

PRIVACY AMPLIFICATION VIA ITERATION FOR SHUFFLED AND ONLINE PNSGD

Matteo Sordello

Department of Statistics
Wharton School, University of Pennsylvania
sordello@wharton.upenn.edu

Zhiqi Bu

Applied Maths and Computational Science
University of Pennsylvania
zbu@sas.upenn.edu

Jinshuo Dong

CS Theory Group
Northwestern University
jinshuo@sas.upenn.edu

Weijie J. Su

Department of Statistics
Wharton School, University of Pennsylvania
suw@wharton.upenn.edu

ABSTRACT

We consider the framework of privacy amplification via iteration, originally proposed by Feldman et al. and then simplified by Asoodeh et al. in their analysis via contraction coefficient. This line of work studies the privacy guarantees obtained by the projected noisy stochastic gradient descent (PNSGD) algorithm with hidden intermediate updates. A limitation in the existing literature is that only the randomly-stopped PNSGD has been studied, while no result has been proved on the more widely-used PNSGD applied on a shuffled dataset. Moreover, no scheme has been yet proposed regarding how to decrease the injected noise when new data are received in an online fashion. Here, we first prove a privacy guarantee for shuffled PNSGD, which is investigated asymptotically when the noise is fixed for each individual but reduced as the sample size n grows. We then provide a faster decaying scheme for the magnitude of the injected noise that also guarantees the convergence of privacy loss when new data are received in an online fashion.

1 INTRODUCTION

Differential privacy (DP) Dwork et al. (2006b), Dwork et al. (2006a) is a strong standard to guarantee the privacy for algorithms that have been widely applied to modern machine learning (Abadi et al., 2016). When multiple operations on the data are involved and each intermediate step is revealed, composition theorems can be used to keep track of the privacy loss (Kairouz et al., 2015). However, because such results are required to be general, their associated privacy bounds are inevitably loose. In contrast, privacy amplification provides a privacy budget for a composition of mechanisms that is less than the budget of each individual operation, which strengthens the bound the more operations are concatenated. Classic examples of this feature are privacy amplification by subsampling (Chaudhuri and Mishra, 2006; Balle et al., 2018), by shuffling (Erlingsson et al., 2019) and by iteration (Feldman et al., 2018; Asoodeh et al., 2020). In this paper, we focus on the setting of privacy amplification by iteration, and extend the analysis via contraction coefficient proposed by Asoodeh et al. (2020) to prove results that apply to an algorithm commonly used in practice, in which the entire dataset is shuffled before training a model with PNSGD.

We refer to Asoodeh et al. (2020) for a careful description of their setting, and only briefly introduce the quantities that are relevant for our analysis. We consider a convex function $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ that satisfies $f(1) = 0$. Ali and Silvey (1966) and Csiszár and Shields (2004) define the f -divergence between two probability distribution μ and ν as

$$D_f(\mu \parallel \nu) = \mathbb{E}_\nu \left[f \left(\frac{d\mu}{d\nu} \right) \right] = \int f \left(\frac{d\mu}{d\nu} \right) d\nu$$

For a Markov kernel $K : \mathcal{W} \rightarrow \mathcal{P}(\mathcal{W})$, where $\mathcal{P}(\mathcal{W})$ is the space of probability measures over \mathcal{W} , we let $\eta_f(K)$ be the contraction coefficient of K under the f -divergence, defined as

$$\eta_f(K) = \sup_{\mu, \nu: D_f(\mu||\nu) \neq 0} \frac{D_f(\mu K || \nu K)}{D_f(\mu || \nu)}$$

Let now $\{K_n\}$ be a sequence of Markov kernels, and the two sequences of measures $\{\mu_n\}$ and $\{\nu_n\}$ be generated starting from μ_0 and ν_0 by applying $\mu_n = \mu_{n-1}K_n$ and $\nu_n = \nu_{n-1}K_n$. The strong data processing inequality (Raginsky, 2016) for the f -divergence tells us that

$$D_f(\mu_n || \nu_n) \leq D_f(\mu_0 || \nu_0) \prod_{t=1}^n \eta_f(K_t) \quad (1)$$

Among the f -divergences, we focus on the E_γ -divergence, which is the f -divergence associated with $f(t) = (t - \gamma)_+ = \max(0, t - \gamma)$. We do so because of its connection with the concept of (ϵ, δ) differential privacy, which is now the state-of-the-art technique to analyze the privacy loss that we incur when releasing information from a dataset. It is in fact easy to prove that a mechanism \mathcal{M} is (ϵ, δ) -DP if and only if the distributions that it generates on two neighboring datasets D and D' (write $D \sim D'$) are close with respect to the E_γ -divergence. In particular, for $D \sim D'$ and \mathbb{P}_D being the output distribution of mechanism \mathcal{M} on D , then \mathcal{M} is (ϵ, δ) -DP if and only if

$$E_{e^\epsilon}(\mathbb{P}_D || \mathbb{P}_{D'}) \leq \delta. \quad (2)$$

The PNSGD is defined with respect to a loss function $\ell : \mathcal{W} \times \mathcal{X} \rightarrow \mathbb{R}$ that takes as inputs a parameter in the space $\mathbb{K} \subseteq \mathcal{W}$ and an observation $x \in \mathcal{X}$. Common assumptions made on the loss functions are the following: for each $x \in \mathcal{X}$, $\ell(\cdot, x)$ is L -Lipschitz and ρ -strongly convex, and $\nabla_w \ell(\cdot, x)$ is β -Lipschitz. The PNSGD algorithm works by combining three steps: (i) a stochastic gradient descent (SGD) step with learning rate η (ii) an injection of i.i.d. noise sampled from a known distribution to guarantee privacy and (iii) a projection $\Pi_{\mathbb{K}} : \mathcal{W} \rightarrow \mathbb{K}$ onto the subspace \mathbb{K} . Combined, these steps give the following update rule: $w_{t+1} = \Pi_{\mathbb{K}}(w_t - \eta(\nabla_w \ell(w_t, x_{t+1}) + Z_{t+1}))$. Such update can be defined as a Markov kernel by assuming that $w_0 \sim \mu_0$ and $w_t \sim \mu_t = \mu_0 K_{x_1} \dots K_{x_t}$, where K_x is the kernel associated to a single PNSGD step when observing the data point x .

