

Supplementary Material for “Inference for Optimal Differential Privacy Procedures for Frequency Tables”

CHENGCHENG LI¹, NAISYIN WANG¹, AND GONGJUN XU^{*1}

¹*Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA*

The supplementary materials include additional simulation results validating the inter- and intra-table inference procedures in Section 1 and the proofs for the theoretical results developed in the main article in Section 2.

1. Additional Simulation Results

This section includes additional simulation results for validating inference procedures developed in Theorem 3 for inter-table merging and Theorem 4 for intra-table merging.

1.1 Inter-Table Merging

We seek to numerically examine the results given in Theorem 3. We consider the same setting as in the main article, $H_0 : P_1 = 0.1, P_2 = 0.1, P_3 = 0.8$, and $D_j \sim \text{Multinomial}(n_j, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$ for $j = 1, 2$ and $n_1/n_2 = 3/7$. Consider merging two tables D_1^* and D_2^* , with total counts n_1 and n_2 respectively. Let the merged private table $D^* = D_1^* + D_2^*$ and consider sample sizes: $n = n_1 + n_2 = 100, 1000$. Setting the significance level to be 0.05, we evaluate 500 empirical test statistics. For private data obtained by five DP methods, Opt, Lap, TLap, GDP and TGDP, we check whether the empirical type I errors can be controlled as indicated by the results in Theorem 3. The reported average empirical type I error rates under the “Inter-Table Merging” scenario in Table 1 are controlled fairly well at around 5% for all the mechanisms. We note that even for small sample sizes of $n_1 = 30$ and $n_2 = 70$ and a high privacy requirement of $\epsilon = 0.25$, the type I errors can be well controlled.

		$n = 100$			$n = 1000$		
		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Inter-Table Merging	Opt	0.040	0.036	0.054	0.050	0.050	0.052
	Lap	0.056	0.042	0.050	0.060	0.044	0.050
	TLap	0.024	0.026	0.030	0.060	0.044	0.050
	GDP	0.042	0.052	0.054	0.034	0.040	0.042
	TGDP	0.020	0.018	0.024	0.034	0.040	0.038
Intra-Table Merging	Opt	0.031	0.050	0.052	0.051	0.051	0.053
	Lap	0.058	0.052	0.051	0.052	0.050	0.053
	TLap	0.037	0.049	0.050	0.052	0.050	0.053
	GDP	0.051	0.051	0.052	0.050	0.050	0.051
	TGDP	0.026	0.037	0.047	0.050	0.050	0.051

Table 1: Mean empirical type I errors out of 500 simulated samples under two study scenarios across different sample sizes $n = 100, 1000$ and different privacy regimes $\epsilon = 0.25, 0.5, 0.75$. The two scenarios are “Inter-Table Merging”: the inference methods proposed for the merging circumstances considered in Section 4.1 of the main article and “Intra-Table Merging”: the inference methods proposed for the merging circumstances considered in Section 4.2 of the main article.

*Corresponding author. Email: gongjun@umich.edu

Similarly, empirical powers are evaluated and compared. The results are summarized in Figure 1 below. Overall, the testing procedures proposed in Section 4.1 of the main article for the Opt and Lap methods yield good statistical power when we merge the tables. Furthermore, we note that when sample sizes and ϵ are small ($n_1 + n_2 = 100$, $\epsilon = 0.25$), the Opt procedure yields visibly better power than the field standard Lap/TLap mechanisms. In practice, data sets with smaller sample sizes are prevalent and, more importantly, they are more prone to privacy risks and the possibility of being merged is high, and as such might require a smaller and stricter ϵ . The Opt procedure stands out in these settings.

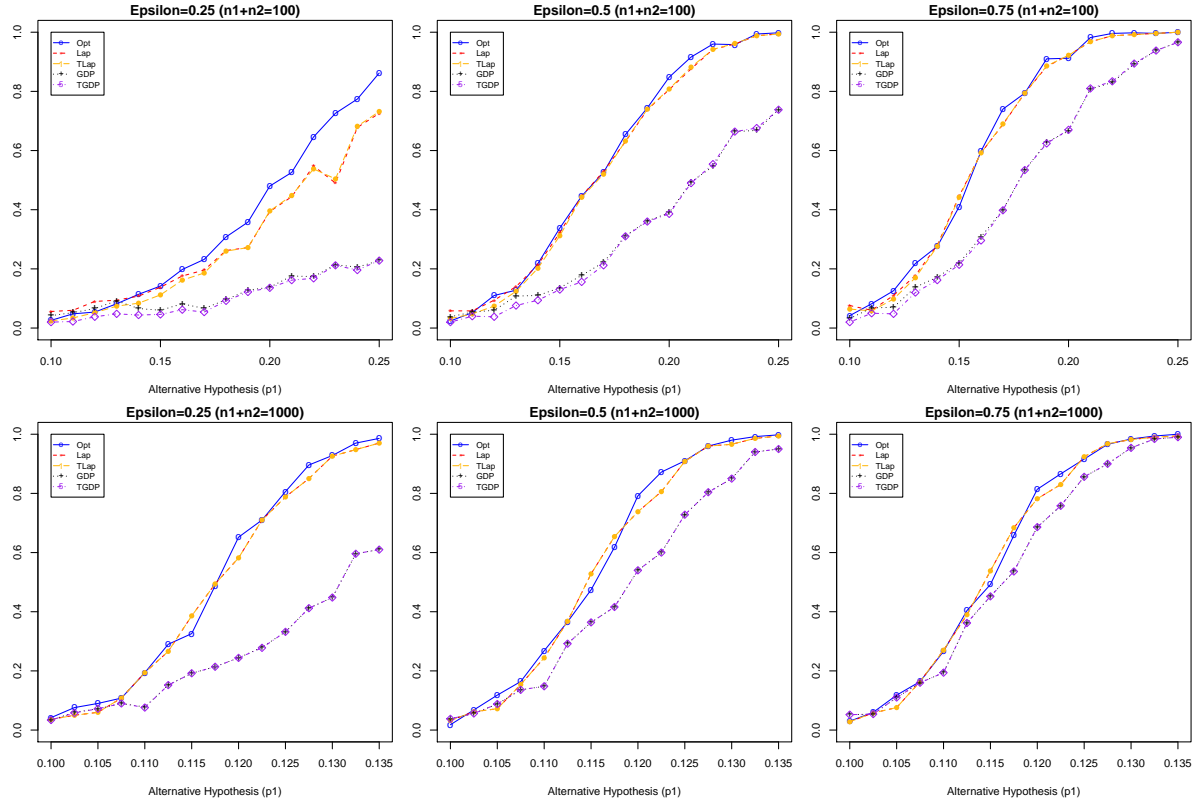


Figure 1: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP, on the inter-table merged private data sets.

1.2 Intra-Table Merging

Next, we examine the numerical performance of the procedures presented in Section 4.2 of the main article. Assuming $D \sim \text{Multinomial}(n, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$, we merge the first two cells of D^* and consider the corresponding goodness-of-fit test of $H_0 : P_{m1} = 0.2, P_{m2} = 0.8$. Setting the significance level to be 0.05 and using 500 each per generated data sets with $n = 100$ and $n = 1000$, we check whether the empirical type I errors can be controlled for the proposed procedures applying to the corresponding D^* . From the results reported in Table 1 “Intra-Table Merging” scenario, we see that the empirical type I errors are controlled fairly well at around 5% for all settings.

