Supplementary Material: Photonic Differential Privacy with Direct Feedback Alignment

A Complete proof of the Differential Privacy parameters

A.1 Extended proof of Proposition 3

As a reminder, we would like to compute the Rényi divergence of the following Gaussian mechanism, where all the quantities are clipped as in Eq. 8:

$$\frac{1}{m}\sum_{i=1}^{m}((\mathbf{B}\boldsymbol{e}_i)\odot\phi'(\boldsymbol{z}_i))h_{ik}+\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{g}_i\odot\phi'(\boldsymbol{z}_i))h_{ik}=\boldsymbol{f}_k(D)+\mathcal{N}(0,\boldsymbol{\Sigma}_k) \tag{1}$$

where $\boldsymbol{\Sigma}_k=\frac{\sigma^2}{m^2}\mathbf{diag}(\boldsymbol{a}_k)^2$ and $(\boldsymbol{a}_k)_j=\sqrt{\sum_{i=1}^{m}(\phi'_{ij}h_{ik})^2},\forall j=1,\dots,n_{\ell-1}$. As explained in the main text, we will focus on column k and will drop the k indices. The proposition we want to prove is the following:

Proposition 3 (Photonic Differential Privacy parameters)**.** *Given two probability distributions $P\sim\mathcal{N}(\boldsymbol{f}(D),\boldsymbol{\Sigma})$ and $Q\sim\mathcal{N}(\boldsymbol{f}(D'),\boldsymbol{\Sigma}')$ corresponding to the Gaussian mechanisms depicted in* (1) *on neighboring datasets D and D', the Rényi divergence of order α between these mechanisms is:*

$$\begin{aligned}\mathbb{D}_\alpha(P\|Q)&\leq\frac{2.\alpha}{m.\sigma^2}\frac{(\gamma^{max}\tau^{max}\tau_B)^2}{(\gamma_\ell^{min}\tau_h^{min})^2}+\frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{m(\gamma_\ell^{min}\tau_h^{min})^2}{(m+1)(\gamma_\ell^{min}\tau_h^{min})^2-(\gamma_\ell^{max}\tau_h^{max})^2}\right]\\&=\varepsilon_{PDFA}\end{aligned} \tag{2}$$

Our mechanism is therefore $(\alpha,T\varepsilon_{PDFA})$-RDP with T the number of training epochs. We can deduce that the mechanism on the weight matrix with $n_{\ell-1}$ columns is $(\alpha,Tn_{\ell-1}\varepsilon_{PDFA})$-RDP. Then the mechanism of the whole network composed of L layers is $(\alpha,LTn_{\ell-1}\varepsilon_{PDFA})$-RDP. We can then convert our bound to DP parameters using Theorem 2 to obtain a $(LTn_{\ell-1}\varepsilon_{PDFA}+\frac{\log 1/\delta}{\alpha-1},\delta)$-DP mechanism for all $\delta\in(0,1)$.

Proof. In the following, the variables with a prime correspond to the ones built upon dataset D'. According to Eq. 13, the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are diagonal and any of their weighted sum is diagonal, as well as their inverse. Moreover, the determinant of a diagonal matrix is the product of its diagonal elements. Using this in Eq. 7 yields:

$$\mathbb{D}_\alpha(P\|Q)=\sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{\alpha\,a_j'^2+(1-\alpha)a_j^2}-\frac{1}{2(\alpha-1)}\log\left[\frac{(1-\alpha)a_j^2+\alpha a_j'^2}{a_j^{2(1-\alpha)}a_j'^{2\alpha}}\right]\right)$$

Using the fact that we are studying neighboring datasets, the sums composing a_j and a'_j differ by only one element at element $i=I$. This implies that

$$\begin{aligned}\alpha\,a_j'^2+(1-\alpha)a_j^2&=\alpha.\sum_{i=1}^{m}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2+(1-\alpha).\sum_{i=1}^{m}(\phi'_{ij}h_{ik})^2\\&=\sum_{i=1}^{m}(\phi'_{ij}h_{ik})^2+\alpha.(\sum_{i=1}^{m}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2-\sum_{i=1}^{m}(\phi'_{ij}h_{ik})^2)\\&=a_j^2+\alpha[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2)]\end{aligned}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

where $\tilde{\phi}'_{Ij}$ and $\tilde{h}^2_{Ik}$ are taken on dataset D'. Inserting this in the Rényi divergence yields:

$$\mathbb{D}_\alpha(P\|Q) = \sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{a_j^2+\alpha[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2)]} - \frac{1}{2(\alpha-1)}\log\left[\frac{a_j^2+\alpha[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2)]}{a_j^{2(1-\alpha)}a_j'^{2\alpha}}\right]\right)$$

By choosing D and D' such that $[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2)] \geq 0$, the Rényi divergence is upper bounded as follow:

$$\mathbb{D}_\alpha(P\|Q) \leq \sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{a_j^2} - \frac{1}{2(\alpha-1)}\log\left[\frac{a_j^{2\alpha}}{a_j'^{2\alpha}}\right]\right)$$