Asoodeh et al. (2020) investigate the privacy guarantees obtained bounding the left hand side of (2) making use of (1) in the setting of PNSGD with Laplace or Gaussian noise. The privacy bound depends on the index at which the neighboring datasets D and D' differ and the distribution of the noise injected in the PNSGD. We report their result.

Theorem 1 (Theorem 3 and 4 in Asoodeh et al. (2020)). *Define*

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{u^2}{2}} du = 1 - \Phi(t) \quad \text{and} \quad \theta_\gamma(r) = Q\left(\frac{\log(\gamma)}{r} - \frac{r}{2}\right) - \gamma Q\left(\frac{\log(\gamma)}{r} + \frac{r}{2}\right)$$

where Φ is the cumulative density function of the standard normal. Let $M = (1 - 2\eta\beta\rho/(\beta + \rho))^{1/2}$. The PNSGD algorithm is (ϵ, δ) -DP for its i -th entry where $\epsilon \geq 0$ and $\delta = A \cdot B^{n-i}$, where

$$A = \theta_{e^\epsilon} \left(\frac{2L}{\sigma} \right), \quad B = \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma} \right) \quad (3)$$

if we consider $\mathbb{K} \subset \mathbb{R}^d$ compact and convex and Gaussian noise $N(0, \sigma^2)$, while

$$A = \left(1 - e^{\frac{\epsilon}{2} - \frac{L}{v}} \right)_+, \quad B = \left(1 - e^{\frac{\epsilon}{2} - \frac{M(b-a)}{2\eta v}} \right)_+ \quad (4)$$

if instead we consider $\mathbb{K} = [a, b]$ for $a < b$ and Laplace noise $\mathcal{L}(0, v)$.

To get a bound that does not depend on the index of the entry on which the two datasets differ, the authors later consider the randomly-stopped PNSGD, which simply consist of picking a random stopping time for the PNSGD uniformly from $\{1, \dots, n\}$. The bound that they obtain for δ in the Gaussian case is $\delta = A/[n(1 - B)]$. Based on their proof, it is clear that the actual bound contains a term $(1 - B^{n-i+1})$ at the numerator and that the same result can be obtained if we consider the Laplace noise. In Section 3 we prove that

a better bound than the one obtained via randomly-stopped PNSGD can be obtained by first shuffling the dataset and then applying the simple PNSGD. In Section 4 we study the asymptotic behavior of such bound and find the appropriate decay rate for the variability of the noise level that guarantees convergence for δ to a non-zero constant. All the proofs of our results can be found in the Appendix.

2 RELATED WORK

In the DP regime, (ϵ, δ) -DP is arguably the most popular definition, which is oftentimes achieved by an algorithm which contains Gaussian or Laplacian noises. For example, in NoisySGD and NoisyAdam in Abadi et al. (2016); Bu et al. (2020), and PNSGD in this paper, a certain level of random noise is injected into the gradient to achieve DP. Notably, as we use more datapoints (or more iterations during the optimization) during the training procedure, the privacy loss accumulates at a rate that depends on the magnitude of the noise. It is remarkably important to characterize, as tightly as possible, the privacy loss at each iteration. An increasing line of works have proposed to address this difficulty (Dong et al., 2019; Bun and Steinke, 2016; Dwork and Rothblum, 2016; Balle et al., 2018; Mironov, 2017; Wang et al., 2019; Koskela et al., 2020; Asoodeh et al., 2020; Abadi et al., 2016), which bring up many useful notions of DP, such as Rényi DP, Gaussian DP, f -DP and so on. Our paper extends Asoodeh et al. (2020) by shuffling the dataset first rather than randomly stopping the PNSGD (see Theorem 5 in Asoodeh et al. (2020)), in order to address the non-uniformity of privacy guarantee. As a consequence, we obtain a strictly better privacy bound and better loss than Asoodeh et al. (2020), Abadi et al. (2016), and an additional online result of the privacy guarantee.

3 SHUFFLED PNSGD WITH FIXED NOISE

We now prove the bound on δ that we can obtain by first shuffling the dataset and then applying the PNSGD algorithm. The simple underlying idea here is that, when shuffling the dataset, the index at which the two neighboring datasets differ has equal probability to end up in each position. This is a key difference compared to the randomly-stopped PNSGD, and allows us to get a better bound that do not depend on the initial position of that index.

Theorem 2. *Let $D \sim D'$ be of size n . Then the shuffled PNSGD is (ϵ, δ) -DP with*

$$\delta = \frac{A \cdot (1 - B^n)}{n(1 - B)} \quad (5)$$

and the constants A and B are defined in (3) for Gaussian noise and (4) for Laplace noise.