Empirical powers are also evaluated by considering alternative hypotheses $H_1 : P_{m1} = p_{m1} = p_1 + p_2, P_{m2} = 1 - p_{m1}$. We explore the cases with $p_1 = p_2 = 0.1, 0.11, \dots, 0.25$ for $n = 100$ and

$p_1 = p_2 = 0.1, 0.1025, \dots, 0.135$ for $n = 1000$. The results are summarized in Figure 2. We observe small improvements from the Opt procedure over the Lap/TLap procedure when the sample sizes are small at $n = 100$ and the privacy requirement is high at $\epsilon = 0.25$; and the statistical powers on private data generated from the Opt and Lap/TLap are superior to those from the GDP/TGDP.

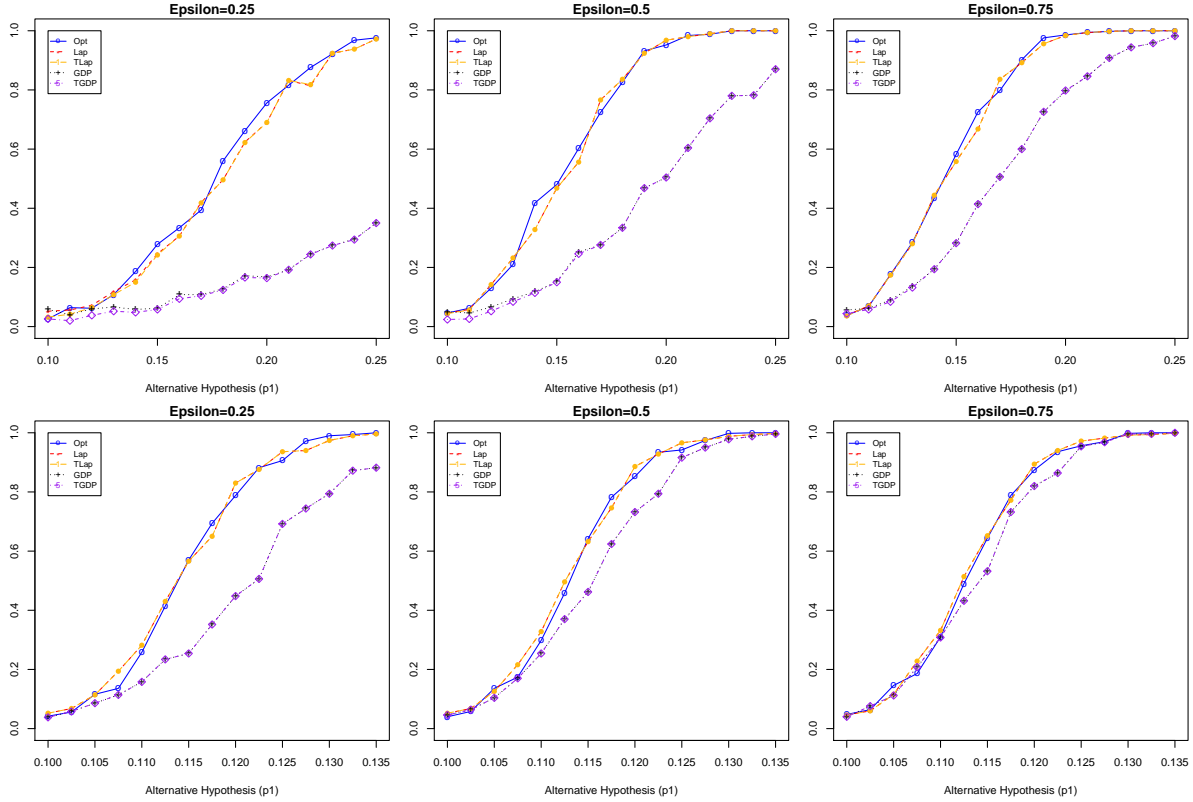


Figure 2: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP, on the intra-table merged private data sets.

2. Proofs of the Main Results

We first give four supporting lemmas that will be used in the proofs of the theorems.

Lemma 1. *For any $\epsilon_n > 0$, if $\tilde{e}_i = r - i$ for $r = 0, \dots, n$ with probability mass g_{ir} as defined in Section 3 of the main article. Then for any $i = 0, 1, \dots, n$, and for any $a > 0$,*

$$P\left(|\tilde{e}_i/\sqrt{n}| > a\right) = O\left(e^{-a\epsilon_n\sqrt{n}}\right) \quad \text{as } n \rightarrow \infty.$$

Proof. Note that in fact \tilde{e}_i has a discretized $\text{Lap}(0, 1/\epsilon_n)$ distribution truncated at 0 and n . We start with the non-truncated but discretized Laplace distribution. Let $Y \sim \text{discretized Lap}(0, 1/\epsilon_n)$ with

$$P(Y = k) = e^{-\epsilon_n|k|}/C, \quad k = \dots, -2, -1, 0, 1, 2, \dots$$

where $C = \sum_{m \in \mathbb{Z}} e^{-\epsilon_n|m|} = (1 + 2e^{-\epsilon_n}/(1 - e^{-\epsilon_n}))$. Now consider the truncated version, \tilde{e}_i , where the truncated tail probabilities are simply added to the corresponding boundaries. For

any $i = 0, \dots, n$,

$$P\left(\left|\frac{\tilde{e}_i}{\sqrt{n}}\right| > a\right) = P(|\tilde{e}_i| > a\sqrt{n}) \leq P(|Y| > a\sqrt{n}) = O\left(e^{-a\epsilon_n\sqrt{n}}\right) \quad \text{as } n \rightarrow \infty.$$

□

Lemma 2 below characterizes the L_1 distance between the remapped index and the input index under the optimal remap y introduced in Section 3 of the main article.

Lemma 2. Let $P(i \mid r') = \alpha^{|i-r'|} / (\sum_{i'=0}^n \alpha^{|i'-r'|})$ for i taking values in $\{0, 1, \dots, n\}$ with $0 < \alpha < 1$ and for some fixed $r' \in \{0, 1, \dots, n\}$. Let r^* be defined as

$$r^* = \min_k \left\{ k = 0, \dots, n : \sum_{i=0}^k \alpha^{|i-r'|} / \left(\sum_{i'=0}^n \alpha^{|i'-r'|} \right) \geq 1/2 \right\}. \quad (1)$$

Then,

$$|r' - r^*| = O\left(\max\left\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha)\right\}\right) \quad \text{as } n \rightarrow \infty.$$

Proof. Note that

$$\begin{aligned} \sum_{i'=0}^n \alpha^{|i'-r'|} &= \sum_{i'=0}^{n-r'} \alpha^{i'} + \sum_{i'=0}^{r'} \alpha^{i'} - 1 \\ &= \frac{1 - \alpha^{n-r'+1}}{1 - \alpha} + \frac{1 - \alpha^{r'+1}}{1 - \alpha} - 1 \\ &= \frac{1 + \alpha - \alpha^{n-r'+1} - \alpha^{r'+1}}{1 - \alpha}. \end{aligned}$$

Consider $\sum_{i=0}^{r^*} \alpha^{|i-r'|}$. We can have either $r^* \leq r'$ or $r^* > r'$. When $r^* \leq r'$, we have

$$\begin{aligned} \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= \sum_{i=0}^{r'} \alpha^i - \sum_{i=0}^{r'-r^*-1} \alpha^i \\ &= \frac{1 - \alpha^{r'+1}}{1 - \alpha} - \frac{1 - \alpha^{r'-r^*}}{1 - \alpha} \\ &= \frac{\alpha^{r'-r^*} - \alpha^{r'+1}}{1 - \alpha} \end{aligned}$$

