CAN VARIABLE NOISE BE DATA-INDEPENDENT? DIFFERENTIALLY PRIVATE BIMODAL SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

DP-SGD serves as the de facto baseline for private deep learning, yet its constant noise scale imposes an unfavorable dilemma: small noise rapidly depletes privacy budgets, while large noise degrades gradient signals. While adaptive noise mitigates this issue, its data dependence violates essential data independence, risking leakage. We propose Differentially Private Bimodal SGD (DP-BiSGD), a strictly data-independent baseline that stochastically alternates between high-probability small perturbations for local exploitation and low-probability large perturbations for global exploration. To translate this concept into a viable baseline, we address two challenges inherent in this bimodal formulation. Specifically, we develop a bypass-scaled update scheme that shields the momentum buffer from extreme-variance corruption, coupled with a dynamic GMM-PLD accountant to overcome loose standard accounting and ensure tight privacy bounds. Extensive experiments demonstrate that DP-BiSGD consistently outperforms DP-SGD in model utility under identical privacy guarantees, preserves strong resilience against inference attacks, and serves as a seamless drop-in replacement.

1 INTRODUCTION

Deep neural networks extract rich representations from massive datasets to power applications such as generative artificial intelligence (Wairimu (2025)), recommender systems (Dasgupta et al. (2025)), and autonomous agents (Li et al. (2025)). However, training these models relies on vast amounts of sensitive personal data. Through iterative gradient updates, they risk memorizing individual training samples, exposing them to privacy vulnerabilities such as membership inference (Nguyen et al. (2025); Kong et al. (2024)) and data inversion attacks (Peng et al. (2024)). Differential Privacy (DP) Dwork (2006) offers a rigorous mathematical framework to quantify and mitigate these risks. In practice, Differentially Private Stochastic Gradient Descent (DP-SGD) Abadi et al. (2016) has served as the de facto standard for private deep learning for nearly a decade. DP-SGD enforces privacy by clipping per-sample gradient ℓ_2-norms to bound sensitivity and injecting calibrated Gaussian noise into each parameter update.

Despite its ubiquity, DP-SGD suffers from a rigid perturbation paradigm that injects constant-variance Gaussian noise across all iterations. In non-convex loss landscapes, this uniform noise creates an unresolvable dilemma: low noise rapidly depletes the privacy budget, whereas high noise permanently degrades the gradient signal-to-noise ratio (SNR) and traps optimization in sub-optimal local minima. To address this, recent studies attempt to patch DP-SGD via adaptive noise scaling, dynamically adjusting perturbation according to data characteristics (Xue et al. (2024)) or real-time gradient variance (Zhu et al. (2025); Räisä et al. (2024)). However, relying on intermediate data representations exposes these heuristics to implicit privacy leakage and precludes rigorous, data-independent theoretical guarantees.

Consequently, rather than patching DP-SGD with incremental data-dependent heuristics, we rethink its perturbation paradigm and ask: *Can a strictly data-independent, variable noise mechanism structurally replace DP-SGD's uniform perturbation to establish a superior baseline?* From an optimization perspective, uniform perturbation is structurally flawed for deep neural networks, where escaping flat valleys requires large stochastic jumps, yet converging to sharp minima demands low-variance gradients. This duality motivates us to investigate a step-wise bimodal Gaussian mechanism. By stochastically alternating between high-probability small noise for local exploitation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and low-probability large noise for global exploration, this approach theoretically achieves a better privacy-utility trade-off than uniform noise. This data-independent mechanism establishes a new foundational baseline, Differentially Private Bimodal Stochastic Gradient Descent (DP-BiSGD), designed as a seamless drop-in replacement for standard DP-SGD.

However, translating this intuition into a viable baseline presents two critical challenges. On the optimization front, standard recursive state accumulation in momentum-based optimizers causes the extreme variance of large-noise steps to propagate forward, permanently polluting the momentum buffer and destroying the SNR advantage of subsequent small-noise steps. On the theoretical front, because the bimodal perturbation forms a Gaussian Mixture Model (GMM), standard Rényi Differential Privacy Mironov (2017) accountants yield overly pessimistic bounds, while continuous multi-step convolutions across training iterations remain analytically and computationally intractable.

To overcome these challenges, we introduce two architectural designs complementing the bimodal mechanism. First, to eliminate momentum corruption, we design a bypass-scaled update mechanism: small-noise gradients update the momentum buffer normally, while large-noise updates bypass the buffer entirely, enabling exploratory jumps without disrupting subsequent optimization trajectories. Second, to ensure precise privacy accounting, we develop a dynamic GMM-PLD accountant. By representing step-wise perturbations as a unified GMM and computing multi-step compositions via discrete Fast Fourier Transform (FFT), this accountant bypasses loose RDP relaxations to deliver tight privacy bounds that maximize allowable training iterations.

In summary, the contributions of this paper can be concluded as follows.

- We propose DP-BiSGD, departing from the constant-noise paradigm to establish a new foundational baseline for private deep learning. While preserving the standard DP-SGD pipeline, it achieves superior utility under identical privacy guarantees without data-dependent leakage.
- We develop two mechanisms to resolve the optimization and accounting challenges in DP-BiSGD: a bypass-scaled update scheme that prevents the momentum buffer from large-noise corruption, and a dynamic GMM-PLD accountant to provide tight bounds beyond loose RDP relaxations.
- We establish rigorous theoretical guarantees for DP-BiSGD and comprehensively validate its superior utility, seamless algorithmic compatibility, and robust resilience against inference attacks.

2 DP-BiSGD

In this section, we present the Differentially Private Bimodal SGD (**DP-BiSGD**) framework.

2.1 Framework Overview

As illustrated in Figure 1, the DP-BiSGD framework consists of three interdependent modules. Their relationships and specific functions are outlined as follows:

- **Step-Wise Bimodal Gaussian Perturbation:** This module executes data perturbation to enforce privacy guarantees and feeds the resulting data into the model training process. It dynamically determines the noise scale of the current training step via a data-independent Bernoulli trial, alternating between fine-grained local descent (small noise) and exploratory jumps (large noise).
- **Bypass-Scaled Update Mechanism:** This module handles parameter updates, utilizing the perturbed gradients from the first module. It isolates large-noise gradients from the momentum buffer, ensuring that exploratory jumps do not contaminate subsequent high-fidelity updates.
- **Dynamic GMM-PLD Accountant:** This component operates synchronously with the perturbation module to track cumulative privacy loss. By modeling the step-wise noise as a unified, hidden-coin Gaussian Mixture Model, it exploits the tail-flattening effect to guarantee a rigorous and tight privacy bound, which serves as a dynamic stopping criterion to determine the maximum allowable training steps.

