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 $S \subseteq \mathcal{Y}$:*

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

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

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 $R(M) \subseteq R(\epsilon, \delta)$.*
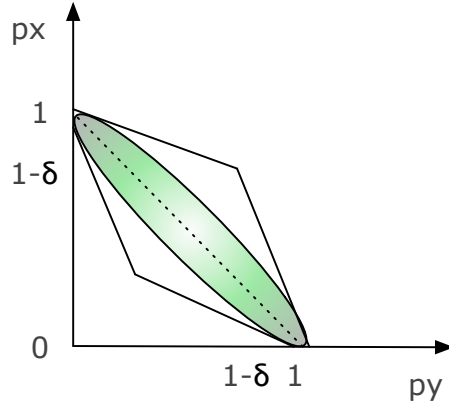

Figure 1: Every tangent line of the privacy region forms a pair of (ϵ, δ).

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 $\mathcal{E} = \{0 \leq \epsilon_i < \infty \mid Pr[M(x) = y] = e^{\epsilon_i} Pr[M(x') = y]$ 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 $S \subseteq \mathcal{Y}$:*

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

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

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 and definitions
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 $M : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$. (We assume $\mathcal{X}$ and $\mathcal{Y}$ are finite discrete for the simplicity of analysis.). M's auxiliary information consists of **the last** $(1 - c)n$ **rows** of the database X_i, i.e., $X_{i\downarrow}$ is given.

Assumptions

3

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

2. According to section 1, M can be characterized as a convex region. It should have a set of (ϵ_j, δ_j), whose values are computable by the slope and shift of the tangent lines of M on some y_j. For each $y_j \in \mathcal{W}$ and for all z, we denote the slope of the tangent line on y_j as

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

It's corresponding shift $\delta_{j|z}$ is the follows.

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

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

Let $\mathcal{E} = \{(\epsilon_{j|z}, \delta_{j|z})\}$ be the set of all slope-shift pairs. 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.

3. For each $y_j \in \mathcal{W}$, we assume the following bound.

$$\forall z, z', \frac{Pr[M(X_i) = y_j \mid X_{i\downarrow} = z]}{Pr[M(alt(X_i)) = y_j \mid X_{i\downarrow} = z']} \leq e^{\gamma_j}, \tag{3}$$

where $\gamma_j > 0$ is a small constant. We denote $\gamma = \max_j \{\gamma_j\}$.

Additional notations. For the simplicity of the proofs, we use $P^0(y_1, \ldots, y_t \mid z)$ to denote $Pr[M(X_1) = y_1, \ldots M(X_t) = y_t \mid X_{t\downarrow} = z]$, and use $P^1(y_1, \ldots, y_t \mid z)$ to denote $Pr[M(alt(X_1)) = y_1, \ldots M(alt(X_t)) = y_t \mid X_{t\downarrow} = z']$.

Lemma 1. *For any $[y_1, \ldots, y_t] \in \mathcal{W}^t$, we have the following bound*

$$\forall z, z', \frac{P^0(y_1, \ldots, y_t \mid z)}{P^1(y_1, \ldots, y_t \mid z')} \leq e^{\Sigma_{j=1}^{t} \gamma_j}$$

4

Proof. We prove this inductively. The base case is trivially true. Assume the lemma is true for $t-1$. For any z, z', conditioned on the first block of X_t (i.e., X_{t1}), we have

$$\frac{P^0(y_1, \ldots, y_t \mid z)}{P^1(y_1, \ldots, y_t \mid z')}$$

$$= \frac{\Sigma_{z*} P^0(y_1, \ldots, y_{t-1} \mid X_{t1} = z*, X_{t\downarrow} = z) P^0(y_t \mid X_{t1} = z*, X_{t\downarrow} = z) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_1, \ldots, y_{t-1} \mid X_{t1} = z*, X_{t\downarrow} = z') P^1(y_t \mid X_{t1} = z*, X_{t\downarrow} = z') Pr[X_{t1} = z*]}$$

$$= \frac{\Sigma_{z*} P^0(y_1, \ldots, y_{t-1} \mid [z*, z]_\uparrow) P^0(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_1, \ldots, y_{t-1} \mid [z*, z']_\uparrow) P^1(y_t \mid X_t = [z*, z']) Pr[X_{t1} = z*]}$$

where $[z*, z]_\uparrow$ denote the top (1/c -1) blocks of database $[z*, z]$.

$$\leq \max_{z_1, z_2}\left\{\frac{P^0(y_1, \ldots, y_{t-1} \mid z_1)}{P^1(y_1, \ldots, y_{t-1} \mid z_2)}\right\} \cdot \frac{\Sigma_{z*} P^0(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_t \mid X_t = [z*, z']) Pr[X_{t1} = z*]}$$

$$\leq e^{\Sigma_{j=1}^{t-1} \gamma_j} \cdot \frac{\Sigma_{z*} P^0(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_t \mid X_t = [z*, z']) Pr[X_{t1} = z*]}$$

$$= e^{\Sigma_{j=1}^{t-1} \gamma_j} \cdot \frac{P^0(y_t \mid z)}{P^1(y_t \mid z')} \leq e^{\Sigma_{j=1}^{t} \gamma_j} \quad \blacksquare$$

