Composition Theorem for streaming CW-pricacy

1 Characterize Privacy as Regions

In this section, we show how to characterize Differential Privacy (DP) and CW-Privacy (CW-P) in terms of convex regions, and how to compute the (ϵ, δ) values for each mechanism in terms of tangent lines of such convex regions. We focus on the context of finite discrete states for the simplicity of analysis.

1.1 Case of DP

Let $\mathcal{X}$ be the set of all databases. Let $\mathcal{Y}$ be the set of all out comes of mechanism M. M is a probability measure from $\mathcal{X}$ to $\mathcal{Y}$. For simplicity of analysis, assume $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$. Then, M corresponds to an $n \times m$ Markov matrix $M = [M(x_1), \ldots, M(x_n)]^\top$.

Now, it is easy to see the following is an equivalent definition of DP.

Theorem 1. *For any $\epsilon \geq 0$ and $\delta \in [0, 1]$, a mechanism M is (ϵ, δ)-differentially private if and only if the following conditions are satisfied for all pairs of neighboring databases x and x', and all region $S \subseteq \mathcal{Y}$:*

$$Pr[M(x) \in S] + e^\epsilon Pr[M(x') \in \bar{S}] \geq 1 - \delta, \quad \textit{and}$$
$$e^\epsilon Pr[M(x) \in S] + Pr[M(x') \in \bar{S}] \geq 1 - \delta.$$

This gives a graphical representation (region) of DP:

$$R(\epsilon, \delta) = \{(p_x, p_y) \,|\, p_x + e^\epsilon p_y \geq 1 - \delta, e^\epsilon p_x + p_y \geq 1 - \delta\}.$$

For any two databases x and x', define

$$R(M, x, x') = convex\{(Pr[M(x) \in S], Pr[M(x') \in \bar{S}]) \,|\, \text{for all } S \subseteq \mathcal{Y}\}$$

$R(M, x, x')$ has the following equivalent form.

$$R(M, x, x') = \{(M(x) \cdot \alpha, M(x') \cdot \beta) \,|\, 0 \leq \alpha_i, \beta_i \leq 1,\ \alpha + \beta = \mathbf{1^m}\},$$

where $\cdot$ denote the dot production of vectors.

Definition 1. *For any mechanism M, we define its privacy region $R(M) = \bigcup_{(x,x')} R(M, x, x')$, where (x, x') is a pair of neighboring databases.*

Immediately, it should not hard to see the following theorem.

Theorem 2. *M is (ϵ, δ)-differentially private iff $R(M) \subseteq R(\epsilon, \delta)$.*

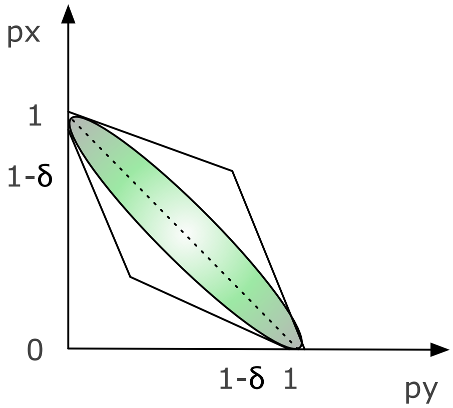

Figure 1: Every tangent line of the privacy region forms a pair of (ϵ, δ).

As it is shown in Figure 1, every tangent line of the privacy region corresponds to a pair of (ϵ, δ). In general, the privacy region $R(M, x, x')$ of any mechanism M can be represented by the intersections of all such regions $\{R(\epsilon_i, \delta_i)\}$, which is completely described by the set of slopes and shifts $\{(\epsilon_i, \delta_i)\}$.

1. For the set slopes, let $\mathcal{E} = \{0 \leq \epsilon_i < \infty \,|\, Pr[M(x) = y] = e^{\epsilon_i} Pr[M(x') = y]$ for some $y \in \mathcal{Y}\}$
2. For each ϵ_i, $\delta_i = \max_{S \subseteq \mathcal{Y}} \{\Sigma_{y \in S} Pr[M(x) = y] - e^{\epsilon_i} \Sigma_{y \in S} Pr[M(x) = y]\}$

1.2 Case of CW-P

In the case of CW-P, we use X to denote a database variable that follows some distribution D. Let the pmf of D be $f_D = [f_{x_1}, \ldots, f_{x_n}]$. Let $alt(X)$ denote a scrubbed version of X. Assume $alt(X)$ follows some distribution D' with pmf f'_D.

