A Proofs, Additional Theoretical Results & Discussion

A.1 (ε, δ)-DP and Rényi DP for Propose-Test-Release

We consider two adjacent datasets S, S' where $S' = S \cup \{x\}$. We denote the threshold $B = \log(1/(2\delta_0))b$. Note that we have $\Pr[\mathrm{Lap}(0, b) > B] = \delta_0$. The output of Algorithm 1 on dataset S is a sample from a joint distribution $(\widehat{\Delta}, \mathcal{M})(S)$ where $\widehat{\Delta}(S) = \mathrm{Lap}(\Delta(S), b)$ and $\mathcal{M}(S)|_{\widehat{\Delta}} = \mathcal{N}(f_1(S), \sigma_1^2)\mathbb{1}[\widehat{\Delta} \le B] + \mathcal{N}(f_2(S), \sigma_2^2)\mathbb{1}[\widehat{\Delta} > B]$.

Theorem 4.2 (restated). Suppose $\mathbb{GS}_{f_1} = \mathbb{GS}_{f_2} = 1$ and $\sigma_1 = \sigma_2/\tau$, then Algorithm 1 is $(\varepsilon_{\mathrm{Lap}}^{(b)} + \varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta), \delta_0 + \delta)$-DP.

Proof. We consider two cases for $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$.

Case 1: both $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are greater than τ. In this case, we have $\Delta(S) = \Delta(S') = 0$ (recall that Δ refers to the minimum amount of data addition/removal to make the local sensitivity $> \tau$). Therefore, there are no privacy loss in $\widehat{\Delta}$. Besides, the probability that PTR releases $f_2(S) + \mathcal{N}(0, \sigma_2^2)$ is at most $\Pr[\widehat{\Delta} > B] = \Pr[\mathrm{Lap}(0, b) > B] = \delta_0$. Therefore, with probability at least $1 - \delta_0$, the PTR is $(\varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta), \delta)$-DP, and overall it is $(\varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta), \delta + \delta_0)$-DP.

Case 2: at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are smaller than τ. In this case, we know that $\Pr[\mathcal{M}(S) \in T] \le e^{\varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta)} \Pr[\mathcal{M}(S') \in T] + \delta$ regardless of the value of $\widehat{\Delta}$. Thus, by basic composition theorem, PTR in this case is $\left(\varepsilon_{\mathrm{Lap}}^{(b)} + \varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta), \delta\right)$-DP.

Therefore, PTR is $(\varepsilon_{\mathrm{Lap}}^{(b)} + \varepsilon_{\mathcal{N}}^{(\sigma_1)}(\delta), \delta_0 + \delta)$-DP overall. $\square$

Comparison between (ε, δ)-DP and RDP Analysis: A motivating example (expanded). Suppose we have two mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ who are $(\varepsilon_1, \delta_1)$-DP and $(\varepsilon_2, \delta_2)$-DP, respectively. Consider a simple PTR-like mechanism $\mathcal{M}$ that randomly picks one of mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ to run, each with probability $1 - \delta_0$ and δ_0[7]. A straightforward (ε, δ)-DP analysis for $\mathcal{M}$ can be given as follows: for any possible event T,

$$\begin{align}
\Pr[\mathcal{M}(S) \in T] &= (1 - \delta_0)\Pr[\mathcal{M}_1(S) \in T] + \delta_0 \Pr[\mathcal{M}_2(S) \in T] \tag{3}\\
&\le (1 - \delta_0)[e^{\varepsilon_1} \Pr[\mathcal{M}_1(S') \in T] + \delta_1] + \delta_0[e^{\varepsilon_2} \Pr[\mathcal{M}_2(S') \in T] + \delta_2] \tag{4}\\
&= e^{\varepsilon_1}(1 - \delta_0)\Pr[\mathcal{M}_1(S') \in T] + e^{\varepsilon_2}\delta_0 \Pr[\mathcal{M}_2(S') \in T] + (1 - \delta_0)\delta_1 + \delta_0\delta_2 \tag{5}\\
&\le e^{\max(\varepsilon_1, \varepsilon_2)} \Pr[\mathcal{M}(S) \in T] + (1 - \delta_0)\delta_1 + \delta_0\delta_2 \tag{6}
\end{align}$$

That is, $\mathcal{M}$ is $(\max(\varepsilon_1, \varepsilon_2), (1 - \delta_0)\delta_1 + \delta_0\delta_2)$-DP. Without further information, this bound is the best we can do since it is tight when there exists event T such that $\Pr[\mathcal{M}_1(S') \in T] = 0$ while $\Pr[\mathcal{M}_2(S') \in T] > 0$. Alternatively, if we know $\varepsilon_2 \gg \varepsilon_1$ we can also move the probability δ_0 to the δ term and obtain $(\varepsilon_1, \delta_0 + \delta_1)$ (which is the case for Theorem 4.2).

However, if we know the RDP guarantee of $\mathcal{M}_1$ and $\mathcal{M}_2$ as $E_\alpha(\mathcal{M}_1(S) \| \mathcal{M}_1(S')) \le f_\alpha(\varepsilon_1)$ and $E_\alpha(\mathcal{M}_2(S) \| \mathcal{M}_2(S')) \le f_\alpha(\varepsilon_2)$[8], then $E_\alpha(\mathcal{M}(S) \| \mathcal{M}(S'))$ can be simply bounded as

$$\begin{align}
\mathbb{E}_{\mathcal{M}(S)}\left[\left(\frac{\mu_{\mathcal{M}(S')}}{\mu_{\mathcal{M}(S)}}\right)^\alpha\right] &= (1 - \delta_0)\mathbb{E}_{\mathcal{M}_1(S)}\left[\left(\frac{\mu_{\mathcal{M}_1(S')}}{\mu_{\mathcal{M}_1(S)}}\right)^\alpha\right] + \delta_0\mathbb{E}_{\mathcal{M}_2(S)}\left[\left(\frac{\mu_{\mathcal{M}_2(S')}}{\mu_{\mathcal{M}_2(S)}}\right)^\alpha\right] \tag{7}\\
&\le (1 - \delta_0)f_\alpha(\varepsilon_1) + \delta_0 f_\alpha(\varepsilon_2) \tag{8}
\end{align}$$

Compared with (ε, δ)-DP analysis, there are no extra inequalities used in RDP analysis of $\mathcal{M}$ except for the RDP guarantee of $\mathcal{M}_1$ and $\mathcal{M}_2$. Thus, RDP is more favorable in for PTR's privacy analysis, especially when δ_0 is close to the target δ.

---

[7] For the actual PTR, the δ_0 is not fixed but depends on the input dataset.

[8] Recall that $f_\alpha(\varepsilon) = \exp((\alpha - 1)\varepsilon)$ where if $\mathcal{M}$ is (α, ε)-RDP then $E_\alpha(\mathcal{M}(S) \| \mathcal{M}(S')) \le f_\alpha(\varepsilon)$.

Theorem 4.3 (restated). Suppose $\mathbb{GS}_{f_1} = \mathbb{GS}_{f_2} = 1$ and $\sigma_1 = \sigma_2/\tau$. Then for any $\alpha > 1$, Algorithm 1 is $(\alpha, \varepsilon_{\mathrm{PTR}}(\alpha))$-RDP for

$$\varepsilon_{\mathrm{PTR}}(\alpha) \le \max\left(f_\alpha^{-1}\left((1-\delta_0) f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R}-\mathcal{N}}(\alpha)\right) + \delta_0 f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R}-\mathcal{N}}(\alpha)\right)\right), \varepsilon^{(\sigma_1)}_{\mathrm{R}-\mathcal{N}}(\alpha) + \varepsilon^{(b)}_{\mathrm{R}-\mathrm{Lap}}(\alpha)\right)$$

Proof. We will denote the density of $(\widehat{\Delta}, \mathcal{M})(S)$ as μ and that of $(\widehat{\Delta}, \mathcal{M})(S')$ as μ'. We will use $\mu(s,t)$ to denote the joint density on the pair of outputs (s,t), where $s \sim \widehat{\Delta}(S)$ and $t \sim \mathcal{M}(S)|_{\widehat{\Delta}}$. Furthermore, when we write $\mu(s)$ it refers to the marginal density of μ on s, and $\mu(t|s)$ refers to the conditional density on t given s.

In order to bound RDP of PTR with order α, it suffices to bound the moments $\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^\alpha\right]$ and $\mathbb{E}_{(s,t)\sim\mu'}\left[\left(\frac{\mu(s,t)}{\mu'(s,t)}\right)^\alpha\right]$ then take the bigger of the two bounds. For readability, we may abbreviate the two quantities as $\mathbb{E}_\mu\left[\left(\frac{\mu'}{\mu}\right)^\alpha\right]$ and $\mathbb{E}_{\mu'}\left[\left(\frac{\mu}{\mu'}\right)^\alpha\right]$. We do the following to decompose $\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^\alpha\right]$:

$$\begin{align}
&\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^\alpha\right] \tag{9}\\
&= \mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s)\mu'(t|s)}{\mu(s)\mu(t|s)}\right)^\alpha\right] \tag{10}\\
&= \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu'(s)}{\mu(s)}\right)^\alpha \mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^\alpha\right]\right] \tag{11}\\
&= \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu'(s)}{\mu(s)}\right)^\alpha \left(\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^\alpha\right]\mathbb{1}[s\le B] + \mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^\alpha\right]\mathbb{1}[s > B]\right)\right] \tag{12}\\
& \tag{13}
\end{align}$$

When $s \le B$, we know that

$$\begin{align}
\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^\alpha\right] &= \mathbb{E}_{t\sim\mathcal{N}(f_1(S),\sigma_1^2\mathbf{1}_d)}\left[\left(\frac{\mathcal{N}(t; f_1(S'), \sigma_1^2\mathbf{1}_d)}{\mathcal{N}(t; f_1(S), \sigma_1^2\mathbf{1}_d)}\right)^\alpha\right] \tag{14}\\
&= \mathbb{E}_{t\sim\mathcal{N}(\mathbf{0},\sigma_1^2\mathbf{1}_d)}\left[\left(\frac{\mathcal{N}(t; f_1(S') - f_1(S), \sigma_1^2\mathbf{1}_d)}{\mathcal{N}(t; \mathbf{0}, \sigma_1^2\mathbf{1}_d)}\right)^\alpha\right] \tag{15}\\
&= \mathbb{E}_{t\sim\mathcal{N}(0,\sigma_1^2)}\left[\left(\frac{\mathcal{N}(t; \|f_1(S') - f_1(S)\|, \sigma_1^2)}{\mathcal{N}(t; 0, \sigma_1^2)}\right)^\alpha\right] \tag{16}\\
&\le \mathbb{E}_{t\sim\mathcal{N}(0,\sigma_1^2)}\left[\left(\frac{\mathcal{N}(t; 1, \sigma_1^2)}{\mathcal{N}(t; 0, \sigma_1^2)}\right)^\alpha\right] \tag{17}\\
&= f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R}-\mathcal{N}}(\alpha)\right) \tag{18}
\end{align}$$

where (15) is due to the translation invariance of Rényi divergence, (16) is due to the rotation trick, (17) is because of $\|f_1(S') - f_1(S)\| \le 1$.

