Composition Theorem for streaming CW-pricacy

Problem setting
Let X be a database in the streaming setting. Let X_k 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_k. 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 **??**.

Notations and definitions
For each X_k of size n, we represent it by $1/c$ blocks (each has cn rows). Namely, let $X_k = [X_{k1}, X_{k2}, \ldots, X_{k\frac{1}{c}}]^\top$, where X_{kj} denote the jth block in X_k. Let $X_{k\downarrow} = [X_{k2}, \ldots, X_{k\frac{1}{c}}]^\top$ and $X_{k\uparrow} = [X_{k1}, X_{k2}, \ldots, X_{k(\frac{1}{c}-1)}]^\top$. Namely, $X_{k\downarrow}$ represent the bottom $(1/c-1)$ blocks of X_k while $X_{k\uparrow}$ represent the top $(1/c-1)$ blocks of X_k.

Assumptions

1. Each row x_j in database X follows the following i.i.d. distribution

$$f(x;p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = -1 \end{cases}$$

2. We consider sum function M.

3. We consider the case of DDP with auxiliary information. $X \sim X'$ denotes a pair of neighboring databases, such that in X' the private row x_i is always set to -1. The auxiliary information is $X_{k\downarrow}$ for each database X_k.

Preliminaries

1. Consider the distribution of $\Sigma_{j=1}^{n} x_j$, with $f_{(n;p)}$ representing its pmf. It is obvious that $f_{(n;p)}$ has integer supports $\{-n, -n+2, \ldots, n-2, n\}$. Notice that $f_{(n;p)} = y$ if and only if there are $(y+n)/2$ many $x_j = 1$.

$$f_{(n;p)}(y) = \binom{n}{(n+y)/2} p^{(n+y)/2}(1-p)^{(n-y)/2} \qquad (1)$$

2. Consider two functions $\frac{f_{(n;p)}(y+1)}{f_{(n;p)}(y-1)}$ and its reverse $\frac{f_{(n;p)}(y-1)}{f_{(n;p)}(y+1)}$, where $y \in \{-n-1, -n+1, \ldots, n-1, n+1\}$. We have $\frac{f_{(n;p)}(y+1)}{f_{(n;p)}(y-1)} = \frac{(n+1+y)p}{(n+1-y)(1-p)}$. Given $|y| \leq \beta$, both $\frac{(n+1+y)}{(n+1-y)}$ and $\frac{(n+1-y)}{(n+1+y)}$ will be $\leq \frac{(n+1+\beta)}{(n+1-\beta)}$. Let $g_{(n;p)}(y) = \log(\frac{n+1+|y|}{n+1-|y|} \cdot \max\{\frac{p}{1-p}, \frac{1-p}{p}\})$.

3. By the Chernoof's inequality, its cdf $F_{(n;p)}(\beta) = Pr[y \geq \beta]$ has the following tail bound.

$$F_{(n;p)}(\beta) \leq h_{(n;p)}(\beta) \qquad (2)$$

Theorem 1. *For any X_k and an arbitrary $\beta > 0$, $M(X_k)$ is $(\epsilon(\beta), \delta(\beta), \Delta)$-DDP, where $\epsilon(\beta) = g_{(cn;p)}(\beta)$, $\delta(\beta) = h_{(cn-1;p)}(\beta-1) + h_{(cn-1;1-p)}(\beta-1)$ and Δ has the auxiliary information of $X_{k\downarrow} = z$.*

Proof. Pick an arbitrary $\beta > 0$.

