Supplementary Material: Bayesian Differential Privacy for Machine Learning

Aleksei Triastcyn [1] **Boi Faltings** [1]

A. Proofs

A.1. Proof of Propositions

Proposition 1. *$(\varepsilon_\mu, \delta_\mu)$-strong Bayesian differential privacy implies $(\varepsilon_\mu, \delta_\mu)$-Bayesian differential privacy.*

Proof. Let us define a set of outcomes for which the privacy loss variable exceeds the ε threshold: $F(x') = \{w : L_{\mathcal{A}}(w, D, D') > \varepsilon\}$, and its compliment $F^c(x')$.

We have,

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] = \int \Pr[\mathcal{A}(D) \in \mathcal{S}, x'] \, dx' \tag{1}$$

$$= \int \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}^c(x'), x'] \tag{2}$$

$$+ \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}(x'), x'] \, dx' \tag{3}$$

$$= \int \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}^c(x') | x'] \mu(x') \tag{4}$$

$$+ \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}(x'), x'] \, dx' \tag{5}$$

$$\leq \int e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S} \cap \mathcal{F}^c(x') | x'] \mu(x') \tag{6}$$

$$+ \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}(x'), x'] \, dx' \tag{7}$$

$$\leq \int e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S}, x'] \tag{8}$$

$$+ \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}(x'), x'] \, dx' \tag{9}$$

$$\leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S}] + \delta_\mu, \tag{10}$$

where we used the observation that $L \leq \varepsilon$ implies $\Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}^c(x')] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S} \cap \mathcal{F}^c(x')]$, and therefore, $\Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}^c(x') \mid x'] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S} \cap \mathcal{F}^c(x') \mid x']$, because $\mathcal{A}(D)$ does not depend on x', and $\mathcal{A}(D')$ is already conditioned on x' through D'. Additionally, in the first line we used marginalisation, and the last inequality is due to the fact that

$$\int \Pr[\mathcal{A}(D) \in \mathcal{S} \cap \mathcal{F}(x'), x'] \, dx' \tag{11}$$

$$\leq \int \Pr[\mathcal{A}(D) \in \mathcal{F}(x'), x'] \, dx' \tag{12}$$

$$= \int \mu(x') \Pr[\mathcal{A}(D) \in \mathcal{F}(x') \mid x'] \, dx' \tag{13}$$

$$= \int \mu(x') \int_{w \in \mathcal{F}(x')} p_{\mathcal{A}}(w|D, x') \, dw \, dx' \tag{14}$$

$$= \mathbb{E}_{x'} \left[\mathbb{E}_w \left[\mathbb{1}\{L > \varepsilon\} \right] \right] \tag{15}$$

$$\leq \delta_\mu \tag{16}$$

□

Proposition 2 (Post-processing)**.** *Let $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ be a $(\varepsilon_\mu, \delta_\mu)$-Bayesian differentially private algorithm. Then for any arbitrary randomised data-independent mapping $f : \mathcal{R} \to \mathcal{R}'$, $f(\mathcal{A}(D))$ is $(\varepsilon_\mu, \delta_\mu)$-Bayesian differentially private.*

Proof. First, by Proposition 1, $(\varepsilon_\mu, \delta_\mu)$-strong BDP implies the weak sense of BDP:

$$\Pr\left[\mathcal{A}(D) \in \mathcal{S}\right] \leq e^{\varepsilon_\mu} \Pr\left[\mathcal{A}(D') \in \mathcal{S}\right] + \delta_\mu, \tag{17}$$

for any set of outcomes $\mathcal{S} \subset \mathcal{R}$.