The idea behind the proof of this theorem is the following: once we shuffled the datasets, the element x_i at which they differ has $1/n$ probability of ending up in each position. When it ends up in position with index j , the privacy bound that we obtain for δ is $A \cdot B^{n-j}$. The overall bound is then $A \cdot \sum_{j=1}^n B^{n-j}/n$, giving the result in (5).

This bound is indeed better than the one found in Asoodeh et al. (2020) for the randomly stopped PNSGD since it contains an extra term $(1 - B^n)$ at the numerator which does not depend on i and is smaller than 1. When the injected noise is reduced at the appropriate rate we will see that that $B \approx 1 - O(1/n)$, so the extra term ends up having an impact in the final bound. It is also important to notice that shuffled PNSGD achieves in general better performance than randomly stopped PNSGD and it is much more commonly used in practice. In the next section we look at the asymptotic behavior of (5) when n grows and the variance of the injected noise is properly reduced to guarantee convergence.

Remark 3. *Notice that the result in Theorem 2 provides a one epoch privacy bound for shuffled PNSGD. In real experiments, e.g. when training deep neural networks, usually multiple passes over the data are necessary to learn the model. In such scenarios, the updates are not kept secret for the whole duration of the training, but are instead released at the end of each epoch. The one epoch result can easily be combined with state-of-the-art composition tools, such as the Moments Accountant (Abadi et al., 2016), f -DP and Gaussian DP (Dong*

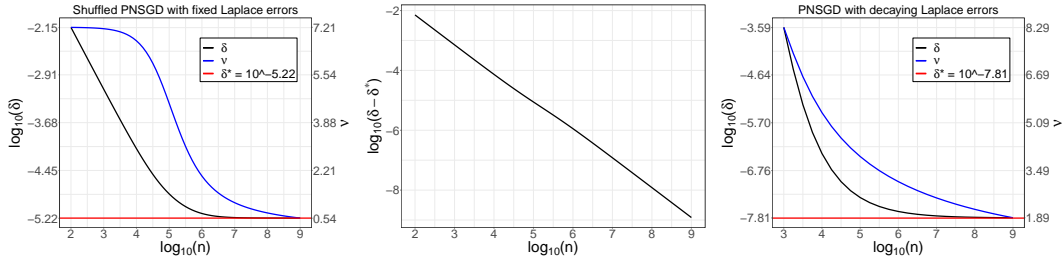


Figure 1: (left) Convergence of δ to δ_L^* (center) The convergence rate of $1/n$ (right) Convergence of δ to $\delta_L^{*,o}$. The parameters are described in Appendix G.

et al., 2019). The way to do that is to migrate from (ϵ, δ) -DP to other regimes, Gaussian DP or Rényi DP, at the first epoch, then compose in those specific regimes until the end of training procedure and at last map from the other regimes back to (ϵ, δ) -DP.

4 ASYMPTOTIC ANALYSIS OF SHUFFLED PNSGD

In this section we analyze the PNSGD algorithm applied to a shuffled dataset where for each update the noise level is fixed. In order to get a convergence result for δ as the size n of the dataset grows, the level of noise that we use should be targeted to n .

In Theorem 4 we consider Laplace noise $\mathcal{L}(0, v(n))$, while in Theorem 5 we use Gaussian noise $N(0, \sigma^2(n))$. The decay of $v(n)$ and $\sigma(n)$ is regulated by two parameters, C_1 and C_2 . While C_1 is set to be large, so that δ converges to a small value, the use of C_2 is simply to allow the noise level not to be too large for small n , but does not appear in the asymptotic bound.

Theorem 4. Consider the shuffled PNSGD with fixed Laplace noise $\mathcal{L}(0, v(n))$. Let

$$v(n) = \frac{M(b-a)}{2\eta \log(n/C_1 + C_2)} \quad \text{and} \quad \delta_L^* = \frac{1 - e^{-C_1 \exp(\epsilon/2)}}{C_1 e^{\frac{\epsilon}{2}}}. \quad (6)$$

Then, for n sufficiently large the procedure is (ϵ, δ) -DP with $\delta = \delta_L^* + O(1/n)$.

The convergence result in Theorem 4 is confirmed by left and center panels of Figure 1, where we see that δ converges to δ_L^* at the correct convergence rate $1/n$.

Theorem 5. Consider the shuffled PNSGD algorithm with fixed Gaussian noise $N(0, \sigma^2(n))$. Let

$$\sigma(n) = \frac{MD_{\mathbb{K}}}{2\eta \sqrt{W\left(\frac{n^2}{2C_1^2\pi} + C_2\right)}} \quad \text{and} \quad \delta_G^* = \frac{1 - e^{-2C_1 e^{\frac{\epsilon}{2}}}}{2C_1 e^{\frac{\epsilon}{2}}} \quad (7)$$

where W is the Lambert W function. Then, for n sufficiently large, the procedure is (ϵ, δ) -DP with $\delta = \delta_G^* + O\left(\frac{1}{\log(n)}\right)$.

This result is similar to the Laplace case with the main differences being the slightly more complicated form for the decay of $\sigma(n)$ and the fact that the convergence happens more slowly, at a rate of $1/\log(n)$. In the left and center panels of Figure 2 we get an empirical confirmation of this result, and in Figure 3 in the Appendix we show different convergence patterns, from above and below δ_G^* .

