F Proofs for Theorem 3, 7, and 8

F.1 Proof of Theorem 3

To compute the Privacy Loss Random Variable (PLRV), we must determine the probability that the privacy loss is exactly ε over all possible outcomes $o \in \mathsf{R}$ and neighboring datasets $d, d' \subset \mathcal{D}$, assuming that d and d' are selected uniformly from $\mathcal{D}$.

First, consider the universe of items $\mathcal{D}$ with $|\mathcal{D}| = N$. Two datasets $d, d' \subset \mathcal{D}$ are neighboring if they differ by at most one element. For a subset d of size k, there are $N-k$ ways to add an element (forming d' of size $k+1$) and k ways to remove an element (forming d' of size $k-1$). The number of ways to choose a subset d of size k is $\binom{N}{k}$. Thus, the total number of neighboring pairs (d, d') is:

$$\sum_{k=0}^{N}\left((N-k)\binom{N}{k} + k\binom{N}{k}\right) = N + \sum_{k=1}^{N} k\binom{N}{k}.$$

Using the identity $\sum_{k=1}^{N} k\binom{N}{k} = N \cdot 2^{N-1}$, we obtain:

$$\text{Total pairs} = N + N \cdot 2^{N-1} = N(1 + 2^{N-1}).$$

Therefore, the denominator $|\mathcal{D}|(1 + 2^{|\mathcal{D}|-1})$ represents the total number of neighboring dataset pairs.

Next, under the assumption that d and d' are chosen uniformly, the PLRV $C_{q,\text{aux}}(\varepsilon)$ represents the probability distribution of the privacy loss value ε. For a fixed outcome o and neighboring datasets d, d', the indicator function $\mathbf{1}_\varepsilon(c(o; M_q, \text{aux}, d, d'))$ is 1 if the privacy loss is ε and 0 otherwise. The sum $\sum_{\forall d, d' \subset \mathcal{D}} \mathbf{1}_\varepsilon(c(o; M_q, \text{aux}, d, d'))$ counts the number of neighboring pairs for which the privacy loss equals ε for a given outcome o. Dividing by the total number of pairs $|\mathcal{D}|(1 + 2^{|\mathcal{D}|-1})$ gives the proportion of such pairs.

Finally, to compute the overall probability that the privacy loss equals ε, we integrate over all possible outcomes $o \in \mathsf{R}$, weighted by the probability of each outcome $\mathbb{P}[M_q(\text{aux}, d) = o]$:

$$C_{q,\text{aux}}(\varepsilon) = \int_{\mathsf{R}} \frac{\sum_{\forall d, d' \subset \mathcal{D}} \mathbf{1}_\varepsilon(c(o; M_q, \text{aux}, d, d'))}{|\mathcal{D}|(2 + 2^{|\mathcal{D}|})} \mathbb{P}[M_q(\text{aux}, d) = o]\, do.$$

This integral represents the expected proportion of outcomes for which the privacy loss equals ε, thereby defining the PLRV as a probability distribution over $\mathsf{R}_{\geq 0}$. □

F.2 Proof of Theorem 7

The KL divergence of the model parameters on two neighboring datasets $q(D)$ and $q(D')$ can be directly obtained from the moment accounting function $\alpha(\lambda)$ by taking its derivative with respect to λ and evaluating it at $\lambda = 0$. From the definition of the moment accounting function (MAF):

$$\alpha(\lambda) = \log \mathbb{E}_{o \sim M_q(\text{aux}, d)}\left[\exp\left(\lambda c(o; M_q, \text{aux}, d, d')\right)\right],$$

where $c(o; M_q, \text{aux}, d, d') = \log \frac{M_q(\text{aux}, d)}{M_q(\text{aux}, d')}$. The KL divergence between the output distributions of the mechanism $M_q(\text{aux}, d)$ and $M_q(\text{aux}, d')$ is defined as:

$$\text{KL}(M_q(\text{aux}, d) \| M_q(\text{aux}, d')) = \mathbb{E}_{o \sim M_q(\text{aux}, d)}[c(o; M_q, \text{aux}, d, d')].$$

From the above expressions, the KL divergence can be interpreted as the derivative of $e^{\alpha(\lambda)}$ with respect to λ, evaluated at $\lambda = 0$:

$$\text{KL}(M_q(\text{aux}, d) \| M_q(\text{aux}, d')) = \left.\frac{\partial e^{\alpha(\lambda)}}{\partial \lambda}\right|_{\lambda=0}.$$

