BEYOND LAPLACE AND GAUSSIAN: EXPLORING THE GENERALIZED GAUSSIAN MECHANISM FOR PRIVATE MACHINE LEARNING

Roy Rinberg*
Harvard University
royrinberg@g.harvard.edu

Ilia Shumailov
University of Oxford
ilia.shumailov@chch.ox.ac.uk

Rachel Cummings
Columbia University
rac2239@columbia.edu

Nicolas Papernot
University of Toronto & Vector Institute
nicolas.papernot@utoronto.ca

ABSTRACT

Differential privacy (DP) is obtained by randomizing a data analysis algorithm, which necessarily introduces a tradeoff between its utility and privacy. Many DP mechanisms are built upon one of two underlying tools: Laplace and Gaussian additive noise mechanisms. We expand the search space of algorithms by investigating the Generalized Gaussian (GG) mechanism, which samples the additive noise term x with probability proportional to $e^{-\frac{|x|}{\sigma}^{\beta}}$ for some $\beta \geq 1$ (denoted $GG_{\beta,\sigma}(f, D)$). The Laplace and Gaussian mechanisms are special cases of GG for $\beta = 1$ and $\beta = 2$ respectively.

We find two compelling negative results: 1. for ML settings (DP-SGD and PATE), generally $\beta = 2$ (Gaussian) has a privacy-accuracy tradeoff that is at least nearly as good as any other option of β, 2. for mechanisms $GG_{\beta,\sigma}(f, D)$ with ℓ_β sensitivity, privacy accounting in fact dimension independent, dramatically simplifying the computational complexity of privacy accounting, which appears untrue for any other mechanism. This leads to the following conclusion: improving private ML will likely require a conceptual breakthrough, not new privacy mechanisms.

1 Introduction

As applications of machine learning (ML) often involve sensitive information, there is an increasing need to provide privacy protections for the individuals whose data are included in the training datasets. Privacy concerns have prompted the development of privacy-preserving ML techniques, which aim to prevent the leakage of private information analyzed during training. One of the primary frameworks for achieving this goal is differential privacy (DP), a rigorous mathematical framework that provides quantifiable privacy guarantees [Dwork et al., 2006].

Two popular techniques for implementing DP in ML are Differentially Private Stochastic Gradient Descent (DP-SGD) [Abadi et al., 2016] and Private Aggregation of Teacher Ensembles (PATE) [Papernot et al., 2017]. Traditionally, DP-SGD entails Poisson sampling from the dataset, gradient clipping, and then the addition of Gaussian noise to the gradient. PATE, on the other hand, involves training an ensemble of teacher models on disjoint subsets of the data, then privately aggregating the votes of the teacher models on a public dataset, in order to train a student model on the privately labeled public dataset. PATE achieves private vote aggregation through a private variant of Argmax, which is obtained by adding noise to the vote counts, and then finding the Argmax of the noisy histogram.

Both DP-SGD and PATE achieve privacy protection through mechanisms that add noise drawn from specific probability distributions, namely Laplace or Gaussian. The choice of the noise distribution plays a crucial role in determining the privacy-accuracy tradeoffs of an algorithm, and algorithm designers often make problem-dependent decisions in choosing between Laplace and Gaussian Mechanisms. However, many of the underlying tradeoffs between these two

*Part of this work completed while R.R. was a student at Columbia University and visiting University of Toronto

discrete choices remain unclear. In this work, we explore a continuum of private mechanisms that extends these two special cases of noise distributions. We investigate the Generalized Gaussian Mechanism (GG) [Liu, 2019], denoted $GG_{\beta,\sigma}(f, D)$, which adds noise to the true function value $f(D)$ sampled from the Generalized Gaussian distribution,[2] denoted $\mathcal{N}_\beta(\mu, \sigma)$, with probability density function (PDF),

$$p(x|\mu, \sigma, \beta) \propto e^{-\frac{|x-\mu|^\beta}{\sigma}}. \tag{1}$$

Figure 1 illustrates these PDFs with $\mu = 0$ for different β and σ values on both linear and log scales. Interestingly, all of the noise distributions illustrated in Figure 1 correspond to equivalent DP guarantees when used in the GG Mechanism (see Appendix B.4 for more details). We observe that a larger β value corresponds to a PDF that is more concentrated around the mean. In order to satisfy the same (ϵ, δ)-DP, σ must be simultaneously increased to compensate for the lighter tail.

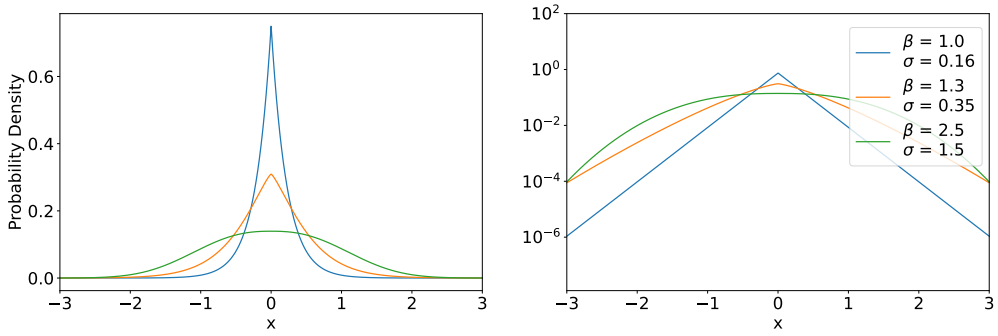

Figure 1: Linear (left) and log-scale (right) PDFs of the Generalized Gaussian distribution for GG mechanisms of varying β, which all satisfy the same (ϵ, δ)-DP privacy guarantee.

We focus on this family of mechanisms because it generalizes both the Laplace and Gaussian Mechanisms, which are are special cases of the GG Mechanism for $\beta = 1$ and $\beta = 2$, respectively.

We explore two settings relevant to ML: PATE and DP-SGD. Our findings show that the the choice of β has a relatively small effect on test-accuracy, but that values around $\beta = 2$ perform near-optimally, although many other choices of β also perform well. This helps explain why Gaussian noise may be so popular, but suggests that the choice of Gaussian or Laplace Mechanism plays a smaller role than anticipated. However, for settings where small improvements are critical, hyperparameter search over β can provide incremental gains. We do find the Gaussian mechanism does not appear to be completely optimal, and thus our work also suggests future directions for further hyperparameter search across a more general range of noise distributions in differentially private algorithm design. However, more importantly, we find that for mechanisms $GG_{\beta,\sigma}(f, D)$, the ℓ_β sensitivity is actually independent of the dimension, significantly reducing the computational complexity of privacy accounting.

Thus, we have uncovered two significant negative findings: 1. For the majority of mechanisms, privacy accounting depends on the dimensionality, making it computationally challenging to apply to high-dimensional machine learning models. 2. And for $GG_{\beta,\sigma}(f, D)$ mechanisms with ℓ_β sensitivity (the small subset of mechanisms with dimension-independent privacy accounting), there is minimal or no improvement in model utility compared to the Gaussian mechanism.

1.1 Related Work

A number of prior works explored alternative DP mechanisms that extend the Laplace and Gaussian Mechanisms. The Staircase Mechanism [Geng and Viswanath, 2014] was derived as an alternative to the Laplace Mechanism by minimizing the variance of the noise distribution in order to improve utility guarantees. The Podium Mechanism [Pihur, 2019] improved upon the Staircase Mechanism by changing the noise distribution to be defined over finite (truncated) support, instead of the infinite support used in other mechanisms like Laplace. While both mechanisms have rigorous analysis of utility in the single-shot regime, they are neither studied nor optimized for high-composition regimes, and thus are not used in ML, which is the key use-case we investigate.

Alghamdi et al. [2022] developed the Cactus Mechanism to specifically address the high-composition regime by numerically computing a mechanism that minimizes the Kullback-Leibler divergence between the conditional output

[2]Sometimes referred to in the literature as the Exponential Power Distribution or Generalized Normal Distribution

distributions of a mechanism given two different inputs, under a high number of compositions. Alghamdi et al. [2022] derived a near-optimal divergence loss in the high-composition regime; however, they only considered privacy for 1-dimensional outputs, making the result not directly applicable to ML models. Further, they focused on the optimal DP mechanism in the limit of a large number of compositions, explicitly ignoring the low-composition regime.

Liu [2019] introduced and partially analyzed the Generalized Gaussian Mechanism, which adds noise from the Generalized Gaussian distribution (Definition 8). They provided probabilistic-DP (a variation of approximate DP) guarantees for $\beta = 1$ and $\beta \geq 2$, and gave empirical results for integer β-values in the GG mechanism applied to training Support Vector Machines on sanitized data from several tabular datasets. In a separate vein of research, Liu et al. [2022], provides a relatively weak bound for the Rényi Differential Privacy (RDP, Definition 7) when $\beta > 2$, for the limited use case of $\alpha = 1$, which restricts much of the usefulness of the RDP formulation. To best of our knowledge, no previous work provides tight (ϵ, δ)-DP guarantees that are useful for the high composition regime. In our work, we consider all $\beta \geq 1$, and explore the privacy-accuracy tradeoff of the GG mechanism and its applications to PATE and DP-SGD, and empirically provide tighter privacy guarantees than those provided by prior works, for any number of compositions.

Awan and Dong [2024] address the multivariate setting by deriving a family of multivariate log-concave canonical noise distributions [Awan and Vadhan, 2023], allowing addition of minimal noise for a particular f-DP guarantee. However, their work is constrained to ℓ_1 and ℓ_∞ sensitivity settings and does not analyze the privacy-utility tradeoff of these mechanisms for private ML. Unfortunately, the proposed mechanisms are either only for Gaussian DP (known to underestimate privacy [Gopi et al., 2024]), or are not easily integrated with existing privacy accountants in the subsampled regime.

1.2 Our Contributions

In this work, we investigate the Generalized Gaussian Mechanism, a little-explored family of mechanisms that satisfies DP. We also introduce the Sampled Generalized Gaussian Mechanism (SGG), which is a variant of the GG-Mechanism that first subsamples the database to make use of privacy amplification. In particular, we have the following contributions:

1. **GG mechanism in ML**: In Section 4, we introduce the GGNMax algorithm for computing a private argmax, and in Section 5, we show how to use the Sampled GG Mechanism for DP-SGD in a new mechanism that we name β-Differentially Private Stochastic Gradient Descent (β-DP-SGD). We show that all 4 mechanisms satisfy differential privacy (Theorems 1, 5, 2, and 3). We also show how to extend an existing privacy accountant (the PRV accountant) to track privacy budget over many compositions of these mechanisms.
2. **Utility of GG mechanism for ML**: For guidance on optimally choosing β, in Section 4, we empirically find that in privately computing an argmax in PATE, the choice of β has a weak relationship with the accuracy. In Section 5, we empirically find that β-DP-SGD performs similarly for most $\beta \in [1, 4]$ when hyperparameters are tuned based on the choice of β, and that the Gaussian mechanism ($\beta = 2$) is nearly optimal, as it performs approximately as well as the best choice of β under full hyperparameter tuning. Small improvements in performance do still exist, typically around the neighborhood of $\beta \approx 2$, which can be valuable when small improvements are critical; the effect of β is more noticeable when the learning algorithm is constrained, and specific hyperparameters (like learning rate η or batch-size) are fixed.
3. **Numerical accountant for arbitrary mechanisms**: Previously, explorations of new mechanisms were confined to those noise distributions that exhibited well-behaved mathematical properties (such as analytically derivable Rényi Divergence values between distributions); this work develops a robust framework for using the PRV accountant [Gopi et al., 2024] to explore new mechanisms and to apply them to private ML tasks. This method enables new directions for future research in novel DP mechanisms that cannot be directly analyzed analytically.
4. **Privacy accounting for GG mechanism is dimension independent**: We find that privacy accounting is dimension-independent $GG_{\beta,\sigma}(f, D)$ mechanisms with ℓ_β sensitivity (e.g. the Gaussian mechanism with ℓ_2 sensitivity). While we do not prove the converse, we suspect that all other mechanisms have dimension-dependent privacy accounting, making it very computationally challenging for high-dimensional ML models. We introduce this in Section 3.2, we prove that the Generalized Gaussian mechanism with ℓ_β sensitivity has dimension independent privacy accounting in Appendix B.2, and we provide an analytic solution for the one dimensional Generalized Gaussian mechanism in Appendix B.1.

Thus, we identified two key negative findings: 1. Privacy accounting is dimension-dependent for most mechanisms, making it computationally challenging for high-dimensional ML models. 2. For the few mechanisms with dimension-independent privacy accounting, such as $GG_{\beta,\sigma}(f, D)$, there is little to no utility improvement over the Gaussian

mechanism. This suggests that advancing private ML will likely require a conceptual breakthrough rather than new privacy mechanisms.

2 Preliminaries

2.1 Differential Privacy

Differential privacy (DP) is a framework for designing privacy-preserving data analysis algorithms that protect the privacy of individuals in a dataset while allowing accurate statistical analysis. Informally, DP provides a mathematical guarantee that an individual's data will have only a limited affect on the result of analysis on a large database. Two datasets are said to be *neighboring* if they differ only in a single data record.

Definition 1 (Differential privacy [Dwork et al., 2006])**.** *A mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for any two neighboring datasets $D, D' \in \mathcal{D}$ and for any $S \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta.$$

Smaller values of the parameters ϵ and δ correspond to stronger privacy guarantees. The Laplace and Gaussian Mechanisms are examples of *output perturbation* mechanisms, which first evaluate a function on the input dataset, and then add mean-zero noise to the result. The variance of the noise scales with the *sensitivity* of the function Δf, defined as the maximum change in the function's value due to the removal or addition of a single database entry.

Definition 2 (ℓ_β sensitivity [Dwork and Roth, 2014])**.** *The ℓ_β-sensitivity of a function $f : \mathbb{N}^{|x|} \to \mathbb{R}^k$ is:*

$$\Delta_\beta(f) = \max_{D, D' \text{ neighbors}} \|f(D) - f(D')\|_\beta.$$

We emphasize that while differently-normed sensitivity denote that the function is bounded for a different ℓ_β norm, the definition of a neighboring dataset remains the same – two datasets are neighboring if they differ by a single entry.

Two of the most common mechanisms in differential privacy are the Laplace Mechanism and the Gaussian Mechanism.

Definition 3. *The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function:*

$$Lap(x|b) = \frac{1}{2b} \exp(-\frac{|x|}{b}).$$

Definition 4 (Laplace Mechanism [Dwork et al., 2006])**.** *For any $\epsilon > 0$, given a real-valued function $f : \mathcal{D} \to \mathbb{R}$, the Laplace mechanism is defined as*

$$\mathcal{M}_L(D, f, \epsilon) = f(D) + Y,$$

where $Y \sim Lap(\Delta f/\epsilon)$. The Laplace Mechanism is $(\epsilon, 0)$-DP.

Definition 5 (Gaussian Mechanism [Dwork and Roth, 2014])**.** *For any $\epsilon > 0$ and $\delta \in (0, 1]$, given a real-valued function $f : \mathcal{D} \to \mathbb{R}$, the Gaussian mechanism is defined as*

$$\mathcal{M}_G(D, f, \epsilon) = f(D) + Y,$$

where $Y \sim \mathcal{N}(0, \sigma)$ for $\sigma > \Delta f \sqrt{2 \log(1.25/\delta)}/\epsilon$. The Gaussian Mechanism is (ϵ, δ)-DP.

One main feature of differential privacy is that the guarantees *compose*, meaning that the overall privacy loss (as measured by ϵ and δ) of running multiple DP mechanisms can be bounded as a function of the privacy parameters of the individual mechanisms. In the simplest version of composition [Dwork et al., 2006], the privacy parameters "add up" so that running two (ϵ, δ)-DP mechanisms results in $(2\epsilon, 2\delta)$-DP overall. In practice, however, this naive composition dramatically overestimates the incurred privacy risk, and more advanced composition algorithms [Dwork et al., 2010] and privacy accountants [Abadi et al., 2016] are used to more accurately bound privacy risk.

A privacy accountant is a tool to track the privacy budget of a system by recording the privacy cost associated with each query; accountants are particularly important for applications like DP-SGD, where DP mechanisms are composed a large number of times, e.g., as many as the number of steps in gradient descent. The introduction of the Moments Accountant [Abadi et al., 2016] enabled the first use of DP-SGD with reasonable privacy guarantees on common datasets like MNIST [Lecun et al., 1998]. This was later replaced in many settings by accountants that rely on Renyi Differential Privacy (RDP), introduced by Mironov [2017].

2.2 Rényi Differential Privacy

Rényi Differential Privacy (RDP) generalizes pure differential privacy ($\delta = 0$) and is closely related to the moments accountant. Defined below, the RDP guarantee of a mechanism is stated in terms of Rényi divergence.

Definition 6 (Rényi Divergence)**.** *The Rényi divergence of order α between two distributions P and Q is defined as:*

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim Q} \left[(P(x)/Q(x))^\alpha\right] = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim P} \left[(P(x)/Q(x))^{\alpha-1}\right].$$

Definition 7. *(Rényi Differential Privacy [Mironov, 2017]). A randomized mechanism $\mathcal{M}$ satisfies (α, ϵ)-RDP with $\alpha \geq 1$ if for any neighboring datasets D and D':*

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim\mathcal{M}(D)} \left[\left(\frac{\Pr[\mathcal{M}(D) = x]}{\Pr[\mathcal{M}(D') = x]}\right)^{\alpha-1}\right] \leq \epsilon.$$

RDP is desirable in ML applications because of its straightforward composition properties: the adaptive composition of mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where each $\mathcal{M}_i$ satisfies (α, ϵ_i)-RDP, will together satisfy $(\alpha, \sum_{i=1}^k \epsilon_i)$-RDP.

Pure $(\epsilon, 0)$-DP corresponds to (∞, ϵ)-RDP; Mironov [2017] provided more general guarantees for converting between RDP and DP: if $\mathcal{M}$ is an (α, ϵ)-RDP mechanism, it also satisfies $(\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $\delta \in (0, 1)$.