5 ONLINE PNSGD WITH DECAYING NOISE

We go back to the original framework of Asoodeh et al. (2020) and consider the PNSGD algorithm applied to the non-shuffled dataset and a different level of noise for each update. The definition of A_i and B_i is the same as in (3) and (4) but the noise level v and σ is

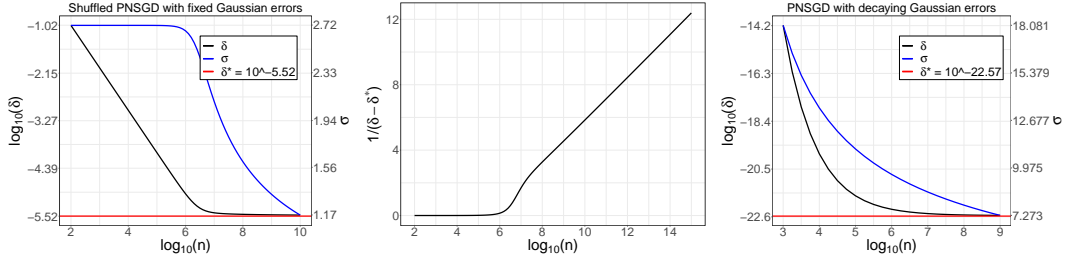


Figure 2: (left) Convergence of δ to δ_G^* (center) The convergence rate is $1/\log(n)$ (right) Convergence of δ to $\delta_G^{*,o}$. The parameters are described in Appendix G.

now dependent on the position of each element in the dataset. When $D \sim D'$ differ on index i , we get $\delta = A_i \cdot \prod_{t=i+1}^n B_t$. In this scenario we can add new data to D in an online fashion, without having to restart the procedure to recalibrate the noise level used for the first entries. It is interesting to notice that for both the Laplace and Gaussian noise the only difference needed with the decay rate for $v(n)$ and $\sigma(n)$ defined before is an exponent $\alpha > 1$. In Theorem 6 we consider for the entry with index j an injected Laplace noise $\mathcal{L}(0, v_j)$, while in Theorem 7 we use Gaussian noise $N(0, \sigma_j^2)$.

Theorem 6. Consider the PNSGD where for update j we use Laplace noise $\mathcal{L}(0, v_j)$. Let

$$v_j = \frac{M(b-a)}{2\eta \log(j^\alpha/C_1 + C_2)} \quad \text{and} \quad \delta_L^{*,o} = \left(1 - e^{\frac{\epsilon}{2} - \frac{2L\eta \log(i^\alpha/C_1 + C_2)}{M(b-a)}}\right)_+ e^{\int_{i+1}^{\infty} \log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{x^\alpha + C_1 C_2}\right) dx}$$

for $\alpha > 1$. Then as $n \rightarrow \infty$ the procedure is $(\epsilon, \delta_L^{*,o})$ -DP.

Theorem 7. Consider the PNSGD where for update j we use Gaussian noise $N(0, \sigma_j^2)$. Let

$$\sigma_j = \frac{MD_{\mathbb{K}}}{2\eta \sqrt{W\left(\frac{j^{2\alpha}}{2\pi C_1^2} + C_2\right)}} \quad \text{and} \quad \delta_G^{*,o} = \theta_{e^\epsilon} \left(\frac{2L}{\sigma_i}\right) e^{\int_{i+1}^{\infty} \log\left(\theta_{e^\epsilon} \left(2\sqrt{W\left(\frac{x^{2\alpha}}{2\pi C_1^2} + C_2\right)}\right)\right) dx}$$

for $\alpha > 1$. Then as $n \rightarrow \infty$ the procedure is $(\epsilon, \delta_G^{*,o})$ -DP.

These two convergence results are confirmed respectively in the right panel of Figure 1 and in the right panel of Figure 2. Notice that $\delta_L^{*,o}$ and $\delta_G^{*,o}$ are upper bounds for the actual limit of δ , since they are obtained bounding from above a sum with an integral. However, we discuss in Appendix E why the convergence appears to be impeccable. Notice that for the last elements in the dataset the injected noise in the online setting is smaller than the fixed noise in the shuffled setting. This makes sense, since in the online setting the noise level for the first entries of the dataset is not adjusted to n , and is instead kept large.

6 CONCLUSION

In this work, we have studied the setting of privacy amplification by iteration in the formulation proposed by Asoodeh et al. (2020), and proved that their analysis of PNSGD also applies to the case where the data are shuffled first. This is a much more common practice than the randomly-stopped PNSGD, originally proposed, because of a clear advantage in terms of accuracy of the algorithm. We proved two asymptotic results on the decay rate of noises that we can use, either the Laplace or the Gaussian injected noise, in order to have asymptotic convergence to a non-trivial privacy bound when the size of the dataset grows. Finally we also showed two results, again for Laplace or Gaussian noise, that can be obtained in an online setting when the noise does not have to be recalibrated for the whole dataset but just decayed for the new data. In addition, our single-epoch analysis can be easily combined with standard composition tools to derive multi-epoch privacy guarantees.

REFERENCES

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.
- Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.
- Asoodeh, S., Diaz, M., and Calmon, F. P. (2020). Privacy amplification of iterative algorithms via contraction coefficients. *arXiv preprint arXiv:2001.06546*.
- Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, pages 6277–6287.
- Bu, Z., Dong, J., Long, Q., and Su, W. J. (2020). Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23).
- Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer.
- Chaudhuri, K. and Mishra, N. (2006). When random sampling preserves privacy. In *Annual International Cryptology Conference*, pages 198–213. Springer.
- Csiszár, I. and Shields, P. C. (2004). *Information theory and statistics: A tutorial*. Now Publishers Inc.
- Dong, J., Roth, A., and Su, W. J. (2019). Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.
- Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.
- Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.
- Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. (2019). Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM.
- Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. (2018). Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE.
- Kairouz, P., Oh, S., and Viswanath, P. (2015). The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.
- Koskela, A., Jälkö, J., and Honkela, A. (2020). Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR.
- Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE.
- Raginsky, M. (2016). Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389.
- Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR.