For a data-independent function $f(\cdot)$:

$$\Pr\left[f(\mathcal{A}(D)) \in \mathcal{T}\right] = \Pr\left[\mathcal{A}(D) \in \mathcal{S}\right] \tag{18}$$

$$\leq e^{\varepsilon_\mu} \Pr\left[\mathcal{A}(D') \in \mathcal{S}\right] + \delta_\mu, \tag{19}$$

$$= e^{\varepsilon_\mu} \Pr\left[f(\mathcal{A}(D')) \in \mathcal{T}\right] + \delta_\mu \tag{20}$$

where $\mathcal{S} = f^{-1}[\mathcal{T}]$, i.e. $\mathcal{S}$ is the preimage of $\mathcal{T}$ under f. □

Proposition 3 (Basic composition)**.** *Let $\mathcal{A}_i : \mathcal{D} \to \mathcal{R}_i$, $\forall i = 1..k$, be a sequence of $(\varepsilon_\mu, \delta_\mu)$-Bayesian differentially private algorithms. Then their combination, defined as $\mathcal{A}_{1:k} : \mathcal{D} \to \mathcal{R}_1 \times \ldots \times \mathcal{R}_k$, is $(k\varepsilon_\mu, k\delta_\mu)$-Bayesian differentially private.*

[1]Artificial Intelligence Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Correspondence to: Aleksei Triastcyn <aleksei.triastcyn@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Proof. Let us denote $L = \log \frac{p(w_1,\ldots,w_k|D)}{p(w_1,\ldots,w_k|D')}$.

Also, let $L_i = \log \frac{p(w_i|D,w_{i-1},\ldots,w_1)}{p(w_i|D',w_{i-1},\ldots,w_1)}$. Then,

$$\Pr\left[L \geq k\varepsilon_\mu\right] = \Pr\left[\sum_{i=1}^{k} L_i \geq k\varepsilon_\mu\right] \tag{21}$$

$$\leq \sum_{i=1}^{k} \Pr[L_i \geq \varepsilon_\mu] \tag{22}$$

$$\leq \sum_{i=1}^{k} \delta_\mu \tag{23}$$

$$\leq k\delta_\mu \tag{24}$$

For the weak sense of BDP, the proof follows the steps of Dwork et al. (2014, Appendix B).

□

Proposition 4 (Group privacy)**.** *Let $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ be a $(\varepsilon_\mu, \delta_\mu)$-Bayesian differentially private algorithm. Then for all pairs of datasets $D, D' \in \mathcal{D}$, differing in k data points $x_1, \ldots, x_k$ s.t. $x_i \sim \mu(x)$ for $i = 1..k$, $\mathcal{A}(D)$ is $(k\varepsilon_\mu, ke^{k\varepsilon_\mu}\delta_\mu)$-Bayesian differentially private.*

Proof. Let us define a sequence of datasets D^i, $i = 1..k$, s.t. $D = D^0$, $D' = D^k$, and D^i and D^{i-1} differ in a single example. Then,

$$\frac{p(w|D)}{p(w|D')} = \frac{p(w|D^0)p(w|D^1)\ldots p(w|D^{k-1})}{p(w|D^1)p(w|D^2)\ldots p(w|D^k)} \tag{25}$$

Denote $L_i = \log \frac{p(w|D^{i-1})}{p(w|D^i)}$ for $i = 1..k$.

Finally, applying the definition of $(\varepsilon_\mu, \delta_\mu)$-Bayesian differential privacy,

$$\Pr\left[L \geq k\varepsilon_\mu\right] = \Pr\left[\sum_{i=1}^{k} L_i \geq k\varepsilon_\mu\right] \tag{26}$$

$$\leq \sum_{i=1}^{k} \Pr[L_i \geq \varepsilon_\mu] \tag{27}$$

$$\leq k\delta_\mu \tag{28}$$

For the weak sense of BDP,

$$Pr\left[\mathcal{A}(D) \in \mathcal{S}\right] \leq e^{\varepsilon_\mu} Pr\left[\mathcal{A}(D^1) \in \mathcal{S}\right] + \delta_\mu \tag{29}$$

$$\leq e^{\varepsilon_\mu}\left(e^{\varepsilon_\mu} Pr\left[\mathcal{A}(D^2) \in \mathcal{S}\right] + \delta_\mu\right) + \delta_\mu \tag{30}$$

$$\leq e^{2\varepsilon_\mu} Pr\left[\mathcal{A}(D^2) \in \mathcal{S}\right] + e^{\varepsilon_\mu}\delta_\mu + \delta_\mu \tag{31}$$