2.3 PRV Accountant

The privacy guarantees of a DP mechanism are commonly defined as a single tuple (ϵ, δ), but can also be described as a function, since the probability of failure (δ) can depend on the required ϵ-bound. This naturally leads to the definition of a privacy curve, such that for every $\epsilon \in \mathbb{R}$, $\mathcal{M}$ is $(\epsilon, \delta(\epsilon))$-DP for the appropriate function $\delta(\epsilon)$. Gopi et al. [2024] provided an efficient method for composing privacy curves directly that gave much tighter privacy guarantees, using an accountant called the Privacy Random Variable accountant (PRV). This relies on a connection between a DP mechanism's privacy curve $\delta(\epsilon)$ and its uniquely defined privacy loss random variables (X, Y), which represent the likelihood of returning a particular outcome on two neighboring databases, respectively defined as:

$$X = \log(\tfrac{Q(\omega)}{P(\omega)}) \text{ where } \omega \sim P; \quad Y = \log(\tfrac{Q(\omega)}{P(\omega)}) \text{ where } \omega \sim Q,$$

where *P* and *Q* are the distribution of the mechanism's output over two neighboring datasets. Intuitively, the privacy loss random variables can be thought of as the *actual* ϵ value for a specific output; it is a random variable because the output $\mathcal{M}(D)$ is itself a random function.

Gopi et al. [2024] introduced the algorithm ComposePRV, which efficiently computes the privacy guarantees for the composition of multiple DP mechanisms. ComposePRV takes as input the CDFs of PRVs $Y_1, \ldots, Y_k$ (as well as a few other hyperparameters), and returns an estimate of the privacy curve for all the mechanisms composed, represented by $\delta(\epsilon)$, allowing for the direct computation of ϵ.

3 Generalized Gaussian Mechanism and Privacy Guarantees

In this section, we first introduce the Generalized Gaussian (GG) Mechanism and show that it satisfies DP (Section 3.1). Since these privacy results are existential, rather than descriptive – i.e., we show that *there exists* some ϵ and δ values, rather than providing a closed form relationship between (β, σ) and (ϵ, δ) – we also present a PRV-based privacy accounting method (Section 3.2) that can be used to measure explicit (ϵ, δ)-DP guarantees in the applications of this mechanism to ML tasks, as studied in Sections 4 and 5.

3.1 Generalized Gaussian Mechanism

We first formally define the Generalized Gaussian distribution and introduce the Generalized Gaussian Mechanism (Algorithm 1), which is an output perturbation mechanism that adds noise sampled from the Generalized Gaussian distribution.

Definition 8 ([Dytso et al., 2018])**.** *The Generalized Gaussian distribution, denoted $\mathcal{N}_\beta(\mu, \sigma)$, is specified by the pdf $p(x|\mu, \sigma, \beta) \propto e^{-\frac{|x-\mu|^\beta}{\sigma}}$ with normalizing constant $\frac{\beta}{2\sigma^{\frac{1}{\beta}}\Gamma(\frac{1}{\beta})}$.*

The Generalized Gaussian Mechanism was introduced by [Ganesh and Zhao, 2020] for ℓ_1 sensitivity and by [Liu, 2019] for ℓ_β sensitivity; we use the latter, more general version.

Algorithm 1 Generalized Gaussian Mechanism, $GG_{\beta,\sigma}(f, D)$. [Liu, 2019]

Input: noise parameters $\beta \geq 1$, $\sigma > 0$, vector-valued function $f : \mathcal{D} \to \mathbb{R}^d$, database $D \in \mathcal{D}$
Let $\Delta_\beta f = \max_{D,D' \text{ neighbors}} \|f(D) - f(D')\|_\beta$
for $i = 1$ to d **do**
    Sample $Y_i \sim \mathcal{N}_\beta(0, \sigma \cdot \Delta_\beta f)$
end for
Output: $f(D) + (Y_1, \ldots Y_d)$

Next, we show that the GG Mechanism satisfies DP (Theorem 1). As stated, the result is existential, and does not provide an explicit relationship between the (β, σ) noise parameters of the mechanism and the resulting DP parameters (ϵ, δ). Critically, this guarantee that the mechanism is DP for *some* ϵ and δ is sufficient to apply the PRV accountant described in Section 3.2.

Theorem 1. *For any $\beta \geq 1$, $\sigma > 0$, $\delta > 0$ there exists a finite value ϵ such that $GG_{\beta,\sigma}(\cdot, \cdot)$ satisfies (ϵ, δ)-DP.*

To prove Theorem 1, we first show that for all $\mu \geq 0$ (corresponding to the difference in the value of f on neighboring databases) and all $\alpha > 1$, the α-Rényi divergence between $\mathcal{N}_\beta(0, \sigma)$ and $\mathcal{N}_\beta(\mu, \sigma)$ is bounded by some finite r; this step is shown formally in Lemma 14 in Appendix C.1. Bounded Rényi divergence means that the $GG_{\beta,\sigma}(f, D)$ mechanism satisfies (α, r)-RDP, and by the RDP-to-DP conversion of Mironov [2017], it also satisfies $(r + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any $\delta \in (0, 1)$. A formal proof of Theorem 1 is given in Appendix C.2.

In Appendix A, we also introduce and analyze the Sampled Generalized Gaussian Mechanism, SGG (Algorithm 4), which is a variant of the GG Mechanism that first applies Poisson subsampling to the input database, evaluates the function f on the sample, and then adds Generalized Gaussian noise to the result. This mechanism is motivated by *privacy amplification by subsampling*, which strengthens privacy guarantees without increasing the level of noise added by the mechanism, by subsampling the database before applying a DP mechanism. It is particularly popular in ML applications, such as the Sampled Gaussian Mechanism (SGM) [Mironov et al., 2019] and DP-SGD, because it formalizes the privacy gains intuitively brought by computing each update on the small subset of training examples selected to form each minibatch.

3.2 Privacy Accounting for GG Mechanisms

We focus on the PRV accountant because it is implemented in common codebases such as Opacus [Yousefpour et al., 2021] and it empirically provides tighter guarantees than other accountants [Gopi et al., 2024]. In this work we extend the PRV accountant to work for privacy accounting of arbitrary DP mechanisms such as the GG mechanism, which do not typically exist in closed-form, providing a framework for investigating alternative DP mechanisms. In order to calculate the privacy consumed using a PRV accountant, one must simply compute the CDF of the PRV. In Appendix B.1, we extend the known PRVs for Laplace and Gaussian Mechanisms and compute a closed-form expression for the PRV of the Generalized Gaussian Mechanism, which enables us to apply the PRV accountant.

We compute the privacy guarantees of $GG_{\beta,\sigma}(f, D)$ using the PRV accountant by sampling from the appropriate $\mathcal{N}_\beta(\mu, \sigma)$ distribution, and numerically computing the CDF of Y. [Gopi et al., 2024] provides a tight estimate of the error in the PRV accountant's estimate; our work relies on this computation to provide a similarly tight error analysis bounding the contribution of error from sampling to the estimate, which is included in Appendix B.3. We then plug in this CDF as input to the ComposePRV algorithm of Gopi et al. [2024], which takes the CDFs of the PRVs of the composed mechanisms, and returns a composed privacy curve $\delta(\epsilon)$, providing the unique ϵ value for a specified choice of δ. The implementation of the PRV accountant is described in greater detail in Appendix B.3.

Figure 2 illustrates the resulting value of ϵ as a function of σ for different values of β and fixed values of $\delta = 1e - 5$ and $\Delta f = 1$. Observe that all curves have a similar structural form to the known Gaussian ($\beta = 2$) and Laplace ($\beta = 1$) Mechanisms, and that as β grows, the same (ϵ, δ)-DP guarantee necessitates a larger value of σ. In the figure, we show the privacy curves for a single composition of the GG mechanism, but importantly, these privacy curves and their relative differences change as a function of the number of times the mechanisms are composed.

Remarkably, when using sensitivity defined in the ℓ_β norm we observe that the privacy cost of the multi-dimensional mechanism is equivalent to the single dimension mechanism. This makes privacy accounting for $GG_{\beta,\sigma}(f, D)$ dimension independent when some ℓ_β sensitivity is used. We prove this in result Appendix B.2, and also provide an

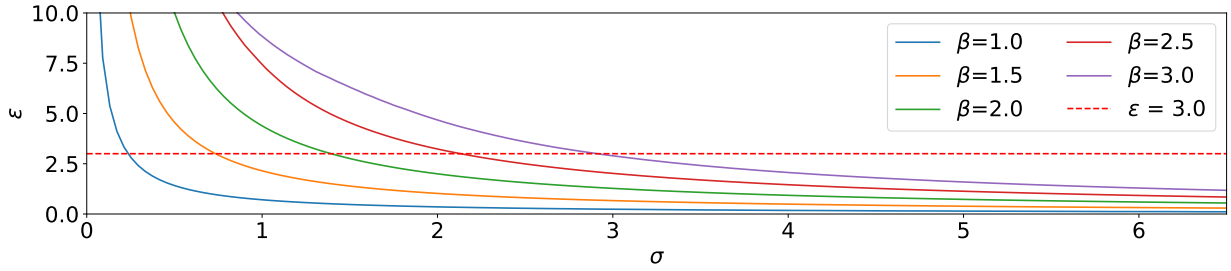

Figure 2: DP parameter ϵ as a function of noise parameter σ with fixed $\delta = 10^{-5}$ and $\Delta f = 1$, calculated using the PRV accountant. Mechanisms with equivalent DP guarantees can be identified by computing the privacy curve's intersection with a horizontal line, illustrated here with a red line for (arbitrarily chosen) $\epsilon = 4$.

analytic solution for the one dimensional GG mechanism in Appendix B.1. This dimension-independence is crucial for machine learning, as it allows us consider numerical, sampling-based approaches. If privacy accounting depended on dimension, sampling numerically would be computationally infeasible, as even small machine learning models regularly have over 1,000,000 parameters (dimensions), requiring us to sample 1,000,000+ distributions millions of times.

For a given privacy budget (ϵ, δ), varying β will change the paired σ and subsequently vary the weight of the tails of the distributions. As a result, for a particular definition of "tail", it is possible to derive the optimal GG mechanism that minimizes the likelihood of an outlier; we explore this in Appendix B.5 and believe it is of independent interest.

4 GG Mechanism for PATE and Private Argmax

Private Aggregation of Teacher Ensembles (PATE) [Papernot et al., 2017] is an algorithm for training a private machine learning model. In PATE, a dataset is partitioned and a model is trained on each partition, then the models in the ensemble privately vote on the labels of an unlabeled dataset, and finally a model is trained on the privately labeled dataset. A more detailed explanation is presented in Appendix D. The core step in PATE that provides formal privacy is the private computation of the Argmax over the votes of an ensemble of models, returning the plurality of a noisy histogram of votes. This is done based on a variation of the ReportNoisyMax algorithm, and is the center of our focus in this section.

PATE has been extensively studied with variations of Laplace and Gaussian Mechanisms; Papernot et al. [2017] employed LNMax, which privately aggregates votes from an ensemble of models by taking the argmax of a histogram after adding Laplace noise. Later, Papernot et al. [2018] developed several variations based on adding Gaussian noise, including GNMax, and empirically found them to be superior to their Laplace counterpart. We introduce a new algorithm, GGNMax, generalizing the GNMax and LNMax algorithms, which adds noise from the Generalized Gaussian Distribution $\mathcal{N}_\beta(0, \sigma)$. We show that the effect of β on average label accuracy is relatively weak and Laplace and Gaussian noise work produce nearly equivalent privacy-accuracy tradeoffs. We supplement these findings with simulations in Appendix D.1, and we empirically show that the Gaussian mechanism is near-optimal when the correct label of the histogram aligns with the non-private majority vote.

4.1 Private Argmax and the GGNMax Mechanism

We present our Generalized Gaussian Private Argmax algorithm (GGNMax), which takes in noise parameters (β, σ), a set of real-valued functions with sensitivity Δ, and a database. The algorithm adds noise sampled from $\mathcal{N}_\beta(0, \sigma\Delta)$ to each function value and then returns the index of the function with the largest noisy value.

Algorithm 2 Generalized Gaussian Private Argmax, GGNMax$(\beta, \sigma, \{f\}, \Delta, D)$

1: **Input** noise parameters $\beta \geq 1$, $\sigma > 0$, functions $f_1, \ldots, f_N : \mathcal{D} \to \mathbb{R}$ each of sensitivity Δ, database $D \in \mathcal{D}$
2: **for** $i = 1$ to N **do**
3: Compute $f_i(D)$
4: Sample $Y_i \sim \mathcal{N}_\beta(0, \sigma\Delta)$
5: **end for**
6: **Output** $\arg\max_{i \in [N]} \{f_i(D) + Y_i\}$

The privacy guarantees of GGNMax rely on the privacy of the GG Mechanism, which is used as a subroutine. Note that the GG Mechanism is known to be private for *some* (ϵ, δ) pair (Theorem 1), so this theorem relates the parameters of the two mechanisms.

Theorem 2. *If the (β, σ)-Generalized Gaussian Mechanism is (ϵ, δ)-DP for a fixed $\epsilon > 0$ and $\delta \geq 0$, then (β, σ)-Generalized Gaussian Private Argmax is also (ϵ, δ)-DP.*

A proof of Theorem 2 is given in Appendix C.3. Note that as with the Laplace-based Report Noisy Max algorithm [Dwork and Roth, 2014] and the Gaussian-based GNMax algorithm [Papernot et al., 2018], the privacy guarantee does not depend on number of functions N. This makes it a desirable method for computing the argmax of the output of the GG Mechanism, which consumes privacy proportional to the number of queries; this property allows GGNMax to have privacy-accuracy tradeoffs that scale well in high dimensions.

4.2 PATE Experiments

As an intermediate step in PATE, the algorithm produces a histogram of votes in order to privately label an unlabeled dataset.

In order to isolate the effect of the specific privacy mechanism used for PATE, we focus our attention to study the effect of the β parameter on the label accuracy of the private aggregation step in PATE. To study this for a realistic setting, we use the histograms generated on the MNIST and the Street View House Numbers (SVHN) dataset [Netzer et al., 2011], produced by Papernot et al. [2017], which introduced the PATE algorithm. For each of these datasets, we start with a collection of $10,000$ histograms; each histogram is the collection of 250 models trained on a partition of the dataset, and evaluated on an unlabeled datapoint x. Then, for each histogram, we compute the private label produced by the GGNMax mechanism for 20 evenly spaced values $\beta \in [1, 4]$ and 200 values of $\sigma \in [0.01, 7]$. For each fixed (ϵ, δ) and value of β, we compute the average label accuracy with respect to the ground truth labels provided by the dataset, averaged across 25 trials for each datapoint. [3] [4]

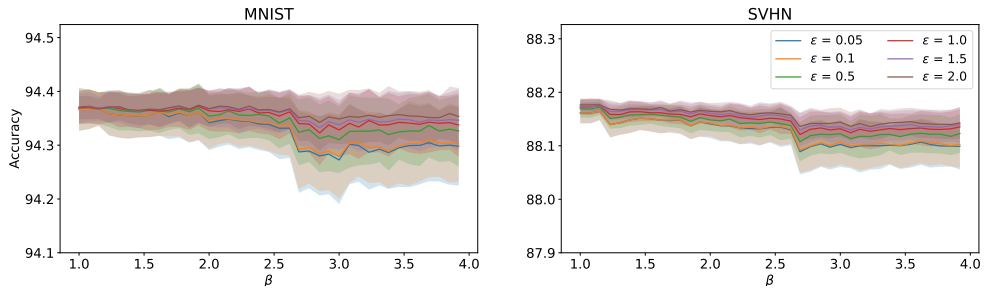

Figure 3: Average Label Accuracy of GGNMax mechanisms with equivalent privacy guarantees and varying values of β, evaluated on histograms which were generated by Papernot et al. [2017] as an intermediate state produced by 250 teachers trained on MNIST (left) and SVHN (right).

Figure 3 shows the average label accuracy for the GGNMax mechanism applied to histograms generated as part of PATE for MNIST and SVHN, as a function of β. We observe that values of β in the region of $\beta \in [1, 2.5]$ perform roughly equivalently, and for larger choices of β, the label accuracy decreases slightly.[5]

To further extend these findings to more settings, we investigate the privacy-accuracy tradeoff of the GGNMax algorithm on simulated histograms in Appendix D.1. We find that when the correct label is the argmax of the unnoised histograms, the GGNMax with β values close to 2 perform near-optimally.

[3] We find that changing the number of trials has minimal effect on the mean or standard deviation of the average accuracy

[4] We note that we use sensitivity ℓ_β for the GGNMax mechanism with $\mathcal{N}_\beta$ noise - which varies across trials. This is to achieve simpler privacy accounting (as discussed in section 3.2. However, we argue this is not problematic because the sensitivity of a histogram is adding-or-removing a single element, and the ell_p norm of a 1-hot vector is 1, for any value of p.

[5] Empirically, we observe a small drop in performance around $\beta = 2.6$, which may be due to an artifact of how the mechanisms of equivalent DP guarantees are generated. First we add noise from an evenly distributed grid of $\beta \in [1, 4]$ and $\sigma \in [0.01, 7]$ values, and then we compute the privacy for those (β, σ) tuples and compute the corresponding mechanisms of approximately equal DP guarantees. This may cause us to overestimate the ϵ when solving for mechanisms with equivalent DP guarantees. We find nothing particularly unique about the value $\beta = 2.6$.

5 GG Mechanism for Differentially Private Stochastic Gradient Descent

We now turn to applications of the GG Mechanism in DP-SGD, which is one of the most commonly used mechanisms for private ML. We propose two simple changes to DP-SGD: replacing the Gaussian noise used in DP-SGD with Generalized Gaussian noise, and using the ℓ_β norm for clipping rather than ℓ_2[6] As articulated in Section 2.1, this change in sensitivity measure does not change the notion of neighboring databases, so optimization over β shoulve be viewed as additional hyperparameter tuning to improve performance, rather than changing the nature of the privacy guarantees.

