ISSN:2790-1688 Volume-15-(2025)

Research on Exploring Fairness Challenges in Differential Privacy Machine Learning

Zhenshan Wu

Data Science and Intelligent Media Institute; Communication University of China; Beijing; 100000; China

zs-wu21@cuc.edu.cn

Abstract. Differential Privacy (DP), the gold standard for privacy protection in machine learning, provides strict mathematical guarantees but also exhibits a "double-edged sword" effect by posing a potential threat to algorithmic fairness. This study systematically reveals that DP, particularly its mainstream implementation paradigm—Differentially Private Stochastic Gradient Descent (DP-SGD)—disproportionately impairs model performance on minority or underrepresented subgroups in its application, thereby exacerbating, rather than mitigating, existing algorithmic biases. This paper first elucidates the core concepts and mechanisms of DP, then analyzes the empirical evidence and intrinsic mechanisms leading to its disparate impact, identifying gradient clipping and noise injection as the key contributing factors. Building on this analysis, the paper comprehensively reviews various state-of-the-art techniques designed to mitigate this issue. It charts an evolutionary path from initial problem identification, through the remediation of existing mechanisms (e.g., group-wise and adaptive clipping), to the principled co-design of privacy and fairness (e.g., FairDP). This research aims to provide a comprehensive theoretical perspective and a technical roadmap for constructing trustworthy artificial intelligence systems that concurrently ensure privacy protection and social equity.

Keywords: Differential Privacy(DP); Machine Learning; Algorithmic Fairness; Differential Privacy Stochastic Gradient Descent; Privacy-Utility-Fairness Trade-off.

1. Introduction

In the era of data-driven decision-making, machine learning technology has been widely applied across various fields, but the collection of massive sensitive data has raised privacy concerns. As the gold standard for privacy protection, Differential Privacy (DP), proposed by Dwork et al. in 2006, ensures privacy through mathematically rigorous noise injection, making analysis outcomes independent of any single individual's data[1]. It has been adopted by major organizations like Apple and Google. However, deeper integration of DP in machine learning reveals its impact on model fairness: mechanisms like DP-SGD disproportionately harm minority groups' model performance, exacerbating algorithmic bias. For instance, DP degrades recommendation accuracy more for inactive (minority) users than active ones. Given DP's growing prevalence and its potential equity risks, this review systematically analyzes fairness challenges in differentially private machine learning, clarifying the "double-edged sword" effects—how gradient clipping and noise injection in DP-SGD systematically degrade minority group performance. It aims to provide theoretical insights and technical roadmaps for constructing trustworthy AI systems that balance strict privacy guarantees with social equity.

2. Differential Privacy: Concepts, Mechanisms and Applications

Differential Privacy (DP) provides a rigorous mathematical framework for privacy protection by ensuring an algorithm's output is insensitive to any single record in its input dataset[2]. This is formalized by the (ϵ,δ)-DP definition, which guarantees that for any two adjacent datasets (differing by one record), the probability of obtaining any output is nearly identical, bounded by a privacy budget ϵ and a failure probability δ.

To achieve this, DP mechanisms inject calibrated noise into a computation, with the noise scale determined by the function's sensitivity—its maximum possible change in output from a single record's alteration. The two primary mechanisms are the Laplace mechanism, which uses L1 sensitivity to achieve pure ϵ-DP and is ideal for low-dimensional queries, and the Gaussian mechanism, which uses L2 sensitivity to achieve (ϵ,δ)-DP. The Gaussian mechanism dominates in machine learning because the L2 norm of high-dimensional vectors (like model gradients) is often much smaller than the L1 norm, permitting less noise and thus preserving greater utility.

In practice, DP is deployed via two main architectures. Centralized DP (CDP) relies on a trusted curator to collect raw data and apply noise once to the aggregated result, maximizing data utility. Local DP (LDP) requires each user to add noise on-device before transmission, offering stronger privacy by removing trust in a central entity but at the cost of significant utility loss. This fundamental trade-off dictates deployment choices: LDP is used for simple statistics (e.g., Apple's telemetry), while complex tasks like model training almost exclusively use CDP to maintain viability[3].