$$\leq \ldots \tag{32}$$

$$\leq e^{k\varepsilon_\mu} \Pr\left[\mathcal{A}(D^k) \in \mathcal{S}\right] + \frac{e^{k\varepsilon_\mu} - 1}{e^{\varepsilon_\mu} - 1}\delta_\mu \tag{33}$$

$$\leq e^{k\varepsilon_\mu} \Pr\left[\mathcal{A}(D^k) \in \mathcal{S}\right] + \frac{k\varepsilon_\mu e^{k\varepsilon_\mu}}{\varepsilon_\mu}\delta_\mu \tag{34}$$

$$\leq e^{k\varepsilon_\mu} \Pr\left[\mathcal{A}(D') \in \mathcal{S}\right] + ke^{k\varepsilon_\mu}\delta_\mu, \tag{35}$$

where in (33) we use the formula for the sum of a geometric progression; in (34), the facts that $e^x - 1 \leq xe^x$, for $x > 0$, and $e^x \geq x + 1$. □

A.2. Proof of Theorem 1

Let us restate the theorem:

Theorem 1 (Advanced Composition)**.** *Let a learning algorithm run for T iterations. Denote by $w^{(1)} \ldots w^{(T)}$ a sequence of private learning outcomes at iterations $1, \ldots, T$, and $L^{(1:T)}$ the corresponding total privacy loss. Then,*

$$\mathbb{E}\left[e^{\lambda L^{(1:T)}}\right] \leq \prod_{t=1}^{T} \mathbb{E}_x\left[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t\|q_t)}\right]^{\frac{1}{T}},$$

where $p_t = p(w^{(t)}|w^{(t-1)}, D)$, $q_t = p(w^{(t)}|w^{(t-1)}, D')$.

Proof. The proof closely follows (Abadi et al., 2016).

First, we can write

$$L^{(1:T)} = \log \frac{p(w^{(1)} \ldots w^{(T)} \mid D)}{p(w^{(1)} \ldots w^{(T)} \mid D')} \tag{36}$$

$$= \log \prod_{t=1}^{T} \frac{p(w^{(t)} \mid w^{(t-1)} \ldots p(w^{(1)}, D)}{p(w^{(t)} \mid w^{(t-1)} \ldots p(w^{(1)}, D')} \tag{37}$$

$$= \log \prod_{t=1}^{T} \frac{p(w^{(t)} \mid w^{(t-1)}, D)}{p(w^{(t)} \mid w^{(t-1)}, D')} \tag{38}$$

$$= \sum_{t=1}^{T} L^{(t)} \tag{39}$$

Unlike the composition proof of the moments accountant by Abadi et al. (2016), we cannot simply swap the product and the expectation in our proof, because the additional example

x' remains the same in all applications of the privacy mechanism and probability distributions will not be independent. However, we can use generalised Hölder's inequality:

$$\left\| \prod_{t=1}^{T} f_t \right\|_r \leq \prod_{t=1}^{T} \|f_t\|_{p_t}, \tag{40}$$

where p_t are such that $\sum_{t=1}^{T} \frac{1}{p_t} = \frac{1}{r}$, and $\|f\|_r = \left(\int_S |f|^r dx\right)^{1/r}$.

Choosing $r = 1$ and $p_t = T$:

$$\mathbb{E}\left[e^{\lambda L^{1:T}}\right] = \mathbb{E}\left[\prod_{t=1}^{T} e^{\lambda \log \frac{p(w^{(t)} \mid w^{(t-1)}, X)}{p(w^{(t)} \mid w^{(t-1)}, X')}}\right] \tag{41}$$

$$= \mathbb{E}_x\left[\mathbb{E}_w\left[\prod_{t=1}^{T} e^{\lambda \log \frac{p(w^{(t)} \mid w^{(t-1)}, X)}{p(w^{(t)} \mid w^{(t-1)}, X')}} \,\middle|\, x'\right]\right] \tag{42}$$

$$= \mathbb{E}_x\left[\prod_{t=1}^{T} \mathbb{E}_w\left[e^{\lambda \log \frac{p(w^{(t)} \mid w^{(t-1)}, X)}{p(w^{(t)} \mid w^{(t-1)}, X')}} \,\middle|\, x'\right]\right] \tag{43}$$