Then by (1), we have for some $0 \leq c \leq 1/2$, such that

$$\begin{aligned} c &= \sum_{i=0}^{r^*} \alpha^{|i-r'|} / \left(\sum_{i'=0}^n \alpha^{|i'-r'|} \right) - \frac{1}{2}, \\ \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= (0.5 + c) \left(\sum_{i'=0}^n \alpha^{|i'-r'|} \right), \\ \alpha^{r'-r^*} - \alpha^{r'+1} &= (0.5 + c)(1 + \alpha - \alpha^{n-r'+1} - \alpha^{r'+1}), \\ |r' - r^*| &= \left| \log \left\{ (0.5 + c)(1 + \alpha - \alpha^{n-r'+1} - \alpha^{r'+1}) + \alpha^{r'+1} \right\} / \log \alpha \right|. \end{aligned} \quad (2)$$

Now we discuss the order of (2). As $n \rightarrow \infty$,

$$(2) = \begin{cases} O\left(\log\{0.5(1+\alpha)\}/\log(\alpha)\right) & \text{if } r' = \omega(1) \text{ and } r' = o(n) \\ O\left(\log\{0.5(1+\alpha-\alpha^{r'+1})+\alpha^{r'+1}\}/\log(\alpha)\right) & \text{if } r' = O(1) \\ O\left(\log\{0.5(1+\alpha-\alpha^{n-r'+1})\}/\log(\alpha)\right) & \text{if } r' = \Theta(n). \end{cases}$$

$$= O\left(\log(0.5)/\log(\alpha)\right).$$

Note that $c \rightarrow 0$ as $n \rightarrow \infty$. Now we consider $r^* > r'$,

$$\begin{aligned} \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= \sum_{i=0}^{r'} \alpha^i + \sum_{i=0}^{r^*-r'} \alpha^i - 1. \\ &= \frac{1-\alpha^{r'+1}}{1-\alpha} + \frac{1-\alpha^{r^*-r'+1}}{1-\alpha} - 1 \\ &= \frac{1+\alpha-\alpha^{r'+1}-\alpha^{r^*-r'+1}}{1-\alpha}. \end{aligned}$$

Also from (1), we have

$$\begin{aligned} \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= (0.5+c) \left(\sum_{i'=0}^n \alpha^{|i'-r'|} \right) \\ 1+\alpha-\alpha^{r'+1}-\alpha^{r^*-r'+1} &= (0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}) \\ \alpha^{r^*-r'+1} &= -(0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}) + 1+\alpha-\alpha^{r'+1} \\ \alpha^{r^*-r'} &= -(0.5+c)(1+\alpha^{-1}-\alpha^{n-r'}-\alpha^{r'}) + 1+\alpha^{-1}-\alpha^{r'} \\ |r'-r^*| &= \left| \log \left\{ -(0.5+c)(1+\alpha^{-1}-\alpha^{n-r'}-\alpha^{r'}) + 1+\alpha^{-1}-\alpha^{r'} \right\} / \log \alpha \right|. \end{aligned} \tag{3}$$

Now we discuss the order of (3). As $n \rightarrow \infty$,

$$(3) = \begin{cases} O\left(\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1})/\log(\alpha)\right) & \text{if } r' = \omega(1) \text{ and } r' = o(n) \\ O\left(\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1} - \frac{1}{2}\alpha^{r'})/\log(\alpha)\right) & \text{if } r' = O(1) \\ O\left(\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1} - \frac{1}{2}\alpha^{n-r'})/\log(\alpha)\right) & \text{if } r' = \Theta(n). \end{cases}$$

$$= O\left(-\log(1+0.5\alpha^{-1})/\log(\alpha)\right).$$

Hence, $|r'-r^*| = O\left(\max\left\{-\log(1+1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\right\}\right)$ as $n \rightarrow \infty$. \square

Lemma 3. Assume X_k^* and X_{Tk}^* are from the ϵ_n -DP Laplace and the truncated ϵ_n -DP Laplace (at zero) mechanisms respectively, with the same underlying $X_k \sim \text{Bin}(n, p_k)$ for $p_k \in (0, 1)$. Then for any $0 < \Delta_k < p_k$,

$$P(X_k^* \neq X_{Tk}^*) = O\left(e^{n(\Delta_k - p_k)\epsilon_n} + e^{-2n\Delta_k}\right), \quad \text{as } n \rightarrow \infty.$$

Proof. Note that

$$\begin{aligned} X_k^* &= X_k + err_k, \\ X_{T_k}^* &= X_k + err_k + b_k, \end{aligned}$$

where $err_k \sim \text{Lap}(0, 1/\epsilon_n)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (4)$$

$$(5)$$

Note that for any $0 < \Delta_k < p_k$, we have

$$\begin{aligned} P(X_k^* \neq X_{T_k}^*) &= P(err_k < -X_k) \\ &= P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} > \Delta_k) \cdot P(\frac{|X_k - np_k|}{n} > \Delta_k) \\ &\quad + P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) \cdot P(\frac{|X_k - np_k|}{n} \leq \Delta_k) \\ &\leq 2P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} > \Delta_k) \cdot e^{-2n\Delta_k} + P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) \cdot 1 \\ &= P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) + O(e^{-2n\Delta_k}) \\ &= P(err_k < -X_k \mid np_k - n\Delta_k \leq X_k \leq np_k + n\Delta_k) + O(e^{-2n\Delta_k}). \end{aligned} \quad (6)$$

where (6) follows from Hoeffding's inequality. Further note that

$$\begin{aligned} P(err_k < -X_k \mid np_k - n\Delta_k \leq X_k \leq np_k + n\Delta_k) &\leq P(err_k < n(\Delta_k - p_k)) \\ &= \frac{1}{2} e^{n(\Delta_k - p_k)\epsilon_n} \\ &= O(e^{n(\Delta_k - p_k)\epsilon_n}). \end{aligned}$$

Hence, we have

$$P(X_k^* \neq X_{T_k}^*) = O(e^{n(\Delta_k - p_k)\epsilon_n} + e^{-2n\Delta_k}).$$

□

Lemma 4. Assume X_k^* and $X_{T_k}^*$ are from the (ϵ_n, δ) -DP Gaussian and the truncated (ϵ_n, δ) -DP Gaussian (at zero) mechanisms respectively, with the same underlying $X_k \sim \text{Bin}(n, p_k)$ for $p_k \in (0, 1)$. Then for any $0 < \Delta_k < p_k$,

$$P(X_k^* \neq X_{T_k}^*) = O\left(\exp\left\{-2n\Delta_k\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_k)^2}{4\ln\{1.25/\delta\}}\right\}\right).$$

Proof. Note that

$$\begin{aligned} X_k^* &= X_k + err_k, \\ X_{T_k}^* &= X_k + err_k + b_k, \end{aligned}$$

where $err_k \sim N(0, 2\ln\{1.25/\delta\}/\epsilon_n^2)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (7)$$

$$(8)$$

Note that for any $Y \sim N(0, \sigma^2)$, we have for any $t \in (0, \infty)$,

$$P(Y > t) \leq \frac{\sigma}{t\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}. \quad (9)$$

To see this, note

$$\begin{aligned} P(Y > t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx \\ &\leq \frac{1}{t\sqrt{2\pi}\sigma^2} \int_t^\infty x e^{-\frac{x^2}{2\sigma^2}} dx \quad \text{since } \frac{x}{t} \geq 1 \text{ for } x \in [t, \infty) \\ &= \frac{1}{t\sqrt{2\pi}\sigma^2} \left[-\sigma^2 e^{-\frac{x^2}{2\sigma^2}} \right]_t^\infty \\ &= \frac{\sigma}{t\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}. \end{aligned}$$