3. Applications of Differential Privacy in Machine Learning

3.1 Differentially Private Stochastic Gradient Descent (DP-SGD)

Deep neural networks, due to their vast number of parameters, possess a powerful capacity for memorization. This characteristic makes them highly susceptible to two severe types of privacy attacks: Membership Inference Attacks (MIAs) and data extraction attacks. MIAs can infer whether a sample exists in the training set by analyzing differences in a model's predictive behavior on specific data, potentially revealing sensitive information such as an individual's health status; the pioneering work of Shokri et al.[4]. confirmed the feasibility of such attacks. The more severe data extraction attacks can directly recover the original text of training samples, with Carlini et al.[5]. demonstrating that large language models can leak real data, including personally identifiable information and private conversations.

To counter these threats, Differentially Private Stochastic Gradient Descent (DP-SGD), proposed by Abadi et al., has become the gold standard for privacy protection in deep learning[6]. This mechanism achieves strict differential privacy guarantees by implementing two core modifications to the standard SGD algorithm. Per-sample gradient clipping first computes the independent gradient for each sample in a mini-batch and constrains its L2 norm with a threshold—if it exceeds a preset value C, it is scaled down proportionally to C. This operation limits the maximum influence of any single sample on the gradient update, providing a mathematical upper bound for the sensitivity of the overall update step. Noise injection occurs during the aggregation of clipped gradients (typically through summation or averaging), where Gaussian noise of a specific intensity, $N(0,\sigma 2C2)$, is injected into each dimension of the aggregated gradient, with the noise multiplier σ being directly related to the privacy budget ϵ. Finally, the model parameters are updated using the clipped and noised gradients, forming the core barrier against privacy attacks.

3.2 Other Differential Privacy Learning Paradigms

In addition to the mainstream DP-SGD, alternative paradigms such as PATE, DP-GANs, and DP-FL offer important pathways for privatized machine learning, each exhibiting distinct trade-offs in privacy, utility, and fairness. The Private Aggregation of Teacher Ensembles (PATE)[7], a mature alternative framework, employs a multi-stage training strategy: multiple "teacher" models are independently trained on private data subsets, and a "student" model is subsequently trained using the noisy voting results of the teacher ensemble on public, unlabeled data. Its core innovation lies in placing the privacy perturbation at the model output level (the aggregation of predicted labels), rather than interfering with the gradient update process as in DP-SGD. This high-level abstract perturbation naturally circumvents the systematic biases caused by fine-grained gradient operations. Other notable approaches include Differentially Private Generative Adversarial Networks (DP-GANs), which focus on data synthesis by imposing DP-SGD constraints on the discriminator , and Differentially Private

Federated Learning (DP-FL), which addresses distributed scenarios by applying noise to model updates.

The fundamental difference among these paradigms lies in the architectural level of privacy injection: DP-SGD operates at the finest granularity of per-sample gradient computation, with its clipping and noise directly affecting model parameter updates; PATE, on the other hand, perturbs the higher-level collective decisions of the teachers, and the internal learning process of the teacher models is not directly affected by noise. This architectural isolation theoretically endows PATE with stronger robustness against the gradient-level biases of DP-SGD, providing a structural basis for the empirical conclusion discussed later that PATE has a milder disparate impact.

4. Differential Privacy and the Fairness Dilemma

4.1 Metrics for Algorithmic Fairness

To rigorously discuss "fairness," it must first be transformed from a vague concept into quantifiable technical indicators. In the field of algorithmic fairness, researchers have proposed a series of metrics, primarily focusing on Group Fairness, which requires that a model's performance across different protected groups (e.g., groups divided by race or gender) should exhibit some form of equality.