We call the resulting mechanism the β-Generalized Gaussian Differentially Private SGD (β-DP-SGD). This is also equivalent to changing the underlying mechanism in DP-SGD from the Sampled Gaussian Mechanism to the Sampled Generalized Gaussian Mechanism (SGG), as DP-SGD performs Poisson subsampling on the dataset before computing the gradient update and adding Gaussian noise. The β-DP-SGD algorithm is presented formally in Algorithm 3.

rcsomewhere describe the role of the accountant vs T in text here. as written, the algorithm looks like it takes in a number of rounds T as input, but this is actually computed online via privacy accountant. this leads to confusion when presenting the results bc you don't report T.

Algorithm 3 β-Generalized Gaussian Differentially Private SGD, β-DP-SGD($\beta, \sigma, D, l, \eta, L, C, T$)

Input: noise parameters $\beta \geq 1$, $\sigma > 0$, database $D = \{x_1, \ldots, x_N\}$ of points in $\mathbb{R}^d$, loss function $l(\theta, x_i)$, learning rate η, average group size L, clip norm C, and training epoch length T.
Initialize $\theta_1 \in \mathbb{R}^d$ randomly
for $t = 1$ to T **do**
    Construct $L_t \subseteq D$ such that each $x_i \in D$ is included with probability $q = L/|D|$ (Poisson sampling)
    for each $i \in L_t$ **do**
        Compute $G_t(x_i) = \nabla_{\theta_t} l(\theta_t, x_i)$
        $\bar{G}_t(x_i) \leftarrow G_t(x_i) / \max\left(1, \frac{\|G_t(x_i)\|_\beta}{C}\right)$
    end for
    Sample $Y_1, \ldots, Y_d \sim_{i.i.d.} \mathcal{N}_\beta(0, \sigma \cdot C)$
    $\tilde{G}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{G}_t(x_i) + \vec{Y}\right)$
    $\theta_{t+1} \leftarrow \theta_t - \eta \tilde{G}_t$
end for
Output: θ_T

The following theorem states that β-DP-SGD is differentially private, which we prove in Appendix A.2.

Theorem 3. *For any $\delta > 0$, $\beta \geq 1$, $\sigma > 0$, $f : \mathcal{D} \to \mathbb{R}^d$, database $D \in \mathcal{D}$, for any loss function of the form $l(\theta, x_i)$, learning rate $\eta \geq 0$, average group size L, clipping norm $C \geq 0$, there exists $\epsilon > 0$ such that the algorithm β-DP-SGD$(\beta, \sigma, D, l, \eta, L, C)$ satisfies (ϵ, δ)-DP.*

5.1 DP-SGD Experiments

We seek to find a relationship between β and DP-SGD's privacy-accuracy trade-off for non-convex optimization tasks by comparing test-accuracy as a function of ϵ, for different β.

To provide a robust evaluation of the role of β in the β-DP-SGD algorithm, we focus on 4 datasets in different domains: CIFAR-10 [Krizhevsky, 2009] and Street View House Numbers (SVHN) [Netzer et al., 2011], two common computer vision datasets; the Adult dataset [Becker and Kohavi, 1996], a tabular dataset with a binary classification task; and the IMDB dataset [Maas et al., 2011], a collection of movie reviews meant for binary sentiment classification. We train four different architectures: for the vision classification tasks, we use the models described in Tramèr and Boneh [2021], which previously achieved SOTA results for the $\epsilon \leq\sim 2.5$ regime (ScatterNet CNNs) For the the Adult Dataset we train a 2-layer Fully Connected Network (FCN), and for the IMDB dataset, we train a Long-Short Term Memory (LSTM) network with $\sim$ 1M parameters.

A full description of the hyperparameters, datasets, and models is included in Appendix E.1. Each experiment is run 3 times, which we found sufficient given standard deviations that generally fell below 0.3%.[7] We find that when fixing a

[6] We choose to use ℓ_β clipping rather than a fixed choice like ℓ_2 because when using the $GG_{\beta,\sigma}(f, D)$ mechanism with ℓ_β sensitivity, privacy accounting is dimension-independent. See Appendix B.2.

[7] Although 3 iterations would typically be considered a small number, each run of the mechanism is already aggregating over T rounds of the GG mechanism and corresponding gradient updates.

choice of β and allowing for hyperparameter tuning along all other hyperparameters, we see a weak but noticeable relationship with final test accuracy.

Results: We report the maximum test-accuracy for each dataset, i.e., the maximum across all hyperparameters under the fixed architecture, and the standard deviation across 3 trials for each set of hyperparameters. We observe a relatively weak relationship between β and test-accuracy, although this relationship is more noticeable in lower ϵ regimes (high privacy).

Figure 4 presents the maximum test-accuracy for 3 different values of ϵ, evaluated for 3 different models on 4 different datasets (ScatterNet on CIFAR-10 and SVHN, a FCN on the Adult dataset, and an LSTM on the IMDB dataset), across different values of β. Similar to our results with the GGNMax, we find that for all ϵ values tested, the choice of β has a weak relationship with test-accuracy across most values of ϵ. However, unlike the results in Section 4, β-DP-SGD seems to perform worse for larger values of β, particularly for values larger than $\beta \geq 3.0$. In Appendix E.2, we explore the relationship of individual hyperparameters with β and find a weak, but more noticeable effect of β on the final test-accuracy, particularly for larger ϵ.

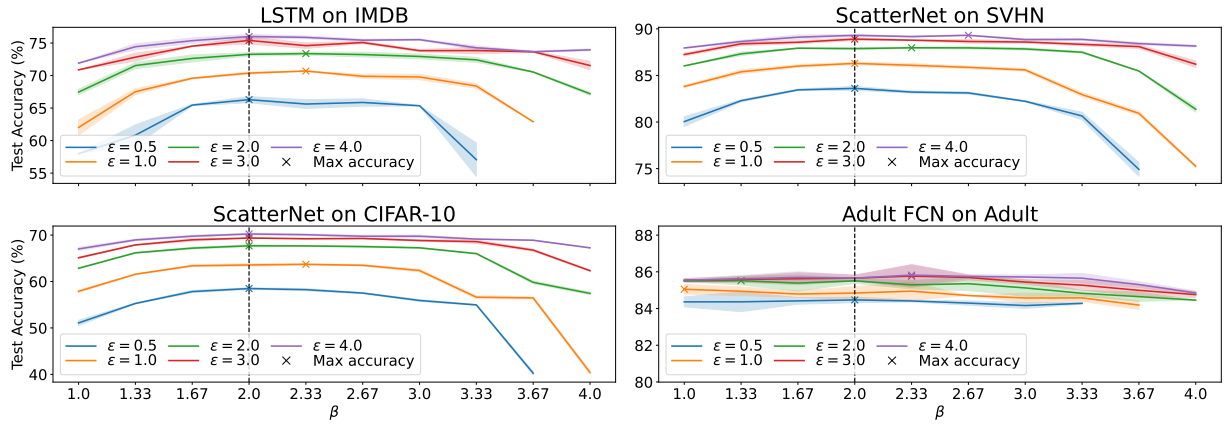

Figure 4: β-DP-SGD results for the corresponding architectures trained on CIFAR-10, SVHN, Adult, and IMDB, for $\delta = 10^{-6}$. The test-accuracy is reported for 3 values of ϵ, computed for each architecture-dataset pair. A vertical dashed line denotes the Gaussian mechanism. Note: Some values are not presented (for lower ϵ), because larger β tends to consume more privacy per step, and the model's privacy budget exceeds the target in the first step.

Table 1 provides numerical results on the performance of β-DP-SGD compared to that of regular DP-SGD across different ϵ values for the example case of CIFAR-10. We observe that despite a relatively weak relationship between β and test-accuracy, β-DP-SGD is able to reach – and slightly surpass – SOTA results. Most existing SOTA results use privacy guarantees provided by an RDP accountant, which can overestimate privacy loss relative to PRV accounting. In order to disambiguate empirical differences due to improved accounting versus the GG mechanism, we recreate existing SOTA results using both the PRV and RDP accountants. As expected, we see that methods using the PRV accountant outperform those using the RDP accountant.

The comparison of β-DP-SGD versus DP-SGD in Table 1 reveals that some minor improvements in performance result from optimizing β, although these must be traded off against the privacy cost of hyperparameter search [Papernot and Steinke, 2022].

6 Conclusion

We studied the Generalized Gaussian Mechanism, its privacy guarantees, and its applications to private ML, particularly for PATE (via private Argmax) and DP-SGD. This work reveals that the choice of β has a relatively modest influence on test accuracy, and the difference between Gaussian, Laplace, and other GG Mechanisms is smaller than anticipated. Interestingly, values close to $\beta = 2$ exhibit near-optimal performance, which provides insight into the popularity of Gaussian noise in DP-SGD, PATE, and other private ML applications. Our observations that the Gaussian is not always exactly optimal suggests new opportunities for the design of DP mechanisms.

We admit that given the relatively little existing results on alternative DP mechanisms for private ML, upon beginning this work we expected that exploring alternative DP mechanisms for private ML would yield greater gains. To us at

Accountant	Training Algorithm	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	β value
RDP	DP-SGD	60.3 (-)	67.2 (-)	69.3 (-)	-
PRV	DP-SGD	63.5 (0.4)	67.6 (0.1)	69.2 (0.3)	-
PRV	β-DP-SGD (*ours*)	**63.7** (0.1)	**67.7** (0.1)	69.4 (0.1)	(2.33, 2.33, 2)

Table 1: SOTA Results for private ML, evaluated on CIFAR-10, grouped by privacy accountant. The model used is a CNN model ($\sim$ 5e5 parameters) trained on Scatternet features, which we refer to as ScatterNet. The β value column is set to '-' if trained with traditional DP-SGD, otherwise it reports a tuple of β values that achieve max accuracy for $\epsilon = 1, 2, 3$, respectively (ties broken by smaller std values). We bold the experiments that achieve SOTA results. Note; SOTA results for $\epsilon = 3$ are achieved by a WRN-40-4 (a Wide Resnet model with $\sim$ 1e7 parameters), which achieves (56.4 (0.6), 65.9 (0.5), **70.7** (0.2)) for ϵ $(1, 2, 3)$ respectively. We exclude WRN-40-4 from our experiments due to difficulty recreating their results in our regime with the PRV accountant

least, our mostly negative results are in fact quite surprising and warrant consideration by the field at large. While there is of course room for new research, we believe that this work partially closes the door on alternative DP mechanisms for DPSGD, at least until new privacy accountants are introduced.

This is because of our finding that the GG mechanism with ℓ_β sensitivity has privacy that is dimension independent (in Appendix B.2); otherwise the sampled-PRV accountant requires $D \cdot N$ samples from the GG distribution, where D is the dimension of the neural network, and N is the number of empirical samples required to accurately approximate a distribution, (for our work, this is on the order of $1e7$).

This finding about dimension-independence constrains explorations of alternative mechanism to the family of GG mechanism, because we believe (but do not prove) that alternative mechanisms do not have this quality of dimension-independence. This observation, coupled with the results in Section 5.1, that we see little to no utility improvement over the Gaussian mechanism with a change in β, seems to provide a mostly negative result that alternative mechanisms will provide much gains in DPSGD and private ML.

However, this field may open back up if a new privacy accountant is introduced which can measure the privacy of arbitrary DP mechanisms without sampling.

An interesting extension for future work is that our GG Mechanism – as well as its variants SGG, GGNmax, and β-DP-SGD – sample noise independently across dimensions. For the Laplace Mechanism ($\beta = 1$), it is known that sampling from a high-dimensional Laplace variant can improve performance in private ML settings such as private empirical risk minimization [Chaudhuri et al., 2011]. Multi-dimensional Gaussian distributions are the only spherically symmetric distribution where all the component random variables are independent Ali [1980], so such high-dimensional variants would not improve performance for $\beta = 2$. This suggests that for $\beta \in [1, 2)$, it may be possible to improve utility for the same privacy guarantee by sampling from a single high-dimensional distribution rather than sampling independently for each coordinate (see Appendix B.2).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318. URL `https://doi.org/10.1145%2F2976749.2978318`.

John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 census disclosure avoidance system topdown algorithm, 2022.

Wael Alghamdi, Shahab Asoodeh, Flavio P. Calmon, Oliver Kosut, Lalitha Sankar, and Fei Wei. Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, June 2022. doi: 10.1109/isit50566.2022.9834438. URL `http://dx.doi.org/10.1109/ISIT50566.2022.9834438`.

Mir M. Ali. Characterization of the normal distribution among the continuous symmetric spherical class. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):162–164, 1980. ISSN 00359246. URL `http://www.jstor.org/stable/2984955`.

Jordan Awan and Salil Vadhan. Canonical noise distributions and private hypothesis tests. *The Annals of Statistics*, 51 (2), April 2023. ISSN 0090-5364. doi: 10.1214/23-aos2259. URL `http://dx.doi.org/10.1214/23-AOS2259`.

Jordan A. Awan and Jinshuo Dong. Log-concave and multivariate canonical noise distributions for differential privacy. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, September 2013. doi: 10.1007/s10994-013-5404-1. URL https://doi.org/10.1007/s10994-013-5404-1.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12(null):1069–1109, July 2011. ISSN 1532-4435.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, aug 2014. ISSN 1551-305X. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, October 2010. doi: 10.1109/focs.2010.12. URL https://doi.org/10.1109/focs.2010.12.

Alex Dytso, Ronit Bustin, H. Vincent Poor, and Shlomo Shamai. Analytical properties of generalized gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1), December 2018. doi: 10.1186/s40488-018-0088-5. URL https://doi.org/10.1186/s40488-018-0088-5.

Arun Ganesh and Jiazheng Zhao. Privately answering counting queries with generalized gaussian mechanisms, 2020. URL https://arxiv.org/abs/2010.01457.

Quan Geng and Pramod Viswanath. The optimal mechanism in differential privacy. In *2014 IEEE International Symposium on Information Theory*, pages 2371–2375, 2014. doi: 10.1109/ISIT.2014.6875258.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately?, 2010.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Ao Liu, Xiaoyu Chen, Sijia Liu, Lirong Xia, and Chuang Gan. Certifiably robust interpretation via rényi differential privacy. *Artif. Intell.*, 313(C), December 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103787. URL https://doi.org/10.1016/j.artint.2022.103787.

Fang Liu. Generalized gaussian mechanism for differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):747–756, apr 2019. doi: 10.1109/tkde.2018.2845388. URL https://doi.org/10.1109%2Ftkde.2018.2845388.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 142–150, USA, 2011. Association for Computational Linguistics. ISBN 9781932432879.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, aug 2017. doi: 10.1109/csf.2017.11. URL https://doi.org/10.1109%2Fcsf.2017.11.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism, 2019. URL https://arxiv.org/abs/1908.10530.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification, 2015.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy, 2022. URL https://arxiv.org/abs/2110.03620.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.05755.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1802.08908.

Vasyl Pihur. The podium mechanism: Improving on the laplace and staircase mechanisms, 2019. URL https://arxiv.org/abs/1905.00191.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YTWGvpFOQD-.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. URL https://openreview.net/forum?id=EopKEYBoI-.

A Sampled Generalized Gaussian Mechanism

Privacy amplification by subsampling is a technique to strengthen DP guarantees without increasing the level of noise, by randomly sampling a subset of the input dataset before applying a DP mechanism; it is commonly used in ML applications. The DP parameters improve proportionally to the subsampling rate, as formalized in Theorem 4. Intuitively, each point is less likely to be used in the analysis, and the noise from sampling can be "counted" toward the privacy budget.

For our mechanism, we will focus on *Poisson subsampling*, a sampling process where each element of a population is included in a set according to the outcome of an independent Bernoulli trial. We use $S(q)$ to refer to a Poisson sampling procedure with sampling rate q.

Theorem 4 (Privacy Amplification by Poisson Subsampling [Kasiviswanathan et al., 2010] [Beimel et al., 2013])**.** *Let $\mathcal{M}$ be an (ϵ, δ)-DP mechanism, and let $\mathcal{S}(q)$ be a Poisson sampling procedure with sampling rate q. Then $\mathcal{M} \circ \mathcal{S}(q)$ is $(O(\log(q)\epsilon), q\delta)$-DP.*

A.1 Sampled Generalized Gaussian Mechanism

Next, we present the Sampled Generalized Gaussian Mechanism, SGG, which is a sampled variant of the GG Mechanism. It generalizes the Sampled Gaussian Mechanism [Mironov et al., 2019], which is a common mechanism in private ML. This mechanism relies on privacy amplification by subsampling (Theorem 4) to attain improved privacy guarantees relative to its non-sampled counterpart. We state the SGG Mechanism in terms of Poisson sampling because the PRV accountant is defined only for Poisson sampling; the mechanism can immediately be extended other types of sampling and the privacy guarantees would still hold under the appropriate accountant.

Algorithm 4 Sampled Generalized Gaussian Mechanism, $SGG_{\beta,\sigma,q}(f, D)$

Input noise parameters $\beta \geq 1$, $\sigma > 0$, sample rate $q \in (0, 1]$ a vector-valued function $f : \mathcal{D} \to \mathbb{R}^d$, database $D \in \mathcal{D}$
Compute l_2 sensitivity of f: $\Delta_2 f = \max_{D,D' \text{ neighbors}} \|f(D) - f(D')\|_2$
$S = \emptyset$
for each data element $x_j \in D$ **do**
    With probability q, add x_j to S
end for
for $i = 1$ to d **do**
    Sample $Y_i \sim \mathcal{N}_\beta(0, \sigma \cdot \Delta_2 f)$
end for
Output $f(S) + (Y_1, \dots Y_d)$

The privacy guarantees of the SGG Mechanism follow nearly immediately from privacy of the GG Mechanism (Theorem 1) and privacy amplification by sampling (Theorem 4).

Theorem 5. *For any $\beta \geq 1$, $\sigma > 0$, $\delta > 0$, $q \in (0, 1]$, there exists a value ϵ such that $SGG_{\beta,\sigma,q}(\cdot, \cdot)$ satisfies (ϵ, δ)-DP.*

Proof. Theorem 1 states that for any $\beta \geq 1$, $\sigma > 0$, $\delta > 0$, there exists an $\epsilon > 0$ for which $GG_{\beta,\sigma,q}(\cdot, \cdot)$ is (ϵ, δ)-DP. The SGG Mechanism simply applies Poisson Subsampling before running this GG Mechanism on the subsampled dataset, so by Theorem 4, SGG satisfies $(O(\log(q)\epsilon), q\delta)$-DP for the same ϵ. □

A.2 β-DP-SGD is Differentially Private

The β-DP-SGD algorithm presented in Algorithm 3 also relies on privacy amplification by subsampling for its privacy guarantees. At each time t, the algorithm performs Poisson subsampling with $q = L/|D|$ before its gradient measurement and update steps. We first provide the existential privacy result that β-DP-SGD is (ϵ, δ)-DP for *some* privacy parameters (Theorem 3), and then provide Corollary 1, which rephrases the privacy guarantee in a way that enables use of the PRV accountant across all T rounds.

