Fundamental Limitations of Favorable Privacy–Utility Guarantees for DP-SGD

Murat Bilgehan Ertan[✉*]
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands
bilgehan.ertan@cwi.nl

Marten van Dijk[*]
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands
mevd@cwi.nl

Abstract

Differentially Private Stochastic Gradient Descent (DP-SGD) is the dominant paradigm for private training, but its fundamental limitations under worst-case adversarial privacy definitions remain poorly understood. We analyze DP-SGD in the f-differential privacy framework, which characterizes privacy via hypothesis-testing trade-off curves, and study shuffled sampling over a single epoch with M gradient updates. We derive an explicit suboptimal upper bound on the achievable trade-off curve. This result induces a geometric lower bound on the separation κ, which is the maximum distance between the mechanism's trade-off curve and the ideal random-guessing line. Because a large separation implies significant adversarial advantage, meaningful privacy requires small κ. However, we prove that enforcing a small separation imposes a strict lower bound on the Gaussian noise multiplier σ, which directly limits the achievable utility. In particular, under the standard worst-case adversarial model, shuffled DP-SGD must satisfy

$$\sigma \geq \frac{1}{\sqrt{2 \ln M}} \quad \text{or} \quad \kappa \geq \frac{1}{\sqrt{8}} \left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right),$$

thus cannot simultaneously achieve strong privacy and high utility. Although the noise threshold vanishes asymptotically as $M \to \infty$, the convergence is extremely slow. Even for practically relevant numbers of updates the required noise magnitude remains substantial. We further show that the same limitation extends to Poisson subsampling up to constant factors. Our experiments confirm that the noise levels implied by this bound lead to significant accuracy degradation at realistic training settings, thus showing a bottleneck in DP-SGD under standard worst-case adversarial assumptions.

CCS Concepts

• **Security and privacy → Privacy protections**.

Keywords

differential privacy, DP-SGD, shuffling, Poisson subsampling

ACM Reference Format:
Murat Bilgehan Ertan and Marten van Dijk. 2026. Fundamental Limitations of Favorable Privacy–Utility Guarantees for DP-SGD. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830454.3832629

[*]Also with Vrije Universiteit Amsterdam.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CCS '26, The Hague, Netherlands
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2871-6/2026/11
https://doi.org/10.1145/3830454.3832629

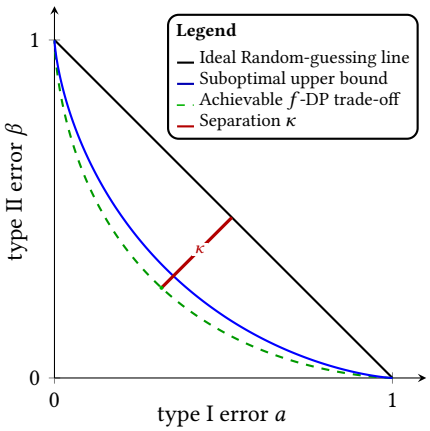


Figure 1: Trade-off view of privacy in the f-DP framework [19]. The black line shows the ideal random-guessing trade-off between type I and type II errors. The red segment κ denotes the *maximum* distance between the achievable f-DP trade-off and the ideal limit.

1 Introduction

The study of differentially private stochastic gradient descent (DP-SGD) [1] has long been motivated by the need to reconcile data privacy with the accuracy of deep neural networks. However, despite significant theoretical and empirical progress, the tension between privacy and utility remains unresolved. In particular, the degradation of training stability and test accuracy due to injected DP noise fundamentally conflicts with the requirement to add sufficient noise for rigorous differential privacy guarantees.[1]

Existing analyses of DP-SGD have primarily relied on the *Poisson subsampling* model [1, 8], in which each record is independently included in a mini-batch with a fixed probability. This assumption, introduced by Abadi et al. [1], enables privacy accounting techniques like Rényi Differential Privacy [37] and privacy-loss distributions for numerically tight composition [31, 36]. However, it is primarily a modeling convenience. In practice, modern deep learning systems do not sample examples independently but instead shuffle the entire dataset and partition it into fixed-size mini-batches each epoch. This shuffling-based batching is far more efficient computationally but breaks the independence assumptions exploited in Poisson

[1]The full version of this paper, containing all appendices referred to here, is available at `arXiv:2601.10237`. All appendix references in this paper point to that version.

analyses, making theoretical guarantees harder to derive. As a result, most practical DP-SGD implementations train using shuffled batches while reporting privacy parameters as if Poisson subsampling were used [13, 14, 39]. This persistent mismatch between analytical assumptions and real implementations motivates our study of *random shuffling* as a more realistic sampling mechanism for understanding DP-SGD's privacy behavior.

Building on these observations, our analysis is carried out entirely within the f-DP framework [19], which provides a complete hypothesis-testing description of privacy and allows us to reason directly about the trade-off curves induced by shuffled DP-SGD. We focus on the single-epoch shuffling regime under the standard *worst-case adversarial model*, in which each data record appears in exactly one batch per epoch. In the standard worst-case adversarial model, the adversary may use the observable noisy batch updates together with any auxiliary information (such as batch size) available to infer the contribution of the differing data record; the formal model is given in Section 4. This setting eliminates the independence structure that underpins Poisson-based analyses and introduces fundamental dependencies in the adversary's observation model. We explicitly formalize this *adversarial observation model*, which is typically left implicit in prior work, and analyze its implications through the geometry of the f-DP trade-off curve. Finally, while our primary results concern shuffled DP-SGD, we show via a rigorous mixture argument that the same limitations extend, up to constant factors, to Poisson subsampling. We additionally derive an asymptotic separation bound for Poisson subsampling within the μ-GDP framework [19] which relies on asymptotic convergence without an explicit rate. Together, these results yield a *unified worst-case picture* covering both the theoretical (Poisson) and practical (shuffling) sampling schemes, and demonstrate that our findings propagate to the privacy parameters reported in standard DP-SGD practice.

Before presenting our main claim, we briefly recall how privacy is parameterized in DP-SGD. In the classical (ε, δ)-DP formulation, a randomized algorithm is private when its output distributions on two neighboring datasets are nearly indistinguishable to any adversary; smaller values of ε therefore correspond to stronger privacy guarantees. In DP-SGD, this effect is achieved by two mechanisms: each per-example gradient is clipped to an ℓ_2 radius C, which bounds its maximum possible influence, and Gaussian noise with standard deviation $C\sigma$ is added to every batch update. The clipping constant C controls sensitivity, while the noise multiplier $\sigma > 0$ is the dominant driver of the resulting privacy parameters (ε, δ). Smaller σ improves test accuracy but necessarily weakens privacy, whereas larger σ strengthens privacy at the expense of utility.

In the f-DP framework [19], an adversary attempting to distinguish two neighboring datasets forms a test with *false positive rate* (FP) α and *false negative rate* (FN) β. The resulting *trade-off function* $f(\alpha)$ gives the minimum achievable FN error among all tests with FP error at most α. Thus, f captures the fundamental FP–FN trade-off faced by the strongest possible adversary. The ideal privacy baseline is the random-guessing line $\beta = 1 - \alpha$, corresponding to zero leakage. Section 3.3 provides further background on the f-DP formalism.

A key geometric quantity in this picture is the *separation* between the trade-off curve and the random-guessing line (perfect privacy). We define

$$\mathrm{sep}(f) \;=\; \max_{\alpha\in[0,1]} \min_{\gamma\in[0,1]} \big\| (\alpha, f(\alpha)) - (\gamma, 1-\gamma) \big\|_2,$$

that is, the *maximum* distance between a point on the trade-off curve $(\alpha, f(\alpha))$ and the ideal line $(\gamma, 1-\gamma)$. When f is symmetric and convex, this separation is attained at the fixed point $\hat{a}$ satisfying $f(\hat{a}) = \hat{a}$, and we can write $\kappa = \mathrm{sep}(f) = (1 - 2\hat{a})/\sqrt{2}$. Smaller values of $\mathrm{sep}(f)$ (and hence smaller κ) correspond to stronger privacy, as the entire trade-off curve lies closer to random guessing. This privacy interpretation will be formalized and analyzed rigorously later.

This notion of separation follows the f-DP interpretation introduced by Dong et al. [19], applied to deep learning by Bu et al. [11], and later formalized by Zhu et al. [52]. Figure 1 provides an early geometric view of this hypothesis-testing landscape for shuffled DP-SGD. The vertical segment κ illustrates the separation $\mathrm{sep}(f)$ for the corresponding trade-off curve. In the remainder of the paper, we show that under the standard worst-case adversarial model, *our explicit upper bound on the shuffled trade-off curve implies a lower bound on* $\kappa = \mathrm{sep}(f)$. Combining these elements yields our main conceptual consequence, an unavoidable *or* constraint for one-epoch shuffled DP-SGD with M rounds. Under the standard worst-case adversarial model, the mechanism must satisfy:

$$\sigma \;\geq\; \frac{1}{\sqrt{2\ln M}} \qquad \textbf{or} \qquad \kappa \;\geq\; \frac{1}{\sqrt{8}}\left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right)$$

Both cannot be made small simultaneously. A large noise multiplier σ hurts accuracy, while a large separation κ indicates substantial privacy leakage. While the quantity $1/\sqrt{\ln M}$ vanishes asymptotically as $M \to \infty$, this decay is extremely slow. Even for astronomically large values of M, the required noise level remains far from zero. Our lower bound thus translates into a privacy–utility tension. In the shuffled setting and under the standard worst-case adversary of the DP framework, DP-SGD cannot operate in a regime where both privacy and utility are simultaneously strong. As our analysis confirms, this tension is not an artifact of shuffling; the same lower bound applies to Poisson subsampling (up to a constant factor), suggesting that the bottleneck originates in the worst-case adversarial framework, which both sampling schemes share. We complement our theoretical results with experiments illustrating the practical severity of this trade-off. In particular, we show that the noise levels implied by our lower bound already lead to substantial degradation in test accuracy under realistic training budgets. These empirical results confirm that the theoretical limitations we identify also manifest concretely at finite parameter values encountered in practice. In classical terms, this threshold has an immediate (ε, δ)-DP consequence. Whenever $\sigma < 1/\sqrt{2\ln M}$, any (ε, δ)-DP guarantee reported for the same mechanism must satisfy $\varepsilon \geq \ln\big((1 + \kappa\sqrt{2} - 2\delta)/(1 - \kappa\sqrt{2})\big)$ for the corresponding separation lower bound κ (Corollary 6.2). Already at the one-epoch baseline with $\delta = 1/N$ and $N = 10^8$ this forces $\varepsilon \gtrsim 1$, and standard privacy accountants report $\varepsilon \gg 10$ at realistic multi-epoch budgets (Section 7).

Our result is a *converse*. Accounting-based analyses fix σ and ask how private the mechanism is. We ask how small σ can be if worst-case privacy is to remain nontrivial. Numerical privacy-loss lower bounds for shuffled DP-SGD are known at specific parameter instances [13, 14], but to our knowledge none gives a *closed-form* bound on the minimum admissible σ as a function of M. We emphasize that the limitation we prove is not inherent to DP-SGD as an algorithm, but rather to the adversarial assumption framework under which its privacy is analyzed. In particular, the standard definition implicitly grants the adversary full visibility into every component of the stochastic updates, making certain forms of privacy leakage unavoidable. Our main contributions are as follows:

- We analyze single-epoch shuffled DP-SGD under the standard worst-case adversarial model in the f-DP framework and derive an explicit suboptimal upper bound on its trade-off curve.
- This yields a geometric lower bound on the separation κ, exposing a fundamental privacy–utility trade-off: the noise multiplier σ and separation cannot be driven below explicit lower bounds at the same time. Then, via a mixture argument, we show that Poisson subsampling inherits the same limitation up to a constant factor, providing a unified worst-case characterization of both sampling schemes.
- We translate our f-DP bounds into consequences for classical (ε, δ)-DP, obtaining a closed-form lower bound on the minimal admissible ε for fixed δ (Corollary 6.2), and verify the implication independently with the numerically tight shuffled privacy accountant of Chua et al. [14] (Section 7).
- Empirically, we validate the predicted threshold behavior, showing that the noise levels implied by our bounds already cause substantial accuracy degradation across models, datasets, and batch sizes.