We now analyze the upper bound of $\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^\alpha\right]$ when $s > B$ by considering two separate cases: when both $\mathbb{LS}_{f_2}(S) > \tau$ and $\mathbb{LS}_{f_2}(S') > \tau$, and when there is at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ is greater than τ.

Case 1: both $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are greater than τ. In this case, the only known upper bound of $\|f_2(S) - f_2(S')\|$ is the global sensitivity $\mathbb{GS}_{f_2} = 1$. Therefore, we only have

$\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right] \leq f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)$ when $s > B$. Therefore, in this case we have

$$\begin{aligned}
&\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\mathbb{1}[s \leq B] + \mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\mathbb{1}[s > B] && (19)\\
&= f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s \leq B] + f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s > B] && (20)
\end{aligned}$$

However, note that when both $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ is greater than τ, we have $\Delta(S) = \Delta(S') = 0$, which means that there is no privacy loss by releasing the result of $\widehat{\Delta}(S)$ or $\widehat{\Delta}(S')$. Therefore, we have $\mu(s) = \mu'(s) = \mathrm{Lap}(s; 0, b)$, and thus

$$\begin{aligned}
&\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^{\alpha}\right] && (21)\\
&= \mathbb{E}_{s\sim\mu}\left[f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s \leq B] + f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s > B]\right] && (22)\\
&= f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\Pr[\mathrm{Lap}(0,b) \leq B] + f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\Pr[\mathrm{Lap}(0,b) > B] && (23)\\
&= (1-\delta_0) f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) + \delta_0 f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) && (24)
\end{aligned}$$

Case 2: at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are smaller than τ. In this case, we know that we have $\|f_2(S) - f_2(S')\| \leq \tau$. Thus, when $s \geq B$, we have

$$\begin{aligned}
\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right] &= \mathbb{E}_{t\sim\mathcal{N}(0,\sigma_2^2)}\left[\left(\frac{\mathcal{N}(t; \|f_2(S') - f_2(S)\|, \sigma_2^2)}{\mathcal{N}(t; 0, \sigma_2^2)}\right)^{\alpha}\right] && (25)\\
&\leq \mathbb{E}_{t\sim\mathcal{N}(0,\sigma_2^2)}\left[\left(\frac{\mathcal{N}(t; \tau, \sigma_2^2)}{\mathcal{N}(t; 0, \sigma_2^2)}\right)^{\alpha}\right] && (26)\\
&= f_\alpha\left(\varepsilon^{(\sigma_2/\tau)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) && (27)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^{\alpha}\right] && (28)\\
&\leq \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu'(s)}{\mu(s)}\right)^{\alpha}\left(f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s \leq B] + f_\alpha\left(\varepsilon^{(\sigma_2/\tau)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{1}[s > B]\right)\right] && (29)\\
&= \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu'(s)}{\mu(s)}\right)^{\alpha} f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\right] && (30)\\
&= f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu'(s)}{\mu(s)}\right)^{\alpha}\right] && (31)\\
&\leq f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) f_\alpha(\varepsilon^{(b)}_{\mathrm{R-Lap}}(\alpha)) && (32)
\end{aligned}$$

where (30) is because by our condition, $\sigma_1 = \sigma_2/\tau$.

Therefore, we have

$$\begin{aligned}
&D_\alpha(\mu'\|\mu)\\
&\leq \frac{1}{\alpha-1}\log\left(\max((1-\delta_0) f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) + \delta_0 f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right), f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) f_\alpha(\varepsilon^{(b)}_{\mathrm{R-Lap}}(\alpha)))\right)\\
&= \max\left(\frac{1}{\alpha-1}\log\left((1-\delta_0) f_\alpha\left(\varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right) + \delta_0 f_\alpha\left(\varepsilon^{(\sigma_2)}_{\mathrm{R-}\mathcal{N}}(\alpha)\right)\right), \varepsilon^{(\sigma_1)}_{\mathrm{R-}\mathcal{N}}(\alpha) + \varepsilon^{(b)}_{\mathrm{R-Lap}}(\alpha)\right)
\end{aligned}$$

Since we did not use any condition that depends on the fact that $S' = S \cup \{x\}$, we know that $D_\alpha(\mu\|\mu')$ also has the exactly the same upper bound, which leads to the conclusion. □

A.1.1 Discussion: can we improve privacy analysis by not releasing $\widehat{\Delta}$?

One may wonder if we can further improve the privacy analysis of PTR by not releasing $\widehat{\Delta}$. However, releasing $\widehat{\Delta}$ is essential for the applications of PTR. The rationale behind PTR is to exploit the fact that, while a function's global sensitivity may be large, its local sensitivity may be much smaller for most of the "common inputs". Thus, such a mechanism will only be preferred over a regular output perturbation mechanism when the local sensitivity of data drawn from input data distribution rarely exceeds the threshold. Without knowing about $\widehat{\Delta}$, the user **cannot** know whether they are actually enjoying the benefits from PTR or simply wasting privacy budgets on private sensitivity tests. Furthermore, the user cannot adjust the hyperparameters or switch algorithms accordingly. Notably, in Section 5 (the application of PTR in privatizing robust SGD), we also use the information from $\widehat{\Delta}$ to dynamically adjust the number of gradients to be trimmed (note that this does not affect privacy analysis since the adjustment is post-processing of $\widehat{\Delta}$).

Besides, we gave an attempt to directly analyze the variant of PTR that does not release $\hat{\Delta}$, and we do not see an easy way to obtain a better privacy bound than we have in Theorem 4.3.

We follow the same notations as in the proof of Theorem 4.3: Given a pair of neighboring dataset S, S', we denote the density of $\mathcal{M}(S)$ as μ and that of $\mathcal{M}(S')$ as μ'. Given $s \sim \widehat{\Delta}(S)$, we denote $\mu(t|s \le B)$ the density of $\mathcal{N}(f_1(S), \sigma_1^2)$, and $\mu(t|s > B)$ the density of $\mathcal{N}(f_2(S), \sigma_2^2)$. $\mu'(t|s \le B)$ and $\mu'(t|s > B)$ are defined analogously.

Similar to the proof of Theorem 4.3, we consider two separate cases: when both $\mathbb{LS}_{f_2}(S) > \tau$ and $\mathbb{LS}_{f_2}(S') > \tau$, and when there is at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ is greater than τ.

Case 1. For the case that both $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are greater than τ, from the proof of Theorem 4.3 we know that $\widehat{\Delta}(S)$ and $\widehat{\Delta}(S')$ has exactly the same distribution since $\Delta(S) = \Delta(S') = 0$. Thus, the exactly the same proof in Theorem 4.3 applies for the case of not releasing $\widehat{\Delta}$.

Case 2. For the case that at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ is smaller than τ, here's our attempt:

$$E_\alpha(\mathcal{M}(S) \| \mathcal{M}(S')) = \mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t)}{\mu(t)}\right)^\alpha\right] \tag{33}$$

$$= \mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t|s \le B)\Pr[\widehat{\Delta}(S') \le B] + \mu'(t|s > B)\Pr[\widehat{\Delta}(S') > B]}{\mu(t|s \le B)\Pr[\widehat{\Delta}(S) \le B] + \mu(t|s > B)\Pr[\widehat{\Delta}(S) > B]}\right)^\alpha\right] \tag{34}$$

As we can see, while the distribution of $\mathcal{M}(S)$ is a Gaussian mixture, the probability for different components is also depending on S, which introduce more challenge in bounding $E_\alpha(\mathcal{M}(S)\|\mathcal{M}(S'))$. One relatively simple way to bound the above expression is by noticing that since $\widehat{\Delta} = \Delta + \mathrm{Lap}(0, b)$, by the privacy guarantee of Laplace mechanism we have $\Pr[\widehat{\Delta}(S') \le B] \le e^{1/b}\Pr[\widehat{\Delta}(S) \le B]$ and $\Pr[\widehat{\Delta}(S') > B] \le e^{1/b}\Pr[\widehat{\Delta}(S) > B]$. Thus, we have

$$(34) \le \exp\left(\frac{\alpha}{b}\right)\mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t|s \le B)\Pr[\widehat{\Delta}(S) \le B] + \mu'(t|s > B)\Pr[\widehat{\Delta}(S) > B]}{\mu(t|s \le B)\Pr[\widehat{\Delta}(S) \le B] + \mu(t|s > B)\Pr[\widehat{\Delta}(S) > B]}\right)^\alpha\right] \tag{35}$$

$$\le \exp\left(\frac{\alpha}{b}\right)\mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t|s \le B)}{\mu(t|s \le B)}\right)^\alpha\right] \tag{36}$$

where the last inequality is due to the quasi-convexity of Renyi divergence [VEH14] (note that $\mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t|s\le B)}{\mu(t|s\le B)}\right)^\alpha\right] = \mathbb{E}_{t\sim\mu}\left[\left(\frac{\mu'(t|s>B)}{\mu(t|s>B)}\right)^\alpha\right]$ by construction for this case). Thus, we have

$$R_\alpha(\mathcal{M}(S)\|\mathcal{M}(S')) \le \varepsilon_{\mathrm{R-}\mathcal{N}}^{(\sigma_1)}(\alpha) + \frac{1}{\alpha - 1}\left(\frac{\alpha}{b}\right) \tag{37}$$

for the case of at least one of $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ are smaller than τ.