$$= \mathbb{E}_x\left[\prod_{t=1}^{T} e^{\lambda \mathcal{D}_{\lambda+1}(p_t \| q_t)}\right] \tag{44}$$

$$\leq \prod_{t=1}^{T} \mathbb{E}_x\left[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t \| q_t)}\right]^{\frac{1}{T}}, \tag{45}$$

where (42) is by the law of total expectation; (43) is due to independence of noise between iterations, similarly to (Abadi et al., 2016); and (45) is by Hölder's inequality. □

A.3. Proof of Theorem 3

Let us restate the theorem:

Theorem 3. *Given the Gaussian noise mechanism with the noise parameter σ and subsampling probability q, the privacy cost for $\lambda \in \mathbb{N}$ at iteration t can be expressed as*

$$c_t(\lambda) = \max\{c_t^L(\lambda), c_t^R(\lambda)\},$$

where

$$c_t^L(\lambda) = \log \mathbb{E}_x\left[\mathbb{E}_{k\sim B(\lambda+1,q)}\left[e^{\frac{k^2-k}{2\sigma^2}\|g_t - g_t'\|^2}\right]\right],$$

$$c_t^R(\lambda) = \log \mathbb{E}_x\left[\mathbb{E}_{k\sim B(\lambda,q)}\left[e^{\frac{k^2+k}{2\sigma^2}\|g_t - g_t'\|^2}\right]\right],$$

and $B(\lambda, q)$ is the binomial distribution with λ experiments and the probability of success q.

Proof. Without loss of generality, assume $D' = D \cup \{x'\}$. For brevity, let $d_t = \|g_t - g_t'\|$.

Let us first consider $\mathcal{D}_{\lambda+1}(p(w|D')\|p(w|D))$:

$$\mathbb{E}\left[\left(\frac{p(w|D')}{p(w|D)}\right)^{\lambda+1}\right]$$

$$= \mathbb{E}\left[\left(\frac{(1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(d_t,\sigma^2)}{\mathcal{N}(0,\sigma^2)}\right)^{\lambda+1}\right] \tag{46}$$

$$= \mathbb{E}\left[\left((1-q) + q\frac{\mathcal{N}(d_t,\sigma^2)}{\mathcal{N}(0,\sigma^2)}\right)^{\lambda+1}\right] \tag{47}$$

$$= \mathbb{E}\left[\left((1-q) + qe^{\frac{(w-d_t)^2 - w^2}{2\sigma^2}}\right)^{\lambda+1}\right] \tag{48}$$

$$= \mathbb{E}\left[\left((1-q) + qe^{\frac{2dw - d_t^2}{2\sigma^2}}\right)^{\lambda+1}\right] \tag{49}$$

$$= \mathbb{E}\left[\sum_{k=0}^{\lambda+1}\binom{\lambda+1}{k} q^k (1-q)^{\lambda+1-k} e^{\frac{2d_t kw - kd_t^2}{2\sigma^2}}\right] \tag{50}$$

$$= \sum_{k=0}^{\lambda+1}\binom{\lambda+1}{k} q^k (1-q)^{\lambda+1-k} \mathbb{E}\left[e^{\frac{2d_t kw - kd_t^2}{2\sigma^2}}\right] \tag{51}$$

$$= \sum_{k=0}^{\lambda+1}\binom{\lambda+1}{k} q^k (1-q)^{\lambda+1-k} e^{\frac{k^2-k}{2\sigma^2}d_t^2} \tag{52}$$

$$= \mathbb{E}_{k\sim B(\lambda+1,q)}\left[e^{\frac{k^2-k}{2\sigma^2}\|g_t - g_t'\|^2}\right], \tag{53}$$

Here, in (50) we used the binomial expansion, in (51) the fact that the factors in front of the exponent do not depend on w, and in (52) the property $\mathbb{E}_w\left[\exp(2aw/(2\sigma^2))\right] = \exp(a^2/(2\sigma^2))$ for $w \sim \mathcal{N}(0, \sigma^2)$. Plugging the above in the privacy cost formula (Eq. 10 in the main paper), we get the expression for $c_t^L(\lambda)$.

