Lecture : Approximate DP (Contd.)

Last time, we discussed a relaxation of (pure) ϵ-DP, which we called (approximate) (ϵ,δ)-DP. In this lecture, we shall discuss some properties of (ϵ,δ)-DP and present a "silver-bullet" mechanism for achieving it for the release of aggregate statistics.

Properties of (ϵ,δ)-DP :

1. **Basic Composition**: The composition of k (ϵ,δ)-DP mechanisms satisfies $(k\epsilon, k\delta)$-DP.

Proof : Fix neighbouring $\underline{x}, \underline{x}'$ and let the outputs of the k composed mechanisms M be $(a_1,\dots,a_k)$. We wish to show that

$$\Pr[M(\underline{x}) = (a_1,\dots,a_k)] \le e^{\epsilon}\cdot \Pr[M(\underline{x}') = (a_1,\dots,a_k)] + \delta.$$

Now, note that $\Pr[M(\underline{x}) = (a_1,\dots,a_k)] = \prod_{i=1}^{k} \Pr[M_j(\underline{x}) = a_j]$, where M_j implicitly takes as arguments a_1^{j-1}, for $j\in[k]$.

We know that $\Pr[M_j(\underline{x}) = a_j] \le e^{\epsilon}\Pr[M_j(\underline{x}') = a_j] + \delta$.

Infact, we have that

$$\Pr[M_j(\underline{x}) = a_j] \le \min\{e^{\epsilon}\Pr[M_j(\underline{x}') = a_j] + \delta, 1\}.$$

$$\leq \min\{e^{\epsilon} \cdot \Pr[M_j(\underline{x}') = a_j], 1\} + \delta .$$

The following sequence of inequalities then holds:

$$\prod_{j=1}^{k} \Pr[M_j(\underline{x}) = a_j] \leq \min\{e^{\epsilon} \cdot \Pr[M_1(\underline{x}) = a_1], 1\} \cdot \prod_{j=2}^{k} \Pr[M_j(\underline{x}) = a_j] + \delta$$

$$\leq \min\{e^{\epsilon} \Pr[M_1(\underline{x}') = a_1], 1\} \cdot \min\{e^{\epsilon} \Pr[M_2(\underline{x}') = a_2], 1\} \times \prod_{j=3}^{k} \Pr[M_j(\underline{x}) = a_j] + 2\delta$$

$$\vdots$$

$$\leq \prod_{j=1}^{k} \min\{e^{\epsilon} \Pr[M_j(\underline{x}') = a_j], 1\} + k\delta$$

$$\leq e^{k\epsilon} \cdot \Pr[M(\underline{x}') = \underline{a}] + k\delta . \quad \blacksquare$$

2. <u>Post-processing</u>: The post-processing of the output of an (ϵ, δ)-DP mechanism preserves (ϵ, δ)-DP.

<u>Proof</u>: Similar to that for ϵ-DP.

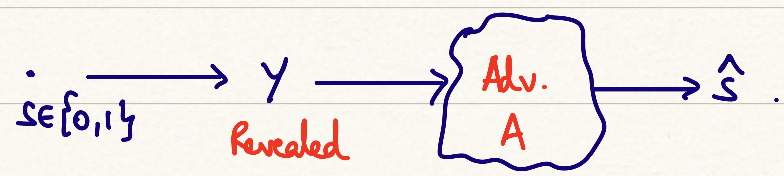
In fact, we shall soon see that (ϵ, δ)-DP obeys a much stronger "Advanced Composition Theorem". But before we do so, we shall view the two notions of DP discussed until now from a slightly different lens.

Aside : An information-theoretic lens on DP

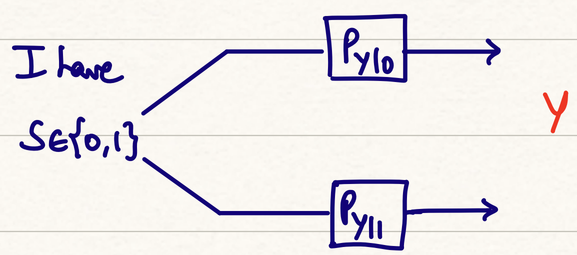
We shall first review the definition and properties of DP from an information-theoretic angle [Cf. A. Sarwate, F.P. Calmon, O. Kosut, L. Sankar, "An information theorist's tour of DP," IEEE BITS 2025].

