Appendix A: Differential Privacy Background

Here we provide background information on the definition of algorithmic privacy and a composition method that we will use in our algorithm, as well as the general formulation of the MAC algorithm.

Differential privacy

Differential privacy (DP) is a formal definition of the privacy properties of data analysis algorithms [5]. Given an algorithm $\mathcal{M}$ and neighbouring datasets $\mathcal{D}$, $\mathcal{D}'$ differing by a single entry. Here, we focus on the inclusion-exclusion[1] case, i.e., the dataset $\mathcal{D}'$ is obtained by excluding one datapoint from the dataset $\mathcal{D}$. The *privacy loss* random variable of an outcome o is $L^{(o)} = \log \frac{Pr(\mathcal{M}_{(\mathcal{D})}=o)}{Pr(\mathcal{M}_{(\mathcal{D}')}=o)}$. The mechanism $\mathcal{M}$ is called ε-DP if and only if $|L^{(o)}| \leq \varepsilon, \forall o$. A weaker version of the above is (ε, δ)-DP, if and only if $|L^{(o)}| \leq \varepsilon$, with probability at least $1 - \delta$. What the definition states is that a single individual's participation in the data do not change the output probabilities by much, which limits the amount of information that the algorithm reveals about any one individual.

The most common form of designing differentially private algorithms is by adding noise to a quantity of interest, e.g., a deterministic function $h : \mathcal{D} \mapsto \mathbb{R}^p$ computed on sensitive data $\mathcal{D}$. See [5] and [17] for more forms of designing differentially-private algorithms. For privatizing h, one could use the *Gaussian mechanism* [16] which adds noise to the function, where the noise is calibrated to h's *sensitivity*, S_h, defined by the maximum difference in terms of L2-norm, $\|h(\mathcal{D}) - h(\mathcal{D}')\|_2$, $\tilde{h}(\mathcal{D}) = h(\mathcal{D}) + \mathcal{N}(0, S_h^2\sigma^2\mathbf{I}_p)$, where $\mathcal{N}(0, S_h^2\sigma^2\mathbf{I}_p)$ means the Gaussian distribution with mean 0 and covariance $S_h^2\sigma^2\mathbf{I}_p$. The perturbed function $\tilde{h}(\mathcal{D})$ is (ε, δ)-DP, where $\sigma \geq \sqrt{2\log(1.25/\delta)}/\varepsilon$. In this paper, we use the Gaussian mechanism to achieve differentially private network weights. Next, we describe how the cumulative privacy loss is calculated when we use the Gaussian mechanism repeatedly during training.

The moments accountant

In the moments accountant, a cumulative privacy loss is calculated by bounding the moments of $L^{(o)}$, where the λ-th moment is defined as the log of the moment generating function evaluated at λ [6]: $\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \log \mathbb{E}_{o\sim\mathcal{M}(\mathcal{D})}\left[e^{\lambda L^{(o)}}\right]$. By taking the maximum over the neighbouring datasets, we obtain the worst case λ-th moment $\alpha_{\mathcal{M}}(\lambda) = \max_{\mathcal{D},\mathcal{D}'} \alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}')$, where the form of $\alpha_{\mathcal{M}}(\lambda)$ is determined by the mechanism of choice. The moments accountant compute $\alpha_{\mathcal{M}}(\lambda)$ at each step. Due to the composability theorem which states that the λ-th moment composes linearly (See the composability theorem: Theorem 2.1 in [6] when independent noise is added in each step, we can simply sum each upper bound on $\alpha_{\mathcal{M}_j}$ to obtain an upper bound on the total λ-th moment after T compositions, $\alpha_{\mathcal{M}}(\lambda) \leq \sum_{j=1}^{T} \alpha_{\mathcal{M}_j}(\lambda)$. Once the moment bound is computed, we can convert the λ-th moment to the (ε, δ)-DP, guarantee by, $\delta = \min_\lambda \exp\left[\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon\right]$, for any $\varepsilon > 0$. See Appendix A in [6] for the proof.

[1] This is for using the moments accountant method when calculating the cumulative privacy loss.

Appendix B: Experiment Results

	DP-SGD	DP-CDBN	**DP-MAC**
$\varepsilon = 0.5$	0.90	0.92	0.90
# epochs	16	162	10
$\varepsilon = 2$	0.95	0.95	0.95
# epochs	120	162	30
$\varepsilon = 8$	0.97		0.97
# epochs	700		30

(a) Test classification accuracy on MNIST

	DP-SGD	**DP-MAC**
$\varepsilon = 1$	11.8	12.7
$\varepsilon = 2$	9.6	10.9
$\varepsilon = 4$	7.9	9.6
$\varepsilon = 8$	6.4	8.4
$\varepsilon = \infty$	3.6	4.4

(b) Test reconstruction MSE on USPS

Table 1: Test performance of DP-MAC compared to [6] DP-SGD and DP-CDBN [11] at $\delta = 10^{-5}$

