Subsampling is not Magic: Why Large Batch Sizes Work for Differentially Private Stochastic Optimisation

Ossi Räisä [*1] **Joonas Jälkö** [*1] **Antti Honkela** [1]

Abstract

We study how the batch size affects the total gradient variance in differentially private stochastic gradient descent (DP-SGD), seeking a theoretical explanation for the usefulness of large batch sizes. As DP-SGD is the basis of modern DP deep learning, its properties have been widely studied, and recent works have empirically found large batch sizes to be beneficial. However, theoretical explanations of this benefit are currently heuristic at best. We first observe that the total gradient variance in DP-SGD can be decomposed into subsampling-induced and noise-induced variances. We then prove that in the limit of an infinite number of iterations, the effective noise-induced variance is invariant to the batch size. The remaining subsampling-induced variance decreases with larger batch sizes, so large batches reduce the effective total gradient variance. We confirm numerically that the asymptotic regime is relevant in practical settings when the batch size is not small, and find that outside the asymptotic regime, the total gradient variance decreases even more with large batch sizes. We also find a sufficient condition that implies that large batch sizes similarly reduce effective DP noise variance for one iteration of DP-SGD.

1. Introduction

As deep learning models are being trained on ever larger datasets, the privacy of the subjects of these training datasets is a growing concern. *Differential privacy* (DP) (Dwork et al., 2006b) is a property of an algorithm that formally quantifies the privacy leakage that can result from releasing the output of the algorithm. Due to the formal guarantee provided by DP, there is a great deal of interest in training deep learning models with a DP variant of stochastic gradient descent (DP-SGD) (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016).

One of the key properties of DP is so called *subsampling amplification* (Li et al., 2012; Beimel et al., 2014). Broadly speaking, subsampling the data before applying a DP mechanism adds an additional layer of protection to the data samples, leading to stronger privacy guarantees for a fixed amount of added noise. Quantifying the gains from subsampling amplification (Abadi et al., 2016; Zhu & Wang, 2019; Koskela et al., 2020; Zhu et al., 2022) has been a crucial component in making algorithms such as DP-SGD work under strict privacy guarantees.

In non-DP gradient descent, subsampling the gradients has been shown to provide better generalization on many occasions (Hardt et al., 2016; Mou et al., 2018; Kleinberg et al., 2018). The reason for the improved generalization is that the subsampling-induced noise allows the optimization to avoid collapsing into a local optimum early on (Keskar et al., 2017). However, when optimizing with DP-SGD, many works have shown that the large batch sizes actually provide better performance compared to heavy subsampling of the training data (McMahan et al., 2018; De et al., 2022; Mehta et al., 2023). This suggests that there are some fundamental differences in how the level of subsampling affects the performance of non-private and private SGD.

We focus on the Poisson subsampled Gaussian mechanism, which is the basis for the privacy analysis of DP-SGD. The privacy analysis can be done with several different upper bounds (Abadi et al., 2016; Zhu & Wang, 2019; Koskela et al., 2020), but we only look at the tight analysis that gives the smallest possible privacy bounds for the subsampled Gaussian mechanism. The tight analysis is what modern numerical privacy accountants (Koskela et al., 2020; 2021; Gopi et al., 2021; Doroshenko et al., 2022; Alghamdi et al., 2023) approximate.

In Poisson subsampling, each datapoint is included with probability q, called the *subsampling rate*, which is proportional to the expected batch size. The total gradient variance

[*]Equal contribution [1]Department of Computer Science, University of Helsinki. Correspondence to: Ossi Räisä <ossi.raisa@helsinki.fi>, Joonas Jälkö <joonas.jalko@helsinki.fi>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

in DP-SGD can be decomposed into two parts: the subsampling variance and the Gaussian noise variance σ^2. A larger q reduces the subsampling variance, but the effect on the noise variance is not as clear. On one hand, a larger q reduces the privacy amplification effect, necessitating a larger σ^2, but on the other hand, an unbiased gradient estimate must be divided by q, so the effective noise variance is $\frac{\sigma^2}{q^2}$.

We study how the effective noise variance scales with the subsampling rate, making the following contributions:

1. In Section 5, we prove that in the limit of an infinite number of iterations, there is a linear relationship $\sigma = cq$ between q and σ, meaning that the two effects q has on the effective noise variance cancel each other, leaving the effective noise variance invariant to the subsampling rate. This means that a larger subsampling rate always reduces the effective total gradient variance, since the subsampling-induced variance decreases with a larger subsampling rate.

2. In Section 6, we consider the case of a single iteration of the subsampled Gaussian mechanism. We find a sufficient condition which implies that large subsampling rates always reduce the effective injected DP noise variance, hence also reducing the effective total gradient variance. We check this condition numerically for a wide grid of hyperparameter values, and find that the condition holds amongst these hyperparameters.

3. We look at the relationship between the subsampling rate and noise standard deviation empirically in Section 5.1, and find that the asymptotic regime from our theory is reached quickly with small privacy parameters. Moreover, we find that when we are not in the asymptotic regime, the effective injected DP noise variance decreases even more with a large subsampling rate.

1.1. Related Work

Several works have empirically observed that large batch sizes are useful in DP-SGD (McMahan et al., 2018; De et al., 2022; Mehta et al., 2023). Indeed, the linear relationship between the subsampling rate and noise standard deviation we study has been suggested as a heuristic rule-of-thumb in previous work (Li et al., 2021; Sander et al., 2023) to explain these results.

The linear relationship also appears in several works that study Rényi DP (Mironov, 2017) accounting of the subsampled Gaussian mechanism (Abadi et al., 2016; Bun et al., 2018; Mironov et al., 2019), though these works do not make the connection with large subsampling rates. These Rényi DP-based analyses of the subsampled Gaussian mechanism do not provide tight privacy bounds anyway (Zhu et al., 2022), so these results do not imply that the linear relationship holds even asymptotically with tight accounting.

Our work focuses on Poisson subsampling, which is widely used and implemented in libraries like Opacus (Yousefpour et al., 2021). There are other subsampling schemes like subsampling with replacement (WR) and subsampling without replacement (WOR), which take a fixed-size uniformly random subsample, either with or without replacement. Previous work has done tight privacy accounting for both of them (Koskela et al., 2020; Zhu et al., 2022). However, recent work has questioned the validity of these works for WOR subsampling (Lebeda et al., 2023), pointing out issues which should be resolved before studying WOR subsampling further. WR subsampling is rarely used in practice, and the accounting for it is much more complex (Koskela et al., 2020) than Poisson subsampling.

Sommer et al. (2019) and Dong et al. (2022) prove central limit theorems for privacy accounting, which essentially show that the privacy loss of any sufficiently well-behaved mechanism after many compositions is asymptotically like the privacy loss of the Gaussian mechanism. As we study the asymptotic behaviour of the subsampled Gaussian mechanism after many compositions, it is possible that these theorems could be used to prove our result. However, we opted for another route in our proof. Instead of showing that privacy accounting for the subsampled Gaussian mechanism is asymptotically similar to accounting for the Gaussian mechanism, we show that the subsampled mechanism itself is asymptotically similar to the Gaussian mechanism.

2. Background

In this section, we go through some background material on differential privacy. We start with the definition and basics in Section 2.1, look at composition in Section 2.2 and finally introduce DP-SGD in Section 2.3.

2.1. Differential privacy

Differential privacy (DP) (Dwork et al., 2006b; Dwork & Roth, 2014) is a property of an algorithm that quantifies the privacy loss resulting from releasing the output of the algorithm. In the specific definition we use, the privacy loss is quantified by two numbers: $\epsilon \geq 0$ and $\delta \in [0, 1]$.

Definition 2.1. Let $\mathcal{M}$ be a randomised algorithm. $\mathcal{M}$ is (ϵ, δ)-DP if, for all measurable A and all neighbouring $x, x' \in \mathcal{X}$,

$$\Pr(\mathcal{M}(x) \in A) \leq e^{\epsilon} \Pr(\mathcal{M}(x') \in A) + \delta. \quad (1)$$

DP algorithms are also called *mechanisms*. We use $x \sim x'$ to denote x and x' being neighbouring. The meaning of neighbouring is domain specific. The most common definitions are *add/remove* neighbourhood, where $x \sim x'$ if they

differ in only adding or removing one row, and *substitute* neighbourhood, where $x \sim x'$ if they differ in at most one row, and have the same number of rows. We focus on the add/remove neighbourhood in this work.

Differential privacy is immune to post-processing, so applying any function to the output of a DP mechanism cannot weaken the privacy guarantee (Dwork & Roth, 2014).

Theorem 2.2. *Let $\mathcal{M}$ be an (ϵ, δ)-DP mechanism and let f be any randomised algorithm. Then $f \circ \mathcal{M}$ is (ϵ, δ)-DP.*

The basic mechanism we look at is the Gaussian mechanism, which adds Gaussian noise to the output of a function (Dwork et al., 2006a):

Definition 2.3. The Gaussian mechanism G for a function $f\colon \mathcal{X} \to \mathbb{R}^d$ and noise variance σ^2 is $G(x) = f(x) + \eta$, where $\eta \sim \mathcal{N}_d(0, \sigma^2 I_d)$.

The privacy bounds of the Gaussian mechanism depend on the sensitivity of f.

Definition 2.4. The sensitivity of a function f is $\Delta = \sup_{x \sim x'} ||f(x) - f(x')||_2$.

Tight privacy bounds for the Gaussian mechanism were derived by Balle & Wang (2018). We omit the exact expression here, as it is not important for this work.

2.2. Composition of Differential Privacy

Another useful property of DP is composition, which means that running multiple DP algorithms in succession degrades the privacy bounds in a predictable way. The tight composition theorem we are interested in is most easily expressed through privacy loss random variables (Sommer et al., 2019) and dominating pairs (Zhu et al., 2022), which we introduce next.

Definition 2.5. For $\alpha \geq 0$ and random variables P, Q, the hockey-stick divergence is

$$H_\alpha(P, Q) = \mathbb{E}_{t \sim Q}\left(\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(t) - \alpha\right)_+\right), \tag{2}$$

where $(x)_+ = \max\{x, 0\}$ and $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative, which simplifies to the density ratio if P and Q are continuous.

Differential privacy can be characterised with the hockey-stick divergence: $\mathcal{M}$ is (ϵ, δ)-DP if and only if $\sup_{x \sim x'} H_{e^\epsilon}(\mathcal{M}(x), \mathcal{M}(x')) \leq \delta$ (Zhu et al., 2022).

Definition 2.6 (Zhu et al. 2022)**.** A pair of random variables P, Q is called a dominating pair for mechanism $\mathcal{M}$ if for all $\alpha \geq 0$,

$$\sup_{x \sim x'} H_\alpha(\mathcal{M}(x), \mathcal{M}(x')) \leq H_\alpha(P, Q). \tag{3}$$

Dominating pairs allow computing privacy bounds for a mechanism from simpler distributions than the mechanism itself. A dominating pair for the Gaussian mechanism is $P = \mathcal{N}(0, \sigma^2)$, $Q = \mathcal{N}(\Delta, \sigma^2)$ (Koskela et al., 2020).

Definition 2.7 (Sommer et al. 2019)**.** For random variables P, Q, the privacy loss function is $\mathcal{L}(t) = \ln \frac{\mathrm{d}P}{\mathrm{d}Q}(t)$ and the privacy loss random variable (PLRV) is $L = \mathcal{L}(T)$ where $T \sim P$.

The PLRV of a mechanism is formed by setting P and Q to be a dominating pair for the mechanism. For the Gaussian mechanism, the PLRV is $\mathcal{N}(\mu, 2\mu)$ with $\mu = \frac{\Delta^2}{2\sigma^2}$ (Sommer et al., 2019).

The PLRV allows computing the privacy bounds of the mechanism with a simple expectation (Sommer et al., 2019):

$$\delta(\epsilon) = \mathbb{E}_{s \sim L}((1 - e^{\epsilon - s})_+). \tag{4}$$

PLRVs also allow expressing the tight composition theorem in a simple way.