Further note that for any $0 < \Delta_k < p_k$, we have

$$\begin{aligned} P(X_k^* \neq X_{Tk}^*) &= P(err_k < -X_k) \\ &= P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} > \Delta_k) \cdot P(\frac{|X_k - np_k|}{n} > \Delta_k) \\ &\quad + P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) \cdot P(\frac{|X_k - np_k|}{n} \leq \Delta_k) \\ &\leq 2P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} > \Delta_k) \cdot e^{-2n\Delta_k} + P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) \cdot 1 \\ &= P(err_k < -X_k \mid \frac{|X_k - np_k|}{n} \leq \Delta_k) + O(e^{-2n\Delta_k}) \\ &= P(err_k < -X_k \mid np_k - n\Delta_k \leq X_k \leq np_k + n\Delta_k) + O(e^{-2n\Delta_k}). \end{aligned}$$

where the inequality in the third step can be obtained from Hoeffding's inequality. By applying (9), we have

$$\begin{aligned} P(err_k < -X_k \mid np_k - n\Delta_k \leq X_k \leq np_k + n\Delta_k) &\leq P(err_k < n(\Delta_k - p_k)) \\ &\leq \frac{\sqrt{\ln\{1.25/\delta\}}}{\sqrt{\pi}n\epsilon_n(p_k - \Delta_k)} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_k)^2}{4\ln\{1.25/\delta\}}\right\} \\ &= O\left(\frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_k)^2}{4\ln\{1.25/\delta\}}\right\}\right). \end{aligned}$$

Hence, we have

$$P(X_k^* \neq X_{Tk}^*) = O\left(\exp\left\{-2n\Delta_k\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_k)^2}{4\ln\{1.25/\delta\}}\right\}\right).$$

□

Next we give a proposition specifying the rate of convergence of root- n -scaled random errors injected by the optimal mechanism with L_1 loss.

Proposition 1. *For $0 < \epsilon_n < \infty$ and some fixed $i \in \{0, 1, \dots, n\}$, let $err_i = r - i$ for r taking values in $\{0, \dots, n\}$ with probability mass p_{ir}^* as defined in Section 3 of the main article. Then for any $a > 0$,*

$$P\left(\left|err_i/\sqrt{n}\right| > a\right) = O\left(\max\{1 + e^{\epsilon_n}/2, 2\}e^{-\epsilon_n a \sqrt{n}}\right) \quad \text{as } n \rightarrow \infty.$$

Proof. Consider

$$\begin{aligned} P\left(\left|\frac{err_i}{\sqrt{n}}\right| > a\right) &= P(|err_i| > a\sqrt{n}) = \sum_{r: |r-i| > a\sqrt{n}, r \in \{0, \dots, n\}} p_{ir}^* \\ &= \sum_{r: |r-i| > a\sqrt{n}, r \in \{0, \dots, n\}} \left(\sum_{r'=0}^n g_{ir'} y_{r'r} \right) \\ &= \sum_{r'=0}^n g_{ir'} \left(\sum_{r: |r-i| > a\sqrt{n}, r \in \{0, \dots, n\}} y_{r'r} \right), \end{aligned} \quad (10)$$

where $y_{r'r}$ is the (r', r) entry of the optimal remap as defined in Section 3 of the main article. Note that from the computation of the optimal remap y , for any given $r \in \{0, \dots, n\}$, $y_{r,r^*} = 1$ if r^* equals the conditional median of $i \in \{0, \dots, n\}$ with probability mass $P(i \mid r')$, and $y_{r,r'} = 0$ for all $r' \neq r^*$. Write $\alpha = e^{-\epsilon_n}$. Note that for any given $r' \in \{0, 1, \dots, n\}$,

$$P(i \mid r') = \frac{\alpha^{|i-r'|}}{\sum_{i'=0}^n \alpha^{|i'-r'|}}.$$

By Lemma 2, we know that $|r' - r^*| = O\left(\max\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\}\right)$.

$$\begin{aligned} |r^* - i| &= |r^* - r' + r' - i| \leq |r^* - r'| + |r' - i| \\ &\leq |r' - i| + O\left(\max\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\}\right) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (11)$$

Note that there exists n' such that for all $n > n'$, $|r^* - r'| \leq C$, where $C = \max\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\}$. Also note that whenever $|r' - i| \leq a\sqrt{n} - C$, we have $|r^* - i| \leq a\sqrt{n}$ for $n > n'$. It follows that for all $n > n'$,

$$\left(\sum_{r: |r-i| > a\sqrt{n}, r \in \{0, \dots, n\}} y_{r'r} \right) = 0.$$

Hence for any $n > n'$,

$$(10) = \sum_{r': |r' - i| > a\sqrt{n} - C} g_{ir'} \leq 2P(Y > a\sqrt{n} - C) = O(e^{-a\epsilon_n \sqrt{n} + C\epsilon_n}),$$

where Y has discretized $\text{Lap}(0, 1/\epsilon_n)$ distribution. Hence, it follows that

$$P\left(\left|\frac{err_i}{\sqrt{n}}\right| > a\right) = O\left(e^{\epsilon_n(C - a\sqrt{n})}\right) \quad \text{as } n \rightarrow \infty. \quad (12)$$

Since $\alpha = e^{-\epsilon_n}$, further note that

$$C = \max \left\{ -\log(1 + 1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha) \right\} = \max \left\{ \frac{\log(1 + 1/2e^{\epsilon_n})}{\epsilon_n}, \frac{\log(2)}{\epsilon_n} \right\}.$$

Substitute into (12), we have

$$P\left(\left|\frac{err_i}{\sqrt{n}}\right| > a\right) = O\left(e^{\epsilon_n \left(\max \left\{ \frac{\log(1+1/2e^{\epsilon_n})}{\epsilon_n}, \frac{\log(2)}{\epsilon_n} \right\} - a\sqrt{n}\right)}\right) = O\left(\max \left\{ 1 + \frac{e^{\epsilon_n}}{2}, 2 \right\} e^{-\epsilon_n a\sqrt{n}}\right) \text{ as } n \rightarrow \infty.$$

□

Next, we give proofs for the theorems. We start with proving Theorem 2 first.

Proof of Theorem 2. Note that

$$T_k = \begin{cases} \frac{1}{\sqrt{np_k}}(X_k^* - np_k) = \sqrt{\frac{n}{p_k}}\left(\frac{X_k - np_k}{n}\right) + \frac{err_k}{\sqrt{np_k}} & \text{if } X_k^* \text{ from Lap,} \\ \frac{1}{\sqrt{np_k}}(X_k^* - np_k) = \sqrt{\frac{n}{p_k}}\left(\frac{X_k - np_k}{n}\right) + \frac{err_k}{\sqrt{np_k}} + \frac{b_k}{\sqrt{np_k}} & \text{if } X_k^* \text{ from TLap,} \end{cases} \quad (13)$$

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (14)$$

where $err_k \sim \text{Lap}(0, 1/\epsilon_n)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (15)$$

First, we seek to show that $P(\sum_k (13)^2 \neq \sum_k (14)^2) = o(1)$ as $n \rightarrow \infty$, so that we can ignore the bias term b_k in the test statistic in the remaining proofs. Note that for any $0 < \Delta_k < p_k$, by Lemma 3, we have

$$P(b_k \neq 0) = O\left(e^{n(\Delta_k - p_k)\epsilon_n} + e^{-2n\Delta_k}\right). \quad (17)$$