Computing $\mathcal{D}_{\lambda+1}(p(w|D)\|p(w|D'))$ is a little more challenging. Let us first change to $\mathcal{D}_{\lambda}(p(w|D)\|p(w|D'))$, so that the expectation is taken over $\mathcal{N}(0, \sigma^2)$. Then, we can bound it observing that $f(x) = \frac{1}{x}$ is convex for $x > 0$ and using the definition of convexity, and apply the same steps as above:

$$\mathbb{E}\left[\left(\frac{p(w|D)}{p(w|D')}\right)^{\lambda}\right]$$

$$= \mathbb{E}\left[\left(\frac{\mathcal{N}(0,\sigma^2)}{(1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(d_t,\sigma^2)}\right)^{\lambda}\right] \tag{54}$$

$$\leq \mathbb{E}\left[\left((1-q) + qe^{\frac{d_t^2 - 2dw}{2\sigma^2}}\right)^{\lambda}\right] \tag{55}$$

$$= \mathbb{E}_{k\sim B(\lambda,q)}\left[e^{\frac{k^2+k}{2\sigma^2}\|g_t - g_t'\|^2}\right] \tag{56}$$

In practice, we haven't found any instance of $\mathcal{D}_{\lambda+1}(p(w|D')\|p(w|D)) < \mathcal{D}_{\lambda+1}(p(w|D)\|p(w|D'))$ when the latter was computed using numerical integration,

although it may happen when using this theoretical upper bound. □

A.4. Proof of Theorem 4

Let us restate the theorem:

Theorem 4. *Estimator $\hat{c}_t(\lambda)$ overestimates $c_t(\lambda)$ with probability $1-\gamma$. That is,*

$$\Pr\left[\hat{c}_t(\lambda) < c_t(\lambda)\right] \leq \gamma.$$

Proof. First of all, we can drop the logarithm from our consideration because of its monotonicity.

Now, assuming that samples $e^{\lambda \hat{D}^{(t)}_{\lambda+1}}$ have a common mean and a common variance, and applying the maximum entropy principle in combination with an uninformative (flat) prior, one can show that the quantity $\frac{M(t)-\mathbb{E}\left[e^{\lambda \hat{D}^{(t)}_{\lambda+1}}\right]}{S(t)}\sqrt{m-1}$ follows the Student's t-distribution with $m-1$ degrees of freedom (Oliphant, 2006).

Finally, we use the inverse of the Student's t CDF to find the value that this random variable would only exceed with probability γ. The result follows by simple arithmetical operations. □

B. Evaluation

B.1. Experimental setting

All experiments were performed on a machine with Intel Xeon E5-2680 (v3), 256 GB of RAM, and two NVIDIA TITAN X graphics cards. We train a classifier represented by a neural network on MNIST (LeCun et al., 1998) and on CIFAR10 (Krizhevsky, 2009) using DP-SGD. The first dataset contains 60,000 training examples and 10,000 testing images. We use large batch sizes of 1024, clip gradient norms to $C = 1$, and $\sigma = 0.1$. We also experimented with the idea of dropping updates for a random subset of weights, and achieved the best performance with updating 10% of weights at each iteration. The second dataset consists of 50,000 training images and 10,000 testing images of objects split in 10 classes. For this dataset, we use the batch size of 512, $C = 1$, and $\sigma = 0.8$. We fix $\delta = 10^{-5}$ in all experiments, and $\delta_\mu = 10^{-10}$ to achieve $(\varepsilon, 10^{-5})$ bound for 99.999% of data distribution using Markov inequality.

MNIST experiments are performed with the CNN model from Tensorflow tutorial (the same as in (Abadi et al., 2016), except we do not use PCA), trained using SGD with the learning rate 0.02. In case of CIFAR10, in order for our results to be comparable to (Abadi et al., 2016), we pre-train convolutional layers of the model on a different dataset and retrain a fully-connected layer in a privacy-preserving way. We were unable to reproduce the experiment exactly as specified in (Abadi et al., 2016) and chose a different model (VGG-16 pre-trained on ImageNet), guided by maintaining a similar or lower non-private accuracy. The model was trained using Adam with the learning rate of 0.001. Since the goal of these experiments is to show relative performance of private methods, we did not perform an exhaustive search for hyperparameters, either using default or previously published values or values that yield reasonable training behaviour.