2.2 Step-Wise Bimodal Gaussian Perturbation

DP-BiSGD dynamically adjusts the noise scale at each training step via bimodal perturbation to generate private gradients. Let D denote a training dataset of n samples. At each step t, we apply

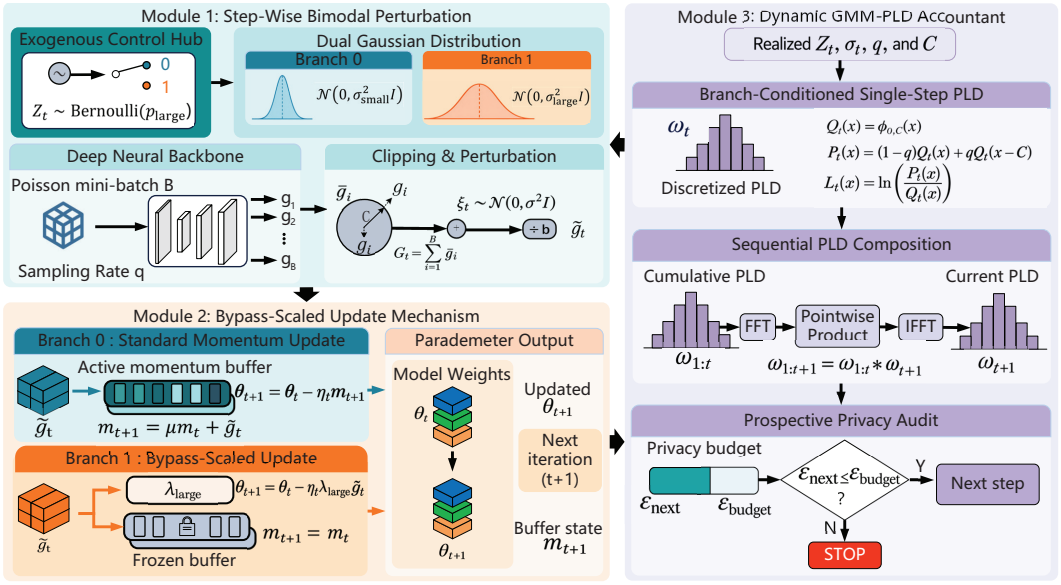


Figure 1: Overview of the framework.

Poisson sampling with rate q to form a mini-batch $\mathcal{B}_t$. For each instance $(x_i, y_i) \in \mathcal{B}_t$, where x_i and y_i denote the feature vector and label, respectively, we compute the per-sample gradient $g_t(x_i, y_i)$ and clip its ℓ_2 norm to a threshold C to obtain the clipped gradient $\overline{g}_t(x_i, y_i)$. We then compute the clipped gradient sum:

$$G_t = \sum_{(x_i, y_i) \in \mathcal{B}_t} \overline{g}_t(x_i, y_i). \tag{1}$$

Under standard differential privacy definitions, the ℓ_2 sensitivity of G_t is strictly bounded by C. Additionally, we define the expected batch size as $b = qn$ and use it instead of the fluctuating empirical batch size as the fixed normalization factor for the gradient update.

Unlike conventional approaches that maintain a static noise scale, our method dynamically perturbs gradients using three predefined structural hyperparameters, which include a large-noise branch probability p_{large} and two ordered noise multipliers satisfying $0 < \sigma_{\text{small}} < \sigma_{\text{large}}$. For every training step t, we independently draw a binary branch indicator Z_t from a Bernoulli distribution:

$$Z_t \sim \text{Bernoulli}(p_{\text{large}}). \tag{2}$$

The indicator Z_t determines the active noise multiplier for the current step, setting $\sigma_t = \sigma_{\text{large}}$ if $Z_t = 1$, and $\sigma_t = \sigma_{\text{small}}$ otherwise. We then sample a Gaussian noise vector ξ_t calibrated to the sensitivity C and the selected multiplier σ_t:

$$\xi_t \sim \mathcal{N}\left(0, \sigma_t^2 C^2 I\right), \tag{3}$$

where $I \in \mathbb{R}^{d \times d}$ is an identity matrix matching the dimension of the model parameters. Finally, we perturb the clipped gradient sum G_t with ξ_t and normalize the result by the expected batch size b to obtain the private gradient $\widetilde{g}_t$:

$$\widetilde{g}_t = \frac{1}{b}\left(G_t + \xi_t\right). \tag{4}$$

2.3 Bypass-Scaled Update Mechanism

While bimodal perturbation facilitates stochastic exploration, naively accumulating large-noise gradients into the recursive momentum buffer propagates high variance across steps, severely degrading the signal-to-noise ratio of future small-noise updates. To prevent this momentum contamination,

DP-BiSGD adopts a bypass-scaled update mechanism that leverages the global branch indicator Z_t to precisely control the optimizer behavior.

Specifically, when $Z_t = 0$, DP-BiSGD executes a standard momentum update to exploit fine-grained gradient directions. We first compute the regularized descent direction $d_t \in \mathbb{R}^d$:

$$d_t = \widetilde{g}_t + \gamma P_w \theta_t, \tag{5}$$

where $\gamma \geq 0$ is the weight decay coefficient, and $P_w \in \{0,1\}^{d\times d}$ is a diagonal projection matrix that applies weight decay exclusively to the weights, excluding bias and normalization parameters. We then update the momentum buffer and model parameters according to:

$$m_{t+1} = \mu m_t + d_t, \quad \theta_{t+1} = \theta_t - \eta_t m_{t+1}, \tag{6}$$

where $\mu \in [0,1)$ denotes the momentum decay factor, and η_t represents the learning rate.

Conversely, when $Z_t = 1$, DP-BiSGD bypasses the recursive momentum buffer to prevent the extreme-variance perturbations from poisoning future iterations. Rather than accumulating the heavily perturbed gradient, we freeze the momentum state and update the parameters directly via the scaled private gradient:

$$\theta_{t+1} = \theta_t - \eta_t \lambda_{\text{large}} \widetilde{g}_t, \quad m_{t+1} = m_t, \tag{7}$$

where $\lambda_{\text{large}} \in (0,1]$ is a scaling factor designed to further dampen the physical magnitude of the exploratory jumps.

2.4 Dynamic GMM-PLD Accountant

To track the cumulative privacy loss, we design a dynamic Privacy Loss Distribution (PLD) accountant. Since the perturbation is probabilistically drawn from two Gaussian distributions, the overall injected noise forms a Gaussian Mixture Model (GMM). As this GMM noise is uniformly applied across all dimensions, the worst-case privacy leakage depends on the maximum ℓ_2 gradient distance, bounded by the clipping threshold C. Our dynamic accounting process executes in four steps.

Step 1. Constructing Unified Mixture Densities. Due to the isotropy of the injected GMM noise, the worst-case privacy leakage in $\mathbb{R}^d$ is equivalent to projecting the distributions onto a 1D scalar space along the maximum gradient deviation. Given the ℓ_2 sensitivity C, the maximum shift between adjacent projected distributions is C. Let $x \in \mathbb{R}$ denote the 1D observation, the baseline null distribution (representing the absence of the target sample) is a unified 1D Gaussian mixture:

$$Q(x) = \frac{1-p_{\text{large}}}{\sqrt{2\pi}\sigma_{\text{small}}C} \exp\left(-\frac{x^2}{2\sigma_{\text{small}}^2 C^2}\right) + \frac{p_{\text{large}}}{\sqrt{2\pi}\sigma_{\text{large}}C} \exp\left(-\frac{x^2}{2\sigma_{\text{large}}^2 C^2}\right). \tag{8}$$

In the worst-case scenario, the gradient of the alternative dataset (including the target sample) shifts the center by the maximum distance C. Incorporating the Poisson subsampling rate q, the alternative distribution $P(x)$ becomes a mixture of the null distribution and its C-shifted counterpart:

$$P(x) = (1-q)Q(x) + qQ(x-C). \tag{9}$$

Step 2. Single-Step PLD. The single-step privacy loss is quantified by the log-likelihood ratio (LLR) between $P(x)$ and $Q(x)$:

$$\mathcal{L}(x) = \ln \frac{P(x)}{Q(x)} = \ln\left(1 - q + q\frac{Q(x-C)}{Q(x)}\right). \tag{10}$$