Now we show that this bound is not as good as the corresponding bound in Theorem 4.3. The corresponding bound in Theorem 4.3 for this case is $\varepsilon_{\mathrm{R-}\mathcal{N}}^{(\sigma_1)}(\alpha) + \varepsilon_{\mathrm{R-Lap}}^{(b)}(\alpha)$, so we only need to show $\varepsilon^{(b)}_{\mathrm{R-Lap}}(\alpha) < \frac{1}{\alpha-1}\left(\frac{\alpha}{b}\right)$.

$$\varepsilon^{(b)}_{\mathrm{R-Lap}}(\alpha) = \frac{1}{\alpha-1}\log\left(\frac{\alpha}{2\alpha-1}\exp\left(\frac{\alpha-1}{b}\right)+\frac{\alpha-1}{2\alpha-1}\exp\left(-\frac{\alpha}{b}\right)\right) \tag{38}$$

$$< \frac{1}{\alpha-1}\log\left(\frac{\alpha}{2\alpha-1}\exp\left(\frac{\alpha}{b}\right)+\frac{\alpha-1}{2\alpha-1}\exp\left(\frac{\alpha}{b}\right)\right) \tag{39}$$

$$= \frac{1}{\alpha-1}\log\left(\exp\left(\frac{\alpha}{b}\right)\right) \tag{40}$$

$$= \frac{1}{\alpha-1}\left(\frac{\alpha}{b}\right) \tag{41}$$

where the first inequality is due to $\exp\left(\frac{\alpha-1}{b}\right) < \exp\left(\frac{\alpha}{b}\right)$ and $\exp\left(-\frac{\alpha}{b}\right) < \exp\left(\frac{\alpha}{b}\right)$.

Thus, we think at least there are no simple solution for improving the privacy bound for PTR by not releasing $\widehat{\Delta}$. However, even if there are a better way to derive the privacy bound, this variant of PTR may not be user-friendly as the counterpart who release $\widehat{\Delta}$.

A.2 Rényi DP for Subsampled Propose-Test-Release

We consider two adjacent datasets S, S' where $S' = S \cup \{x\}$. We denote the threshold $B = \log(1/(2\delta_0))b$. Note that we have $\Pr[\mathrm{Lap}(0,b) > B] = \delta_0$. The output of Algorithm 1 on dataset S is a sample from a joint distribution $(\widehat{\Delta}, \mathcal{M})(S)$ where $\widehat{\Delta}(S) = \mathrm{Lap}(\Delta(S), b)$ and $\mathcal{M}(S)|_{\widehat{\Delta}} = \mathcal{N}(f_1(S), \sigma_1^2)\mathbb{1}[\widehat{\Delta} \le B] + \mathcal{N}(f_2(S), \sigma_2^2)\mathbb{1}[\widehat{\Delta} > B]$.

Theorem 4.4 (full version). Let q be the subsampling probability. Suppose $\mathbb{GS}_{f_1} = \mathbb{GS}_{f_2} = 1$ and $\sigma_1 = \sigma_2/\tau$. If $q \le \frac{\exp(-1/b)}{4+\exp(-1/b)}$ and $\sigma_1 \ge \sigma_2 \ge 4$, and α satisfy $1 < \alpha \le \frac{1}{2}\sigma_2^2 L - 2\ln\sigma_2, \alpha \le \frac{\frac{1}{2}\sigma_2^2 L^2 - \ln 5 - 2\ln\sigma_2}{L + \ln(q'\alpha) + 1/\left(2\sigma_2^2\right)}$, where $L = \ln\left(1 + \frac{1}{q'(\alpha-1)}\right)$ and $q' = \frac{q}{q+(1-q)\exp(-1/b)}$, we have

$$\varepsilon_{\mathrm{PTR}\circ\mathrm{PoissonSample}}(\alpha) \le f_\alpha^{-1}\left(\max(\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2)\right)$$

where

$$\mathcal{B}_0 = 1 + 2q^2\alpha(\alpha-1)\left(\frac{1-\delta_0}{\sigma_1^2} + \frac{\delta_0}{\sigma_2^2}\right) \tag{42}$$

$$\mathcal{B}_1 = \mathrm{R}_q^{(\alpha)} + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\mathrm{R}_q^{(\alpha)} - 2(1-q)\mathrm{R}_q^{(\alpha-1)} + (1-q)^2\mathrm{R}_q^{(\alpha-2)}\right] \tag{43}$$

$$\mathcal{B}_2 = \widetilde{\mathrm{R}}_q^{(\alpha)} + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\widetilde{\mathrm{R}}_q^{(\alpha)} - 2(1-q)\widetilde{\mathrm{R}}_q^{(\alpha+1)} + (1-q)^2\widetilde{\mathrm{R}}_q^{(\alpha+2)}\right] \tag{44}$$

with $\mathrm{R}_q^{(\alpha)} = \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^\alpha\right]$ and $\widetilde{\mathrm{R}}_q^{(\alpha)} = \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu_0(s)}{\mu(s)}\right)^\alpha\right]$ for $\mu_0 \sim \mathrm{Lap}(0,b)$ and $\mu \sim (1-q)\mathrm{Lap}(0,b) + q\mathrm{Lap}(1,b)$.

Proof. Let T denote a set-valued random variable defined by taking a random subset of S, where each element of S is independently placed in T with probability q. Conditioned on T, the PTR outputs $(\widehat{\Delta}, \mathcal{M})(T)$. Thus,

$$\left(\widehat{\Delta}, \mathcal{M}\right)(S) = \sum_{T\subseteq S} p_T \cdot \left(\widehat{\Delta}, \mathcal{M}\right)(T) \tag{45}$$

$$\left(\widehat{\Delta}, \mathcal{M}\right)(S') = \sum_{T\subseteq S} p_T \cdot \left((1-q)\cdot(\widehat{\Delta}, \mathcal{M})(T) + q\cdot(\widehat{\Delta}, \mathcal{M})(T\cup\{x\})\right) \tag{46}$$

where p_T denotes the probabilty of sampling the subset T.

$$
\begin{aligned}
& D_\alpha\left(\left(\widehat{\Delta}, \mathcal{M}\right)(S') \| \left(\widehat{\Delta}, \mathcal{M}\right)(S)\right) \\
&= D_\alpha\left(\sum_T p_T \cdot \left((1-q)\cdot(\widehat{\Delta}, \mathcal{M})(T) + q \cdot (\widehat{\Delta}, \mathcal{M})(T \cup \{x\})\right) \| \sum_T p_T \cdot \left(\widehat{\Delta}, \mathcal{M}\right)(T)\right) \\
&\le \sup_T D_\alpha\left((1-q)\cdot(\widehat{\Delta}, \mathcal{M})(T) + q \cdot (\widehat{\Delta}, \mathcal{M})(T \cup \{x\}) \| \left(\widehat{\Delta}, \mathcal{M}\right)(T)\right)
\end{aligned}
$$

where the last step is due to the quasi-convexity of Rényi divergence ([VEH14], Theorem 13). Symmetrically, we also have

$$
\begin{aligned}
& D_\alpha\left(\left(\widehat{\Delta}, \mathcal{M}\right)(S') \| \left(\widehat{\Delta}, \mathcal{M}\right)(S)\right) && (47) \\
& \le \sup_T D_\alpha\left(\left(\widehat{\Delta}, \mathcal{M}\right)(T) \| (1-q)\cdot(\widehat{\Delta}, \mathcal{M})(T) + q \cdot (\widehat{\Delta}, \mathcal{M})(T \cup \{x\})\right) && (48)
\end{aligned}
$$

Fix a subset T and denote $T' = T \cup \{x\}$. We use μ_0 to denote the density function of $(\widehat{\Delta}, \mathcal{M})(T)$, where $\mu_0(s,t)$ refers to the density on (s,t). We use μ_1 to denote the density function of $(\widehat{\Delta}, \mathcal{M})(T')$, where $\mu_1(s,t)$ refers to the density on (s,t). Let $\mu = (1-q)\mu_0 + q\mu_1$. We want to bound $\mathbb{E}_{\mu_0}\left[\left(\frac{\mu}{\mu_0}\right)^\alpha\right]$ and $\mathbb{E}_{\mu}\left[\left(\frac{\mu_0}{\mu}\right)^\alpha\right]$.

We first bound $\mathbb{E}_{\mu_0}\left[\left(\frac{\mu}{\mu_0}\right)^\alpha\right]$, which is usually considered as an easier one.