	DP-MAC Classifier	DP-MAC Autoencoder	DP-SGD Autoencoder
layer-sizes	300	300-100-20-100-300-256	300-100-20-100-300-256
batch size	1000	500 (250 if $\varepsilon \leq 2$)	500 (250 if $\varepsilon \leq 2$)
train epochs	30 (10 if $\varepsilon = 0.5$)	50	100
optimizer	Adam	Adam	SGD
$\mathbf{W}$ learning rate	0.01 (0.03 if $\varepsilon = 0.5$)	0.003	0.03
$\mathbf{z}$ learning rate	0.003	0.001	
$\mathbf{W}$ lr-decay	0.95 (0.7 if $\varepsilon = 0.5$)	0.97	100 (50 if $\varepsilon \leq 2$)
$\mathbf{z}$-steps	30	30	
$\mathbf{W}$-steps	1	1	
$\Theta_{\partial T}$	0.3	0.001	
Θ_g			0.01
σ values	1.0, 2.8, 8.0	1.8, 3.1, 4.1, 7.8	2.4, 4.3, 5.7, 11.0
σ_{DP-PCA}	4.0, 8.0, 16.0		

Table 2: Training parameters choices for both DP-MAC experiments and the DP-SGD autoencoder comparison

Appendix C: Differences from the Previous Version

The previous version of this paper contained an error in the implementation which mistakenly lowered the necessary amount of noise for a given privacy guarantee and, as a result reported wrong test results. In this version we have corrected this error and made a number of additional changes which are listed below:

- Gradient update
  - Fixed faulty gradient computation, which had reduced effective noise by up to 99% during training.
  - Improved clipping sensitivity by clipping ∂T rather than the norms of the coefficients $\mathbf{b}_k, \mathbf{C}_k$.
  - Reduced analytic sensitivity by 50% by excluding $\frac{1}{2S}$ term from coefficients and making better use of inclusion/exclusion DP.
- Experiments
  - Removed histogram-based layer-wise clipping bound search, which had turned out to be costly in terms of the privacy budget and yield relatively little improvement. Instead all layers now use the same bound.
  - Classifier experiment now uses DP-PCA to reduce input dimensionality as in [6]. overall results stay roughly the same.
  - Autoencoder: Worse results than DP-SGD, likely due to vanishing gradient issues.
  - Replaced softplus activations with ReLUs.
  - Significantly increased batch sizes.
- Notation
  - Denoted Clipping thresholds as Θ to avoid confusion with T_{nkh} terms.
  - Defined coefficients $a_k, \mathbf{b}_k, \mathbf{C}_k$ excluding $\frac{1}{2S}$ term due to changes in sensitivity analysis.

Appendix D: Additional Figures

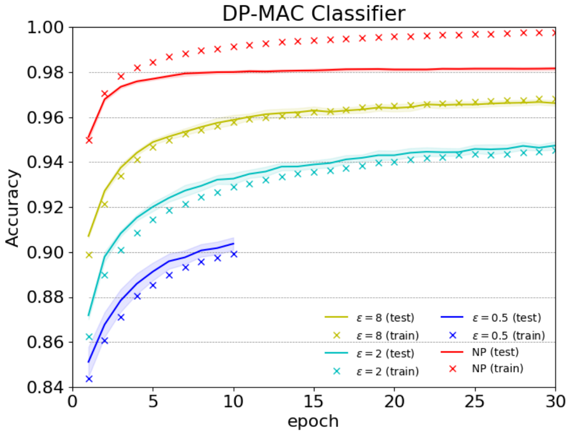

Figure 2: Classifier training and test errors. (± 1 stdev. of the latter) averaged over 10 runs each.

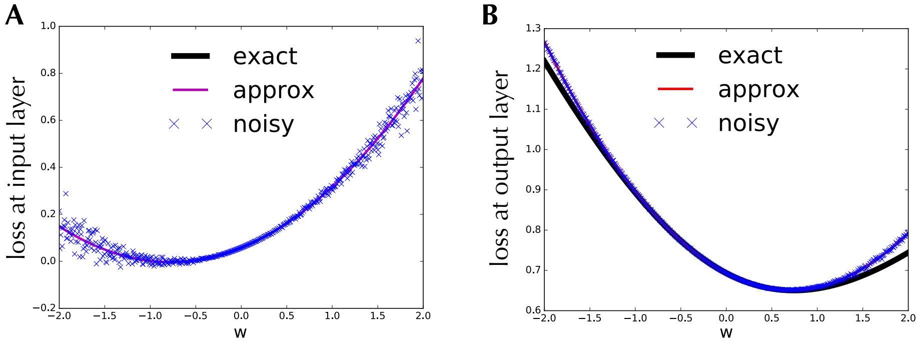

Figure 3: The input and output objective functions (black) are well approximated by the 2nd-order approximations (red). In both cases, approximation is made at 0, where the true w at the input layer is -0.7, and 0.7 at the output layer. The blue crosses depict additive noise centered around the approximated loss and the noise variance is determined by the sensitivities of the coefficients and privacy parameter σ^2.

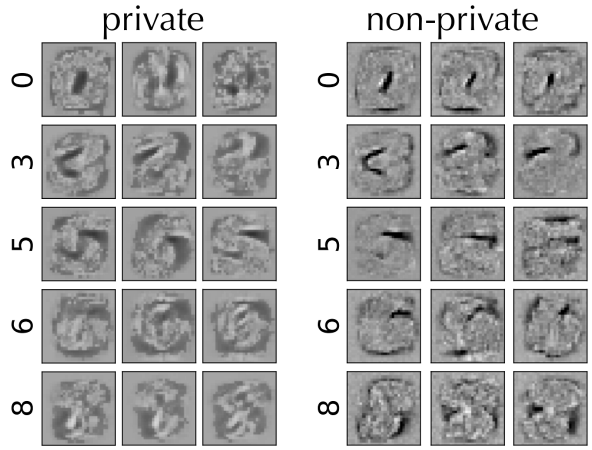

Figure 4: Learned significant features for labels 0,3,5,6,8 respectively. The non-private features show higher contrast and more characteristics in the high frequencies, whereas the private features become smoothed out and lose contrast.