It is easy to see that CW-P has the following equivalent definition.

Theorem 3. *For any $\epsilon \geq 0$ and $\delta \in [0,1]$, a mechanism M is $(\epsilon, \delta, \Delta, \Gamma)$-differentially private if and only if the following conditions are satisfied for all distributions on D on (X, Z), all $(priv, alt) \in \Gamma$ pairs, and all region $S \subseteq \mathcal{Y}$:*

$$Pr[M(X) \in S \,|\, priv(X), Z] + e^{\epsilon} Pr[M(alt(X)) \in \bar{S} \,|\, priv(X), Z] \geq 1 - \delta, \quad \textit{and}$$

$$e^{\epsilon} Pr[M(X) \in S \,|\, priv(X), Z] + Pr[M(alt(X)) \in \bar{S} \,|\, priv(X), Z] \geq 1 - \delta,$$

where $M(X) = f_D M$ and $M(alt(X)) = f'_D M$.

Notice mechanism $M' = [f_D M, f'_D M]^\top$ can be seen as a DP version of M. Hence, CW-P also has form of privacy regions, and everything else described above follows.

2 Composition Theorem of CW-P in the streaming setting

Problem setting

Let X be a database in the streaming setting. Let X_i represent the portion of X that is currently held at time step i. We assume that at each time step, a fraction of c of the database is replaced. We assume the oldest rows are always the ones replaced, and that X has row drown i.i.d. from some distribution D. Let n be the size of each X_i. This means that the first n rows of X constitute X_1, rows $cn + 1$ through $cn + n$ constitute X_2, and so forth. X has size $n + cn(t - 1)$, where t is the total number of time steps being considered. $1/c$ is the total number of time steps a given row will be present for. See (1) and (2) in Figure 2.

Notations

For each X_i of size n, we represent it by $1/c$ blocks (each has cn rows). Namely, let $X_i = [x_{i1}, x_{i2}, \ldots, x_{i\frac{1}{c}}]^\top$, where x_{ij} denote the jth block in X_i. Let $X_{i\downarrow} = [x_{i2}, \ldots, x_{i\frac{1}{c}}]^\top$ and $X_{i\uparrow} = [x_{i1}, x_{i2}, \ldots, x_{i(\frac{1}{c}-1)}]^\top$. Namely, $X_{i\downarrow}$ represent the bottom $(1/c - 1)$ blocks of X_i while $X_{i\uparrow}$ represent the top $(1/c - 1)$ blocks of X_i.

Consider a mechanism $M : \mathcal{X} \to \mathcal{Y}$ on each X_i. We assume $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$ for the simplicity of analysis. Let each X_i follows some distribution D with pmf f_D. Let each $alt(X_i)$ follows some distribution D' with pmf f'_D.

Assumptions

1. M's auxiliary information consists of **the last** $(1-c)n$ **rows** of the database X_i, i.e., $X_{i\downarrow}$ is given. We denote the pmf of X_i conditioned on $X_{i\downarrow}=z$ as $f_{D|X_{i\downarrow}=z}$. We denote the pmf of $alt(X_i)$ conditioned on $X_{i\downarrow}=z$ as $f'_{D|X_{i\downarrow}=z}$. For all z and all $x\in\mathcal{X}$, we assume

   $$\frac{f_{D|X_{i\downarrow}=z}(x)}{f'_{D|X_{i\downarrow}=z}(x)} \le e^{\gamma},$$

   where $\gamma > 0$ is a small constant.