By decomposition, we have

$$
\begin{aligned}
\mathbb{E}_{s,t\sim\mu_0}\left[\left(\frac{\mu(s,t)}{\mu_0(s,t)}\right)^\alpha\right] &= \mathbb{E}_{s,t\sim\mu_0}\left[\left(\frac{\mu(s)\mu(t|s)}{\mu_0(s)\mu_0(t|s)}\right)^\alpha\right] && (49) \\
&= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^\alpha \mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^\alpha\right]\right] && (50)
\end{aligned}
$$

For the density of conditional distribution $\mu(t|s)$, we have

$$
\begin{aligned}
\mu(t|s) &= \frac{\mu(s,t)}{\mu(s)} && (51) \\
&= \frac{(1-q)\mu_0(s,t) + q\mu_1(s,t)}{\mu(s)} && (52) \\
&= I[s \le B] \cdot \frac{(1-q)\mu_0(s)\cdot\mathcal{N}(t; f_1(T), \sigma_1^2) + q\mu_1(s)\cdot\mathcal{N}(t; f_1(T'), \sigma_1^2)}{\mu(s)} && (53) \\
&\quad + I[s > B] \cdot \frac{(1-q)\mu_0(s)\cdot\mathcal{N}(t; f_2(T), \sigma_2^2) + q\mu_1(s)\cdot\mathcal{N}(t; f_2(T'), \sigma_2^2)}{\mu(s)} && (54)
\end{aligned}
$$

Denote $A(s) = \frac{q\mu_1(s)}{\mu(s)}$. Recall that $\mu(s) = (1-q)\mu_0(s) + q\mu_1(s)$, so we have $\frac{(1-q)\mu_0(s)}{\mu(s)} = 1 - A$. Then we have

$$
\begin{aligned}
\mu(t|s) = I[s \le B]&\left((1-A)\cdot\mathcal{N}(t; f_1(T), \sigma_1^2) + A\cdot\mathcal{N}(t; f_1(T'), \sigma_1^2)\right) && (55) \\
+ I[s > B]&\left((1-A)\cdot\mathcal{N}(t; f_2(T), \sigma_2^2) + A\cdot\mathcal{N}(t; f_2(T'), \sigma_2^2)\right) && (56)
\end{aligned}
$$

and we know that

$$
\mu_0(t|s) = I[s \le B]\mathcal{N}(t; f_1(T), \sigma_1^2) + I[s > B]\mathcal{N}(t; f_2(T), \sigma_2^2) \quad (57)
$$

Therefore we have

$$
\begin{aligned}
&\mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha}\right] && (58)\\
&= \mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha} I[s\le B] + \left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha} I[s>B]\right] && (59)\\
&= \mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha}\right] I[s\le B] + \mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha}\right] I[s>B] && (60)\\
&= I[s\le B]\mathbb{E}_{t\sim\mathcal{N}(f_1(T),\sigma_1^2)}\left[\left((1-A)+A\cdot\frac{\mathcal{N}(t;f_1(T'),\sigma_1^2)}{\mathcal{N}(t;f_1(T),\sigma_1^2)}\right)^{\alpha}\right] && (61)\\
&\quad + I[s>B]\mathbb{E}_{t\sim\mathcal{N}(f_2(T),\sigma_2^2)}\left[\left((1-A)+A\cdot\frac{\mathcal{N}(t;f_2(T'),\sigma_2^2)}{\mathcal{N}(t;f_2(T),\sigma_2^2)}\right)^{\alpha}\right] && (62)
\end{aligned}
$$

Note that $\mathbb{E}_{t\sim\mathcal{N}(f_1(T),\sigma_1^2)}\left[\left((1-A)+A\cdot\frac{\mathcal{N}(t;f_1(T'),\sigma_1^2)}{\mathcal{N}(t;f_1(T),\sigma_1^2)}\right)^{\alpha}\right]$ can be exactly bounded by the Rényi DP of subsampled RDP with sampling probability A.

Lemma A.1 ([MTZ19], Theorem 11)**.** *If* $q\le\frac{1}{5}, \sigma\ge 4$, *and* α *satisfy* $1<\alpha\le\frac{1}{2}\sigma^2 L - 2\ln\sigma$, $\alpha\le\frac{\frac{1}{2}\sigma^2L^2-\ln 5-2\ln\sigma}{L+\ln(q\alpha)+1/(2\sigma^2)}$ *where* $L=\ln\left(1+\frac{1}{q(\alpha-1)}\right)$, *then for any function* f *with* ℓ_2*-sensitivity* τ *satisfies*

$$
\mathbb{E}_{t\sim\mathcal{N}(f(T),\sigma^2)}\left[\left((1-q)+q\cdot\frac{\mathcal{N}(t;f(T'),\sigma^2)}{\mathcal{N}(t;f_2(T),\sigma^2)}\right)^{\alpha}\right]\le 1+2q^2\tau^2\alpha(\alpha-1)/\sigma^2 \qquad (63)
$$

Similar to the proof of Theorem 4.3, we consider two cases:

Case 1: both $\mathbb{LS}_{f_2}(T)$ and $\mathbb{LS}_{f_2}(T')$ are greater than τ. In this case, the only known upper bound of $\|f_2(T)-f_2(T')\|$ is the global sensitivity $\mathbb{GS}_{f_2}=1$. Therefore, we only have $\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\le 1+2A^2\alpha(\alpha-1)/\sigma_2^2$ when $s>B$. Therefore, in this case we have

$$
\begin{aligned}
&\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\mathbb{1}[s\le B] + \mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\mathbb{1}[s>B] && (64)\\
&= (1+2A^2\alpha(\alpha-1)/(\sigma_1)^2)\mathbb{1}[s\le B] + (1+2A^2\alpha(\alpha-1)/(\sigma_2)^2)\mathbb{1}[s>B] && (65)\\
&= 1+2A^2\alpha(\alpha-1)\left(\frac{\mathbb{1}[s\le B]}{\sigma_1^2}+\frac{\mathbb{1}[s>B]}{\sigma_2^2}\right) && (66)
\end{aligned}
$$

However, note that when both $\mathbb{LS}_{f_2}(S)$ and $\mathbb{LS}_{f_2}(S')$ is greater than τ, we have $\Delta(S)=\Delta(S')=0$. Therefore, we have $\mu(s)=\mu'(s)=\mathrm{Lap}(s;0,b)$, $A=q$, and thus

$$
\begin{aligned}
&\mathbb{E}_{(s,t)\sim\mu}\left[\left(\frac{\mu'(s,t)}{\mu(s,t)}\right)^{\alpha}\right] && (67)\\
&= \mathbb{E}_{s\sim\mu}\left[1+2q^2\alpha(\alpha-1)\left(\frac{\mathbb{1}[s\le B]}{\sigma_1^2}+\frac{\mathbb{1}[s>B]}{\sigma_2^2}\right)\right] && (68)\\
&= 1+2q^2\alpha(\alpha-1)\left(\frac{1-\delta_0}{\sigma_1^2}+\frac{\delta_0}{\sigma_2^2}\right) && (69)
\end{aligned}
$$

Case 2: at least one of $\mathbb{LS}_{f_2}(T)$ and $\mathbb{LS}_{f_2}(T')$ are smaller than τ. In this case, we know that we have $\|f_2(T)-f_2(T')\|\le\tau$. Since $A=\frac{q\mu_1(s)}{(1-q)\mu_0(s)+q\mu_1(s)}\le\frac{q}{(1-q)+q\exp(-1/b)}$ which satisfy the conditions in Lemma A.1 by our assumption, when $s\ge B$, we have $\mathbb{E}_{t\sim\mu|s}\left[\left(\frac{\mu'(t|s)}{\mu(t|s)}\right)^{\alpha}\right]\le$

$1 + 2A^2\alpha(\alpha-1)/(\sigma_2/\tau)^2$. Thus we have

$$\mathbb{E}_{s,t\sim\mu_0}\left[\left(\frac{\mu(s,t)}{\mu_0(s,t)}\right)^{\alpha}\right] \tag{70}$$

$$= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\mathbb{E}_{t\sim\mu_0(\cdot|s)}\left[\left(\frac{\mu(t|s)}{\mu_0(t|s)}\right)^{\alpha}\right]\right] \tag{71}$$

$$\leq \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\left(1 + 2A^2\alpha(\alpha-1)\left(\frac{\mathbb{1}[s\leq B]}{\sigma_1^2} + \frac{\mathbb{1}[s>B]}{(\sigma_2/\tau)^2}\right)\right)\right] \tag{72}$$

$$= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\right] + 2\alpha(\alpha-1)\mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}A^2\left(\frac{I[s\leq B]}{\sigma_1^2} + \frac{I[s>B]}{(\sigma_2/\tau)^2}\right)\right] \tag{73}$$

$$= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\right] + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}A^2\right] \tag{74}$$

Denote $\mathrm{R}_q^{(\alpha)} = \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\right]$, which is the RDP of subsampled Laplace mechanism with sampling rate q (note that $\mu(s) = (1-q)\mu_0(s) + q\mu_1(s)$).

Since

$$A = \frac{q\mu_1(s)}{\mu(s)} = 1 - \frac{(1-q)\mu_0(s)}{\mu(s)} \tag{75}$$

$$A^2 = 1 - \frac{2(1-q)\mu_0(s)}{\mu(s)} + \frac{(1-q)^2\mu_0^2(s)}{\mu^2(s)} \tag{76}$$

Plug this back to the second term, we have

$$\mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}A^2\right] \tag{77}$$

$$= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\left(1 - \frac{2(1-q)\mu_0(s)}{\mu(s)} + \frac{(1-q)^2\mu_0^2(s)}{\mu^2(s)}\right)\right] \tag{78}$$

$$= \mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha}\right] - 2(1-q)\mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha-1}\right] + (1-q)^2\mathbb{E}_{s\sim\mu_0}\left[\left(\frac{\mu(s)}{\mu_0(s)}\right)^{\alpha-2}\right] \tag{79}$$

$$= \mathrm{R}_q^{(\alpha)} - 2(1-q)\mathrm{R}_q^{(\alpha-1)} + (1-q)^2\mathrm{R}_q^{(\alpha-2)} \tag{80}$$

Note that this bound is independent on T due to translation invariance, hence it is an upper bound for arbitrary T (that satisfy case 2). Thus, the overall bound becomes

$$\mathbb{E}_{s,t\sim\mu_0}\left[\left(\frac{\mu(s,t)}{\mu_0(s,t)}\right)^{\alpha}\right] \leq \mathrm{R}_q^{(\alpha)} + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\mathrm{R}_q^{(\alpha)} - 2(1-q)\mathrm{R}_q^{(\alpha-1)} + (1-q)^2\mathrm{R}_q^{(\alpha-2)}\right] \tag{81}$$