First, We show that for all y such that $|y - M(z)| \leq \beta$, $Pr[M(X_k = y) \mid X_{k\downarrow} = z] \leq e^{\epsilon(\beta)} Pr[M(X'_k = y) \mid X_{k\downarrow} = z]$.

$$\frac{Pr[M(X_k = y) \mid X_{k\downarrow} = z]}{Pr[M(X'_k = y) \mid X_{k\downarrow} = z]} = \frac{f_{(cn-1;p)}(y - M(z) - x_i)}{f_{(cn-1;p)}(y - M(z) + 1)}$$

It is trivial when $x_i = -1$. Consider the case when $x_i = 1$. Then,

$$\frac{f_{(cn-1;p)}(y - M(z) - x_i)}{f_{(cn-1;p)}(y - M(z) + 1)} = \frac{(cn - (y - M(z))(1-p)}{(cn + (y - M(z))p} \leq \frac{(cn+\beta)(1-p)}{(cn-\beta)p}$$

$$\leq e^{g_{(cn;p)}(\beta)} = e^{\epsilon(\beta)}$$

Next, we show $Pr[|y - M(z)| > \beta \mid X_{k\downarrow} = z] \leq \delta(\beta)$. Since $y - M(z) = M(X_{k1} + x_i)$, we have

$$Pr[|y - M(z)| > \beta \mid X_{k\downarrow} = z] = Pr[|M(X_{k1} + x_i)| > \beta]$$

$$\leq Pr[|M(X_{k1}| > \beta - 1] = h_{(cn-1;p)}(\beta - 1) + h_{(cn-1;1-p)}(\beta - 1)$$
$$= \delta(\beta)$$

From the above, it can be easily shown that for any set S and an arbitrary $\beta > 0$,

$$Pr[M(X_k \in S) \,|\, X_{k\downarrow} = z] \leq e^{\epsilon(\beta)} Pr[M(X'_k \in S) \,|\, X_{k\downarrow} = z] + \delta(\beta)$$

The other direction is similar. ■

Theorem 2. *For any X_1, X_2 and arbitrary $\beta_1, \beta_2 > 0$, $G(X) = (M(X_1), M(X_2))$ is $(\epsilon_2(\beta_1, \beta_2), \delta_2(\beta_1, \beta_2), \Delta)$-DDP, where $\epsilon_2(\beta_1, \beta_2) = \epsilon(\beta_1 + \beta_2)$, $\delta_2(\beta_1, \beta_2) = \max\{\delta(\beta_1 + \beta_2 + 1) + \delta(\beta_2), \delta(\beta_1 + \beta_2) + \delta(\beta_2 + 1)\}$ and Δ has the auxiliary information of $X_{2\downarrow} = z$.*

Proof. Let $z = [\hat{z}, z^{\#}]^\top$, where $\hat{z} = [X_{22}, \ldots, X_{2\frac{1}{c}-1}]^\top$ and $z^{\#} = X_{2\frac{1}{c}}$. Pick arbitrary $\beta_1, \beta_2 > 0$.

First, we show for all y_1 such that $|y_1 - M(\hat{z})| \leq \beta_1$, and all y_2 such that $|y_2 - M(z)| \leq \beta_2$, we have

$$Pr[G(X) = (y_1, y_2) \,|\, X_{2\downarrow} = z] \leq e^{\epsilon(\beta_1 + \beta_2)} Pr[G(X') = (y_1, y_2) \,|\, X_{2\downarrow} = z].$$

Consider the ratio

$$\frac{Pr[G(X) = (y_1, y_2) \,|\, X_{2\downarrow} = z]}{Pr[G(X') = (y_1, y_2) \,|\, X_{2\downarrow} = z]}$$

$$= \frac{\Sigma_{M(z^*)} Pr[M(X_1 = y_1) \,|\, z, M(z^*)] \cdot Pr[M(X_2 = y_2) \,|\, M(z^*), z] \cdot Pr[M(X_{21}) = M(z^*))]}{\Sigma_{M(z^\$)} Pr[M(X'_1 = y_1) \,|\, z, M(z^\$)] \cdot Pr[M(X'_2 = y_2) \,|\, M(z^\$), z] \cdot Pr[M(X'_{21}) = M(z^\$))]} \tag{3}$$

Consider the two cases on the position of private row x_i.