Theorem 2.8 (Sommer et al. 2019)**.** *If $L_1, \ldots, L_T$ are the PLRVs for mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$, the PRLV of the adaptive composition of $\mathcal{M}_1, \ldots, \mathcal{M}_T$ is the convolution of $L_1, \ldots, L_T$, which is the distribution of $L_1 + \cdots + L_T$.*

2.3. Differentially private SGD

Differentially private SGD (DP-SGD) (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016) is one of the most important DP algorithms. Making SGD private requires *clipping* the gradients to bound their sensitivity to a threshold C: $\mathrm{clip}_C(x) = \frac{x}{||x||_2} \min\{||x||_2, C\}$, $\mathrm{clip}_C(0) = 0$. Then noise is added to the sum of the clipped gradients with the Gaussian mechanism, so DP-SGD uses

$$G_{DP} = \sum_{i \in \mathcal{B}} \mathrm{clip}_C(g_i) + \mathcal{N}(0, \sigma^2 I_d) \tag{5}$$

in place of the non-private sum of gradients. G_{DP} can also be used in adaptive versions of SGD like Adam.

To compute the privacy bounds for DP-SGD, we need to account for the *subsampling amplification* that comes from the subsampling in SGD. This requires fixing the subsampling scheme. We consider Poisson subsampling, where each datapoint in $\mathcal{B}$ is included in the subsample with probability q, independently of any other datapoints.

When the neighbourhood relation is add/remove, the Poisson subsampled Gaussian mechanism, of which DP-SGD is an instance of, has the dominating pair (Koskela et al., 2020)

$$\begin{aligned} P &= q\mathcal{N}(\Delta, \sigma^2) + (1 - q)\mathcal{N}(0, \sigma^2), \\ Q &= \mathcal{N}(0, \sigma^2). \end{aligned} \tag{6}$$

With this dominating pair, we can form the PLRV for the mechanism, take a T-fold convolution for T iterations, and compute the privacy bounds from the expectation in (4). The computation is not trivial, but can be done using numerical privacy accountants (Koskela et al., 2020; 2021; Gopi et al., 2021; Doroshenko et al., 2022; Alghamdi et al., 2023). These accountants are used by libraries implementing DP-SGD like Opacus (Yousefpour et al., 2021) to compute privacy bounds in practice.

2.4. Kullback-Leibler Divergence and Total Variation Distance

Our proofs use two notions of distance between random variables, *total variation distance* and *Kullback-Leibler (KL) divergence* (Kullback & Leibler, 1951).

Definition 2.9. Let P and Q be random variables.

1. The total variation distance between P and Q is
$$\mathrm{TV}(P, Q) = \sup_A |\Pr(P \in A) - \Pr(Q \in A)|. \quad (7)$$
The supremum is over all measurable sets A.
2. The KL divergence between P and Q with densities $p(t)$ and $q(t)$ is
$$\mathrm{KL}(P, Q) = \mathbb{E}_{t\sim P}\left(\ln \frac{p(t)}{q(t)}\right). \quad (8)$$

The two notions of distance are related by Pinsker's inequality.

Lemma 2.10 (Kelbert, 2023)**.** *For random variables P, Q,*
$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)}. \quad (9)$$

3. Accounting Oracles

We use the concept of *accounting oracles* (Tajeddine et al., 2020) to make formalising properties of privacy accounting easier. The accounting oracle is the ideal privacy accountant that numerical accountants aim to approximate.

Definition 3.1. For a mechanism $\mathcal{M}$, the accounting oracle $\mathrm{AO}_{\mathcal{M}}(\epsilon)$ returns the smallest δ such that $\mathcal{M}$ is (ϵ, δ)-DP.

In case $\mathcal{M}$ has hyperparameters that affect its privacy bounds, these hyperparameters will also be arguments of $\mathrm{AO}_{\mathcal{M}}$. We write the accounting oracle for the T-fold composition of the Poisson subsampled Gaussian mechanism as $\mathrm{AO}_S(\sigma, \Delta, q, T, \epsilon)$, and the accounting oracle of a composition of the Gaussian mechanism as $\mathrm{AO}_G(\sigma, \Delta, T, \epsilon)$.

Accounting oracles make it easy to formally express symmetries of privacy accounting. For example, privacy bounds are invariant to post-processing with a bijection.

Lemma 3.2. *Let $\mathcal{M}$ be a mechanism and f be a bijection. Then $\mathrm{AO}_{\mathcal{M}}(\epsilon) = \mathrm{AO}_{f\circ\mathcal{M}}(\epsilon)$.*

Proof. This follows by using post-processing immunity for f to show that $\mathrm{AO}_{f\circ\mathcal{M}}(\epsilon) \leq \mathrm{AO}_{\mathcal{M}}(\epsilon)$ and for f^{-1} to show that $\mathrm{AO}_{f\circ\mathcal{M}}(\epsilon) \geq \mathrm{AO}_{\mathcal{M}}(\epsilon)$. □

We can also formalise the lemma that considering $\Delta = 1$ is sufficient when analysing the (subsampled) Gaussian mechanism.

Lemma 3.3.
$$\mathrm{AO}_G(\sigma, \Delta, T, \epsilon) = \mathrm{AO}_G(\sigma/\Delta, 1, T, \epsilon), \quad (10)$$
$$\mathrm{AO}_S(\sigma, \Delta, q, T, \epsilon) = \mathrm{AO}_S(\sigma/\Delta, 1, q, T, \epsilon). \quad (11)$$

Proof. Let $\mathcal{M}$ be the (subsampled) Gaussian mechanism. Then
$$\mathcal{M}(x) = f(x) + \eta = \Delta\left(\frac{1}{\Delta}f(x) + \frac{1}{\Delta}\eta\right) \quad (12)$$
with $\eta \sim \mathcal{N}_d(0, \sigma^2 I_d)$. The sensitivity of $\frac{1}{\Delta}f(x)$ is 1, so the part inside parenthesis in the last expression is a (subsampled) Gaussian mechanism with sensitivity 1 and noise standard deviation σ/Δ. Multiplying by Δ is bijective post-processing, so that mechanism must have the same privacy bounds as the original mechanism. In a composition, this transformation can be done separately for each individual mechanism of the composition. □

Since considering $\Delta = 1$ when analysing the subsampled Gaussian mechanism is enough by Lemma 3.3, we occasionally shorten $\mathrm{AO}_S(\sigma, 1, q, T, \epsilon)$ to $\mathrm{AO}_S(\sigma, q, T, \epsilon)$.

The next lemma and corollary show that mechanisms close in total variation distance also have similar privacy bounds.

Lemma 3.4. *Let $\mathcal{M}$ be an (ϵ, δ)-DP mechanism, and let $\mathcal{M}'$ be a mechanism with*
$$\sup_x \mathrm{TV}(\mathcal{M}(x), \mathcal{M}'(x)) \leq d, \quad (13)$$
for some $d \geq 0$. Then $\mathcal{M}'$ is $(\epsilon, \delta + (1 + e^\epsilon)d)$-DP.

Proof. For any measurable set A,
$$\begin{aligned}\Pr(\mathcal{M}'(x) \in A) &\leq \Pr(\mathcal{M}(x) \in A) + d \\ &\leq e^\epsilon \Pr(\mathcal{M}(x') \in A) + \delta + d \\ &\leq e^\epsilon(\Pr(\mathcal{M}'(x') \in A) + d) + \delta + d \\ &= e^\epsilon \Pr(\mathcal{M}'(x') \in A) + \delta + (1 + e^\epsilon)d.\end{aligned} \quad (14)$$
□

Lemma 3.4 can also be expressed with accounting oracles.

Corollary 3.5. *Let $\mathrm{AO}_{\mathcal{M}}$ and $\mathrm{AO}_{\mathcal{M}'}$ be the accounting oracles for mechanisms $\mathcal{M}$ and $\mathcal{M}'$, respectively. If*

$$\sup_x \mathrm{TV}(\mathcal{M}(x), \mathcal{M}'(x)) \leq d. \tag{15}$$

then

$$|\mathrm{AO}_{\mathcal{M}}(\epsilon) - \mathrm{AO}_{\mathcal{M}'}(\epsilon)| \leq (1 + e^{\epsilon})d. \tag{16}$$

Proof. Let $\delta = \mathrm{AO}_{\mathcal{M}}(\epsilon)$. By Lemma 3.4, $\mathcal{M}'$ is $(\epsilon, (1 + e^{\epsilon})d + \delta)$-DP, so

$$\mathrm{AO}_{\mathcal{M}'}(\epsilon) \leq \delta + (1 + e^{\epsilon})d. \tag{17}$$

If $\mathrm{AO}_{\mathcal{M}'}(\epsilon) < \delta - (1 + e^{\epsilon})d$,

$$\mathrm{AO}_{\mathcal{M}}(\epsilon) < \delta - (1 + e^{\epsilon})d + (1 + e^{\epsilon})d = \delta \tag{18}$$

by Lemma 3.4, which is a contradiction, so

$$\mathrm{AO}_{\mathcal{M}'}(\epsilon) \geq \delta - (1 + e^{\epsilon})d. \tag{19}$$

Combining (17) and (19),

$$|\mathrm{AO}_{\mathcal{M}}(\epsilon) - \mathrm{AO}_{\mathcal{M}'}(\epsilon)| \leq (1 + e^{\epsilon})d. \tag{20}$$

□

4. DP-SGD noise decomposition

For now, let us denote the sum on the right in Equation (5), the sum of clipped gradients with subsampling rate q, with G_q:

$$G_q = \sum_{i \in \mathcal{B}} \mathrm{clip}_C(g_i). \tag{21}$$

In each step of DP-SGD, we are releasing this sum using Gaussian perturbation. However, due to the subsampling, our gradient sum can comprise of any number of terms between 0 and N. Therefore, before we do the step, we want to upscale the summed gradient to approximate the magnitude of the full data gradient G_1. This is typically done by scaling the G_q with $N/|B|$ where $|B|$ denotes the size of the minibatch. However, using the $|B|$ as such might leak private information. Instead, we will use $1/q$ scaling for G_q, which also gives an unbiased estimator of G_1:

$$\mathbb{E}\left[\frac{1}{q}G_q - G_1\right] = \mathbb{E}\left[\frac{1}{q}\sum_{i \in [N]} b_i \tilde{g}_i - \sum_{i \in [N]} \tilde{g}_i\right] \tag{22}$$

$$= \sum_{i \in [N]} \tilde{g}_i \, \mathbb{E}\left[\frac{b_i}{q} - 1\right] = 0, \tag{23}$$

where $b_i \sim \mathrm{Bernoulli}(q)$ and $\tilde{g}_i$ denote the clipped per-example gradients. Using Lemma 3.3, we can decouple the noise-scale and the clipping threshold C and write the DP gradient $\tilde{G}$ used to update the parameters as

$$\tilde{G} = \frac{1}{q}\left(G_q + C\sigma\eta\right), \tag{24}$$

where $\eta \sim \mathcal{N}(0, I_d)$. The clipping threshold C affects the update as a constant scale independent of q, and therefore for notational simplicity we set $C = 1$. Since the subsampled gradient G_q and the DP noise are independent, we can decompose the total gradient variance of the jth element of $\tilde{G}$ as

$$\underbrace{\mathrm{Var}(\tilde{G}_j)}_{\text{Total}} = \underbrace{\frac{1}{q^2}\mathrm{Var}(G_{q,j})}_{\text{Subsampling}} + \underbrace{\left(\frac{\sigma}{q}\right)^2}_{\text{Effective DP Noise}}. \tag{25}$$

The first component on the right in Equation (25), the subsampling variance, arises from the subsampling, so it is easy to see, as we show in Appendix A.1, that it is decreasing w.r.t. q, which holds regardless of the clipping threshold assuming $C > 0$. For the rest of the paper, we will use $\sigma_{\text{eff}}^2 := \sigma^2/q^2$ to denote the second component of the variance decomposition in Equation (25), the effective noise variance.

In order to guarantee (ϵ, δ)-DP, the σ term in σ_{eff} needs to be adjusted based on the number of iterations T and the subsampling rate q. Hence we will treat the σ as a function of q and T, and denote the smallest σ that provides (ϵ, δ)-DP as $\sigma(q, T)$.