A PROOF OF THEOREM 2

Let's start by considering the simple case $n = 2$, so that $D = \{x_1, x_2\}$ and $D' = \{x'_1, x'_2\}$ and let $i \in \{1, 2\}$ be the index at which they differ. Let μ be the output distribution of the shuffled PNSGD on D , and ν be the corresponding distribution from D' . If we define $S(D)$ and $S(D')$ to be the two dataset after performing the same shuffling, then we can only have either $S(D) = \{x_1, x_2\}$ or $S(D) = \{x_2, x_1\}$, both with equal probability $1/2$. The outcomes of the shuffled PNSGD on D and D' are then

$$\begin{aligned}\mu &= \frac{1}{2}\mu_0 K_{x_1} K_{x_2} + \frac{1}{2}\mu_0 K_{x_2} K_{x_1} \\ \nu &= \frac{1}{2}\mu_0 K_{x'_1} K_{x'_2} + \frac{1}{2}\mu_0 K_{x'_2} K_{x'_1}\end{aligned}$$

By convexity and Jensen's inequality we have that

$$\begin{aligned}E_\gamma(\mu\|v) &\leq \frac{1}{2}E_\gamma(\mu_0 K_{x_1} K_{x_2} \| \mu_0 K_{x'_1} K_{x'_2}) \\ &\quad + \frac{1}{2}E_\gamma(\mu_0 K_{x_2} K_{x_1} \| \mu_0 K_{x'_2} K_{x'_1})\end{aligned}$$

and now we have two options, based on where the two original datasets differ. If $i = 1$, in the first term the privacy is stronger than in the second one (because x_1 is seen earlier), and we have

$$E_\gamma(\mu\|v) \leq \frac{1}{2}A \cdot B + \frac{1}{2}A = \frac{1}{2}A(B+1)$$

If $i = 2$, now the privacy is stronger in the second term, and

$$E_\gamma(\mu\|v) \leq \frac{1}{2}A + \frac{1}{2}A \cdot B = \frac{1}{2}A(B+1)$$

Since in both cases the bound is the same, this means that for any $i \in \{1, 2\}$ the privacy guarantee of the shuffled PNSGD algorithm is equal to $A(B+1)/2$. From here we see that, when $n > 2$, the situation is similar. Instead of just two, we have $n!$ possible permutations for the elements of D , each one happening with the same probability $1/n!$. For each fixed index i on which the two neighboring datasets differ, we have $(n-1)!$ permutations in which element x_i appears in each of the n positions. When, after the permutation, element x_i ends up in last position, the bound on $E_\gamma(\mu\|v)$ is the weakest and just equals A . When it ends up in first position, the bound is the strongest and is equal to $A \cdot B^{n-1}$. We then have that, irrespectively of the index i ,

$$E_\gamma(\mu\|v) \leq \frac{1}{n!}(n-1)!A \sum_{j=0}^{n-1} B^j = \frac{A \cdot (1 - B^n)}{n(1 - B)}$$

B PROOF OF THEOREM 4

We use the result in Theorem 2 combined with (4), and get that

$$\delta = \frac{\left(1 - e^{\frac{\epsilon}{2} - \frac{L}{v(n)}}\right)_+ \cdot \left[1 - \left(1 - e^{\frac{\epsilon}{2} - \frac{M(b-a)}{2\eta v(n)}}\right)_+^n\right]}{n \cdot e^{\frac{\epsilon}{2} - \frac{M(b-a)}{2\eta v(n)}}$$

Once we plug in the $v(n)$ defined in (6) we have that, when n is sufficiently large,

$$\begin{aligned}\delta &= \frac{\left(1 - e^{\frac{\epsilon}{2} - \frac{2L\eta \log(\frac{n}{C_1} + C_2)}{M(b-a)}}\right)_+ \cdot \left[1 - \left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{n + C_1 C_2}\right)_+^n\right]}{n \cdot e^{\frac{\epsilon}{2} - \log(\frac{n}{C_1} + C_2)}} \\ &= \frac{\left[1 - \left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{n + C_1 C_2}\right)_+^n\right]}{n \cdot \frac{C_1 e^{\frac{\epsilon}{2}}}{n + C_1 C_2}} \cdot \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= \frac{1 - e^{-C_1 \exp(\epsilon/2)}}{C_1 e^{\frac{\epsilon}{2}}} + O\left(\frac{1}{n}\right)\end{aligned}$$

C AN IMPORTANT LEMMA FOR THE GAUSSIAN CASE

Lemma 8. *For a sufficiently small σ and two constants c and ϵ , we have*

$$\theta_{e^\epsilon} \left(\frac{c}{\sigma} \right) = 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{c^2}{8\sigma^2}} \left(\frac{4\sigma}{c} + O(\sigma^3) \right).$$

Recall from the definition of the function $\theta_\gamma(r)$ that

$$\theta_{e^\epsilon} \left(\frac{c}{\sigma} \right) = Q \left(\frac{\epsilon\sigma}{c} - \frac{c}{2\sigma} \right) - e^\epsilon Q \left(\frac{\epsilon\sigma}{c} + \frac{c}{2\sigma} \right) \quad (8)$$