In binary classification tasks, assume A is a protected attribute (e.g., A=a represents a disadvantaged group, A=b represents an advantaged group), Y is the true label (Y=1 for a positive outcome), and $\hat{Y}$is the model's predicted outcome. The following are several core group fairness metrics:

Table1 Key Group Fairness Metrics[8][9][10]

Metric Name	Definition and Explanation	Mathematical Formula
Demographic Parity	Requires that different groups receive positive predicted outcomes at equal rates. It focuses solely on the distribution of prediction outcomes, ignoring prediction accuracy. This is a strong 'equality of outcome' requirement.	P(Y^=1)
Equalized Odds	Requires that, for a given true outcome (positive or negative), different groups receive positive predicted outcomes at equal rates. This means that both the true positive rate (TPR) and the false positive rate (FPR) must be equal across groups, ensuring the model's predictive power is balanced.	P(Y^=1)
Accuracy Parity	Requires that the model's overall accuracy is the same for different groups. This is a relatively weaker fairness standard but is intuitive and easy to understand.	P(Y^=Y)

4.2 The Disparate Impact of Differential Privacy: A Review of Empirical Studies

The academic community first systematically revealed the adverse effects of differential privacy on fairness in the pioneering work published by Bagdasaryan, Poursaeed, and Shmatikov[11]. Their research, along with a substantial body of subsequent work, has empirically demonstrated that the accuracy cost of DP-SGD is not uniformly distributed across all data subgroups but instead exhibits a significant disparate impact. Their core finding is a "the rich get richer, and the poor get poorer" effect: for minority groups or hard-to-classify categories that already have lower accuracy in non-private models, the drop in accuracy after applying DP-SGD is far greater than for advantaged groups. This effect means that differential privacy not only fails to mitigate but actually exacerbates the model's pre-existing unfairness.

This finding has been confirmed in two key use cases: racial bias in facial recognition and dialect bias in natural language processing. In gender classification tasks based on facial images, researchers found that models trained with DP-SGD exhibited a significantly larger drop in classification accuracy for individuals with darker skin tones (a minority group) compared to those with lighter skin tones (the majority group). This worsened the pre-existing racial bias in the privatized models. In sentiment analysis of tweets, DP-SGD models showed a drastic performance decline when processing African American Vernacular English (AAVE), an underrepresented dialect, while their performance on Standard American English was much less affected. Similar issues have been observed in tasks like hate speech detection.

These laboratory findings are mirrored in the high-stakes, real-world case of the U.S. 2020 Census, where the Bureau's adoption of a DP-based algorithm was a landmark deployment[12]. Subsequent analyses revealed that the injected noise had a disproportionate impact, introducing greater errors into the statistical data for rural areas, non-white communities, and other ethnic minority groups[13]. This consequence is severe, as census data determines the allocation of congressional seats, electoral districts, and trillions in federal funding. When a technology designed to protect privacy systematically undermines the representation and resources of marginalized groups, a technical issue becomes a profound problem of social justice, making the search for mitigation strategies an urgent ethical task.

4.3 Investigating the Mechanisms of Impact: Why DP-SGD Exacerbates Unfairness

To understand why Differential Privacy (DP) systematically amplifies model unfairness, it is necessary to deeply analyze how the core operations of its machine learning implementation paradigm, DP-SGD—gradient clipping and noise injection—interact with data imbalance to produce structural bias[11]. Gradient clipping, as the primary step, is a key source of this disparate impact. In model training, the norm of a sample's gradient often reflects its learning difficulty: majority group samples, being well-fitted by the model, generally have smaller gradient norms; whereas minority group samples, outliers, or complex examples produce larger-norm gradients due to higher prediction uncertainty, and these gradients contain crucial learning signals. The uniform clipping threshold C imposed by DP-SGD, however, disproportionately suppresses these large-norm gradients. As minority group samples are clipped more frequently and more severely, their effective learning rate is systematically weakened, and their contribution to model updates is significantly reduced. This process introduces a structural clipping bias into the mini-batch aggregated gradient, causing its direction to deviate from the true gradient and skew towards the smaller, less-clipped gradients of the majority group samples.

Noise injection, as the second core step, further exacerbates the fairness imbalance. Its root cause lies in the inherent difference in signal strength correlated with group size: in a random mini-batch, majority group samples are naturally more numerous, and their contributed aggregated gradient signal is significantly stronger than that of the minority group. The group-agnostic, equal-magnitude Gaussian noise added by DP-SGD has a non-uniform destructive effect on these two types of signals—it has a limited impact on the strong signal from the majority group but can easily overwhelm the already weak signal from the minority group, leading to a sharp deterioration in their signal-to-noise ratio.