For Gaussian-based mechanisms, the likelihood ratio $\frac{Q(x-C)}{Q(x)}$ mathematically constitutes a strictly monotonically increasing function of x. Thus, the worst-case privacy leakage is defined by the supremum of $\mathcal{L}(x)$, which occurs as $x \to \infty$. Since $\sigma_{\text{small}} < \sigma_{\text{large}}$, the small-variance component in Equation 8 decays exponentially faster, meaning the large-variance component strictly dominates $Q(x)$ in the tail region. This yields the asymptotic likelihood ratio:

$$\lim_{x\to\infty} \frac{Q(x-C)}{Q(x)} = \lim_{x\to\infty} \exp\left(\frac{x^2 - (x-C)^2}{2\sigma_{\text{large}}^2 C^2}\right) \approx \exp\left(\frac{x}{\sigma_{\text{large}}^2 C}\right). \tag{11}$$

Substituting Equation 11 into Equation 10, the LLR's asymptotic growth is linearly bounded by σ_{large}:

$$\lim_{x\to\infty} \mathcal{L}(x) \approx \frac{x}{\sigma_{\text{large}}^2 C} + \ln(q). \tag{12}$$

Since the observation $x \sim P(x)$ is random, $\mathcal{L}(x)$ is also a random variable. We denote its probability distribution as ω_{step}, which serves as the base unit for tracking cumulative leakage.

Step 3. Composing Cumulative Distributions. Due to the independence of training steps, the cumulative privacy expenditure at iteration t is the t-fold convolution of the single-step distribution:

$$\omega_{\text{dom},t} = \omega_{\text{step}}^{\otimes t}, \tag{13}$$

where $\otimes t$ denotes convolution. To compute this efficiently for large t, we discretize the distribution and apply frequency-domain acceleration.

(1) Discretization and Pessimistic Bounding. We define an equidistant grid over $y = \mathcal{L}(x) \in [-L, L]$ with K bins and step size $\Delta y = 2L/(K-1)$. The continuous density $\omega_{\text{step}}(y)$ is mapped to a discrete vector $\mathbf{u} \in \mathbb{R}^K$. To preserve the worst-case guarantee, mass exceeding L is pessimistically accumulated into the terminal bin:

$$\mathbf{u}[k] = \begin{cases} \int_{y_k-\Delta y/2}^{y_k+\Delta y/2} \omega_{\text{step}}(y)\, dy, & \text{for } 0 \le k < K-1, \\ \int_{L-\Delta y/2}^{\infty} \omega_{\text{step}}(y)\, dy, & \text{for } k = K-1, \end{cases} \tag{14}$$

where $y_k = -L + k\Delta y$ is the k-th bin's center.

(2) Frequency-Domain Convolution. Using the Convolution Theorem, the t-fold convolution equates to element-wise exponentiation in the frequency domain. Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the forward and inverse Discrete Fourier Transforms. The discretized cumulative distribution vector $\mathbf{v}_t \in \mathbb{R}^K$ is computed as:

$$\mathbf{v}_t = \mathcal{F}^{-1}\left(\left(\mathcal{F}(\mathbf{u})\right)^{\circ t}\right), \tag{15}$$

where $(\cdot)^{\circ t}$ denotes element-wise power.

Step 4. Dynamic Stopping Criterion. To determine the stopping iteration, we define the discrete privacy failure function $H(\epsilon)$, mapping a candidate budget ϵ to its failure probability δ:

$$H(\epsilon) = \sum_{k:y_k \ge \epsilon} \left(1 - \exp(\epsilon - y_k)\right) \mathbf{v}_t[k]. \tag{16}$$

Because $\mathbf{v}_t[K-1]$ pessimistically absorbs the tail mass, this summation strictly upper-bounds the continuous integral. For a target failure probability δ_{target}, the consumed budget $\epsilon(t)$ at iteration t is the unique root of:

$$H(\epsilon(t)) = \delta_{\text{target}}. \tag{17}$$

Since $H(\epsilon)$ is monotonically decreasing, $\epsilon(t)$ is efficiently isolated via binary search. Finally, we evaluate the stopping condition against the total budget ϵ_{budget}:

$$\text{Action}_t = \begin{cases} \text{Halt Training}, & \text{if } \epsilon(t) > \epsilon_{\text{budget}}, \\ \text{Proceed to } t+1, & \text{otherwise.} \end{cases} \tag{18}$$

Remark 1. *The theoretical analysis of DP-BiSGD is deferred to Appendix A, including privacy guarantees, utility and convergence guarantees, and complexity analysis.*

3 Experiments

In this section, we conduct extensive experiments to evaluate the proposed DP-BiSGD framework. The code repository for this paper is available online[1].

[1] https://anonymous.4open.science/r/DP-BiSGD-D46D

Table 1: Impact of σ_{small} under Different σ_{base} Ratios on the Prediction Accuracy (%) of DP-BiSGD

Dataset	Metric	Relative Ratio Factor (α)													
		0.50	0.60	0.70	0.80	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
CIFAR-10	σ_{small}	5.50	6.60	7.70	8.80	9.90	10.01	10.12	10.23	10.34	**10.45**	10.56	10.67	10.78	10.89
	Accuracy	33.34	41.01	44.79	46.96	46.68	48.09	48.45	47.95	49.10	**49.73**	48.29	46.27	49.07	47.87
FMNIST	σ_{small}	2.00	2.40	2.80	3.20	3.60	3.64	**3.68**	3.72	3.76	3.80	3.84	3.88	3.92	3.96
	Accuracy	74.17	76.31	79.36	81.11	81.50	80.71	**81.52**	80.74	80.94	80.64	79.85	80.71	80.38	80.96

3.1 Experimental Setting

Research Questions. To systematically assess the performance of DP-BiSGD, our experimental evaluation is designed to answer four research questions. **RQ1:** What is the impact of parameters on the performance of DP-BiSGD? **RQ2:** Under what scenarios are the advantages of DP-BiSGD over DP-SGD most pronounced? **RQ3:** Can DP-BiSGD be effectively applied to improved models of DP-SGD? **RQ4:** How resilient is DP-BiSGD against inference attacks?

Remark 2. *The experimental analysis for RQ4 is deferred to Appendix B.*

Datasets. Mainstream privacy-preserving models are typically evaluated on the MNIST LeCun et al. (1998), Fashion-MNIST Xiao et al. (2017), and CIFAR-10 Krizhevsky et al. (2009) datasets. However, classification accuracy on MNIST remains above 95% even under stringent privacy guarantees such as $\epsilon = 1$. Such high accuracy yields limited insights for addressing our research questions. Therefore, we only report experimental results for the Fashion-MNIST and CIFAR-10 datasets.

Evaluation Metrics. We evaluate model utility using standard classification accuracy (%), where higher values indicate better performance.

Parameter Settings. Following standard private learning benchmarks (Fu et al. (2024); Wei et al. (2022)), we train a CNN using momentum SGD (momentum 0.9, learning rate 4.0, microbatch size 1), drawing batches via Poisson subsampling at a ratio of $q = B/N$, with batch sizes $B = 2048$ for Fashion-MNIST and 8192 for CIFAR-10. For privacy parameters, we fix the clipping bound to $C = 0.1$ and set $\delta = 10^{-5}$. Across $\epsilon \in \{1, 2, 3, 4\}$, we set the baseline noise scales σ_{base} to $\{4.0, 2.15, 2.15, 2.15\}$ on Fashion-MNIST and $\{11.0, 9.0, 5.67, 5.67\}$ on CIFAR-10, respectively.