Noting that $a_j^2 = \sum_{i=1}^m(\phi'_{ij}h_{ik})^2+(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2 = a_j'^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]$ yields:

$$\begin{aligned}
\mathbb{D}_\alpha(P\|Q) &\leq \sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{a_j^2} - \frac{\alpha}{2(\alpha-1)}\log\left[\frac{a_j^2}{a_j'^2}\right]\right)\\
&\leq \sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{a_j^2} - \frac{\alpha}{2(\alpha-1)}\log\left[\frac{a_j'^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]}{a_j'^2}\right]\right)\\
&\leq \sum_{j=1}^{n_\ell}\left(\frac{\alpha m^2}{2\sigma^2}\frac{(f_j(D)-f_j(D'))^2}{a_j^2} + \frac{\alpha}{2(\alpha-1)}\log\left[\frac{a_j'^2}{a_j'^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]}\right]\right)\\
&\leq \frac{\alpha m^2}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{m(\gamma_\ell^{min}\tau_h^{min})^2} + \sum_{j=1}^{n_\ell}\frac{\alpha}{2(\alpha-1)}\log\left[\frac{a_j'^2}{a_j'^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]}\right]\\
&\leq \frac{2.\alpha}{m.\sigma^2}\frac{(\gamma^{max}\tau^{max}\tau_B)^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{m(\gamma_\ell^{min}\tau_h^{min})^2}{(m+1)(\gamma_\ell^{min}\tau_h^{min})^2-(\gamma_\ell^{max}\tau_h^{max})^2}\right]\\
&= \varepsilon_{\text{PDFA}}
\end{aligned}$$

where we used the upper bounds on the sensitivity $\Delta_{\boldsymbol{f}}^2$ and $a_j'^2$. This is the result of Proposition 3. $\square$

Note that an alternative expression is:

$$\begin{aligned}
\mathbb{D}_\alpha(P\|Q) &\leq \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \sum_{j=1}^{n_\ell}\frac{\alpha}{2(\alpha-1)}\log\left[\frac{a_j'^2}{a_j'^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]}\right]\\
&= \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \sum_{j=1}^{n_\ell}\frac{\alpha}{2(\alpha-1)}\log\left[\frac{\sum_{i=1}^m(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2}{\sum_{i=1}^m(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2-[(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2-(\phi'_{Ij}h_{Ik})^2]}\right]\\
&= \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \sum_{j=1}^{n_\ell}\frac{\alpha}{2(\alpha-1)}\log\left[\frac{\sum_{i\neq I}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2+(\tilde{\phi}'_{Ij}\tilde{h}_{Ik})^2}{\sum_{i\neq I}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2+(\phi'_{Ij}h_{Ik})^2}\right]\\
&\leq \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{\sum_{i\neq I}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2+(\gamma_\ell^{max}\tau_h^{max})^2}{\sum_{i\neq I}(\tilde{\phi}'_{ij}\tilde{h}_{ik})^2+(\gamma_\ell^{min}\tau_h^{min})^2}\right]\\
&\leq \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{(m-1).(\gamma_\ell^{min}\tau_h^{min})^2+(\gamma_\ell^{max}\tau_h^{max})^2}{(m-1).(\gamma_\ell^{min}\tau_h^{min})^2+(\gamma_\ell^{min}\tau_h^{min})^2}\right]\\
&\leq \frac{\alpha m}{2\sigma^2}\frac{n_\ell.\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{(m-1).(\gamma_\ell^{min}\tau_h^{min})^2+(\gamma_\ell^{max}\tau_h^{max})^2}{m.(\gamma_\ell^{min}\tau_h^{min})^2}\right]\\
&\leq \frac{2.\alpha}{m.\sigma^2}\frac{(\gamma^{max}\tau^{max}\tau_B)^2}{(\gamma_\ell^{min}\tau_h^{min})^2} + \frac{n_\ell.\alpha}{2(\alpha-1)}\log\left[\frac{m-1}{m}+\frac{(\gamma_\ell^{max}\tau_h^{max})^2}{m.(\gamma_\ell^{min}\tau_h^{min})^2}\right]
\end{aligned}$$

A.2 Equal covariance matrices

First, we can notice that when the covariance matrices are equal, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma}' = \frac{\sigma^2}{m^2}\mathbf{diag}(\boldsymbol{a}_k)^2$, the log-term in Eq. 7 is equal to 0. Then, we have:

$$\begin{aligned}
\mathbb{D}_\alpha(P\|Q) &= \sum_{j=1}^{n_\ell} \frac{\alpha m^2}{2\sigma^2} \frac{(f_j(D) - f_j(D'))^2}{a_j^2} \\
&\leq \frac{n_\ell.\alpha.m}{2\sigma^2} \sum_{j=1}^{n_\ell} \frac{(f_j(D) - f_j(D'))^2}{(\gamma_\ell^{min}\tau_h^{min})^2} \\
&\leq \frac{n_\ell.\alpha.m}{2\sigma^2} \frac{\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell^{min}\tau_h^{min})^2} \\
&\leq \frac{2\alpha}{m\sigma^2} \frac{(\tau_B\gamma_\ell^{max}\tau_h^{max})^2}{(\gamma_\ell^{min}\tau_h^{min})^2} \\
&\doteq \varepsilon_2
\end{aligned}$$