Now take $\Delta_k = \min\{p_1, \dots, p_K\}/2$. Further note that

$$\begin{aligned} P(\sum_k (13)^2 \neq \sum_k (14)^2) &= P(\cup_k \{b_k \neq 0\}) \\ &\leq \sum_{k=1}^K P(b_k \neq 0) \\ &= O\left(\exp\left\{-\frac{1}{2}n\epsilon_n \min\{p_1, \dots, p_K\}\right\} + \exp\{-n \min\{p_1, \dots, p_K\}\}\right). \end{aligned}$$

where the last step follows directly from (17) and the fact that $K < \infty$. Since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, it follows $P(\sum_k (13)^2 \neq \sum_k (14)^2) = o(1)$ as $n \rightarrow \infty$. Therefore, we can ignore the bias term b_k in the following proofs. Note that marginally, $X_k \sim \text{Binomial}(n, p_k)$. By the Central Limit Theorem (CLT), we know that $(X_k - np_k)/n$ will converge to a Gaussian variable as $n \rightarrow \infty$ since X_k can be viewed as a sum of n i.i.d. $\text{Ber}(p_k)$ under H_0 . As a direct consequence of Lemma 1, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, the second term in (13), $err_k/\sqrt{np_k} \xrightarrow{P} 0$ as $n \rightarrow \infty$. Therefore, overall, T_k will converge to a Gaussian distribution as $n \rightarrow \infty$. However, instead of treating err_k/\sqrt{n} as 0, we take its first

and second moments into account to have better finite sample approximations while maintaining correct asymptotic distribution. Note $err_k \sim \text{Lap}(0, 1/\epsilon_n)$, we have $\mathbb{E}[T_k] = 0$ and

$$\begin{aligned} \text{Var}(T_k) &= \frac{1}{np_k} \text{Var}(X_k^*) = \frac{1}{np_k} \{ \text{Var}(X_k) + \text{Var}(err_k) \} \\ &= 1 - p_k + 2/(np_k \epsilon_n^2). \end{aligned}$$

Note also that correlations amongst X_k induce correlations amongst T_k . For some $k \neq j$, consider,

$$\begin{aligned} \text{Cov}(T_k, T_j) &= \text{Cov}\left(\sqrt{\frac{n}{p_k}}\left(\frac{X_k^*}{n} - p_k\right), \sqrt{\frac{n}{p_j}}\left(\frac{X_j^*}{n} - p_j\right)\right) = \frac{1}{n\sqrt{p_k p_j}} \text{Cov}(X_k + err_k, X_j + err_j) \\ &= \frac{1}{n\sqrt{p_k p_j}} \text{Cov}(X_k, X_j) \\ &= -\sqrt{p_k p_j}. \end{aligned}$$

Let Σ be the covariance matrix of $\mathbf{T} = (T_1, T_2, \dots, T_K)$. Consider a matrix $O = [v_1, \dots, v_K] \in \mathbb{R}^{K \times K}$ consisting of the orthonormal eigenvectors of Σ as columns. So we must have $\Sigma O = \Lambda O$, where Λ is a diagonal matrix with diagonal elements Λ_k being the eigenvalues of Σ with respect to v_k . We require $\|v_k\| = 1$ for all $k = 1, \dots, K$. So we must have $OO^T = O^T O = I_K$. Consider transformed vector $\mathbf{T}' = O\mathbf{T} = (T'_1, \dots, T'_K)$. First note that each T'_k is asymptotically normal since it is a linear combination of normal distributions. Further, we also have $\text{Cov}(\mathbf{T}') = \text{Cov}(O\mathbf{T}) = \Lambda$. Hence T'_k are independent $N(0, \Lambda_k)$ for $k = 1, \dots, K$. So, $T^* = \sum_{k=1}^K T_k^2 = \mathbf{T}^T \mathbf{T} = \mathbf{T}'^T \mathbf{T}' \rightarrow \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1.

Now for part (b) of the theorem, again we can express

$$T_k = \begin{cases} \frac{1}{\sqrt{np_k}}(X_k^* - np_k) = \sqrt{\frac{n}{p_k}}\left(\frac{X_k - np_k}{n}\right) + \frac{err_k}{\sqrt{np_k}} & \text{if } X_k^* \text{ from GDP,} \\ \frac{1}{\sqrt{np_k}}(X_k^* - np_k) = \sqrt{\frac{n}{p_k}}\left(\frac{X_k - np_k}{n}\right) + \frac{err_k}{\sqrt{np_k}} + \frac{b_k}{\sqrt{np_k}} & \text{if } X_k^* \text{ from TGDP,} \end{cases} \quad (18)$$

where $err_k \sim N(0, 2 \ln\{1.25/\delta\}/\epsilon_n^2)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (20)$$

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -X_k, \\ -(X_k + err_k), & \text{otherwise.} \end{cases} \quad (21)$$

Similarly, we seek to show that $P(\sum_k (18)^2 \neq \sum_k (19)^2) = o(1)$ as $n \rightarrow \infty$, so that we can ignore the bias term b_k in the test statistic in the remaining proofs. Note that for any $0 < \Delta_k < p_k$, by Lemma 4, we have

$$P(b_k \neq 0) = O\left(\exp\left\{-2n\Delta_k\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_k)^2}{4\ln\{1.25/\delta\}}\right\}\right). \quad (22)$$

Again we can take $\Delta_k = \min\{p_1, \dots, p_K\}/2$. Further note that

$$\begin{aligned} P\left(\sum_k (18)^2 \neq \sum_k (19)^2\right) &= P(\cup_k \{b_k \neq 0\}) \\ &\leq \sum_k P(b_k \neq 0) \\ &= O\left(\exp\left\{-n \min\{p_1, \dots, p_K\}\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2 \min\{p_1, \dots, p_K\}^2}{16\ln\{1.25/\delta\}}\right\}\right), \end{aligned}$$

where the last step follows directly from (22) and $K < \infty$. Again, since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, it follows $P(\sum_k (18)^2 \neq \sum_k (19)^2) = o(1)$ as $n \rightarrow \infty$. Therefore, we can ignore the bias term b_k in the test statistic. The remaining proof is similar to the Laplace case, except for the difference in the covariance Σ matrix. Note that in the case of (ϵ_n, δ) -Gaussian mechanism,

$$\begin{aligned} \text{Var}(T_k) &= \frac{1}{np_k} \left(\text{Var}(X_k) + \text{Var}(\text{err}_k) \right) \\ &= \frac{1}{np_k} \left(np_k(1 - p_k) + (2 \log(1.25/\delta) - 1)/\epsilon_n^2 \right) \\ &= 1 - p_k + \{2 \log(1.25/\delta) - 1\} / \{np_k \epsilon_n^2\}, \end{aligned}$$

and $\text{Cov}(T_k, T_j) = -\sqrt{p_k p_j}$. Again, by considering the orthonormal eigenvectors and corresponding eigenvalues Λ_k of Σ , we can arrive at the same conclusion that as $n \rightarrow \infty$, $T^* \rightarrow \sum_{k=1}^K \Lambda_k Z_k$. Therefore, the results of the theorem follow. \square

Proof of Theorem 1. The proof is similar as the proof of Theorem 2 except that there is an additional de-bias term in the test statistic and err_k has probability mass p_k^* in this case.

We focus on T'_k first,

$$T'_k = \frac{1}{\sqrt{np_k}} (X_k^* - np_k - b(x_k^*)) = \sqrt{\frac{n}{p_k}} \left(\frac{X_k - np_k}{n} \right) + \frac{\text{err}_k - b(x_k^*)}{\sqrt{np_k}}. \quad (23)$$