2. According to section 1, M can be characterized as a convex region. It should have a set of (ϵ_j, δ_j) values that are computable by its tangent lines. We assume that there is a subset $\mathcal{W}\subseteq\mathcal{Y}$ such that (we ignore the $priv(X_i)$ in the expression)

   (a) $\forall y\in\mathcal{W}$, $\forall z : Pr[M(X_i) = y \mid X_{i\downarrow} = z] \neq 0\ ,\ Pr[M(alt(X_i)) = y \mid X_{i\downarrow} = z] \neq 0\ ;$

   (b) $\forall y\in\mathcal{Y}-\mathcal{W}$, $\forall z : Pr[M(X_i) = y \mid X_{i\downarrow} = z] = 0\ ,\ Pr[M(alt(X_i)) = y \mid X_{i\downarrow} = z] = 0\ .$

   For each $y_j\in\mathcal{W}$ and for all z, let

   $$e^{\epsilon_{j|z}} = \frac{Pr[M(X_i) = y_j \mid X_{i\downarrow} = z]}{Pr[M(alt(X_i)) = y_j \mid X_{i\downarrow} = z]} \ge 1. \tag{1}$$

   For each $\epsilon_{j|z}$, it has a corresponding $\delta_{j|z}$ such that

   $$\delta_{j|z} = \max_{S\subseteq\mathcal{W}}\{\Sigma_{y\in S} Pr[M(X_i) = y_j \mid X_{i\downarrow} = z] - e^{\epsilon_i}\Sigma_{y\in S} Pr[M(alt(X_i)) = y_j \mid X_{i\downarrow} = z]\} \tag{2}$$

   For each j, let $\min_z\{\epsilon_{j|z}\} = \epsilon_{j|min}$ and $\max_z\{\epsilon_{j|z}\} = \epsilon_{j|max}$. Obviously, the minimal shift $\delta_{j|min} = \min_z\{\delta_{j|z}\}$ is computed by $\epsilon_{j|max}$. The maximal shift $\delta_{j|max} = \max_z\{\delta_{j|z}\}$ is computed by $\epsilon_{j|min}$. Let $\epsilon_{max} = \max_j\{\epsilon_{j|max}\}$, which corresponds to $\delta_{min} = \min_j\{\delta_{j|min}\}$.

   Let $\mathcal{E} = \{(\epsilon_{j|min}, \delta_{j|max}), (\epsilon_{j|max}, \delta_{j|min})\}$. Then, M is $(\epsilon, \delta, \Delta, \Gamma)$-CW private is CW-private for all $(\epsilon, \delta)\in\mathcal{E}$, where Δ specifies the distribution on X_i and the auxiliary information as stated above.

Theorem 4. *Let $G(X) = (M(X_1), M(X_2), \ldots, M(X_t))$ be the composite query that runs M at each time step. Then, G is $(\epsilon_t, \delta_t, \Delta, \Gamma)$ -CW private, where $\epsilon_t \le \epsilon_{max} + (t-1)\gamma$ and $\delta_t \le t\delta_{min}$.*

Proof. The base case is obvious when $t = 1$. Next, we show the case when $t = 2$, which can be easily generalized to an inductive proof.

For any $[y_1, y_2] \in \mathcal{W}^2$, consider the following expression.

$$\frac{Pr[M(X_1) = y_1, M(X_2) = y_2 \,|\, X_{2\downarrow} = z]}{Pr[M(alt(X_1)) = y_1, M(alt(X_2)) = y_2 \,|\, X_{2\downarrow} = z]}$$

$$= \frac{\Sigma_{z'} Pr[M(X_1) = y_1 \,|\, X_{21} = z', X_{2\downarrow} = z] Pr[M(X_2) = y_2 \,|\, X_{21} = z', X_{2\downarrow} = z] Pr[X_{21} = z']}{\Sigma_{z'} Pr[M(alt(X_1)) = y_1 \,|\, X_{21} = z', X_{2\downarrow} = z] Pr[M(alt(X_2)) = y_2 \,|\, X_{21} = z', X_{2\downarrow} = z] Pr[X_{21} = z']}$$

Assume the theorem is true for $t-1$. That is, for any $[y_1, \ldots, y_{(t-1)}] \in \mathcal{W}^{t-1}$ and for all z,

$$e^{\epsilon_{t-1}} = \frac{Pr[M(X_1) = y_1, \ldots, M(X_{(t-1)}) = y_{(t-1)} \,|\, X_{t-1\downarrow} = z]}{Pr[M(alt(X_1)) = y_1, \ldots, M(alt(X_{(t-1)})) = y_{(t-1)} \,|\, X_{t-1\downarrow} = z]} \leq e^{\epsilon + (t-2)\gamma}$$