Algorithm 2 DP-MAC with learning T_{b_k}

```
Require: $\mathcal{D}$, $T$, $\sigma^2$, $\sigma^2_{hist}$, $q$, initial threshold $T_{b_k}$
Ensure: $(\varepsilon, \delta)$-DP weights $\{\mathbf{W}_k\}_{k=1}^{K+1}$
  1. Pre-training using DP-MAC (Algo. 1)
  2. DP-histogram release which determines $T_{b_k}$
  3. DP-MAC (Algo. 1) training using learned $T_{b_k}$
```

Appendix E: sensitivity of a_k

We are using a few assumptions and facts to derive sensitivities below.

- $\|\mathbf{z}_{k,s}\|_2 \leq T_z$ for a predefined threshold T_z for all k, s.
- Due to Cauchy-Schwarz inequality: $\mathbf{w}_{kh}^T \mathbf{z}_{k-1,s} \leq \|\mathbf{w}_{kh}\|_2 T_z$
- Using a monotonic nonlinearity (e.g., softplus): $f(\mathbf{w}_{kh}^T \mathbf{z}_{k-1,S}) \leq f(\|\mathbf{w}_{kh}\|_2 T_z)$ and $f'(\mathbf{w}_{kh}^T \mathbf{z}_{k-1,S}) \leq f'(\|\mathbf{w}_{kh}\|_2 T_z)$
- For softplus, $0 < f'(\mathbf{w}_{kh}^T \mathbf{z}_{k-1,S}) \leq f'(\|\mathbf{w}_{kh}\|_2 T_z) \leq 1$ and $0 < f^{''}(\mathbf{w}_{kh}^T \mathbf{z}_{k-1,s}) \leq \frac{1}{4}$
- $\|\mathbf{a}\|_2 \leq \|\mathbf{a}\|_1 \leq \|\mathbf{a}\|_2 \sqrt{D}$ for $\mathbf{a} \in \mathbb{R}^D$
- Direct application of above : $|\sum_{h=1}^{D_{out}^k} z_{kh,s}| \leq \sum_{h=1}^{D_{out}^k} |z_{kh,s}| = \|\mathbf{z}_{k,s}|^\top \mathbf{1}| \leq T_z \sqrt{D_{out}^k}$

Denote

$$\alpha_{\hat{\mathbf{w}}_{kh}} := f(\|\hat{\mathbf{w}}_{kh}\|_2 T_z)$$
$$\beta_{\hat{\mathbf{w}}_{kh}} := f'(\|\hat{\mathbf{w}}_{kh}\|_2 T_z)$$

which we will further denote as vectors $\boldsymbol{\alpha}_{\hat{\mathbf{w}}_k}, \boldsymbol{\beta}_{\hat{\mathbf{w}}_k}$.

Without loss of generality, we further assume that (1): the neighbouring datasets are in the form of $\mathcal{D} = \{\mathcal{D}', (\mathbf{x}_S, \mathbf{y}_S)\}$.

$$
\begin{aligned}
\Delta a_k &= \max_{|\mathcal{D}\backslash\mathcal{D}'|=1} |a_k(\mathcal{D}) - a_k(\mathcal{D}')|, \\
&= \max_{|\mathcal{D}\backslash\mathcal{D}'|=1} | \sum_{h=1}^{D_{out}^k} \{ \sum_{s=1}^{S} (T_n(\hat{\mathbf{w}}_{kh}) - \hat{\mathbf{w}}_{kh}{}^\top \partial T_{skh} + \tfrac{1}{2}\hat{\mathbf{w}}_{kh}{}^\top \partial^2 T_{skh} \hat{\mathbf{w}}_{kh}) - \sum_{s=1}^{S-1} (T_n(\hat{\mathbf{w}}_{kh}) - \hat{\mathbf{w}}_{kh}{}^\top \partial T_{skh} + \tfrac{1}{2}\hat{\mathbf{w}}_{kh}{}^\top \partial^2 T_{skh} \hat{\mathbf{w}}_{kh}) \} | \\
&= \max_{|\mathcal{D}\backslash\mathcal{D}'|=1} | \sum_{h=1}^{D_{out}^k} (T_S(\hat{\mathbf{w}}_{kh}) - \hat{\mathbf{w}}_{kh}{}^\top \partial T_{Skh} + \tfrac{1}{2}\hat{\mathbf{w}}_{kh}{}^\top \partial^2 T_{Skh} \hat{\mathbf{w}}_{kh}) |
\end{aligned}
$$

Now the sensitivity can be divided into three terms due to triangle inequality as

$$
\Delta a_k \leq \underbrace{\max_{|\mathcal{D}\backslash\mathcal{D}'|=1} \left| \sum_{h=1}^{D_{out}^k} T_S(\hat{\mathbf{w}}_{kh}) \right|}_{\Delta a_{k_1}} + \underbrace{\max_{|\mathcal{D}\backslash\mathcal{D}'|=1} \left| \sum_{h=1}^{D_{out}^k} \hat{\mathbf{w}}_{kh}{}^\top \partial T_{Skh} \right|}_{\Delta a_{k_2}} + \underbrace{\max_{|\mathcal{D}\backslash\mathcal{D}'|=1} \left| \sum_{h=1}^{D_{out}^k} \tfrac{1}{2}\hat{\mathbf{w}}_{kh}{}^\top \partial^2 T_{Skh} \hat{\mathbf{w}}_{kh} \right|}_{\Delta a_{k_3}}.
$$