From Theorem 6, we have:

$$e^{\alpha(\lambda)} = \frac{(\lambda+1) \cdot \mathcal{M}_u(\lambda \cdot \Delta_1 q) + \lambda \cdot \mathcal{M}_u(-(\lambda+1) \cdot \Delta_1 q)}{2\lambda + 1}.$$

Define:

$$g(\lambda) = (\lambda+1) \cdot \mathcal{M}_u(\lambda \cdot \Delta_1 q) + \lambda \cdot \mathcal{M}_u(-(\lambda+1) \cdot \Delta_1 q),$$

$$h(\lambda) = 2\lambda + 1.$$

The derivative is given by:

$$\frac{\partial e^{\alpha(\lambda)}}{\partial \lambda} = \frac{g'(\lambda)h(\lambda) - g(\lambda)h'(\lambda)}{g(\lambda)h(\lambda)}.$$

Evaluate $g(0)$, $g'(0)$, $h(0)$, and $h'(0)$ At $\lambda = 0$:

$$g(0) = (0+1) \cdot \mathcal{M}_u(0 \cdot \Delta_1 q) + 0 \cdot \mathcal{M}_u(-1 \cdot \Delta_1 q) = \mathcal{M}_u(0),$$

$$h(0) = 1, \quad h'(0) = 2.$$

The derivative of $g(\lambda)$ is:

$$g'(\lambda) = \mathcal{M}_u(\lambda \cdot \Delta_1 q) + (\lambda+1) \cdot \mathcal{M}'_u(\lambda \cdot \Delta_1 q) \cdot \Delta_1 q + \mathcal{M}_u(-(\lambda+1) \cdot \Delta_1 q) - \lambda \cdot \mathcal{M}'_u(-(\lambda+1) \cdot \Delta_1 q) \cdot \Delta_1 q.$$

At $\lambda = 0$:

$$g'(0) = \mathcal{M}_u(0) + \mathcal{M}'_u(0) \cdot \Delta_1 q + \mathcal{M}_u(-\Delta_1 q).$$

Substitute $g(0) = \mathcal{M}_u(0)$, $g'(0) = \mathcal{M}_u(0) + \mathcal{M}'_u(0) \cdot \Delta_1 q + \mathcal{M}_u(-\Delta_1 q)$, $h(0) = 1$, and $h'(0) = 2$:

$$\frac{\partial e^{\alpha(\lambda)}}{\partial \lambda} = \frac{(\mathcal{M}_u(0) + \mathcal{M}'_u(0) \cdot \Delta_1 q + \mathcal{M}_u(-\Delta_1 q)) \cdot 1 - (\mathcal{M}_u(0) \cdot 2)}{\mathcal{M}_u(0) \cdot 1}.$$

Simplify:

$$\frac{\partial e^{\alpha(\lambda)}}{\partial \lambda} = \frac{-\mathcal{M}_u(0) + \mathcal{M}'_u(0) \cdot \Delta_1 q + \mathcal{M}_u(-\Delta_1 q)}{\mathcal{M}_u(0)}.$$

Since for MGF, we always have $\mathcal{M}_u(0) = 1$:

$$\text{KL}(M_q(\text{aux}, d) \| M_q(\text{aux}, d')) = -1 + \mathcal{M}'_u(0) \cdot \Delta_1 q + \mathcal{M}_u(-\Delta_1 q).$$

F.3 Proof of Theorem 8

We aim to find an upper bound for the ratio:

$$\frac{\mathbb{E}_X(X)}{\mathbb{E}_X(e^{-\Delta q \cdot X})},$$

where $X > 0$ is a positive random variable. Using Jensen's inequality on the exponential function, we know:

$$\mathbb{E}_X(e^{-\Delta q \cdot X}) \geq e^{-\Delta q \cdot \mathbb{E}_X(X)}.$$

Hence:

$$\frac{\mathbb{E}_X(X)}{\mathbb{E}_X(e^{-\Delta q \cdot X})} \leq \mathbb{E}_X(X) \cdot e^{\Delta q \cdot \mathbb{E}_X(X)}.$$

$$\frac{\mathbb{E}_X(X)}{\mathbb{E}_X(e^{-\Delta q \cdot X})} \leq \sqrt{\mathbb{E}_X(X^2)} \cdot e^{\Delta q \cdot \mathbb{E}_X(X)}.$$

A tighter bound can be given by Hölder's inequality. Hölder's inequality states that for a random variable X and Y, and any $p > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$:

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{1/p} \cdot (\mathbb{E}[|Y|^q])^{1/q}.$$