We apply the following approximation of the normal cumulative density function, valid for large positive x ,

$$Q(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} + \dots \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{1}{x} + O\left(\frac{1}{x^3}\right) \right) \quad (9)$$

and similarly, for large negative values of x

$$Q(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du = 1 + \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{1}{x} + O\left(\frac{1}{x^3}\right) \right).$$

Therefore (8) can be reformulated as

$$\begin{aligned} \theta_{e^\epsilon} \left(\frac{c}{\sigma} \right) &= 1 + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\epsilon^2\sigma^2}{c^2} + \frac{c^2}{4\sigma^2} \right) + \frac{\epsilon}{2}} \left(\frac{1}{\frac{\epsilon\sigma}{c} - \frac{c}{2\sigma}} + O\left(\frac{1}{\left(\frac{\epsilon\sigma}{c} - \frac{c}{2\sigma}\right)^3}\right) \right) \\ &\quad - \frac{1}{\sqrt{2\pi}} e^{\epsilon} e^{-\frac{1}{2} \left(\frac{\epsilon^2\sigma^2}{c^2} + \frac{c^2}{4\sigma^2} \right) - \frac{\epsilon}{2}} \left(\frac{1}{\frac{\epsilon\sigma}{c} + \frac{c}{2\sigma}} + O\left(\frac{1}{\left(\frac{\epsilon\sigma}{c} + \frac{c}{2\sigma}\right)^3}\right) \right) \\ &= 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{1}{2} \left(\frac{\epsilon^2\sigma^2}{c^2} + \frac{c^2}{4\sigma^2} \right)} \left(\frac{4\sigma}{c} + O(\sigma^3) \right) \\ &= 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{c^2}{8\sigma^2}} \left(\frac{4\sigma}{c} + O(\sigma^3) \right) \end{aligned}$$

D PROOF OF THEOREM 5

From Theorem 2 we know that

$$\delta = \frac{\theta_{e^\epsilon} \left(\frac{2L}{\sigma(n)} \right) \cdot \left[1 - \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma(n)} \right)^n \right]}{n \cdot \left[1 - \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma(n)} \right) \right]} \quad (10)$$

We show that with $\sigma(n)$ that decays according to (7) we have that

$$\theta_{e^\epsilon} \left(\frac{2L}{\sigma(n)} \right) \rightarrow 1 \quad \text{and} \quad \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma(n)} \right) \rightarrow 1 - \frac{2C_1 e^{\frac{\epsilon}{2}}}{n}.$$

Let's first focus briefly on the behavior of the Lambert W function. Formally, the Lambert W function is an implicit function defined as the inverse of $f(w) = we^w$, meaning that for any x one has $W(x)e^{W(x)} = x$. As an interesting fact, we note that the Lambert W function's behavior is approximately logarithmic, e.g. $\log(x) > W(x) > \log_4(x)$, where by \log we denote the natural logarithm. We also denote the argument of the W Lambert function in $\sigma(n)$ as $x = \frac{n^2}{2C_1^2\pi} + C_2$. Using this fact, an immediate consequence of Lemma 8 is that, when plugging in the $\sigma(n)$ from (7), we get

$$\theta_{e^\epsilon} \left(\frac{2L}{\sigma(n)} \right) = 1 - o(\sigma^3) = 1 - o\left(\frac{1}{\sqrt{W^3(x)}}\right) = 1 - o\left(\frac{1}{(\log n)^{3/2}}\right)$$

since $e^{-\frac{c^2}{8\sigma^2}} \cdot \frac{1}{\sigma^2} \rightarrow 0$ as the exponential decays faster than the polynomial. Next, we study $\theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma(n)} \right)$. Again by Lemma 8, we have

$$\begin{aligned}
\theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma(n)} \right) &= 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{M^2 D_{\mathbb{K}}^2}{8\eta^2 \sigma(n)^2}} \left(\frac{4\eta\sigma(n)}{MD_{\mathbb{K}}} + O(\sigma(n)^3) \right) \\
&= 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{W(x)}{2}} \left(\frac{2}{\sqrt{W(x)}} + O\left(\frac{1}{W(x)^{3/2}}\right) \right) \\
&= 1 - \frac{2e^{\frac{\epsilon}{2}}}{\sqrt{2\pi}} \frac{1}{\sqrt{e^{W(x)}W(x)}} + O\left(\frac{1}{\sqrt{e^{W(x)}W(x)^3}}\right) \\
&= 1 - \frac{2e^{\frac{\epsilon}{2}}}{\sqrt{2\pi x}} + O\left(\frac{1}{\sqrt{x \log(x)}}\right) \\
&= 1 - \frac{2C_1 e^{\frac{\epsilon}{2}}}{n} + O\left(\frac{1}{n \log(n)}\right)
\end{aligned} \tag{11}$$