We compute the sensitivity of each of these terms below. The sensitivity of a_{k_1} is given by

$$
\begin{aligned}
\Delta a_{k_1} &= \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} z_{kh,S}^2 - 2 z_{kh,S} f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) + \big(f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})\big)^2 \right| \\
&\leq \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} z_{kh,S}^2 \right| + \left| \sum_{h=1}^{D_{out}^k} 2 z_{kh,S} f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right| + \left| \sum_{h=1}^{D_{out}^k} \big(f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})\big)^2 \right| \\
&\leq \big(T_z^2 + 2T_z \|\boldsymbol{\alpha}_{\hat{\mathbf{w}}_k}\|_2 + \|\boldsymbol{\beta}_{\hat{\mathbf{w}}_k}\|_2^2\big)
\end{aligned}
$$

The sensitivity of a_{k_2} is given by

$$
\begin{aligned}
\Delta a_{k_2} &= \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} \big(-2 z_{kh,S} f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) + 2 f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})\big) \hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right| \\
&\leq \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} \big(2 z_{kh,S} f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})\big) \hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right| + \left| \sum_{h=1}^{D_{out}^k} \big(2 f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})\big) \hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right| \\
&\leq \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} \|2 z_{kh,S} f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \hat{\mathbf{w}}_{kh}\|_2 \cdot \|\mathbf{z}_{k-1,S}\|_2 \right| + \left| \sum_{h=1}^{D_{out}^k} \|2 f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \hat{\mathbf{w}}_{kh}\|_2 \cdot \|\mathbf{z}_{k-1,S}\|_2 \right| \\
&\leq 2T_z \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left(\left| \sum_{h=1}^{D_{out}^k} |z_{kh,S}| \cdot \|f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \hat{\mathbf{w}}_{kh}\|_2 \right| + \left| \sum_{h=1}^{D_{out}^k} \|f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \hat{\mathbf{w}}_{kh}\|_2 \right| \right) \\
&\leq 2T_z \left(T_z \cdot \left(\sum_{h=1}^{D_{out}^k} \beta_{\hat{\mathbf{w}}_{kh}}^2 \|\mathbf{w}_{kh}\|_2^2 \right)^{1/2} + \sum_{h=1}^{D_{out}^k} |\alpha_{\hat{\mathbf{w}}_{kh}} \beta_{\hat{\mathbf{w}}_{kh}}| \cdot \|\hat{\mathbf{w}}_{kh}\|_2 \right),
\end{aligned}
$$

The sensitivity of a_{k_3} is given by

$$\begin{aligned}
\Delta a_{k_3} &= \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} \left(-z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) + \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 + f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right) \left(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right)^2 \right| \\
&\leq \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} \left(z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right) \left(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right)^2 \right| + \left| \sum_{h=1}^{D_{out}^k} \left(\left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 \right) \left(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right)^2 \right| \\
&\qquad + \left| \sum_{h=1}^{D_{out}^k} \left(f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right) \left(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S} \right)^2 \right| \\
&\leq \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \left| \sum_{h=1}^{D_{out}^k} z_{kh,S} \cdot f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 \cdot \|\mathbf{z}_{k-1,S}\|_2^2 \right| + \left| \sum_{h=1}^{D_{out}^k} \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 \cdot \|\mathbf{z}_{k-1,S}\|_2^2 \right| \\
&\qquad + \left| \sum_{h=1}^{D_{out}^k} f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 \cdot \|\mathbf{z}_{k-1,S}\|_2^2 \right| \\
&\leq T_z^2 \max_{\mathbf{z}_{k,S}} \left(\sum_{h=1}^{D_{out}^k} 1/4 \cdot |z_{kh,S}| \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 + \sum_{h=1}^{D_{out}^k} \left(\beta_{\hat{\mathbf{w}}_{kh}} \right)^2 \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 + \sum_{h=1}^{D_{out}^k} 1/4 \alpha_{\hat{\mathbf{w}}_{kh}} \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 \right) \\
&\leq \frac{T_z^2}{4} \left(T_z \left(\sum_{h=1}^{D_{out}^k} \left(\|\mathbf{w}_{kh}\|_2^2 \right)^2 \right)^{1/2} + \sum_{h=1}^{D_{out}^k} \left(4 \left(\beta_{\hat{\mathbf{w}}_{kh}} \right)^2 + \alpha_{\hat{\mathbf{w}}_{kh}} \right) \cdot \|\hat{\mathbf{w}}_{kh}\|_2^2 \right)
\end{aligned}$$