Reviewing (ϵ,δ)-DP :

Consider the following setting : I have a secret $s \in \{0,1\}$ and some information about s in the form of a RV Y is revealed to a client. The client, if malicious, may use Y to deduce information about $s \in \{0,1\}$.



This situation is reminiscent of a hypothesis test:



with $\mathcal{H}_0 : Y \sim P_{Y|0}$ and $\mathcal{H}_1 : Y \sim P_{Y|1}$. One can verify that the definition of (ϵ,δ)-DP when the dataset is $\{s\}$ is equivalent to requiring that

$$\begin{aligned} P_{FA} + e^{\epsilon} P_{MD} &\geq 1-\delta, \\ P_{MD} + e^{\epsilon} P_{FA} &\geq 1-\delta. \end{aligned} \quad \text{—— (1)}$$

The "mechanism" I choose for releasing Y is simply a choice of conditional distributions $\{P_{Y|0}, P_{Y|1}\}$.

HW1: Plot the curves P_{FA} vs P_{MD}, for when the mechanism adds Gaussian Noise: $P_{Y|0} \sim \mathcal{N}(0, \sigma^2)$ and $P_{Y|1} \sim \mathcal{N}(1, \sigma^2)$, for an adversary that declares $\hat{s} = 0$, if $P_{Y|0} \geq P_{Y|1}$, and 1, o.w, for varying σ values.

Another characterization for $\mathcal{D} = \{s\}$ is obtained via the distribution of the "privacy loss RV" $L \triangleq \log \frac{P_{Y|1}(Y)}{P_{Y|0}(Y)}$. [Here and elsewhere, recall that $P_{Y|0}, P_{Y|1}$ are densities]

Namely, if $|L| \leq \epsilon$ almost surely, then the mechanism $(P_{Y|0}, P_{Y|1})$ is ϵ-DP.

HW2: Verify that if $\Pr_{Y|1}[L \leq \epsilon] \geq 1-\delta$, then $(P_{Y|0}, P_{Y|1})$ is (ϵ,δ)-DP.

Generalizing the notions above to the standard setting of "item-level DP" then follows easily: the mechanism I employ is a family of distributions $\{P_{Y|\mathcal{D}} : \mathcal{D} \in \mathbb{D}_n\}$, where $\mathbb{D}_n$ is the collection of datasets with n users.

Standard DP then requires that (1) hold, when $s \in \{0,1\}$ is suitably replaced by neighbouring datasets (either containing or missing some fixed user $i \in [n]$) $\mathcal{D}, \mathcal{D}'$. Likewise, the privacy loss RV can now be defined for database pairs as

$$L_{\mathcal{D},\mathcal{D}'} = \log \frac{P_{Y|\mathcal{D}}(Y)}{P_{Y|\mathcal{D}'}(Y)} .$$

<u>Privacy loss distribution (PLD) and composition</u>

Suppose that we have a sequence of mechanisms (corresponding to potentially adaptive queries) $M = (M_1, M_2, \cdots, M_k)$, with corresponding "released" RVs $\underline{Y} = (Y_1, \cdots, Y_k)$. Then, the privacy loss RV (PLRV) for the "composed" mechanism M for neighbouring datasets $\mathcal{D}, \mathcal{D}' \in \mathbb{D}_n$ is:

$$L_{\mathcal{D},\mathcal{D}'} = \log \frac{P(\underline{Y}|\mathcal{D})}{P(\underline{Y}|\mathcal{D}')}$$

$$= \log \prod_{i=1}^{k} \frac{P(Y_i|\mathcal{D}, y^{i-1})}{P(Y_i|\mathcal{D}', y^{i-1})} = \sum_{i=1}^{k} L^{(i)}_{\mathcal{D},\mathcal{D}'}, \quad \text{—(2)}$$

where $L^{(j)}_{\mathcal{D},\mathcal{D}'}$, $1 \le j \le k$, is the PLRV corresponding to the mechanism $M_j : \mathbb{D}_n \times \mathcal{Y}^{j-1} \longrightarrow \mathbb{R}$.

<u>HW3</u>: Via (2), argue that if each of M_j, $j \in [k]$, satisfies the condition in HW 2 (and is hence (ϵ, δ)-DP), then M is a $(k\epsilon, k\delta)$-DP mechanism.