In summary, the causal chain through which DP-SGD exacerbates unfairness is clear: in the context of data imbalance, minority group samples produce larger gradient norms; a uniform clipping threshold systematically suppresses these information-rich gradients, constituting the first major blow to fairness; the weak aggregated signal from the minority group in the mini-batch is then overwhelmed by an equal amount of noise, forming the second blow. During the training process, this dual, systematic suppression of learning signals from minority groups continuously accumulates, ultimately leading to a significant degradation in the model's predictive performance for them, thereby amplifying pre-existing biases. This mechanistic explanation profoundly reveals how the "double-

edged sword" of differential privacy, while providing privacy protection, unintentionally yet inevitably harms fairness[14].

5. Alleviating Injustice: The Path to Fairness through Differential Privacy Machine Learning

To address DP's systematic harm to fairness, researchers have developed three core strategies: improving DP-SGD, exploring alternative paradigms, and fairness-aware co-design.

Table2 Overview of Major Mitigation Strategies for Disparate Impact of Differential Privacy[15-21]

Strategy Category	Representative Algorithm/Method	Core Mechanism
Improving DP-SGD Mechanism	Group-wise Clipping	Set independent clipping thresholds Cgroup for different groups.
	Adaptive Clipping	Dynamically adjust the global clipping threshold C during training.
	Bounded Adaptive Clipping	Add a tunable lower bound to adaptive clipping.
Alternative DP Paradigms	PATE	Train a student model through noisy aggregation of votes from teacher models.
Fairness-Aware Co-Design	FairDP	Train models for each group independently, then integrate them in a private manner.

Improving DP-SGD focuses on correcting the bias from gradient clipping. Group-wise Clipping sets independent thresholds for different demographic groups, but requires group labels. Adaptive Clipping dynamically adjusts a global threshold but can still shrink excessively and harm minority groups. The state-of-the-art solution, Bounded Adaptive Clipping, introduces a tunable lower bound to prevent this collapse, significantly protecting the learning signal for minority groups and improving worst-group accuracy by 5-10 percentage points.

The PATE paradigm offers a structural alternative that avoids gradient clipping altogether. By training a student model on the noisy aggregated votes of a teacher ensemble, it applies privacy at a higher level of abstraction. Empirical evidence confirms PATE's negative impact on fairness is significantly weaker than DP-SGD's. However, its reliance on public unlabeled data and its potential to fail on extremely sparse groups limit its applicability.

Fairness-aware co-design marks a paradigm shift by integrating fairness and privacy as primary design goals. The FairDP mechanism, a major breakthrough, trains separate models for each group and then privately fuses their knowledge. This architecture not only allows for fine-grained control over noise but also leverages the noise distribution to generate a formally provable fairness certificate for the final model. This evolution from problem identification to mechanism repair and now to principled co-design advances the field from empirical improvements toward an era of certifiable guarantees

6. Conclusion

This study has systematically revealed the profound fairness dilemma triggered by differential privacy, particularly its machine learning implementation paradigm, DP-SGD, while providing strict mathematical privacy guarantees. This 'double-edged sword' effect is not a random perturbation but a systematic and disproportionate impairment of model performance for minority and underrepresented subgroups in the data. Its intrinsic root lies in the disparate impact caused by the two core operations of DP-SGD: gradient clipping and noise injection. Gradient clipping disproportionately suppresses the more informative gradients from minority group samples, introducing a significant 'clipping bias'; noise injection further overwhelms the already weak 'learning

signals' of these groups, drastically reducing their signal-to-noise ratio. The synergistic effect of these two factors systematically amplifies the model's pre-existing biases during training. This disparate impact is by no means confined to theoretical discussion; its real-world consequences have become apparent in high-stakes scenarios such as the U.S. 2020 Census, where greater distortion of statistical data for rural and ethnic minority communities directly threatens their political representation and fair allocation of public resources, highlighting the urgency of resolving this issue. In response to this fundamental conflict, the research field has undergone an evolutionary process from problem identification, to the mitigation and repair of existing mechanisms (such as group-wise and adaptive clipping), and finally to a principled co-design approach. The latest algorithms (such as FairDP) can now provide formally provable fairness certificates, marking a substantial breakthrough in reconciling the goals of privacy protection and fairness.