Now, the interesting question is, at what rate does the $\sigma(q, T)$ grow w.r.t. q? In Equation (25), we saw that the effective variance arising from the DP noise addition scales as $(\sigma(q, T)/q)^2$. In Section 5, we show that as $T \to \infty$, $\sigma(q, T)$ becomes a linear function w.r.t. q. Since $\sigma(1, T) = 1 \cdot \sigma(1, T)$, this means that

$$\sigma(q, T) \approx q\sigma(1, T). \tag{26}$$

This result implies that subsampling does not reduce the effective level of injected DP noise, while we would have more subsampling-induced noise arising from the first term of Equation (25) for small values of q.

In Section 5.1 we study σ_{eff} numerically, and find that as T grows, the σ_{eff} decreases monotonically towards $\sigma(1, T)$.

Furthermore, in Section 6 we show that when $T = 1$, the subsampling can in fact be harmful and smaller values q incur a proportionally larger effective noise variance in the gradient updates.

5. Subsampled Gaussian Mechanism in the Limit of Many Compositions

Since privacy accounting for DP-SGD is based on the accounting oracle $\mathrm{AO}_S(\sigma, \Delta, q, T, \epsilon)$ for the Poisson subsam-

pled Gaussian mechanism, it suffices to study the behaviour of AO_S, instead of DP-SGD directly. This can be done in a much simpler setting.

In the setting we use, there is a single datapoint $x \in \{0, 1\}$, which is released T times with the Poisson subsampled Gaussian mechanism:

$$\mathcal{M}_i(x) \sim xB_q + \mathcal{N}(0, \sigma_T^2) \tag{27}$$

for $1 \leq t \leq T$, where B_q is a Bernoulli random variable. Since $\mathcal{M}$ is a Poisson subsampled Gaussian mechanism, its accounting oracle is $\mathrm{AO}_S(\sigma, 1, q, T, \epsilon)$.[1] We have

$$\mathbb{E}(\mathcal{M}_i(x)) = qx \tag{28}$$

$$\mathrm{Var}(\mathcal{M}_i(x)) = x^2 q(1-q) + \sigma_T^2. \tag{29}$$

As $\sigma_T^2 \to \infty$, the variance and σ_T^2 approach each other. As a result, we can approximate $\mathcal{M}_i$ with

$$\mathcal{M}_i'(x) \sim \mathcal{N}(qx, \sigma_T^2). \tag{30}$$

To prove this, we first need to find a lower bound on how quickly σ_T^2 must grow as $T \to \infty$.

Theorem 5.1. *Let σ_T be such that $\mathrm{AO}_S(\sigma_T, \Delta, q, T, \epsilon) \leq \delta$ for all T, with $\delta < 1$. Then $\sigma_T^2 = \Omega(T)$.*

Proof. The idea of the proof is to show that unless $\sigma_T^2 = \Omega(T)$, the mean of the PLRV can obtain arbitrarily large values, and then show that this leads to violating the δ bound. We defer the full proof to Appendix A.5. □

Now we can prove that the approximation is sound.

Theorem 5.2. *For $1 \leq i \leq T$, let*

$$\mathcal{M}_i(x) \sim xB_q + \mathcal{N}(0, \sigma_T^2), \tag{31}$$

$$\mathcal{M}_i'(x) \sim \mathcal{N}(qx, \sigma_T^2). \tag{32}$$

be independent for each i. Let $\mathcal{M}_{1:T}$ be the composition of $\mathcal{M}_i$, and let $\mathcal{M}_{1:T}'$ be the composition of $\mathcal{M}_i'$. Then

$$\sup_x \mathrm{TV}(\mathcal{M}_{1:T}(x), \mathcal{M}_{1:T}'(x)) \to 0 \tag{33}$$

as $T \to \infty$.

Proof. It suffices to show

$$\sup_x T \cdot \mathrm{KL}(\mathcal{M}_i'(x), \mathcal{M}_i(x)) \to 0 \tag{34}$$

due to Pinsker's inequality and the additivity of KL divergence for products of independent random variables. When $x = 0$, the two mechanism are the same, so it suffices to look at $x = 1$. The idea for the rest of the proof is to first reparameterise σ in terms of a variable u with $u \to 0$ as $T \to \infty$ and then use Taylor's theorem at $u = 0$ to find the convergence rate of the KL divergence. For the complete proof see Appendix A.3. □

The mechanism $\mathcal{M}_i'$ is nearly a Gaussian mechanism, since

$$\frac{1}{q}\mathcal{M}_i'(x) \sim x + \mathcal{N}\left(0, \frac{\sigma_T^2}{q^2}\right). \tag{35}$$

On the right is a Gaussian mechanism with noise standard deviation $\frac{\sigma_T}{q}$, which has the linear relationship between σ_T and q we are looking for. In the next theorem, we use this to prove our main result.

Theorem 5.3. *For any σ, q_1, q_2, Δ and ϵ*

$$|\mathrm{AO}_S(\sigma, \Delta, q_1, T, \epsilon) - \mathrm{AO}_S(\sigma \cdot q_2/q_1, \Delta, q_2, T, \epsilon)| \to 0 \tag{36}$$

as $T \to \infty$.

Proof. It suffices to look at $\Delta = 1$ by Lemma 3.3. Let $\mathrm{AO}'(\sigma, q, T, \epsilon)$ be the accounting oracle for $\mathcal{M}_{1:T}'$. By Lemma 3.2 and (35),

$$\mathrm{AO}'(\sigma, q, T, \epsilon) = \mathrm{AO}_G\left(\frac{\sigma}{q}, T, \epsilon\right). \tag{37}$$

Let $\sigma_2 = \sigma \cdot q_2/q_1$,

$$d_T^{(1)} = \sup_x \mathrm{TV}(\mathcal{M}_{1:T}(x, q_1, \sigma), \mathcal{M}_{1:T}'(x, q_1, \sigma)), \tag{38}$$

and

$$d_T^{(2)} = \sup_x \mathrm{TV}(\mathcal{M}_{1:T}(x, q_2, \sigma_2), \mathcal{M}_{1:T}'(x, q_2, \sigma_2)). \tag{39}$$

Now

$$|\mathrm{AO}_S(\sigma, q_1, T, \epsilon) - \mathrm{AO}'(\sigma, q_1, T, \epsilon)| \leq (1 + e^\epsilon) d_T^{(1)} \tag{40}$$

and

$$|\mathrm{AO}_S(\sigma_2, q_2, T, \epsilon) - \mathrm{AO}'(\sigma_2, q_2, T, \epsilon)| \leq (1 + e^\epsilon) d_T^{(2)} \tag{41}$$

by Corollary 3.5. Since $\frac{\sigma}{q_1} = \frac{\sigma_2}{q_2}$,

$$\begin{aligned} \mathrm{AO}'(\sigma, q_1, T, \epsilon) &= \mathrm{AO}_G\left(\frac{\sigma}{q_1}, T, \epsilon\right) \\ &= \mathrm{AO}_G\left(\frac{\sigma_2}{q_2}, T, \epsilon\right) \\ &= \mathrm{AO}'(\sigma_2, q_2, T, \epsilon). \end{aligned} \tag{42}$$

Now

$$\begin{aligned} &|\mathrm{AO}_S(\sigma, q_1, T, \epsilon) - \mathrm{AO}_S(\sigma \cdot q_2/q_1, q_2, T, \epsilon)| \\ &\leq |\mathrm{AO}_S(\sigma, q_1, T, \epsilon) - \mathrm{AO}'(\sigma, q_1, T, \epsilon)| \\ &\quad + |\mathrm{AO}_S(\sigma_2, q_2, T, \epsilon) - \mathrm{AO}'(\sigma_2, q_2, T, \epsilon)| \\ &\leq (1 + e^\epsilon)(d_T^{(1)} + d_T^{(2)}) \to 0 \end{aligned} \tag{43}$$

by Theorem 5.2. □

[1] Considering $\Delta = 1$ suffices due to Lemma 3.3.

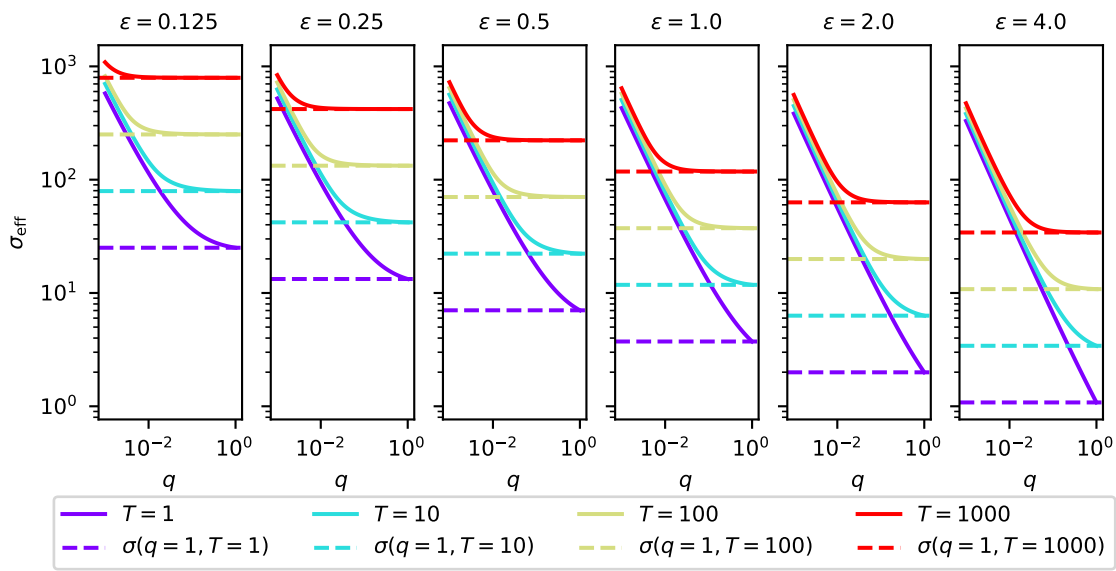

Figure 1. The $\sigma_{\text{eff}} := \sigma(q,T)/q$ decreases as q grows for all the ϵ and T values. As T grows, σ_{eff} approaches the $\sigma(1,T)$, as the asymptotic theory predicts. The privacy parameter δ was set to 10^{-5} when computing the $\sigma(q,T)$.

By setting $q_2 = 1$, we see that $\text{AO}_S(\sigma, \Delta, q, T, \epsilon)$ behaves like $\text{AO}_S(\sigma/q, \Delta, 1, T, \epsilon)$ in the $T \to \infty$ limit, so $\frac{\sigma}{q}$ must be a constant that does not depend on q to reach a given (ϵ, δ)-bound. We formalise this in the following corollary.

Corollary 5.4. *Let $\sigma(q,T)$ be the smallest σ such that $\text{AO}_S(\sigma, \Delta, q, T, \epsilon) \leq \delta$. Then*

$$\lim_{T\to\infty} \frac{\sigma(q,T)}{q\sigma(1,T)} = 1. \tag{44}$$

Proof. The claim follows from Theorem 5.3 and the continuity of the inverse of $\text{AO}_S(\sigma, \Delta, 1, T, \epsilon) = \text{AO}_G(\sigma, \Delta, T, \epsilon)$ when considered only a function of σ, with other parameters fixed. We defer the full proof to Appendix A.4. □

5.1. Empirical results

In order to demonstrate how the σ_{eff} converges to $q\sigma(1,T)$, we use Google's open source implementation[2] of the PLD accountant (Doroshenko et al., 2022) to compute the $\sigma(q,T)$.

Figure 1 shows that as the the number of iterations grows, the σ_{eff} approaches the $\sigma(1,T)$ line. We can also see that for smaller values of ϵ, the convergence happens faster. This behaviour can be explained by the larger level of noise needed for smaller ϵ values, which will make the components of the Gaussian mixture in Equation (31) less distinguishable from each other.

