Privacy Amplification by Subsampling
Tight Analyses via Couplings and Divergences

Borja Balle[1] Gilles Barthe[2] Marco Gaboardi[3]

[1]Amazon Research, Cambridge, UK [2]IMDEA Software Institute, Madrid, Spain [3]University at Buffalo (SUNY), Buffalo, USA

Differential Privacy and Subsampling

Subsampling (Informal Definition) A *subsampling mechanism* is a randomized algorithm $S : X^n \to X^m$ that given as input a tuple $x = (x_1, \dots, x_n)$ outputs a random tuple $y = (y_1, \dots, y_m)$ obtained by "subsampling" x.

Subsampled Mechanisms

- Given a mechanism $M : X^m \to Z$ and a subsampling $S : X^n \to X^m$ we consider the subsampled mechanism $M^S(x)$ that first obtains $y \sim S(x)$ and then outputs $M(y)$.
- Privacy amplification intuition: M^S should provide more privacy than M because when the subsample $y \sim S(x)$ does not contain the individual we are trying to protect no leakage occurs.
- The output distribution of $M^S(x)$ is a *mixture*:

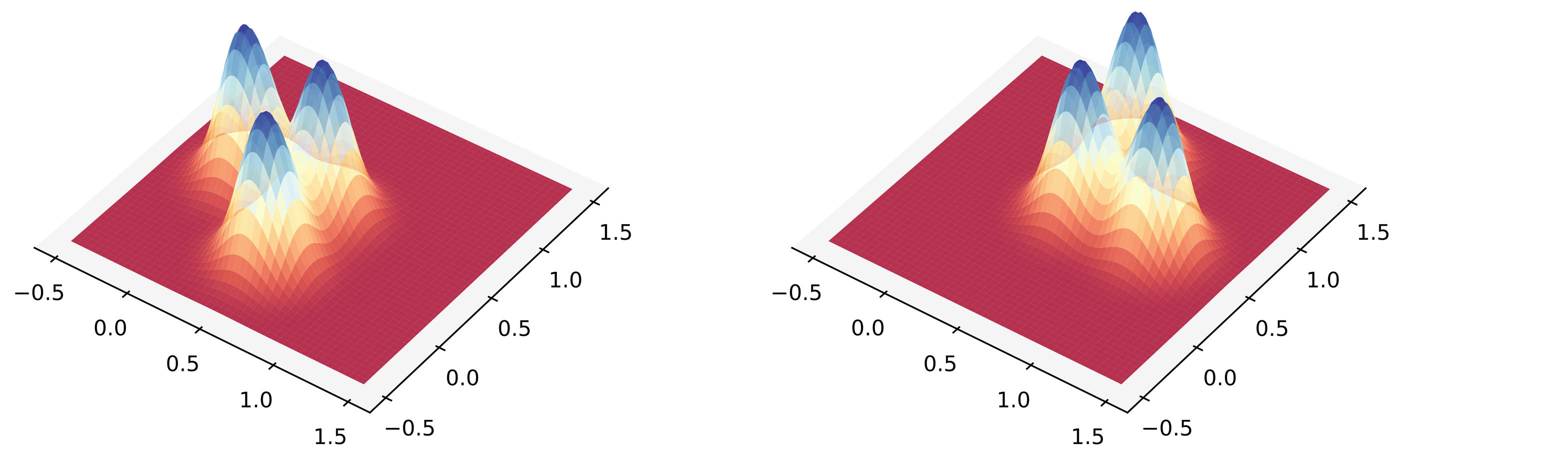
$$\Pr[M^S(x) = z] = \sum_{y \in X^m} \Pr[S(x) = y] \cdot \Pr[M(y) = z] = \sum_y \omega_x(y) \cdot \mu_y(z)$$

- Technical challenge: analyze differential privacy guarantees of mechanisms whose output distribution is a mixture (with a large number of components).

Example Subsampled Gaussian Mechanism $(n = 3, m = 2)$

$$x = \{(0,0), (1,0), (0,1)\} \quad \to \quad M^S(x) \equiv \frac{\mathcal{N}(p_0, \sigma^2 I) + \mathcal{N}(p_1, \sigma^2 I) + \mathcal{N}(p_2, \sigma^2 I)}{3}$$

$$x' = \{(1,1), (1,0), (0,1)\} \quad \to \quad M^S(x) \equiv \frac{\mathcal{N}(p_0, \sigma^2 I) + \mathcal{N}(p'_1, \sigma^2 I) + \mathcal{N}(p'_2, \sigma^2 I)}{3}$$



Examples of Subsampling Mechanisms

Sampling Without replacement Given a set $x = \{x_1, \dots, x_n\}$, $y \sim S(x)$ is uniform among all $\binom{n}{m}$ subsets of x if size m. Can also be defined for multisets with indistinguishable copies.

Poisson Sampling Given a set $x = \{x_1, \dots, x_n\}$, $y \sim S(x)$ is obtained by adding to y each element from x with fixed probability γ.

