose there was an (ε, δ)-DP mechanism $\mathcal{M}$ that *did* overfit its data with probability p (over the choice of sample and choice of internal randomness). We'll design a new mechanism $\mathcal{M}_B$ that "monitors" roughly $T \approx 1/p$ independent runs of $\mathcal{M}$ on independent datasets and outputs the result with the worst generalization error. We'll design the monitor $\mathcal{M}_B$ to also be differentially private, and show that it must overfit its data with constant probability. This will contradict the "in expectation" generalization guarantees we've already proved.

Today we'll build up some of the machinery we'll need to construct the monitor $\mathcal{M}_B$. This monitor will need to privately select the worst query output by several runs on $\mathcal{M}$, or something comparably bad. We'll show how to do this privately via the *exponential mechanism*.

The exponential mechanism takes as input

- A dataset $S \in \mathcal{X}^m$
- A set of candidate outputs $\mathcal{Y}$
- A score function $f : \mathcal{Y} \times \mathcal{X}^m \to \mathcal{R}$ which scores how good each candidate output is with respect to dataset S

- A sensitivity bound δ for f on neighboring datasets that is worst-case over $y \in \mathcal{Y}$. That is, it must hold that

$$\sup_{y \in \mathcal{Y}} \sup_{S, S' \in \mathcal{X}^m} |f(y, S) - f(y, S')| \leq \Delta$$

- Privacy parameter ϵ

Algorithm 1 Exponential Mechanism $\mathcal{E}(S, \mathcal{Y}, f, \Delta, \varepsilon)$

1: Define distribution $D_Y(y) \propto e^{\frac{\varepsilon}{2\Delta} f(y,S)}$
2: **return** $y \sim D_Y$

This is somewhat underspecified, as algorithms go. How do we sample from D_Y? How do we actually *define* D_Y? Note that, so long as there exists a measure such that $\int_{y \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} f(y,S)} dy$ is finite, we're all set! Then we can set

$$D_Y(y) = \frac{e^{\frac{\varepsilon}{2\Delta} f(y,S)}}{\int_{y' \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} f(y',S)} dy'}$$

and we have a well-defined distribution. Sampling from it might still be a challenge though. Let's imagine the case where $\mathcal{Y}$ is finite. Then our normalization factor is

$$\sum_{y \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} f(y,S)}$$

and we can use the following rejection sampling procedure to sample from D_Y.

Algorithm 2 Rejection Sampler Samp$(S, \mathcal{Y}, f, \Delta, \varepsilon)$

1: **while** True **do**
2: Select an element $y \sim Unif(\mathcal{Y})$ uniformly at random
3: Compute $D_Y(y) = \frac{e^{\frac{\varepsilon}{2\Delta} f(y,S)}}{\sum_{y' \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} f(y',S)}}$
4: Sample a threshold $t \sim Unif([0, 1])$
5: **if** $D_Y(y) \geq t$ **then**
6: **return** y
7: **end if**
8: **end while**

Theorem 0.1. *[McSherry and Talwar, 2007] The exponential mechanism $\mathcal{E}$ is $(\epsilon, 0)$-DP.*

Proof. Let's continue under the assumption that $\mathcal{Y}$ is finite. For any $y \in \mathcal{Y}$,

$$\begin{aligned}\frac{e^{\frac{\varepsilon}{2\Delta}f(y,S)}}{e^{\frac{\varepsilon}{2\Delta}f(y,S')}} &= e^{\frac{\varepsilon}{2\Delta}(f(y,S)-f(y,S'))} \\ &\leq e^{\frac{\varepsilon\Delta}{2\Delta}} \\ &= e^{\frac{\varepsilon}{2}}\end{aligned}$$

Now we'll need an analogous bound for the normalization factors.

$$\begin{aligned}\frac{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S')}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} &\leq \frac{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}(f(y',S)+\Delta)}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} \\ &= e^{\frac{\varepsilon\Delta}{2\Delta}} \frac{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} \\ &= e^{\frac{\varepsilon}{2}}\end{aligned}$$

We can now put it all together:

$$\begin{aligned}\frac{\Pr[\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon)=y]}{\Pr[\mathcal{E}(S',\mathcal{Y},f,\Delta,\varepsilon)=y]} &= \frac{e^{\frac{\varepsilon}{2\Delta}f(y,S)}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} \frac{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S')}}{e^{\frac{\varepsilon}{2\Delta}f(y,S')}} \\ &\leq e^{\frac{\varepsilon}{2}} e^{\frac{\varepsilon}{2}} \\ &= e^{\varepsilon}\end{aligned}$$

It follows that for all $y \in \mathcal{Y}$

$$\Pr[\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon)=y] \leq e^{\varepsilon}\Pr[\mathcal{E}(S',\mathcal{Y},f,\Delta,\varepsilon)=y].$$

