EN.601.774 Theory of Replicable ML Spring 2025

Lecture 18

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

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** $y \sim D_Y$

Theorem 0.1. *Let $f : \mathcal{Y} \times S \to \mathbb{R}$ be the score function given as input to $\mathcal{E}$. Let $OPT(S) = \max_{y \in \mathcal{Y}} f(y, S)$ be the largest score obtainable by any output $y \in \mathcal{Y}$. Then except with probability β,*

$$f(\mathcal{E}(S, \mathcal{Y}, f, \Delta, \varepsilon), S) > OPT(S) - \frac{2\Delta}{\varepsilon} (\ln |\mathcal{Y}| + \log(1/\beta))$$

Theorem 0.2. *Bassily et al. [2016] Let $\varepsilon \in [\sqrt{\frac{12}{n}}, \frac{1}{8}]$ and $\delta \leq \frac{\varepsilon}{16}$. Let $\mathcal{M} : \mathcal{X}^m \to \mathcal{Q}$ be an (ε, δ)-private algorithm, where $\mathcal{Q}$ is the class of all queries such that $|q(S) - q(S')| \leq \frac{1}{m}$ for $|S| = m$. Then for any distribution D on $\mathcal{X}$:*

$$\Pr_{S \sim D^m q \leftarrow \mathcal{M}(S)} [|q(S) - q(D)| \geq 6\varepsilon] \leq \max\{\frac{4\delta}{\varepsilon}, e^{\frac{-\varepsilon^2 m}{8}}\}$$

Algorithm 2 Monitor($\{S_t\}_{t=1}^T$)
Parameters: Number of datasets T
Privacy parameters $\varepsilon, \varepsilon', \delta$
(ε, δ)-DP Mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Q}$
Distribution D on $\mathcal{X}$

1: $Y = \emptyset$
2: **for** $t \in [T]$ **do**
3: $\quad q_t \leftarrow \mathcal{M}(S_t)$
4: $\quad q_{-t}(x) = 1 - q_t$
5: $\quad Y = Y \cup \{(t, q_t), (t, q_{-t})\}$
6: **end for**
7: define score $f : Y \to \mathbb{R}$, $f((t, q), S) = q(S_t) - q(D)$
8: $\Delta = \frac{1}{|S_1|}$
9: $(t^*, q^*) \leftarrow \mathcal{E}(S, Y, f, \Delta, \varepsilon')$
10: **return** (t^*, q^*)

Proof Plan

1. Show Monitor is $(\varepsilon+\varepsilon', \delta)$-DP
2. Lower bound the expected generalization error of the query output by the monitor as a function of how often $\mathcal{M}$ overfits its data by more than 6ε
3. Upper bound the expected generalization error of the query output by the monitor using a (modified) version of results we've seen already (stability $\Rightarrow$ expected generalization guarantees)
4. Use the upper bound on generalization error to show that the probability of overfitting by more than 6ε must be small, obtaining high probability generalization guarantees!

Step 2

Claim 0.3. *Let*

$$p_\alpha = \Pr_{\substack{S\sim D^m \\ q\leftarrow \mathcal{M}(S)}} [q(S) - q(D) \geq \alpha].$$

Let $S_{max} = \max_{(t,q)\in Y} f((t,q),S) = \max_{(t,q)\in Y} q(S_t) - q(D)$. *Then*

$$\mathop{\mathbb{E}}_{\substack{S\sim (D^m)^T \\ Monitor(S)}} [S_{max}] \geq \alpha(1-(1-p_\alpha)^T).$$

Step 3

Lemma 0.4. *(DP Monitor $\Rightarrow$ Expected Generalization) For every $\varepsilon > 0, \delta > 0, T \in \mathbb{N}$, and distribution D on $\mathcal{X}$: If the algorithm Monitor: $(\mathcal{X}^m)^T \to [T]\times\mathcal{Q}$ is (ε,δ)-DP and $S = (S_1,\ldots,S_T) \sim (D^m)^T$, then*

$$\mathop{\mathbb{E}}_{\substack{S\sim (D^m)^T \\ (t,q)\sim Monitor(S)}} [q(S_t) - q(D)] \leq (e^\varepsilon - 1) + T\delta$$