Despite significant progress in the field of privacy and fairness co-design, building truly fair and private machine learning systems still faces multidimensional challenges that require in-depth exploration in future research. The primary direction is to move beyond the limitations of group fairness. Most current work focuses on this, while the intersection of more granular fairness concepts—such as individual fairness (requiring similar individuals to be treated similarly) and casual fairness (aiming to eliminate causal discrimination based on protected attributes)—with differential privacy remains to be explored. Secondly, the precise quantitative characterization of the complex trade-offs among privacy, fairness, and utility remains a core theoretical problem; there is an urgent need to derive tighter, quantifiable bounds to definitively answer questions like, 'What is the optimal fairness guarantee achievable for a given privacy budget ε?'. Thirdly, improving the scalability and efficiency of algorithms is crucial. Many theoretically superior fairness-aware differential privacy algorithms are difficult to apply to large-scale models and datasets due to significant computational or memory overhead; developing efficient solutions, especially in distributed environments like federated learning, is a pressing need. Finally, emerging AI paradigms, such as Large Language Models (LLMs) and generative AI, present unprecedented challenges. These models are often trained on vast web-scale data imbued with societal biases and pose enormous privacy leakage risks. Extending existing theories of fair and private learning to these new paradigms is an urgent and challenging research frontier.

References

[1] Dwork C. Differential privacy[C]//International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1-12.

[2] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of cryptography conference. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 265-284.

[3] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and trends® in theoretical computer science, 2014, 9(3–4): 211-407.

[4] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

[5] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//30th USENIX security symposium (USENIX Security 21). 2021: 2633-2650.

[6] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016: 308-318.

[7] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data[J]. arXiv preprint arXiv:1610.05755, 2016.

[8] Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness[C]//Proceedings of the 3rd innovations in theoretical computer science conference. 2012: 214-226.

[9] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning[J]. Advances in neural information processing systems, 2016, 29.

[10] Uniyal A, Naidu R, Kotti S, et al. Dp-sgd vs pate: Which has less disparate impact on model accuracy?[J]. arXiv preprint arXiv:2106.12576, 2021.

[11] Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy[J]. Advances in neural information processing systems, 2019, 32.

[12] Kenny C T, Kuriwaki S, McCartan C, et al. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census[J]. Science advances, 2021, 7(41): eabk3283.

[13] Santos-Lozada A R, Howard J T, Verdery A M. How differential privacy will affect our understanding of health disparities in the United States[J]. Proceedings of the National Academy of Sciences, 2020, 117(24): 13405-13412.

[14] Mangold P, Perrot M, Bellet A, et al. Differential privacy has bounded impact on fairness in classification[C]//International Conference on Machine Learning. PMLR, 2023: 23681-23705.

[15] Andrew G, Thakkar O, McMahan B, et al. Differentially private learning with adaptive clipping[J]. Advances in Neural Information Processing Systems, 2021, 34: 17455-17466.

[16] Christ L, Amiriparian S, Kathan A, et al. Towards Multimodal Prediction of Spontaneous Humor: A Novel Dataset and First Results[J]. IEEE Transactions on Affective Computing, 2024.

[17] Zhao L, Rehn A, Heikkilä M A, et al. Mitigating Disparate Impact of Differentially Private Learning through Bounded Adaptive Clipping[J]. arXiv preprint arXiv:2506.01396, 2025.

[18] Tran K, Fioretto F, Khalil I, et al. Fairdp: Certified fairness with differential privacy[J]. arXiv preprint arXiv:2305.16474, 2023.

[19] Jagielski M, Kearns M, Mao J, et al. Differentially private fair learning[C]//International Conference on Machine Learning. PMLR, 2019: 3000-3008.

[20] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016: 308-318.

[21] Zhu T, Li G, Zhou W, et al. Differential privacy and applications[M]. Cham, Switzerland: Springer International Publishing, 2017.