3.2 RQ1: Impact of Parameters on DP-BiSGD Performance

In these experiments, we fix $\epsilon = 1$. To ensure consistency across datasets with different noise baselines, we evaluate the impact of σ_{small} and σ_{large} on DP-BiSGD by expressing them proportionally as $\alpha\sigma_{\text{base}}$, where α represents the relative ratio factor.

3.2.1 Impact of the Small Noise Scale σ_{small} on DP-BiSGD

As shown in Table 1, we vary the relative ratio factor α of σ_{small} with a coarse step of 0.1 from 0.5 to 0.9 and a fine-grained step of 0.01 from 0.9 to 0.99. The experimental results reveal several key insights.

First, as the ratio α increases from 0.50 to 0.90, the prediction accuracy exhibits a significant upward trend, climbing from 33.34 percent to 46.68 percent on CIFAR-10 and from 74.17 percent to 81.50 percent on FMNIST. This continuous growth highlights a critical trade-off between the per-step noise scale and the total allowable training iterations. Although a smaller σ_{small} injects less noise into the gradients and provides cleaner signals, it consumes the fixed privacy budget much faster. This rapid depletion forces the dynamic privacy accountant to terminate the training prematurely with significantly fewer epochs. Therefore, a careful balance between the noise scale and the optimization steps is essential to avoid premature termination.

Furthermore, as α enters the fine-grained search range, the model achieves peak accuracies of 49.73 percent at α equal to 0.95 on CIFAR-10 and 81.52 percent at α equal to 0.92 on FMNIST. However,

when α further approaches 1.0, the prediction accuracy begins to degrade because the perturbation degenerates toward the uniform noise scale of standard DP-SGD. This performance contrast demonstrates that the static noise paradigm of DP-SGD does not represent the optimal privacy-utility trade-off. Instead, our bimodal mechanism builds upon the baseline scale to discover a superior balance point. It maintains sufficient training iterations while leveraging the low-variance gradient advantage required for fine-grained local descent and stable convergence.

3.2.2 Impact of the Large Noise Scale σ_{large} on DP-BiSGD

Since model utility is primarily governed by the small noise scale while σ_{large} essentially acts as a privacy compensator, the prediction performance is relatively insensitive to the precise fine-tuning of this large noise scale. We thus restrict its evaluation to a coarse-grained grid where the large noise ratio factor α takes integer values from 2 to 5. To provide a comprehensive evaluation, we pair each large noise setting with two distinct configurations of α_{small}, which represents the relative ratio factor for σ_{small}. The first configuration fixes α_{small} at 0.9 for both datasets. The second configuration utilizes the optimal small noise ratios identified in our previous analysis, setting α_{small} to 0.95 for CIFAR-10 and 0.92 for FMNIST.

Table 2: Impact of σ_{large} on DP-BiSGD (%)

α	**CIFAR-10**		**FMNIST**	
	α_{small}=0.9	α_{small}=0.95	α_{small}=0.9	α_{small}=0.92
2	50.14	**49.34**	**81.00**	**81.63**
3	48.33	48.51	80.27	80.75
4	47.88	47.18	80.40	79.54
5	**50.21**	48.48	80.68	80.12

As shown in Table 2, the experimental results demonstrate remarkable stability across all tested configurations. Whether utilizing the fixed baseline or the optimal small noise ratio, the prediction accuracy exhibits minimal fluctuation despite the massive increase in the large noise scale. Specifically, as α increases from 2 to 5, the accuracy on CIFAR-10 remains within a narrow band between approximately 47 percent and 50 percent. Similarly, the performance on FMNIST maintains a steady level ranging from 79 percent to 81 percent without experiencing any catastrophic degradation.

This high robustness perfectly validates the core architectural design of our bimodal mechanism. Because the large noise steps occur with an extremely low probability governed by the exogenous Bernoulli trial, they function strictly as occasional exploratory jumps rather than dominant optimization steps. More importantly, our bypass-scaled update scheme completely isolates these extreme variance steps from the momentum buffer. Consequently, even when σ_{large} grows significantly to compensate for the privacy deficit, this architectural isolation prevents the severe noise from polluting the historical optimization trajectory. The high-probability local exploitation steps can therefore continue to leverage clean gradient signals for stable and accurate convergence.

3.2.3 Impact of the Large Noise Branch Probability p_{large} on DP-BiSGD

To assess the influence of the large noise branch probability p_{large}, we vary its value from 0.01 to 0.30 under two distinct paired configurations of the noise ratios α_{small} and α_{large} for each dataset. Specifically for CIFAR-10, we evaluate the settings where these two parameters are paired at 0.90 and 5 as well as 0.95 and 2. Similarly for FMNIST, we test the parameter pairs at 0.90 and 2 alongside 0.92 and 2.

Table 3: Impact of p_{large} on DP-BiSGD (%)

p_{large}	**CIFAR-10**		**FMNIST**	
	α_{small}=0.90 α_{large}=5	α_{small}=0.95 α_{large}=2	α_{small}=0.90 α_{large}=2	α_{small}=0.92 α_{large}=2
0.01	48.19	48.52	**82.24**	80.99
0.05	48.03	**49.25**	81.13	81.09
0.10	**50.29**	47.91	80.74	80.86
0.15	49.33	47.64	80.07	78.95
0.20	47.29	48.78	81.61	**81.62**
0.25	49.14	48.25	80.31	79.55
0.30	46.17	48.14	80.32	79.49

As shown in Table 3, the prediction accuracy exhibits moderate fluctuations as the branch probability increases. Across the tested configurations, the optimal prediction accuracies consistently emerge at relatively low probability values ranging between 0.01 and 0.20. When p_{large} further increases toward 0.30, the classification performance experiences a slight but observable degradation across both datasets.

This performance trend aligns with the dual mechanisms of our bimodal design. First, the absence of catastrophic performance drops at higher probabilities strongly validates the robustness of our bypass scaled update scheme. It insulates the momentum buffer from extreme variance injections even when these noisy steps occur more frequently. However, the slight performance degradation at larger probabilities occurs because an excessively high p_{large} disrupts the continuity of the local descent. If the model triggers the exploratory branch too often, it loses the continuous clean gradient steps required to stably converge to sharp minima. Therefore, maintaining a low branch probability ensures that the occasional exploratory jumps effectively assist the model in escaping poor local minima while preserving the steady momentum necessary for precise local exploitation.

3.3 Utility Comparison Against DP-SGD

We compare the model utility of DP-BiSGD against the standard DP-SGD baseline across four different privacy budgets ranging from 1 to 4. For these evaluations, we configure the small noise ratio at 0.90 and the large noise ratio at 2 alongside a large noise branch probability of 0.1 for both the CIFAR-10 and FMNIST datasets.

Table 4: Accuracy of DP-BiSGD and DP-SGD

Method	**FMNIST** (ϵ)				**CIFAR-10** (ϵ)			
	1	2	3	4	1	2	3	4
DP-SGD	80.52	**81.46**	84.45	84.71	46.73	52.58	53.07	59.16
DP-BiSGD	**82.24**	81.39	**84.33**	**85.10**	**50.29**	**54.56**	**55.71**	**61.27**

As shown in Table 4, DP-BiSGD consistently outperforms the uniform noise baseline across almost all evaluated regimes. More importantly, the experimental results reveal two distinct trends regarding the performance gap between the two methods. First, the classification accuracy improvement becomes significantly more pronounced when the privacy budget is small. On CIFAR-10, our method surpasses DP-SGD by a substantial margin of 3.56 percent at a strict privacy budget of 1, whereas this advantage narrows to 2.11 percent at a looser budget of 4. Second, the overall performance advantage on CIFAR-10 is substantially greater than that on FMNIST. While CIFAR-10 exhibits the aforementioned robust gains ranging from 2.11 percent to 3.56 percent, the accuracy increase on FMNIST remains comparatively modest. Specifically, the FMNIST improvement peaks at only 1.72 percent at a budget of 1 and drops to a marginal 0.39 percent at a budget of 4.