Organization. The remainder of the paper is structured as follows. Section 2 reviews the most relevant prior work. Section 3 introduces the differential privacy framework, the sampling mechanisms used in practice, and the BatchSGD update model. Section 4 formalizes the worst-case adversarial observation model that underlies our analysis. In Section 5, we introduce a suboptimal hypothesis testing framework for DP-SGD and define the separation metric κ formally. Next, Section 6 develops the geometric tools needed to analyze this metric, establishes our main separation lower bound for shuffled DP-SGD, and extends the result to Poisson subsampling via a mixture argument and yielding a unified worst-case picture for both sampling schemes. Section 7 presents empirical illustrations of our theoretical findings, and Sections 8 and 9 conclude with broader implications and future directions.

2 Related Work

Differential Privacy (DP) [1, 21, 22]has emerged as a principled framework for preserving individual information in data analysis and model training. Among its applications, Differentially Private Stochastic Gradient Descent (DP-SGD) [1] is the most widely used privacy-preserving method in deep learning, injecting calibrated noise into gradient updates to satisfy a target privacy budget. It has also been studied for various machine learning domains [3, 15, 35], and adopted in real-world systems from the U.S. Census [2] to Microsoft telemetry [16] and Google's recent initiative to train large language models from scratch under DP [45]. It is now supported by major privacy libraries such as JAX Privacy and Opacus [7, 49], which provide practical implementations and privacy accounting tools for large-scale training.

Analytically, the privacy guarantees of DP-SGD have been refined through composition frameworks such as Rényi Differential Privacy (RDP) [37] and Privacy-Loss Distributions (PLD) [31, 36], which enable tighter accounting of cumulative privacy loss across training iterations. The classical (ε, δ)-DP [21, 22] formulation summarizes privacy using two scalar parameters, whereas the emerging f-DP framework [11, 19]describes privacy through the entire *trade-off curve* of optimal type I/type II errors, providing a complete and geometrically interpretable characterization of indistinguishability. Importantly, the trade-off curve is the most informative DP representation. As shown in [19] (Appendix A and B), all standard DP notions, including (ε, δ)-DP and divergence-based relaxations, can be *derived* from it. Every f-DP trade-off curve induces a corresponding family of (ε, δ)-DP guarantees, and conversely each (ε, δ)-DP guarantee imposes a linear constraint on the trade-off curve [19]. This relationship allows techniques developed for (ε, δ)-DP to be imported into the f-DP framework. This motivates our decision to conduct all analysis directly in the f-DP framework.

Privacy amplification by shuffling, introduced by Erlingsson et al. [23] and later refined by Feldman et al. [24, 25], showed that randomly permuting locally privatized reports can strengthen privacy guarantees in the central model. However, this amplification applies to locally randomized mechanisms rather than to DP-SGD's centralized setting.

The most closely related analysis of shuffled DP-SGD is due to Chua et al. [13, 14], who study the mechanism under their *Adaptive Batch Linear Queries* (ABLQ) framework. For the shuffled sampler, their results provide *lower bounds on the privacy loss* ε for single- and multi-epoch training. In the f-DP viewpoint, these correspond to *upper bounds* on the achievable trade-off function, the same direction as our analysis, in which an explicit upper bound on the trade-off curve yields a lower bound on the separation κ. Importantly, they also show that random shuffling in DP-SGD leads to *weaker provable privacy guarantees* than Poisson subsampling, highlighting a discrepancy with the common practice of reporting Poisson-based parameters for shuffled implementations [39]. Our analysis agrees in direction, and differs in form and scope. Their lower bounds are given by analytic expressions that must be evaluated, and inverted, numerically for each parameter setting, whereas our max-statistic construction yields a simple *closed-form* upper bound on the shuffled trade-off curve and, in turn, an explicit threshold $\sigma \geq 1/\sqrt{2 \ln M}$ as a function of M alone. Moreover, the separation formulation applies uniformly over all f-DP families, transfers to Poisson subsampling via an explicit mixture argument (Section 6.3), and translates to closed-form (ε, δ) consequences (Corollary 6.2).

Concurrently, a growing line of work proposes allocation-style samplers that retain the implementability of shuffling while recovering Poisson-like accounting. Examples include the Balls-and-Bins sampler [12] and the k-out-of-t allocation scheme, in which each

record is assigned to k of the t rounds, introduced as balanced iteration subsampling [17] (random allocation [44]), with analytic guarantees and efficient privacy-loss-distribution accounting [26] and a two-limit optimality result [18]. These are achievability analyses, and they are complementary to ours. At $k = 1$ these samplers place each record in exactly one uniformly random round, so the tightly dominating pair identified for the Balls-and-Bins sampler in [12, Thm. 3.1] coincides, after rescaling by σ, with the hypothesis pair we analyze in (3)–(4), so the dichotomy of Theorem 6.1 applies to them unchanged under the same assumptions (which does not imply that their trade-off curves coincide with f_{shuf}).

The random-shuffling regime remains the standard sampling strategy in modern deep learning practice. It has been extensively studied from an optimization perspective, where several works analyze its convergence behavior and compare it to sampling with replacement [27, 28]. Its widespread adoption instead reflects training conventions and practical considerations. Poisson subsampling does not scale well to large datasets because random access is inefficient for data that do not fit in memory, and the resulting variable batch sizes are inconvenient to handle in standard deep learning frameworks [13]. As a result, shuffling-based batching remains the default in modern training systems, even though the corresponding privacy analyses continue to rely on Poisson-based assumptions [39].

Apart from privacy accounting and sampling considerations, it is now well established that training deep neural networks with DP-SGD often incurs a substantial loss in utility, with prior work consistently reporting degradation in accuracy and generalization under meaningful privacy guarantees [4, 9, 15, 33, 34, 43, 46]. A common insight across these works is that the injected Gaussian noise, controlled by the noise multiplier σ, is one of the dominant sources of this degradation, acting as a barrier to effective training.

3 Preliminaries

Differential Privacy (DP) guarantees that a randomized mechanism produces statistically similar outputs on any two neighboring datasets differing in a single individual's record. Within stochastic gradient descent (SGD), the DP-SGD algorithm [1] enforces this requirement by first bounding the effect of each example via gradient clipping, and then injecting Gaussian noise to obscure the contribution of any individual.

DP-SGD modifies the standard mini-batch update in two ways: (i) each per-sample gradient is clipped to a fixed threshold, bounding sensitivity, and (ii) Gaussian noise scaled to this threshold is added to the aggregated gradient before updating the model parameters. The precise mechanism is formalized later in Section 3.2. Moreover, the privacy of DP-SGD depends critically on how mini-batches are sampled. We therefore introduce the two canonical sampling mechanisms, namely *Poisson Subsampling* and *Random Shuffling* in the following subsection.

Scope of analysis. We analyze the privacy guarantees of a *single epoch* of DP-SGD under the standard worst-case adversarial model (see Section 4). In this setting the sampler produces index sets $S_1, \ldots, S_M$, each used exactly once to form a mini-batch. Extensions to multiple epochs require additional composition and are beyond the scope of our core lower bound.

3.1 Sampling Mechanisms in DP-SGD

The sampling mechanism plays a central role in the privacy analysis of DP-SGD because subsampling provably amplifies differential privacy guarantees. When each update is computed on a random mini-batch, a given record influences the mechanism only in the rounds where it is sampled, thereby reducing the likelihood of leaking information about that individual [6]. While various sampling strategies have been proposed [39], two have become canonical in both theory and practice: *Poisson subsampling* and *random shuffling*. The former is analytically convenient and therefore ubiquitous in theoretical work; the latter is the standard choice in practical deep learning systems. Both determine how mini-batches are formed from a dataset $d = \{\xi_1, \ldots, \xi_N\}$, but they exhibit fundamentally different statistical dependencies and therefore lead to different privacy-amplification behaviors. We formalize both mechanisms below, with pseudocode in Appendix B.

Definition 3.1 (Poisson Subsampling). Given a dataset d of size N, Poisson subsampling includes each element independently with probability $q \in (0, 1)$ at each round $j = 1, \ldots, M$. This produces random index sets $S_j \subseteq [N]$ with expected size $\mathbb{E}[|S_j|] = qN$, and independence is maintained both across records and across rounds.

Definition 3.2 (Random Shuffling). Given a dataset d of size N and a fixed number of rounds M, random shuffling samples a uniform permutation π of the dataset indices and then forms M disjoint batches of equal size $b = \lfloor N/M \rfloor$ by taking consecutive blocks of π. Concretely, for each $j = 1, \ldots, M$, we have $S_j = \{\pi_{(j-1)b+1}, \ldots, \pi_{jb}\}$. So each record appears in at most one batch per epoch, and the last $N - Mb = N \bmod M$ permuted indices are discarded.

In the remainder of this work, we focus on the random-shuffling mechanism, as it captures practical DP-SGD implementations and introduces inter-round dependencies that are absent under Poisson subsampling. As we later show, our main limitation results extend to Poisson subsampling up to constant factors, yielding a unified worst-case characterization for both sampling schemes.

3.2 DP-SGD Mechanism

We formalize the update rule used throughout this work. Let $d = \{\xi_1, \ldots, \xi_N\}$ denote the training dataset of size N, and let $S_1, \ldots, S_M \subseteq [N]$ be index sets produced by a sampling mechanism (Poisson subsampling or random shuffling). The corresponding mini-batch at iteration j is $B_j = \{\xi_i : i \in S_j\}$.

Given a loss function $\ell(w; \xi)$, the empirical risk is

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(w; \xi_i),$$

and the per-sample gradient at iteration j is

$$g(w_j; \xi_i) = \nabla_w \ell(w_j; \xi_i).$$

Definition 3.3 (Gradient Clipping). Given a clipping threshold $C > 0$, the clipped version of a per-sample gradient $g(w; \xi)$ is

$$[g(w; \xi)]_C = g(w; \xi) \cdot \min\left(1, \frac{C}{\|g(w; \xi)\|_2}\right).$$

Thus $\|[g(w; \xi)]_C\|_2 \leq C$ for all ξ, ensuring bounded sensitivity of the aggregated update.

Algorithm $\mathcal{B}$ DP-SGD Mechanism [1]

Input: Dataset d, number of rounds M, learning rate η, clipping bound C, noise multiplier σ (optional)
for each epoch **do**
  Generate mini-batches $\{S_1, \ldots, S_M\}$ using either Poisson subsampling (Def. 3.1) or random shuffling (Def. 3.2)
  for each S_j **do**
    Compute per-sample gradients $g(w_j; \xi_i)$ for all $i \in S_j$
    Clip each gradient:

$$[g(w_j; \xi_i)]_C = g(w_j; \xi_i) \cdot \min\Big(1, \frac{C}{\|g(w_j; \xi_i)\|_2}\Big)$$

    Compute clipped batch sum:

$$G_j = \sum_{i \in S_j} [g(w_j; \xi_i)]_C$$

    Add Gaussian noise and average:

$$\mathbf{z}_j \sim \mathcal{N}\big(0, (C\sigma)^2 I\big), \qquad \tilde{G}_j = \frac{1}{|S_j|}\Big(G_j + \mathbf{z}_j\Big)$$

    Update model: $w_{j+1} \leftarrow w_j - \eta\, \tilde{G}_j$
  end for
end for

Definition 3.4 (Gaussian Noise Mechanism). Given a clipping threshold C and noise multiplier $\sigma \geq 0$, the DP-SGD mechanism adds Gaussian noise

$$\mathbf{z} \sim \mathcal{N}\big(0, (C\sigma)^2 I\big),$$

where I is the identity matrix (i.e., independent Gaussian noise per coordinate). The standard deviation $C\sigma$ scales proportionally with both the clipping threshold and the noise multiplier.