Appendix F: sensitivity of b_k

$$\begin{aligned}
\Delta \mathbf{b}_k = \max_{|\mathcal{D} \backslash \mathcal{D}'|=1} \|\mathbf{b}_k(\mathcal{D}) - \mathbf{b}_k(\mathcal{D}')\|_F &= \max_{|\mathcal{D} \backslash \mathcal{D}'|=1} \left(\sum_{h=1}^{D_{out}^k} \|\mathbf{b}_{kh}(\mathcal{D}) - \mathbf{b}_{kh}(\mathcal{D}')\|_2^2 \right)^{\frac{1}{2}}, \\
&\leq \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S},\mathbf{z}'_{k,S},\mathbf{z}'_{k-1,S}} \left(\sum_{h=1}^{D_{out}^k} \|(\partial T_S(\hat{\mathbf{w}}_{kh}) - \partial^2 T_S(\hat{\mathbf{w}}_{kh}) \hat{\mathbf{w}}_{kh})\|_2^2 \right)^{\frac{1}{2}} \\
&\leq \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S},\mathbf{z}'_{k,S},\mathbf{z}'_{k-1,S}} \left(\sum_{h=1}^{D_{out}^k} \|\partial T_{Skh}\|_2^2 + \sum_{h=1}^{D_{out}^k} \|\partial^2 T_{Skh} \hat{\mathbf{w}}_{kh}\|_2^2 \right)^{\frac{1}{2}} \\
&\leq (\Delta b_{k_1} + \Delta b_{k_2})^{\frac{1}{2}},
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Delta}\mathbf{b}_{k_1} &= \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} \| \left(-2z_{kh,S}f'(\hat{\mathbf{w}}_{kh}^T\mathbf{z}_{k-1,S}) + 2f(\hat{\mathbf{w}}_{kh}^T\mathbf{z}_{k-1,S})f'(\hat{\mathbf{w}}_{kh}^T\mathbf{z}_{k-1,S})\right)\mathbf{z}_{k-1,S}\|_2^2 \\
&\leq 2T_z^2 \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} |(f(\hat{\mathbf{w}}_{kh}^T\mathbf{z}_{k-1,S}) - z_{kh,S})\beta_{\hat{\mathbf{w}}_{kh}}|^2 \\
&\leq 2T_z^2 \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} |\alpha_{\hat{\mathbf{w}}_{kh}}\beta_{\hat{\mathbf{w}}_{kh}} - z_{kh,S}\beta_{\hat{\mathbf{w}}_{kh}}|^2 \\
&\leq 2T_z^2 \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} \left(\alpha_{\hat{\mathbf{w}}_{kh}}^2\beta_{\hat{\mathbf{w}}_{kh}}^2 + 2\alpha_{\hat{\mathbf{w}}_{kh}}\beta_{\hat{\mathbf{w}}_{kh}}^2|z_{kh,S}| + z_{kh,S}^2\beta_{\hat{\mathbf{w}}_{kh}}^2\right) \\
&\leq 2T_z^2\left(||\boldsymbol{\alpha}_{\hat{\mathbf{w}}_k} \odot \boldsymbol{\beta}_{\hat{\mathbf{w}}_k}||_2^2 + 2\min\left(T_z\sum_{h=1}^{D_{out}^k}\alpha_{\hat{\mathbf{w}}_{kh}}, \max_{\mathbf{z}_{k,S}\mathbf{z}_{k-1,S}}\max_i\left(\alpha_{\hat{\mathbf{w}}_{ki}}\beta_{\hat{\mathbf{w}}_{ki}}^2\right)\sum_{h=1}^{D_{out}^k} z_{kh,S}\right) + T_z^2\right), \text{ since z may be negative} \\
&\leq 2T_z^2\left(||\boldsymbol{\alpha}_{\hat{\mathbf{w}}_k} \odot \boldsymbol{\beta}_{\hat{\mathbf{w}}_k}||_2^2 + 2T_z\min\left(\sum_{h=1}^{D_{out}^k}\alpha_{\hat{\mathbf{w}}_{kh}}, \sqrt{D_{out}^k}\max_i\left(\alpha_{\hat{\mathbf{w}}_{ki}}\beta_{\hat{\mathbf{w}}_{ki}}^2\right)\right) + T_z^2\right)
\end{aligned}$$

$$
\begin{aligned}
\Delta \mathbf{b}_{k_2} &= \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} \| \Big(\Big(-2 z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) + 2 \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 \\
&\qquad + 2 f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \Big) \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \Big) \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\leq 2 \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} \| z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \hat{\mathbf{w}}_{kh} \|_2^2 + \| \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\qquad + \| f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\leq 2 \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} \| 1/4 \cdot z_{kh,S} \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\qquad + \| \left(\left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 + 1/4 \cdot f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right) \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\leq 2 \max_{\mathbf{z}_{k,S},\mathbf{z}_{k-1,S}} \sum_{h=1}^{D_{out}^k} | 1/4 \cdot z_{kh,S} |^2 \cdot T_z^2 \cdot T_z^2 \| \hat{\mathbf{w}}_{kh} \|_2^2 \\
&\qquad + \| \left(\left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 + 1/4 \cdot f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right) \hat{\mathbf{w}}_{kh} \|_2^2 \cdot T_z^2 \cdot T_z^2 \\
&\leq \frac{T_z^4}{8} \left(T_z^2 \| \mathbf{w}_k \|_F^2 + \sum_{h=1}^{D_{out}^k} \| \left(4 \beta_{\hat{\mathbf{w}}_{kh}}^2 + \alpha_{\hat{\mathbf{w}}_{kh}} \right) \hat{\mathbf{w}}_{kh} \|_2^2 \right)
\end{aligned}
$$