Theorem 3. *For any $\delta > 0$, $\beta \geq 1$, $\sigma > 0$, $f : \mathcal{D} \to \mathbb{R}^d$, database $D \in \mathcal{D}$, for any loss function of the form $l(\theta, x_i)$, learning rate $\eta \geq 0$, average group size L, clipping norm $C \geq 0$, there exists $\epsilon > 0$ such that the algorithm β-DP-SGD$(\beta, \sigma, D, l, \eta, L, C)$ satisfies (ϵ, δ)-DP.*

Proof. Algorithm 3 can be viewed as an adaptive T-fold composition of the SGG Mechanism, with post-processing on its results. Specifically, the algorithm's final output θ_T can be written as $\theta_T = \theta_0 + \eta \sum_t^T \tilde{G}_t$, where the $\tilde{G}_t$ is the postprocessed output of an instantiation of SGG with $q = L/|D|$ and function $f(D) = \sum_i \nabla_{\theta_t} l(\theta_t, x_i) / \max\left(1, \frac{\|\nabla_{\theta_t} l(\theta_t, x_i)\|_\beta}{C}\right)$ and noise vector $\vec{Y}$ sampled with each entry $Y_i \sim \mathcal{N}_\beta(0, \sigma C)$. Since – given β, δ, σ, q – SGG is (ϵ, δ)-DP for some ϵ (Theorem 5), then the adaptive composition and postprocessing of these mechanisms' outputs that is performed in Algorithm 3 will also be (ϵ, δ)-DP for some ϵ. □

Corollary 1 restates this result in a way that directly allows the PRV accountant to be applied for tighter composition across all T rounds.

One final minor step is necessary to connect an accountant for the SGG mechanism to $DP - SGD$ – this is straightforward as $DP - SGD$ is the composition of many SGG mechanisms, but we make it explicit in Appendix A.2 for completeness. This enables us to use the PRV accountant to provide privacy accounting for β-DP-SGD.

Corollary 1. *If the $SGG_{\beta,\sigma,q}(f, D)$ mechanism composed T times on function f with sensitivity Δf satisfies (ϵ, δ)-DP, then for any $L \leq |D|$, $C = \Delta f$, and loss function $l(\theta, x_i)$, the β-DP-SGD$(\beta, \sigma, D, l, \eta, L, C)$ also satisfies (ϵ, δ)-DP.*

Proof. As shown in the proof of Theorem 3, the β-DP-SGD mechanism consists of the T-fold adaptive composition (with postprocessing) of the SGG Mechanism, with the clipping norm C specified by the sensitivity Δf of the input function f, so that $C = \Delta f$. Thus by Theorem 3 and the post-processing and composition guarantees of DP, the β-DP-SGD mechanism also satisfies (ϵ, δ)-DP for the same (ϵ, δ).

□

B Empirically Computing Privacy Guarantees

Due to the complexity that arises from composing multiple DP mechanisms together, any statement about the privacy-accuracy tradeoff of a highly-composed mechanism (like DP-SGD) is typically caveated by the specific implementation of the privacy accountant used for privacy accounting. We demonstrate an example of this in Table 1 where we present a clear improvement in the privacy-accuracy tradeoff of existing methods by simply changing RDP to PRV accounting.

At the time of writing, numerical accountants such as the PRV accountant achieve the tightest privacy guarantees Gopi et al. [2024]. However, as a result of using a numerical accountant, closed form solutions for the privacy consumed do not typically exist. Specifically, Gopi et al. [2024] introduced the algorithm ComposePRV (presented here in Appendix B.3), which efficiently computes privacy guarantees for the composition of multiple DP mechanisms based on the PRV accountant. It takes as input the CDFs of PRVs $Y_1, \ldots, Y_k$, a mesh size h, and truncation parameter L, and returns an estimate of the privacy curve for all the mechanisms composed, represented by $\delta(\epsilon)$, enabling the direct computation of ϵ. Given ComposePRV, [Gopi et al., 2024] also shows that the PRV accountant can be directly used to compute a mechanism's privacy loss, including for Poisson subsampled variants of mechanisms such as SGG.

However, when using the PRV accountant one must compute the CDF of the PRV, which is not a simple task for all differentially private mechanisms. In this work, we get around this by estimating the CDF of the PRV numerically; while this does introduce error, it also makes the privacy accounting of arbitrary mechanisms possible. In this appendix we show how to use the PRV accountant for the β-DP-SGD mechanism by replacing the true CDF with an empirical estimate, and provide error bounds for the resulting privacy guarantee. In short, we replace the CDF of the PRV with the CDF of a (binned) histogram of the PRV sampled n times, and account for the error introduced by using numerically computed histograms.

In this appendix, we first show that how to compute the PRVs for the single-dimensional GG mechanism, and the multi-dimensional GG mechanism, and show that the GG mechanism is dimension-independent for specific sensitivity bounds (Appendix B.1 and Appendix B.2). We then show how to use these PRVs to do privacy accounting for the GG mechanism and the sampled GG mechanism, and provide error bounds under the modified PRV accountant that estimates the CDF (Appendix B.3). Lastly, we explore how to find mechanisms that satisfy equivalent (ϵ, δ)-DP guarantees, which can be used to define a new search space of comparable mechanisms (Appendix B.4); we end by proposing a niche, but potentially valuable use-case where the Generalized Gaussian for $\beta \notin \{1, 2\}$ surpasses the Laplace or Gaussian Mechanisms, for the goal of preventing outliers (Appendix B.5).

In this paper, our privacy accounting for β-DP-SGD depends on numerically computing the CDF of the PRV for the Generalized Gaussian Mechanism by sampling from the Generalized Gaussian. For most distributions, computing the PRVs for a multi-dimensional output perturbation DP mechanism (e.g., a private mean in multiple dimensions) would depend on dimension. In typical machine learning applications (the primary use-case of DP-SGD), even small models

regularly exceed 1 million parameters, so dependence on dimension could be catastrophic if estimating a histogram involved sampling from $d > 1,000,000$ distributions. Instead, we find that the PRVs for the Generalized Gaussian Mechanism do not depend on dimension when noise is sampled from $\mathcal{N}_\beta$ and the function has $||\cdot||_\beta$-sensitivity. We note that this retroactively provides some rationale for why the Laplace and Gaussian Mechanism respectively use ℓ_1- and ℓ_2-norms for their noise distributions.

B.1 Analytic PRV for a Single-Dimensional Generalized Gaussian Mechanism

In order to compute the PRV for $GG_{\beta,\sigma}(f, D)$ we consider the privacy loss random variables for two distributions shifted by $\mu = \Delta f$, corresponding to the outputs of the mechanism on two neighboring datasets: $P \sim \mathcal{N}_\beta(\mu, \sigma)$ and $Q \sim \mathcal{N}_\beta(0, \sigma)$.

Proposition 1. *For $\sigma, \beta > 0$, let $Z \sim \mathcal{N}_\beta(0, \sigma)$, and let $\mu = \Delta f$. Then the PRVs for $GG_{\beta,\sigma}(f, D)$ are $X = \left(\frac{1}{\sigma}\right)^\beta \left(|Z|^\beta - |Z - \mu|^\beta\right)$ and $Y = \left(\frac{1}{\sigma}\right)^\beta \left(|Z - \mu|^\beta - |Z|^\beta\right)$.*

Proof. Given a noise distribution Z, and the associated differentially private mechanism, we derive the PRVs X, Y. This proof is of the same form as a similar derivation in Gopi et al. [2024], specifically deriving the PRVs for the Gaussian mechanism. The maximum difference in output of f between two neighboring datasets is $\mu = \Delta f$, so we can abstract the worst-case pair of neighboring datasets with distributions $P \sim \mathcal{N}_\beta(\mu, \sigma)$ and $Q \sim \mathcal{N}_\beta(0, \sigma)$.

Then for $t \sim Q = \mathcal{N}_\beta(0, \sigma)$, we can define the PRV Y as the following distribution:

$$Y \sim \log\left(\frac{Q(t)}{P(t)}\right) = \log\left(\frac{\exp(-t^\beta)}{-|t - \mu|^\beta}\right) = |t - \mu|^\beta - |t|^\beta.$$

By symmetry for X, for $t \sim P = \mathcal{N}_\beta(\mu, \sigma)$, we can define the PRV X as the distribution:

$$X \sim \log\left(\frac{Q(t)}{P(t)}\right) = \log\left(\frac{-|t - \mu|^\beta}{\exp(-t^\beta)}\right) = |t|^\beta - |t - \mu|^\beta.$$

□

We note that while it is not necessarily true that $X = -Y$ for all private mechanisms, it is true for the Generalized Gaussian mechanism, as the proof above holds for all values of β or σ.

B.2 Dimension Independence in Multi-Dimensional PRVs

We first derive the PRVs for the multi-dimensional Generalized Gaussian Mechanism, and then show that they are dimension independent.

As shown in Proposition 1, the PRVs for the single dimensional GG Mechanism are $Y = |Z - \mu|^\beta - |Z|^\beta$ and $X = |Z|^\beta - |Z - \mu|^\beta$ where $Z \sim \mathcal{N}_\beta(\mu = \Delta f, \sigma = 1)$.

As shown in Section 2.3, the definition of a PRV is random variable generated from the probability of sampling a particular sample t,

$$Y \sim \log\left(\frac{Q(t)}{P(t)}\right) \text{where } t \sim Q \quad \text{and} \quad X \sim \log\left(\frac{Q(t)}{P(t)}\right) \text{where } t \sim P.$$

This definition of a PRV does not change if t is a scalar or a vector — denoted $\vec{t} \in \mathbb{R}^d$, where t_i is a scalar from i-th dimension of $\vec{t}$. We now observe that the multi-dimensional probability distribution is equal to the product distribution of the single-dimensional PDFs, since the multidimensional GG mechanism is sampled independently for each dimension:

$$\Pr(\vec{t}) = \prod_{i=1}^{d} \Pr(\vec{t_i}) \propto \prod_{i=1}^{d} \exp(-|\vec{t_i}|^\beta / \sigma).$$

Using this fact, we can then compute the PRVs of the multi-dimensional GG Mechanism. Let $\vec{\mu}$ be a d-dimensional vector such that $||\vec{\mu}||_\beta = \Delta f$, and let μ_i be the i-th dimension of $\vec{\mu}$. Since each dimension of the noise is sampled independently in the GG mechanism, we denote the PRVs for a d dimensional Generalized Gaussian mechanism as Y_d and X_d. For $\vec{t} \sim Q$, these can be written as:

$$Y_d \sim \log \prod_{i=1}^{d} \left(\frac{\exp(-|t_i|^\beta/\sigma)}{\exp(-|t_i-\mu_i|^\beta/\sigma)}\right) = \sum_{i=1}^{d} (|t_i-\mu_i|^\beta - |t_i|^\beta)/\sigma,$$

and by symmetry,

$$X_d \sim \sum_{i=1}^{d} (|t_i|^\beta - |t_i-\mu_i|^\beta)/\sigma.$$

For a one-dimensional mechanism, μ in the PRV calculation must equal the sensitivity Δf of the function being privately evaluated. However, for multi-dimensional mechanisms, $\vec{\mu}$ must instead be a vector with norm $||\vec{\mu}||_\beta = \Delta f$ that maximizes the privacy loss random variables X and Y. As we will see, there are many possible $\vec{\mu}$ vectors which meet this requirement.

Below, we show that if sensitivity is measured with respect to the ℓ_β norm, then the PRVs for a multi-dimensional Generalized Gaussian mechanism are simply the PRVs of the corresponding the single-dimensional GG mechanism; this motivates our choice of $||\cdot||_\beta$ sensitivity, rather than fixing the sensitivity for all mechanisms.

Recall the definition of the ℓ_β norm: $||\vec{x}||_\beta^\beta = (\sum_i |x_i|^\beta)^{1/\beta}$ To start, we observe that $Y_d = \sum_{i=1}^{d} (|t_i-\mu_i|^\beta - |t_i|^\beta)/\sigma$ can be rewritten as

$$Y_d = \frac{1}{\sigma}\left[||\vec{t}-\vec{\mu}||_\beta^\beta - ||\vec{t}||_\beta^\beta\right].$$

Since differential privacy requires a worst-case bound over all pairs of neighboring databases, it requires bounding $\max\{\frac{Q(\vec{t})}{Q(\vec{t}-\vec{\mu})}, \frac{Q(\vec{t}-\vec{\mu})}{Q(\vec{t})}\}$, for all difference vectors μ satisfying the sensitivity bound $||\vec{\mu}||_\beta = \Delta f$. If the mechanism used ℓ_2-sensitivity, then the set of $\vec{\mu}$ over which this difference must be maximized is all points on an ℓ_2-ball of radius Δf, i.e., those satisfying $||\vec{\mu}||_2 = \Delta f$.

However, we observe that any $\vec{\mu}$ in the ℓ_β ball of radius Δf, i.e., $\vec{\mu} \in \{\vec{x} : ||\vec{x}||_\beta = \Delta f\}$, will satisfy this maximal-difference constraint because all such points are exactly Δf away from the origin in ℓ_β-distance. Thus $Y = \|\vec{t}-\vec{\mu}\|_\beta^\beta - \|\vec{t}\|_\beta^\beta$ will be identical for any $\vec{\mu}$ on the ℓ_β-ball. This includes, for example, the one-hot vector $\vec{\mu} = \langle 1, 0, \ldots, 0\rangle$. For this particular choice of $\vec{\mu}$, we see that the PRVs for the multidimensional GG mechanism are the same as the single-dimensional GG mechanism:

$$\begin{aligned}
Y_d &\sim \sum_{i=1}^{d} (|t_i-\mu_i|^\beta - |t_i|^\beta)/\sigma \\
&= \Big(|t_1-\Delta f|^\beta - |t_1|^\beta + 0 + \ldots + 0\Big)/\sigma \\
&= (|(t_1-\Delta f)|^\beta - |t_1|^\beta)/\sigma \\
&= Y.
\end{aligned}$$

The same holds for X_d and X:

$$\begin{aligned}
X_d &\sim \sum_{i=1}^{d} (|t_i|^\beta - |t_i-\mu_i|^\beta)/\sigma \\
&= \Big(|t_1|^\beta - |t_1-\Delta f|^\beta + 0 + \ldots + 0\Big)/\sigma \\
&= (|t_1|^\beta - |t_1-\Delta f|^\beta)/\sigma \\
&= X.
\end{aligned}$$

Thus for β-GG mechanism with sensitivity measured with a ℓ_β-norm, the PRV for the multi-dimensional GG distribution is equivalent to the PRV for a single-dimensional GG distribution, when the sensitivity is also measured with the ℓ_β-norm.

B.3 Privacy of the Sampled PRV Accountant

In this section we formally introduce the Sampled PRV accountant, and provide theoretical guarantees for its privacy and accuracy as a PRV accountant. Informally, the Sampled PRV accountant functions the same as the PRV accountant except whereas the PRV accountant takes the CDFs of the PRVs of the DP mechanisms as input, the Sampled PRV accountant takes Empirical Distribution Functions (EDF) of the PRVs, computed from samples of the PRVs.

Our approach builds upon the ComposePRV algorithm of Gopi et al. [2024], which takes as input the CDF of multiple PRVs (Y_i), and returns an estimate of the composed PRV $\tilde{Y}$ and its privacy curve $\delta_{\tilde{Y}}(\cdot)$. It uses as a subroutine DiscretizePRV, which discretizes the PRV. For completeness, we reproduce both algorithms below in Algorithms 5 and 6, respectively.

Our Sampled PRV accountant (SampledComposePRV, Algorithm 7) extends the ComposePRV algorithm to work even when the CDF of the PRVs is not known, but the algorithm instead has sample-access to the distribution. This process works by generating a histogram of samples from the PRV and using the empirical CDF an estimate for the CDF of the PRV, which can then be plugged into the ComposePRV algorithm in place of the true CDFs. The original ComposePRV algorithm produces an estimate of the composed privacy curve, along with error bars for the estimate; in Theorem 13 we extend the analysis to provide error bounds for ComposePRV when using a sampled CDF.

Algorithm 5 ComposePRV [Gopi et al., 2024]

Input: CDFs of PRVs $Y_1, Y_2, \dots Y_k$, mesh size h, truncation parameter L that is a multiple of h
output PDF of an approximation $\tilde{Y}$ of $Y = \sum_{i=1}^{k} Y_i$, supported on a grid over $[-L, L]$ with bin-width h
for $i = 1$ to k do **do**
    $\tilde{Y}_i \leftarrow \text{DiscretizePRV}(Y_i, L, h)$ ▷ Algorithm 6
end for
Compute PDF of $\tilde{Y} = \tilde{Y_1} \oplus_L \tilde{Y_2} \oplus_L \dots \oplus_L \tilde{Y_k}$ by convolving PDFs of $\tilde{Y_1}, \tilde{Y_2}, \dots, \tilde{Y_k}$ using FFT, where $\oplus_L$ is a convolution.
Compute $\delta_{\tilde{Y}}(\epsilon) = \mathbb{E}_{\tilde{Y}}\left[\left(1 - e^{\epsilon - \tilde{Y}}\right)_+\right]$ for all $\epsilon \in [0, L]$.
Return $\tilde{Y}, \delta_{\tilde{Y}}(\cdot)$.