1. $x_i \in X_{21}$. The middle term $Pr[M(X_2 = y_2) \,|\, M(z^*), z] = 1$ if only if $M(z^*) = y_2 - M(z) - x_i$) (Similarly, $M(z^\$) = y_2 - M(z) + 1$)). Equation 3 can be simplified as the follows.

$$\frac{Pr[(M(X_1) = y_1) \,|\, z, M(X_{21}) = y_2 - M(z) - x_i] \cdot Pr[M(X_{21} = y_2 - M(z) - x_i)]}{Pr[(M(X'_1) = y_1) \,|\, z, M(X'_{21}) = y_2 - M(z) + 1] \cdot Pr[M(X'_{21} = y_2 - M(z) + 1)]}$$

$$= \frac{Pr[M(X_{11} = y_1 - y_2 + M(z^{\#}))] \cdot Pr[M(X_{21} = y_2 - M(z) - x_i)]}{Pr[M(X'_{11} = y_1 - y_2 + M(z^{\#}))] \cdot Pr[M(X'_{21} = y_2 - M(z) + 1)]}$$

$$= \frac{Pr[M(X_{21} = y_2 - M(z) - x_i)]}{Pr[M(X'_{21} = y_2 - M(z) + 1)]}$$

The last equality comes from the fact that $X_{11} = X'_{11}$ in this case. By the conclusion in Theorem 1, we have

$$\frac{Pr[M(X_{21} = y_2 - M(z) - x_i)]}{Pr[M(X'_{21} = y_2 - M(z) + 1)]} \le e^{\epsilon(\beta_2)}.$$

2. $x_i \in X_{11}$. The middle term $Pr[M(X_2 = y_2) \,|\, M(z^*), z] = 1$ if only if $M(z^*) = y_2 - M(z))$ (Similarly, $M(z^{\$}) = y_2 - M(z))$). Equation 3 can be simplified as the follows.

$$\frac{Pr[(M(X_1) = y_1) \,|\, z, M(X_{21}) = y_2 - M(z)] \cdot Pr[M(X_{21} = y_2 - M(z))]}{Pr[(M(X'_1) = y_1) \,|\, z, M(X'_{21}) = y_2 - M(z)] \cdot Pr[M(X'_{21} = y_2 - M(z))]}$$
$$= \frac{Pr[M(X_{11} = y_1 - y_2 + M(z^{\#}) - x_i)]}{Pr[M(X'_{11} = y_1 - y_2 + M(z^{\#}) + 1)]}$$

The first equality comes from the fact that $X_{21} = X'_{21}$ in this case. The second equality comes from the fact that $M(X_{11} + X_{21} + \hat{z} + x_i) = y_1$ (Similarly, $M(X'_{11} + X'_{21} + \hat{z} - 1) = y_1$).

Notice that $|y_1 - y_2 + M(z^{\#})| = |y_1 - M(\hat{z}) - (y_2 - M(z))| \le |y_1 - M(\hat{z})| + |y_2 - M(z))|$. Since $|y_1 - M(\hat{z})| \le \beta_1$ and $|y_2 - M(z)| \le \beta_2$, we have $|y_1 - y_2 + M(z^{\#})| \le \beta_1 + \beta_2$. By the conclusion in Theorem 1, we have

$$\frac{Pr[M(X_{11} = y_1 - y_2 + M(z^{\#}) - x_i)]}{Pr[M(X'_{11} = y_1 - y_2 + M(z^{\#}) + 1)]} \le e^{\epsilon(\beta_1 + \beta_2)}.$$

Function ϵ is monotonically increasing. Hence, we have

$$\frac{Pr[G(X) = (y_1, y_2) \,|\, X_{2\downarrow} = z]}{Pr[G(X') = (y_1, y_2) \,|\, X_{2\downarrow} = z]} \le e^{\epsilon(\beta_1 + \beta_2)}.$$