A.3 Equal saturating covariance matrices

In this subsection, we will suppose that the covariance matrices are equal and saturating, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma}' = \frac{\sigma^2}{n_\ell.m}(\gamma_\ell\tau_h)^2\mathbf{I}$ with $\gamma_\ell = \{\gamma_\ell^{min}, \gamma_\ell^{max}\}$ and $\tau_h = \{\tau_h^{min}, \tau_h^{max}\}$. Then we can start by noticing that $a_j^2 = a_j'^2 = \frac{m}{n_\ell}\tau_h^2\gamma_\ell^2$. In that case, the sensitivity of the function can be written as:

$$\begin{aligned}
\Delta_{\boldsymbol{f}}^\ell &= \sup_{D\sim D'} \|\boldsymbol{f}(D) - \boldsymbol{f}(D')\|_2 \leq \frac{2}{m} \left\|(\mathbf{B}^\ell \boldsymbol{e}_i) \odot \phi_\ell'(\boldsymbol{z}_i^\ell))h_{ik}^{\ell-1}\right\|_2 \\
&\leq \frac{2}{m}\tau_B\gamma_\ell\frac{\tau_h}{\sqrt{n_\ell}}
\end{aligned}$$

This implies that:

$$\begin{aligned}
\mathbb{D}_\alpha(P\|Q) &= \sum_{j=1}^{n_\ell} \frac{\alpha m^2}{2\sigma^2} \frac{(f_j(D) - f_j(D'))^2}{a_j^2} \\
&\leq \frac{n_\ell.\alpha.m}{2\sigma^2} \sum_{j=1}^{n_\ell} \frac{(f_j(D) - f_j(D'))^2}{(\gamma_\ell\tau_h)^2} \\
&\leq \frac{n_\ell.\alpha.m}{2\sigma^2} \frac{\Delta_{\boldsymbol{f}}^2}{(\gamma_\ell\tau_h)^2} \\
&\leq \frac{2\alpha}{m\sigma^2} \frac{(\tau_B\gamma_\ell\tau_h)^2}{(\gamma_\ell\tau_h)^2} = \frac{2\alpha}{m\sigma^2}\tau_B^2 \\
&\doteq \varepsilon_3
\end{aligned}$$

B Additional numerical results on MNIST and CIFAR-10

MNIST – We provide below results on MNIST, obtained with the same code, procedure, and hyper-parameters as for the FashionMNIST experiments. These results are in line with Table 1 of our paper, with photonics results always close to ternarized ones. We notice that this "default" choice of hyper-parameters on MNIST results in ternarized DFA outperforming vanilla DFA. (We only lightly tune hyperparameters on BP, to demonstrate that our approach does not require any specific expensive fine-tuning search.)

σ	**non-private**	**0.01**	**0.05**	**0.1**
τ_f			**1**	
BP	97,94	62,63	58,42	48,33
DFA	96,36	92,99	92,68	92,45
TDFA	97,09	93,67	93,57	93,28
PDFA	96,95	93,60	93,57	93,12

Table 1: **Test accuracy on MNIST with our DP mechanism.** We find our approach to be robust to increasing DP noise σ. In particular, photonic DFA results (PDFA) are always within 1% of the corresponding DFA run.

CIFAR-10 – We chose to use a pre-trained network on ImageNet and extract its trained convolutional layers. Since these convolutions can be seen as feature extractors of the images and are not re-trained, they do not need to be taken into account into the Differential Privacy mechanism. We fine-tune only the fully-connected layers of the classifier using our Photonic DFA+DP mechanism.

We choose this expreiment to demonstrate the scalability of our scheme. We do not seek to achieve state-of-the-art performance or to exhaustively explore the dynamics/impact of different differentially private configuration (as we did with MNIST), but simply to show our scheme can scale to such harder tasks.

We used a VGG16 network pre-trained on ImageNet. We leave the convolutions untouched, and fine-tune the classifier layers (25088 –> 4096 –> 4096 –> 10) with differentially private photonic training. We do not use any data augmentation, and simply resize the CIFAR-10 images to 224x224. We fine-tune for 15 epochs, using SGD with learning rate 5.10^{-3}, momentum 0.9, and batch size 256. Hyperparameters are kept identical across all methods and hardware. We obtain results both in a vanilla (no DP) setting as a comparison baseline, and in a differentially private setting yielding the following accuracies:
Vanilla (no differential privacy): 83.17% (BP), 81.34% (DFA), 83.36% (TDFA).
DP ($\sigma = 0.05$, $\tau_f = 1$): 60.45% (BP), 79.68% (DFA), 79.33% (TDFA), 78.64% (PDFA).

We note that this result shows good scalability, with performance in line with our MNIST/FashionMNIST results. Over all the experiments we have performed, the DFA algorithms seem much more resilient to adding noise and clipping (i.e. the DP alogirithmical modification) than Backpropagation, which could open new research directions.