Algorithm 6 DiscretizePRV [Gopi et al., 2024]

Input: $\text{CDF}_Y(\cdot)$ of a PRV Y, mesh size h, truncation parameter L that is a multiple of h
Output: PDF of an approximation $\tilde{Y}$ of Y, supported on a grid over $[-L, L]$ with bin-width h

Set $n = \frac{L - \frac{h}{2}}{h}$
for $i = -n$ to n **do**
    Set $q_i = CDF_Y(ih + \frac{h}{2}) - CDF_Y(ih - \frac{h}{2})$
end for
Normalize $q = \frac{q}{\sum_{i=-n}^{n} q_i}$ ▷ Ensures q is a valid probability distribution
Define $Y_L = Y_{|Y| \le L}$ ▷ i.e., Y conditioned on $|Y| \le L$
Set $\mu = \mathbb{E}[Y_L] - \sum_{i=-n}^{n} ih \cdot q_i$
Define $\tilde{Y}$ as the distribution that produces $ih + \mu$ with probability q_i for all $-n \le i \le n$
return $\tilde{Y}$

Finally, we can present our SampledComposePRV algorithm (Algorithm 7). This algorithm takes in sample-access mechanisms to k PRVs $Y_1, \dots, Y_k$, where each Y_i is the PRV for a single mechanism. For each Y_i, the algorithm generates n samples and constructs an empirical (PRV) distribution $Y_{i,n}$ from the samples. $Y = \sum_{i=1}^{k} Y_i$ is the true composed PRV, and $Y_n = \sum_{i=1}^{k} Y_{i,n}$ is the true composed empirical PRV. This latter term is estimated by the algorithm as $\tilde{Y}_{n,L,h}$, by applying ComposePRV to the empirical PRVs $Y_{i,n}$ using bins of width h over the domain $[-L, L]$.[8] We use $\tilde{Y}_{L,h}$ to denote the output of ComposePRV on the true PRVs Y_i, also with bin width h and domain $[-L, L]$; this will be used as an intermediary point of comparison in the analysis of Algorithm 7.

[8] Note, in practice, computing the empirical CDF involves binning and truncating as well. We choose an L such that with high probability the support of the empirical CDF is contained with $[-L, L]$ and we use the same bin width h for the empirical CDF as we do for the ComposePRV step.

Algorithm 7 SampledComposePRV

Input: Sample-access to PRVs $Y_1, Y_2, \dots Y_k$, mesh size h, truncation parameter L that is a multiple of h, number of samples n

output PDF of an approximation $\tilde{Y}$ of $Y = \sum_{i=1}^{k} Y_i$, supported on a grid over $[-L, L]$ with bin-width h

1: **for** $i = 1$ to k **do**
2: Generate n samples from Y_i. Bin and discretize each sample into equal sized bins of size h over the domain $[-L, L]$.
3: Let $Y_{i,n}$ be a RV defined by the empirical PDF generated by n samples from Y_i
4: **end for**
5: Compute $\tilde{Y}_{n,L,h} = \text{ComposePRV}((Y_{1,n}, \dots Y_{k,n}), h, L)$
6: Return $\tilde{Y}_{n,L,h}$

We also define a function $\texttt{sample}_n(Y)$ which takes an integer n and a random variable Y as input, and returns a random variable $Y_{n,L,h}$ as output. $\texttt{sample}_n(Y)$ generates n samples from Y, then constructs a random variable with the empirical CDF of the samples.

Algorithm 8 Sample Function $\texttt{sample}_n(Y)$

Input: Integer n, random variable Y

Output: Random variable $Y_{n,L,h}$ constructed from the empirical CDF of n samples from Y

1: Generate n independent samples from Y: $\{y_1, y_2, \dots, y_n\} \sim Y$.
2: Construct the empirical CDF F_n from the samples $\{y_1, y_2, \dots, y_n\}$.
3: Define $Y_{n,L,h}$ as the random variable corresponding to F_n.
4: Return $Y_{n,L,h}$.

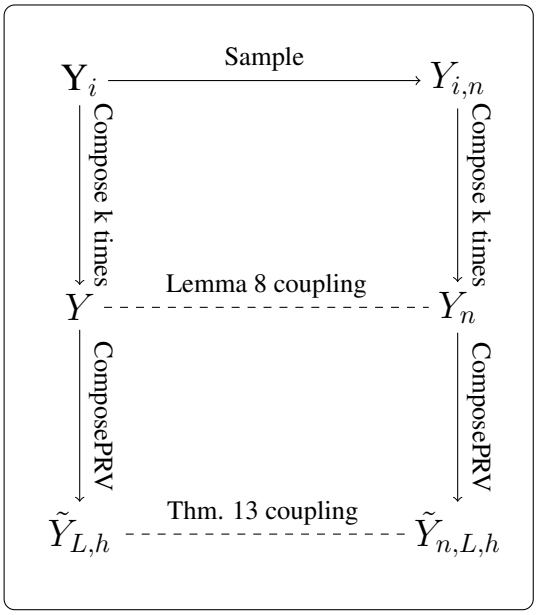

Figure 5: Visualization of some of our variables, in order to assist with tracking our couplings and our goal with the proof from Theorem 13.

To prove performance of this algorithm as a privacy accountant, we must bound the error in privacy guarantees estimated through Algorithm 7 and the true privacy guarantees of the composed mechanisms. The remainder of this subsection will explore the relationships between Y, $Y_{n,L,h}$, and $\tilde{Y}_{n,L,h}$, culminating in the final guarantee in Theorem 13. A rough outline of the analysis is as follows:

1. Couple the total variation distance of a sampled RV and the RV from which it was generated (Lemma 6). Given a total variation distance, produce a coupling between the sampled RV and the RV (Lemma 7). These combine in Corollary 2 to give a coupling between the true PRV and its sampled PRV.
2. Couple the composed PRV $\tilde{Y}_{L,h}$ to the sampled composed PRV $\tilde{Y}_{n,L,h}$, with high probability (Lemma 9)
3. Combine a result of Gopi et al. [2024] on the error of ComposePRV (Theorem 10) with the previous couplings to produce an error bound for SampledComposePRV (Theorem 13)

First we bound the total variation distance between the sampled-and-discretized PRV $Y_{n,L,h}$ and true (bounded) PRV Y_L over bounded domain $[-L, L]$ that it seeks to approximate.

Lemma 6. *Given a random variable Y_L with support over $[-L, L]$, define another random variable $Y_{n,L,h}$, defined as the empirical distribution of n samples from Y_L discretized into bins of width h. Then with probability at least $1 - \beta$, the total variation distance between Y_L and $Y_{n,L,h}$ is at most $2L \cdot \alpha$, as long as $n \geq \frac{1}{\alpha^2} \ln\left(\frac{4L}{h\beta}\right)$.*

Proof. First, we bound the probability that any one bin of the $Y_{n,L,h}$ is off by more than α. Using an additive Chernoff bound, for the ith bucket, we can bound the probability that the empirical frequency in that bucket with n samples ($\hat{p}_i$) – i.e., the frequency under $Y_{n,L,h}$ – differs from its mean – i.e., the frequency under Y_L by more than α. Specifically, $\Pr[|p_i - \hat{p}_i| \leq \alpha] \leq 2\exp(-2n\alpha^2)$. Taking a union bound over all $2L/h$ buckets, we can bound the probability of any bucket having error larger than α:

$$\Pr\left[\max_i |p_i - \hat{p}_i| \leq \alpha\right] \leq \sum_{i=1}^{2L/h} \Pr[|p_i - \hat{p}_i| \leq \alpha] = (2L/h) \cdot 2\exp(-2n\alpha^2).$$

To ensure that this failure probability is at most β requires $\beta \geq \frac{4L}{h}\exp(-2n\alpha^2)$, or equivalently, that $n \geq \frac{1}{\alpha^2}\ln(\frac{4L}{h\beta})$.

Thus with probability at least $1 - \beta$, the total variation distance between Y_L and $Y_{n,L,h}$ is $2L\alpha$, for n sufficiently large. □

To analyze the error between the true PRV and the Sampled PRV, we will use the following notion of coupling approximation, which is a particular measure of closeness between two distributions.

Definition 9 (Coupling Approximation [Gopi et al., 2024])**.** *Given two random variables Y_1, Y_2, we write that $|Y_1 - Y_2| \leq_\eta c$ if there exists a coupling between Y_1, Y_2 such that $\Pr[|Y_1 - Y_2| > c] \leq \eta$.*

Gopi et al. [2024], also provided the following lemma, which relates total variation distance to a coupling.

Lemma 7 ([Gopi et al., 2024])**.** *For random variables X, Y, if their total variation distance is at most η: $d_{TV}(X, Y) \leq \eta$, then $|X - Y| \leq_\eta 0$.*

Combining Lemmas 6 and 7, we immediately get a coupling approximation between the bounded PRV Y_L and the sampled PRV $Y_{n,L,h}$ for a single (non-composed) DP mechanism.

Corollary 2. *Given a random variable Y_L with support over $[-L, L]$, define random variable $Y_{n,L,h}$ with PDF equal to the empirical distribution of n samples from Y_L discretized into bins of width h. Then if $n \geq \frac{1}{\alpha^2}\ln\left(\frac{4L}{h\beta}\right)$, then the following coupling holds: $|Y_L - Y_{n,L,h}| \leq_\beta \alpha$.*

Lemma 8. *Given a random variable Y_L with support over $[-L, L]$, define random variable $Y_{n,L,h}$ with PDF equal to the empirical distribution of n samples from Y_L discretized into bins of width h. With probability $1 - \beta$, there exists a coupling such that $|Y_L - Y_{n,L,h}| \leq_0 \alpha$, as long as $n \geq \frac{1}{\alpha^2}\ln\left(\frac{4L}{h\beta}\right)$.*

Proof. Look to the definition provided by Corollary 2, if $n \geq \frac{1}{\alpha^2}\ln\left(\frac{4L}{h\beta}\right)$, then the following coupling holds: $|Y_L - Y_{n,L,h}| \leq_\beta \alpha$. Which equivalently means $\Pr[|Y_L - Y_{n,L,h}| > \alpha] \leq \beta$. This coupling is equivalent to stating that with probability $1 - \beta$, $|Y_L - Y_{n,L,h}| <_0 \alpha$.

We do note that one must be careful about what you sample over, as Y_L and $Y_{n,L,h}$ are both random variables, but for $Y_{n,L,h}$ the associated PDF is also a random variable (in other words, $Y_{n,L,h}$ depends on the specific n samples from Y_L that are drawn). Thus, this randomness is over draws from sample$_n(Y_L)$. □

Next, we use Corollary 2 to show a coupling approximation between the composed PRV and the sampled composed PRV.

The following lemma from Gopi et al. [2024] shows that the sum of $(0, c)$-coupled random variables is also coupled.

Lemma 9 ([Gopi et al., 2024])**.** *Suppose $Y_1, Y_2, \ldots, Y_k$ and $Y_1', Y_2', \ldots, Y_k'$ are two collections of independent random variables such that $|Y_i - Y_i'| \leq_0 c$ and $\mathbb{E}[Y_i] = \mathbb{E}[Y_i']$ for all i, then*

$$\left|\sum_{i=1}^{k} Y_i - \sum_{i=1}^{k} Y_i'\right| \leq_\eta c\sqrt{2k \log \frac{2}{\eta}}.$$

Importantly, we note that while lemma 9 requires a coupling of the form $|Y_i - \tilde{Y}_i| \leq_0 \alpha$. However, the output of Corollary 2 only returns a coupling of the form $|Y_i - \tilde{Y}_i| \leq_\beta \alpha$.

Finally, we translate the coupling approximation property into a guarantee on the accuracy of the SampledComposePRV algorithm. We begin by stating the result of Gopi et al. [2024], which gives an accuracy guarantee on the ComposePRV algorithm, and will later extend this to a similar bound for SampledComposePRV.

Theorem 10 ([Gopi et al., 2024])**.** *Let $\alpha, \beta > 0$ be some fixed error terms. Let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ be DP algorithms with privacy curves $\delta_{\mathcal{M}_i}(\epsilon)$. Let Y_i denote the PRV corresponding to $\mathcal{M}_i$ (where all $\mathcal{M}_i$ are identical) so that $\delta_{\mathcal{M}_i}(\epsilon) = \delta_{Y_i}(\epsilon)$ for $\epsilon \geq 0$. Let $\mathcal{M}$ be the (adaptive) composition of $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ and let $\delta_{\mathcal{M}}(\epsilon)$ be its privacy curve. Let $L \geq 2 + \alpha$ be sufficiently large such that $\sum_{i=1}^{k} \delta_{\mathcal{M}_i}(L-2) \leq \frac{\beta}{8}$ and $\delta_{\mathcal{M}}(L-2-\alpha) \leq \frac{\beta}{4}$. Let $\tilde{Y}$ be the approximation of $Y = \sum_{i=1}^{k} Y_i$ produced by ComposePRV with mesh size $h = \frac{\alpha}{\sqrt{\frac{k}{2}\log(12/\beta)}}$ and truncation parameter L. Then,*

$$\delta_{\tilde{Y}}(\epsilon + \alpha) - \beta \leq \delta_Y(\epsilon) = \delta_{\mathcal{M}}(\epsilon) \leq \delta_{\tilde{Y}}(\epsilon - \alpha) + \beta.$$

Furthermore, ComposePRV takes $O\left(b\frac{L}{h}\log(\frac{L}{h})\right)$ time, where b is the number of distinct algorithms among $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$.

Theorem 10 gives accuracy bounds for the ComposePRV algorithm, but to bound the accuracy of SampledComposePRV, we must also take into account the additional error from the sampling process. Lemma 11 below is useful for translating the coupling approximation guarantees between the sampled composed PRV and the composed PRV.

Lemma 11 ([Gopi et al., 2024])**.** *If Y and $\tilde{Y}$ are two random variables such that $|Y - \tilde{Y}| \leq_\eta c$, then for every $\epsilon > 0$,*

$$\delta_{\tilde{Y}}(\epsilon + c) - \eta \leq \delta_Y(\epsilon) \leq \delta_{\tilde{Y}}(\epsilon - c) + \eta.$$

For completeness, we now recount the triangle inequality for couples from Gopi et al. [2024]:

Lemma 12 ([Gopi et al., 2024])**.** *Suppose X, Y, Z are random variables such that $|X - Y| \leq_{\eta_1} c_1$ and $|Y - Z| \leq_{\eta_2} c_2$. Then,*

$$|X - Z| \leq_{\eta_1+\eta_2} c_1 + c_2.$$

Finally, we are ready to state and proof our main result on the accuracy of SampledComposePRV as a privacy accountant.

Theorem 13. *Let $\alpha, \beta > 0$ be some fixed error terms. Let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ be DP algorithms with privacy curves $\delta_{\mathcal{M}_i}(\epsilon)$. Let Y_i be the PRV corresponding to $\mathcal{M}_i$ so that $\delta_{\mathcal{M}_i}(\epsilon) = \delta_{Y_i}(\epsilon)$ for $\epsilon \geq 0$. Let $\mathcal{M}$ be the (adaptive) composition of $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ and let $\delta_{\mathcal{M}}(\epsilon)$ be its privacy curve. Let $L \geq 2 + \alpha$ be sufficiently large such that $\sum_{i=1}^{k} \delta_{\mathcal{M}_i}(L-2) \leq \frac{\beta}{8}$ and $\delta_{\mathcal{M}}(L-2-\alpha) \leq \frac{\beta}{4}$. Let $\tilde{Y}$ be the approximation of $Y = \sum_{i=1}^{k} Y_i$ produced by SampledComposePRV with mesh size $h = \frac{\alpha}{\sqrt{\frac{k}{2}\log(12/\beta)}}$, truncation parameter L, and $n \geq \frac{1}{\alpha^2}\ln\left(\frac{4L}{h\beta}\right)$. Then, with probability at least $(1-\beta)^b$,*

$$\delta_{\tilde{Y}}(\epsilon + 2\alpha) - 2\beta \leq \delta_Y(\epsilon) = \delta_{\mathcal{M}}(\epsilon) \leq \delta_{\tilde{Y}}(\epsilon - 2\alpha) + 2\beta.$$

Furthermore, SampledComposePRV takes $O\left(b \cdot \frac{L}{h}\log(\frac{L}{h})\right)$ time, where b is the number of distinct algorithms among $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$. [9]

Proof. From Lemma 2, we know that for each individual DP mechanism, there is a coupling approximation between the truncated PRV Y_L and the sampled PRV $Y_{n,L,h}$: $|Y_L - Y_{n,L,h}| \leq_\beta \alpha$, as long as $n \geq \frac{1}{\alpha^2}\ln(\frac{4L}{h\beta})$. By Lemma 8, we

[9] Most accountants compose a single mechanism many times (e.g., inside DP-SGD), so typically the success probability is $1 - \beta$, and the runtime does not incur this linear dependence on b.

know that with probability $1-\beta$, there is a coupling between $Y_{n,L,h}$ and Y_L such that $|Y_L - Y_{n,L,h}| \leq_0 \alpha$, as long as $n \geq \frac{1}{\alpha^2} \ln\left(\frac{4L}{h\beta}\right)$. We have now coupled the composed the PRVs and the sampled PRVs for an individual mechanism. Next we look at the coupled versions.

From Theorem 10, we know that the composed PRV $\tilde{Y}_{L,h}$ produced by ComposePRV is coupled with the true composed PRV Y. Thus Y is coupled to Y_n, and Y is coupled to ComposePRV(Y), and Y_n is coupled to ComposePRV(Y_n) (which we denote $Y_{n,L,h}$). By the triangle inequality for couplings stated in Lemma 12 we have coupled $Y_{n,L,h}$ with Y with probability $(1-\beta)$.

As a final note, observe that in this process we have coupled $Y_{n,L,h}$ which is a truncated and sampled random variable, to Y which is not truncated, by choosing L for truncation to be the same L used in ComposePRV.

□

We note that the α, β used in the sample size requirement from Lemma 2 need not be the same as the α, β used in the accuracy of ComposePRV in Theorem 10, although we set them to be equal here for easy of presentation.

B.4 Mechanisms with Equivalent Privacy Guarantees

For any privacy accountant, it is generally possible to run the accounting algorithm many times to compute the hyperparemeters required to achieve a particular degree of privacy. We introduce the following simple but effective algorithm for using the PRV accountant as part of a binary search over possible values of σ in order to compute the minimal σ value that $GG_{\beta,\sigma}$ satisfies (ϵ, δ)-DP for a given β.