Going back to the expression in (10) we finally have that

$$\begin{aligned}
\delta &= \frac{\left(1 - o\left(\frac{1}{(\log(n))^{3/2}}\right)\right) \left[1 - \left(1 - \frac{2C_1 e^{\frac{\epsilon}{2}}}{n} + O\left(\frac{1}{n \log(n)}\right)\right)^n\right]}{n \cdot \left[1 - \left(1 - \frac{2C_1 e^{\frac{\epsilon}{2}}}{n} + O\left(\frac{1}{n \log(n)}\right)\right)\right]} \\
&= \frac{\left(1 - o\left(\frac{1}{(\log(n))^{3/2}}\right)\right) \left[1 - \left(1 - \frac{2C_1 e^{\frac{\epsilon}{2}} + O(1/\log(n))}{n}\right)^n\right]}{2C_1 e^{\frac{\epsilon}{2}} + O\left(\frac{1}{\log(n)}\right)} \\
&= \frac{\left(1 - o\left(\frac{1}{(\log(n))^{3/2}}\right)\right) \left[1 - e^{-2C_1 e^{\frac{\epsilon}{2}} + O(1/\log(n))}\right]}{2C_1 e^{\frac{\epsilon}{2}} + O\left(\frac{1}{\log(n)}\right)} \\
&= \left(1 - o\left(\frac{1}{(\log(n))^{3/2}}\right)\right) \left[1 - e^{-2C_1 e^{\frac{\epsilon}{2}}} + O\left(\frac{1}{\log(n)}\right)\right] \left(\frac{1}{2C_1 e^{\frac{\epsilon}{2}}} - O\left(\frac{1}{\log(n)}\right)\right) \\
&= \frac{1 - e^{-2C_1 e^{\frac{\epsilon}{2}}}}{2C_1 e^{\frac{\epsilon}{2}}} + O\left(\frac{1}{\log(n)}\right)
\end{aligned}$$

where the last equality holds because $f(n) = O\left(\frac{1}{\log(n)}\right) + o\left(\frac{1}{(\log(n))^{3/2}}\right) = O\left(\frac{1}{\log(n)}\right)$.

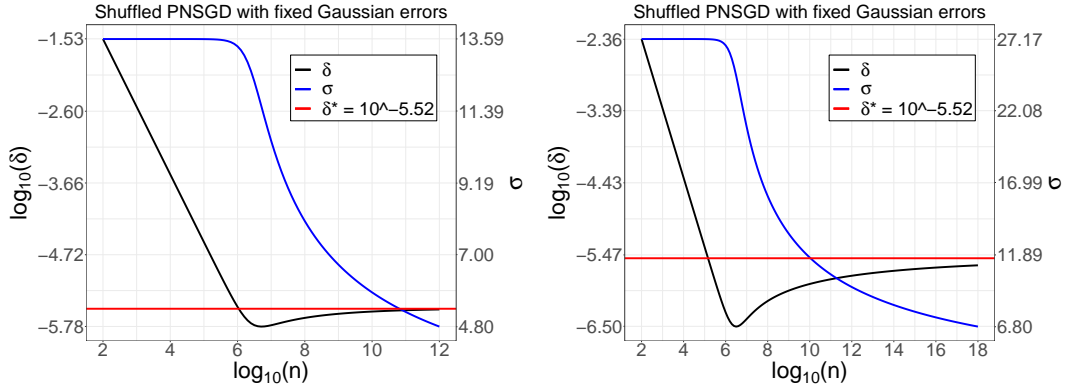


Figure 3: Convergence of δ to δ^*G . We consider $\eta \in \{0.02, 0.01\}$ and the other parameters are listed in Appendix G.

E PROOF OF THEOREM 6

We show again that δ converges to a non-zero value as n goes to ∞ . In fact, again following the proof of (Asoodeh et al. (2020) Theorem 3), we get that,

$$\begin{aligned}\delta &= \left(1 - e^{\frac{\epsilon}{2} - \frac{L}{v_i}}\right)_+ \cdot \prod_{t=i+1}^n \left(1 - e^{\frac{\epsilon}{2} - \frac{M(b-a)}{2\eta v_t}}\right)_+ \\ &= \left(1 - e^{\frac{\epsilon}{2} - \frac{2L\eta \log(\frac{i\alpha}{C_1} + C_2)}{M(b-a)}}\right)_+ \prod_{t=i+1}^n \left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right)_+\end{aligned}$$

We know that, for a sequence a_t of positive values, $\prod_{t=1}^{\infty} (1 - a_t)$ converges to a non-zero number if and only if $\sum_{t=1}^{\infty} a_t$ converges. Here we have that

$$\sum_{t=i+1}^{\infty} \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2} \leq \sum_{t=i+1}^{\infty} \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha}$$

and, since $\alpha > 1$ the right hand side converges, hence δ converges to a non-zero number. Let now $f(n) = \prod_{t=i+1}^n \left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right)_+$. To find the limit $f(\infty)$ we can first log-transform this function, and then upper bound the infinite sum with an integral before transforming back. Since $\log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right)$ is monotonically increasing in t , we have

$$\begin{aligned}\log(f(n)) &= \sum_{t=i+1}^n \log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right) \\ &< \int_{i+1}^n \log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right) dt \\ &\rightarrow \int_{i+1}^{\infty} \log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right) dt.\end{aligned}$$

This integral can be written in closed form using the hypergeometric function, or approximated numerically.

The convergence result that we get is slightly conservative, since we found an upper bound for $\log(f(n))$. However, we can find an easy lower bound by just noticing that

$$\log(f(\infty)) > \int_i^{\infty} \log\left(1 - \frac{C_1 e^{\frac{\epsilon}{2}}}{t^\alpha + C_1 C_2}\right) dt.$$

When i is not too small, the difference between the upper and lower bound is negligible, as it is confirmed by what we see in the right panel of Figure 1, where the convergence to the upper bound appears to be impeccable. Since the convergence is not exactly to δ^* , we cannot find an explicit convergence rate the same way we did in Section 4. However, we see in the left panel of Figure 4 that the convergence rate empirically appears to be $1/\log(n)$.

F PROOF OF THEOREM 7

This proof combines elements of the proofs of Theorem 5 and Theorem 6 to show that asymptotically the terms $B_t = \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma_t}\right)$ behave approximately as $1 - O(1/t^\alpha)$ as t grows.