[2]Available at https://github.com/google/differential-privacy/tree/main/python

We can also see that for all of the settings, the σ_{eff} stays above the $\sigma(1,T)$ line, and it is the furthest away when q is the smallest. This would suggest, that our observation in Section 6 hold also for $T > 1$, and that smaller values of q incur a disproportionally large DP-induced variance component in Equation (25).

The source code for reproducing our experiments can be found in https://github.com/DPBayes/subsampling-is-not-magic.

6. Subsampled Gaussian Mechanism without Compositions

For now, let us consider $T = 1$, and denote $\sigma(q) := \sigma(q, 1)$. We can express the δ using the hockey-stick divergence and the dominating pair (P, Q) defined in Equation (6) as

$$\begin{aligned}\delta(q) =& q \Pr\left(Z \geq \sigma(q) \log\left(\frac{h(q)}{q}\right) - \frac{1}{2\sigma(q)}\right) \\ &- h(q) \Pr\left(Z \geq \sigma(q) \log\left(\frac{h(q)}{q}\right) + \frac{1}{2\sigma(q)}\right),\end{aligned} \tag{45}$$

where $h(q) := e^{\epsilon} - (1 - q)$. Recall that we have assumed $\sigma(q,T)$ to be a function that returns a noise-level matching to a particular (ϵ, δ) privacy level for given q and T. Therefore, the derivative of $\delta(q)$ in Equation (45) w.r.t. q is 0, and we can solve the derivative of the RHS for $\sigma'(q)$. Let us denote

$$a := \frac{1}{2\sqrt{2}\sigma(q)} \tag{46}$$

$$b := \frac{\sigma(q)}{\sqrt{2}} \log\left(\frac{e^{\epsilon} - (1-q)}{q}\right). \tag{47}$$

We have the following Lemma

Lemma 6.1. *For the smallest $\sigma(q)$ that provides (ϵ, δ)-DP for the Poisson subsampled Gaussian mechanism, we have*

$$\sigma'(q) = \frac{\sigma(q)}{q} \frac{1}{2a} \frac{1}{\mathrm{erf}'(a-b)} (\mathrm{erf}(a-b) - \mathrm{erf}(-a-b)). \tag{48}$$

Proof. See Appendix A.6. □

Now, Lemma 6.1 allows us to establish following result.

Theorem 6.2. *If $a < b$ for a and b defined in Equations (46) and (47), then $\frac{\mathrm{d}}{\mathrm{d}q}\frac{\sigma(q)}{q} < 0$.*

Proof. The $\mathrm{erf}(x)$ is a convex function for $x \in \mathbb{R}_{<0}$. Since $a, b \geq 0$ we have $-a - b \leq 0$ and if $a - b < 0$ we get from the convexity that

$$\mathrm{erf}(a-b) - \mathrm{erf}(-a-b) < 2a\,\mathrm{erf}'(a-b). \tag{49}$$

Substituting this upper bound into Equation (48) gives

$$\sigma'(q) < \frac{\sigma(q)}{q} \Leftrightarrow \frac{q\sigma'(q) - \sigma(q)}{q^2} = \frac{\mathrm{d}}{\mathrm{d}q}\frac{\sigma(q)}{q} < 0. \tag{50}$$

□

Now, Theorem 6.2 implies that σ_{eff} is a decreasing function w.r.t. q, and therefore larger subsampling rates should be preferred when $a < b$. So now the remaining question is, when is a smaller than b. It is easy to see that if $a > b$, we have an upper bound on the $\sigma(q)$, and therefore we cannot obtain arbitrarily strict privacy levels.

However, analytically solving the region where $a < b$ is intractable as we do not have a closed form expression for $\sigma(q)$. Therefore we make the following Conjecture, which we study numerically.

Conjecture 6.3. *For $\epsilon, q \geq 4\delta$, we have $a - b < 0$.*

Verifying the Conjecture 6.3 numerically requires computing the $\sigma(q)$ values for a range of q and ϵ values, which would be computationally inefficient. However, computing a, b and δ can be easily parallelized over multiple values of q and σ. We set $\delta_{\text{target}} = 10^{-5}$ and compute the a, b and δ for $q \in [4\delta_{\text{target}}, 1.0]$, $\sigma \in [\min(q)\sigma(1,\epsilon), \sigma(1,\epsilon)]$ and $\epsilon \in [4\delta_{\text{target}}, 4.0]$ which would be a reasonable range of ϵ values for practical use. The $\sigma(1,\epsilon)$ is a noise-level matching $(\epsilon, \delta_{\text{target}})$-DP guarantee for $q = 1$, and the lower bound $\min(q)\sigma(1,\epsilon)$ for σ was selected based on hypothesis that within $[\min(q)\sigma(1,\epsilon), \sigma(1,\epsilon)]$ we can find a noise-level closely matching the δ_{target}.

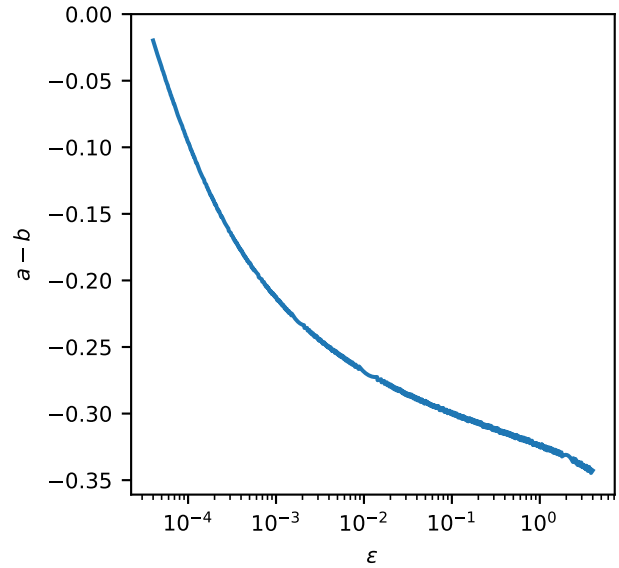

Figure 2. The largest $a - b$ computed for multiple (q, σ) pairs stays negative for a broad range of ϵ values. The a and b were selected so that the corresponding (q, σ) pair satisfies (ϵ, δ)-DP with $\delta \leq 10^{-5}$.

Figure 2 shows the largest $a - b$ value for our target ϵ values among the q and σ pairs that resulted into $\delta \leq \delta_{\text{target}}$. We can see that among these values there are no cases of $a > b$. While our evaluation covers a range of ϵ values, $a - b$ seems to be monotonically decreasing w.r.t. ϵ, which suggests that the conjecture should hold even for larger values of ϵ. Based on our numerical evaluation, the σ range resulted into δ values differing from $\delta_{\text{target}} = 10^{-5}$ at most $\approx 2 \times 10^{-7}$ in absolute difference. As a final remark, the constant 4 in the Conjecture 6.3 was found numerically. With smaller values of this constant we obtained cases for which $a - b > 0$. Furthermore, values of $q \approx \delta$ empirically produce results that even fail to satisfy the claim of Theorem 6.2.

Limitations Our analysis for the $T = 1$ case reduced the monotonicity of effective DP noise standard deviation σ_{eff} to a sufficient condition $a < b$. We were unable to provide a formal proof of when this condition is satisfied, but verified numerically that it appears valid for a broad range of practically relevant parameters with $\delta = 10^{-5}$, a standard value suggested by NIST[3].

[3] https://doi.org/10.6028/NIST.SP.800-226.ipd

7. Discussion

In this paper, we have studied how the subsampling rate affects the level of Gaussian perturbation in Poisson subsampled DP-SGD. An important question is how this affects the convergence of DP-SGD. Recently, Bu et al. (2023) studied the convergence of DP-SGD using gradient normalisation instead of clipping. In their work, they proved a convergence result for DP-SGD with expected batch size B, number of iterations T and Gaussian noise with standard deviation of σ. Their Theorem 4 gives an upper bound on the expected gradient norm as

$$\min_{0\leq t\leq T} \mathbb{E}(||g_t||) \leq \mathcal{G}\left(\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0 - \mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}; \xi, \gamma\right)$$

where the $\mathcal{L}_0, \mathcal{L}*$ and L are regularity condition constants for the loss function, γ is a stability parameter for the gradient normalisation, and ξ^2 is the sum of the total gradient variances for each dimension from (25). $\mathcal{G}$ is a complicated function that arises from the gradient normalisation.

The bound is affected by the noise-level and the batch size in terms of the factor $\frac{\sigma^2}{B^2}$. This term is linearly proportional to the $\frac{\sigma(q)^2}{q^2}$ which we study in this paper. Thus, given the same number of iterations, our results that $\frac{\sigma(q)}{q}$ is non-increasing w.r.t. q would also imply that the $\frac{\sigma^2}{B^2}$ term in the above bound is also non-increasing w.r.t q. The function $\mathcal{G}$ in this Theorem is increasing with regards to the first parameter, which is the only place the number of iterations T appears. This means that the batch size does not affect the convergence rate in the asymptotic regime, and a large batch size improves the constant for the convergence rate in the non-asymptotic regime.

The effect of the batch size through ξ is harder to reason about due to the complicated nature of $\mathcal{G}$. $\mathcal{G}$ appears to be increasing with regards to ξ based on numerical results of Bu et al. (2023), which would mean that a larger batch size also improves the convergence through reducing ξ. When $\gamma = 0$, Bu et al. (2023) show that $\mathcal{G} \geq \xi$, so a larger batch size improves the quality of the solution DP-SGD can find in this case.

8. Conclusion

We studied the relationship between the effective noise variance and the subsampling rate in Poisson subsampled DP-SGD, and proved that as the number of iterations approaches infinity, the relationship becomes linear, which cancels the effect of the subsampling rate in the effective noise variance. This means that a large subsampling rate always reduces the effective total gradient variance. Furthermore, we demonstrated that under a wide range of ϵ values, a single application of the Poisson subsampled Gaussian mechanism actually incurs a monotonically decreasing effective noise variance w.r.t. subsampling rate. Our numerical experiments show that the asymptotic regime is relevant in practice, and that outside the asymptotic regime, smaller subsampling rates lead to increasingly large effective total gradient variances. This explains the observed benefits of large batch sizes in DP-SGD, which has so far had only empirical and heuristic explanations, furthering the theoretical understanding of DP-SGD.

For future work, it would be important to theoretically study how to interpolate our results between the $T = 1$ and the asymptotic case. Based on our numerical evaluations however, we expect our main conclusion, that the large batches provide smaller effective noise, to hold even for finite $T > 1$.

In all cases we have studied, less subsampling (larger q) always leads to better privacy–utility trade-off, at the cost of more compute. Thus the magic of subsampling amplification lies in saving compute, not in achieving higher utility than without subsampling.

Acknowledgements

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI as well as Grants 356499 and 359111), the Strategic Research Council at the Research Council of Finland (Grant 358247) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Impact Statement

This paper presents work whose goal is to advance the field of trustworthy machine learning. Our results improve the understanding of DP-SGD and help designing better DP algorithms. As such, we believe work like ours has an overall positive effect on machine learning.

References

Abadi, M., Chu, A., Goodfellow, I. J., 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.

Alghamdi, W., Gomez, J. F., Asoodeh, S., Calmon, F., Kosut, O., and Sankar, L. The Saddle-Point Method in Differential Privacy. In *International Conference on Machine*

Learning, 2023.

Balle, B. and Wang, Y.-X. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th annual symposium on foundations of computer science*, FOCS '14, pp. 464–473. IEEE Computer Society, 2014.

Beimel, A., Brenner, H., Kasiviswanathan, S. P., and Nissim, K. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014.

Bu, Z., Wang, Y., Zha, S., and Karypis, G. Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems*, volume 34, 2023.

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. Composable and versatile privacy via truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pp. 74–86. ACM, 2018.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. arXiv: 2204.13650, 2022. https://arxiv.org/abs/2204.13650.

Dong, J., Roth, A., and Su, W. J. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 02 2022.

Doroshenko, V., Ghazi, B., Kamath, P., Kumar, R., and Manurangsi, P. Connect the dots: Tighter discrete approximations of privacy loss distributions. In *Proceedings on Privacy Enhancing Technologies*, volume 4, pp. 552–570, 2022.