Definition 0.5. Let X, Y be random variables over a shared domain $\mathcal{O}$. We write $X \approx_{\varepsilon,\delta}$ to indicate that X, Y are ϵ,δ indistinguishable. That is, for all $T \subset \mathcal{O}$,

$$\Pr[X\in\mathcal{O}] \leq e^\varepsilon \Pr[Y\in\mathcal{O}] + \delta$$

Lemma 0.6. *Let X, Y be distributions on a set $\mathcal{O}$ such that $X \approx_{\varepsilon,\delta} Y$, and let $f : \mathcal{O} \to [0,1]$ be a bounded real-valued function. Then*

$$\mathbb{E}[f(X)] \leq e^\varepsilon\, \mathbb{E}[f(Y)] + \delta$$

Proof. We'll use the fact that for non-negative r.v.'s $\mathbb{E}[X] = \int_{x=0}^{\infty} \Pr[f(X) \leq x] dx$

$$\begin{aligned}\mathbb{E}[f(X)] &= \int_{z=0}^{1} \Pr[f(X) \leq z] dz \\ &\leq \int_{z=0}^{1} e^{\varepsilon} \Pr[f(Y) \leq z] + \delta dz \\ &= e^{\varepsilon}\, \mathbb{E}[f(Y)] + \delta\end{aligned}$$

□

Proof. (DP Monitor ⇒ Expected Generalization) We write $S = (S_1, \ldots, S_t)$. Then we can express the expected value of the query q output by the Monitor on the associated subsample S_t:

$$\mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [q^*(S_{t^*})] = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(S_{t^*})]$$

Similarly, we have

$$\mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [q^*(D)] = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(D)]$$

Ultimately we want to be able to relate the expectation of q^* on S_{t*} to the expectation of q^* on D. Let's start by relating it to the expectation of q^* on a neighboring dataset. We won't just consider any worst-case neighboring dataset in this instance, however, we'll consider a neighboring dataset S' in which the new element $x' \sim D$, and $S' = (S_1, \ldots, S_{t,j\to x'}, \ldots, S_T)$. We know that the Monitor is (ε, δ)-DP, so its outputs (q^*, t^*) and $(q^{*'}, t^{*'})$ given S, S' must satisfy $(q^*, t^*) \approx_{\varepsilon,\delta} (q^{*'}, t^{*'})$ for any $t \in [T]$ and any $j \in [m]$.

We can follow the techniques from previous lectures and define the bounded function $f_{S,j}(q, t) = q(S_{t,j})$. For fixed j, let $S'_{t,j} = (S_1, \ldots, S_{t,j\to x'}, \ldots, S_T)$. Then using Lemma 0.6 and our distribution swapping trick, it follows that

$$
\begin{aligned}
\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(S_t)] &= \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(S_{t,j})] \\
&\leq \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{T} e^{\varepsilon} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T, x'\sim D \\ (q^*,t^*)\leftarrow Monitor(S'_{t,j})}} [\mathbb{1}_{t=t^*} \cdot q^*(S_{t,j})] + \delta \\
&= \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{T} e^{\varepsilon} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T, x'\sim D \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(x')] + \delta \\
&= \frac{1}{m} \sum_{j=1}^{m} \sum_{t=1}^{T} e^{\varepsilon} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(D)] + \delta \\
&= \sum_{t=1}^{T} e^{\varepsilon} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(D)] + \delta
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (t,q)\sim Monitor(S)}} [q(S_t) - q(D)] &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(S_{t^*}) - \mathbb{1}_{t=t^*} \cdot q^*(D)] \\
&\leq \sum_{t=1}^{T} (e^{\varepsilon} - 1) \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (q^*,t^*)\leftarrow Monitor(S)}} [\mathbb{1}_{t=t^*} \cdot q^*(D)] + \delta \\
&\leq (e^{\varepsilon} - 1) + T\delta
\end{aligned}
$$

□

Step 4 (For the remainder of the analysis, we'll make the simplifying assumption that $\mathcal{E}$ always returns an output with nearly optimal score, rather than "except with probability β).