From an optimization perspective, both observational trends point to the same fundamental conclusion: the superiority of our bimodal mechanism scales directly with the magnitude of the baseline noise. A stricter privacy budget inherently necessitates a substantially larger baseline noise scale. Similarly, under identical privacy guarantees, a complex dataset such as CIFAR-10 demands much heavier noise compared to that required for FMNIST. Standard DP-SGD injects this overwhelming noise into every iteration, completely drowning the true gradient and stalling convergence. In contrast, DP-BiSGD mitigates this issue by dynamically suppressing noise variance during high-probability steps, ensuring that true gradient signals remain discernible from the injected noise. Consequently, our bimodal perturbation yields substantially greater utility gains under stricter privacy regimes.

3.4 Applicability of DP-BiSGD to Improved DP-SGD Models

To evaluate the generalizability of DP-BiSGD, we select DP-SGD-HF, a representative enhanced variant of standard DP-SGD that employs ScatterNet for handcrafted feature extraction to avoid training convolutional layers under strict privacy constraints. Specifically, we integrate our bimodal mechanism into this framework by replacing its standard uniform perturbation and privacy accounting with our counterparts. For comparison, we configure the small noise ratio, large noise ratio, and branch probability as $(0.8, 4, 0.2)$ for FMNIST

Table 5: Accuracy of DP-BiSGD-HF and DP-SGD-HF

Method	**FMNIST** (ϵ)				**CIFAR-10** (ϵ)			
	1	2	3	4	1	2	3	4
DP-SGD-HF	85.82	**88.64**	89.06	88.80	60.57	62.40	66.67	67.44
DP-BiSGD-HF	**87.15**	88.40	**89.30**	**89.50**	**61.98**	**64.83**	**67.02**	**68.45**

and $(0.8, 3, 0.3)$ for CIFAR-10, respectively. Under these configurations, we evaluate the resulting prediction performance across a spectrum of privacy budgets $\epsilon \in \{1, 2, 3, 4\}$.

As shown in Table 5, the comparative results between DP-BiSGD-HF and DP-SGD-HF exhibit a highly similar trend to our previous observations on standard DP-SGD. DP-BiSGD-HF successfully preserves the utility advantage across almost all evaluated privacy regimes. More importantly, the classification accuracy improvement follows the exact same pattern where the performance gap becomes more pronounced under stricter privacy constraints. On the CIFAR-10 dataset, the accuracy improvement achieved by our bimodal integration reaches 1.41 percent at a strict budget of 1, while this advantage narrows to 1.01 percent at a looser budget of 4. A similarly consistent enhancement pattern emerges on the FMNIST dataset. The consistency of these identical experimental trends firmly validates that our bimodal mechanism is not limited to basic frameworks but is highly applicable to advanced improved models of DP-SGD.

4 Related Work

Existing works improve the privacy-utility trade-off in DP-SGD across four dimensions: data-dependent clipping, private optimizer design, structural perturbation, and privacy accounting.

Data-Dependent Clipping. To counter signal-to-noise ratio degradation in DP-SGD, adaptive methods adjust clipping dynamically. Thakkar et al. (2019) track online quantiles to adjust thresholds, while Wei et al. (2025) propose DC-SGD using DP histograms to optimize clipping bounds. To mitigate clipping distortion, Kalarde et al. (2025) introduce Adaptive Sigmoid Clipping (AdaSig), and Gilani et al. (2025) propose GeoClip for geometry-aware transformation. However, relying on intermediate data breaches required data independence and risks privacy leakage, demanding a purely data-independent alternative.

Private Optimizer Design. Injected DP noise accumulates over time, persistently corrupting momentum buffers and moment statistics. Tang et al. (2024) show that DP noise biases the second-moment estimation in Adam and propose DP-AdamBC for explicit debiasing. Zhang et al. (2024) introduce DOPPLER to suppress high-frequency perturbation via low-pass filtering. These findings confirm that high-variance noise corrupts momentum buffers. Instead of correcting corrupted momentum states afterward, our bypass-scaled update proactively isolates large-noise steps.

Structural Perturbation. Beyond standard Gaussian perturbation, recent works explore noise distributions tailored to specific optimization structures. Jang et al. (2024) argue that Gaussian noise mismatches discrete optimizers such as signSGD (Bernstein et al. (2018)) and introduce a tailored Logistic mechanism. Kalinin & Lampert (2024) shift from independent noise to cross-iteration correlated perturbation via Banded Square Root factorization to improve convergence. However, existing works overlook data-independent exploration-exploitation trade-offs, resolved by our bimodal formulation.

Privacy Accounting. Privacy accounting has advanced from analytical approximations to exact numerical tracking. Abadi et al. (2016) introduce the Moments Accountant and Mironov (2017) formalize Rényi Differential Privacy (RDP), while Gopi et al. (2021) propose numerical Privacy Loss Distributions (PLD). Furthermore, Chua et al. (2024); Birrell et al. (2024); Shenfeld & Feldman (2025) enforce fidelity to underlying sampling schemes, and Schuchardt et al. (2024) show that explicit probability structures tighten bounds beyond generic RDP. Since standard RDP is overly loose for Gaussian Mixture Models (GMM), we develop an FFT-based dynamic GMM-PLD accountant to track exact distributions and maximize training iterations.

5 Conclusion

In this paper, we rethink the uniform perturbation paradigm of DP-SGD and propose DP-BiSGD, a strictly data-independent bimodal baseline for private deep learning. By stochastically alternating between small and large noise scales, DP-BiSGD balances local exploitation and global exploration without relying on vulnerable heuristics. Supported by a bypass-scaled momentum update and an FFT-based GMM-PLD accountant, it provides strict privacy guarantees while preventing buffer corruption. Extensive theoretical and empirical results demonstrate that DP-BiSGD achieves superior utility and attack resilience, serving as a robust drop-in baseline for private learning.

AI USE STATEMENT

In this work, we used generative AI tools solely to enhance language quality, specifically for grammar correction and vocabulary refinement. We did not use these tools to generate research ideas, design experiments, or analyze data. We take full responsibility for the scientific content and integrity of the final manuscript.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their time and effort in evaluating our manuscript.

REFERENCES

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, 2016.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SIGNSGD: compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018.

Jeremiah Birrell, Reza Ebrahimi, Rouzbeh Behnia, and Jason Pacheco. Differentially private stochastic gradient descent with fixed-size minibatches: Tighter RDP guarantees with or without replacement. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, 2022.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. How private are DP-SGD implementations? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Shib Sankar Dasgupta, Michael Boratko, and Andrew McCallum. A geometric approach to personalized recommendation with set-theoretic constraints using box embeddings. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, 2025.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, Lecture Notes in Computer Science, 2006.