Dwork, C. and Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Advances in Cryptology - EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating Noise to Sensitivity in Private Data Analysis. In *Third Theory of Cryptography Conference*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006b.

Gopi, S., Lee, Y. T., and Wutschitz, L. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48, pp. 1225–1234, 2016.

Kelbert, M. Survey of Distances between the Most Popular Distributions. *Analytics*, 2(1):225–245, 2023.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

Kleinberg, B., Li, Y., and Yuan, Y. An Alternative View: When Does SGD Escape Local Minima? In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Koskela, A., Jälkö, J., and Honkela, A. Computing tight differential privacy guarantees using FFT. In *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2560–2569. PMLR, 2020.

Koskela, A., Jälkö, J., Prediger, L., and Honkela, A. Tight Differential Privacy for Discrete-Valued Mechanisms and for the Subsampled Gaussian Mechanism Using FFT. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2021.

Kullback, S. and Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.

Lebeda, C. J., Regehr, M., and Kamath, G. Avoiding Pitfalls for Privacy Accounting of Subsampled Mechanisms under Composition. https://openreview.net/forum?id=fj5SqqXfn1, 2023.

Li, N., Qardaji, W. H., and Su, D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *7th ACM symposium on information, compuer and communications security*, pp. 32–33. ACM, 2012.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large Language Models Can Be Strong Differentially Private

Learners. In *International Conference on Learning Representations*, 2021.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International conference on learning representations*, 2018.

Mehta, H., Thakurta, A. G., Kurakin, A., and Cutkosky, A. Towards Large Scale Transfer Learning for Differentially Private Image Classification. *Transactions on Machine Learning Research*, 2023.

Mironov, I. Rényi Differential Privacy. In *30th IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017.

Mironov, I., Talwar, K., and Zhang, L. Rényi Differential Privacy of the Sampled Gaussian Mechanism. arXiv:1908.10530, 2019. https://arxiv.org/abs/1908.10530.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on learning theory, COLT 2018*, volume 75, pp. 605–638, 2018.

Sander, T., Stock, P., and Sablayrolles, A. TAN Without a Burn: Scaling Laws of DP-SGD. In *International Conference on Machine Learning*, 2023.

Sommer, D. M., Meiser, S., and Mohammadi, E. Privacy Loss Classes: The Central Limit Theorem in Differential Privacy. *PoPETs*, 2019(2):245–269, 2019.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Tajeddine, R., Jälkö, J., Kaski, S., and Honkela, A. Privacy-preserving Data Sharing on Vertically Partitioned Data. arXiv:2010.09293, 2020. https://arxiv.org/abs/2010.09293.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-Friendly Differential Privacy Library in PyTorch. arXiv:2109.12298, 2021. https://arxiv.org/abs/2109.12298v4.

Zhu, Y. and Wang, Y.-X. Poisson subsampled Rényi differential privacy. In *International Conference on Machine Learning*, pp. 7634–7642, 2019.

Zhu, Y., Dong, J., and Wang, Y.-X. Optimal accounting of differential privacy via characteristic function. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4782–4817. PMLR, 2022.

A. Missing proofs

A.1. The subsampling induced variance decreases w.r.t. q

Recall from Section 4, that we denote the sum of clipped subsampled gradients with G_q:

$$G_q = \sum_{i \in \mathcal{B}} \text{clip}_C(g_i). \tag{51}$$

Now for the subsampling induced variance in noise decomposition of Equation (25) we have

$$\begin{aligned}
\text{Var}\left(\frac{1}{q} G_{q,j}\right) &= \frac{1}{q^2} \text{Var}\left(\sum_{i \in [N]} b_i g_{i,j}\right) && (52)\\
&= \frac{1}{q^2} \sum_{i \in [N]} g_{i,j}^2 \, \text{Var}\left(b_i\right) && (53)\\
&= \frac{q(1-q)}{q^2} \sum_{i \in [N]} g_{i,j}^2 && (54)\\
&= \frac{1-q}{q} \sum_{i \in [N]} g_{i,j}^2. && (55)
\end{aligned}$$

Now, it is easy to see that the sum in Equation (55) is a constant w.r.t. q, and the term $(1-q)/q$ is decreasing w.r.t. q. Thus the subsampling induced variance is decreasing w.r.t. q.

A.2. Useful Lemmas

Lemma A.1 (Kullback & Leibler, 1951)**.** *Properties of KL divergence:*

1. *If P and Q are joint distributions over independent random variables $P_1, \ldots, P_T$ and $Q_1, \ldots, Q_T$,*

$$\text{KL}(P, Q) = \sum_{i=1}^{T} \text{KL}(P_i, Q_i). \tag{56}$$

2. *If f is a bijection,*

$$\text{KL}(f(P), f(Q)) = \text{KL}(P, Q). \tag{57}$$

We will need to analyse the following function:

$$f_x(u) = \ln \frac{\mathcal{N}(x; 0, 1)}{q\mathcal{N}(x; u - qu, 1) + (1-q)\mathcal{N}(x; -qu, 1)}. \tag{58}$$

In particular, we need the fourth-order Taylor approximation of $\mathbb{E}_x(f_x(u))$ for $x \sim \mathcal{N}(0, 1)$ at $u = 0$. We begin by looking at the Taylor approximation of $f_x(u)$ without the expectation, and then show that we can differentiate under the expectation.

Lemma A.2. *The fourth-order Taylor approximation of f_x at $u = 0$ is*

$$\begin{aligned}
f_x(u) = &\frac{1}{2}(q-1)q(x^2-1)u^2 \\
&- \frac{1}{6}(q-1)q(2q-1)x(x^2-3)u^3 \\
&+ \frac{1}{24} q \left(-3 + 6x^2 - x^4 - 12q^2(2 - 4x^2 + x^4) + 6q^3(2 - 4x^2 + x^4) + q(15 - 30x^2 + 7x^4)\right) u^4 \\
&+ r_x(u) u^4,
\end{aligned} \tag{59}$$

with $\lim_{u \to 0} r_x(u) = 0$.

Proof. The claim follows from Taylor's theorem after computing the first four derivatives of f_x at $u = 0$. We computed the derivatives with Mathematica. The notebook together with the corresponding pdf file can be found in https://github.com/DPBayes/subsampling-is-not-magic/tree/main/notebooks under the names LEMMA_A.2_COMPUTATION.{NB,PDF}. □

Lemma A.3. *When $x \sim \mathcal{N}(0, 1)$,*

$$\mathbb{E}_x\left(\frac{1}{2}(q-1)q(x^2-1)u^2\right) = 0 \tag{60}$$

$$\mathbb{E}_x\left(-\frac{1}{6}(q-1)q(2q-1)x(x^2-3)u^3\right) = 0 \tag{61}$$

$$\mathbb{E}_x\left(\frac{f_x^{(4)}(0)}{4!}u^4\right) = \frac{1}{4}(q-1)^2q^2u^4. \tag{62}$$

Proof. It is well-known that $\mathbb{E}(x) = 0$, $\mathbb{E}(x^2) = 1$, $\mathbb{E}(x^3) = 0$ and $\mathbb{E}(x^4) = 3$. The first expectation:

$$\begin{aligned}\mathbb{E}_x\left(\frac{1}{2}(q-1)q(x^2-1)u^2\right) &= \frac{1}{2}(q-1)qu^2\,\mathbb{E}_x\left(x^2-1\right)\\ &= \frac{1}{2}(q-1)qu^2\left(\mathbb{E}_x(x^2)-1\right)\\ &= 0.\end{aligned} \tag{63}$$

The second expectation:

$$\begin{aligned}\mathbb{E}_x\left(-\frac{1}{6}(q-1)q(2q-1)x(x^2-3)u^3\right) &= -\frac{1}{6}(q-1)q(2q-1)u^3\,\mathbb{E}_x\left(x(x^2-3)\right)\\ &= -\frac{1}{6}(q-1)q(2q-1)u^3\left(\mathbb{E}_x(x^3)-3\,\mathbb{E}_x(x)\right)\\ &= 0.\end{aligned} \tag{64}$$

The third expectation:

$$\begin{aligned}&\mathbb{E}_x\left(\frac{f_x^{(4)}(0)}{4!}u^4\right)\\ &= \mathbb{E}_x\left(\frac{1}{24}q\left(-3+6x^2-x^4-12q^2(2-4x^2+x^4)+6q^3(2-4x^2+x^4)+q(15-30x^2+7x^4)\right)u^4\right)\\ &= \frac{1}{24}q\left(-3+\mathbb{E}_x(6x^2)-\mathbb{E}_x(x^4)-12q^2\,\mathbb{E}_x(2-4x^2+x^4)+6q^3\,\mathbb{E}_x(2-4x^2+x^4)+q\,\mathbb{E}_x(15-30x^2+7x^4)\right)u^4\\ &= \frac{1}{24}q\left(-3+6-3-12q^2(2-4+3)+6q^3(2-4+3)+q(15-30+21)\right)u^4\\ &= \frac{1}{24}q\left(-12q^2+6q^3+6q\right)u^4\\ &= \frac{1}{4}q^2\left(-2q+q^2+1\right)u^4\\ &= \frac{1}{4}(q-1)^2q^2u^4.\end{aligned} \tag{65}$$

□

Lemmas A.2 and A.3 show that the Taylor approximation of $\mathbb{E}_x(f_x(u))$ is

$$\mathbb{E}_x(f_x(u)) = \frac{1}{4}(q-1)^2q^2u^4 + r(u)u^4 \tag{66}$$

if we can differentiate under the expectation. Next, we show that this is possible in Lemma A.7, which requires several preliminaries.

Definition A.4. A function $g(x, u)$ is a polynomial-exponentiated simple polynomial (PESP) if

$$g(x, u) = \sum_{i=1}^{n} P_i(x, u) e^{Q_i(x,u)} \tag{67}$$

for some $n \in \mathbb{N}$ and polynomials $P_i(x, u)$ and $Q_i(x, u)$, $1 \leq i \leq n$, with $Q_i(x, u)$ being first-degree in x.

Lemma A.5. *If g_1 and g_2 are PESPs,*

1. *$g_1 + g_2$ is a PESP,*
2. *$g_1 \cdot g_2$ is a PESP,*
3. *$\frac{\partial}{\partial u} g_1$ is a PESP.*

Proof. Let

$$g_j(x, u) = \sum_{i=1}^{n_j} P_{i,j}(x, u) e^{Q_{i,j}(x,u)} \tag{68}$$

for $j \in \{1, 2\}$. (1) is clear by just writing the sums in g_1 and g_2 as a single sum. For (2),

$$\begin{aligned} g_1(x, u) \cdot g_2(x, u) &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P_{i,1}(x, u) P_{j,2}(x, u) e^{Q_{i,1}(x,u)} e^{Q_{j,2}(x,u)} \\ &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P_{i,1}(x, u) P_{j,2}(x, u) e^{Q_{i,1}(x,u) + Q_{j,2}(x,u)} \\ &= \sum_{i=1}^{n_3} P_{i,3}(x, u) e^{Q_{i,3}(x,u)} \end{aligned} \tag{69}$$

since the product of two polynomials is a polynomial, and the sum of two polynomials is a polynomial of the same degree. For (3)

$$\begin{aligned} \frac{\partial}{\partial u} g_1(x, u) &= \sum_{i=1}^{n_1} \left(\frac{\partial}{\partial u} P_{i,1}(x, u) \right) e^{Q_{i,1}(x,u)} + \sum_{i=1}^{n_1} P_{i,1}(x, u) \left(\frac{\partial}{\partial u} Q_{i,1}(x, u) \right) e^{Q_{i,1}(x,u)} \\ &= \sum_{i=1}^{n_1} P_{i,4}(x, u) e^{Q_{i,1}(x,u)} \end{aligned} \tag{70}$$

since the partial derivatives, products and sums of polynomials are polynomials. □

Lemma A.6. *When $x \sim \mathcal{N}(0, 1)$, for $a > 0, b \in \mathbb{R}$ and $k \in \mathbb{N}$,*

$$\mathbb{E}_x(a|x|^k e^{b|x|}) < \infty. \tag{71}$$

Proof.