For the Monitor algorithm, we have $\Delta = \frac{1}{m}$, so we know that the exponential mechanism, and therefore the Monitor algorithm, will return a pair (t, q) such that

$$\begin{aligned} f(\mathcal{E}(S, Y, f, \Delta, \varepsilon), S) &> S_{max} - \frac{2\ln T}{\varepsilon m} \\ \Rightarrow q(S_t) - q(D) &> S_{max} - \frac{2\ln T}{\varepsilon m} && \text{by def of } f \\ \Rightarrow \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (t,q)\sim Monitor(S)}} [q(S_t) - q(D)] &> \alpha(1-(1-p_\alpha)^T) - \frac{2\log T}{\varepsilon m} && \text{by Step 2} \end{aligned}$$

We also have from Lemma 0.4 that

$$\mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (t,q)\sim Monitor(S)}} [q(S_t) - q(D)] \leq \mathop{\mathbb{E}}_{\substack{S\sim(D^m)^T \\ (t,q)\sim Monitor(S)}} [q(S_t) - q(D)] \leq (e^{\varepsilon+\varepsilon'} - 1) + T\delta$$

It follows that

$$\alpha(1-(1-p_\alpha)^T) - \frac{2\log T}{\varepsilon m} \leq (e^{\varepsilon+\varepsilon'} - 1) + T\delta$$

Take $\varepsilon' = \varepsilon$ and $T = \lfloor 1/p_\alpha \rfloor$, so that $1 - p_\alpha \leq p_\alpha T$ and note that for any $p_\alpha \leq 1/4$ we have

$$(1-p_\alpha)^T \leq e^{-p_\alpha T} \leq e^{-1+p_\alpha} \leq 1/2.$$

Then

$$\begin{aligned} \alpha(1-(1-p_\alpha)^T) &\geq \frac{\alpha}{2} \\ \Rightarrow \frac{\alpha}{2} - \frac{2\ln T}{\varepsilon m} &\leq (e^{2\varepsilon} - 1) + T\delta \\ \Rightarrow \alpha - 2(e^{2\varepsilon} - 1) &\leq \frac{4\ln T}{\varepsilon m} + 2T\delta \\ &\leq \frac{4\ln\frac{1}{p_\alpha}}{\varepsilon m} + \frac{2\delta}{p_\alpha} \end{aligned}$$

We want to show that $p_\alpha \leq \max\{\frac{4\delta}{\varepsilon}, e^{\frac{-\varepsilon^2 m}{8}}\}$ for $\alpha = 6\varepsilon$, and $\varepsilon \leq 1/8$. Recalling the fun math fact $e^{2x} \leq 1 + 2x + 4x^2$ when $x \leq 1/2$, it follows that $e^{2\varepsilon} - 1 \leq 2\varepsilon + 4\varepsilon^2$. Then

$$\begin{aligned} \alpha - 2(e^{2\varepsilon} - 1) &\geq 6\varepsilon - 4\varepsilon - 8\varepsilon^2 \geq 2\varepsilon - \varepsilon = \varepsilon \\ \Rightarrow \varepsilon &\leq \frac{4\ln\frac{1}{p_\alpha}}{\varepsilon m} + \frac{2\delta}{p_\alpha} \end{aligned}$$

For this to be true, at least one of $\frac{4\ln\frac{1}{p_\alpha}}{\varepsilon m}$ or $\frac{2\delta}{p_\alpha}$ must be at least than $\varepsilon/2$. So

- either $\frac{2\delta}{p_\alpha} \geq \frac{\varepsilon}{2} \Rightarrow p_\alpha \leq \frac{4\delta}{\varepsilon}$
- or $\frac{4\ln\frac{1}{p_\alpha}}{\varepsilon m} \geq \frac{\varepsilon}{2} \Rightarrow p_\alpha \leq e^{\frac{\varepsilon^2 m}{8}}$.

References

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.