Appendix G: sensitivity of $\mathbf{C}_k$

The sensitivity of $\Delta \mathbf{C}_k$ is given by

$$
\begin{aligned}
\Delta \mathbf{C}_k &= \max_{|\mathcal{D} \setminus \mathcal{D}'|=1} \| \mathbf{C}_k(\mathcal{D}) - \mathbf{C}_k(\mathcal{D}') \|_F, \\
&= \max_{|\mathcal{D} \setminus \mathcal{D}'|=1} \left(\sum_{h=1}^{D_{out}^k} \| \mathbf{C}_{kh}(\mathcal{D}) - \mathbf{C}_{kh}(\mathcal{D}') \|_F^2 \right)^{\frac{1}{2}}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{C}_{kh}(\mathcal{D}) &= \sum_{s=1}^{S} \tfrac{1}{2} \partial^2 T_{skh}, \\
&= \sum_{s=1}^{S} \left[-2 z_{kh,s} f''(\hat{\mathbf{w}}_{kh}{}^\top \mathbf{z}_{k-1,s}) + 2 \{ f'(\hat{\mathbf{w}}_{kh}{}^\top \mathbf{z}_{k-1,s}) \}^2 + 2 f(\hat{\mathbf{w}}_{kh}{}^\top \mathbf{z}_{k-1,s}) f''(\hat{\mathbf{w}}_{kh}{}^\top \mathbf{z}_{k-1,s}) \right] \mathbf{z}_{k-1,s} \mathbf{z}_{k-1,s}{}^\top .
\end{aligned}
$$

Due to the triangle inequality,

$$\begin{aligned}
\Delta \mathbf{C}_k &\leq \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left(\sum_{h=1}^{D_{out}^k} \|\Delta \mathbf{C}_{kh}\|_F^2 \right)^{1/2} \\
&= \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left(\sum_{h=1}^{D_{out}^k} \|\Big(-2 z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) + 2 \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 + 2 f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \Big) \mathbf{z}_{k-1,S} \mathbf{z}_{k-1,S}^T \|_F^2 \right)^{1/2} \\
&\leq 2 T_z^2 \max_{\mathbf{z}_{k,S}, \mathbf{z}_{k-1,S}} \left(\sum_{h=1}^{D_{out}^k} |z_{kh,S} f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})|^2 + | \left(f'(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) \right)^2 + f(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S}) f''(\hat{\mathbf{w}}_{kh}^T \mathbf{z}_{k-1,S})|^2 \right)^{1/2} \\
&\leq \frac{T_z^2}{2} \left(T_z^2 + \|4 \left(\boldsymbol{\beta}_{\hat{\mathbf{w}}_k} \right)^2 + \boldsymbol{\alpha}_{\hat{\mathbf{w}}_k} \|_2^2 \right)^{1/2}
\end{aligned}$$

Appendix H: sensitivity of coefficients in the output layer objective function

$$\begin{aligned}
\Delta a_o &\leq \frac{1}{2S} \max_{\mathbf{z}_{K,S}} \left| \sum_{h=1}^{D_{out}^o} f(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) - \mathbf{w}_{K+1h}^T f'(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) \mathbf{z}_{K,S} + 1/2 \mathbf{w}_{K+1h}^T f''(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) \mathbf{z}_{K,S} \mathbf{z}_{K,S}^T \mathbf{w}_{K+1h} \right| \\
&\leq \frac{1}{2S} \max_{\mathbf{z}_{K,S}} \sum_{h=1}^{D_{out}^o} \left| f(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) \right| + \sum_{h=1}^{D_{out}^o} \left| \mathbf{w}_{K+1h}^T f'(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) \mathbf{z}_{K,S} \right| + \sum_{h=1}^{D_{out}^o} \left| 1/2 \mathbf{w}_{K+1h}^T f''(\hat{\mathbf{w}}_{K+1h}^T \mathbf{z}_{K,S}) \mathbf{z}_{K,S} \mathbf{z}_{K,S}^T \mathbf{w}_{K+1h} \right| \\
&\leq \frac{1}{2S} \left(\|\boldsymbol{\alpha}_{\mathbf{w}_{K+1}}\|_1 + T_z \sum_{h=1}^{D_{out}^o} \|\mathbf{w}_{K+1h} \cdot \beta_{\mathbf{w}_{K+1h}}\|_2 + \frac{T_z^2}{8} \|\mathbf{w}_{K+1}\|_F^2 \right)
\end{aligned}$$

$$\begin{aligned}
\Delta\mathbf{b}_o &\leq \frac{1}{2S}\max_{\mathbf{y},\mathbf{z}_{K,S}}\left(\sum_{h=1}^{D^o_{out}}\|-y_h\mathbf{z}_{K,S}+f'(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}_{K,S}-f''(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}_{K,S}\mathbf{z}^T_{K,S}\mathbf{w}_{K+1h}\|_2^2\right)^{1/2} \\
&\leq \frac{1}{2S}\max_{\mathbf{y},\mathbf{z}_{K,S}}\left(\sum_{h=1}^{D^o_{out}}|(f'(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})-y_h)-f''(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}^T_{K,S}\mathbf{w}_{K+1h}|^2\|\mathbf{z}_{K,S}\|_2^2\right)^{1/2} \\
&\leq \frac{T_z}{2S}\max_{\mathbf{y},\mathbf{z}_{K,S}}\left(\sum_{h=1}^{D^o_{out}}(f'(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})-y_h)^2+2|(f'(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})-y_h)f''(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}^T_{K,S}\mathbf{w}_{K+1h}|\right. \\
&\qquad\qquad\left.+(f''(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}^T_{K,S}\mathbf{w}_{K+1h})^2\right)^{1/2} \\
&\leq \frac{T_z}{2S}\left(D^o_{out}+2T_z\sum_{h=1}^{D^o_{out}}[\beta_{\mathbf{w}_{K+1}h}\|\mathbf{w}_{K+1h}\|_2]+1/16\cdot T_z^2\cdot\|\mathbf{W}_{K+1}\|_F^2\right)^{1/2}
\end{aligned}$$