At each iteration j, the mechanism processes the mini-batch S_j by computing clipped per-sample gradients, adding Gaussian noise as in Definition 3.4, averaging by the batch size, and then performing an SGD update step. The Algorithm $\mathcal{B}$ summarizes the procedural implementation, including batch generation. For $\sigma = 0$ and $C = \infty$, this procedure reduces to standard mini-batch SGD; for $\sigma > 0$ with finite C, it corresponds to the DP-SGD update of [1].

3.3 Differential Privacy and f-DP

We briefly recall the classical (ε, δ) formulation of differential privacy and its refinement through the f-DP framework, both of which rely on a notion of adjacency between datasets.

Two datasets that differ only in the contribution of a single individual are called *adjacent*, written $d \sim d'$. This notion can be formalized in several ways, depending on how one models the presence, replacement, or removal of a record. We will introduce the standard adjacency relations shortly, but for any such choice, differential privacy is defined as follows.

Definition 3.5 (Differential Privacy (DP) [21]). A randomized mechanism $\mathcal{M}$ is (ε, δ)*-DP* with respect to an adjacency relation $\sim$ if, for all adjacent datasets $d \sim d'$ and all measurable events E,

$$\Pr[\mathcal{M}(d) \in E] \;\leq\; e^{\varepsilon} \Pr[\mathcal{M}(d') \in E] + \delta.$$

The f-DP framework characterizes privacy through the entire *trade-off function* $f(\alpha)$, which captures the best achievable Type I and Type II error trade-off when distinguishing d from d'. This viewpoint provides a complete geometric representation of privacy, supports tight composition, and matches the Gaussian mechanism central to DP-SGD. In addition, DP relies on a formal notion of *adjacency* between datasets. Several adjacency models appear in the literature; we record them formally below.

Definition 3.6 (Substitution Adjacency). Let $\mathcal{X}$ denote the space of valid data records. Datasets d and d' are *substitution neighbors* if they have equal cardinality and differ in exactly one record, i.e., if there exist records $\xi \in d$ and $\xi' \in \mathcal{X}$ with $\xi' \neq \xi$ such that

$$d' = \big(d \setminus \{\xi\}\big) \cup \{\xi'\}.$$

Thus $|d| = |d'|$, and ξ' takes the place of the differing individual's record ξ. Throughout, ξ and ξ' denote individual data records (elements of $\mathcal{X}$), whereas mini-batches are denoted B_j (cf. Section 3.2).

Definition 3.7 (Add/Remove Adjacency). Datasets d and d' are *add/remove neighbors* if one is obtained from the other by inserting or deleting a single record. Equivalently, $|d'| = |d| \pm 1$ and

$$d' = d \cup \{\xi\} \quad \text{for some } \xi \in \mathcal{X}, \quad \text{or}$$
$$d' = d \setminus \{\xi\} \quad \text{for some } \xi \in d.$$

Definition 3.8 (Zero-Out Adjacency [14, 30]). Let $\mathcal{X}_\perp := \mathcal{X} \cup \{\perp\}$ be the augmented record space, where $\perp$ denotes a ghost record whose contribution to any query or gradient is defined to be zero.

For a dataset $d = (d_1, \ldots, d_N) \in \mathcal{X}_\perp^N$ and an index $i \in [N]$, let d_{-i} denote the dataset obtained by removing the i-th entry. Two datasets $d, d' \in \mathcal{X}_\perp^N$ are said to be *zero-out adjacent* if there exists an index $i \in [N]$ such that

$$d_{-i} = d'_{-i}, \qquad \{d_i,\, d'_i\} = \{\perp,\, \xi\} \text{ for some } \xi \in \mathcal{X},$$

that is, the datasets are identical except at a single position where one contains a genuine record and the other contains the ghost record $\perp$.

When DP-SGD uses Poisson subsampling, each element is included in a batch independently with probability q, so the batch size itself is random. The privacy question is therefore whether a given record is *present or absent*, which corresponds to *add/remove adjacency* (Definition 3.7). This is the pairing adopted in the subsampling-amplification literature [6], and in subsequent f-DP and privacy-accounting analyses of DP-SGD [11]. Under add/remove adjacency, the worst-case ℓ_2-sensitivity of the clipped gradient sum is at most C, since the two neighboring sums differ by a single clipped gradient of norm at most C (with equality when a record's gradient reaches the clipping threshold), and the Gaussian mechanism accordingly uses noise of scale $C\sigma$.

In contrast, the original f-DP analysis of DP-SGD [19] is carried out under *substitution adjacency* (Definition 3.6), where the two datasets have equal size and differ in one record. In this case the worst-case ℓ_2-sensitivity of the clipped gradient sum is at most $2C$, the maximum possible distance between two clipped gradients. This imposes an extra factor of 2 on the required Gaussian noise scale in substitution-based f-DP analyses. We stress that sensitivity is a statement about the worst case over neighboring datasets. It places no requirement on how any individual record's gradient behaves.

A third notion, *zero-out adjacency* (Definition 3.8) [14, 30], replaces the differing record with a ghost record $\perp$ whose gradient is zero. The worst-case sensitivity is then at most C (the distance between a clipped gradient and the zero vector, attained when the clipping threshold is reached), and, unlike add/remove adjacency, the total dataset cardinality N is preserved. The latter property is precisely what our shuffling analysis requires. The ghost-record coupling keeps the index sets $S_1, \ldots, S_M$ identically distributed under both neighboring datasets, so the two hypotheses differ only in the contribution of a single position.

For the remainder of this work we therefore analyze DP-SGD under **zero-out adjacency**, so that our bounds reflect the tightest possible worst-case guarantees. We emphasize that zero-out adjacency is an *analysis device*, not an algorithmic modification. The DP-SGD mechanism (Algorithm $\mathcal{B}$) is unchanged, and only the pair of neighboring datasets compared in the privacy analysis changes. All of our results translate to substitution adjacency at the cost of a factor-of-two increase in the required noise scale, which would only strengthen our conclusions.

4 Adversarial Model

A key element in any differential privacy analysis is the choice of adversary and the information they are assumed to observe. In DP-SGD, the adversary's view is determined by what portion of the stochastic updates (gradients, batch structure, and added noise) are observable. We consider two notions: a *practical* adversary corresponding to realistic communication constraints, and a *worst-case* adversary required for proving formal differential privacy guarantees.

4.1 Practical Adversary

In many distributed or federated training settings, an adversary can observe only the messages transmitted to the server. These messages are the *noisy aggregated gradients* produced by DP-SGD. The following assumption captures this realistic eavesdropper model.

Assumption 4.1 (Practical Adversary). In each round $j = 1, \ldots, M$, the adversary observes only the (noisy, averaged) update

$$\tilde{G}_j = \frac{1}{|S_j|}\left(\sum_{i\in S_j}[g(w_j;\xi_i)]_C + \mathcal{N}(0,(C\sigma)^2 I)\right),$$

and has no access to individual gradients or intermediate computations.

4.2 Worst-Case Adversary

To enable standard DP arguments, analyses in the literature traditionally endow the adversary with additional knowledge that is not realistically available but simplifies the analysis under the adjacency models defined in Section 3.3.

We adopt *zero-out adjacency* (Definition 3.8): neighboring datasets d and d' both have size N and share the first $N-1$ records. They differ only in the N-th element. In d it is a ghost record $\perp$ whose gradient is defined to be zero ($[g(w_j;\perp)]_C = 0$), while in d' it is a valid record ξ_N. Then, we give the adversary access to the sum of all clipped gradients *except* the potentially differentiating one (index N). Furthermore, we assume the batch size $|S_j|$ is public knowledge (or observable from metadata).

Assumption 4.2 (Auxiliary Knowledge Under Zero-out Adjacency). Under zero-out adjacency (Definition 3.8), for each round j the adversary knows the batch size $|S_j|$ and the partial sum of clipped gradients excluding the potentially differing record:

$$G_j^{(-)} = \sum_{i\in S_j\setminus\{N\}}[g(w_j;\xi_i)]_C.$$

Given the observable update $\tilde{G}_j$ from Assumption 4.1 and the auxiliary information from Assumption 4.2, the adversary can deterministically recover the noisy contribution of the differing record.

Proposition 4.3 (Isolation of the Individual Contribution). *Let $\delta_{N\in S_j}$ denote the indicator that N belongs to batch S_j. Under Assumptions 4.1 and 4.2, the adversary can compute*

$$Z_j := |S_j|\,\tilde{G}_j - G_j^{(-)}.$$

Consequently, when the dataset is d (ghost record $\perp$ present), the contribution is zero and the adversary reconstructs pure noise:

$$Z_j = \mathcal{N}(0,(C\sigma)^2 I), \tag{1}$$

and when the dataset is d' (valid record ξ_N present), the reconstruction becomes

$$Z_j' = \delta_{N\in S_j}[g(w_j;\xi_N)]_C + \mathcal{N}(0,(C\sigma)^2 I). \tag{2}$$

Consequence. Since the worst-case adversary can mathematically reconstruct these differential terms from the public observations and the auxiliary knowledge under Assumptions 4.1 and 4.2, analyzing the indistinguishability of (1) and (2) is sufficient to bound the privacy leakage of the entire protocol. Hence our analysis may focus entirely on the reconstructed variables Z_j and Z_j' for the remainder of this work. In essence, we take the worst-case adversary in our analysis to be the one that distinguishes Z_j and Z_j', and the present subsection motivates this definition. Other functions of the adversary's view may also carry information about the differing record, and the auxiliary knowledge of Assumption 4.2 already includes the batch cardinality $|S_j|$. Since our results are lower bounds obtained by exhibiting an explicit hypothesis test on Z_j and Z_j', restricting attention to these variables can only understate the adversary, and the resulting bounds remain valid.

5 Hypothesis Testing and f-DP Characterization

Having defined the adversarial view, we now characterize privacy leakage as a statistical hypothesis testing problem between neighboring datasets d and d'. This perspective aligns naturally with the f-DP framework and reveals structural limitations of DP-SGD under the standard worst-case adversarial model. Throughout this section and the next we work entirely under random shuffling. The Poisson subsampling case will be derived later, in Section 6.3.

5.1 Projection onto a One-Dimensional Statistic

The adversary observes all transmitted model parameters $\{w_j\}$. Under zero-out adjacency, the two datasets differ in the N-th record: dataset d contains a ghost record $\perp$ (zero gradient), while dataset

d' contains the valid record ξ_N. The adversary can compute the clipped gradient of the valid record:

$$v = [g(w_j; \xi_N)]_C.$$

Under zero-out adjacency, the DP interpretation requires considering the *worst-case sensitivity* relative to the zero vector. Because the clipped gradient satisfies $\|v\|_2 \le C$, the maximum possible distance between the valid gradient and the ghost gradient is $\|v - 0\|_2 \le C$. Consequently, the noise added to the batch gradient in round j is $\mathcal{N}(0, (C\sigma)^2 I)$ under both datasets. Using terms in Equations (1) and (2), the adversary's observations in round j are:

$$Z_j = \mathcal{N}(0, (C\sigma)^2 I) \qquad \text{(dataset } d\text{, ghost record)},$$

$$Z'_j = \delta_{N \in S_j}\, v + \mathcal{N}(0, (C\sigma)^2 I) \qquad \text{(dataset } d'\text{, valid record)}.$$

Because Gaussian noise is rotationally symmetric, all information relevant to distinguishing d (pure noise) from d' (signal plus noise) lies along the direction of the signal vector v. Hence, we can project both Z_j and Z'_j onto this direction. Now, invoking the sensitivity-attainment assumption of Theorem 6.1 (a record with $\|v\|_2 = C$, the worst case over neighboring datasets) and normalizing by C, we obtain a scalar statistic. Conditioned on the round where the record is sampled ($N \in S_j$), the distributions differ by a shift of exactly 1:

$$\mathcal{N}(0, \sigma^2) \quad \text{vs.} \quad \mathcal{N}(1, \sigma^2).$$

Finally, dividing by σ for convenience produces the canonical form of the hypothesis test:

$$\mathcal{N}(0, 1) \quad \text{vs.} \quad \mathcal{N}(\sigma^{-1}, 1).$$

This projection reduces the adversary's task to a one-dimensional Gaussian hypothesis test, fully capturing the information relevant to determining whether the differentiating record was included in the batch (signal) or replaced by a ghost record (noise). It serves as the basis for the hypothesis-testing formulation presented in the next section.