Note that by CLT, we know that $(X_k - np_k)/n$ will converge to a Gaussian variable as $n \rightarrow \infty$. As a direct consequence of Proposition 1 and the fact that $b(x_k^*) = O(1)$, the second term in (23) converges in probability to 0. Therefore, T_k converges in distribution to a normal random variable as $n \rightarrow \infty$. We just need to derive its asymptotic mean and variance to pinpoint its distribution. Note that $\mathbb{E}[T'_k] \rightarrow 0$. Before giving an estimate for the variance, we first derive the order of $\text{Var}(\text{err}_k)$. Let $\alpha_n = e^{-\epsilon_n}$. Note that $0 < \alpha_n < 1$. First consider $\tilde{e}_k = j - k$ for $j = 0, 1, \dots, n$ with probability mass $g_{kj} = \frac{1-\alpha_n}{1+\alpha_n} \alpha_n^{|j-k|}$ for $j = 1, \dots, (n-1)$, $g_{kj} = \alpha_n^{|j-k|}/(1+\alpha_n)$ for $j = 0, n$, with some fixed $k \in \{0, \dots, n\}$. Note that

$$\begin{aligned} \tilde{\mu}_k &:= \mathbb{E}[\tilde{e}_k] = \sum_{j=0}^n g_{kj} (j - k) = \frac{1 - \alpha_n}{1 + \alpha_n} \sum_{j=1}^{n-1} \alpha_n^{|j-k|} (j - k) + \frac{\alpha_n^{|n-k|}}{1 + \alpha_n} (n - k) - \frac{k \alpha_n^k}{1 + \alpha_n} \\ &= O(1) \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Consider variance of \tilde{e}_k ,

$$\begin{aligned} \text{Var}(\tilde{e}_k) &= \sum_{j=0}^n g_{kj} (j - k - \tilde{\mu}_k)^2 = \frac{1 - \alpha_n}{1 + \alpha_n} \sum_{j=1}^{n-1} \alpha_n^{|j-k|} (j - k)^2 - 2\tilde{\mu}_k^2 + \frac{1 - \alpha_n}{1 + \alpha_n} \sum_{j=1}^{n-1} \alpha_n^{|j-k|} \tilde{\mu}_k^2 \\ &\quad + \frac{k^2 \alpha_n^k}{1 + \alpha_n} + \frac{\alpha_n^{|n-k|}}{1 + \alpha_n} (n - k)^2 + \frac{\tilde{\mu}_k^2 \alpha_n^k}{1 + \alpha_n} + \frac{\tilde{\mu}_k^2 \alpha_n^{n-k}}{1 + \alpha_n} \\ &= O(1) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (24)$$

Now denote $\mu_k := \mathbb{E}[\text{err}_k]$. As a direct consequence of Lemma 2, there exists a constant $C_1 < \infty$ such that $|\tilde{\mu}_k - \mu_k| < C_1$ and $C_2 < \infty$ such that the input index r' and optimally remapped

index r^* satisfying $|r' - r^*| < C_2$. Then,

$$\begin{aligned}
\text{Var}(\text{err}_k) &= \sum_{j=0}^n p_{kj}^* (j - k - \mu_k)^2 \\
&\leq \frac{1 - \alpha_n}{1 + \alpha_n} \sum_{j=1}^{n-1} \alpha_n^{C_2+|j-k|} (C_1 + C_2 + j - k - \tilde{\mu}_k)^2 \\
&\quad + \frac{\alpha_n^{C_2+k}}{1 + \alpha_n} (C_1 + C_2 + j - k - \tilde{\mu}_k)^2 + \frac{\alpha_n^{C_2+n-k}}{1 + \alpha_n} (C_1 + C_2 + j - k - \tilde{\mu}_k)^2 \\
&= \alpha_n^{C_2} \sum_{j=0}^n g_{kj} \left\{ (C_1 + C_2)^2 + 2(C_1 + C_2)(j - k - \tilde{\mu}_k) + (j - k - \tilde{\mu}_k)^2 \right\} \\
&= \alpha_n^{C_2} \left\{ (C_1 + C_2)^2 + 2(C_1 + C_2)(\tilde{\mu}_k - \tilde{\mu}_k) + \text{Var}(\tilde{e}_k) \right\} \\
&= O(1) \quad \text{as } n \rightarrow \infty,
\end{aligned} \tag{25}$$

where the last step follows from (24). Now we can evaluate the variance term.

$$\begin{aligned}
\text{Var}(T'_k) &= \frac{1}{np_k} \text{Var}(X_k^*) = \frac{1}{np_k} (\text{Var}(X_k) + \text{Var}(\text{err}_k)) \\
&= 1 - p_k + \text{Var}(\text{err}_k)/(np_k) \\
&= 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty.
\end{aligned} \tag{26}$$

The last step follows directly from (25). Note also that

$$1 - p_k + v(x_k^*)/(np_k) = 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \tag{27}$$

We can see this from the construction of the variance estimate $v(x_k^*)$ in Algorithm 3 in the main article. First note that in the step 2 of Algorithm 3, $v_i = \text{Var}(\text{err}_i) = O(1)$. $\{f_{ix_k^*} : i = 1, \dots, n\}$ is a probability distribution such that $\sum_{i=1}^n f_{ix_k^*} = 1$. Hence, $v(x_k^*) = \sum_{i=1}^n f_{ix_k^*} v_i = O(1)$ as $n \rightarrow \infty$. From Equations (26) and (27), we have

$$\left| \text{Var}(T'_k) - (1 - p_k + v(x_k^*)/(np_k)) \right| = O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty.$$

Lastly, since X_k are correlated, T'_k are correlated. For any $k \neq j$, consider,

$$\begin{aligned}
\text{Cov}(T'_k, T'_j) &= \text{Cov}\left(\sqrt{\frac{n}{p_k}} \left(\frac{X_k^*}{n} - p_k\right), \sqrt{\frac{n}{p_j}} \left(\frac{X_j^*}{n} - p_j\right)\right) = \frac{1}{n\sqrt{p_k p_j}} \text{Cov}(X_k + \text{err}_k, X_j + \text{err}_j) \\
&= \frac{1}{n\sqrt{p_k p_j}} \text{Cov}(X_k, X_j) \\
&= -\sqrt{p_k p_j}.
\end{aligned}$$

Let $\Sigma \in \mathbb{R}^{K \times K}$ be a matrix with diagonal entries $\Sigma_{kk} = 1 - p_k + v(x_k^*)/(np_k)$ and off-diagonal entries $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\mathbf{T}' = (T'_1, T'_2, \dots, T'_K)$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \rightarrow \infty$. Consider a matrix $O = [v_1, \dots, v_K] \in \mathbb{R}^{K \times K}$ consisting of the orthonormal eigenvectors of Σ as columns. So we must have $\Sigma O = \Lambda O$, where Λ is a

diagonal matrix with diagonal elements Λ_k being the eigenvalues of Σ with respect to v_k . We require $\|v_k\| = 1$ for all $k = 1, \dots, K$. Following a similar argument as in the proof of Theorem 2, we can derive that $T_{opt}^* \rightarrow \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1. \square

Proof of Theorem 3. We start with proving part (a). Consider T_{mk} ,

$$T_{mk} = \frac{1}{\sqrt{np_k}} \left(\sum_{j=1}^C X_{jk}^* - np_k - b_M(\{x_{jk}^*\}_{j=1}^C) \right) = \sqrt{\frac{n}{p_k}} \left(\frac{\sum_{j=1}^C X_{jk} - np_k}{n} \right) + \frac{\sum_{j=1}^C (err_{jk} - b(x_{jk}^*))}{\sqrt{np_k}}. \quad (28)$$

Assume H_0 is true, by CLT, we know that $(\sum_{j=1}^C X_{jk} - np_k)/n$ will converge to a Gaussian distribution with mean 0 as $n \rightarrow \infty$ since $\sum_{j=1}^C X_{jk}$ can be viewed as a sum of $n = n_1 + n_2 + \dots + n_C$ i.i.d. $\text{Ber}(p_k)$ random variables. As a direct consequence of Proposition 1 and the facts that $b(x_{jk}^*), C < \infty$, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, the second term of (28) will converge in probability to 0 as $n \rightarrow \infty$. Therefore, overall, T_{mk} will converge to a Gaussian random variable. Denote $X_k = \sum_{j=1}^C X_{jk}$ and $X_k^* = \sum_{j=1}^C X_{jk}^*$. Also

$$\begin{aligned} \text{Var}(T_{mk}) &= \frac{1}{np_k} \text{Var}(X_k^*) = \frac{1}{np_k} \left(\text{Var}(X_k) + \text{Var}\left(\sum_{j=1}^C err_{jk}\right) \right) \\ &= \frac{1}{np_k} \left(\text{Var}(X_k) + \sum_{j=1}^C \text{Var}(err_{jk}) \right) \\ &= 1 - p_k + \frac{\sum_{j=1}^C \text{Var}(err_{jk})}{np_k} \\ &= 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (29)$$