$$\begin{aligned}
\mathbb{E}_x(a|x|^k e^{b|x|}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} a|x|^k e^{b|x|} e^{-\frac{1}{2}x^2} \mathrm{d}x \\
&\propto \int_{-\infty}^{\infty} |x|^k e^{b|x|} e^{-\frac{1}{2}x^2} \mathrm{d}x \\
&= 2\int_{0}^{\infty} x^k e^{bx} e^{-\frac{1}{2}x^2} \mathrm{d}x \\
&= 2\int_{0}^{\infty} x^k e^{-\frac{1}{2}(x^2-2bx)} \mathrm{d}x \\
&\propto \int_{0}^{\infty} x^k e^{-\frac{1}{2}(x^2-2bx+b^2)} \mathrm{d}x \\
&= \int_{0}^{\infty} x^k e^{-\frac{1}{2}(x-b)^2} \mathrm{d}x \\
&\leq \int_{-\infty}^{\infty} |x|^k e^{-\frac{1}{2}(x-b)^2} \mathrm{d}x \\
&< \infty
\end{aligned} \tag{72}$$

since all absolute moments of Gaussian distributions are finite. □

Lemma A.7. *For any $k \in \mathbb{N}$, $k \geq 1$, there is a function $g_k(x)$ such that $|f_x^{(k)}(u)| \leq g_k(x)$ for all $u \in [-1, 1]$ and $x \in \mathbb{R}$, and $\mathbb{E}_x(g_k(x)) < \infty$.*

Proof. We start by computing the first derivative of $f_x(u)$

$$\begin{aligned}
f_x(u) &= \ln \frac{e^{-\frac{1}{2}x^2}}{qe^{-\frac{1}{2}((q-1)u+x)^2} + (1-q)e^{-\frac{1}{2}(qu+x)^2}} \\
&= -\ln\left(qe^{-\frac{1}{2}((q-1)u+x)^2} + (1-q)e^{-\frac{1}{2}(qu+x)^2}\right) - \frac{x^2}{2}
\end{aligned} \tag{73}$$

$$f_x'(u) = -\frac{-q(q-1)((q-1)u+x)e^{-\frac{1}{2}((q-1)u+x)^2} - (1-q)q(qu+x)e^{-\frac{1}{2}(qu+x)^2}}{qe^{-\frac{1}{2}((q-1)u+x)^2} + (1-q)e^{-\frac{1}{2}(qu+x)^2}} \tag{74}$$

Since

$$e^{-\frac{1}{2}((q-1)u+x)^2} = e^{-\frac{1}{2}((q-1)^2u^2+2(q-1)xu+x^2)} = e^{-\frac{1}{2}x^2}e^{-\frac{1}{2}((q-1)^2u^2+2(q-1)xu)} \tag{75}$$

and

$$e^{-\frac{1}{2}(qu+x)^2} = e^{-\frac{1}{2}(q^2u^2+2qux+x^2)} = e^{-\frac{1}{2}x^2}e^{-\frac{1}{2}(q^2u^2+2qux)} \tag{76}$$

we can write the first derivative in the following form:

$$f_x'(u) = \frac{P_1(x,u)e^{Q_1(x,u)} + P_2(x,u)e^{Q_2(x,u)}}{qe^{Q_1(x,u)} + (1-q)e^{Q_2(x,u)}} \tag{77}$$

for polynomials $P_1(x,u)$, $P_2(x,u)$, $Q_1(x,u)$ and $Q_2(x,u)$, with $Q_1(x,u)$ and $Q_2(x,u)$ being first-degree in x.

We show by induction that further derivatives have a similar form:

$$f_x^{(k)}(u) = \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{(qe^{Q_1(x,u)} + (1-q)e^{Q_2(x,u)})^{2^{k-1}}} \tag{78}$$

We also require the Q-polynomials to still be first degree in x, so the numerator must be a PESP. The claim is clearly true for $k = 1$. If the claim is true for k, then

$$f_x^{(k+1)}(u) = \frac{G(x,u)\frac{\partial}{\partial u}F(x,u) - F(x,u)\frac{\partial}{\partial u}G(x,u)}{(qe^{Q_1(x,u)} + (1-q)e^{Q_2(x,u)})^{2^k}} \tag{79}$$

where

$$F(x,u) = \sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)} \tag{80}$$

$$G(x,u) = (qe^{Q_1(x,u)} + (1-q)e^{Q_2(x,u)})^{2^{k-1}} \tag{81}$$

by the quotient differentiation rule. The denominator has the correct form, so it remains to show that the numerator is a PESP. F is clearly a PESP, and so is G due to Lemma A.5. The numerator is a sum of products of F, G and their derivatives, so it is a PESP by Lemma A.5, which concludes the induction proof.

We have

$$qe^{Q_1(x,u)} + (1-q)e^{Q_2(x,u)} \geq \min\left(e^{Q_1(x,u)}, e^{Q_2(x,u)}\right). \tag{82}$$

We can split $\mathbb{R}$ into measurable subsets A_1 and A_2 such that

$$\min\left(e^{Q_1(x,u)}, e^{Q_2(x,u)}\right) = \begin{cases} e^{Q_1(x,u)} & x \in A_1 \\ e^{Q_2(x,u)} & x \in A_2 \end{cases} \tag{83}$$

Now

$$|f_x^{(k)}(u)| \leq \left| \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{\min\left(e^{Q_1(x,u)}, e^{Q_2(x,u)}\right)^{2^{k-1}}} \right| \tag{84}$$

$$= \left| I_{A_1} \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{(e^{Q_1(x,u)})^{2^{k-1}}} \right| + \left| I_{A_2} \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{(e^{Q_2(x,u)})^{2^{k-1}}} \right| \tag{85}$$

$$\leq \left| \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{e^{2^{k-1}Q_1(x,u)}} \right| + \left| \frac{\sum_{i=1}^{n_k} P_{i,k}(x,u)e^{Q_{i,k}(x,u)}}{e^{2^{k-1}Q_2(x,u)}} \right| \tag{86}$$

$$\leq \sum_{i=1}^{m_k} |R_{i,k}(x,u)|e^{S_{i,k}(x,u)} \tag{87}$$

where $R_{i,k}(x,u)$ and $S_{i,k}(x,u)$ are further polynomials, with the S-polynomials being of first degree in x.

Since $u \in [-1,1]$, for a monomial $ax^{k_1}u^{k_2}$

$$ax^{k_1}u^{k_2} \leq |ax^{k_1}u^{k_2}| \leq |a| \cdot |x|^{k_1}. \tag{88}$$

Using this inequality on each monomial of $R_{i,k}(x,u)$ and $S_{i,k}(x,u)$ gives upper bound polynomials of $|x|$ $\hat{R}_{i,k}(x)$ and $\hat{S}_{i,k}(x)$ such that

$$|f_x^{(k)}(u)| \leq \sum_{i=1}^{m_k} |R_{i,k}(x,u)|e^{S_{i,k}(x,u)} \leq \sum_{i=1}^{m_k} |\hat{R}_{i,k}(x)|e^{\hat{S}_{i,k}(x)}, \tag{89}$$

with the $\hat{S}$-polynomials being first degree.

Let

$$g_k(x) = \sum_{i=1}^{m_k} |\hat{R}_{i,k}(x)|e^{\hat{S}_{i,k}(x)}. \tag{90}$$

We have shown that $|f_x(u)| \leq g_k(x)$ for $u \in [-1,1]$ and $x \in \mathbb{R}$. The integrability of $g_k(x)$ against a standard Gaussian follows from Lemma A.6, as we can first push the absolute value around $\hat{R}_{i,k}(x)$ to be around each monomial of $\hat{R}_{i,k}(x)$ with the triangle inequality, and then write the resulting upper bound as a sum with each term of the form $a|x|^k e^{b|x|}$, with $a > 0$, $b \in \mathbb{R}$ and $k \in \mathbb{N}$. □

Now we can put the preliminaries together to use the Taylor approximation of of $\mathbb{E}_x(f_x(u))$ to find its order of convergence.

Lemma A.8. *When $x \sim \mathcal{N}(0,1)$, $\mathbb{E}_x(f_x(u)) = O(u^4)$ as $u \to 0$.*

Proof. Since $u \to 0$ in the limit, it suffices to consider $u \in [-1, 1]$. First, we find the fourth-order Taylor approximation of $\mathbb{E}_x(f_x(u))$. Lemma A.7 allows us to differentiate under the expectation four times. Then Taylor's theorem, and Lemmas A.2 and A.3 give

$$\mathbb{E}_x(f_x(u)) = \frac{1}{4}(q-1)^2 q^2 u^4 + r(u)u^4 \tag{91}$$

where $\lim_{u\to 0} r(u) = 0$. Now

$$\begin{aligned}\lim_{u\to 0} \frac{1}{u^4}\,\mathbb{E}(f_x(u)) &= \lim_{u\to 0} \frac{1}{u^4}\left(\frac{1}{4}(q-1)^2 q^2 u^4 + r(u)u^4\right)\\ &= \frac{1}{4}(q-1)^2 q^2 + \lim_{u\to 0} r(u)\\ &= \frac{1}{4}(q-1)^2 q^2\\ &< \infty,\end{aligned} \tag{92}$$

which implies the claim. □

A.3. Proof of Theorem 5.2

Theorem 5.2. *For $1 \le i \le T$, let*

$$\mathcal{M}_i(x) \sim xB_q + \mathcal{N}(0, \sigma_T^2), \tag{31}$$

$$\mathcal{M}_i'(x) \sim \mathcal{N}(qx, \sigma_T^2). \tag{32}$$

be independent for each i. Let $\mathcal{M}_{1:T}$ be the composition of $\mathcal{M}_i$, and let $\mathcal{M}'_{1:T}$ be the composition of $\mathcal{M}'_i$. Then

$$\sup_x \mathrm{TV}(\mathcal{M}_{1:T}(x), \mathcal{M}'_{1:T}(x)) \to 0 \tag{33}$$

as $T \to \infty$.

Proof. It suffices to show

$$\sup_x T \cdot \mathrm{KL}(\mathcal{M}_i'(x), \mathcal{M}_i(x)) \to 0 \tag{93}$$

due to Pinsker's inequality and the additivity of KL divergence for products of independent random variables (Lemma A.1). When $x = 0$, the two mechanism are the same, so it suffices to look at $x = 1$.

KL-divergence is invariant to bijections, so

$$\begin{aligned}\mathrm{KL}(\mathcal{M}_i'(1), \mathcal{M}_i(1)) &= \mathrm{KL}\big(\mathcal{N}(q, \sigma_T^2), q\mathcal{N}(1, \sigma_T^2) + (1-q)\mathcal{N}(0, \sigma_T^2)\big)\\ &= \mathrm{KL}\left(\mathcal{N}\left(q\frac{1}{\sigma_T}, 1\right), q\mathcal{N}\left(\frac{1}{\sigma_T}, 1\right) + (1-q)\mathcal{N}(0,1)\right)\\ &= \mathrm{KL}\left(\mathcal{N}(qu, 1), q\mathcal{N}(u, 1) + (1-q)\mathcal{N}(0,1)\right)\\ &= \mathrm{KL}\left(\mathcal{N}(0, 1), q\mathcal{N}(u - qu, 1) + (1-q)\mathcal{N}(-qu, 1)\right)\end{aligned} \tag{94}$$

where we first divide both distributions by σ_T, then set $u = \frac{1}{\sigma_T}$, and finally subtract qu. As $\sigma_T^2 = \Omega(T)$, $u = O\left(\frac{1}{\sqrt{T}}\right)$.