$$\begin{aligned}
\Delta\mathbf{C}_o &\leq \frac{1}{2S}\max_{\mathbf{z}_{K,S}}\left(\sum_{h=1}^{D^o_{out}}\|1/2f''(\hat{\mathbf{w}}^T_{K+1h}\mathbf{z}_{K,S})\mathbf{z}_{K,S}\mathbf{z}^T_{K,S}\|_F^2\right)^{1/2} \\
&\leq \frac{1}{2S}\max_{\mathbf{z}_{K,S}}\left(\sum_{h=1}^{D^o_{out}}\|1/8\mathbf{z}_{K,S}\mathbf{z}^T_{K,S}\|_F^2\right)^{1/2} \\
&\leq \frac{1}{16S}\sqrt{D^o_{out}}T_z^2
\end{aligned}$$

Appendix I: Computing a cumulative privacy loss

Preliminary

We first address how the level of perturbation in the coefficients affects the level of privacy in the resulting estimate. Suppose we have an objective function that's quadratic in w, i.e.,

$$E(w) = a + bw + cw^2,$$

where only the coefficients a, b, c contain the information on the data (not anything else in the objective function is relevant to data). We perturb the coefficients to ensure the coefficients are collectively (ϵ, δ)-differentially private.

$$\begin{aligned}
\tilde{a} &= a + n_a, \text{ where } n_a \sim \mathcal{N}(0, \Delta_a^2\sigma^2), \\
\tilde{b} &= b + n_b, \text{ where } n_b \sim \mathcal{N}(0, \Delta_b^2\sigma^2), \\
\tilde{c} &= c + n_c, \text{ where } n_c \sim \mathcal{N}(0, \Delta_c^2\sigma^2),
\end{aligned}$$

where $\Delta_a, \Delta_b, \Delta_c$ are the sensitivities of each term, and σ is a function of ϵ and δ. Here "collectively" means composing the perturbed $\tilde{a}, \tilde{b}, \tilde{c}$ results in (ϵ, δ)-DP. For instance, if one uses the linear composition method (privacy degrades with the number of compositions), and perturbs each of these with ϵ_a, ϵ_b, and ϵ_c, then the total privacy loss should match the sum of these losses, i.e., $\epsilon = \epsilon_a + \epsilon_b + \epsilon_c$. In this case, if one allocates the same privacy budget to perturb each of these coefficients, then $\epsilon_a = \epsilon_b = \epsilon_c = \epsilon/3$. The same holds for δ.

However, if one uses more advanced composition methods and allocates the same privacy budget for each perturbation, per-perturbation budget becomes some function (denoted by g) of total privacy budget ϵ, i.e., $\epsilon_a = \epsilon_b = \epsilon_c = g(\epsilon)$, where $g(\epsilon) \geq \epsilon/3$. So, per-perturbation for a, b, c has a higher privacy budget to spend, resulting in adding less amount of noise.

Whatever composition methods one uses to allocate the privacy budget in each perturbation of those coefficients, since the objective function is a simple quadratic form in w, the resulting estimate of w is some function of those perturbed coefficients, i.e., $\hat{w} = h(\tilde{a}, \tilde{b}, \tilde{c})$. Since the data are summarized in the coefficients and the coefficients are (ϵ, δ)-differentially private, the function of these coefficients is also (ϵ, δ)-differentially private.

One could write the perturbed objective as

$$\begin{aligned}\tilde{E}(w) &= \tilde{a} + \tilde{b}w + \tilde{c}w^2, \\ &= (a + bw + cw^2) + (n_a + n_b w + n_c w^2), \\ &= E(w) + n(w).\end{aligned}$$

Note that we write down the noise term as $n(w)$ to emphasize that when we optimize this objective function, the noise term also contributes to the gradient with respect to w (not just the term $E(w)$).

If we denote some standard normal noise $\alpha \sim \mathcal{N}(0, 1)$, we can rewrite the noise term as

$$n(w) = (\Delta_a + \Delta_b w + \Delta_c w^2)\sigma\alpha,$$

which is equivalent to

$$\begin{aligned}n(w) &\sim \mathcal{N}(0, (\Delta_a + \Delta_b w + \Delta_c w^2)^2\sigma^2), \\ &\sim \mathcal{N}(0, \Delta_{E(w)}^2\sigma^2)\end{aligned}$$

where we denote $\Delta_{E(w)} = \Delta_a + \Delta_b w + \Delta_c w^2$.

Extending the preliminary to DP-MAC

In the framework of DP-MAC, given a mini-batch of data $\mathcal{D}_q$ with a sampling rate q, the DP-mechanism we introduce first computes coefficients for layer-wise objective functions (K layer-wise objective functions for a model with K layers, including the output layer), then noise up the coefficients using Gaussian noise, and outputs the vector of perturbed coefficients for each layer, given by:

$$\mathcal{M}_k(\mathcal{D}_q) = \begin{bmatrix} a_k \\ \mathbf{b}_k \\ \mathbf{C}_k \end{bmatrix} + \begin{bmatrix} n^*_{a,k}(\mathbf{W}_k, \Delta_{a_k}) \\ n^*_{\mathbf{b},k}(\mathbf{W}_k, \Delta_{\mathbf{b}_k}) \\ n^*_{\mathbf{C},k}(\mathbf{W}_k, \Delta_{\mathbf{C}_k}) \end{bmatrix}.$$

We denote the noise terms by $n^*(\mathbf{W}, \Delta)$ and the sensitivities of each coefficient by $\Delta_{a_k}, \cdots, \Delta_{\mathbf{C}_k}$.