Let $PRV(\beta, \sigma, \delta)$ be a subroutine that runs the PRV accountant for the $GG_{\beta,\sigma}$ mechanism, and returns the ϵ value associated, such that $GG_{\beta,\sigma}$ satisfies (ϵ, δ)-DP.

Algorithm 9 Binary-search σ-solver

Input: $\beta \geq 1$, $\epsilon > 0$, $\delta > 0$, tolerance $\tau > 0$
Output: σ, such that $GG_{\beta,\sigma}$ satisfies (ϵ, δ)-DP
$\sigma_{min} = \sigma_{max} = 1$
while $PRV(\beta, \sigma_{min}, \delta) > \epsilon$ **do**
    $\sigma_{min} = \sigma_{min}/2$.
end while
while $PRV(\beta, \sigma_{max}, \delta) < \epsilon$ **do**
    $\sigma_{max} = \sigma_{max} * 2$.
end while
while $PRV(\beta, \sigma_{max}, \delta) - \epsilon > \tau$ **do**
    $\sigma_{mid} = \frac{\sigma_{max}+\sigma_{min}}{2}$
    if $PRV(\beta, \sigma_{mid}, \delta) > \epsilon$ **then**
        $\sigma_{min} = \sigma_{mid}$
    else $\sigma_{max} = \sigma_{mid}$
    end if
end while
return σ_{max}

For our empirical results in Sections 4 and 5, we compare how the value of β impacts accuracy for a fixed privacy guarantee.

B.5 Outliers for Equivalently Private Mechanisms

Using the sampled-PRV privacy accountant and the Binary-search σ-solver algorithm described in Appendix B.4, we are able to compute the σ-value — as function of ϵ, δ, and β — such that $GG_{\beta,\sigma}(f, D)$ satisfies (ϵ, δ)-DP. Such a computation is currently not possible with other popular privacy accountants, such as the RDP accountant and GDP accountant,[10], and no such analytic privacy bound currently exists for the GG mechanism for non-integer values of β.

[10]These are the only two other privacy accountants supported by Opacus, the most popular private machine learning library Yousefpour et al. [2021]

Combining this empirical privacy accountant with the known CDF of the Generalized Gaussian distribution $\mathcal{N}_\beta(0, \sigma)$ [Dytso et al., 2018], we can compute the weight w of the tail of the distribution, as a function of the mechanism's parameters and a cutoff parameter τ, which specifies the threshold for defining the tail: $w(\beta, \epsilon, \delta, \tau) = \int_{-\infty}^{-\tau} GG_{\beta,\epsilon,\delta}(x)dx + \int_{\tau}^{\infty} GG_{\beta,\epsilon,\delta}(x)dx$, where $GG_{\beta,\epsilon,\delta}(x)$ is a probability distribution associated with a Generalized Gaussian for a particular value of β that satisfies (ϵ, δ)-DP.[11] Using this formalism, given a tail threshold τ, any sample that falls in the tail can be labeled as an outlier, and the weight w described how likely such an outlier is to be observed, given the mechanism's parameters.

Figure 6 plots the weight in the tail of a GG distribution that satisfies a particular (ϵ, δ)-DP guarantee, as a function of β, and varying $\tau = \{1, 2, 4\}$. The left shows the tail weight for the GG Mechanism satisfying $(1.5, 10^{-5})$-DP, and the right shows tail weight for the SGG Mechanism satisfying $(1.5, 10^{-5})$-DP using Poisson sampling rate $q = 0.01$ and composed over 100 rounds. Given a desired (ϵ, δ)-DP guarantee, a practitioner that wishes to minimize the probability of outliers in any of their computations could generate such a plot and then pick the choice β which provides a minimum weight, thus minimizing the likelihood of outliers.

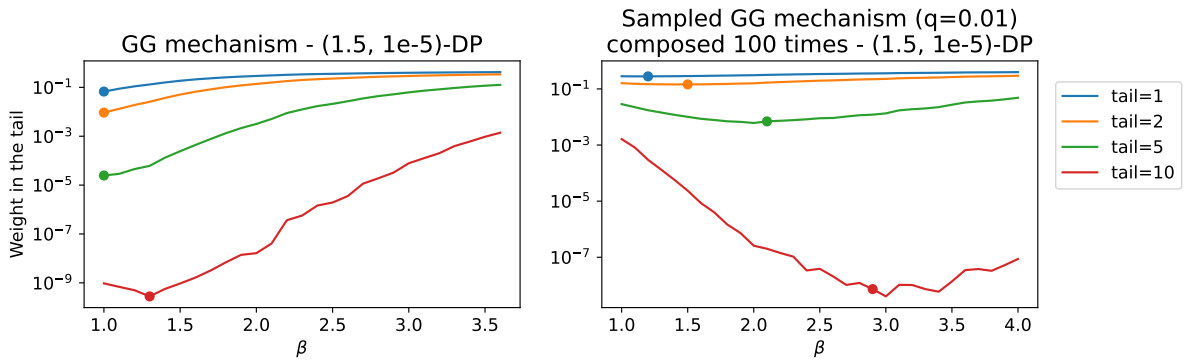

Figure 6: Likelihood of outliers for the GG Mechanism (left) and SGG Mechanism with Poisson sampling probability $q = 0.01$ composed 100 rounds (right), both that satisfy $(1.5, 10^{-5})$-DP . In the legend, *tail* refers to the τ tail cutoff parameter. The jaggedness in the plotting comes from the contributions of sampling error methods and integration error for very small weights; values are slightly smoothed with a Savitzky-Golay filter with polynomial order 2 and window size 5.

While for the single-shot GG mechanism (left), the Laplace mechanism appears optimal (i.e., has the least weight in the tail), Figure 6 shows that that there are regimes where the tails of Laplace and Gaussian are heavier (i.e., outliers are more likely) than other equivalently private mechanisms in the GG family. This shows it is possible for that there to exist values of $\beta \notin \{1, 2\}$ for which the associated GG Mechanism outperforms Laplace and Gaussian mechanism on this particular metric. This observation provides a potential direction for new methods to evaluate and search for alternative DP mechanisms: minimizing outliers was one of the main considerations cited by both the US Census [Abowd et al., 2022] and researchers behind PATE [Papernot et al., 2018] when deciding to use the Gaussian mechanism instead of the Laplace mechanism. Of course, the likelihood of outliers should be integrated with other organizational objectives and constraints, since in general, an algorithm designer's goal is not simply to minimize outliers, but rather to minimize some loss function (e.g., maximize accuracy) that is impacted by outliers

C Omitted Privacy Proofs

C.1 Bounded Rényi Divergence of the Generalized Gaussian Mechanism

Here we prove that the Renyi divergence between two Generalized Gaussian random variables is bounded, which is used in the proof of Theorem 1. The two distributions in the statement of Lemma 14 correspond to the maximally different output distributions of the GG Mechanism on neighboring databases, where $\mu > 0$ is the sensitivity of the

[11] We are explicitly not specifying the value of σ for ease of presentation. Specifying the input β parameter and the desired privacy parameters ϵ, δ will uniquely determine the minimum σ value that achieves the privacy parameters.

input function. Note that this bound would be unchanged if we shifted the means of both distributions by the same amount, so we do not need to consider the values of the function, only the difference in the values.

Lemma 14. *The Renyi Divergence $D_\alpha(\mathcal{N}_\beta(0,\sigma)||\mathcal{N}_\beta(\mu,\sigma))$ is bounded for all $\alpha > 1$, $\beta \geq 1$, $\sigma > 0$, and $\mu > 0$.*

Proof. **Definition of Rényi Divergence**
The Rényi divergence of order α between two probability density functions P and Q is defined as:

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty} P(x)^\alpha Q(x)^{1-\alpha}\,dx.$$

Proof Overview To show that $D_\alpha(P\|Q)$ is bounded, we will:

1. **Compute the Integrand:** Find the explicit form of $P(x)^\alpha Q(x)^{1-\alpha}$.
2. **Analyze the Integral:** Demonstrate that the integral is finite by examining the behavior of the integrand at $x \to \infty$ and $x \to 0$.
3. **Conclude Boundedness:** Use the finiteness of the integral to conclude that $D_\alpha(P\|Q)$ is bounded.

Detailed Proof
1. Compute the Integrand
Let $P(x) = f(x;0,\sigma)$ and $Q(x) = f(x;\mu,\sigma)$. Then,

$$\begin{aligned}P(x)^\alpha Q(x)^{1-\alpha} &= \left(\frac{\beta}{2\sigma\Gamma(1/\beta)}\right)^\alpha \exp\left(-\alpha\left(\frac{|x|}{\sigma}\right)^\beta\right) \times \left(\frac{\beta}{2\sigma\Gamma(1/\beta)}\right)^{1-\alpha} \exp\left(-(1-\alpha)\left(\frac{|x-\mu|}{\sigma}\right)^\beta\right)\\ &= \left(\frac{\beta}{2\sigma\Gamma(1/\beta)}\right)\exp\left(-\alpha\left(\frac{|x|}{\sigma}\right)^\beta - (1-\alpha)\left(\frac{|x-\mu|}{\sigma}\right)^\beta\right).\end{aligned}$$

Simplifying the exponent:

$$\begin{aligned}&-\alpha\left(\frac{|x|}{\sigma}\right)^\beta - (1-\alpha)\left(\frac{|x-\mu|}{\sigma}\right)^\beta\\ &= -\alpha\left(\frac{|x|}{\sigma}\right)^\beta + (\alpha-1)\left(\frac{|x-\mu|}{\sigma}\right)^\beta\\ &= (\alpha-1)\left(\left(\frac{|x-\mu|}{\sigma}\right)^\beta - \left(\frac{|x|}{\sigma}\right)^\beta\right) - \left(\frac{|x|}{\sigma}\right)^\beta.\end{aligned}$$

So the integrand becomes:

$$P(x)^\alpha Q(x)^{1-\alpha} = \left(\frac{\beta}{2\sigma\Gamma(1/\beta)}\right)\exp\left((\alpha-1)\left(\left(\frac{|x-\mu|}{\sigma}\right)^\beta - \left(\frac{|x|}{\sigma}\right)^\beta\right) - \left(\frac{|x|}{\sigma}\right)^\beta\right).$$

2. Analyze the Integral
We need to evaluate:

$$I = \int_{-\infty}^{\infty} P(x)^\alpha Q(x)^{1-\alpha}\,dx = \left(\frac{\beta}{2\sigma\Gamma(1/\beta)}\right)\int_{-\infty}^{\infty}\exp\left(F(x)\right)\,dx,$$

where

$$F(x) = (\alpha-1)\left(\left(\frac{|x-\mu|}{\sigma}\right)^\beta - \left(\frac{|x|}{\sigma}\right)^\beta\right) - \left(\frac{|x|}{\sigma}\right)^\beta.$$

Behavior at Infinity ($x \to \pm\infty$)
As $|x| \to \infty$:

- $|x|^\beta$ dominates $|x-\mu|^\beta$ since μ is finite.
- $\left(\frac{|x-\mu|}{\sigma}\right)^\beta \approx \left(\frac{|x|}{\sigma}\right)^\beta$.

Therefore,

$$F(x) \approx (\alpha - 1)(0) - \left(\frac{|x|}{\sigma}\right)^{\beta} = -\left(\frac{|x|}{\sigma}\right)^{\beta}.$$

Thus, the integrand behaves like:

$$\exp\left(-\left(\frac{|x|}{\sigma}\right)^{\beta}\right),$$

which decays exponentially as $|x| \to \infty$. Hence, the integral converges at infinity.
Thus we know that there is a value a for which the integral of the Renyi Divergence integrand over the domain $[a, \infty] \cup [-\infty, -a]$ is bounded, as it is bounded by an exponential function $e^{-C|x|}$ for some value of C and the Laplace function has a bounded integral for all values of C.
For the rest of the proof we need to show that the integrand is finite for all values in the range $[-a, a]$ and we will have proven that the integral is bounded.
Behavior Near Zero ($x \approx 0$)
As $x \to 0$:

- $|x| \to 0$, so $\left(\frac{|x|}{\sigma}\right)^{\beta} \to 0$.
- $|x - \mu| \to |\mu|$, so $\left(\frac{|x-\mu|}{\sigma}\right)^{\beta} \to \left(\frac{\mu}{\sigma}\right)^{\beta}$.

Therefore,

$$F(x) \to (\alpha - 1)\left(\left(\frac{\mu}{\sigma}\right)^{\beta} - 0\right) - 0 = (\alpha - 1)\left(\frac{\mu}{\sigma}\right)^{\beta}.$$

So near $x = 0$, the integrand behaves like:

$$\exp\left((\alpha - 1)\left(\frac{\mu}{\sigma}\right)^{\beta}\right),$$

which is a finite constant since $\alpha > 1$, $\mu > 0$, and $\sigma > 0$.
From here we conclude that the integrand is bounded at all values of x including the domain $[-a, a]$ introduced in the previous subsection.
Conclusion on Integrability
Since the integrand is exponentially decaying at infinity and finite near zero, the integral I converges:

$$I = \int_{-\infty}^{\infty} P(x)^{\alpha} Q(x)^{1-\alpha}\, dx < \infty.$$

3. Conclude Boundedness
The Rényi divergence is:

$$D_{\alpha}(P\|Q) = \frac{1}{\alpha - 1} \log I,$$

and since I is finite and positive, $\log I$ is finite. Therefore, $D_{\alpha}(P\|Q)$ is bounded for all $\alpha > 1$, $\beta \geq 1$, $\sigma > 0$, and $\mu > 0$.
Final Statement
Thus, we have proven that the Rényi divergence $D_{\alpha}(\mathcal{N}_{\beta}(0, \sigma)\|\mathcal{N}_{\beta}(\mu, \sigma))$ is bounded under the given conditions.
□

C.2 DP Guarantees of Generalized Gaussian Mechanism

Theorem 1. *For any $\beta \geq 1$, $\sigma > 0$, $\delta > 0$ there exists a finite value ϵ such that $GG_{\beta,\sigma}(\cdot, \cdot)$ satisfies (ϵ, δ)-DP.*

Proof. By Lemma 14 we know that the Rényi divergence of the GG Mechanism on any two neighboring databases $D_{\alpha}(\mathcal{N}_{\beta}(0, \sigma)||\mathcal{N}_{\beta}(\Delta f, \sigma))$ is bounded. By the definition of Rényi Differential Privacy [Mironov, 2017] (Definition 7), then the GG Mechanism must satisfy (α, ϵ)-RDP for the finite ϵ corresponding to the upper bound on $D_{\alpha}(\mathcal{N}_{\beta}(0, \sigma)||\mathcal{N}_{\beta}(\Delta f, \sigma))$. Then by the known translation between RDP and DP [Mironov, 2017], the GG Mechanism also satisfies $(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP for the same ϵ and for any $\delta \in (0, 1)$. □

C.3 DP Guarantees of GGNmax

Theorem 2. *If the (β, σ)-Generalized Gaussian Mechanism is (ϵ, δ)-DP for a fixed $\epsilon > 0$ and $\delta \geq 0$, then (β, σ)-Generalized Gaussian Private Argmax is also (ϵ, δ)-DP.*

The proof of Theorem 2 follows closely to the proof of privacy of the ReportNoisyMax algorithm with Laplace noise, as presented in Dwork and Roth [2014]. It is included here for completeness.

Proof. Fix neighboring databases D, D', where $D = D' \cup \{a\}$. Let c and c' respectively denote the vector of function values when the database is D and D'. We use two properties:

1. *Monotonicity of counts*. For all $j \in [N]$, $c_j \geq c'_j$
2. *Lipschitz property*. For all $j \in [N]$, $\Delta + c'_j \geq c_j$

Fix any $i \in [N]$. We will bound from above and below ratio of the probabilities that i is selected with D and with D'. Fix r_{-i}, a draw from $[\mathcal{N}_\beta(0, \sigma\Delta]^{N-1}$ used for all the noisy function evaluation except for the ith function. We use the notation $Pr[i|\zeta]$ to mean the probability that the output of the GGNmax algorithm is i conditioned on ζ.

We first argue that $\Pr[i|D, r_{-i}] \leq e^\epsilon \Pr[i|D', r_{-i}] + \delta$. Define

$$r^* = \min_{r_i} : c_i + r_i > c_j + r_j \;\; \forall j \neq i.$$

Note that having fixed r_{-i}, i is will be the output of the GGNmax mechanism (i.e., the ith function will have the largest noisy value) when the database is D if and only if $r_i \geq r^*$. Then for all $1 \leq j \neq i \leq N$:

$$\begin{aligned} c_i + r^* &> c_j + r_j \\ \Rightarrow (\Delta + c'_j) + r^* \geq c_i + r^* &> c_j + r_j \geq c'_j + r_j \\ \Rightarrow c'_i + (r^* + 1) &> c'_j + r_j \end{aligned}$$

Thus if $r_i \geq r^* + \Delta$, then i will be the output of GGNmax, meaning that the ith function had the largest noisy value, when the database is D' and the noise vector is (r_i, r_{-i}). We now wish to relate the probability of $\Pr[r_i \geq \Delta + r^*]$ to the probability $\Pr[r_i \geq r^*]$. We note that this is the step where the proof diverges from the privacy proof of Report Noisy Max, as given in Dwork and Roth [2014].