Denote $\widetilde{\mathrm{R}}_q^{(\alpha)} = \mathbb{E}_{s\sim\mu}\left[\left(\frac{\mu_0(s)}{\mu(s)}\right)^{\alpha}\right]$. Since we know that

$$\widetilde{\mathrm{R}}_q^{(\alpha)} = \mathbb{E}_{\mu}\left[\left(\frac{\mu_0}{\mu}\right)^{\alpha}\right] = \mathbb{E}_{\mu_0}\left[\left(\frac{\mu_0}{\mu}\right)^{\alpha-1}\right] = \mathbb{E}_{\mu_0}\left[\left(\frac{\mu}{\mu_0}\right)^{1-\alpha}\right] = \mathrm{R}_q^{(1-\alpha)} \tag{82}$$

Thus, by setting $\alpha \leftarrow 1-\alpha$, we have

$$\mathbb{E}_{s,t\sim\mu}\left[\left(\frac{\mu_0(s,t)}{\mu(s,t)}\right)^{\alpha}\right] \tag{83}$$

$$= \mathbb{E}_{s,t\sim\mu_0}\left[\left(\frac{\mu(s,t)}{\mu_0(s,t)}\right)^{1-\alpha}\right] \tag{84}$$

$$= \widetilde{\mathrm{R}}_q^{(1-\alpha)} + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\widetilde{\mathrm{R}}_q^{(1-\alpha)} - 2(1-q)\widetilde{\mathrm{R}}_q^{(-\alpha)} + (1-q)^2\widetilde{\mathrm{R}}_q^{(-\alpha-1)}\right] \tag{85}$$

$$= \widetilde{\mathrm{R}}_q^{(\alpha)} + \frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\widetilde{\mathrm{R}}_q^{(\alpha)} - 2(1-q)\widetilde{\mathrm{R}}_q^{(\alpha+1)} + (1-q)^2\widetilde{\mathrm{R}}_q^{(\alpha+2)}\right] \tag{86}$$

So overall, the RDP of subsampled PTR is

$$\frac{1}{\alpha-1}\ln\Bigg(\max\Bigg(1+2q^2\alpha(\alpha-1)\left(\frac{1-\delta_0}{\sigma_1^2}+\frac{\delta_0}{\sigma_2^2}\right), \tag{87}$$

$$\mathrm{R}_q^{(\alpha)}+\frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\mathrm{R}_q^{(\alpha)}-2(1-q)\mathrm{R}_q^{(\alpha-1)}+(1-q)^2\mathrm{R}_q^{(\alpha-2)}\right], \tag{88}$$

$$\widetilde{\mathrm{R}}_q^{(\alpha)}+\frac{2\alpha(\alpha-1)}{\sigma_1^2}\left[\widetilde{\mathrm{R}}_q^{(\alpha)}-2(1-q)\widetilde{\mathrm{R}}_q^{(\alpha+1)}+(1-q)^2\widetilde{\mathrm{R}}_q^{(\alpha+2)}\right]\Bigg)\Bigg) \tag{89}$$

□

A.3 Bound of the Local sensitivity after r adding/removal for Trimmed Sum

Recall that we denote a dataset $S=\{x_1,\ldots,x_m\}$, and $x_{(k)}$ denote the kth smallest data point among S in ℓ_2 norm, i.e., $x_{(1)}\leq x_{(2)}\leq\ldots\leq x_{(m)}$. $\mathrm{TSUM}_F(S)=\sum_{i=1}^{m-F}x_{(i)}$ if $m>F$, or 0 if $m\leq F$.

Theorem 5.1 (Restate). $\mathbb{LS}^{(r)}_{\mathrm{TSUM}_F}(S)=\left\|x_{(m-F+1+r)}\right\|$ if $r\leq F-1$, or $\mathbb{GS}_{\mathrm{TSUM}_F}$ if $r>F-1$.

Proof. The $\mathbb{GS}_{\mathrm{TSUM}_F}$ for $r>F-1$ is trivial as the local sensitivity can never be larger than global sensitivity. When $r\leq F-1$, it is easy to see that the local sensitivity of TSUM_F is just $x_{(m-F+1)}$, as we can add the element with the maximum possible norm x_∞ in the data space to S, so that $\mathrm{TSUM}_F(S\cup\{x_\infty\})=\sum_{i=1}^{m+1-F}x_{(i)}$, and $\|\mathrm{TSUM}_F(S\cup\{x_\infty\})-\mathrm{TSUM}_F(S)\|=\left\|x_{(i)}\right\|$. If the added element has norm smaller than $\left\|x_{(m+1-F)}\right\|$, we will always have $\|\mathrm{TSUM}_F(S\cup\{x_\infty\})-\mathrm{TSUM}_F(S)\|<\left\|x_{(i)}\right\|$. We can easily see that single element removal will also not change $\mathrm{TSUM}_F(S)$ that much. Thus $\mathbb{LS}^{(0)}_{\mathrm{TSUM}_F}(S)=x_{(m-F+1)}$.

To maximize the local sensitivity of S with r elements addition/removal, it's trivial to see that the best strategy is simply adding element with the maximum possible norm x_∞ in the data space to S, and the local sensitivity for the changed dataset $\tilde{S}=S\cup(\{x_\infty\}\times r)$ has local sensitivity $\left\|x_{(m+r)-F+1}\right\|=\|x_{m-F+1+r}\|$ as long as $r\leq F-1$. □

A.4 Convergence Guarantee of PTR-based Gradient Aggregation under Byzantine Failure

Settings of Robust Training. We denote the target loss function as $\mathcal{L}(w)=\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(w,z)\right]$, where $w\in\mathbb{R}^d$ is the model parameters and z is a data point randomly drawn from some distribution $\mathcal{D}$. We assume ℓ is R-Lipschitz, β-smooth and α-strongly convex. We have n stochastic gradient oracles $g_1,\ldots,g_n$, where at each iteration t, for every non-corrupted gradient oracles i, it is an unbiased estimator $g_i^{(t)}$ for the gradient of the global expected loss function with respect to the current model parameters w_t, i.e., $\mathbb{E}[g_i^{(t)}]=\nabla\mathcal{L}(w^{(t)})$. We additionally assume that non-corrupted stochastic gradients have bounded variance, i.e., for some $\sigma>0$ we have

$$\mathbb{E}_{g_i^{(t)}}\left[\left\|g_i^{(t)}-\mathbb{E}[g_i^{(t)}]\right\|^2\right]\leq\sigma^2 \tag{90}$$

at every step t.

Remark A.2. We do not consider the effect of subsampling here for clean presentation. The effect from subsampling could be easily handled by deriving a high probability upper bound for the number of corrupted gradient oracles that will be sampled.

SGD with PTR works as

$$\widehat{\Delta},\tilde{g}^{(t)}\leftarrow\mathrm{PTR}(\{g_1,\ldots,g_n\}) \tag{91}$$

$$w^{(t+1)}\leftarrow w^{(t)}-\left(\eta_A I[\widehat{\Delta}<\log(1/(2\delta_0))]+\eta_B I[\widehat{\Delta}\geq\log(1/(2\delta_0))]\right)\tilde{g}^{(t)} \tag{92}$$

Further, we call it Routine A if $\widehat{\Delta}$ is small and $\mathrm{PTR}(\{g_1,\ldots,g_n\})=\sum_{i=1}^{n}g_{(i)}^{(t)}+\mathcal{N}(0,\sigma_1\mathbf{1}_d)$, and call it Routine B if $\widehat{\Delta}$ is large and $\mathrm{PTR}(\{g_1,\ldots,g_n\})=\sum_{i=1}^{n-F}g_{(i)}^{(t)}+\mathcal{N}(0,\sigma_2\mathbf{1}_d)$. It makes sense to use a smaller learning rate η_A when the PTR goes to Routine A, and use a larger learning rate η_B for Routine B, since the two routines use different amount of gradient information.

Theorem 5.2 (formal version). Let $w^* \in \operatorname{argmin}_w \mathcal{L}(w)$. If there are at most F gradients being corrupted at each iteration, and if we set $\eta_A = \frac{n-F}{n}\eta_B$ and $\sigma_2^2 \geq \frac{(n-F)(n+1)}{n^2}\left(\frac{(n-F)\sigma^2+FR}{d} + \sigma_1^2\right)$, then as $t \to \infty$, we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t)} - w^*\right\|^2\right] \leq \frac{M_B}{1-\rho_B} \tag{93}$$

for

$$0 < \eta_B \leq \frac{2\alpha(n-2F)}{n^2 + (n-F+1)(n-F)\beta^2} \tag{94}$$

where

$$\rho_B = 1 - 2\eta_B\alpha(n-2F) + \eta_B^2(n^2 + (n-F+1)(n-F)\beta^2) \tag{95}$$

$$M_B = \eta_B^2(n-F+1)\left((n-F)\sigma^2 + d\sigma_2^2\right) + \left(\frac{F}{n}\right)^2 R^2 \tag{96}$$

Proof. Let $\eta = \eta_A I[\widehat{\Delta} < \log(1/(2\delta_0))] + \eta_B I[\widehat{\Delta} \geq \log(1/(2\delta_0))]$. By the update rule of parameters in SGD, we have

$$\left\|w^{(t+1)} - w^*\right\|^2 = \left\|w^{(t)} - w^*\right\|^2 - 2\eta\left\langle w^{(t)} - w^*, \tilde{g}^{(t)}\right\rangle + \eta^2\left\|\tilde{g}^{(t)}\right\|^2 \tag{97}$$

where

$$\tilde{g}^{(t)} = \mathrm{PTR}(\{g_1, \ldots, g_n\}) \tag{98}$$

If PTR runs Routine B, then

$$\tilde{g}^{(t)} = \sum_{i=1}^{n-F} g_{(i)}^{(t)} + \xi, \xi \sim \mathcal{N}(0, \sigma_2 \mathbf{I}_d) \tag{99}$$

If PTR runs Routine A, then

$$\tilde{g}^{(t)} = \sum_{i=1}^{n} g_{(i)}^{(t)} + \xi, \xi \sim \mathcal{N}(0, \sigma_1 \mathbf{I}_d) \tag{100}$$

In the following proof, we consider the two routines separately. We denote a set of $n-F$ non-corrupted gradients at step t as $\mathcal{H}^{(t)}$. We use $g_{(i)}^{(t)}$ the ith smallest gradient among the set of gradients at step t, i.e., $g_{(1)}^{(t)} \leq \ldots \leq g_{(n)}^{(t)}$.