Sampling With replacement Given a set $x = \{x_1, \dots, x_n\}$, $y \sim S(x)$ is obtained by picking m elements independently and uniformly from x (with replacement). Even when x is a set y can be a multiset.

Divergences and Privacy Profiles

The Hockey-Stick Divergence Given distributions μ, μ' over Z define the divergence:

$$D_{e^\varepsilon}(\mu \| \mu') = \sum_{z \in Z} [\mu(z) - e^\varepsilon \mu'(z)]_+ = \sup_{E \subseteq Z} \mu(E) - e^\varepsilon \mu'(E)$$

This is an f-divergence (in the sense of Csiszár) and therefore satisfies a number of important properties, including joint convexity and data processing inequality.

Differential Privacy with Divergences A randomized mechanism $M : X \to Z$ is (ε, δ)-DP if and only if:

$$\sup_{x \simeq x'} D_{e^\varepsilon}(M(x) \| M(x')) \leq \delta$$

Privacy Profiles Using the divergence point of view allows us to define the *privacy profile* of a mechanism M that gives all the (ε, δ) pairs for which the mechanism is (ε, δ)-DP:

$$\delta(e^\varepsilon) = \sup_{x \simeq x'} D_{e^\varepsilon}(M(x) \| M(x'))$$

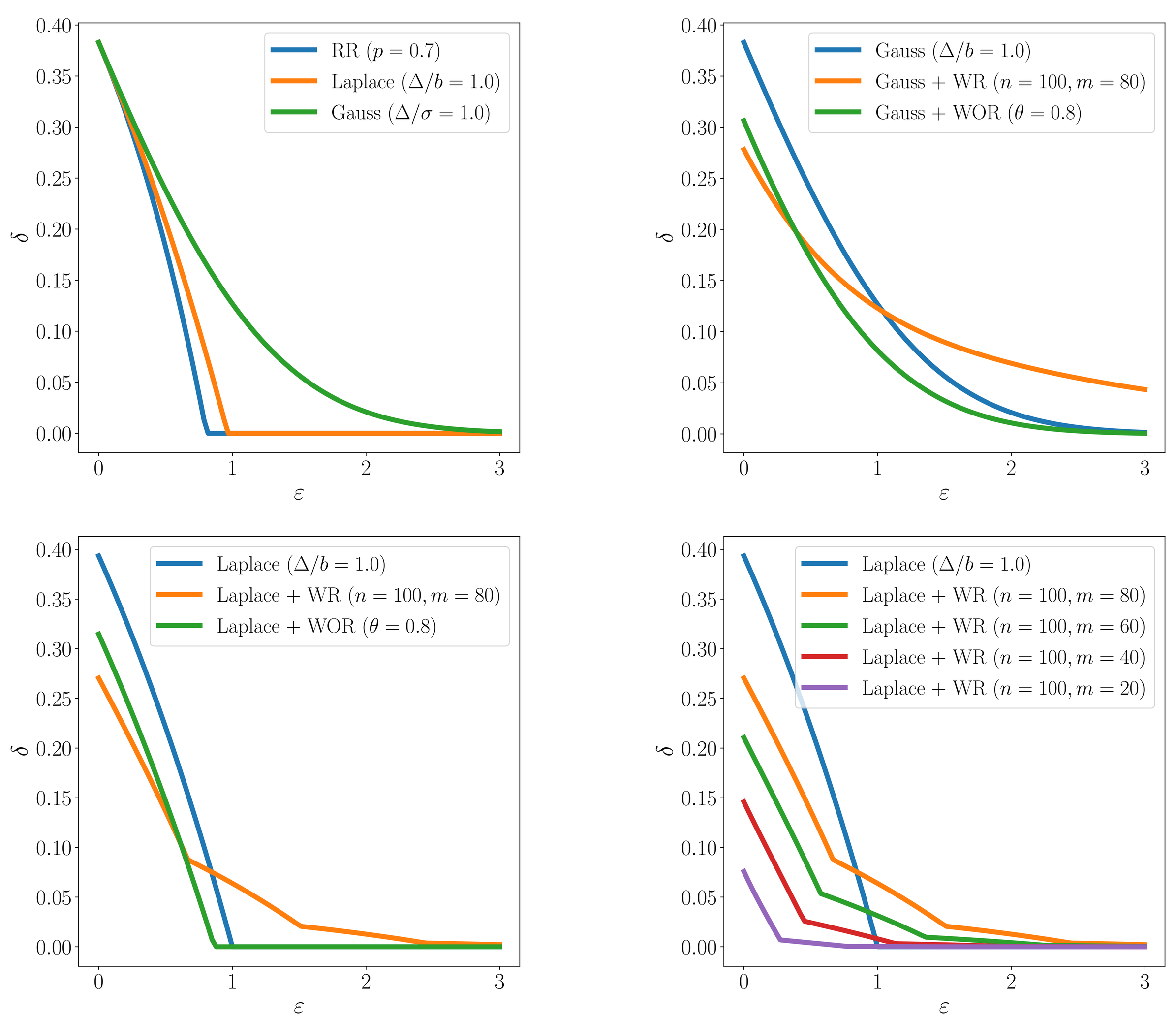
Group Privacy Profiles Using the relation $\simeq^k$ to allow k changes in the dataset we obtain the *group privacy profiles*:

$$\delta_k(e^\varepsilon) = \sup_{x \simeq^k x'} D_{e^\varepsilon}(M(x) \| M(x'))$$