Jie Fu, Qingqing Ye, Haibo Hu, Zhili Chen, Lulu Wang, Kuncan Wang, and Xun Ran. DPSUR: accelerating differentially private stochastic gradient descent using selective update and release. *Proc. VLDB Endow.*, 17(6):1200–1213, 2024.

Atefeh Gilani, Naima Tasnim, Lalitha Sankar, and Oliver Kosut. Geoclip: Geometry-aware clipping for differentially private SGD. 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*, 2025.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Jonggyu Jang, Seongjin Hwang, and Hyun Jong Yang. Rethinking DP-SGD in discrete domain: Exploring logistic distribution in the realm of signsgd. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Faeze Moradi Kalarde, Ali Bereyhi, Ben Liang, and Min Dong. Adaptive sigmoid clipping for balancing the direction-magnitude mismatch trade-off in differentially private learning. 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*, 2025.

Nikita P. Kalinin and Christoph H. Lampert. Banded square root matrix factorization for differentially private model training. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Fei Kong, Jinhao Duan, Ruipeng Ma, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Shanshan Li, Da Huang, Yu He, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. TP-MDDN: task-preferenced multi-demand-driven navigation with autonomous decision-making. 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*, 2025.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, 2017.

Khanh Nguyen, Raouf Kerkouche, Mario Fritz, and Dimosthenis Karatzas. Docmia: Document-level membership inference attacks against docvqa models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Xiong Peng, Bo Han, Feng Liu, Tongliang Liu, and Mingyuan Zhou. Pseudo-private data guided model inversion attacks. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Ossi Räisä, Joonas Jälkö, and Antti Honkela. Subsampling is not magic: Why large batch sizes work for differentially private stochastic optimisation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Jan Schuchardt, Mihail Stoian, Arthur Kosmala, and Stephan Günnemann. Unified mechanism-specific amplification by subsampling and group privacy amplification. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Moshe Shenfeld and Vitaly Feldman. 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*, 2025.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually DP-SGD (unless you apply bias correction). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, 2024.

Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019.

G. Mercy Wairimu. Design and evaluation of a generative artificial intelligence-based tool for students with learning disabilities in kenya. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, 2025.

Chengkun Wei, Weixian Li, Chen Gong, and Wenzhi Chen. DC-SGD: differentially private SGD with dynamic clipping through gradient norm distribution estimation. *IEEE Trans. Inf. Forensics Secur.*, 20:4498–4511, 2025.

Jianxin Wei, Ergute Bao, Xiaokui Xiao, and Yin Yang. DPIS: an enhanced mechanism for differentially private SGD with importance sampling. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Rui Xue, Kaiping Xue, Bin Zhu, Xinyi Luo, Tianwei Zhang, Qibin Sun, and Jun Lu. Differentially private federated learning with an adaptive noise mechanism. *IEEE Trans. Inf. Forensics Secur.*, 19:74–87, 2024.

Xinwei Zhang, Zhiqi Bu, Mingyi Hong, and Meisam Razaviyayn. DOPPLER: differentially private optimizers with low-pass filter for privacy noise reduction. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Zehan Zhu, Yan Huang, Xin Wang, Shouling Ji, and Jinming Xu. Dyn-dˆ2p: Dynamic differentially private decentralized learning with provable utility guarantee. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, 2025.

APPENDIX

A THEORETICAL ANALYSIS

A.1 PRIVACY GUARANTEES

To establish the privacy guarantees of DP-BiSGD, we decouple the analysis into two sequential parts. First, we formally prove that the continuous step-wise bimodal perturbation exhibits a tail-flattening effect, strictly bounding the worst-case single-step privacy loss. Second, we prove that our Dynamic GMM-PLD accountant rigorously preserves this bound during discrete computation.

A.1.1 PRIVACY BOUND OF SINGLE-STEP PERTURBATION

Theorem 1. ***(Asymptotic Privacy Bound of Bimodal Perturbation).*** *Given a clipping threshold C and a Poisson subsampling rate q, the asymptotic supremum of the single-step privacy loss under the continuous bimodal Gaussian mechanism ($\sigma_t \in \{\sigma_{\text{small}}, \sigma_{\text{large}}\}$) is strictly bounded by a linear function of $1/\sigma_{\text{large}}^2$.*

Proof. According to the post-processing property of differential privacy, the privacy leakage depends solely on the sequence of released private gradients. For a single step, the noise ξ_t is drawn from a Gaussian Mixture Model (GMM). Let $Q(x)$ and $P(x)$ be the null and alternative distributions (as defined in Equation 8 and Equation 9). The privacy loss random variable is governed by the log-likelihood ratio (LLR) $\mathcal{L}(x) = \ln(P(x)/Q(x))$.

For Gaussian-based mechanisms, $\mathcal{L}(x)$ monotonically increases with the observation x. Therefore, the supremum defining the worst-case privacy leakage occurs at the tail where $x \to \infty$. We expand

the base likelihood ratio:

$$\frac{Q(x-C)}{Q(x)} = \frac{(1-p_{\text{large}})\sigma_{\text{large}} \exp\left(-\frac{(x-C)^2}{2\sigma_{\text{small}}^2 C^2}\right) + p_{\text{large}}\sigma_{\text{small}} \exp\left(-\frac{(x-C)^2}{2\sigma_{\text{large}}^2 C^2}\right)}{(1-p_{\text{large}})\sigma_{\text{large}} \exp\left(-\frac{x^2}{2\sigma_{\text{small}}^2 C^2}\right) + p_{\text{large}}\sigma_{\text{small}} \exp\left(-\frac{x^2}{2\sigma_{\text{large}}^2 C^2}\right)}. \tag{19}$$

Since $\sigma_{\text{small}} < \sigma_{\text{large}}$, as $x \to \infty$, the exponential decay rate of the small-variance component is significantly faster than that of the large-variance component. Dividing both the numerator and denominator by $\exp(-x^2/2\sigma_{\text{large}}^2 C^2)$, the terms associated with σ_{small} vanish to zero. Thus, the asymptotic behavior is strictly dominated by the large-variance component:

$$\lim_{x\to\infty} \frac{Q(x-C)}{Q(x)} = \exp\left(\frac{x^2-(x-C)^2}{2\sigma_{\text{large}}^2 C^2}\right) = \exp\left(\frac{2xC-C^2}{2\sigma_{\text{large}}^2 C^2}\right) \approx \exp\left(\frac{x}{\sigma_{\text{large}}^2 C}\right). \tag{20}$$

Substituting this asymptotic ratio back into the single-step LLR (Equation 10), we obtain the supremum bound:

$$\lim_{x\to\infty} \mathcal{L}(x) \leq \frac{x}{\sigma_{\text{large}}^2 C} + \ln(q). \tag{21}$$

This mathematical derivation demonstrates the *tail-flattening effect*: the worst-case single-step privacy loss is strictly constrained by the large exploratory noise σ_{large}. □

A.1.2 Dynamic GMM-PLD accountant

Building upon the single-step bound, we establish the framework's global privacy guarantee. We first introduce a foundational lemma proving that our numerical PLD accountant strictly upper-bounds the true privacy loss via stochastic dominance.