5.2 Adversarial Hypothesis Formulation

Let H_0 denote the null hypothesis corresponding to dataset d and let H_1 denote the alternative corresponding to dataset d'. Over an epoch of $M = N/m$ rounds, the adversary observes:

$$(x_1, \ldots, x_M) \sim \begin{cases} (\mathcal{N}(0,1))^M, & H_0, \\ (\mathcal{N}(u_1\sigma^{-1}, 1), \ldots, \mathcal{N}(u_M\sigma^{-1}, 1)), & H_1, \end{cases}$$

where $u_j := \delta_{N \in S_j}$ indicates whether the valid sample ξ_N appears in the j-th batch. Under H_0, no valid record exists, so effectively $u_j = 0$ for all j. Under H_1, since M divides N, the shuffling procedure discards no records (an assumption carried into Theorem 6.1), and the valid record appears in exactly one batch S_j, so exactly one $u_j = 1$, with the index j uniformly distributed over $[M]$.

Under the worst-case adversary, the entire sequence $(x_1, \ldots, x_M)$ is observable, so the adversary can compute the full likelihood ratio between H_0 and H_1. The corresponding probability densities are

$$P[(x_j)_{j=1}^M | H_0] = \prod_{j=1}^{M} \frac{e^{-x_j^2/2}}{\sqrt{2\pi}}, \tag{3}$$

$$\begin{aligned} P[(x_j)_{j=1}^M | H_1] &= \frac{1}{M} \sum_{j=1}^{M} \frac{e^{-(x_j - 1/\sigma)^2/2}}{\sqrt{2\pi}} \prod_{\substack{i=1 \\ i \ne j}}^{M} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}} \\ &= \frac{1}{M} \sum_{j=1}^{M} e^{x_j/\sigma - 1/(2\sigma^2)} \prod_{i=1}^{M} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}}. \end{aligned} \tag{4}$$

5.3 Application of the Neyman–Pearson Lemma

We now recall the classical Neyman–Pearson framework, which provides the optimal hypothesis test distinguishing between datasets d and d' in terms of Type I and Type II errors.

Let ϕ denote a (possibly randomized) rejection rule for the null hypothesis. Given the observed sequence $(x_j)_{j=1}^M$, the rule ϕ outputs the probability in $[0, 1]$ with which H_0 is rejected, the values 1 and 0 being the deterministic case.

The *false positive rate* (Type I error) of a test ϕ is defined as

$$\alpha(\phi) = \mathbf{E}[\phi \mid d] \in [0, 1],$$

and represents the probability of incorrectly rejecting the null hypothesis when d is true. A test of size $\alpha(\phi)$ is said to have level a if $\alpha(\phi) \le a$. The set of all level-a tests is therefore

$$T_a = \{\phi' \ : \ \mathbf{E}[\phi' \mid d] = \alpha(\phi') \le a\}.$$

The *false negative rate* (Type II error) is defined as

$$\beta(\phi) = 1 - \mathbf{E}[\phi \mid d'] \in [0, 1],$$

which represents the probability of incorrectly accepting the null hypothesis when d' is true.

Among all tests with size at most a, the *uniformly most powerful (UMP)* test is the one that minimizes the Type II error. Formally, ϕ is UMP among level-a tests if, for every $\phi' \in T_a$,

$$\beta(\phi) \le \beta(\phi').$$

The optimal trade-off between α and β can thus be summarized by the *trade-off function*

$$f(a) = \inf_{\phi'} \{\beta(\phi') : \alpha(\phi') \le a\}.$$

This function $f(a)$ corresponds directly to the definition of f-DP [19].

According to the Neyman–Pearson lemma, any UMP test of level a must be a *likelihood-ratio test.* That is, there exists a threshold $h \ge 0$, where $\mathbf{P}[\cdot]$ denotes probability:

$$\begin{aligned} \phi((x_j)_{j=1}^M) = 1 \quad &\Rightarrow \quad \mathbf{P}[(x_j)_{j=1}^M \mid d'] > h \cdot \mathbf{P}[(x_j)_{j=1}^M \mid d], \\ \phi((x_j)_{j=1}^M) = 0 \quad &\Rightarrow \quad \mathbf{P}[(x_j)_{j=1}^M \mid d'] < h \cdot \mathbf{P}[(x_j)_{j=1}^M \mid d], \end{aligned} \tag{5}$$

where the boundary set $\{(x_j)_{j=1}^M : \mathbf{P}[(x_j)_{j=1}^M \mid d'] = h \cdot \mathbf{P}[(x_j)_{j=1}^M \mid d]\}$ has probability zero under both d and d'. The lemma guarantees that if a test ϕ satisfies $\alpha(\phi) = a$, then it is UMP among level-a tests; moreover, all such UMP tests share the same likelihood-ratio form in (5) with a common threshold h.

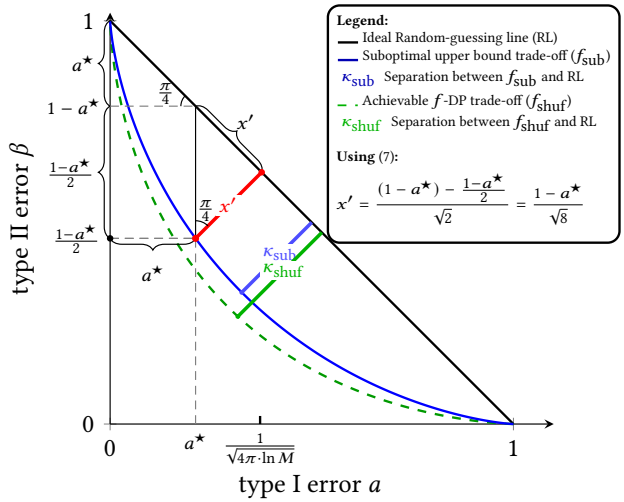


Figure 2: Illustrative geometry of the suboptimal and true trade-off functions in our impossibility argument. The figure compares the *random-guessing line* (black), the *suboptimal upper bound* trade-off f_{sub} (blue), and the *true* f*-DP trade-off* f_{shuf} (green, dashed). The red segment x' is the pointwise separation of the suboptimal curve at the analytically tractable point $a^\star$, which forms an *isosceles right triangle* with base angle $45°$; the blue segment κ_{sub} is the separation of the entire suboptimal curve; and the green segment κ_{shuf} is the separation of the true f-DP trade-off under random shuffling. The separations satisfy $x' \le \kappa_{\text{sub}} \le \kappa_{\text{shuf}}$.

Substituting (3) and (4) derived earlier into (5) yields the concrete rejection rule

$$\begin{aligned} \phi((x_j)_{j=1}^M) = 1 \quad &\Rightarrow \quad \tfrac{1}{M}\textstyle\sum_{j=1}^M e^{x_j/\sigma - 1/(2\sigma^2)} > h, \\ \phi((x_j)_{j=1}^M) = 0 \quad &\Rightarrow \quad \tfrac{1}{M}\textstyle\sum_{j=1}^M e^{x_j/\sigma - 1/(2\sigma^2)} < h. \end{aligned} \tag{6}$$

In the subsequent sections, we will not use this exact form of ϕ, instead we adopt a suboptimal test to derive our bound.

5.4 Geometric Setup and Notation

We conclude this section by introducing the geometric framework used in our limitation argument. Our analysis is based on the notion of *separation* between an f-DP trade-off curve and the random-guessing line, and on the observation that separation can be lower bounded using any explicit pointwise upper bound on the true trade-off curve.

Definition 5.1 (Pointwise separation). For any trade-off curve $g(\alpha)$, its pointwise separation from the ideal random-guessing line $\beta = 1 - \alpha$ is

$$\operatorname{sep}_g(\alpha) := \frac{(1-\alpha) - g(\alpha)}{\sqrt{2}}. \tag{7}$$

Geometric interpretation. For any $\alpha \in [0,1]$, the quantity $\operatorname{sep}_g(\alpha)$ in (7) is exactly the Euclidean distance from the point $(\alpha, g(\alpha))$ to the random-guessing line $\mathcal{L} := \{(\gamma, 1-\gamma) : \gamma \in [0,1]\}$. In Fig. 2, treating the blue curve as the graph of g, the perpendicular dropped from $(\alpha, g(\alpha))$ to $\mathcal{L}$ forms an isosceles right triangle with legs of length x', so that $((1-\alpha) - g(\alpha))^2 = 2(x')^2$, yielding (7).

Definition 5.2 (Global Separation). The separation of a trade-off curve g, denoted $\operatorname{sep}(g)$, is the maximum pointwise separation:

$$\operatorname{sep}(g) = \max_{\alpha \in [0,1]} \operatorname{sep}_g(\alpha). \tag{8}$$

Equivalently, using the geometric interpretation above, this corresponds to the maximum Euclidean distance from the curve to the random-guessing line:

$$\operatorname{sep}(g) = \max_{\alpha \in [0,1]} \min_{\gamma \in [0,1]} \big\| (\alpha, g(\alpha)) - (\gamma, 1-\gamma) \big\|_2. \tag{9}$$

where γ parameterizes the projection of $(\alpha, g(\alpha))$ onto the random-guessing line $\mathcal{L}$. Equations (8) and (9) therefore define the same quantity.

We now characterize where the maximum in (8)–(9) is attained for the class of symmetric and convex trade-off curves.

Lemma 5.3 (Separation). *Let $f\colon [0,1] \to [0,1]$ be a symmetric and convex trade-off function in the f-DP framework. Then the maximum in (9) is attained at the unique fixed point $\hat{a} \in (0, 1/2]$ satisfying $f(\hat{a}) = \hat{a}$, and the separation equals*

$$\operatorname{sep}(f) = \frac{1 - 2\hat{a}}{\sqrt{2}} =: \kappa. \tag{10}$$

Moreover, for any symmetric and convex f that is not identically equal to the random-guessing line, the separation necessarily satisfies

$$0 < \kappa < \frac{1}{\sqrt{2}}. \tag{11}$$

Proof. Because the random-guessing line is the diagonal $\{(\gamma, 1-\gamma)\colon \gamma \in [0,1]\}$, the inner minimization in (9) amounts to orthogonally projecting $(\alpha, f(\alpha))$ onto this line. The projection lands at $\gamma = \big(\alpha - f(\alpha) + 1\big)/2 \in [0,1]$, so the minimum is the perpendicular distance $\big((1-\alpha) - f(\alpha)\big)/\sqrt{2}$, giving the simplified form in (8).

$$\operatorname{sep}(f) = \max_{\alpha \in [0,1]} \frac{(1-\alpha) - f(\alpha)}{\sqrt{2}}. \tag{12}$$

By convexity of optimal trade-off curves [11, 19], the map $\alpha \mapsto (1-\alpha) - f(\alpha)$ is concave, so its maximum is attained where $-1 \in \partial f(\alpha)$, and by symmetry of f about the diagonal that point is the unique fixed point $\hat{a}$ satisfying $f(\hat{a}) = \hat{a}$. The separation is therefore the distance between $(\hat{a}, \hat{a})$ and the midpoint $(\frac{1}{2}, \frac{1}{2})$ of the random-guessing line, and substituting into (12) yields (10).