Setting $p = 2$ and $q = 2$, we get:

$$\mathbb{E}_X(X) \leq \sqrt{\mathbb{E}_X(X^2 e^{2\Delta q \cdot X})} \cdot \sqrt{\mathbb{E}_X(e^{-2\Delta q \cdot X})}.$$

Divide both sides by $\mathbb{E}_X(e^{-\Delta q \cdot X})$:

$$\frac{\mathbb{E}_X(X)}{\mathbb{E}_X(e^{-\Delta q \cdot X})} \leq \sqrt{\mathbb{E}_X(X^2 e^{2\Delta q \cdot X})} \cdot \frac{\sqrt{\mathbb{E}_X(e^{-2\Delta q \cdot X})}}{\mathbb{E}_X(e^{-\Delta q \cdot X})}.$$

By Jensen inequality for Concave functions:

$$\frac{\sqrt{\mathbb{E}_X(e^{-2\Delta q \cdot X})}}{\mathbb{E}_X(e^{-\Delta q \cdot X})} \leq 1.$$

The final bound is given by:

$$\frac{\mathbb{E}_X(X)}{\mathbb{E}_X(e^{-\Delta q \cdot X})} \leq \sqrt{\mathbb{E}_X(X^2 e^{2\Delta q \cdot X})}.$$

G Detailed Answers to Reviewers' Comments

G.1 Algorithm 1 (Review A-B)

Algorithm 1 optimizes over a **simplified** version of Eq.13.

In Eq.13, we minimize the following expression: $(\Delta_1 q \mathcal{M}'_u(0) + \mathcal{M}_u(-\Delta_1 q) - 1)/(\mathcal{M}'_u(0)/\mathcal{M}_u(-\Delta_1 q))$, where $\mathcal{M}'_u(0)$ **represents the "mean of the expected values of the three PDFs"**.

To simplify this, rewrite the expression by moving the denominator to the numerator:

$$\mathcal{M}_u(-\Delta_1 q) \cdot \left(\Delta_1 q + \frac{\mathcal{M}_u(-\Delta_1 q) - 1}{\mathcal{M}'_u(0)}\right).$$

While the privacy engine (see **Theorem.6**) minimizes $\mathcal{M}_u(\cdot)$ during moment accounting, i.e., $\min_\lambda \frac{\mathcal{M}_u(\lambda \Delta q)}{e^{\lambda \varepsilon}}$, the **mean of "u"** is much more flexible to optimize as it directly relates to the derivative of $\mathcal{M}_u$.

Recall that the derivative of the moment-generating function of a random variable "u" at 0, $\mathcal{M}'_u(0)$, is its mean $\mathcal{M}'_u(0) = E(u)$. Thus, our objective simplifies to minimizing $O(1/\mathcal{M}'_u(0))$, which means **maximizing the mean of "u"**.

G.2 Why KL Over Adjacent Data (Review C)

Fine-tuning updates are given by: $W(\text{fine-tuned}) = W(\text{pretrained}) + \Delta W$ with $W(\text{pretrained})$ frozen, ΔW is trained using DPSGD. The noisy update is: $\Delta W = \Delta W(\mathcal{D}) + n, \quad n \sim \text{Lap}(b, b \sim f)$, with n ensuring eps,delta-DP. The divergence between finetuned model and pretrained can be controlled by minimizing **how big** each single NEW sample can modify KL divergence (similar to DP). To minimize the impact of fine-tuning over general features, we need to bound the following divergence: $\text{KL}(\Delta W + W_{\text{pre-trained}} \,\|\, W_{\text{pre-trained}})$.

Specifically, we consider finetuning as |data size| number incremental updates first sample, second sample, ...,|Data size|. By minimizing KL over each two neighboring data (similar to DP) we minimize each of those atomic steps, thus we minimize $\forall\, D, D' \;\; \text{KL}(\Delta W(\mathcal{D}) + n \,\|\, \Delta W(\mathcal{D}') + n)$.

G.3 Clarifying Utility Function (Review C)

We seek to balance two critical objectives: 1) preserving the intrinsic representations learned from the pre-trained model by minimizing the Kullback-Leibler (KL) divergence between the noise-perturbed fine-tuned model distribution and the original pre-trained distribution, thus ensuring that the noise minimally distorts the overall distribution $P_{\text{LM}}(\theta)$ defined by the pre-trained LLM parameters θ and 2) minimizing the task-specific loss on the downstream task by minimizing the task-specific loss over the downstream task via maximizing the **product of the mean and the concentration** of PLRV to ensure sampling smaller b for smaller noise to perturb.