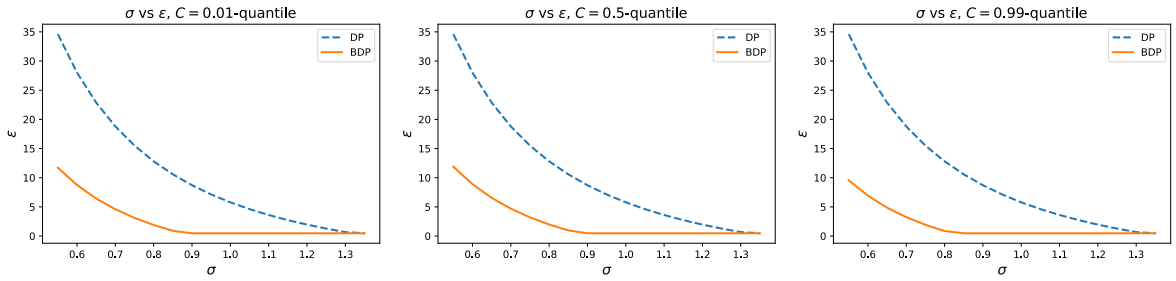
Privacy accounting with DP-SGD works in the following way. The non-private learning outcome at each iteration t is the gradient g_t of the loss function w.r.t. the model parameters, the outcome distribution is the Gaussian $\mathcal{N}(g_t, \sigma^2 C^2)$. Before adding noise, the norm of the gradients is clipped to C. For the moments accountant, the privacy loss is calculated using this C and σ. For the Bayesian accountant, either pairs of examples x_i, x_j or pairs of batches are sampled from the dataset at each iteration, and used to compute $\hat{c}_t(\lambda)$. Although clipping gradients is no longer necessary with the Bayesian accountant, it is highly beneficial for incurring lower privacy loss at each iteration and obtaining tighter composition. Moreover, it ensures the classic DP bounds on top of BDP bounds.

We also run evaluation on two binary classification tasks taken from UCI database: Abalone (Waugh, 1995) (predicting the age of abalone from physical measurements) and Adult (Kohavi, 1996) (predicting income based on a person's attributes). In this setting, we compare differentially private variational inference (DPVI-MA (Jälkö et al., 2016)) to the variational inference with BDP. The datasets have 4,177 and 48,842 examples with 8 and 14 attributes accordingly. We use the same pre-processing and models as (Jälkö et al., 2016). We run experiments using the authors original implementation (`https://github.com/DPBayes/DPVI-code`) with slight modifications (e.g. accounting randomness of sampling from variational distributions, instead of adding noise, using Bayesian accountant, and performing classification with variational samples instead of optimal variational parameters).

B.2. Effect of σ and bounded sensitivity

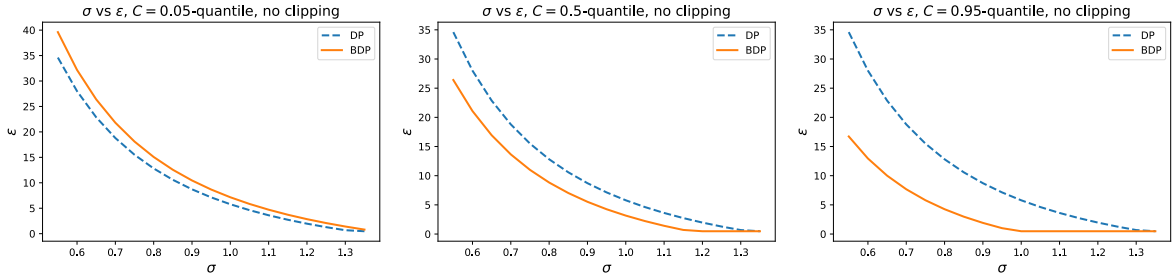
The primary goal of our paper is to obtain more meaningful privacy guarantees sacrificing as little utility as possible. The main factor in the loss of utility is the variance of the noise we add during training. Therefore it is critical to examine how our guarantee behaves compared to the classic DP for the same amount of noise. Or equivalently, how much noise does it require to reach the same ε.