Lemma 1. ***(Strict Upper Bound of the Discrete GMM-PLD Accountant).*** *Let $\delta_{\text{true}}(\epsilon)$ denote the true continuous privacy failure probability for a given budget ϵ. The discrete privacy failure function $H(\epsilon)$ computed by our accountant mathematically guarantees $H(\epsilon) \geq \delta_{\text{true}}(\epsilon)$ for all $\epsilon > 0$.*

Proof. Let $Y_i \sim \omega_{\text{step}}$ be independent continuous privacy loss variables at step i. For cumulative loss $Y_{1:t} = \sum_{i=1}^{t} Y_i$, the exact failure probability is:

$$\delta_{\text{true}}(\epsilon) = \mathbb{E}_{Y_{1:t}}\left[f(Y_{1:t})\right], \quad \text{where } f(y) = \max(0, 1-\exp(\epsilon - y)). \tag{22}$$

The penalty function $f(y)$ is strictly monotonically non-decreasing.

The discretization (Equation 14) acts as a right-shifting projection operator $\mathcal{M} : \mathbb{R} \to \{y_0, \ldots, y_{K-1}\}$. We define the discrete random variable $\hat{Y}_i = \mathcal{M}(Y_i)$. For any continuous observation within a bounded interval $[y_{k-1}, y_k)$, the mapping uniformly outputs the supremum y_k. Concurrently, probability mass from the unbounded right tail $[y_{K-2}, \infty)$ is entirely absorbed into the maximum terminal bin y_{K-1}. Thus, the projection strictly ensures $\hat{Y}_i \geq Y_i$ globally. Consequently, its cumulative distribution function (CDF) satisfies:

$$F_{\hat{Y}_i}(y) \leq F_{Y_i}(y), \forall y \in \mathbb{R} \implies \hat{Y}_i \succeq Y_i, \tag{23}$$

establishing strict first-order stochastic dominance.

The t-fold spatial convolution (Equation 15) yields the sum $\hat{Y}_{1:t} = \sum_{i=1}^{t} \hat{Y}_i$. Because first-order stochastic dominance is closed under the convolution of independent random variables, the spatial composition rigorously maintains the inequality:

$$\hat{Y}_{1:t} \succeq Y_{1:t}. \tag{24}$$

The computational function $H(\epsilon)$ evaluates the expectation over the discrete domain. Because $\hat{Y}_{1:t} \succeq Y_{1:t}$ and $\frac{\partial f(y)}{\partial y} \geq 0$, it strictly follows that:

$$H(\epsilon) = \mathbb{E}_{\hat{Y}_{1:t}}[f(\hat{Y}_{1:t})] \geq \mathbb{E}_{Y_{1:t}}[f(Y_{1:t})] = \delta_{\text{true}}(\epsilon). \tag{25}$$

□

Building upon the strict bounding property established in Lemma 1, we finalize the theoretical privacy guarantee of the entire DP-BiSGD framework.

Theorem 2. ***(Strict Privacy Guarantee of DP-BiSGD).*** *Let the target privacy failure probability be δ_{target} and the predefined budget limit be ϵ_{budget}. By halting training via the dynamic stopping criterion (Equation 18), the DP-BiSGD framework strictly satisfies $(\epsilon_{\text{budget}}, \delta_{\text{target}})$-Differential Privacy.*

Proof. During the optimization loop, the root-finding mechanism (Equation 17) computes the evaluated budget $\epsilon(t)$ satisfying:

$$H(\epsilon(t)) = \delta_{\text{target}}. \tag{26}$$

Applying the established inequality from Lemma 1 (Equation 25) immediately yields:

$$\delta_{\text{true}}(\epsilon(t)) \leq H(\epsilon(t)) = \delta_{\text{target}}. \tag{27}$$

The dynamic stopping condition (Equation 18) executes halting strictly where $\epsilon(t) \leq \epsilon_{\text{budget}}$. Therefore, upon termination, the framework's true privacy parameters mathematically adhere to the predefined $(\epsilon_{\text{budget}}, \delta_{\text{target}})$-DP bound. □

A.2 Utility and Convergence Guarantee

We now demonstrate that DP-BiSGD converges to a stationary point for non-convex objectives and formally analyze how the bypass-scaled update mechanism improves theoretical utility. Recall from Equation 5 that our algorithm applies weight decay directly to the descent direction via $d_t = \widetilde{g}_t + \gamma P_w \theta_t$. Mathematically, this step is equivalent to optimizing an ℓ_2-regularized objective. Specifically, we incorporate the decay penalty into a unified objective function, $\mathcal{F}(\theta) := F(\theta) + \frac{\gamma}{2}\|\theta\|^2_{P_w}$, such that the expected descent direction aligns with $\nabla\mathcal{F}(\theta) = \nabla F(\theta) + \gamma P_w \theta$. To establish the convergence bounds for $\mathcal{F}(\theta)$, we introduce two standard assumptions widely adopted in the stochastic non-convex optimization literature.

Assumption 1. ***(L-smoothness).*** *The objective function $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and L-smooth, meaning there exists a constant $L > 0$ such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$:*

$$\|\nabla\mathcal{F}(\theta_1) - \nabla\mathcal{F}(\theta_2)\| \leq L\|\theta_1 - \theta_2\|. \tag{28}$$

This mathematically implies the standard descent inequality: $\mathcal{F}(\theta_1) \leq \mathcal{F}(\theta_2) + \langle\nabla\mathcal{F}(\theta_2), \theta_1 - \theta_2\rangle + \frac{L}{2}\|\theta_1 - \theta_2\|^2$.

Assumption 2. ***(Bounded Variance of Stochastic Gradients).*** *The stochastic gradient calculated from the mini-batch $\mathcal{B}$ is an unbiased estimator of the full-batch true gradient, and its variance is strictly bounded by a constant σ_g^2. Namely, for any $\theta \in \mathbb{R}^d$:*

$$\mathbb{E}_{\mathcal{B}}[\|\nabla\mathcal{F}(\theta; \mathcal{B}) - \nabla\mathcal{F}(\theta)\|^2] \leq \sigma_g^2. \tag{29}$$

Remark 3. *Both assumptions are standard and mild. Assumption 1 ensures Lipschitz smoothness of the objective, which is well-justified for neural networks equipped with smooth activations (e.g., GELU or Sigmoid) and weight decay. Moreover, under differential privacy, Assumption 2 holds automatically. Because per-sample gradients are explicitly bounded by the ℓ_2 clipping threshold C before aggregation, the mini-batch gradient variance remains strictly bounded, thereby preventing gradient explosion and heavy-tailed behavior.*

To mathematically isolate and evaluate the convergence impact of the bimodal mechanism, we analyze the generalized expected gradient flow corresponding to the effective update step. Specifically, under the Bernoulli random variable $Z_t \sim \text{Bernoulli}(p_{\text{large}})$, we establish three core structural quantities: (i) the expected scaling $\bar{\lambda} = \mathbb{E}[\Lambda_t] = (1 - p_{\text{large}}) + p_{\text{large}}\lambda_{\text{large}}$, (ii) the expected squared scaling $\overline{\lambda^2} = \mathbb{E}[\Lambda_t^2] = (1 - p_{\text{large}}) + p_{\text{large}}\lambda_{\text{large}}^2$, and (iii) the resulting effective DP variance $\Psi_{\text{DP}}^2 = (1 - p_{\text{large}})\sigma_{\text{small}}^2 + p_{\text{large}}\lambda_{\text{large}}^2\sigma_{\text{large}}^2$.

Theorem 3. ***(Convergence Bound of DP-BiSGD).*** *Under Assumptions 1 and 2, let the learning rate satisfy $\eta \leq \frac{1}{L}$. The expected average squared gradient norm of DP-BiSGD over T iterations is rigorously bounded by:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{F}(\theta_t)\|^2] \leq \frac{2[\mathcal{F}(\theta_0) - \mathcal{F}^*]}{\eta\bar{\lambda}T} + \frac{\eta L\overline{\lambda^2}}{\bar{\lambda}}\sigma_g^2 + \frac{\eta L}{\bar{\lambda}}\frac{dC^2}{b^2}\Psi_{\text{DP}}^2, \tag{30}$$

where $\mathcal{F}^*$ is the global infimum, d is the model dimension, and b is the expected batch size.