The last step follows from (25) where $\text{Var}(err_{jk}) = O(1)$ and the fact that $C < \infty$. Note also that

$$1 - p_k + \frac{\sum_{j=1}^C v(x_{jk}^*)}{np_k} = 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \quad (30)$$

Equation (30) follows from the facts that $v(x_{jk}^*) = O(1)$ as $n \rightarrow \infty$ (for details see proof of Theorem 1) and $C < \infty$. From (29) and (30), we must have

$$\left| \text{Var}(T_{mk}) - \left(1 - p_k + \frac{\sum_{j=1}^C v(x_{jk}^*)}{np_k} \right) \right| = O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty.$$

Lastly, since X_k are correlated, T_{mk} are correlated. For some $k \neq j$, consider,

$$\begin{aligned} \text{Cov}(T_{mk}, T_{mj}) &= \text{Cov}\left(\sqrt{\frac{n}{p_k}} \left(\frac{X_k + \sum_{i=1}^C err_{ik}}{n} - p_k \right), \sqrt{\frac{n}{p_j}} \left(\frac{X_j + \sum_{i=1}^C err_{ij}}{n} - p_j \right)\right) \\ &= \frac{1}{n\sqrt{p_k p_j}} \text{Cov}\left(X_k + \sum_{i=1}^C err_{ik}, X_j + \sum_{i=1}^C err_{ij}\right) \\ &= \frac{1}{n\sqrt{p_k p_j}} \text{Cov}(X_k, X_j) \\ &= -\sqrt{p_k p_j}. \end{aligned}$$

Similarly, let $\Sigma \in \mathbb{R}^{K \times K}$ be a matrix with diagonal entries $\Sigma_{kk} = 1 - p_k + (\sum_{j=1}^C v(x_{jk}^*)) / (np_k)$ and off-diagonal entries $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\mathbf{T}_M = \{T_{m1}, \dots, T_{mK}\}$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \rightarrow \infty$. By a similar argument as in the proof of Theorem 1 we can derive that $T_M^* \rightarrow \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1 and Λ_k are the eigenvalues of Σ corresponding to a set of orthonormal eigenvectors of Σ . Hence, we have the result of part (a) follows.

For part (b), note that

$$T_{mk} = \begin{cases} \sqrt{\frac{n}{p_k}} \left(\frac{\sum_{j=1}^C X_{jk} - np_k}{n} \right) + \frac{\sum_{j=1}^C (err_{jk})}{\sqrt{np_k}}, & \text{if } X_{jk}^* \text{ from Lap,} \\ \sqrt{\frac{n}{p_k}} \left(\frac{\sum_{j=1}^C X_{jk} - np_k}{n} \right) + \frac{\sum_{j=1}^C (err_{jk} + b_{jk})}{\sqrt{np_k}}, & \text{if } X_{jk}^* \text{ from TLap,} \end{cases} \quad (31)$$

where $err_{jk} \sim \text{Lap}(0, 1/\epsilon_n)$ and b_{jk} is the bias term which can be expressed as follows,

$$b_{jk} = \begin{cases} 0, & \text{if } err_{jk} \geq -X_{jk}, \\ -(X_{jk} + err_{jk}), & \text{otherwise.} \end{cases} \quad (33)$$

$$(34)$$

It can be shown that $P(\sum_k (\text{31})^2 \neq \sum_k (\text{32})^2) = o(1)$ as $n \rightarrow \infty$. To see this, note for any $0 < \Delta_{jk} < p_k$, by Lemma 3, it follows

$$P(b_{jk} \neq 0) = O\left(e^{n(\Delta_{jk} - p_k)\epsilon_n} + e^{-2n\Delta_{jk}}\right). \quad (35)$$

Take $\Delta_{jk} = \min\{p_1, \dots, p_K\}/2$ for all $j = 1, \dots, C$. Further note that as $n \rightarrow \infty$,

$$\begin{aligned} P\left(\sum_k (\text{31})^2 \neq \sum_k (\text{32})^2\right) &= P(\cup_k \cup_j \{b_{jk} \neq 0\}) \\ &\leq \sum_{k=1}^K \sum_{j=1}^C P(b_{jk} \neq 0) \\ &= O\left(\exp\left\{-\frac{1}{2}n\epsilon_n \min\{p_1, \dots, p_K\}\right\} + \exp\left\{-n \min\{p_1, \dots, p_K\}\right\}\right). \end{aligned}$$

where the last step follows directly from (35) and the fact that $K, C < \infty$. Since the privacy regime ϵ_n satisfies $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, it follows $P(\sum_k (\text{31})^2 \neq \sum_k (\text{32})^2) = o(1)$ as $n \rightarrow \infty$. Therefore, we can ignore all the bias terms b_{jk} and work only with (31) in the following proofs. The remaining proofs for part (b) are similar as in part (a), except for the difference in Σ . In the case of ϵ_n -Laplace mechanism in part (b), $\text{Cov}(T_{mk}, T_{mj})$ remains the same for $j \neq k$ but

$$\begin{aligned} \text{Var}(T_{mk}) &= \frac{1}{np_k} \text{Var}(X_k^*) \\ &= \frac{1}{np_k} \left(\text{Var}(X_k) + \text{Var}\left(\sum_{j=1}^C err_{jk}\right) \right) \\ &= \frac{1}{np_k} \left(\text{Var}(X_k) + \sum_{j=1}^C \text{Var}(err_{jk}) \right) \\ &= \frac{1}{np_k} (np_k(1 - p_k) + 2C/\epsilon_n^2) \\ &= 1 - p_k + 2C/(\epsilon_n^2 np_k). \end{aligned}$$

Now for part (c), again we can express

$$T_{mk} = \begin{cases} \sqrt{\frac{n}{p_k}} \left(\frac{\sum_{j=1}^C X_{jk} - np_k}{n} \right) + \frac{\sum_{j=1}^C (err_{jk})}{\sqrt{np_k}}, & \text{if } X_{jk}^* \text{ from GDP,} \\ \sqrt{\frac{n}{p_k}} \left(\frac{\sum_{j=1}^C X_{jk} - np_k}{n} \right) + \frac{\sum_{j=1}^C (err_{jk} + b_{jk})}{\sqrt{np_k}}, & \text{if } X_{jk}^* \text{ from TGDP,} \end{cases} \quad (36)$$

where $err_{jk} \sim N(0, 2 \ln\{1.25/\delta\}/\epsilon_n^2)$ and b_{jk} is the bias term remains the same as before,

$$b_{jk} = \begin{cases} 0, & \text{if } err_{jk} \geq -X_{jk}, \\ -(X_{jk} + err_{jk}), & \text{otherwise.} \end{cases} \quad (38)$$

$$(39)$$

Similarly, it can be shown that $P(\sum_k (36)^2 \neq \sum_k (37)^2) = o(1)$ as $n \rightarrow \infty$. To see this, note for any $0 < \Delta_{jk} < p_k$, by Lemma 4, we have

$$P(b_{jk} \neq 0) = O\left(\exp\left\{-2n\Delta_{jk}\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2(p_k - \Delta_{jk})^2}{4 \ln\{1.25/\delta\}}\right\}\right). \quad (40)$$