Let $Z \sim \mathcal{N}_\beta(0, \sigma\Delta)$. We are given that $GG_{\beta,\sigma}(f, D)$ satisfies (ϵ, δ)-DP, so,

$$\begin{aligned} \Pr[Z \geq r^*] &\leq e^\epsilon \Pr[Z \geq r^* + \Delta] + \delta \\ \Rightarrow \Pr[i|D, x_{-i}] = \Pr[r_i \geq r^*] &\leq e^\epsilon \Pr[r_i \geq r^* + \Delta] + \delta \leq e^\epsilon \Pr[i|D', x_{-i}] + \delta. \end{aligned}$$

We now argue that $Pr[i|D'] \leq e^\epsilon \Pr[i|D] + \delta$. Define, again,

$$r^* = \min_{r_i} : c'_i + r_i > c'_j + r_j \;\; \forall j \neq i.$$

Note that having fixed r_{-i}, i is will be the output of the GGNmax when the database is D' if and only if $r_i \geq r^*$. For all $1 \leq j \neq i \leq N$,

$$\begin{aligned} c'_i + r^* &> c'_j + r_j \\ \Rightarrow \Delta + c'_i + r^* &> \Delta + c'_j + r_j \\ \Rightarrow c'_i + (\Delta + r^*) &> (\Delta + c'_j) + r_j \\ \Rightarrow c_i + (\Delta + r^*) \geq c'_i + (r^* + \Delta) &> (\Delta + c'_j) + r_j \geq c_j + r_j \end{aligned}$$

Thus, if $r_i \geq r^* + \Delta$, then i will be the output of GGNmax on database D with randomness (r_i, r_{-i}). We again wish to relate the probability of $\Pr[r_i \geq \Delta + r^*]$ to the probability $\Pr[r_i \geq r^*]$, where $Z \sim \mathcal{N}_\beta(0, \sigma\Delta)$. Again using the fact that $GG_{\beta,\sigma}(f, D)$ satisfies (ϵ, δ)-DP,

$$\begin{aligned} \Pr[Z \geq r^*] &\leq e^\epsilon \Pr[Z \geq r^* + \Delta] + \delta \\ \Rightarrow \Pr[i|D', x_{-i}] = \Pr[r_i \geq r^*] &\leq e^\epsilon \Pr[r_i \geq r^* + \Delta] + \delta \leq e^\epsilon \Pr[i|D, x_{-i}] + \delta. \end{aligned}$$

Thus the GGNmax mechanism satisfies the same (ϵ, δ)-DP guarantee as the 1-dimensional $GG_{\beta,\sigma}(f, D)$ mechanism, when the function f has the same sensitivity Δ. □

D Additional Details and Results for Private Argmax and PATE

PATE (Private Aggregation of Teacher Ensembles) is an algorithm to train a private machine learning model. In the first step, the private dataset is partitioned into T datasets, such that a single user's data is only in single partition. A "teacher" model is trained for each partition. Then, the teacher models are collected to privately vote on how to label an unlabeled, public dataset, usually through an algorithm based off of the Report-Noisy-Max algorithm. A "student" model is trained on the privately labeled dataset. We present a high-level overview of the algorithm in Figure 7.

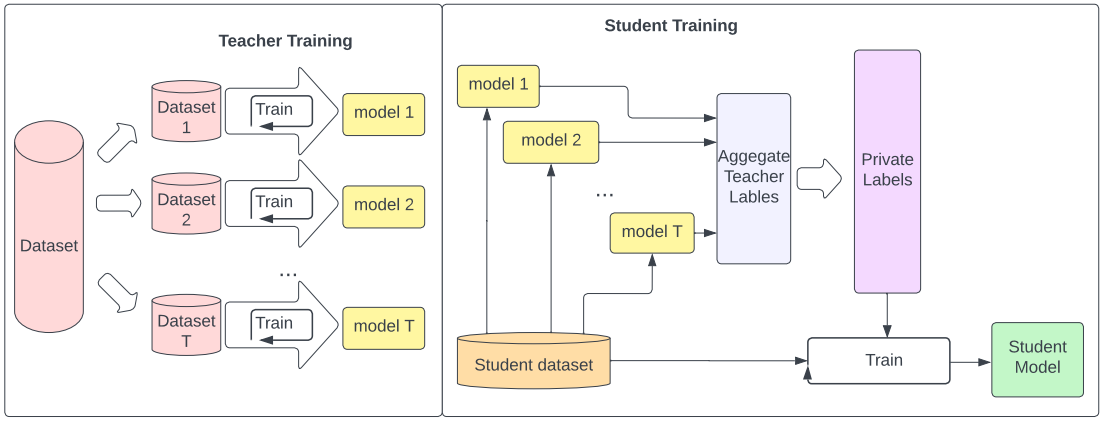

Figure 7: Diagram of PATE implementation [Papernot et al., 2017]

In the main body of the paper we primarily explore the idea of how the choice of β affects the label accuracy of PATE; in order to explore this more fully, we propose a new measure for accuracy for the private Argmax problem, which is better suited to the goals of private ML — measuring the probability of returning the true Argmax, rather than returning an outcome with score similar to the true Argmax. For this utility measure, we empirically find that $\beta = 2$ (Gaussian) is again near-optimal.

D.1 Simulations for Ensemble-Based Private Vote Aggregation

While PATE is one of the primary motivations for the GGNMax mechanism, taking a private Argmax is a very general problem and is particularly important for algorithms which attempt to reconcile beliefs (or votes) across many parties. Classical work in differential privacy on private Argmax considers that for a vector of values, the utility of the mechanism is a function of the probability that the mechanism returns any index that has a value associated with, *close* to the maximum value. However, in a task like classification in ML, there is only a single label that is the "correct" label, and thus we argue that for a task like classification in ML, a mechanism should be evaluated on how often it returns the label that would have been assigned without noise.

Building upon this intuition, we define the *Hardmax Utility* of an Argmax mechanism $\mathcal{M}$ on functions $\{f_i\}$ over a distribution $\mathcal{P}$ of databases as:

$$\text{Hardmax-Utility}_{\mathcal{P}}(\mathcal{M}, \{f_i\}) := \Pr_{D \sim \mathcal{P}}[\mathcal{M}(D, \{f_i\}) = \arg\max_i(f_i(D))].$$

With this utility measure in mind, we wish to measure the impact of β on the Hardmax utility of GG Private Argmax algorithm. Given a vector of function values $\{f_i(D)\}$, noise addition can only change the Argmax if the noise added is larger than the existing gap between the highest function value and all other values. We refer to the difference between the largest and second largest value in $\{f_i(D)\}$, as the *runner-up-gap*.

To empirically evaluate the impact of β on the Hardmax utility, we construct a set of random histograms with varying running-up-gaps; by varying the runner-up-gap, we vary the extend to which the outcome of the mechanism can be changed by outliers in the noise terms.

For our simulations, we construct 500 histograms of votes for each class as follows: for vote count V (we set $V = 1000$), maximum value v, and runner-up parameter $r \in [.001, 0.2]$, we fix the largest number of votes for class 0 at $x_0 = v$,

and the second-largest number of votes for class 1 at $x_1 = v(1-r)$, corresponding to a runner-up-gap of $vr = 100r$. We fill the remainder of the histogram by repeatedly drawing $N-2$ random integers from the range $[1, v(1-r)]$ until $\sum_{i\in[N]}(x_i) = V$. We then instantiate GGNmax on each database (histogram) with counting queries f_i that output the number of votes for each class i. For 2-class histograms, fixing V and r uniquely specifies the exact histogram used, and in that setting we ignore any randomness in the construction of the histograms.

For a given (ϵ, δ)-DP guarantee and for a range of values $\beta_i \in [1, 4]$, we compute σ_i such that (β_i, σ_i)-Generalized Gaussian Private Argmax satisfies (ϵ, δ)-DP (see Appendix B.4 for algorithmic details of this process). For each pair (β_i, σ_i), we compute the Hardmax Utility of each mechanism by computing the likelihood of the (β_i, σ_i)-Generalized Gaussian Private Argmax algorithm returning the true argmax, averaged across 50 trials.

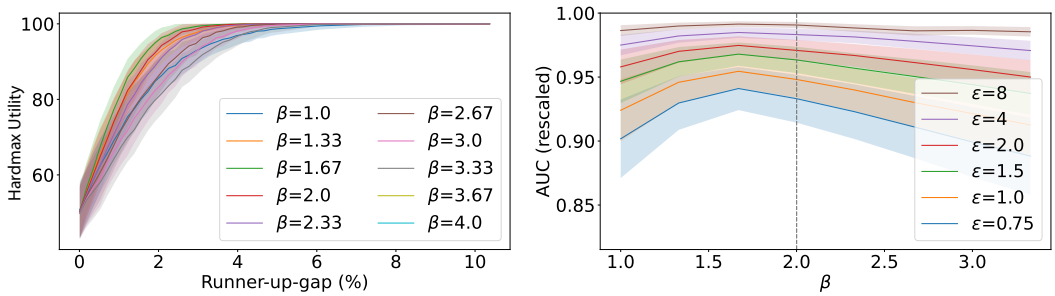

Figure 8: Hardmax Utility for 2-class histogram. Left: Hardmax Utility as a function of runner-up-gap, for mechanisms satisfying $(1, 10^{-5})$-DP. Right: Area-Under-the-Curve (AUC) of the curves on the left for different values of ϵ. AUC is rescaled so that 1.0 is the maximum.

Figure 8 shows the Hardmax Utility of (β_i, σ_i)-Generalized Gaussian Private Argmax for 2-class histograms under varying β values. The left shows the Hardmax utility as a function of the runner up gap for varying β values. We observe that varying β for a fixed runner-up-gap does not have a substantial impact on Hardmax utility, and that the optimal β value typically does not dependent on the runner-up-gap.

The right shows the Area-Under-the-Curve (AUC) of the Hardmax utility curves (similar to those on the left) as a function of β for different values of ϵ. Higher AUC means a better overall accuracy. The AUC is integrated over a range of runner-up-gap values from 0% to 10%, and AUC is rescaled such that 1.0 is the maximum. We observe that empirically, values of β close to $\beta = 2$ have better Hardmax-Utility than ones further away, regardless of the (ϵ, δ)-DP parameters. However, we do note that while $\beta = 2$ is near-optimal, β values slightly smaller than 2 do give even better performance. A key takeaway is that Gaussian does outperform Laplace, although there is room for further improvements over Gaussian by fine-tuning the β parameter.

Figure 9 extends these findings to a multi-class setting with 25 classes. On the left, we see that the choice of β has a larger impact on Hardmax utility in the multi-class setting, but that the optimal choice β is again independent of the runner-up-gap. On the right, we observe that the impact of β is also more pronounced in this multi-class setting, and that the optimal β value is much closer to 2 in this setting.

We also observe that as β grows, the AUC (right) becomes more jagged – we note that this is also true for the Hardmax Utility plot, but it is harder to observe. This is because for a fixed ϵ, the sensitivity of β on σ increase as you increase β (observable through inspection of Figure 2). Let $\sigma'(\beta, \epsilon, \delta)$ be the function that returns the minimum value of σ that satisfies (ϵ, δ)-DP for a given choice of β; the function σ' increases at an increasing rate, for a fixed value of (ϵ, δ). Importantly, in our experiments, we only consider values of σ from a set of evenly spaced values (as with β). Thus because the sensitivity increases, the gap between using the best choice of σ and using the best choice of sigma, from the set of evenly-spaced-values of σ, increases as a function of β.

E Additional β-DP-SGD Results and Implementation Details

E.1 Hyperparameters Used in Training

Hyperparameters We run our β-DP-SGD algorithm for a maximum of 100 epochs for each parameter setting and sweep over the following parameters: β (12 evenly spaced values of $\beta \in [1, 4]$), noise multiplier (6 evenly spaced values of $\sigma \in [0.5, 3.0]$), average batch size $L \in \{128, 256\}$, learning rate $\eta \in \{0.5, 1.0\}$, and clipping norm $C \in \{0.05, 0.1, 0.25, 0.5\}$, for $\delta = 10^{-6}$. Each experiment is run 3 times, which we found sufficient given standard deviations that generally fell below 0.3%.

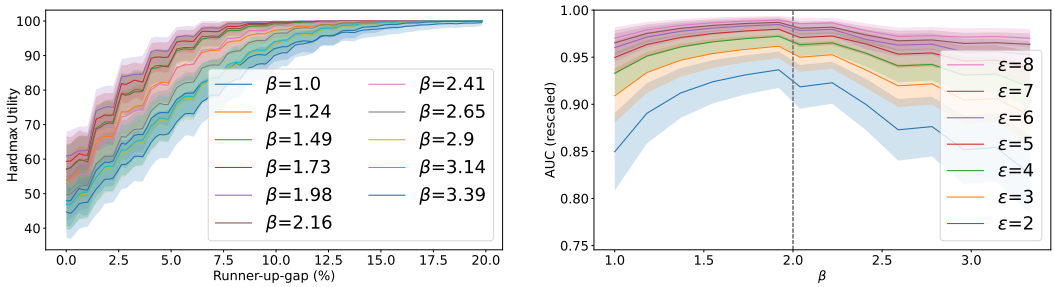

Figure 9: Hardmax Utility for 25-class histogram. Left: Hardmax Utility as a function of runner-up-gap, for mechanisms with equivalent $(2,10^{-5})$-DP. Right: Area-Under-the-Curve (AUC) of the curves on the left for different values of ϵ, where a higher AUC means a better overall accuracy. AUC is rescaled such that 1.0 is the maximum to normalize across different sizes of domain.

Datasets: We train on CIFAR-10 [Krizhevsky, 2009] and Street View House Numbers (SVHN) [Netzer et al., 2011], two common computer vision datasets, which respectively contain 60,000 and 99,289 small, color images split across ten classes; the Adult dataset [Becker and Kohavi, 1996], a tabular dataset with a binary classification task; and the IMDB dataset [Maas et al., 2011], a collection of movie reviews meant for binary sentiment classification.

Models: For the vision classification tasks (CIFAR-10 and SVHN), we use the models described in in Tramèr and Boneh [2021], which previously achieved SOTA results for the $\epsilon \leq\sim 2.5$ regime. Specifically, we train Convolutional Neural Networks (CNNs) on pretrained image features produced by scattering networks Oyallon and Mallat [2015] as described in "handcrafted CNNs". Tramèr and Boneh [2021]. For the the Adult Dataset we train a 2-layer Fully Connected Network (FCN), with 32 neurons in the hidden layer. For the IMDB dataset, we train a Long-Short Term Memory (LSTM) network with 1,081,002 parameters, in order to demonstrate the method on a relatively medium-sized model from scratch.

E.2 The Role of Individual Hyperparameters in β-DP-SGD

Below, we investigate the role of individual hyperparameters on the the final test-accuracy. We replicate the experiments of Section 5.1, but holding fixed individual hyperparameters, and varying those fixed values to determine whether that hyperparameter substantially impacts test-accuracy. Specifically, we evaluate the impact of the learning rate (E.2.1), clipping norm (E.2.2), batch size (E.2.3), and noise multiplier (E.2.4). The rest of the values are optimized over as described in E.1, and we see little effect in varying other values. We report the average maximum test-accuracy for each trial independently – this means that for each plot each value of β hyperparameters are evaluated independently, and the maximum average-test-accuracy is taken for each value of β.

In general, we observe that while some hyperparameters may have some small effect, varying these hyperparameters do not have a substantial effect on test-accuracy. They also do not impact the general relationship between β and test-accuracy: $\beta = 2$ remains near-optimal for most parameter values, although generally not exactly optimal.

As in the presentation of results in Section 5.1, some plots do not have test-accuracy values reported for specific choices of β and ϵ. This is because larger values of β tend to consume more privacy per-step, and the starting ϵ required exceeded the ϵ budget for that curve.

E.2.1 Learning Rate

Here we explore the impact of varying the learning rate on the test-accuracy, as a function of β. Figures 17 and 18 show test-accuracy results of β-DP-SGD on all four databases with respective learning rates 0.5 and 1.0. We observe that performance is relatively similar across these two figures, and thus conclude that learning rate does not have a strong impact.

E.2.2 Clipping Norm

Here we explore the impact of varying the clipping norm on the test-accuracy, as a function of β. Figures 12, 13, and 14 show test-accuracy results of β-DP-SGD on all four databases with respective clipping norms of 0.05, 0.10, and 0.25. We observe that clipping norm appears to have a bigger effect on the relationship between test-accuracy and β than learning rate, but still relatively minor and only at larger (sub-optimal) values of β.