The absolute bound (11) follows directly from (10) and the fact that any valid symmetric convex trade-off curve with $f(0) > 0$ satisfies $0 < \hat{a} < 1/2$ □

Remark. Separation inherits structural properties from f-DP such as post-processing because it is derived from the trade-off curve. However, $\operatorname{sep}(f)$ is *not* a DP metric: it lacks a composition theorem and cannot track privacy under adaptive composition. It is an interpretable geometric property of an f-DP curve. It does, however, admit a direct operational reading. If $\operatorname{sep}(f) \ge \kappa$, then at the fixed point of the trade-off curve the worst-case adversary attains $\alpha + \beta = 1 - \sqrt{2}\operatorname{sep}(f) \le 1 - \sqrt{2}\kappa$, i.e., total error better

than random guessing ($\alpha + \beta = 1$) by at least $\sqrt{2}\,\kappa$. Conversely, if $\mathrm{sep}(f) \le \kappa$, then every test satisfies $\alpha + \beta \ge 1 - \sqrt{2}\,\kappa$, uniformly over the curve. Separation is not intended to replace (ε, δ). It is the geometric intermediate that our proof produces, it is uniform over all f-DP families, and it translates back to (ε, δ) in closed form (Corollary 6.2).

Geometric preview. Figure 2 provides a preview of the geometric structure underlying our limitation argument. The black line corresponds to random guessing, the green curve depicts the true f-DP trade-off f_{shuf} of shuffled DP-SGD, and the blue curve f_{sub} represents a particular explicit trade-off curve that upper-bounds f_{shuf} pointwise. **At this stage, the figure should be interpreted illustratively**: the precise definition and derivation of f_{sub} are deferred to Section 6.1.

The figure highlights three separation quantities. The quantity x' denotes the pointwise separation at a specific point on the upper-bounding curve, κ_{sub} denotes the global separation of that upper bound, and κ_{shuf} denotes the separation of the true trade-off curve. Independently of how the upper bound is constructed, these quantities necessarily satisfy

$$x' \;\le\; \kappa_{\mathrm{sub}} \;\le\; \kappa_{\mathrm{shuf}},$$

whenever one trade-off curve upper bounds another pointwise.

In Section 6.1, we formally construct the suboptimal trade-off curve f_{sub}, verify that it upper bounds f_{shuf}, and make all quantities shown in Figure 2 fully explicit.

6 Limitations of Favorable Privacy–Utility Trade-offs

Theorem 6.1 (Lower Bound on Separation for Shuffled DP-SGD). *Consider one-epoch random shuffling. Let $M \ge 2$ denote the number of rounds, and let $\sigma > 0$ be the noise multiplier. Assume the worst-case sensitivity C is attained and that M divides N, so that shuffling discards no records. Let $\kappa_{\mathrm{shuf}} := \mathrm{sep}(f_{\mathrm{shuf}})$ denote the separation of the true f-DP trade-off curve under shuffling defined in* (22).

Then we have the following statement:

$$\boxed{\kappa_{\mathrm{shuf}} \;\ge\; \frac{1}{\sqrt{8}}\left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right) \quad \textbf{\textit{OR}} \quad \sigma \;\ge\; \frac{1}{\sqrt{2\ln M}}}$$

Thus, unless the assumptions underlying this worst-case analysis are relaxed, improving privacy (i.e., reducing the separation κ) inevitably forces the noise level σ into a regime that degrades utility (Corollary 6.2 below, with the full derivation in Appendix E). This trade-off is reflected in our experimental results and is consistent with empirical observations in private deep learning. In Section 6.3, we show that the same phenomenon persists under Poisson subsampling (up to constant factors). The proof of Theorem 6.1 is given in Section 6.2. The $\sqrt{2\ln M}$ scale reflects Gaussian extreme-value behavior in the observation model itself, since the expected maximum of M i.i.d. standard normals is at most $\sqrt{2\ln M}$ [10, Thm. 2.5], as Appendix G explains.

Corollary 6.2 ((ε, δ)-DP implication). *Under the conditions of Theorem 6.1, suppose $\sigma < 1/\sqrt{2\ln M}$ and let*

$$\kappa_M := \frac{1}{\sqrt{8}}\left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right) \;\le\; \kappa_{\mathrm{shuf}}.$$

If the same mechanism is reported as (ε, δ)-DP, then

$$\varepsilon \;\ge\; \ln\left(\frac{1 + \kappa_M\sqrt{2} - 2\delta}{1 - \kappa_M\sqrt{2}}\right).$$

In particular, under the one-epoch baseline $\delta = 1/N$ with $N = 10^8$, this already forces $\varepsilon \gtrsim 1$ across all practically relevant M (e.g., $\varepsilon \ge 0.96$ at $M = 10^3$, rising to $\varepsilon \ge 1.00$ at $M = 5{\times}10^6$, see the table in Appendix E). The proof follows from Lemmas E.2 and E.3 in Appendix E, together with monotonicity of the map $\kappa \mapsto \varepsilon_{\min}(\kappa; N)$ defined there, which lets us substitute the lower bound κ_M for κ_{shuf}.

6.1 Suboptimal Hypothesis Test

Instead of using the optimal Neyman–Pearson likelihood ratio test, we consider a *suboptimal* rejection rule that thresholds the *maximum* coordinate of the observation vector $(x_1, \ldots, x_M)$:

$$\begin{aligned} \phi((x_j)_{j=1}^M) = 1 \quad &\Rightarrow \quad \max_{j=1}^M e^{x_j/\sigma - 1/(2\sigma^2)} > h, \\ \phi((x_j)_{j=1}^M) = 0 \quad &\Rightarrow \quad \max_{j=1}^M e^{x_j/\sigma - 1/(2\sigma^2)} < h. \end{aligned} \tag{13}$$

This test is equivalent to thresholding the maximum observed statistic:

$$\begin{aligned} \phi((x_j)_{j=1}^M) = 1 \quad &\Rightarrow \quad \max_{j=1}^M x_j > \bar{h}, \\ \phi((x_j)_{j=1}^M) = 0 \quad &\Rightarrow \quad \max_{j=1}^M x_j < \bar{h}, \end{aligned} \tag{14}$$

where $\bar{h} = \sigma \ln h + 1/(2\sigma)$. Although suboptimal, this rule isolates the single coordinate shifted by $1/\sigma$ under H_1, enabling a closed-form f-DP trade-off.

False-negative and false-positive rates. For $\bar{h} \ge 0$, define

$$\bar{\alpha}(\bar{h}) = \Pr\left[\max_{j=1}^M x_j > \bar{h} \mid d\right], \tag{15}$$

$$\bar{\beta}(\bar{h}) = \Pr\left[\max_{j=1}^M x_j < \bar{h} \mid d'\right]. \tag{16}$$

Under d (no shift), all $x_j \sim \mathcal{N}(0,1)$ independently, giving

$$\bar{\alpha}(\bar{h}) = 1 - \Phi(\bar{h})^M, \qquad \bar{\alpha}^{-1}(a) = \Phi^{-1}\big((1-a)^{1/M}\big),$$

where Φ is the CDF of $\mathcal{N}(0,1)$.

Under d', the coordinates are *not* marginally independent. The joint law is the uniform mixture over the M possible positions $j^\star$ of the shifted coordinate (cf. (4)). Conditional on any fixed position $j^\star = j$, however, the coordinates are independent, with $x_j \sim \mathcal{N}(1/\sigma, 1)$ and the remaining $M-1$ coordinates i.i.d. $\mathcal{N}(0,1)$. Hence

$$\bar{\beta}(\bar{h}) = \frac{1}{M}\sum_{j=1}^M \Pr\Big[\max_i x_i < \bar{h} \;\Big|\; j^\star = j\Big] = \Phi(\bar{h} - 1/\sigma)\,\Phi(\bar{h})^{M-1},$$

since, by conditional independence, the conditional probability equals $\Phi(\bar{h} - 1/\sigma)\,\Phi(\bar{h})^{M-1}$ for every position j of the shifted coordinate. Thus, the induced (suboptimal) trade-off function is equal to $\bar{\beta}(\bar{\alpha}^{-1}(a))$:

$$f_{\mathrm{sub}}(a) = \Phi\Big(\Phi^{-1}\big((1-a)^{1/M}\big) - \sigma^{-1}\Big) \cdot (1-a)^{(M-1)/M}. \tag{17}$$

Throughout this section, we work in the regime $\sigma \le 1/\sqrt{2\ln M}$, and write $\sigma = s/\sqrt{\ln M}$ for some $s \le \sqrt{1/2}$.

Lemma 6.3 (Bound under $\sigma \le 1/\sqrt{2\ln M}$). *Let $f_{\text{sub}}(a)$ denote the suboptimal trade-off in* (17)*. For any integer $M \ge 2$ and any noise multiplier $\sigma > 0$ satisfying $\sigma \le 1/\sqrt{2\ln M}$ there exists some $a^\star \le 1/\sqrt{4\pi \ln M}$ such that*

$$f_{\text{sub}}(a^\star) = \tfrac{1}{2}\,(1-a^\star)^{(M-1)/M}.$$

Proof. Under the standing assumption $\sigma = s/\sqrt{\ln M}$ with $s \le \sqrt{1/2}$, we choose $a^\star$ such that the inner argument of Φ in (17) equals zero, i.e.,

$$\Phi^{-1}\big((1-a^\star)^{1/M}\big) = \sigma^{-1}, \tag{18}$$

Substituting into (17) gives

$$f_{\text{sub}}(a^\star) = \Phi(0)\cdot(1-a^\star)^{(M-1)/M} = \frac{1}{2}(1-a^\star)^{(M-1)/M},$$

where we used the symmetry of the standard normal distribution, for which $\Phi(0) = \frac{1}{2}$.

Applying Φ to both sides of (18), and using that Φ is invertible, yields $(1-a^\star)^{1/M} = \Phi(1/\sigma)$, and hence

$$1 - a^\star = \Phi(1/\sigma)^M = (1-\Phi(-1/\sigma))^M. \tag{19}$$

Next, we use the standard Gaussian tail bound

$$\Phi(-t) < \frac{e^{-t^2/2}}{t\sqrt{2\pi}}, \quad t > 0,$$

and substituting $t = 1/\sigma$, we obtain

$$\Phi(-1/\sigma) \;\le\; \frac{\sigma e^{-1/(2\sigma^2)}}{\sqrt{2\pi}}.$$

Exponentiating both sides with M and using (19) yields

$$1 - a^\star \ge \left(1 - \frac{\sigma e^{-1/(2\sigma^2)}}{\sqrt{2\pi}}\right)^M.$$

Substituting $\sigma = s/\sqrt{\ln M}$ yields

$$1 - a^\star \ge \left(1 - \frac{s\,M^{-1/(2s^2)}}{\sqrt{2\pi\ln M}}\right)^M \ge 1 - \frac{s\,M^{1-1/(2s^2)}}{\sqrt{2\pi\ln M}}.$$

For $s \le \sqrt{1/2}$, this implies

$$a^\star \le \frac{1}{\sqrt{4\pi\ln M}}.$$

This establishes the desired bound. □

6.2 Separation Lower Bounds for Shuffled DP-SGD

We now instantiate the geometric framework introduced earlier using the analytically tractable suboptimal trade-off curve derived above (cf. (17)). This yields an explicit geometric lower bound on separation. In Figure 2, we consider:

- the **black line** $\beta = 1-\alpha$ (ideal random guessing);
- the **green curve** $f_{\text{shuf}}(\alpha)$, representing the true trade-off of shuffled DP-SGD under a worst-case adversary;
- the **blue curve** $f_{\text{sub}}(\alpha)$, an explicit suboptimal upper bound derived in (17).