Achieving (ϵ,δ)-DP: The Gaussian Mechanism

Let $\mathcal{D} = (x_1, \dots, x_N)$ as before. Given a function $f: \mathcal{D} \to \mathbb{R}^d$, the Gaussian mechanism simply performs

$$M^G_{\epsilon,\delta}(\underline{x}) = f(\underline{x}) + \mathcal{N}\left(\underline{0}, \frac{2\Delta_2^2 \log(2/\delta)}{\epsilon^2} I_d\right)$$

where $\Delta_2 = \Delta_{2,f} \triangleq \max_{\underline{x} \sim \underline{x}'} \| f(\underline{x}) - f(\underline{x}')\|_2$. In order to prove that this Gaussian mechanism satisfies (ϵ,δ)-DP, we shall invoke HW2 above.

<u>Theorem</u>: For suitable $\epsilon, \delta, \Delta_2$ values $M^G_{\epsilon,\delta}$ satisfies (ϵ,δ)-DP.

<u>Proof</u>: Consider the PLRV $\left(\sigma^2 = \frac{2\Delta_2^2}{\epsilon^2} \log(2/\delta)\right)$

$$L_{\underline{x},\underline{x}'} = \ln \frac{f_{\underline{x}}(\underline{y})}{f_{\underline{x}'}(\underline{y})} = \ln \left(\frac{\exp\left(-\frac{\|\underline{y} - f(\underline{x})\|_2^2}{2\sigma^2}\right)}{\exp\left(-\frac{\|\underline{y} - f(\underline{x}')\|_2^2}{2\sigma^2}\right)} \right)$$

pdf of $M^G_{\epsilon,\delta}(\underline{x})$ at $\underline{y}$

$$= \frac{\|\underline{z} + \underline{v}\|_2^2 - \|\underline{z}\|_2^2}{2\sigma^2},$$

$\sim \mathcal{N}(0, \sigma^2 I_d)$

where $\underline{z} \triangleq \underline{y} - f(\underline{x})$ and $\underline{v} \triangleq f(\underline{x}) - f(\underline{x}')$. Consider the event $\{L_{\underline{x},\underline{x}'} > \epsilon\}$. We wish to show that the probability of this event is at most δ, under the distribution of $\underline{z}$.

Observe that $\left\{ \|\underline{z}+\underline{v}\|_2^2 - \|\underline{z}\|_2^2 > 2\sigma^2\epsilon \right\} = \left\{ 2\langle \underline{z}, \underline{v}\rangle + \|\underline{v}\|_2^2 > 2\sigma^2\epsilon \right\}$

$$\subseteq \left\{ 2\langle \underline{z}, \underline{v}\rangle + \Delta^2 > 2\sigma^2\epsilon \right\}.$$

Consider the RV $W = \langle \underline{z}, \underline{v}\rangle$. Note that $W \sim \mathcal{N}(\underline{0}, \|\underline{v}\|_2^2 \sigma^2)$.

Hence,

$$\Pr\left[W + \frac{\Delta^2}{2} > \sigma^2\epsilon \right] = \Pr\left[W > \left(\sqrt{2\log(2/\delta)} - \frac{\Delta}{2\sigma} \right) \Delta\sigma \right] \quad \text{— (1)}$$

We restrict our attention to the regime where

$$\sqrt{2\log\left(\frac{2}{\delta}\right)} - \frac{\Delta}{2\sigma} > \sqrt{2\log(1/\delta)}.$$

Thus, we have from (1) that

$$\Pr\left[L_{\underline{x},\underline{x}'} > \epsilon \right] \leq \Pr\left[W > \sqrt{2\log\left(\frac{1}{\delta}\right)} \cdot \Delta\sigma \right]$$

$$\leq \Pr\left[W > \sqrt{2\log(1/\delta)} \cdot \|\underline{v}\|_2 \sigma \right]. \quad \text{—(2)}$$

Since $W \sim \mathcal{N}(\underline{0}, \|\underline{v}\|_2^2\sigma^2)$, we have by Gaussian concentration of (2) that

$$\Pr\left[L_{\underline{x},\underline{x}'} > \epsilon \right] \leq \delta. \quad \blacksquare$$