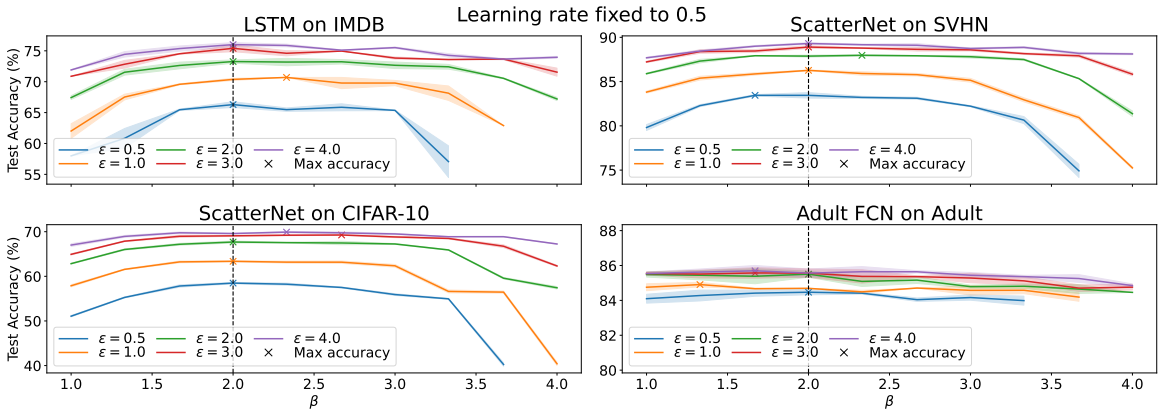

Figure 10: β-DP-SGD results with learning rate 0.5

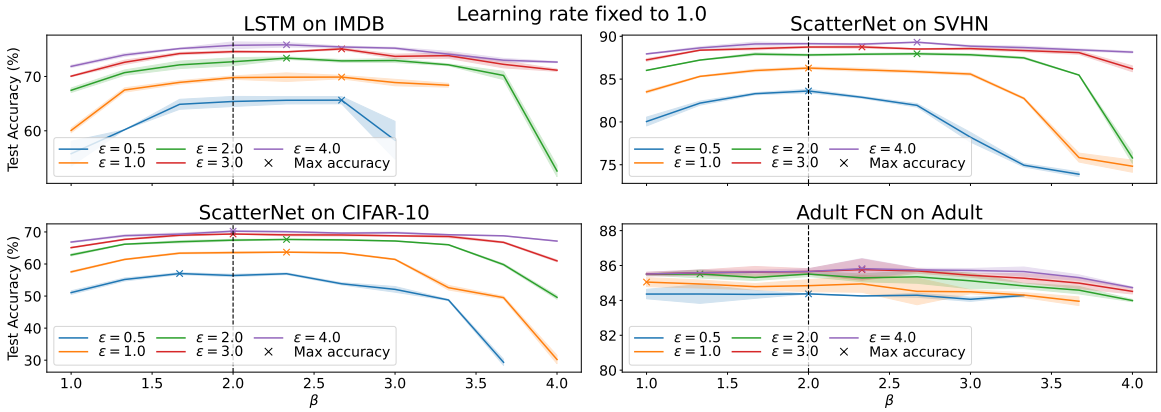

Figure 11: β-DP-SGD results with learning rate 1.0

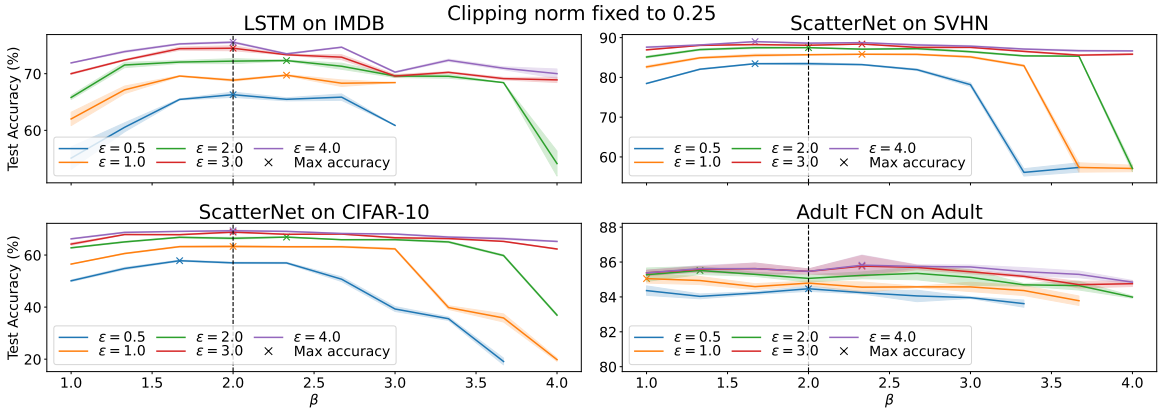

Figure 14: β-DP-SGD results with clipping norm 0.25

E.2.3 Batch Size

Here we explore the impact of varying the batch size on the test-accuracy, as a function of β. Figures 15 and 16 show test-accuracy results of β-DP-SGD on all four databases with respective batch sizes of 128 and 256. As has been

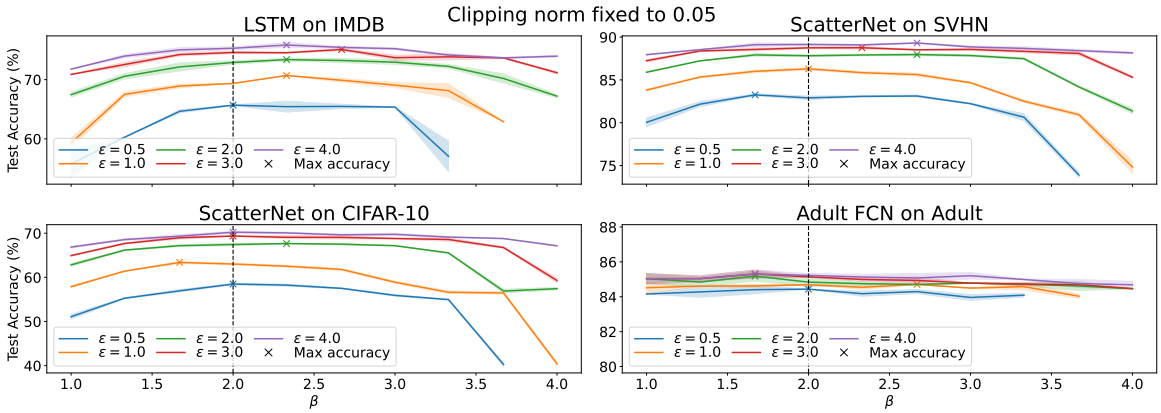

Figure 12: β-DP-SGD results with clipping norm 0.05

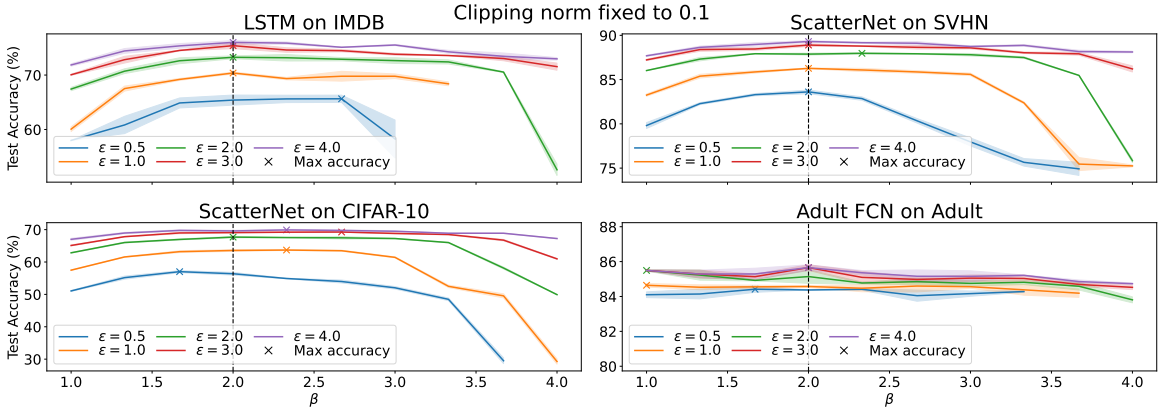

Figure 13: β-DP-SGD results with clipping norm 0.1

observed with non-private training, we observe that batch size does have a substantial effect on test-accuracy. One observation is that the for batch size set to 128, we see a fairly strong dependence on β in some settings, like ScatterNet on SVHN; however, it remains unclear where this dependence on β emerges from, and could be a fruitful direction for future research.

E.2.4 Noise Multiplier

Here we explore the impact of varying the noise multiplier σ on the test-accuracy, as a function of β. Figures 17, 18, 19, and 20 show test-accuracy results of β-DP-SGD on all four databases with noise multipliers of 1.5, 2.0, 2.5, and 3.0. While we see that the maximal performance of $\beta - DP - SGD$ remains similar (as seen in Figure 4), we see that the choice of noise multiplier has an weak effect on how sensitive training is to choice of β. We see that for larger values of σ, the test-accuracy (as a function of β) is flatter, and less dependent on choice of β – particularly for cases like the ScatterNet and LSTM. We believe this is because as one increases the noise multiplier increase, there are more training steps required to reach a specific value of (ϵ, δ), and for some models the dominant effect on test-accuracy is training time (with sufficiently good-enough parameters). We observe that across all values of noise multipliers evaluated, $\beta = 2$ remains near-optimal on all datasets and ϵ values evaluated.

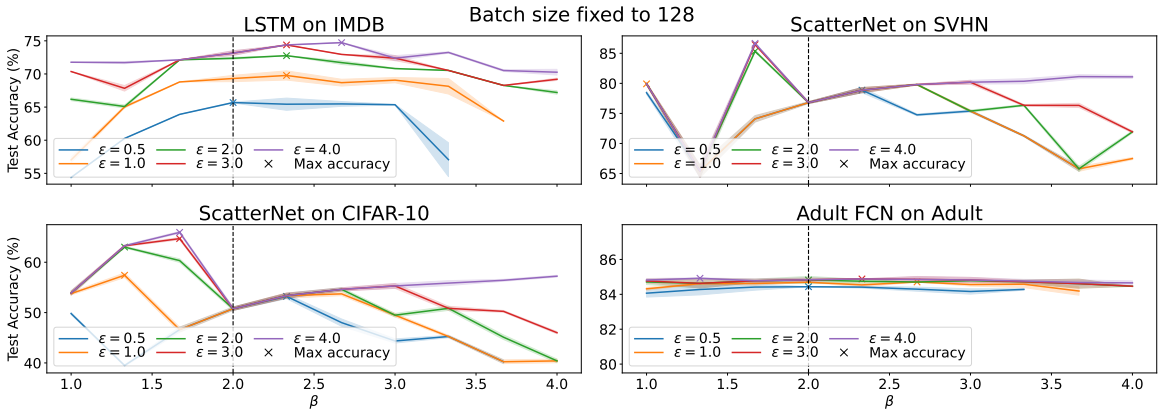

Figure 15: β-DP-SGD results with batch size 128

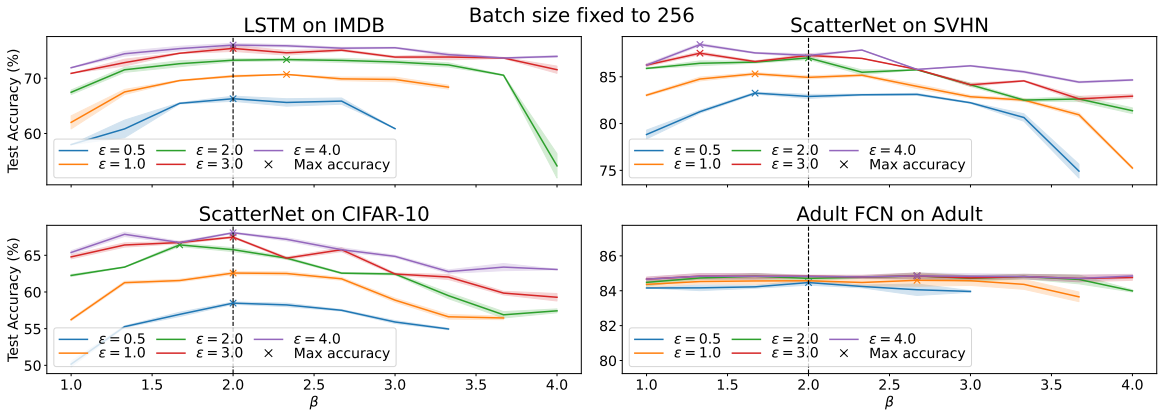

Figure 16: β-DP-SGD results with batch size 256

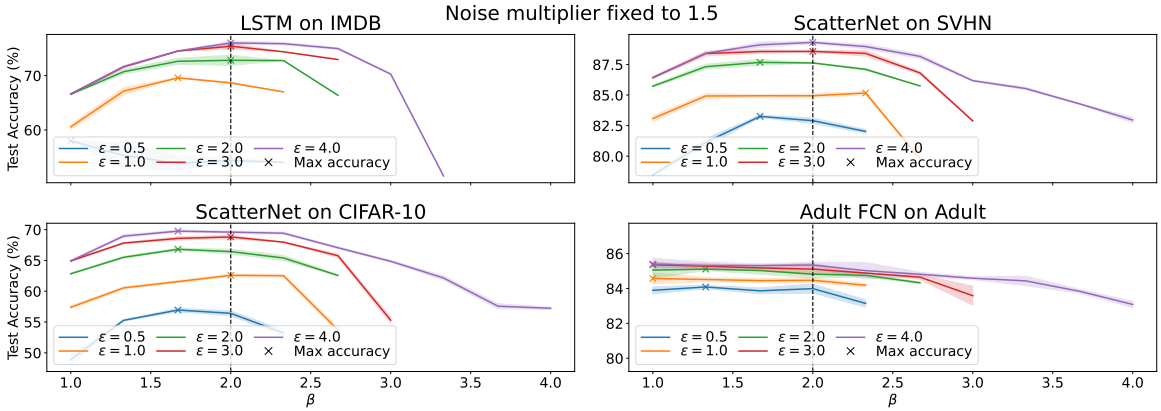

Figure 17: β-DP-SGD results with noise multiplier $\sigma = 1.5$

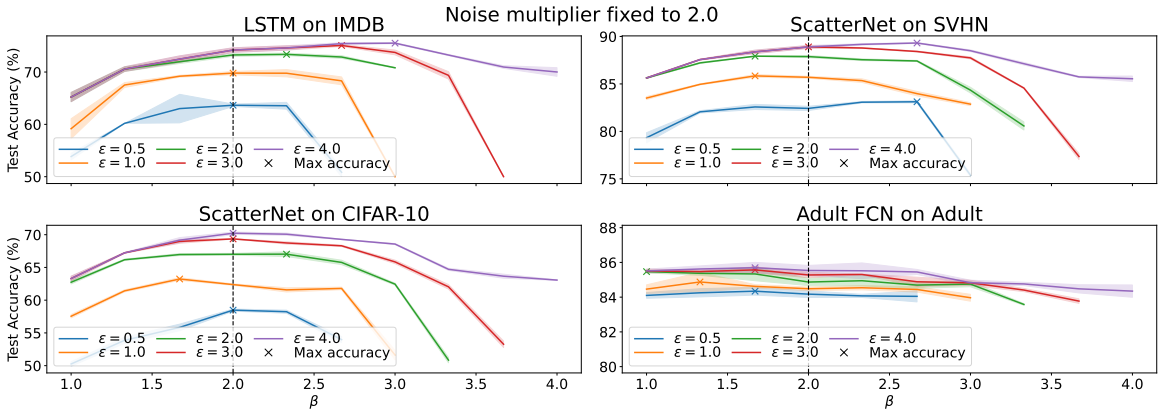

Figure 18: β-DP-SGD results with noise multiplier $\sigma = 2.0$

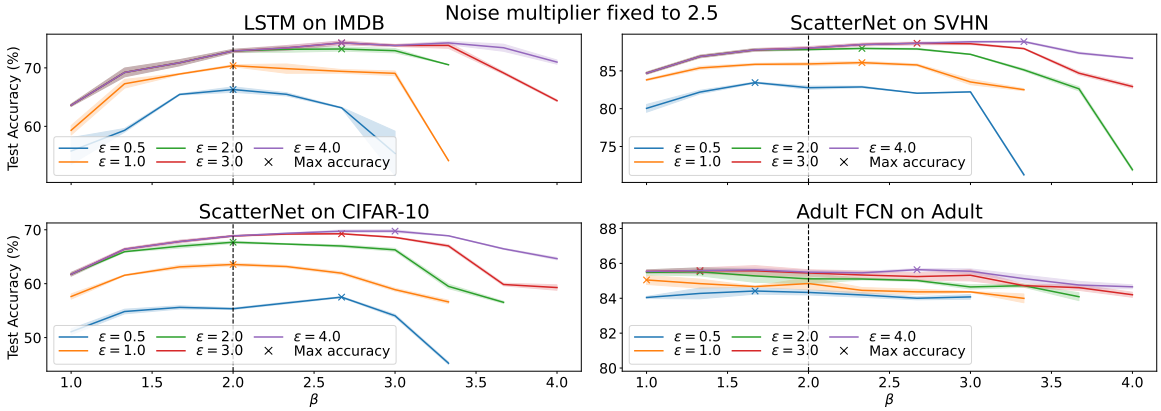

Figure 19: β-DP-SGD results with noise multiplier $\sigma = 2.5$

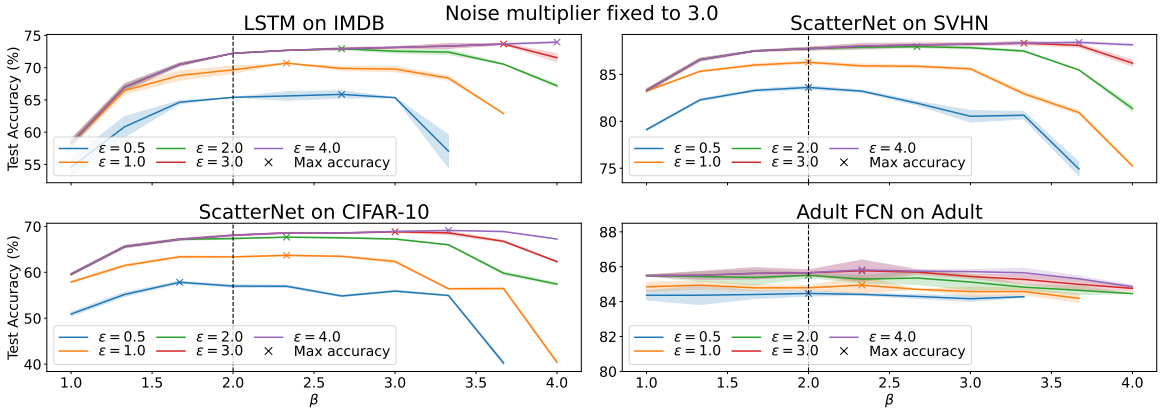

Figure 20: β-DP-SGD results with noise multiplier $\sigma = 3.0$

E.3 Sampling from the Generalized Gaussian Distribution

Unlike the Gaussian and Laplace distributions, sampling from the Generalized Gaussian mechanism is not natively supported by built-in libraries like Python's 'math' library, or the commonly used numpy library. Another commonly used library for statistical computing, SciPy, does have the 'scipy.stats.gennorm' function; however, we found that it regularly takes too long for intensive computations like stochastic gradient descent in practical settings, which involves sampling from high-dimensional gradients thousands of times. Further, the Scipy function is only able to be sampled on a CPU, which makes it ill-suited for DP-SGD, which is regularly performed on a GPU.

We implement a method for sampling from the Generalized Gaussian mechanism included in our code here: `https://github.com/RoyRin/data-aware-dp`.

In our experiments, we can sample from the Generalized Gaussian only ~ 1.3x slower than sampling from a Gaussian directly. It is possible to conduct similar sampling using the method of inverse probability transforms, since the Generalized Gaussian has a known CDF.

E.4 Reproducibility and Computing Resources

For our DP-SGDexperiments, the execution of our techniques does not result in a significant increase in processing time compared to the conventional application of DP-SGD. The only addition to the computation duration comes from increased amounts of hyperparameter searching. All experiments and data analysis are reproducible in the codebase provided `https://github.com/RoyRin/data-aware-dp`.The DP-SGD results in this paper were completed in under 500 hours of GPU time, which was split across 16 machines that were mounted on Nvidia T4 and RTX6000 machines GPUs. All data analysis was conducted on a 8 core machine with 16 GB RAM machine.