Case of Running Routine B. In this case, $\mathbb{E}[\|\xi\|^2] = d\sigma_2^2$. We first upper bound $\left\|\tilde{g}^{(t)}\right\|$:

$$\left\|\tilde{g}^{(t)}\right\| = \left\|\sum_{i=1}^{n-F} g_{(i)}^{(t)} + \xi\right\| \tag{101}$$

$$\leq \sum_{i=1}^{n-F}\left\|g_{(i)}^{(t)}\right\| + \|\xi\| \tag{102}$$

$$\leq \sum_{i \in \mathcal{H}^{(t)}}\left\|g_i^{(t)}\right\| + \|\xi\| \tag{103}$$

Therefore

$$\left\|\tilde{g}^{(t)}\right\|^2 \leq (n-F+1)\left(\sum_{i \in \mathcal{H}^{(t)}}\left\|g_i^{(t)}\right\|^2 + \|\xi\|^2\right) \tag{104}$$

due to AM-QM inequality.

Now we lower bound $\left\langle w^t - w^*, \tilde{g}^{(t)} \right\rangle$. We denote the non-corrupted gradients that are not being trimmed in step t as $\tilde{\mathcal{H}}^{(t)} = \mathcal{H}^{(t)} \cap \{g_{(1)}^{(t)}, \dots, g_{(n-F)}^{(t)}\}$, and we denote the corrupted gradients that are not being trimmed in step t as $\mathcal{B}^{(t)} = \{g_{(1)}^{(t)}, \dots, g_{(n-F)}^{(t)}\} \setminus \tilde{\mathcal{H}}^{(t)}$. Since there are at most F corrupted gradients, we have $|\tilde{\mathcal{H}}^{(t)}| \geq n - 2F$ and $|\mathcal{B}| \leq F$. Since

$$\tilde{g}^{(t)} = \sum_{i \in \tilde{\mathcal{H}}^{(t)}} g_i^{(t)} + \sum_{i \in \mathcal{B}^t} g_i^{(t)} + \xi \tag{105}$$

we have

$$\begin{aligned} \left\langle w^{(t)} - w^*, \tilde{g}^{(t)} \right\rangle &= \sum_{i \in \tilde{\mathcal{H}}^{(t)}} \left\langle w^{(t)} - w^*, g_i^{(t)} \right\rangle + \sum_{i \in \mathcal{B}^t} \left\langle w^{(t)} - w^*, g_i^{(t)} \right\rangle + \left\langle w^{(t)} - w^*, \xi \right\rangle & (106) \\ &\geq \sum_{i \in \tilde{\mathcal{H}}^{(t)}} \left\langle w^{(t)} - w^*, g_i^{(t)} \right\rangle - FR \left\| w^{(t)} - w^* \right\| + \left\langle w^{(t)} - w^*, \xi \right\rangle & (107) \end{aligned}$$

Denote this lower bound as

$$\phi_t = \sum_{i \in \tilde{\mathcal{H}}^{(t)}} \left\langle w^{(t)} - w^*, g_i^{(t)} \right\rangle - FR \left\| w^{(t)} - w^* \right\| + \left\langle w^{(t)} - w^*, \xi \right\rangle \tag{108}$$

Then

$$\left\| w^{(t+1)} - w^* \right\|^2 \leq \left\| w^{(t)} - w^* \right\|^2 - 2\eta\phi_t + \eta^2(n - F + 1) \left(\sum_{i \in \mathcal{H}^{(t)}} \left\| g_i^{(t)} \right\|^2 + \|\xi\|^2 \right) \tag{109}$$

Now this upper bound only include quantities that are independent from corrupted gradients. Take expectation of both sides over $g^{(t)}$ and ξ, we have

$$\begin{aligned} &\mathbb{E}_{g,\xi} \left[\left\| w^{(t+1)} - w^* \right\|^2 \right] & (110) \\ &\leq \left\| w^{(t)} - w^* \right\|^2 - 2\eta \mathbb{E}_{g,\xi}[\phi_t] + \eta^2(n - F + 1)\mathbb{E}_{g,\xi} \left[\sum_{i \in \mathcal{H}^{(t)}} \left\| g_i^{(t)} \right\|^2 + \|\xi\|^2 \right] & (111) \end{aligned}$$

For $\mathbb{E}_{g,\xi}[\phi_t]$, we can obtain its lower bound

$$\begin{aligned} \mathbb{E}_{g,\xi}[\phi_t] &= \sum_{i \in \tilde{\mathcal{H}}^{(t)}} \left\langle w^{(t)} - w^*, \nabla \mathcal{L}(w^{(t)}) \right\rangle - FR \left\| w^{(t)} - w^* \right\| & (112) \\ &\geq \alpha(n - 2F) \left\| w^{(t)} - w^* \right\|^2 - FR \left\| w^{(t)} - w^* \right\| & (113) \end{aligned}$$

since $\mathcal{L}$ is α-strongly convex.

For $\mathbb{E}_{g,\xi} \left[\sum_{i \in \mathcal{H}^{(t)}} \left\| g_i^{(t)} \right\|^2 + \|\xi\|^2 \right]$, since $\mathbb{E}_g \left[\left\| g_i^{(t)} \right\|^2 \right] \leq \sigma^2 + \|\nabla \mathcal{L}(w^t)\|^2$ and $\mathbb{E}[\|\xi\|^2] = d\sigma_2^2$, we have

$$\begin{aligned} &\mathbb{E}_{g,\xi} \left[\sum_{i \in \mathcal{H}^{(t)}} \left\| g_i^{(t)} \right\|^2 + \|\xi\|^2 \right] & (114) \\ &\leq (n - F) \left(\sigma^2 + \left\| \nabla \mathcal{L}(w^{(t)}) \right\|^2 \right) + d\sigma_2^2 & (115) \\ &\leq (n - F) \left(\sigma^2 + \beta^2 \left\| w^{(t)} - w^* \right\|^2 \right) + d\sigma_2^2 & (116) \end{aligned}$$

since $\left\| \nabla \mathcal{L}(w^{(t)}) \right\| = \left\| \nabla \mathcal{L}(w^{(t)}) - \nabla \mathcal{L}(w^*) \right\| \leq \beta \left\| w^{(t)} - w^* \right\|$ by β-smoothness of $\mathcal{L}$.

Plugging in the lower and upper bounds, we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t+1)}-w^*\right\|^2\right] \tag{117}$$

$$\leq \left(1-2\eta\alpha(n-2F)+\eta^2\beta^2(n-F+1)(n-F)\right)\left\|w^{(t)}-w^*\right\|^2 \tag{118}$$

$$+\,2\eta FR\left\|w^{(t)}-w^*\right\| \tag{119}$$

$$+\,\eta^2(n-F+1)\left((n-F)\sigma^2+d\sigma_2^2\right) \tag{120}$$

$$\leq \left(1-2\eta\alpha(n-2F)+\eta^2(n^2+(n-F+1)(n-F)\beta^2)\right)\left\|w^{(t)}-w^*\right\|^2 \tag{121}$$

$$+\,\eta^2(n-F+1)\left((n-F)\sigma^2+d\sigma_2^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{122}$$

where the last step is due to

$$2\eta FR\left\|w^{(t)}-w^*\right\|\leq\left(\frac{F}{n}\right)^2R^2+n^2\eta^2\left\|w^{(t)}-w^*\right\|^2 \tag{123}$$

Let

$$\rho_B=1-2\eta\alpha(n-2F)+\eta^2(n^2+(n-F+1)(n-F)\beta^2) \tag{124}$$

$$M_B=\eta^2(n-F+1)\left((n-F)\sigma^2+d\sigma_2^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{125}$$

Then we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t+1)}-w^*\right\|^2\right]\leq\rho_B\left\|w^{(t)}-w^*\right\|^2+M_B \tag{126}$$

Therefore, as long as $\rho_B<1$, $w^{(t)}$ will eventually

For $\rho_B<1$, we need

$$0<\eta\leq\frac{2\alpha(n-2F)}{n^2+(n-F+1)(n-F)\beta^2} \tag{127}$$

Case of Running Routine A. We follow a similar analysis as for the case of Routine A. In this case, $\mathbb{E}[\|\xi\|^2]=d\sigma_1^2$.