Next, we show that the probability of (y_1, y_2) *does not* fall in the "good region" is at most

$$\max\{\delta(\beta_1 + \beta_2 + 1) + \delta(\beta_2), \delta(\beta_1 + \beta_2) + \delta(\beta_2 + 1)\}.$$

Consider the probability that (y_1, y_2) *does* fall in the 'good region" is $\le \delta(\beta_1 + \beta_2) + 2\delta(\beta_2)$.

$$Pr[|y_1 - \hat{z}| \le \beta_1, |y_2 - M(z)| \le \beta_2 \,|\, X_{2\downarrow} = z] \tag{4}$$
$$= \Sigma_{M(z^*)} Pr[|y_1 - \hat{z}| \le \beta_1 \,|\, M(X_{21} = z^*), X_{2\downarrow} = z] \cdot Pr[|y_2 - M(z)| \le \beta_2 \,|\, M(X_{21}) = z^*, X_{2\downarrow} = z]$$
$$\cdot Pr[M(X_{21}) = M(z^*)]$$

Again, we discuss the two cases on the position of private row x_i.

1. $x_i \in X_{21}$. Then, the middle term $Pr[|y_2 - M(z)| \leq \beta_2 \,|\, M(X_{21}) = z^*, X_{2\downarrow} = z]$

$$= \begin{cases} 1 & \text{if } |M(z^*) + x_i| \leq \beta_2) \\ 0 & \text{otherwise} \end{cases}$$

The above summation can be simplified by only considering the terms such that $|M(X_{21}) + x_i| \leq \beta_2$. We make a little relaxation and get the following lower bound.

$$\geq \Sigma_{M(z^*):|M(z^*)|\leq \beta_2 - 1} Pr[|y_1 - \hat{z}| \leq \beta_1 \,|\, M(X_{21} = z^*), X_{2\downarrow} = z] \cdot Pr[M(X_{21}) = M(z^*)]$$

Notice that $Pr[|y_1 - \hat{z}| \leq \beta_1 \,|\, M(X_{21}) = M(z^*), X_{2\downarrow} = z] = Pr[|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1 \,|\, M(X_{21}) = M(z^*)]$. If $|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1$ and $|M(X_{21}) + x_i| \leq \beta_2$, then $|M(X_{11})| \leq \beta_1 + \beta_2$. Hence, we have the following inequality.

$$Pr[|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1 \,|\, M(X_{21}) = M(z^*)]$$

$$\geq Pr[|M(X_{11})| \leq \beta_1 + \beta_2 \,|\, M(X_{21}) = M(z^*)] = Pr[|M(X_{11})| \leq \beta_1 + \beta_2]$$

Therefore, we have the expression in equation 4

$$\geq Pr[|M(X_{11})| \leq \beta_1 + \beta_2] \cdot \Sigma_{M(z^*):|M(z^*)|\leq \beta_2 - 1} Pr[M(X_{21}) = M(z^*)]$$

$$= Pr[|M(X_{11})| \leq \beta_1 + \beta_2] \cdot Pr[|M(X_{21})| \leq \beta_2 - 1]$$

By the discussion in Theorem 1, we have

$$Pr[|M(X_{11})| \leq \beta_1 + \beta_2] = 1 - Pr[|M(X_{11})| > \beta_1 + \beta_2] \geq 1 - \delta(\beta_1 + \beta_2 + 1),$$

and that

$$Pr[|M(X_{21})| \leq \beta_2 - 1] = 1 - Pr[|M(X_{21})| > \beta_2 - 1] \geq 1 - \delta(\beta_2).$$

Hence, the probability that (y_1, y_2) falls in the “good region” is at least

$$(1 - \delta(\beta_1 + \beta_2 + 1))(1 - \delta(\beta_2)) > 1 - \delta(\beta_1 + \beta_2 + 1) - \delta(\beta_2).$$