From the definition of KL-divergence, $u^4 = O(\frac{1}{T^2})$ and Lemma A.8, when $x \sim \mathcal{N}(0, 1)$ we have

$$\mathrm{KL}(\mathcal{M}_i'(1), \mathcal{M}_i(1)) = \mathbb{E}_x(f_x(u)) = O(u^4) = O\left(\frac{1}{T^2}\right). \tag{95}$$

This implies

$$\lim_{T\to\infty} T \cdot \mathrm{KL}(\mathcal{M}_i'(1), \mathcal{M}_i(1)) = 0, \tag{96}$$

which implies the claim. □

A.4. Proof of Corollary 5.4

With fixed Δ, T, ϵ, the function $\sigma \mapsto \mathrm{AO}_G(\sigma, \Delta, T, \epsilon)$ is strictly decreasing (Balle & Wang, 2018, Lemma 7) and continuous, so it has a continuous inverse $\delta \mapsto \mathrm{AO}_G^{-1}(\delta, \Delta, T, \epsilon)$. To declutter the notation, we omit the ϵ and Δ arguments from AO_S, AO_G and AO_G^{-1} in the rest of this section.

Lemma A.9. *For AO_G and its inverse,*

1. $\mathrm{AO}_G(\sigma, T) = \mathrm{AO}_G\left(\frac{\sigma}{\sqrt{T}}, 1\right)$,
2. $\mathrm{AO}_G^{-1}(\delta, T) = \mathrm{AO}_G^{-1}(\delta, 1)\sqrt{T}$.

Proof. Recall that the PLRV of the Gaussian mechanism is $\mathcal{N}(\mu_1, 2\mu_1)$ with $\mu_1 = \frac{\Delta^2}{2\sigma^2}$ (Sommer et al., 2019). By Theorem 2.8, the PLRV of T compositions of the Gaussian mechanism is $\mathcal{N}(T\mu_1, 2T\mu_1)$. Denoting $\mu_T = T\mu_1 = \frac{T\Delta^2}{2\sigma^2}$, we see that the T-fold composition of the Gaussian mechanism has the same PLRV as a single composition of the Gaussian mechanism with standard deviation $\frac{\sigma}{\sqrt{T}}$, which proves (1).

To prove (2), first we have

$$\mathrm{AO}_G(\mathrm{AO}_G^{-1}(\delta, 1)\sqrt{T}, T) = \mathrm{AO}_G(\mathrm{AO}_G^{-1}(\delta, 1), 1) = \delta \tag{97}$$

by applying (1) to the outer AO_G. Applying $\mathrm{AO}_G^{-1}(\cdot, T)$ to both sides gives

$$\mathrm{AO}_G^{-1}(\delta, 1)\sqrt{T} = \mathrm{AO}_G^{-1}(\delta, T) \tag{98}$$

□

Corollary 5.4. *Let $\sigma(q, T)$ be the smallest σ such that $\mathrm{AO}_S(\sigma, \Delta, q, T, \epsilon) \leq \delta$. Then*

$$\lim_{T\to\infty} \frac{\sigma(q, T)}{q\sigma(1, T)} = 1. \tag{44}$$

Proof. By definition, $\mathrm{AO}_S(\sigma(q, T), q, T) = \delta$, so Theorem 5.3 implies

$$\lim_{T\to\infty} \mathrm{AO}_S\left(\frac{\sigma(q, T)}{q}, 1, T\right) = \delta. \tag{99}$$

Since $\mathrm{AO}_S(\sigma, 1, T) = \mathrm{AO}_G(\sigma, T)$ for any σ, we get from Lemma A.9

$$\left|\mathrm{AO}_G\left(\frac{\sigma(q, T)}{q\sqrt{T}}, 1\right) - \delta\right| = \left|\mathrm{AO}_G\left(\frac{\sigma(q, T)}{q}, T\right) - \delta\right| \to 0 \tag{100}$$

as $T \to \infty$. We have

$$\mathrm{AO}_G^{-1}\left(\mathrm{AO}_G\left(\frac{\sigma(q, T)}{q\sqrt{T}}, 1\right), 1\right) = \frac{\sigma(q, T)}{q\sqrt{T}} \tag{101}$$

and by Lemma A.9,

$$\mathrm{AO}_G^{-1}(\delta, 1) = \frac{1}{\sqrt{T}}\mathrm{AO}_G^{-1}(\delta, T) = \frac{\sigma(1, T)}{\sqrt{T}}. \tag{102}$$

Now, by the continuity of $\mathrm{AO}_G^{-1}(\cdot, 1)$,

$$\frac{1}{\sqrt{T}}\left|\frac{\sigma(q, T)}{q} - \sigma(1, T)\right| \to 0 \tag{103}$$

as $T \to \infty$.

Since $\sigma(1, T) = \Omega(\sqrt{T})$ by Theorem 5.1,

$$\left|\frac{\sigma(q, T)}{q\sigma(1, T)} - 1\right| = \frac{|\sigma(q, T)/q - \sigma(1, T)|}{\sigma(1, T)} \to 0 \tag{104}$$

as $T \to \infty$, which implies the claim. □

A.5. $\sigma_T^2 = \Omega(T)$

Lemma A.10. *Let L be the PLRV of a single iteration of the Poisson subsampled Gaussian mechanism. Then $\mathbb{E}(L) \geq \frac{q^2}{2\sigma^2}$.*

Proof. Recall, that for Poisson subsampled Gaussian mechanism we have the dominating pair $P = q\mathcal{N}(1, \sigma^2) + (1 - q)\mathcal{N}(0, \sigma^2)$ and $Q = \mathcal{N}(0, \sigma^2)$. Let f_P and f_Q be their densities. Now the mean of the PLRV can be written as

$$\begin{aligned}
\mathbb{E}(L) &= \mathbb{E}_{t\sim P}\left[\log \frac{f_P(t)}{f_Q(t)}\right] \\
&= \mathbb{E}_{t\sim P}[\log f_P(t)] - \mathbb{E}_{t\sim P}[\log f_Q(t)] \\
&= -H(P) - \left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\,\mathbb{E}_{t\sim P}[t^2]\right) \\
&= -H(P) + \frac{1}{2}\left(\log 2\pi + \log\sigma^2 + \frac{1}{\sigma^2}(\sigma^2 + q)\right) \\
&= -H(P) + \frac{1}{2}\left(\log 2\pi + \log\sigma^2 + \frac{q}{\sigma^2} + 1\right),
\end{aligned} \tag{105}$$

where H denotes the differential entropy. The entropy term is analytically intractable for the mixture of Gaussians (the P). However, we can upper bound it with the entropy of a Gaussian with the same variance, as the Gaussian distribution maximises entropy among distributions with given mean and variance. The variance of P is $\sigma^2 + q - q^2 = \sigma^2 + q(1 - q)$, and therefore

$$\begin{aligned}
H(P) &\leq H(\mathcal{N}(0, \sigma^2 + q(1 - q))) \\
&= \frac{1}{2}\left(\log 2\pi(\sigma^2 + q(1 - q)) + 1\right) \\
&= \frac{1}{2}\left(\log 2\pi + \log\big(\sigma^2 + q(1 - q)\big) + 1\right)
\end{aligned} \tag{106}$$

Now, substituting this into (105) we get the following lower bound

$$\mathbb{E}(L) \geq \frac{1}{2}\left(\log\frac{\sigma^2}{\sigma^2 + q(1 - q)} + \frac{q}{\sigma^2}\right) = \frac{1}{2}\left(-\log\left(1 + \frac{q - q^2}{\sigma^2}\right) + \frac{q}{\sigma^2}\right). \tag{107}$$

Since for all $x \geq -1$,

$$\log(1 + x) \leq x \Leftrightarrow -\log(1 + x) \geq -x, \tag{108}$$

we have

$$-\log\left(1 + \frac{q - q^2}{\sigma^2}\right) \geq -\frac{q - q^2}{\sigma^2} \tag{109}$$

which gives

$$\mathbb{E}(L) \geq \frac{1}{2}\left(-\frac{q - q^2}{\sigma^2} + \frac{q}{\sigma^2}\right) = \frac{q^2}{2\sigma^2}. \tag{110}$$

□

Lemma A.11. *Let L be the PLRV of a single iteration of the Poisson subsampled Gaussian mechanism. Then $\mathrm{Var}(L) \leq \frac{1}{\sigma^2} + \frac{1}{4\sigma^4}$*

Proof.

$$\begin{aligned}
\ln \frac{f_P(t)}{f_Q(t)} &= \ln \frac{q \exp\left(-\frac{(t-1)^2}{2\sigma^2}\right) + (1-q)\exp\left(-\frac{t}{2\sigma^2}\right)}{\exp\left(-\frac{t^2}{2\sigma^2}\right)} \\
&= \ln\left(q \exp\left(\frac{t^2-(t-1)^2}{2\sigma^2}\right) + (1-q)\right) \\
&= \ln\left(q \exp\left(\frac{2t-1}{2\sigma^2}\right) + (1-q)\right) \\
&\leq \ln \max\left\{\exp\left(\frac{2t-1}{2\sigma^2}\right), 1\right\} \\
&= \max\left\{\frac{2t-1}{2\sigma^2}, 0\right\}
\end{aligned} \tag{111}$$

Similarly, we also have

$$\ln \frac{f_P(t)}{f_Q(t)} \geq \ln \min\left\{\exp\left(\frac{2t-1}{2\sigma^2}\right), 1\right\} = \min\left\{\frac{2t-1}{2\sigma^2}, 0\right\} \tag{112}$$

so

$$\begin{aligned}
\left|\ln \frac{f_P(t)}{f_Q(t)}\right| &\leq \max\left\{\max\left\{\frac{2t-1}{2\sigma^2}, 0\right\}, -\min\left\{\frac{2t-1}{2\sigma^2}, 0\right\}\right\} \\
&= \max\left\{\max\left\{\frac{2t-1}{2\sigma^2}, 0\right\}, \max\left\{-\frac{2t-1}{2\sigma^2}, 0\right\}\right\} \\
&\leq \left|\frac{2t-1}{2\sigma^2}\right|
\end{aligned} \tag{113}$$

Since $\mathbb{E}_{t\sim P}(t) = q$ and $\mathbb{E}_{t\sim P}(t^2) = \sigma^2 + q$,

$$\begin{aligned}
\mathrm{Var}(L^{(T)}) &\leq \mathbb{E}((L^{(T)})^2) \\
&= \mathbb{E}_{t\sim P}\left(\left(\ln \frac{f_P(t)}{f_Q(t)}\right)^2\right) \\
&\leq \mathbb{E}_{t\sim P}\left(\left|\frac{2t-1}{2\sigma^2}\right|^2\right) \\
&= \mathbb{E}_{t\sim P}\left(\left(\frac{2t-1}{2\sigma^2}\right)^2\right) \\
&= \frac{1}{4\sigma^4}\,\mathbb{E}_{t\sim P}\left(4t^2 - 4t + 1\right) \\
&= \frac{\sigma^2+q}{\sigma^4} - \frac{q}{\sigma^4} + \frac{1}{4\sigma^4} \\
&= \frac{1}{\sigma^2} + \frac{1}{4\sigma^4}.
\end{aligned} \tag{114}$$

□

Lemma A.12. *Let L_T be the PLRV of T iterations of the Poisson subsampled Gaussian mechanism, and let $K \in \mathbb{N}$. If $\sigma_T^2 = \Omega(T)$ is not true, for any $\alpha_i > 0$, $b_i > 0$ with $1 \leq i \leq K$, it is possible to find a T such that,*

$$\Pr_{s\sim L_T}(s \leq b_i) \leq \alpha_i \tag{115}$$

holds simultaneously for all i.

Proof. If $\sigma_T^2 = \Omega(T)$ is not true,

$$\liminf_{T\to\infty} \frac{\sigma_T^2}{T} = 0. \tag{116}$$

By Lemma A.10 and the composition theorem, $\mathbb{E}(L_T) \geq T\frac{q^2}{2\sigma_T^2}$. By Lemma A.11, $\mathrm{Var}(L_T) \leq \frac{T}{\sigma_T^2} + \frac{T}{4\sigma_T^4}$.