We first upper bound $\left\|\tilde{g}^{(t)}\right\|$:

$$\left\|\tilde{g}^{(t)}\right\|=\left\|\sum_{i=1}^{n}g_{(i)}^{(t)}+\xi\right\| \tag{128}$$

$$\leq\sum_{i=1}^{n}\left\|g_{(i)}^{(t)}\right\|+\|\xi\| \tag{129}$$

Therefore

$$\left\|\tilde{g}^{(t)}\right\|^2\leq(n+1)\left(\sum_{i\in\mathcal{H}^{(t)}}\left\|g_i^{(t)}\right\|^2+\|\xi\|^2\right) \tag{130}$$

In this case, there are no gradients being corrupted, and thus we have $|\tilde{\mathcal{H}}^{(t)}|\geq n-F$. Therefore, for $\mathbb{E}_{g,\xi}[\phi_t]$ we have

$$\mathbb{E}_{g,\xi}[\phi_t]=\sum_{i\in\tilde{\mathcal{H}}^{(t)}}\left\langle w^{(t)}-w^*,\nabla\mathcal{L}(w^{(t)})\right\rangle-FR\left\|w^{(t)}-w^*\right\| \tag{131}$$

$$\geq\alpha(n-F)\left\|w^{(t)}-w^*\right\|^2-FR\left\|w^{(t)}-w^*\right\| \tag{132}$$

For $\mathbb{E}_{g,\xi}\left[\sum_{i\in\mathcal{H}^{(t)}}\left\|g_{(i)}^{(t)}\right\|^2+\sum_{i\in[n]\setminus\mathcal{H}^{(t)}}\left\|g_{(i)}^{(t)}\right\|^2+\|\xi\|^2\right]$, since $\mathbb{E}_g\left[\left\|g_{(i)}^{(t)}\right\|^2\right] \leq \sigma^2 + \left\|\nabla\mathcal{L}(w^{(t)})\right\|^2$ and $\mathbb{E}[\|\xi\|^2]=d\sigma_1^2$, we have

$$\mathbb{E}_{g,\xi}\left[\sum_{i\in\mathcal{H}^{(t)}}\left\|g_{(i)}^{(t)}\right\|^2+\sum_{i\in[n]\setminus\mathcal{H}^{(t)}}\left\|g_{(i)}^{(t)}\right\|^2+\|\xi\|^2\right] \tag{133}$$

$$\leq (n-F)\left(\sigma^2+\left\|\nabla\mathcal{L}(w^{(t)})\right\|^2\right)+FR+d\sigma_1^2 \tag{134}$$

$$\leq (n-F)\left(\sigma^2+\beta^2\left\|w^{(t)}-w^*\right\|^2\right)+FR+d\sigma_1^2 \tag{135}$$

Plugging in the lower and upper bounds, we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t+1)}-w^*\right\|^2\right] \tag{136}$$

$$\leq \left(1-2\eta\alpha(n-F)+\eta^2(n+1)(n-F)\beta^2\right)\left\|w^{(t)}-w^*\right\|^2 \tag{137}$$

$$+2\eta FR\cdot\left\|w^{(t)}-w^*\right\| \tag{138}$$

$$+\eta^2(n+1)\left((n-F)\sigma^2+FR+d\sigma_1^2\right) \tag{139}$$

$$\leq \left(1-2\eta\alpha(n-F)+\eta^2(n^2+(n+1)(n-F)\beta^2)\right)\left\|w^{(t)}-w^*\right\|^2 \tag{140}$$

$$+\eta^2(n+1)\left((n-F)\sigma^2+FR+d\sigma_1^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{141}$$

Follow similar analysis, but we will have

$$\rho_A=1-2\eta\alpha(n-F)+\eta^2(n^2+(n+1)(n-F)\beta^2) \tag{142}$$

$$M_A=\eta^2(n+1)\left((n-F)\sigma^2+FR+d\sigma_1^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{143}$$

So for every time we run Routine A (with full gradient sum), we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t)}-w^*\right\|^2\right]\leq\rho_A\left\|w^{(t-1)}-w^*\right\|^2+M_A \tag{144}$$

And we have $M_B<M_A$, so more routine B can improve the utility.

Overall, we if there are at most F gradients being corrupted at each iteration, we have

$$\mathbb{E}\left[\left\|w^{(t)}-w^*\right\|^2\right]\leq\rho_A\left\|w^{(t-1)}-w^*\right\|^2+M_A \tag{145}$$

for

$$\rho_A=1-2\eta_A\alpha(n-F)+\eta_A^2(n^2+(n+1)(n-F)\beta^2) \tag{146}$$

$$M_A=\eta_A^2(n+1)\left((n-F)\sigma^2+FR+d\sigma_1^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{147}$$

if PTR runs Routine A, and

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t+1)}-w^*\right\|^2\right]\leq\rho_B\left\|w^{(t)}-w^*\right\|^2+M_B \tag{148}$$

for

$$\rho_B=1-2\eta_B\alpha(n-2F)+\eta_B^2(n^2+(n-F+1)(n-F)\beta^2) \tag{149}$$

$$M_B=\eta_B^2(n-F+1)\left((n-F)\sigma^2+d\sigma_2^2\right)+\left(\frac{F}{n}\right)^2R^2 \tag{150}$$

if PTR runs Routine B.

If we set $\eta_A = \frac{n-F}{n}\eta_B$, since

$$\sigma_2^2 \geq \frac{(n-F)(n+1)}{n^2}\left(\frac{(n-F)\sigma^2 + FR}{d} + \sigma_1^2\right) \tag{151}$$

we have $M_A \leq M_B$, and we can also easily verify that $\rho_A \leq \rho_B$. Thus, as $t \to \infty$, we have

$$\mathbb{E}_{g,\xi}\left[\left\|w^{(t)} - w^*\right\|^2\right] \leq \frac{M_B}{1-\rho_B} \tag{152}$$

for

$$0 < \eta_B \leq \frac{2\alpha(n-2F)}{n^2 + (n-F+1)(n-F)\beta^2} \tag{153}$$

□

A.5 Pseudo-code of TSGD+PTR

The pseudo-code of TSGD+PTR is shown in Algorithm 2, which uses Algorithm 3 (`PTR-TMEAN`) as a subroutine.

Algorithm 2: Private Trimmed-mean SGD with Propose-Test-Release

```
input : Dataset $\{\mathbf{z}^1, \dots, \mathbf{z}^N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, \mathbf{z}^i)$, learning rate $\eta$, batch size
        $B$, sensitivity bound $\tau$, Clipping threshold $R$, noise multiplier $\sigma$.
Initialize $\theta_0$ randomly.
for $t \in [T]$ do
    Random Subsampling.
    Take a random batch $\mathcal{B}_t$ with sampling probability $q$ in Poisson subsampling.
    Obtain Gradients.
    For each $i \in \mathcal{B}_t$, get (potentially faulty) $g_i^{(t)}$.
    Gradient Clipping.
    $g_i^{(t)} \leftarrow C \cdot g_i^{(t)}$ for $C = \min\left(1, R/\left\|g_i^{(t)}\right\|_2\right)$.
    Noisy Gradient Aggregation with PTR.
    $\tilde{g}^{(t)} \leftarrow$ PTR-TMEAN $\left(\{g_i^{(t)}\}\right)$.
    Descent.
    $\theta_{t+1}, \omega \leftarrow \theta_t - \eta\tilde{g}^{(t)}$.
    Adjust $F$.
    if $\omega$ is '+' then
        Increase $F$.
    else
        Decrease $F$.
```

Algorithm 3: `PTR-TMEAN`

```
input : $S$ – Set of (clipped) gradient vectors at step $t$: $\{g_i^{(t)}\} \subseteq \mathbb{R}^d$,
$\Delta \leftarrow \min_{\tilde{S} \in \{\tilde{S}: \mathbb{LS}_{f_2}(\tilde{S}) > \tau\}} d\left(S, \tilde{S}\right)$.
$\widehat{\Delta} \leftarrow \Delta + \mathrm{Lap}(0, b)$.
if $\widehat{\Delta} \leq \log(1/(2\delta_0))b$ then
    return $\mathrm{SUM}(S) + \sigma R \cdot \mathcal{N}(0, \mathbf{1}_d)$, '+'
else
    return $\mathrm{TSUM}_F(S) + \sigma\tau \cdot \mathcal{N}(0, \mathbf{1}_d)$, '−'
```

A.5.1 Why not directly apply PTR to regular SGD?

As we discussed in the main text, PTR typically works with robust statistics such as trimmed mean. Regular SGD use mean as gradient aggregation function. Mean, however, does not have a low local sensitivity on most of the inputs. Therefore, we focus on the application of PTR in privatizing robust statistics.

B Experiment Settings & Additional Results

B.1 Experiment Settings for Table 1

B.1.1 Corruption Simulation.

Following the literature in Byzantine robustness [YCKB18, XKG19, AHJ$^+$21, GLV21], we consider three possible sources of Byzantine failures: corruption in features, labels and communicated gradients. All experiments are repeated for 0% (i.e., clean), 10%, and 20% corruption ratio (CR).

Feature Corruption. Corruption in Features can arise from the process of data collection. Following [AHJ$^+$21], we adopt the additive corruption introduced in [HD18]. Specifically, we add Gaussian noise from $\mathcal{N}(0, 100)$ directly to the corrupted images.

Gradient Corruption. Gradient can be corrupted in distributed SGD, e.g., due to hardware malfunction or malicious users. We consider the gradient corruption following [XKG19, AHJ$^+$21], where we add Gaussian noise from $\mathcal{N}(0, 100)$ to the true gradients.

Label Corruption. Noisy labels are pervasive in the dataset. We randomly flip of label of certain amount of data points.

B.1.2 Datasets & Models.

MNIST [LeC98] is one of the most commonly used benchmark datasets in deep learning containing 70000 handwritten digit images. CIFAR-10 [Kri09] is another classic benchmark for image classification. It consists of 60000 images from 10 different classes with 6000 images each. EMNIST [CATVS17] is similar to MNIST but has a much larger size (145,600 character images and 26 balanced classes).

In Table 1, all models are trained entirely from scratch. For all datasets, we use a small CNN whose architecture is inherited from the official tutorial of tensorflow/privacy[9].

B.1.3 Hyperparameters.

For `TSGD+PTR`, we set $\delta_0 = 10^{-8}, b = 1$. For MNIST and EMNIST, we set gradient clipping bound $R = 1$, $\tau = 0.5$, noise multiplier $\sigma = 1.1$ for `TSGD+PTR`, and $\sigma = 0.7$ for `TSGD+Gaussian`. The noise multiplier for `TSGD+PTR` and `TSGD+Gaussian` are picked differently in order to align their privacy loss in each iteration. For CIFAR10, we set gradient clipping bound $R = 3$, $\tau = 2$, noise multiplier $\sigma = 1.1$ for `TSGD+PTR`, and $\sigma = 0.9$ for `TSGD+Gaussian`.

For MNIST and EMNIST dataset, we set the learning rate as 0.15, batch size as 256; for CIFAR10 dataset, we set the learning rate as 0.1, batch size as 1024.

We set F to be 25% of the batch size for `TSGD+Gaussian`. For `TSGD+PTR`, F is dynamically adjusted based on the value of $\widehat{\Delta}$. If sensitivity test is passed, we increase F by $0.02 \times \mathrm{batchsize}$, and if sensitivity test is failed, we decrease F by the same amount.

