EN.601.774 Theory of Replicable ML Spring 2025

Lecture 17

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

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

Define distribution $D_Y(y) \propto e^{\frac{\varepsilon}{2\Delta} f(y,S)}$
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 $\mathcal{X}$

$Y = \emptyset$
for $t \in [T]$ **do**
  $q_t \leftarrow \mathcal{M}(S_t)$
  $q_{-t}(x) = 1 - q_t$
  $Y = Y \cup \{(t, q_t), (t, q_{-t})\}$
end for
define score $f : Y \to \mathbb{R}$, $f((t, q), S) = q(S_t) - q(D)$
$\Delta = \frac{1}{|S_1|}$
$(t^*, q^*) \leftarrow \mathcal{E}(S, Y, f, \Delta, \varepsilon')$
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!

Claim 0.3. *Let $\mathcal{M}$ be an (ε, δ)-DP algorithm. Then the Monitor algorithm run with $\mathcal{M}$ is $(\varepsilon + \varepsilon', \delta)$-DP.*

Proof. Let $\mathcal{M}_Y(\{S_t\}_{t=1}^T)$ be as in Algorithm 3. Then $\mathcal{M}_Y$ is (ε, δ)-DP. This follows from the fact that for neighboring S, S', there is only a single value of t such that $S_t \neq S'_t$, and therefore there is only a single t such that $\{(t, q_t), (t, q_{-t})\}$ is not identically distributed to $\{(t, q'_t), (t, q'_{-t})\}$. Let $Y = (y_1, y_2, \ldots, y_T)$ and let t^* denote the value of $t*$ such that $S_{t^*} \neq S'_{t^*}$. Then we have

$$\begin{aligned}
\Pr[\mathcal{M}_Y(S) = Y] &= \Pr[\mathcal{M}(S_1) = y_1] \ldots \Pr[\mathcal{M}(S_{t^*}) = y_{t^*}] \ldots \Pr[\mathcal{M}(S_T) = y_T] \\
&= \Pr[\mathcal{M}(S'_1) = y_1] \ldots \Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}] \ldots \Pr[\mathcal{M}(S'_T) = y_T] \cdot \frac{\Pr[\mathcal{M}(S_{t^*}) = y_{t^*}]}{\Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}]} \\
&= \Pr[\mathcal{M}_Y(S') = Y] \cdot \frac{\Pr[\mathcal{M}(S_{t^*}) = y_{t^*}]}{\Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}]} \\
&\leq \Pr[\mathcal{M}_Y(S') = Y] \cdot \frac{e^{\varepsilon} \Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}] + \delta}{\Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}]} \\
&= e^{\varepsilon} \Pr[\mathcal{M}_Y(S') = Y] + \frac{\delta \Pr[\mathcal{M}_Y(S') = Y]}{\Pr[\mathcal{M}(S'_{t^*}) = y_{t^*}]} \\
&\leq e^{\varepsilon} \Pr[\mathcal{M}_Y(S') = Y] + \delta
\end{aligned}$$

Note that the Y given to $\mathcal{E}$ as input in the Monitor algorithm is just a postprocessing of $\mathcal{M}_Y$, as is therefore *also* (ε, δ)-DP. Recall that the exponential mechanism run with privacy parameter ε is ε-DP. Then the composition of $\mathcal{E}$ with the post-processed output of $\mathcal{M}_Y$ is $(\varepsilon + \varepsilon', \delta)$-DP. □

Algorithm 3 $\mathcal{M}_Y(\{S_t\}_{t=1}^T)$
Parameters: Number of datasets T
Privacy parameters $\varepsilon, \varepsilon'$
Mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Q}$
Distribution D on $\mathcal{X}$

1: $\mathcal{Y} = \emptyset$
2: **for** $t \in [T]$ **do**
3: $\quad q_t \leftarrow \mathcal{M}(S_t)$
4: $\quad Y = Y \cup \{(t, q_t)\}$
5: **end for**
6: **return** Y

Claim 0.4. *Let*

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

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

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

Proof. The probability that there exists some $t \in [T]$ such that $q_t(S_t) - q_t(D) \geq \alpha$ is $1 - (1 - p_\alpha)^T$. Furthermore $S_{max} \geq 0$ always, because we include both q and $1 - q$ in Y. Therefore

$$\mathop{\mathbb{E}}_{\substack{S \sim (D^m)^T \\ Monitor(S)}} [S_{max}] \geq \alpha(1 - (1 - p_\alpha)^T) + 0 \cdot (1 - p_\alpha)^T$$

□

Lemma 0.5. *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*

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

(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

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

We also have from Lemma 0.5 that

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

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.