Let $k_i = \frac{1}{\sqrt{\alpha_i}}$. Choose T such that

$$\frac{Tq^2}{\sigma_T^2} - k_i\sqrt{\frac{T}{\sigma_T^2} + \frac{T}{4\sigma_T^4}} \geq b_i. \tag{117}$$

for all i. This is possible by (116) by choosing $\frac{T}{\sigma_T^2}$ to be large enough to satisfy all the inequalities. Now

$$\mathbb{E}(L_T) - k_i\sqrt{\mathrm{Var}(L_T)} \geq \frac{Tq^2}{\sigma_T^2} - k_i\sqrt{\frac{T}{\sigma_T^2} + \frac{T}{4\sigma_T^4}} \geq b_i \tag{118}$$

for all i, so

$$\Pr_{s\sim L_T}(s \leq b_i) \leq \Pr_{s\sim L_T}\left(|s - \mathbb{E}(L_T)| \geq k_i\sqrt{\mathrm{Var}(L_T)}\right) \leq \frac{1}{k_i^2} = \alpha_i \tag{119}$$

for all i by Chebyshev's inequality. □

This means that it is possible to make $\Pr_{s\sim L_T}(s \leq b)$ arbitrarily small for any b by choosing an appropriate T, and to satisfy a finite number of these constraints simultaneously with a single T.

Theorem 5.1. *Let σ_T be such that $\mathrm{AO}_S(\sigma_T, \Delta, q, T, \epsilon) \leq \delta$ for all T, with $\delta < 1$. Then $\sigma_T^2 = \Omega(T)$.*

Proof. By Lemma 3.3, it suffices to consider $\Delta = 1$. To obtain a contradiction, assume that $\sigma_T^2 = \Omega(T)$ is not true. Let L_T be the PLRV for T iterations of the Poisson subsampled Gaussian mechanism.

From (4),

$$\begin{aligned}
\mathrm{AO}_S(\sigma_T, q, T, \epsilon) &= \mathbb{E}_{s\sim L_T}((1 - e^{\epsilon - s})_+) \\
&= \mathbb{E}_{s\sim L_T}(I(s > \epsilon)(1 - e^{\epsilon - s})) \\
&= \mathbb{E}_{s\sim L_T}(I(s > \epsilon)) - \mathbb{E}_{s\sim L}(I(s > \epsilon)e^{\epsilon - s}) \\
&= \Pr_{s\sim L_T}(s > \epsilon) - \mathbb{E}_{s\sim L}(I(s > \epsilon)e^{\epsilon - s})
\end{aligned} \tag{120}$$

By choosing $b_1 = \epsilon$ and $\alpha_1 = 1 - \frac{1}{2}(\delta + 1)$ in Lemma A.12, we get $\Pr_{s\sim L_T}(s > \epsilon) \geq \frac{1}{2}(\delta + 1)$.

To bound the remaining term,

$$\begin{aligned}
\mathbb{E}_{s\sim L_T}(I(s > \epsilon)e^{\epsilon - s}) &= e^{\epsilon}\,\mathbb{E}_{s\sim L_T}(I(s > \epsilon)e^{-s}) \\
&= e^{\epsilon}\sum_{i=0}^{\infty}\mathbb{E}_{s\sim L}(I(\epsilon + i < s \leq \epsilon + i + 1)e^{-s}) \\
&\leq e^{\epsilon}\sum_{i=0}^{\infty}\mathbb{E}_{s\sim L}(I(\epsilon + i < s \leq \epsilon + i + 1)e^{-\epsilon - i}) \\
&\leq \sum_{i=0}^{\infty} e^{-i}\,\mathbb{E}_{s\sim L}(I(\epsilon + i < s \leq \epsilon + i + 1)) \\
&\leq \sum_{i=0}^{\infty} e^{-i}\Pr_{s\sim L}(\epsilon + i < s \leq \epsilon + i + 1) \\
&= \sum_{i=0}^{K} e^{-i}\Pr_{s\sim L}(\epsilon + i < s \leq \epsilon + i + 1) + \sum_{i=K+1}^{\infty} e^{-i}\Pr_{s\sim L}(\epsilon + i < s \leq \epsilon + i + 1) \\
&\leq \sum_{i=0}^{K} e^{-i}\Pr_{s\sim L}(s \leq \epsilon + i + 1) + \sum_{i=K+1}^{\infty} e^{-i}.
\end{aligned} \tag{121}$$

The series $\sum_{i=0}^{\infty} e^{-i}$ converges, so it is possible to make $\sum_{i=K+1}^{\infty} e^{-i}$ arbitrarily small by choosing an appropriate K, which does not depend on T.

If we choose K such that $\sum_{i=K+1}^{\infty} e^{-i} < \frac{1}{4}(1-\delta)$ and then choose T such that $e^{-i}\Pr(s \leq \epsilon + i + 1) < \frac{1}{4(K+1)}(1-\delta)$ for all $0 \leq i \leq K$, we have

$$\mathbb{E}_{s\sim L_T}(I(s>\epsilon)e^{\epsilon-s}) < (K+1)\cdot\frac{1}{4(K+1)}(1-\delta) + \frac{1}{4}(1-\delta) = \frac{1}{2}(1-\delta). \tag{122}$$

Lemma A.12 allows multiple inequalities for a single T, so we can find a T that satisfies all of the $K+2$ inequalities we have required it to satisfy. With this T,

$$\mathrm{AO}_S(\sigma_T, q, T, \epsilon) = \Pr_{s\sim L_T}(s>\epsilon) - \mathbb{E}_{s\sim L}(I(s>\epsilon)e^{\epsilon-s}) > \frac{1}{2}(\delta+1) - \frac{1}{2}(1-\delta) = \delta \tag{123}$$

which is a contradiction. □

A.6. Solving $\sigma'(q)$ for $T=1$ case

Lemma 6.1. *For the smallest $\sigma(q)$ that provides (ϵ,δ)-DP for the Poisson subsampled Gaussian mechanism, we have*

$$\sigma'(q) = \frac{\sigma(q)}{q}\frac{1}{2a}\frac{1}{\mathrm{erf}'(a-b)}(\mathrm{erf}(a-b) - \mathrm{erf}(-a-b)). \tag{48}$$

Proof. Recall Equation (45): for a single iteration Poisson subsampled Gaussian mechanism, we have

$$\begin{aligned}\delta(q) =& q\Pr\left(Z \geq \sigma(q)\log\left(\frac{h(q)}{q}\right) - \frac{1}{2\sigma(q)}\right)\\ &- h(q)\Pr\left(Z \geq \sigma(q)\log\left(\frac{h(q)}{q}\right) + \frac{1}{2\sigma(q)}\right),\end{aligned} \tag{124}$$

where $h(q) = e^{\epsilon} - (1-q)$. Since $\sigma(q)$ is a function that returns a noise-level matching any (ϵ,δ)-DP requirement for subsampling rate q, $\delta'(q) = 0$. Using Mathematica, we can solve the derivative of the RHS in Equation (124) for $\sigma'(q)$ and we get

$$\sigma'(q) = \frac{\sqrt{\frac{\pi}{2}}\sigma(q)^2 e^{\frac{1}{2}\sigma(q)^2\log^2\left(\frac{q+e^{\epsilon}-1}{q}\right)+\frac{1}{8\sigma(q)^2}}\left(\mathrm{erf}\left(\frac{1-2\sigma(q)^2\log\left(\frac{q+e^{\epsilon}-1}{q}\right)}{2\sqrt{2}\sigma(q)}\right) - \mathrm{erf}\left(-\frac{2\sigma(q)^2\log\left(\frac{q+e^{\epsilon}-1}{q}\right)+1}{2\sqrt{2}\sigma(q)}\right)\right)}{q\sqrt{\frac{q+e^{\epsilon}-1}{q}}}. \tag{125}$$

The notebook together with the corresponding pdf file can be found in https://github.com/DPBayes/subsampling-is-not-magic/tree/main/notebooks under the names NO_COMP_DERIVATIVE.{NB,PDF}.

Note that

$$\exp\left(\frac{1}{2}\sigma(q)^2\log^2\left(\frac{q+e^{\epsilon}-1}{q}\right) + \frac{1}{8\sigma(q)^2}\right) \tag{126}$$

$$= \exp\left(\frac{1}{2}\left(\sigma(q)^2\log^2\left(\frac{q+e^{\epsilon}-1}{q}\right) + \frac{1}{4\sigma(q)^2}\right)\right) \tag{127}$$

$$= \exp\left(\frac{1}{2}\left(\left(\sigma(q)\log\left(\frac{q+e^{\epsilon}-1}{q}\right) - \frac{1}{2\sigma(q)}\right)^2 + \log\left(\frac{q+e^{\epsilon}-1}{q}\right)\right)\right) \tag{128}$$

$$= \exp\left(\left(\frac{1}{\sqrt{2}}\sigma(q)\log\left(\frac{q+e^{\epsilon}-1}{q}\right) - \frac{1}{2\sqrt{2}\sigma(q)}\right)^2\right)\sqrt{\frac{q+e^{\epsilon}-1}{q}} \tag{129}$$

$$= \exp\left(\left(\frac{1-2\sigma(q)^2\log\left(\frac{q+e^{\epsilon}-1}{q}\right)}{2\sqrt{2}\sigma(q)}\right)^2\right)\sqrt{\frac{q+e^{\epsilon}-1}{q}} \tag{130}$$

and therefore the derivative becomes

$$\sigma'(q) = \frac{\sqrt{\frac{\pi}{2}}\sigma(q)^2 e^{\frac{1}{2}\left(\sigma(q)\log\left(\frac{q+e^\epsilon-1}{q}\right)-\frac{1}{2\sigma(q)}\right)^2}\left(\text{erf}\left(\frac{1-2\sigma(q)^2\log\left(\frac{q+e^\epsilon-1}{q}\right)}{2\sqrt{2}\sigma(q)}\right)-\text{erf}\left(-\frac{2\sigma(q)^2\log\left(\frac{q+e^\epsilon-1}{q}\right)+1}{2\sqrt{2}\sigma(q)}\right)\right)}{q} \tag{131}$$

$$= \frac{\sqrt{\frac{\pi}{2}}\sigma(q)^2 e^{\left(\frac{1-2\sigma(q)^2\log\left(\frac{q+e^\epsilon-1}{q}\right)}{2\sqrt{2}\sigma(q)}\right)^2}\left(\text{erf}\left(\frac{1-2\sigma(q)^2\log\left(\frac{q+e^\epsilon-1}{q}\right)}{2\sqrt{2}\sigma(q)}\right)-\text{erf}\left(-\frac{2\sigma(q)^2\log\left(\frac{q+e^\epsilon-1}{q}\right)+1}{2\sqrt{2}\sigma(q)}\right)\right)}{q} \tag{132}$$

Lets denote

$$a := \frac{1}{2\sqrt{2}\sigma(q)} \tag{133}$$

$$b := \frac{\sigma(q)}{\sqrt{2}}\log\left(\frac{e^\epsilon-(1-q)}{q}\right). \tag{134}$$

Using this notation we can write

$$q\sigma'(q) = \sqrt{\frac{\pi}{2}}\sigma(q)^2\exp\big((a-b)^2\big)(\text{erf}(a-b)-\text{erf}(-a-b)) \tag{135}$$

$$= \sqrt{\frac{\pi}{2}}\sigma(q)^2\exp\big((a-b)^2\big)(\text{erf}(a+b)-\text{erf}(b-a)). \tag{136}$$

Note that we have

$$\text{erf}'(a-b) = \frac{2}{\sqrt{\pi}}\exp\big(-(a-b)^2\big) \tag{137}$$

$$\Leftrightarrow \exp\big((a-b)^2\big) = \frac{2}{\sqrt{\pi}\text{erf}'(a-b)} \tag{138}$$

Hence

$$q\sigma'(q) = \sigma(q)^2\frac{\sqrt{2}}{\text{erf}'(a-b)}(\text{erf}(a+b)-\text{erf}(b-a)) \tag{139}$$

$$= \sigma(q)^2\frac{\sqrt{2}}{\text{erf}'(a-b)}(\text{erf}(a+b)+\text{erf}(a-b)) \tag{140}$$

$$= \sigma(q)\frac{1}{2a}\frac{1}{\text{erf}'(a-b)}(\text{erf}(a+b)+\text{erf}(a-b)) \tag{141}$$

$$= \sigma(q)\frac{1}{2a}\frac{1}{\text{erf}'(a-b)}(\text{erf}(a-b)-\text{erf}(-a-b)). \tag{142}$$

□