Example (Laplace Mechanism) For $M(x) = f(x) + \mathsf{Lap}(b)$ with $\mathsf{GS}(f) = \Delta$ we have:

$$\delta(e^\varepsilon) = \left[1 - \exp\left(\frac{\varepsilon}{2} - \frac{b}{2\Delta}\right)\right]_+ \qquad \delta_k(e^\varepsilon) \leq \left[1 - \exp\left(\frac{\varepsilon}{2} - \frac{b}{2k\Delta}\right)\right]_+$$

Subsampled Privacy Profiles



Method Overview

Setup Given a subsampled mechanism M^S and inputs $x \simeq x'$ define the distributions

$$\omega \equiv S(x) \qquad \mu = \omega M \equiv M^S(x)$$
$$\omega' \equiv S(x') \qquad \mu' = \omega' M \equiv M^S(x')$$

Decomposing Mixtures via Maximal Couplings Given the *total variation distance* $\theta = \mathsf{TV}(\omega \| \omega')$, the *maximal coupling* between ω and ω' yields the overlapping decompositions:

$$\begin{aligned} \omega &= (1-\theta)\omega_0 + \theta\omega_1 \\ \omega' &= (1-\theta)\omega_0 + \theta\omega'_1 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \mu &= (1-\theta)\mu_0 + \theta\mu_1 \\ \mu' &= (1-\theta)\mu_0 + \theta\mu'_1 \end{aligned}$$

Cancelation for Overlapping Mixtures If $e^{\varepsilon'} = 1 + \theta(e^\varepsilon - 1)$ and $\beta = e^{\varepsilon'}/e^\varepsilon$ then we have

$$\begin{aligned} D_{e^{\varepsilon'}}((1-\theta)\mu_0 + \theta\mu_1 \| (1-\theta)\mu_0 + \theta\mu'_1) &= \theta D_{e^\varepsilon}(\mu_1 \| (1-\beta)\mu_0 + \beta\mu'_1) \\ &\leq (1-\beta)\theta D_{e^\varepsilon}(\mu_1 \| \mu_0) + \beta\theta D_{e^\varepsilon}(\mu_1 \| \mu'_1) \end{aligned}$$

Coupling Conditional Subsamplings By joint convexity, taking a coupling $\pi \in C(\tilde{\omega}, \tilde{\omega}')$ we get a bound in terms of group privacy profiles:

$$D_{e^\varepsilon}(\tilde{\omega} M \| \tilde{\omega}' M) \leq \sum_{y,y'} \pi(y, y') D_{e^\varepsilon}(M(y) \| M(y')) \leq \sum_{y,y'} \pi(y, y') \delta_{d(y,y')}(e^\varepsilon)$$

Distance-compatible Couplings Suppose $\tilde{\omega}$ and $\tilde{\omega}'$ admit a *d-compatible* coupling π with $(y, y') \in \mathsf{supp}(\pi) \Rightarrow d(y, y') = d(y, \mathsf{supp}(\tilde{\omega}'))$. Defining $Y_k = \{y : d(y, \mathsf{supp}(\tilde{\omega}')) = k\}$ and optimizing over couplings yields:

$$\min_{\pi \in C(\tilde{\omega}, \tilde{\omega}')} \sum_{y,y'} \pi(y, y') \delta_{d(y,y')}(e^\varepsilon) = \sum_{k \geq 0} \omega(Y_k) \delta_k(e^\varepsilon)$$

Tightness Results The bounds obtained by this method are attained by the *randomized membership* mechanism $M_{p,u}(x) = \mathsf{RandomizedResponse}_p(\mathbb{I}[u \in x])$.

Results for Typical Subsamplings

Concrete results depend on the neighbouring relations considered for M and M^S: *remove/add-one* (R) or *substitute-one* (S).

Sampling	$\simeq_M$	$\simeq_{M^S}$	θ	δ'
Poisson(γ)	R	R	γ	$\gamma\delta$
WOR(n,m)	S	S	$\frac{m}{n}$	$\frac{m}{n}\delta$
WR(n,m)	S	S	$1 - \left(1 - \frac{1}{n}\right)^m$	$\sum_{k \in [m]} \binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} \delta_k$
WR(n,m)	S	R	$1 - \left(1 - \frac{1}{n}\right)^m$	$\sum_{k \in [m]} \binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} \delta_k$
Poisson(γ)	S	S	(It's complicated, see paper)	