Lemma 2. *Let* $G(X) = (M(X_1), M(X_2), \ldots, M(X_t))$ *be the composite query that runs* M *at each time step. The slope of* G *on* $[y_1, \ldots, y_t] \in \mathcal{W}^t$ *has the following bound.*

$$\forall z, \frac{P^0(y_1, \ldots, y_t \mid z)}{P^1(y_1, \ldots, y_t \mid z)} \leq e^{\Sigma_{j=1}^{t-1} \gamma_j + \epsilon_{t|z}}$$

Proof. For any z, we have

$$\frac{P^0(y_1, \ldots, y_t \mid z)}{P^1(y_1, \ldots, y_t \mid z)} = \frac{\Sigma_{z*} P^0(y_1, \ldots, y_{t-1} \mid [z*, z]_\uparrow) P^0(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_1, \ldots, y_{t-1} \mid [z*, z]_\uparrow) P^1(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}$$

$$\leq \max_{z_1, z_2}\left\{\frac{P^0(y_1, \ldots, y_{t-1} \mid z_1)}{P^1(y_1, \ldots, y_{t-1} \mid z_2)}\right\} \cdot \frac{\Sigma_{z*} P^0(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}{\Sigma_{z*} P^1(y_t \mid X_t = [z*, z]) Pr[X_{t1} = z*]}$$

$$= e^{\Sigma_{j=1}^{t-1} \gamma_j} \cdot \frac{P^0(y_t \mid z)}{P^1(y_t \mid z)} = e^{\Sigma_{j=1}^{t-1} \gamma_j + \epsilon_{j|z}} \quad \blacksquare$$

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* $\epsilon_t \leq \epsilon_{max} + (t-1)\gamma$ *and* $\delta_t \leq t\delta_{max}$.

Proof. Pick ϵ_t that corresponds to the largest slope on G. By Lemma 2, it is easy to see $\epsilon_t \leq \epsilon_{max} + (t-1)\gamma$. Next, we prove shift part of the theorem inductively. Notice that when ϵ_t is the largest slope on G, we have

$$\delta_t = \max_{S \subseteq \mathcal{W}^t}\{\Sigma_{[y_1,\ldots,y_t]\in S}P^0(y_1,\ldots,y_t \mid z) - e^{\epsilon_{t|z}}\Sigma_{[y_1,\ldots,y_t]\in S}P^1(y_1,\ldots,y_t \mid z)\}$$

$$= \Sigma_{[y_1,\ldots,y_t]:P^1(y_1,\ldots,y_t \mid z)=0}P^0(y_1,\ldots,y_t \mid z)$$

$$= 1 - \Sigma_{[y_1,\ldots,y_t]\in S_t}P^0(y_1,\ldots,y_t \mid z)$$

where $S_t = \{[y_1,\ldots,y_t] \mid P^1(y_1,\ldots,y_t \mid z) \neq 0\}$. Notice that

$$P^1(y_1,\ldots,y_t \mid z) = \Sigma_{z*}P^1(y_1,\ldots,y_{t-1} \mid [z*,z]_\uparrow)P^1(y_t \mid X_t = [z*,z])Pr[X_{t1} = z*]$$

From now on, we only consider $z*$ with $Pr[X_{t1} = z*] \neq 0$.

Let $V_{t-1|z*} = \{[y_1,\ldots,y_{t-1}] \mid P^1(y_1,\ldots,y_{t-1} \mid [z*,z]_\uparrow) \neq 0\}$. Let $V_{t|z*} = \{y_t \mid P^1(y_t \mid X_t = [z*,z]) \neq 0\}$. Obviously, we have $V_{t-1|z*} \times V_{t|z*} \subseteq S_t$ for all $z*$.

We next show the lower bound of $\Sigma_{[y_1,\ldots,y_t]\in S_t}P^0(y_1,\ldots,y_t \mid z)$.

$$\Sigma_{[y_1,\ldots,y_t]\in S_t}P^0(y_1,\ldots,y_t \mid z)$$

$$= \Sigma_{[y_1,\ldots,y_t]\in S_t}\Sigma_{z^\#}P^0(y_1,\ldots,y_{t-1} \mid [z^\#,z]_\uparrow)P^0(y_t \mid X_t = [z^\#,z])Pr[X_{t1} = z^\#]$$

$$\geq \Sigma_{z^\#}\Sigma_{[y_1,\ldots,y_t]\in V_{t-1|z^\#}\times V_{t|z^\#}}P^0(y_1,\ldots,y_{t-1} \mid [z^\#,z]_\uparrow)P^0(y_t \mid X_t = [z^\#,z])Pr[X_{t1} = z^\#]$$

$$= \Sigma_{z^\#}\Sigma_{y_t\in V_{t|z^\#}}\Sigma_{[y_1,\ldots,y_{t-1}]\in V_{t-1|z^\#}}P^0(y_1,\ldots,y_{t-1} \mid [z^\#,z]_\uparrow)P^0(y_t \mid X_t = [z^\#,z])Pr[X_{t1} = z^\#]$$