Notation: three separations. Using Definition 5.1 and (8), we introduce the three quantities used in the analysis and in the Figure 2:

$$x' := \text{sep}_{f_{\text{sub}}}(a^\star), \tag{20}$$

$$\kappa_{\text{sub}} := \text{sep}(f_{\text{sub}}) = \max_{\alpha\in[0,1]} \text{sep}_{f_{\text{sub}}}(\alpha), \tag{21}$$

$$\kappa_{\text{shuf}} := \text{sep}(f_{\text{shuf}}) = \max_{\alpha\in[0,1]} \text{sep}_{f_{\text{shuf}}}(\alpha). \tag{22}$$

Here, x' is the pointwise separation at the analytically tractable point $(a^\star, f_{\text{sub}}(a^\star))$, κ_{sub} is the global separation of the suboptimal curve f_{sub}, and κ_{shuf} is the true separation of the shuffled trade-off curve f_{shuf}. Then, since $f_{\text{sub}}(\alpha) \ge f_{\text{shuf}}(\alpha)$ for all α, we have

$$\text{sep}_{f_{\text{sub}}}(\alpha) \;\le\; \text{sep}_{f_{\text{shuf}}}(\alpha) \qquad \text{for all } \alpha,$$

and therefore

$$x' \;\le\; \kappa_{\text{sub}} \;\le\; \kappa_{\text{shuf}}. \tag{23}$$

Thus any lower bound obtained for x' immediately implies the same lower bound for κ_{shuf}, the true separation of the shuffled f-DP trade-off.

We now prove Theorem 6.1 by combining the ingredients developed in the previous sections. The proof proceeds by constructing an explicit, analytically tractable upper bound on the f-DP trade-off curve under one-epoch shuffling using the suboptimal hypothesis test of Section 6.1, and then translating this bound into a geometric lower bound on the separation via the framework of Section 5.4.

Proof of Theorem 6.1. Assume $\sigma \le 1/\sqrt{2\ln M}$, and let f_{sub} denote the induced (suboptimal) trade-off curve defined in (17).

We now convert Lemma 6.3 into a geometric lower bound that reveals an inherent limitation of the privacy–utility trade-off. In Figure 2, the suboptimal point $(a^\star, f_{\text{sub}}(a^\star))$ lies on the blue curve representing the analytically tractable upper bound:

$$(a^\star, f_{\text{sub}}(a^\star)) = \Big(a^\star, \tfrac{1}{2}(1-a^\star)^{(M-1)/M}\Big).$$

Let x' be the perpendicular distance from this point to the random-guessing line $\beta = 1-\alpha$, so that x' is given by (20).

Rewrite

$$f_{\text{sub}}(a^\star) = \tfrac{1}{2}(1-a^\star)^{(M-1)/M} = \frac{1-a^\star}{2}\,(1-a^\star)^{-1/M}.$$

By Lemma 6.3, $a^\star \le 1/\sqrt{4\pi\ln M}$, hence

$$(1-a^\star)^{-1/M} \le \left(1 - \tfrac{1}{\sqrt{4\pi\ln M}}\right)^{-1/M}. \tag{24}$$

Define

$$\varepsilon_M := \frac{2}{M\sqrt{4\pi\ln M}}. \tag{25}$$

Since $1/\sqrt{4\pi\ln M} < 1/2$ for all $M \ge 2$, the elementary inequality $(1-t)^{-1/M} \le 1 + \frac{2t}{M}$ for $t \in (0, 1/2)$ implies (applying it to the right-hand side of (24) with $t = 1/\sqrt{4\pi\ln M}$) that

$$(1-a^\star)^{-1/M} \le 1 + \varepsilon_M. \tag{26}$$

Consequently,

$$f_{\text{sub}}(a^\star) \le \frac{1-a^\star}{2}\,(1+\varepsilon_M). \tag{27}$$

By definition, the perpendicular distance to the line $\beta = 1 - \alpha$ equals

$$x' = \frac{(1 - a^\star) - f_{\text{sub}}(a^\star)}{\sqrt{2}}.$$

Substituting (27) yields

$$x' \geq \frac{1 - a^\star}{\sqrt{8}} (1 - \varepsilon_M). \tag{28}$$

Using again $a^\star \leq 1/\sqrt{4\pi \ln M}$ gives

$$x' \geq \frac{1}{\sqrt{8}} \left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right) (1 - \varepsilon_M). \tag{29}$$

Finally, since f_{sub} arises from a specific (suboptimal) test while f_{shuf} is the infimum over all tests, we have $f_{\text{sub}}(\alpha) \geq f_{\text{shuf}}(\alpha)$ for all α. The monotonicity of pointwise separation under this ordering, established in (23), gives $x' \leq \kappa_{\text{sub}} \leq \kappa_{\text{shuf}}$. Therefore,

$$\kappa_{\text{shuf}} \geq x' \geq \frac{1}{\sqrt{8}} \left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right) (1 - \varepsilon_M) \tag{30}$$

which proves the claimed lower bound. Note that, although the bound above is explicit, the correction factor $(1 - \varepsilon_M)$ converges to 1 at rate $O(1/M)$; we therefore omit it in the theorem statement and retain it only in the proof for full rigor. □

Interpretation. The inequality (30) provides a concrete *lower bound* on the geometric separation κ_{shuf} between the true f-DP trade-off curve and the random-guessing line. Equivalently, it shows that any mechanism whose noise level satisfies $\sigma < 1/\sqrt{2 \ln M}$ must retain at least this amount of separation and therefore cannot make its trade-off curve arbitrarily close to the random-guessing regime. Moreover, although the bound on σ decreases as M grows, the decay is extremely slow. Even for very large M, the required noise level remains far from zero. For example, assuming a batch size of 256, the standard ImageNet-1k dataset (∼1.3M images, $M \approx 5 \times 10^3$) requires $\sigma \gtrsim 0.24$. Scaling to foundation models on LAION-5B [42] (∼5.8B image-text pairs, $M \approx 2.3 \times 10^7$), the limit only drops to $\sigma \approx 0.17$. Thus, the condition on σ remains substantial across practically relevant values of M, revealing a true bottleneck in this setting and simply scaling the dataset is insufficient to bypass this fundamental noise bottleneck.

6.3 Extension to Poisson Subsampling

While our derivation focuses on the shuffled regime which reflects practical implementations, we now show that the same phenomenon extends to Poisson subsampling. A simple mixture argument reveals that Poisson subsampling can be upper bounded by a mechanism that, with some probability, leaks no information at all, and otherwise behaves like the shuffled mechanism analyzed above. This immediately transfers the shuffled lower bound on separation to the Poisson setting up to a constant factor. We first state the resulting bound for Poisson-sampled DP-SGD, deferring the technical argument to a supporting lemma below.

Theorem 6.4 (Lower Bound on Separation for Poisson-sampled DP-SGD). *Consider DP-SGD trained for one epoch under* Poisson subsampling *with $M \geq 2$ rounds, sampling rate $q = 1/M$ (i.e., expected batch size $b = N/M$, so that one epoch is processed in expectation), and noise multiplier $\sigma > 0$, under the assumptions of Theorem 6.1. Let f_{pois} denote the true f-DP trade-off curve under Poisson subsampling, and let $\text{sep}(f_{\text{pois}})$ denote its separation. Then the following dichotomy holds:*

$$\boxed{\kappa_{\text{pois}} \;\geq\; \left(1 - \tfrac{1}{e}\right) \frac{1}{\sqrt{8}} \left(1 - \frac{1}{\sqrt{4\pi \ln M}}\right) \quad \textbf{\textit{OR}} \quad \sigma \;\geq\; \frac{1}{\sqrt{2 \ln M}}}$$

Lemma 6.5 (Transferring shuffling bounds to Poisson subsampling under zero-out adjacency). *Consider DP-SGD trained for one epoch under* Poisson subsampling *with $M \geq 1$ rounds and per-round sampling probability $q \in (0, 1)$, and assume* zero-out adjacency *(cf. Definition 3.8). Let f_{pois} denote the true f-DP trade-off curve under Poisson subsampling, and let f_{shuf} denote the true f-DP trade-off curve under one-epoch random shuffling into M batches of size $b = N/M$, where M divides N as in Theorem 6.1, so that the shuffling procedure discards no records. Assume that*

$$f_{\text{shuf}}(\alpha) \leq f'_{\text{shuf}}(\alpha) \qquad \text{for all } \alpha \in [0, 1].$$

Let $p = (1 - q)^M$ be the probability that the differing record is never sampled during the epoch. Then, for all $\alpha \in [0, 1]$,

$$f_{\text{pois}}(\alpha) \;\leq\; p(1 - \alpha) + (1 - p)\, f'_{\text{shuf}}(\alpha). \tag{31}$$

In particular, if the expected batch size is $b = qN$ and number of rounds is $M = N/b$ and $q = b/N = 1/M$, then for typical batch sizes $b \ll N$ we have

$$p = (1 - q)^M = \left(1 - \tfrac{1}{M}\right)^M \approx e^{-1}.$$

Proof of Lemma 6.5 can be found in Appendix D. We now combine this mixture bound with the separation lower bound established for shuffled DP-SGD to complete the proof of Theorem 6.4.

Proof of Theorem 6.4. Under Poisson subsampling, each sample is independently included in each round with probability $q = 1/M$ over M rounds (one epoch), so that $p = (1 - q)^M$. Let f_{shuf} denote the true f-DP trade-off curve under one-epoch random shuffling. By Lemma 6.5 (applied with $f'_{\text{shuf}} := f_{\text{shuf}}$), for all $\alpha \in [0, 1]$,

$$f_{\text{pois}}(\alpha) \;\leq\; p(1 - \alpha) + (1 - p)\, f_{\text{shuf}}(\alpha).$$

Geometrically, this shows that f_{pois} lies below a convex combination of the random-guessing line (which has zero separation) and the shuffled trade-off curve. Hence,

$$\kappa_{\text{pois}} \;\geq\; (1 - p)\, \kappa_{\text{shuf}}.$$

Finally, using $p \approx e^{-1}$ from Lemma 6.5 and using Theorem 6.1 proves the claim. □

Shuffled and Poisson DP-SGD obey the *same* worst-case limitation up to a constant factor. In particular, Poisson-sampled DP-SGD cannot make its trade-off curve arbitrarily close to random guessing under the standard worst-case DP adversary and the one-epoch regime, unless σ exceeds the same threshold ($\sigma \geq 1/\sqrt{2 \ln M}$) as in the shuffled case. Together, Theorems 6.1 and 6.4 establish a unified worst-case limitation. Under the standard worst-case DP adversary, neither shuffling nor Poisson subsampling can simultaneously achieve a noise multiplier below the σ threshold and a separation below the corresponding lower bounds.

7 Empirical Analysis

Sampling, normalization, and microbatching. We consider both Poisson subsampling and random shuffling. For Poisson subsampling, each example is included independently with probability q at each step, giving a random number of participating examples. To have the DP-SGD update consistent with common practice, we normalize gradients and calibrate Gaussian noise using the *expected* batch size $b = qN$, treating non-sampled examples as contributing zero gradient. This yields one DP-SGD update per Poisson-sampled step with a fixed noise scale, while preserving the independence structure assumed in our theoretical analysis. For random shuffling, we use fixed-size batches throughout by shuffling the dataset at each epoch and discarding any incomplete remainder batch which coincides with the standard shuffled DP-SGD mechanism (Definition 3.2). All experiments are implemented using the `JAX Privacy` DP-SGD framework [7]. We use microbatching with fixed size 32 only to reduce memory usage. Gradients are clipped per example and Gaussian noise is added once per logical batch, so microbatching does not affect the optimization or privacy mechanism. We emphasize that all experiments run the standard DP-SGD mechanism of Algorithm $\mathcal{B}$ without modification. Zero-out adjacency enters only the privacy *analysis* and does not alter the training algorithm (cf. Section 3.3). Under substitution adjacency the same conclusions hold with the noise scale doubled, so the utility figures we report are, if anything, optimistic for the substitution-based reading.