It follows that the probability that (y_1, y_2) does not fall in the “good region” is

$$< \delta(\beta_1 + \beta_2 + 1) + \delta(\beta_2).$$

2. $x_i \in X_{11}$. Then, the middle term $Pr[|y_2 - M(z)| \leq \beta_2 \,|\, M(X_{21}) = z^*, X_{2\downarrow} = z]$

$$= \begin{cases} 1 & \text{if } |M(z^*)| \leq \beta_2) \\ 0 & \text{otherwise} \end{cases}$$

The above summation can be simplified by only considering the terms such that $|M(X_{21})| \leq \beta_2$.

$= \Sigma_{M(z^*):|M(z^*)| \leq \beta_2} Pr[|y_1 - \hat{z}| \leq \beta_1 \,|\, M(X_{21} = z^*), X_{2\downarrow} = z] \cdot Pr[M(X_{21}) = M(z^*)]$

Notice that $Pr[|y_1 - \hat{z}| \leq \beta_1 \,|\, M(X_{21}) = M(z^*), X_{2\downarrow} = z] = Pr[|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1 \,|\, M(X_{21}) = M(z^*)]$. If $|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1$ and $|M(X_{21})| \leq \beta_2$, then $|M(X_{11}) + x_i| \leq \beta_1 + \beta_2$. Hence, we have the following inequality.

$$Pr[|M(X_{11)} + M(X_{21}) + x_i| \leq \beta_1 \,|\, M(X_{21}) = M(z^*)]$$

$$\geq Pr[|M(X_{11} + x_i)| \leq \beta_1 + \beta_2 \,|\, M(X_{21}) = M(z^*)] = Pr[|M(X_{11} + x_i)| \leq \beta_1 + \beta_2]$$

Therefore, we have the expression in equation 4

$$\geq Pr[|M(X_{11} + x_i)| \leq \beta_1 + \beta_2] \cdot \Sigma_{M(z^*):|M(z^*)| \leq \beta_2} Pr[M(X_{21}) = M(z^*)]$$

$$= Pr[|M(X_{11} + x_i)| \leq \beta_1 + \beta_2] \cdot Pr[|M(X_{21})| \leq \beta_2]$$

By the discussion in Theorem 1, we have

$$Pr[|M(X_{11}) + x_i| \leq \beta_1 + \beta_2] = 1 - Pr[|M(X_{11}) + x_i| > \beta_1 + \beta_2]$$

$$\geq 1 - Pr[|M(X_{11})| > \beta_1 + \beta_2 - 1] \geq 1 - \delta(\beta_1 + \beta_2),$$

and that

$$Pr[|M(X_{21})| \leq \beta_2] = 1 - Pr[|M(X_{21})| > \beta_2] \geq 1 - \delta(\beta_2 + 1).$$

Hence, the probability that (y_1, y_2) falls in the "good region" is at least

$$(1 - \delta(\beta_1 + \beta_2))(1 - \delta(\beta_2 + 1)) > 1 - \delta(\beta_1 + \beta_2) - \delta(\beta_2 + 1).$$

It follows that the probability that (y_1, y_2) does not fall in the "good region" is

$$< \delta(\beta_1 + \beta_2) + \delta(\beta_2 + 1).$$

From the above two, it can be easily shown that for any two sets S_1, S_2 and arbitrary $\beta_1, \beta_2 > 0$,

$$Pr[G(X) \in (S_1, S_2) \,|\, X_{2\downarrow} = z] \leq e^{\epsilon(\beta_1 + \beta_2)} Pr[G(X') \in (S_1, S_2) \,|\, X_{2\downarrow} = z]$$

$$+ \max\{\delta(\beta_1 + \beta_2 + 1) + \delta(\beta_2), \delta(\beta_1 + \beta_2) + \delta(\beta_2 + 1)\}.$$

The other direction is similar. ■