Again, take $\Delta_{jk} = \min\{p_1, \dots, p_K\}/2$ for all $j = 1, \dots, C$. Further note that as $n \rightarrow \infty$,

$$\begin{aligned} P\left(\sum_k (36)^2 \neq \sum_k (37)^2\right) &= P(\cup_k \cup_j \{b_{jk} \neq 0\}) \\ &\leq \sum_{k=1}^K \sum_{j=1}^C P(b_{jk} \neq 0) \\ &= O\left(\exp\left\{-n \min\{p_1, \dots, p_K\}\right\} + \frac{1}{n\epsilon_n} \exp\left\{-\frac{n^2\epsilon_n^2 \min\{p_1, \dots, p_K\}^2}{16 \ln\{1.25/\delta\}}\right\}\right). \end{aligned}$$

where the last step follows directly from (40) and the fact that $K, C < \infty$. Since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, it follows $P(\sum_k (36)^2 \neq \sum_k (37)^2) = o(1)$ as $n \rightarrow \infty$. Therefore, we can ignore all the bias terms b_{jk} and work only with (36). The remaining proofs for part (c) are similar to that in part (a), except for the difference in Σ . In the case of (ϵ_n, δ) -Gaussian Mechanism in part (c), $\text{Cov}(T_{mk}, T_{mj})$ remains the same for $j \neq k$ but

$$\begin{aligned} \text{Var}(T_{mk}) &= \frac{1}{np_k} \text{Var}(X_k^*) \\ &= \frac{1}{np_k} \left(\text{Var}(X_k) + \text{Var}\left(\sum_{j=1}^C err_{jk}\right) \right) \\ &= \frac{1}{np_k} \left(\text{Var}(X_k) + \sum_{j=1}^C \text{Var}(err_{jk}) \right) \\ &= \frac{1}{np_k} (np_k(1 - p_k) + (2C \log(1.25/\delta) - 1)/\epsilon_n^2) \\ &= 1 - p_k + C(2 \log(1.25/\delta) - 1)/(np_k \epsilon_n^2). \end{aligned}$$

We have both parts (b) and (c) of the theorem follow. \square

Proof of Theorem 4. We start with part (a). Consider T_{m1} ,

$$T_{m1} = \frac{1}{\sqrt{np_1}} \left(\sum_{k=1}^M X_k^* - np_1 - b_M(\{x_j^*\}_{j=1}^M) \right) = \sqrt{\frac{n}{p_1}} \left(\frac{\sum_{k=1}^M X_k - np_1}{n} \right) + \frac{\sum_{k=1}^M (err_k - b(x_k^*))}{\sqrt{np_1}}. \quad (41)$$

Assume H_0 is true, by CLT, we know that $(\sum_{k=1}^M X_k - np_1)/n$ is Gaussian asymptotically with mean 0 as $n \rightarrow \infty$. Following from Proposition 1 and the facts that $M, b(x_k^*) < \infty$, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \rightarrow 0$ as $n \rightarrow \infty$, the second term of (41) will converge in probability to 0 as $n \rightarrow \infty$. Therefore, overall, T_{m1} will converge to a Gaussian distribution. Following a similar argument as the proof of Theorem 2, we know that T_{mk} for $k = 2, \dots, K$ will also converge to Gaussian random variables with mean 0 asymptotically. Denote $X_{m1} = \sum_{k=1}^M X_k$ and $X_{m1}^* = \sum_{k=1}^M X_k^*$. Note

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(X_{m1}^*) = \frac{1}{np_1} \left(Var(X_{m1}) + Var\left(\sum_{k=1}^M err_k\right) \right) \\ &= \frac{1}{np_1} \left(Var(X_{m1}) + \sum_{k=1}^M Var(err_k) \right) \\ &= 1 - p_1 + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (42)$$

For $k = 2, \dots, K$,

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(X_{mk}^*) = \frac{1}{np_k} \left(Var(X_{mk}) + Var(err_k) \right) \\ &= 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (43)$$

Equations (42) and (43) follow directly from (25) and the fact that $M < K < \infty$. Note also that

$$1 - p_1 + \frac{\sum_{j=1}^M v(x_j^*)}{np_1} = 1 - p_1 + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \quad (44)$$

$$1 - p_k + \frac{v(x_k^*)}{np_k} = 1 - p_k + O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \quad (45)$$

Equations (44) and (45) follow from $v(x_j^*) = O(1)$ as $n \rightarrow \infty$ (for more details see proofs of Theorem 1 and the fact that $M < K < \infty$). From (42) and (44), we have

$$\left| Var(T_{m1}) - \left(1 - p_1 + \frac{\sum_{j=1}^M v(x_j^*)}{np_1} \right) \right| = O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty.$$

Similarly, from Equations (43) and (45), we can derive for $k = 2, \dots, K$,

$$\left| Var(T_{mk}) - \left(1 - p_k + \frac{v(x_k^*)}{np_k} \right) \right| = O\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty.$$

Lastly, we derive covariance amongst T_{mk} , $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$, for any $k \neq j$. Similarly, let $\Sigma \in \mathbb{R}^{(K-M+1) \times (K-M+1)}$ be a matrix with diagonal entries $\Sigma_{11} = 1 - p_1 +$

$(\sum_{j=1}^M v(x_j^*))/(np_1)$ and $\Sigma_{kk} = 1 - p_k + v(x_k^*)/(np_k)$ for $k = 2, \dots, K - M + 1$, and off-diagonal entries $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\mathbf{T}_M = (T_{m1}, \dots, T_{m(K-M+1)})$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \rightarrow \infty$. Then with a similar argument as in the proof of Theorem 2, $T_M^* \rightarrow \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1 and Λ_k are eigenvalues of Σ corresponding to a set of orthonormal eigenvectors. Hence, the result of part (a) follows.

For part (b) and (c), first note that using similar arguments as in the proof for Theorem 2(a) and 2(b), it can be shown that the probability of test statistics obtained from the Laplace/Gaussian mechanisms not equal to their counterparts obtained from the truncated Laplace/Gaussian mechanisms tends to 0 exponentially fast as n goes to infinity. Therefore, we can in fact ignore the truncation effect. The proofs for the remaining are similar to the proof for part (a), except for the difference in Σ . In the case of ϵ_n -Laplace mechanism in part (b), $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$ for $k \neq j$ remains the same but

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(X_{m1}^*) = \frac{1}{np_1} \left(Var(X_{m1}) + Var\left(\sum_{k=1}^M err_k\right) \right) \\ &= \frac{1}{np_1} \left(Var(X_{m1}) + \sum_{k=1}^M Var(err_k) \right) \\ &= 1 - p_1 + 2M/(\epsilon_n^2 np_1). \end{aligned}$$

For $k = 2, \dots, K$,

$$Var(T_{mk}) = \frac{1}{np_k} Var(X_{mk}^*) = \frac{1}{np_k} \left(Var(X_{mk}) + Var(err_k) \right) = (1 - p_k) + 2/(np_k \epsilon_n^2).$$

In the case of (ϵ_n, δ) -Gaussian Mechanism in part (c), $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$ for $k \neq j$ remains the same but

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(X_{m1}^*) = \frac{1}{np_1} \left(Var(X_{m1}) + Var\left(\sum_{k=1}^M err_k\right) \right) \\ &= \frac{1}{np_1} \left(Var(X_{m1}) + \sum_{k=1}^M Var(err_k) \right) \\ &= (1 - p_1) + M(2 \log(1.25/\delta) - 1)/(np_1 \epsilon_n^2). \end{aligned}$$

For $k = 2, \dots, K$,

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(X_{mk}^*) = \frac{1}{np_k} \left(Var(X_{mk}) + Var(err_k) \right) \\ &= (1 - p_k) + (2 \log(1.25/\delta) - 1)/np_k \epsilon_n^2. \end{aligned}$$

The results of both part (b) and (c) follow from a similar argument as in the proof of Theorem 2. \square