Datasets and Model Architectures. We evaluate our bounds on image (CIFAR-10/100 [32], SVHN [38]) and text (AG News [50]) benchmarks. We employ standard architectures including ResNet [29], Vision Transformers (ViT) [20], and encoder-only Transformer architecture [47]. To analyze how model capacity interacts with privacy noise, we evaluate variants ranging from **Tiny** to **Base** by increasing the *embedding dimension* (the width of the internal vector representations) and depth. Detailed architectural specifications for all models are provided in Appendix B.

Experimental protocol and hyperparameter selection. For each epoch budget $E \in \{1, 10, 25\}$, we first fix a model architecture that is compatible with DP training (e.g., replacing non-DP-friendly components such as LayerNorm where necessary). Using this fixed model, we perform a hyperparameter search under *clean training, i.e.,* $\sigma = 0$ to identify the best-performing optimizer configuration for each epoch budget. These hyperparameters are then held fixed across all subsequent experiments. The clipping constant C is selected by maximizing utility under a fixed noise multiplier set to our theoretical lower bound (i.e., $\sigma = 1/\sqrt{2 \ln M}$). We then evaluate two training regimes: (i) $\sigma = 0$ without clipping (clean training) and (ii) full DP-SGD with clipping and Gaussian noise with σ set to our theoretical lower bound. The purpose of these experiments is not to achieve peak accuracy, but to demonstrate that the noise levels implied by our lower bound are already severe in practice. Additional implementation and tuning details are provided in Appendix B.

Multi-view interpretation. Although our bounds are derived for a single epoch, this setting models federated learning scenarios where each client contributes one local epoch and the server observes many such client updates across rounds. Accordingly, we repeat the one-epoch mechanism over E independent views, corresponding to observing E independent client updates. Our aim is not to optimize final accuracy, but to empirically illustrate the severity of the lower bound as the number of observed views increases.

Results and interpretation. Our experiments report test accuracy under shuffling and Poisson subsampling across batch sizes and epoch budgets. Introducing Gaussian noise at our lower bound induces a substantial and persistent utility gap relative to clean training. On CIFAR-10 with ResNet-18 at batch size 128, the noise floor $\sigma = 0.29$ reduces test accuracy from 80.1% to 40.4% at 10 epochs and from 82.2% to 43.4% at 25 (Table 2). On AG News with a 136M-parameter Transformer at the same batch size, accuracy drops from 91.4% to 82.6% at 25 epochs (Table 3). The gap is already pronounced at small epoch budgets and does not close with more epochs, so the noise floor imposes a ceiling on optimization that more training cannot lift. Shuffling and Poisson subsampling suffer comparable losses, consistent with Theorems 6.1 and 6.4. Such degradation is well documented [4, 15, 41, 46]. Our contribution is the explicit noise floor that provably induces it. Extended experimental results on additional model scales (Tiny, Small, and Base Transformers, ViT-Tiny/Small/Base, ResNet-34, and WideResNet-28x10) and datasets (AG News, CIFAR-10, CIFAR-100, SVHN) can be found in Appendix C, which also gives the complete batch-size sweeps for every table reported in this section.

Privacy accounting at the noise floor. Our theoretical results certify the limitation in f-DP terms. We also verify the (ε, δ) implication with the independent shuffled-DP-SGD accountants of Chua et al. [14]. Their shuffle accountant certifies a *lower* bound on ε for the released update transcript and their deterministic-batching accountant a valid *upper* bound, and at our threshold the two agree to within 0.01 for batch size 128. For CIFAR-10 ($N = 50{,}000$, $\delta = 1/N = 2 \times 10^{-5}$) at batch size 128 ($M = 390$) with noise at exactly our threshold $\sigma = 1/\sqrt{2 \ln M} \approx 0.29$, the accountant certifies $\varepsilon \geq 19.5$ after one epoch, $\varepsilon \geq 103.7$ after 10, and $\varepsilon \geq 219.2$ after 25. Table 1 reports these values across batch sizes and extends the same computation to corpora as large as ImageNet-21k ($N = 14{,}197{,}122$), where the bound tightens further, since larger N gives more rounds per epoch at a fixed batch size and a smaller $\delta = 1/N$. Every configuration at the threshold yields $\varepsilon \gg 10$, reaching $\varepsilon \geq 28.8$ after a single epoch on ImageNet-21k, far outside any regime commonly regarded as a meaningful worst-case guarantee [39]. Below our noise floor, then, the DP framework cannot certify meaningful privacy in standard (ε, δ) language, complementing the utility results above.

8 Discussion and Future Directions

Our results expose a structural tension in DP-SGD under the standard worst-case adversarial model. This section discusses the implications of this limitation, how it relates to existing privacy frameworks, and which directions appear promising for overcoming or reinterpreting these constraints.

Is the adversary too strong? The lower bounds we derive are with respect to the standard worst-case DP adversary, who is allowed arbitrary side information and observes all noisy updates. From this perspective, our results may be interpreted less as an indictment

Table 1: (ε, δ)-DP at the noise floor, across dataset scales. Tight lower bounds on ε from the shuffle accountant of Chua et al. [14], at one representative batch size per corpus, with $\delta = 1/N$, $M = \lfloor N/\mathrm{BS} \rfloor$ and $\sigma = 1/\sqrt{2 \ln M}$. N counts records.

Dataset	N	BS	M	σ	ε (1 ep)	ε (25 ep)
AG News [50]	120,000	128	937	0.27	22.1	249.8
SVHN [38]	604,388	128	4,721	0.24	26.9	306.2
ImageNet-1k	1,281,167	1,024	1,251	0.26	24.6	268.1
Places365 [51]	1,803,460	1,024	1,761	0.26	25.7	280.1
ImageNet-21k [40]	14,197,122	4,096	3,466	0.25	28.8	309.1

Table 2: CIFAR-10 / ResNet-18. Rows correspond to batch size (BS) and epoch; columns compare Random Shuffling vs. Poisson Subsampling under σ=0 and DP-SGD at the shown σ.

BS	Epoch	shuffling		Poisson	
		$\sigma = 0$	DP (σ)	$\sigma = 0$	DP (σ)
128	1	48.0	38.0 (0.29)	47.9	38.9 (0.29)
	10	80.1	40.4 (0.29)	79.4	40.5 (0.29)
	25	82.2	43.4 (0.29)	81.8	45.7 (0.29)
512	1	39.9	35.3 (0.33)	42.6	36.2 (0.33)
	10	70.5	59.8 (0.33)	71.9	60.3 (0.33)
	25	79.5	61.9 (0.33)	79.0	62.7 (0.33)
4096	1	15.6	14.4 (0.45)	15.1	14.5 (0.45)
	10	45.9	38.3 (0.45)	46.9	40.9 (0.45)
	25	54.5	48.3 (0.45)	57.6	50.0 (0.45)

Table 3: AG News / Transformer-Base (≈136M parameters). Rows correspond to batch size (BS) and epoch; columns compare Random Shuffling vs. Poisson Subsampling under σ=0 and DP-SGD at the shown σ.

BS	Epoch	shuffling		Poisson	
		$\sigma = 0$	DP (σ)	$\sigma = 0$	DP (σ)
128	1	86.8	74.7 (0.27)	87.6	74.1 (0.27)
	10	90.7	53.9 (0.27)	89.9	56.8 (0.27)
	25	91.4	82.6 (0.27)	91.1	82.7 (0.27)
512	1	84.8	76.6 (0.30)	82.9	75.2 (0.30)
	10	89.1	84.1 (0.30)	89.6	83.6 (0.30)
	25	90.2	82.7 (0.30)	90.3	82.3 (0.30)

of DP-SGD itself and more as a limitation of the adversarial model. This raises the question of whether different privacy notions such as instance-based PAC-privacy [48] can meaningfully relax the trade-off without abandoning rigorous protection. An appealing direction is to understand how separation behaves when privacy is required only with high probability over the data distribution, rather than uniformly over all neighboring datasets. Separation provides a concrete yardstick for such comparisons. If an instance-based or distributional privacy notion yields a smaller separation κ at the same noise level σ, the difference directly quantifies the gain obtained by relaxing the worst-case adversarial model.

Algorithmic alternatives to noise scaling. A key takeaway from both our theory and experiments is that increasing the noise multiplier alone is a blunt instrument. Improving privacy–utility trade-offs may require algorithmic changes rather than purely stronger noise calibration. Promising directions include modifying how gradients are clipped and aggregated, reducing the effective dimensionality or sparsity of updates before noise injection, and reconsidering training schedules such as stopping early or running only partial epochs. An open question is whether such interventions can qualitatively weaken the adversary and yield different privacy–utility behavior rather than merely shifting constants within the same worst-case regime. Recent results substantiate parts of this agenda. Random model and data partitioning yields provable amplification [17], and structured samplers that eliminate participation variance match or exceed Poisson subsampling [12, 18, 44].

Multi-epoch behavior. Although our main lower bound is derived for a single epoch, multi-epoch training is the norm in practice. Existing asymptotic analyses within the μ-GDP framework [19] suggest that privacy loss converges to a Gaussian limit under composition. However, such results control only the asymptotic error, leave the non-asymptotic constants open, and do not directly capture the separation metric that governs worst-case distinguishability (see Appendix F). Understanding how separation evolves under repeated composition in the multi-epoch setting remains an open problem, and extending our separation bounds to this regime is an important direction for future work.

Implications for practice and the broader picture. Empirically, we observe that the noise levels implied by our lower bounds already induce substantial accuracy degradation at realistic batch sizes and model scales. Importantly, this does not imply that private learning is infeasible in practice; rather, it clarifies the concrete costs imposed by insisting on strong worst-case privacy guarantees within standard SGD-based training pipelines. More broadly, our results highlight that the combination of worst-case adversarial definitions and conventional optimization dynamics imposes real and sometimes underappreciated constraints on achievable privacy–utility trade-offs. A central open question is whether future progress will primarily come from new algorithmic designs, weaker but still meaningful privacy notions, or a principled combination of both. We view this work as a step toward understanding the boundary of what DP-SGD can and cannot provide under worst-case semantics, and toward motivating alternative frameworks.

9 Conclusion

We analyzed DP-SGD under realistic shuffling and Poisson subsampling through the f-DP hypothesis-testing framework and established a fundamental limitation under the standard worst-case adversarial model. The noise multiplier and worst-case distinguishability cannot be simultaneously driven below explicit lower bounds. This geometric limitation applies to both sampling schemes with empirical results confirming that the implied noise levels already cause substantial utility degradation at practical scales. Importantly, these findings should not be read as a negative statement about private learning per se; rather, they describe the boundary of what

DP-SGD can achieve under strong worst-case definitions, and motivate the exploration of alternative algorithms, adaptive training mechanisms, or weaker but still meaningful privacy notions when this boundary proves too restrictive.

Ethics Considerations

This is a theoretical paper regarding DP-SGD, so it introduces no attack and weakens no existing guarantee, and all experiments use public benchmarks with no human subjects or personal data.

Open Science

All code, runs, checkpoints, and the scripts are archived on Zenodo at https://doi.org/10.5281/zenodo.21768773.

Acknowledgments

The contribution of the authors is part of the project CiCS of the research program Gravitation which is (partly) financed by the Dutch Research Council (NWO) under the grant 024.006.037. We acknowledge the use of the DAS-6 HPC cluster at VU Amsterdam [5]. ChatGPT and Claude were used for language revision.