We define $x = \frac{t^{2\alpha}}{2\pi C_1^2} + C_2$ so that $\sigma_t = \frac{MD_{\mathbb{K}}}{2\eta\sqrt{W(x)}}$ and get, as in (11),

$$\begin{aligned}\theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma_t}\right) &= 1 - \frac{1}{\sqrt{2\pi}} e^{\frac{\epsilon}{2}} e^{-\frac{M^2 D_{\mathbb{K}}^2}{8\eta^2 \sigma_t^2}} \left(\frac{4\eta\sigma_t}{MD_{\mathbb{K}}} + O(\sigma_t^3)\right) \\ &= 1 - \frac{2e^{\frac{\epsilon}{2}}}{\sqrt{2\pi x}} + O\left(\frac{1}{\sqrt{x} \log(x)}\right) \\ &= 1 - \frac{2C_1 e^{\frac{\epsilon}{2}}}{t^\alpha} + O\left(\frac{1}{t^\alpha \log(t)}\right)\end{aligned}$$

This already confirms us that δ^* converges to a finite non zero value, since the asymptotic behavior of each term in the infinite product is the same as in the Laplace case. To express such limit in a more tractable way we follow the proof of Theorem 6 and write $f(n) = \prod_{t=i+1}^n \theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma_t} \right)$ and approximate the infinite sum $\log(f(\infty))$ with an integral.

$$\begin{aligned} \log(f(n)) &= \sum_{t=i+1}^n \log \left(\theta_{e^\epsilon} \left(\frac{MD_{\mathbb{K}}}{\eta\sigma_t} \right) \right) \\ &= \sum_{t=i+1}^n \log \left(\theta_{e^\epsilon} \left(2\sqrt{W \left(\frac{t^{2\alpha}}{2\pi C_1^2} + C_2 \right)} \right) \right) \\ &< \int_{i+1}^n \log \left(\theta_{e^\epsilon} \left(2\sqrt{W \left(\frac{x^{2\alpha}}{2\pi C_1^2} + C_2 \right)} \right) \right) dx \\ &\rightarrow \int_{i+1}^{\infty} \log \left(\theta_{e^\epsilon} \left(2\sqrt{W \left(\frac{x^{2\alpha}}{2\pi C_1^2} + C_2 \right)} \right) \right) dx \end{aligned}$$

This confirms us that

$$\delta^* = \theta_{e^\epsilon} \left(\frac{2L}{\sigma_i} \right) \cdot \exp \left\{ \int_{i+1}^{\infty} \log \left(\theta_{e^\epsilon} \left(2\sqrt{W \left(\frac{x^{2\alpha}}{2\pi C_1^2} + C_2 \right)} \right) \right) dx \right\}.$$

As before, we see empirically in the right panel of Figure 4 that the convergence rate appears to be $1/\log(n)$.

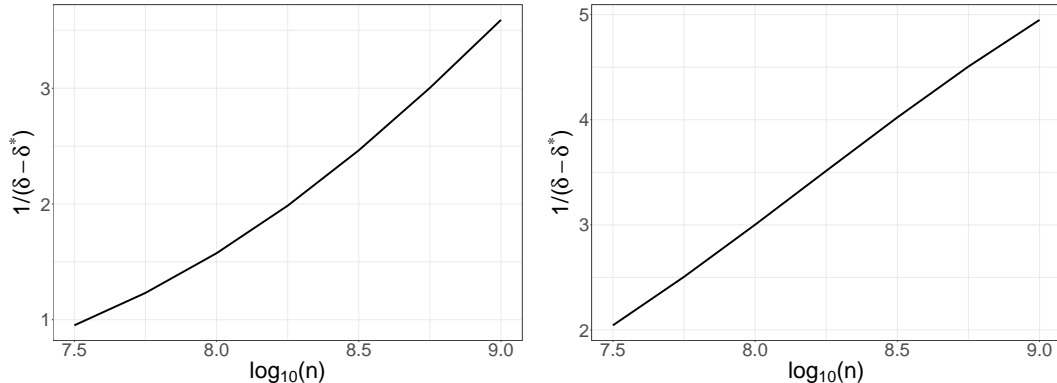


Figure 4: The convergence rates of δ to $\delta_L^{*,o}$ (left) and $\delta_G^{*,o}$ (right) in the online setting appears to be logarithmic for both the Laplace and Gaussian noise.

G DESCRIPTION OF EXPERIMENTAL SETTING

Here we list the parameters that we used in the experiments throughout the paper.

For the shuffled PNSGD with Laplace noise (left and center panels of Figure 1) we used $L = 10, \beta = 0.5, \rho = 0, \eta = 0.1, \epsilon = 1, (a, b) = (0, 1), C_1 = 10^5$ and $C_2 = 2$. For the shuffled PNSGD with Gaussian noise (left and center panels of Figure 2) we used $L = 10, \beta = 0.5, \rho = 0, \eta = 0.1, \epsilon = 1, D_{\mathbb{K}} = 1, C_1 = 10^5$ and $C_2 = 100$ while in Figure 3 we considered different values of $\eta \in \{0.02, 0.01\}$. For the online settings we used the same set of parameters: $L = 10, \beta = 0.5, \rho = 0, \epsilon = 1, \eta = 0.01, \alpha = 1.5, i = 100, C_1 = 100, C_2 = 100$ and as usual we set $(a, b) = (0, 1)$ in the Laplace case and $D_{\mathbb{K}} = 1$ in the Gaussian case.