Here the question is, if we decide to use an advanced composition method such as moments accountant, how the log-moment of the privacy loss random variable composes in this case. To directly use the composition theorem of

Abadi et al, we need to draw a fresh noise whenever we have a new subsampled data. This means, there should be an instance of Gaussian mechanism that affects the these noise terms simultaneously.

To achieve this, we rewrite the vector of perturbed objective coefficients as $\tilde{\mathbf{E}}(\mathbf{w})$ below. For each layer we gather the loss coefficients into one vector $[a_k, \mathrm{vec}(\mathbf{b}_k), \mathrm{vec}(\mathbf{C}_k)]^T$. Then, we scale down each objective function by its own sensitivity times the number of partitions $\sqrt{MK}$, so that the concatenated vector's sensitivity becomes just 1. Then, add the standard normal noise to the vectors with scaled standard deviation, σ. Then, scale up each perturbed quantities by its own sensitivity times $\sqrt{MK}$. In this example we use all three coefficients, so $M = 3$. Note that in the experiments, using linear expansion we would only use $\mathbf{b}_k$ and so M would equal 1 in that case. In the following we use $P_{m,k}$ to denote the a partition of the vector, which may pick out any of the contained coefficients, e.g. $a_k, \mathbf{b}_k$ or $\mathbf{C}_k$ for $m = 1, m = 2$ and $m = 3$ respectively.

$$
\begin{aligned}
\mathcal{M}(\mathcal{D}_q) &= \begin{bmatrix} \tilde{P}_{1,1}(\mathbf{W}_1) \\ \vdots \\ \tilde{P}_{M,1}(\mathbf{W}_1) \\ \vdots \\ \tilde{P}_{M,K}(\mathbf{W}_K) \end{bmatrix} \\
&= \begin{bmatrix} P_{1,1}(\mathbf{W}_1) \\ \vdots \\ P_{M,K}(\mathbf{W}_K) \end{bmatrix} + \begin{bmatrix} n_1^*(\mathbf{W}_1) \\ \vdots \\ n_{M,K}^*(\mathbf{W}_K) \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{MK}\Delta_{P_{1,1}(\mathbf{W}_1)} \cdot \left\{ \frac{P_{1,1}(\mathbf{W}_1)}{\sqrt{MK}\Delta_{P_{1,1}(\mathbf{W}_1)}} + \sigma\mathcal{N}(0,1) \right\} \\ \cdots \\ \sqrt{MK}\Delta_{P_{M,K}(\mathbf{W}_K)} \cdot \left\{ \frac{P_{M,K}(\mathbf{W}_K)}{\sqrt{MK}\Delta_{P_{M,K}(\mathbf{W}_K)}} + \sigma\mathcal{N}(0,1) \right\} \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{MK}\Delta_{P_{1,1}(\mathbf{W}_1)} \\ \cdots \\ \sqrt{MK}\Delta_{P_{M,K}(\mathbf{W}_K)} \end{bmatrix} \cdot \left(\begin{bmatrix} \frac{P_{1,1}(\mathbf{W}_1)}{\sqrt{K}\Delta_{P_{1,1}(\mathbf{W}_1)}} \\ \cdots \\ \frac{P_{M,K}(\mathbf{W}_K)}{\sqrt{K}\Delta_{P_{M,K}(\mathbf{W}_K)}} \end{bmatrix} + \mathcal{N}(0, \sigma^2 I) \right)
\end{aligned}
$$

Since we're adding independent Gaussian noise under each subsampled data, the privacy loss after T steps, is simply following the composibility theorem in the Abadi et al paper.

So compared to the sensitivity for $n(w)$ in the first section, the new noise $n^*(w)$ has a higher sensitivity due to the factor $\sqrt{MK}$.

Moments Calculations

In this case, with a subsampling with rate q, we re-do the calculations in Abadi et al. First, let:

$$\mu_0 = \mathcal{N}(\mathbf{0}_K, \sigma^2 I), \mu_1 = \mathcal{N}(\mathbf{1}_K, \sigma^2 I)$$

and let μ as a mixture of the two Gaussians,

$$\mu = (1-q)\mathcal{N}(\mathbf{0}_K, \sigma^2 I) + q\mathcal{N}(\mathbf{1}_K, \sigma^2 I).$$

Here $\mathbf{0}_K$ is the K-dimensional 0 vector, and $\mathbf{1}_K$ is the K-dimensional all ones vector. Here $\alpha_M(\lambda)$ should be $\log \max(E_1, E_2)$ where

$$E_1 = \mathbb{E}_{z\sim\mu}[(\mu(z)/\mu_0(z))^\lambda], E_2 = \mathbb{E}_{z\sim\mu_0}[(\mu_0(z)/\mu(z))^\lambda]$$

Then, we can compose further mechanisms using this particular $\alpha_M(\lambda)$, which follows the same analysis as in Abadi et al.