References

[1] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM, Vienna, Austria, 308–318. doi:10.1145/2976749.2978318

[2] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2867. doi:10.1145/3219819.3226070

[3] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-Scale Differentially Private BERT. arXiv:2108.01624 [cs.LG] https://arxiv.org/abs/2108.01624

[4] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). NeurIPS, Vancouver, BC, Canada, 15453–15462. https://proceedings.neurips.cc/paper/2019/hash/fc0de4e0396fff257ea362983c2dda5a-Abstract.html

[5] Henri E. Bal, Dick H. J. Epema, Cees de Laat, Rob van Nieuwpoort, John W. Romein, Frank J. Seinstra, Cees Snoek, and Harry A. G. Wijshoff. 2016. A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term. *Computer* 49, 5 (2016), 54–63. doi:10.1109/MC.2016.127

[6] Borja Balle, Gilles Barthe, and Marco Gaboardi. 2018. Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). NeurIPS, Montreal, Canada, 6280–6290. https://proceedings.neurips.cc/paper/2018/hash/3b5020bb891119b9f5130f1fea9bd773-Abstract.html

[7] Borja Balle, Leonard Berrada, Zachary Charles, Christopher A Choquette-Choo, Soham De, Vadym Doroshenko, Dj Dvijotham, Andrew Galen, Arun Ganesh, Sahra Ghalebikesabi, Jamie Hayes, Peter Kairouz, Ryan McKenna, Brendan McMahan, Aneesh Pappu, Natalia Ponomareva, Mikhail Pravilov, Keith Rush, Samuel L Smith, and Robert Stanforth. 2025. JAX-Privacy: Algorithms for Privacy-Preserving Machine Learning in JAX. http://github.com/google-deepmind/jax_privacy

[8] Jeremiah Birrell, Reza Ebrahimi, Rouzbeh Behnia, and Jason Pacheco. 2024. Differentially Private Stochastic Gradient Descent with Fixed-Size Minibatches: Tighter RDP Guarantees with or without Replacement. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). NeurIPS, Vancouver, Canada, 11087–11131. http://papers.nips.cc/paper_files/paper/2024/hash/14fef58f09f2ebe69306e0a322e3be2b-Abstract-Conference.html

[9] Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. 2021. Gradient Masking and the Underestimated Robustness Threats of Differential Privacy in Deep Learning. *CoRR* abs/2105.07985 (2021), 1–13. arXiv:2105.07985 https://arxiv.org/abs/2105.07985

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. doi:10.1093/acprof:oso/9780199535255.001.0001

[11] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J. Su. 2019. Deep Learning with Gaussian Differential Privacy. *CoRR* abs/1911.11607 (2019), 32 pages. arXiv:1911.11607 http://arxiv.org/abs/1911.11607

[12] Lynn Chua, Badih Ghazi, Charlie Harrison, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2025. Balls-and-Bins Sampling for DP-SGD. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025 (Proceedings of Machine Learning Research, Vol. 258)*, Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (Eds.). PMLR, Mai Khao, Thailand, 946–954. https://proceedings.mlr.press/v258/chua25a.html

[13] Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2024. How Private are DP-SGD Implementations?. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, Vienna, Austria, 8904–8918. https://proceedings.mlr.press/v235/chua24a.html

[14] Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2024. Scalable DP-SGD: Shuffling vs. Poisson Subsampling. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). NeurIPS, Vancouver, BC, Canada, 70026–70047. http://papers.nips.cc/paper_files/paper/2024/hash/81c252b45b3bd7d9bf080eb27794b762-Abstract-Conference.html

[15] Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. 2022. Unlocking High-Accuracy Differentially Private Image Classification through Scale. arXiv:2204.13650 [cs.LG] https://arxiv.org/abs/2204.13650

[16] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting Telemetry Data Privately. arXiv:1712.01524 [cs.CR] https://arxiv.org/abs/1712.01524

[17] Andy Dong, Wei-Ning Chen, and Ayfer Özgür. 2025. Leveraging Randomness in Model and Data Partitioning for Privacy Amplification. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025 (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR / OpenReview.net, Vancouver, BC, Canada, 25 pages. https://proceedings.mlr.press/v267/dong25a.html

[18] Andy Dong and Ayfer Özgür. 2026. Less Random, More Private: What is the Optimal Subsampling Scheme for DP-SGD? arXiv:2605.07072 doi:10.48550/ARXIV.2605.07072

[19] Jinshuo Dong, Aaron Roth, and Weijie J. Su. 2019. Gaussian Differential Privacy. *CoRR* abs/1905.02383 (2019), 86 pages. arXiv:1905.02383 http://arxiv.org/abs/1905.02383

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. ICLR, Austria, May 3-7, 2021, 21 pages. https://openreview.net/forum?id=YicbFdNTTy

[21] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 4004)*, Serge Vaudenay (Ed.). Springer, St. Petersburg, Russia, 486–503. doi:10.1007/11761679_29

[22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 3876)*, Shai Halevi and Tal Rabin (Eds.). Springer, New York, NY, USA, 265–284. doi:10.1007/11681878_14

[23] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. 2019. Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego,*

California, USA, January 6-9, 2019, Timothy M. Chan (Ed.). SIAM, San Diego, California, USA, 2468–2479. doi:10.1137/1.9781611975482.151

[24] Vitaly Feldman, Audra McMillan, and Kunal Talwar. 2021. Hiding Among the Clones: A Simple and Nearly Optimal Analysis of Privacy Amplification by Shuffling. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*. IEEE, Denver, CO, USA, 954–964. doi:10.1109/FOCS52979.2021.00096

[25] Vitaly Feldman, Audra McMillan, and Kunal Talwar. 2023. Stronger Privacy Amplification by Shuffling for Renyi and Approximate Differential Privacy. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, Nikhil Bansal and Viswanath Nagarajan (Eds.). SIAM, Florence, Italy, 4966–4981. doi:10.1137/1.9781611977554.CH181

[26] Vitaly Feldman and Moshe Shenfeld. 2026. Efficient privacy loss accounting for subsampling and random allocation. *CoRR* abs/2602.17284 (2026), 37 pages. arXiv:2602.17284 doi:10.48550/ARXIV.2602.17284

[27] Mert Gürbüzbalaban, Asuman E. Ozdaglar, and Pablo A. Parrilo. 2021. Why random reshuffling beats stochastic gradient descent. *Math. Program.* 186, 1 (2021), 49–84. doi:10.1007/S10107-019-01440-W

[28] Jeff Z. HaoChen and Suvrit Sra. 2019. Random Shuffling Beats SGD after Finite Epochs. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 2624–2633. http://proceedings.mlr.press/v97/haochen19a.html

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, Las Vegas, NV, USA, 770–778. doi:10.1109/CVPR.2016.90

[30] Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. 2021. Practical and Private (Deep) Learning Without Sampling or Shuffling. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual, 5213–5225. https://proceedings.mlr.press/v139/kairouz21b.html

[31] Antti Koskela, Joonas Jälkö, and Antti Honkela. 2020. Computing Tight Differential Privacy Guarantees Using FFT. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, Palermo, Sicily, Italy, 2560–2569. http://proceedings.mlr.press/v108/koskela20b.html

[32] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images*. Technical Report. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[33] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. 2022. Toward Training at ImageNet Scale with Differential Privacy. *CoRR* abs/2201.12328 (2022), 1–25. arXiv:2201.12328 https://arxiv.org/abs/2201.12328

[34] Hao Liang, Wanrong Zhang, Xinlei He, Kaishun Wu, and Hong Xing. 2025. An Improved Privacy and Utility Analysis of Differentially Private SGD with Bounded Domain and Smooth Losses. *CoRR* abs/2502.17772 (2025), 1–19. arXiv:2502.17772 doi:10.48550/ARXIV.2502.17772

[35] Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A. Choquette-Choo, Badih Ghazi, George Kaissis, Ravi Kumar, Ruibo Liu, Da Yu, and Chiyuan Zhang. 2025. Scaling Laws for Differentially Private Language Models. arXiv:2501.18914 [cs.LG] https://arxiv.org/abs/2501.18914

[36] Sebastian Meiser and Esfandiar Mohammadi. 2018. Tight on Budget?: Tight Bounds for r-Fold Approximate Differential Privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang (Eds.). ACM, Toronto, ON, Canada, 247–264. doi:10.1145/3243734.3243765

[37] Ilya Mironov. 2017. Rényi Differential Privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*. IEEE Computer Society, Barbara, CA, USA, 263–275. doi:10.1109/CSF.2017.11

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. NeurIPS, Granada, Spain, 9 pages. http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf

[39] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy. *J. Artif. Intell. Res.* 77 (2023), 1113–1201. doi:10.1613/JAIR.1.14649

[40] Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik. 2021. ImageNet-21K Pretraining for the Masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). NeurIPS, Virtual, 12 pages. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/98f13708210194c475687be6106a3b84-Abstract-round1.html

[41] Tom Sander, Pierre Stock, and Alexandre Sablayrolles. 2023. TAN Without a Burn: Scaling Laws of DP-SGD. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, Honolulu, Hawaii, USA, 29937–29949. https://proceedings.mlr.press/v202/sander23b.html

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). NeurIPS, New Orleans, LA, USA, 17 pages. http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[43] Yinchen Shen, Zhiguo Wang, Ruoyu Sun, and Xiaojing Shen. 2021. Towards Understanding the Impact of Model Size on Differential Private Classification. *CoRR* abs/2111.13895 (2021), 1–14. arXiv:2111.13895 https://arxiv.org/abs/2111.13895

[44] Moshe Shenfeld and Vitaly Feldman. 2025. Privacy amplification by random allocation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (Eds.). NeurIPS, San Diego, CA, USA, 43 pages. http://papers.nips.cc/paper_files/paper/2025/hash/8a724af64e891da5c84078af2c308a72-Abstract-Conference.html

[45] Amer Sinha, Thomas Mesnard, Ryan McKenna, Daogao Liu, Christopher A. Choquette-Choo, Yangsibo Huang, Da Yu, George Kaissis, Zachary Charles, Ruibo Liu, Lynn Chua, Pritish Kamath, Pasin Manurangsi, Steve He, Chiyuan Zhang, Badih Ghazi, Borja De Balle Pigem, Prem Eruvbetine, Tris Warkentin, Armand Joulin, and Ravi Kumar. 2025. VaultGemma: A Differentially Private Gemma Model. arXiv:2510.15001 [cs.CR] https://arxiv.org/abs/2510.15001

[46] Nurislam Tursynbek, Aleksandr Petiushko, and Ivan V. Oseledets. 2020. Robustness Threats of Differential Privacy. *CoRR* abs/2012.07828 (2020), 1–16. arXiv:2012.07828 https://arxiv.org/abs/2012.07828

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). NeurIPS, Long Beach, CA, USA, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[48] Hanshen Xiao and Srinivas Devadas. 2023. PAC Privacy: Automatic Privacy Measurement and Control of Data Processing. In *Advances in Cryptology - CRYPTO 2023 - 43rd Annual International Cryptology Conference, CRYPTO 2023, Santa Barbara, CA, USA, August 20-24, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14082)*, Helena Handschuh and Anna Lysyanskaya (Eds.). Springer, Santa Barbara, CA, USA, 611–644. doi:10.1007/978-3-031-38545-2_20

[49] 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. 2021. Opacus: User-Friendly Differential Privacy Library in PyTorch. *CoRR* abs/2109.12298 (2021), 18 pages. arXiv:2109.12298 https://arxiv.org/abs/2109.12298

[50] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). NeurIPS, Montreal, Quebec, Canada, 649–657. https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html

[51] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464. doi:10.1109/TPAMI.2017.2723009

[52] Chaoyi Zhu, Jiayi Tang, Juan F. Pérez, Marten van Dijk, and Lydia Y. Chen. 2025. DP-TLDM: Differentially Private Tabular Latent Diffusion Model. In *Availability, Reliability and Security - 20th International Conference, ARES 2025, Ghent, Belgium, August 11-14, 2025, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15992)*, Mila Dalla Preda, Sebastian Schrittwieser, Vincent Naessens, and Bjorn De Sutter (Eds.). Springer, Ghent, Belgium, 337–357. doi:10.1007/978-3-032-00624-0_17