□

Theorem 0.2. *Let $OPT(S) = \max_{y\in\mathcal{Y}} f(y,S)$ be the largest score obtainable by any output $y \in \mathcal{Y}$. Let $\mathcal{Y}^* = \{y \in \mathcal{Y} : f(y,S) = OPT(S)\}$. Then*

$$\Pr[f(\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon) \leq OPT(S) - \frac{2\Delta}{\varepsilon}\left(\ln\tfrac{|\mathcal{Y}|}{|\mathcal{Y}^*|} + t\right)] \leq e^{-t}$$

Proof.

$$\begin{aligned}
\Pr[f(\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon)\leq c] &= \frac{\sum_{y'\in\mathcal{Y}:f(y',S)\leq c} e^{\frac{\varepsilon}{2\Delta}f(y',S)}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} \\
&\leq \frac{|\mathcal{Y}|e^{\frac{\varepsilon c}{2\Delta}}}{\sum_{y'\in\mathcal{Y}} e^{\frac{\varepsilon}{2\Delta}f(y',S)}} \\
&\leq \frac{|\mathcal{Y}|e^{\frac{\varepsilon c}{2\Delta}}}{|\mathcal{Y}^*|e^{\frac{\varepsilon OPT}{2\Delta}}} \\
&= \frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}e^{\frac{\varepsilon}{2\Delta}(c-OPT)}
\end{aligned}$$

Plugging in $c = OPT(S) - \frac{2\Delta}{\varepsilon}\left(\ln\frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}+t\right)$, we get

$$\begin{aligned}
\Pr[f(\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon)\leq c] &\leq \frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}e^{\frac{\varepsilon}{2\Delta}(-\frac{2\Delta}{\varepsilon}(\ln\frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}+t))} \\
&= \frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}e^{-(\ln\frac{|\mathcal{Y}|}{|\mathcal{Y}^*|}+t)} \\
&= e^{-t}
\end{aligned}$$

□

The exponential mechanism can also be used to privately learn finite hypothesis classes!

Algorithm 3 Private Learner $\mathcal{L}(S,\mathcal{H},\varepsilon)$

1: $f(h,S) = -\frac{1}{|S|}\sum_{(x,y)\in S}\mathbb{1}[h(x)\neq y]$
2: $\Delta = \frac{1}{|S|}$
3: **return** $\mathcal{E}(S,\mathcal{H},f,\Delta,\varepsilon)$

Theorem 0.3. *[Kasiviswanathan et al., 2011] Let $\mathcal{H}$ be a finite hypothesis class. There exists an $(\epsilon,0)$-DP PAC learner for $\mathcal{H}$ with sample complexity*

$$m\in O\left(\frac{\log(|\mathcal{H}|/\beta)}{\alpha^2}+\frac{\log(|\mathcal{H}|/\beta)}{\alpha\varepsilon}\right)$$

where α is the target error and β is the target failure rate.

Proof. We've already showed that the exponential mechanism is private, but we need to show that it finds a good hypothesis to return.

The score is just the empirical error. Note that we've chosen our sample to be large enough such that except with probability δ

$$-f(h,S) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y] \leq \Pr_{(x,y)\sim D}[h(x) \neq y] + \alpha/2$$

for all hypotheses

So as long as we return a hypothesis with score at least $OPT - \frac{\alpha}{2}$ with probability at least $\beta/2$, we'll have a hypothesis that is within α of optimal error.

We just showed that

$$\Pr[f(\mathcal{E}(S,\mathcal{Y},f,\Delta,\varepsilon) \leq OPT(S) - \frac{2\Delta}{\varepsilon}\left(\ln\frac{|\mathcal{Y}|}{|\mathcal{Y}^*|} + t\right)] \leq e^{-t}.$$

We need $e^{-t} < \beta/2$, so putting this together with the accuracy requirement, we want

$$\begin{aligned}
OPT(S) - \frac{2\Delta}{\varepsilon}\left(\ln\frac{|\mathcal{H}|}{|\mathcal{H}^*|} + \ln\frac{2}{\beta}\right) &\geq OPT(S) - \alpha/2 \\
\Rightarrow \Delta\left(\ln\frac{|\mathcal{H}|}{|\mathcal{H}^*|} + \ln\frac{2}{\beta}\right) &\leq \frac{\alpha\varepsilon}{4} \\
\Rightarrow \Delta \leq \frac{\alpha\varepsilon}{4\left(\ln\frac{|\mathcal{H}|}{|\mathcal{H}^*|} + \ln\frac{2}{\beta}\right)} \\
\Rightarrow \Delta \leq \frac{\alpha\varepsilon}{4\left(\ln|\mathcal{H}| + \ln\frac{2}{\beta}\right)} \\
\Rightarrow |S| \geq \frac{4(\ln|\mathcal{H}| + \ln\frac{2}{\beta})}{\alpha\varepsilon}
\end{aligned}$$

□

References

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.