All of our experiments are performed on Tesla P100-PCIE-16GB GPU.

B.2 Additional Results on More Architectures

We experiment with more architectures on CIFAR10 dataset. Specifically, we use two famous, moderately large architecture ResNet18 [HZRS16] and VGG11 [SZ14]. We follow the common procedure in prior works [ACG$^+$16]: we use ResNet18 and VGG11 that are pretrained by ImageNet dataset. The pre-training weight is publicly available from PyTorch. We only finetune the last layer of the model.

We set gradient clipping bound $R = 5$, batch size as 2048, learning rate 0.01. For `TSGD+PTR`, we set $\delta_0 = 10^{-8}, b = 1$, and $\tau = 3$. We set noise multiplier $\sigma = 2.2$ for `TSGD+PTR`, and $\sigma = 1.8$ for `TSGD+Gaussian`. We set F to be 25% of the batch size for `TSGD+Gaussian`. F is dynamically adjusted in the same way as the experiments in the main text.

---

[9] https://github.com/tensorflow/privacy

Archi.	Corruption Type	CR	$\varepsilon = 3.0$		$\varepsilon = 5.0$	
			TSGD+Gaussian	**TSGD+PTR**	**TSGD+Gaussian**	**TSGD+PTR**
VGG11		**0**	50.05%	52.09% (+2.04%)	51.75%	52.85% (+1.1%)
	Label	**0.1**	44.03%	48.73% (+4.7%)	48.78%	50.24% (+1.46%)
		0.2	35.04%	43.63% (+8.59%)	43.17%	46.35% (+3.18%)
	Feature	**0.1**	45.59%	49.57% (+3.98%)	49.60%	50.74% (+1.14%)
		0.2	43.82%	47.95% (+4.13%)	48.08%	48.88% (+0.8%)
	Gradient	**0.1**	45.61%	50.15% (+4.54%)	50.10%	51.15% (+1.05%)
		0.2	45.40%	50.15% (+4.75%)	50.46%	50.82% (+0.36%)
			$\varepsilon = 3.0$		$\varepsilon = 5.0$	
			TSGD+Gaussian	**TSGD+PTR**	**TSGD+Gaussian**	**TSGD+PTR**
ResNet18		**0**	38.97%	43.15% (+3.18%)	46.27%	47.15% (+0.88%)
	Label	**0.1**	35.84%	42.55% (+6.71%)	42.05%	44.77% (+2.72%)
		0.2	26.15%	36.48% (+10.33%)	33.97%	39.71% (+5.74%)
	Feature	**0.1**	37.91%	43.15% (+5.24%)	42.54%	44.93% (+2.39%)
		0.2	36.30%	41.85% (+5.55%)	41.21%	43.84% (+2.63%)
	Gradient	**0.1**	38.42%	44.27% (+5.85%)	43.83%	46.4% (+2.57%)
		0.2	37.79%	44.34% (+6.55%)	43.22%	46.14% (+2.92%)

Table 2: Model Accuracy under different privacy budgets and corruption settings. Every statistic is averaged over 5 runs with different random seed. The improvement of $\mathrm{TSGD} + \mathrm{PTR}$ over $\mathrm{TSGD} + \mathrm{Gaussian}$ is highlighted in the red text.

Corruption Type	CR	$\varepsilon = 2.0$		$\varepsilon = 2.5$	
		TSGD+Gaussian	**TSGD+PTR**	**TSGD+Gaussian**	**TSGD+PTR**
Targeted Label Flip	**0**	72.94%	76.44% (+3.5%)	79.15%	81.02% (+1.87%)
	0.1	72.06%	75.11% (+3.05%)	76.25%	78.97% (+2.72%)
	0.2	70.72%	73.85% (+3.13%)	73.69%	76.67% (+2.98%)
Gradient Bit Flip	**0.1**	69.58%	75.67% (+6.09%)	75.39%	78.82% (+3.43%)
	0.2	65.13%	75.67% (+10.54%)	71.85%	78.85% (+7.0%)

Table 3: Model Accuracy on EMNIST dataset under different privacy budgets on two more severe types of failures.

The results are shown in Table 2. As we can see, `TSGD+PTR` consistently outperforms `TSGD+Gaussian` across different architectures.

B.3 Additional Results on More Corruption Types

Besides the three corruption types we considered in the maintext, we evaluate on two additional possible errors which are considered more severe kinds of failure.

1. **Gradient Bit-flipping failure** where the bits that control the sign of the floating numbers are flipped, e.g., due to some hardware failure. A faulty worker pushes the negative gradient instead of the true gradient to the servers.
2. **Targeted label flipping failure** where the labels are flipped in a "targeted" way, i.e., for any *label* $\in \{0, \ldots, 25\}$, is replaced by $25-$*label*. Such failures/attacks can be caused by data poisoning or software failures.

We experiment on EMNIST dataset and the results are shown in Table 3. The experiment settings are exactly the same as the settings for Table 1. As we can see, `TSGD+PTR` once again outperform `TSGD+Gaussian` significantly.

B.4 Comparison between regular DPSGD and trimmed-mean based robust SGD

We additionally show the comparison between trimmed mean robust SGD with/without PTR and regular DPSGD in Table 4 on EMNIST dataset with the same experiment settings described before. As we can see, the robust SGD performs worse than non-robust counterpart on clean training data. This is because when the training data are clean, the outliers filtered out by robust SGD in the gradient batch are usually corresponding to the data points that are misclassified, which are important for

Corruption Type	CR	$\varepsilon = 2.0$ TSGD+Gaussian	DPSGD	TSGD+PTR	$\varepsilon = 2.5$ TSGD+Gaussian	DPSGD	TSGD+PTR
Label	**0**	72.94%	77.29%	76.44% (-0.85%)	79.15%	83.06%	81.02% (-2.04%)
	0.1	72.60%	74.80%	75.66% ($+0.86\%$)	79.38%	79.30%	80.63% ($+1.33\%$)
	0.2	70.03%	71.42%	72.62% ($+1.2\%$)	77.48%	77.43%	79.19% ($+1.76\%$)
Feature	**0.1**	69.60%	74.58%	74.01% (-0.57%)	77.80%	80.95%	81.04% ($+0.09\%$)
	0.2	70.04%	73.79%	74.76% ($+0.97\%$)	78.99%	79.43%	80.74% ($+1.31\%$)
Gradient	**0.1**	71.75%	72.97%	76.19% ($+3.22\%$)	77.32%	76.59%	77.73% ($+1.14\%$)
	0.2	70.76%	71.22%	74.65% ($+3.43\%$)	76.68%	75.69%	77.17% ($+1.48\%$)

Table 4: Model accuracy comparison with regular DPSGD on EMNIST dataset. The improvement of TSGD + PTR over regular DPSGD is highlighted in the red text.

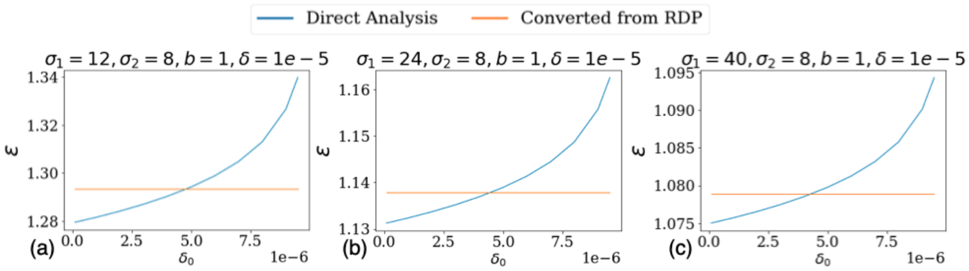

Figure 4: The ε parameter of the (ε, δ)-DP guarantee of PTR when $\delta = 10^{-5}$ for different noise scales. We convert the RDP bound in Theorem 4.3 to (ε, δ)-DP by the RDP-DP conversion formula from [BBG+20], and compare it with the ε obtained from the direct analysis in Theorem 4.2. For the bound converted from RDP, we search for the optimal $\alpha \in [1, 200]$. The bound is constant across different δ_0 since when δ_0 is small, the RDP for PTR will take the second term in (3).

improving model performance. However, `TSGD+PTR` achieves better performance on most of the corruption settings.

B.5 Additional Results on Privacy Analysis Comparison

In this section, we show more numerical results on privacy analysis comparison by varying $\tau = \sigma_2/\sigma_1$. In Figure 4, we numerically compute the privacy bound from direct analysis and the one converted from RDP, with $\tau \in [2/3, 1/3, 0.1]$. In Figure 5, we show the subsampled privacy bound composed with moment account, also with $\tau \in [2/3, 1/3, 0.1]$. As we can see, our Theorem 4.3 and Theorem 4.4 once again provide tighter bounds compared with the baseline.

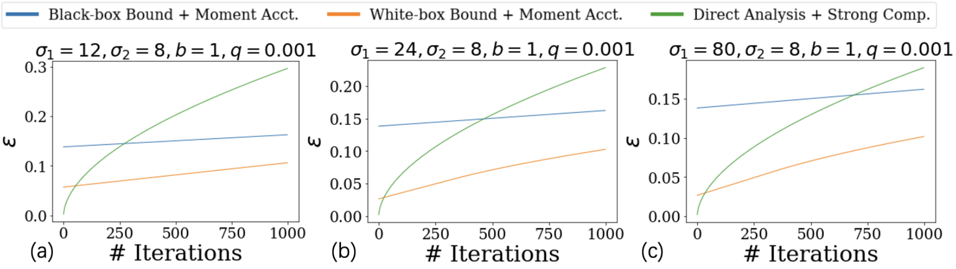

Figure 5: Illustration of the use of our Theorem 4.4 in moments accountant. We plot the the privacy loss ε for $\delta = 10^{-5}$ after different rounds of composition. We set $\delta_0 = 10^{-8}$ here to allow more iterations for Strong Composition of (ε, δ)-DP.