For this ϵ_{t-1}, we have

$$\max_{S \subseteq \mathcal{W}^{t-1}} \{\Sigma_{[y_1, \ldots, y_{t-1}] \in S} Pr[M(X_1) = y_1, \ldots, M(X_{t-1}) = y_{t-1} \,|\, X_{t-1\downarrow} = z]$$

$$-e^{\epsilon_{t-1}} \Sigma_{y \in S} Pr[M(alt(X_1)) = y_1, \ldots, M(alt(X_{t-1})) = y_{t-1} \,|\, X_{t-1\downarrow} = z]\} \leq \delta_{t-1}$$

Now, consider the case for t.

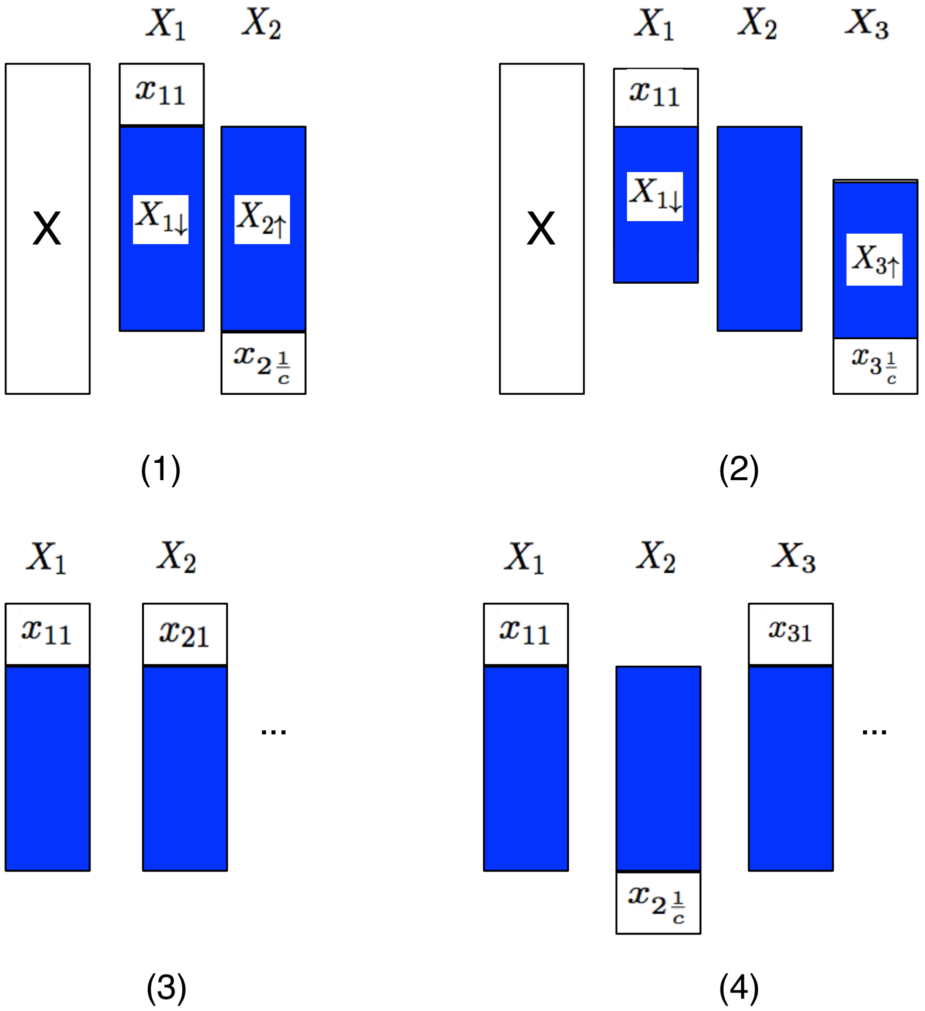

Figure 2: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. (1) Composition of $F(X_1)$ and $F(X_2)$. It has "good" independence since each X_i owns one block. (2) Composition of queries on three streaming databases. Cannot leverage "good" independence, because each block in X_2 is shared by two databases. (3) Composition of queries in the standard DP case. The blue blocks refer to the query while the white blocks refer to the independent noise. (4) It works as long as each streaming database owns one block of "private" block.