As stated above, there are two possible regimes of operation for the Gaussian noise mechanism under Bayesian differential privacy: with bounded sensitivity and with unbounded sensitivity. The first is just like the classic DP: there is a



(a) Clipping at 0.01-quantile of $\|\nabla f\|$. (b) Clipping at 0.50-quantile of $\|\nabla f\|$. (c) Clipping at 0.99-quantile of $\|\nabla f\|$.

Figure 1: Dependency between σ and ε for different C when clipping for both DP and BDP.



(a) Noise at 0.05-quantile of $\|\nabla f\|$. (b) Noise at 0.50-quantile of $\|\nabla f\|$. (c) Noise at 0.95-quantile of $\|\nabla f\|$.

Figure 2: Dependency between σ and ε for different C when clipping for DP and not clipping for BDP.

maximum bound on the contribution of an individual example, and the noise is scaled to it. The second does not have a bound on contribution and mitigates it by taking into account the low probability of extreme contributions.

Figures 1 and 2 demonstrate the dependency between σ and ε for different clipping thresholds C chosen relative to the quantiles of the gradient norm distribution. If we bound sensitivity by clipping the gradients, it ensures that BDP always requires less noise than DP to reach the same ε, as seen in Figure 1. As we decrease the clipping threshold C, more and more gradients get clipped and the BDP curve approaches the DP curve (Figure 1a). However, as we observe in Figure 2 comparing DP with bounded sensitivity and BDP with unbounded sensitivity, using unclipped gradients results in less consistent behaviour. It may require a more thorough search for the right noise variance to reach the same ε.

B.3. Effect of λ

As mentioned in Section 4.2, the privacy cost, and therefore the final value of ε, depend on the choice of λ. We run the Bayesian accountant for the Gaussian mechanism with the fixed pairwise gradient distances (s.t. these results apply exactly to the moments accountant) for different signal-to-noise ratios and different λ.

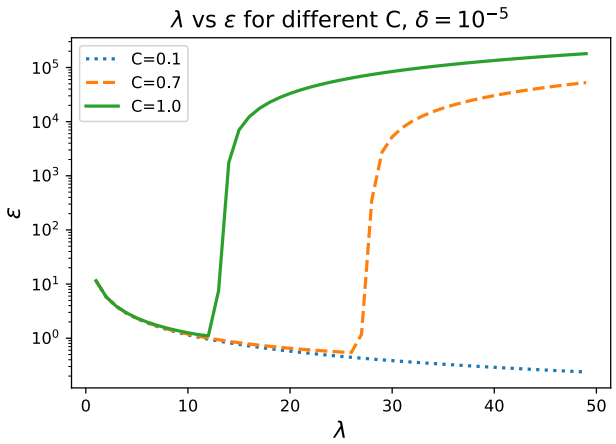


Figure 3: Dependency of λ and ε for different clipping thresholds C, $q = 64/60000$, $\sigma = 1.0$.

Depicted in Figure 3 is ε as a function of λ for 10000 steps. We observe that λ has a clear effect on the final ε value. In some cases this effect is very significant and the change is sharp. It suggests that in practice one should be careful about the choice of λ. We also note that for lower signal-to-noise ratios (e.g. $C = 0.1, \sigma = 1$) the optimal choice of λ is much further on the real line and may well be outside the typically range computed in the literature.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Jälkö, J., Dikmen, O., and Honkela, A. Differentially private variational inference for non-conjugate models. *arXiv preprint arXiv:1610.08749*, 2016.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. Citeseer, 1996.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Oliphant, T. E. A bayesian perspective on estimating mean, variance, and standard-deviation from data. 2006.

Waugh, S. G. *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. PhD thesis, University of Tasmania, 1995.