By the inductive hypothesis,

$$\geq \Sigma_{z^\#}\Sigma_{y_t\in V_{t|z^\#}}(1 - (t-1)\delta_{max})P^0(y_t \mid X_t = [z^\#,z])Pr[X_{t1} = z^\#]$$

$$= (1 - (t-1)\delta_{max})\Sigma_{z^\#}\Sigma_{y_t\in V_{t|z^\#}}P^0(y_t \mid X_t = [z^\#,z])Pr[X_{t1} = z^\#]$$

Let $V_t = \{y_t \mid P^1(y_t \mid z) \neq 0\}$. Then, we have

$$= (1 - (t-1)\delta_{max})\Sigma_{y_t\in V_t}P^0(y_t \mid z)$$

Since $\Sigma_{y_t\in V_t}P^0(y_t \mid z)$ is the minimum shift value of $M(X_t)$ given $X_{t\downarrow} = z$, we have

$$\geq (1 - (t-1)\delta_{max})(1 - \delta_{z|min}) \geq 1 - t\delta_{max}.$$

It follows that

$$\delta_t = 1 - \Sigma_{[y_1,\ldots,y_t]\in S_t}P^0(y_1,\ldots,y_t \mid z) \leq t\delta_{max}. \quad \blacksquare$$
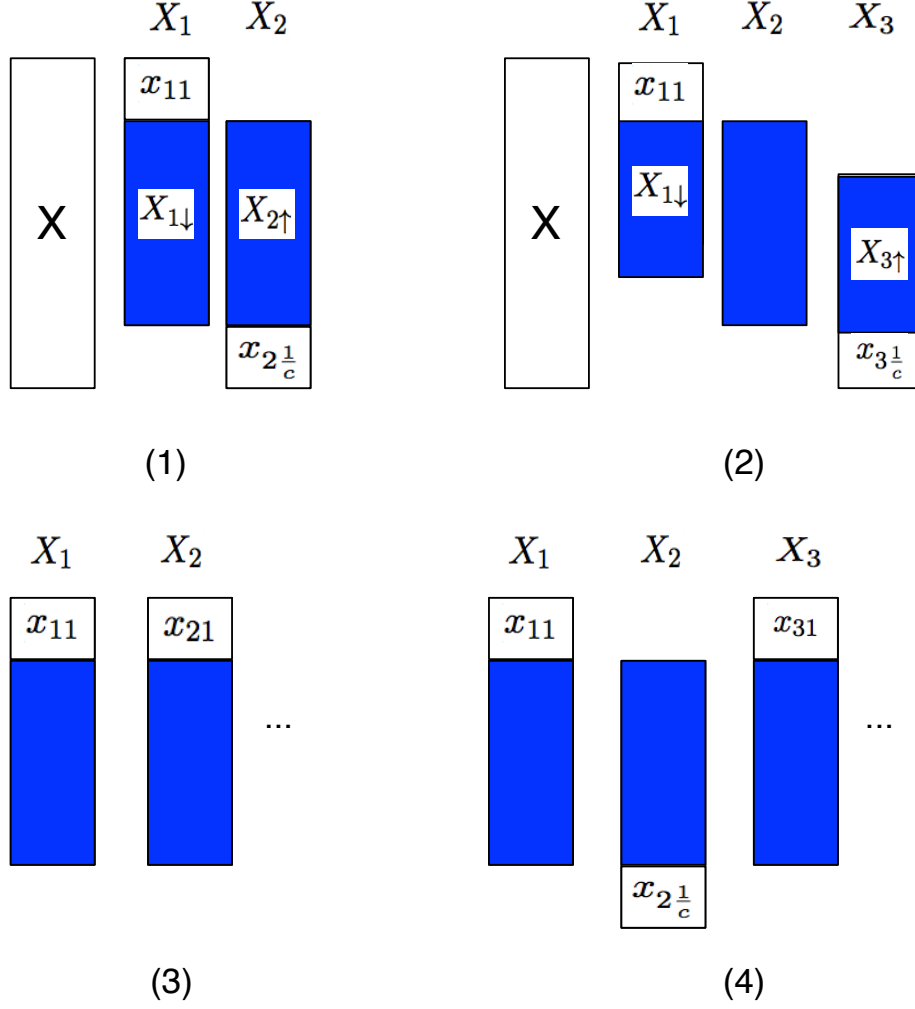
Figure 2: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. (1) Composition of $F(X_1)$ and $F(X_2)$. It has "good" independence since each X_i owns one block. (2) Composition of queries on three streaming databases. Cannot leverage "good" independence, 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.