Proof. Based on the algorithmic design of the bypass-scaled mechanism, the effective parameter update at iteration t can be compactly generalized as $\theta_{t+1} = \theta_t - \eta\Lambda_t\widetilde{g}_t$, where the perturbed gradient is $\widetilde{g}_t = \nabla\mathcal{F}(\theta_t;\mathcal{B}_t) + \frac{1}{b}\xi_t$. To analytically bound the convergence of this trajectory, we must decompose the expected descent and variance across the three mutually independent sources of randomness introduced at each step: the mini-batch data sampling $\mathcal{B}_t$, the injected Gaussian DP noise ξ_t, and the bimodal branch indicator Z_t.

By invoking the L-smoothness condition (Assumption 1) on the effective parameter update, we establish the fundamental descent inequality:

$$\mathcal{F}(\theta_{t+1}) \leq \mathcal{F}(\theta_t) - \eta\Lambda_t\langle\nabla\mathcal{F}(\theta_t),\widetilde{g}_t\rangle + \frac{L\eta^2\Lambda_t^2}{2}\|\widetilde{g}_t\|^2. \tag{31}$$

To simplify this inequality, we evaluate the expectations of both its first-order inner product term and its second-order variance term. We first focus on the first-order term in Equation 31, which dictates the expected descent. Because the DP noise ξ_t is zero-mean and the stochastic gradient is an unbiased estimator of the true gradient (Assumption 2), the expectation over the data and noise simplifies cleanly to $\mathbb{E}_{\mathcal{B}_t,\xi_t}[\widetilde{g}_t] = \nabla\mathcal{F}(\theta_t)$. Furthermore, since the scaling factor Λ_t is governed solely by the exogenous Bernoulli trial Z_t, it acts independently of the gradient:

$$\mathbb{E}\left[-\eta\Lambda_t\langle\nabla\mathcal{F}(\theta_t),\widetilde{g}_t\rangle\right] = -\eta\mathbb{E}_{Z_t}[\Lambda_t]\|\nabla\mathcal{F}(\theta_t)\|^2 = -\eta\bar{\lambda}\|\nabla\mathcal{F}(\theta_t)\|^2. \tag{32}$$

Having established the expected descent direction, we turn our attention to the second-order term $\mathbb{E}[\Lambda_t^2\|\widetilde{g}_t\|^2]$, which encapsulates the optimization variance. Due to the strict independence between the stochastic gradient and the DP noise, their cross-term expectation vanishes. This allows us to directly decouple the base variance into two independent components:

$$\mathbb{E}_{\mathcal{B}_t,\xi_t}[\|\widetilde{g}_t\|^2] = \mathbb{E}_{\mathcal{B}_t}[\|\nabla\mathcal{F}(\theta_t;\mathcal{B}_t)\|^2] + \frac{1}{b^2}\mathbb{E}_{\xi_t}[\|\xi_t\|^2]. \tag{33}$$

Applying the bounded variance definition (Assumption 2) to the first component, we bound the data-induced gradient term as $\mathbb{E}[\|\nabla\mathcal{F}(\theta_t;\mathcal{B}_t)\|^2] \leq \|\nabla\mathcal{F}(\theta_t)\|^2 + \sigma_g^2$. Multiplying the entire expression by the independent bimodal scaling factor Λ_t^2 yields:

$$\mathbb{E}[\|\Lambda_t\widetilde{g}_t\|^2] \leq \mathbb{E}_{Z_t}[\Lambda_t^2]\left(\|\nabla\mathcal{F}(\theta_t)\|^2 + \sigma_g^2\right) + \frac{1}{b^2}\mathbb{E}_{Z_t,\xi_t}\left[\Lambda_t^2\|\xi_t\|^2\right]. \tag{34}$$

Thus, Equation 34 isolates the inherent stochastic variance from the injected DP noise. To resolve the remaining DP noise term under our step-wise mechanism, we condition it on the branch indicator. The DP noise is sampled as $\xi_t \sim \mathcal{N}(0,\sigma_t^2C^2I_{d\times d})$, meaning that conditioned on Z_t, the expected squared norm is exactly $dC^2\sigma_t^2$. Taking the full expectation over the Bernoulli trial explicitly evaluates the effective variance introduced by the bimodal injection:

$$\begin{aligned}\mathbb{E}_{Z_t,\xi_t}\left[\Lambda_t^2\|\xi_t\|^2\right] &= dC^2\mathbb{E}_{Z_t}\left[\Lambda_t^2\sigma_t^2\right] \\ &= dC^2\Big((1-p_{\text{large}})(1)^2\sigma_{\text{small}}^2 + p_{\text{large}}(\lambda_{\text{large}})^2\sigma_{\text{large}}^2\Big) \\ &= dC^2\Psi_{\text{DP}}^2.\end{aligned} \tag{35}$$

Now we synthesize these intermediate results by substituting Equation 32, Equation 34, and Equation 35 back into the master inequality (Equation 31):

$$\mathbb{E}[\mathcal{F}(\theta_{t+1})] \leq \mathbb{E}[\mathcal{F}(\theta_t)] - \eta\bar{\lambda}\|\nabla\mathcal{F}(\theta_t)\|^2 + \frac{L\eta^2}{2}\overline{\lambda^2}\left(\|\nabla\mathcal{F}(\theta_t)\|^2 + \sigma_g^2\right) + \frac{L\eta^2}{2}\frac{dC^2}{b^2}\Psi_{\text{DP}}^2. \tag{36}$$

Rearranging this consolidated expression to isolate the target gradient norm, we obtain:

$$\left(\eta\bar{\lambda} - \frac{L\eta^2\overline{\lambda^2}}{2}\right)\|\nabla\mathcal{F}(\theta_t)\|^2 \leq \mathbb{E}[\mathcal{F}(\theta_t)] - \mathbb{E}[\mathcal{F}(\theta_{t+1})] + \frac{L\eta^2\overline{\lambda^2}}{2}\sigma_g^2 + \frac{L\eta^2}{2}\frac{dC^2}{b^2}\Psi_{\text{DP}}^2. \tag{37}$$

Before proceeding to the final accumulation, we must ensure that the leading coefficient remains valid. Because the dampening scaling strictly falls within $\lambda_{\text{large}} \in (0,1]$, it mathematically guarantees $\overline{\lambda^2} \leq \bar{\lambda}$. Thus, assuming a standard learning rate bounded by $\eta \leq \frac{1}{L}$, the leading coefficient remains strictly positive:

$$\eta\bar{\lambda} - \frac{L\eta^2\overline{\lambda^2}}{2} \geq \eta\bar{\lambda} - \frac{\eta\bar{\lambda}}{2} = \frac{\eta\bar{\lambda}}{2} > 0. \tag{38}$$

Since the leading coefficient is positive, we divide the entire inequality by $\frac{\eta\bar{\lambda}}{2}$ to yield the single-step descent bound:

$$\|\nabla\mathcal{F}(\theta_t)\|^2 \leq \frac{2}{\eta\bar{\lambda}}\Big(\mathbb{E}[\mathcal{F}(\theta_t)] - \mathbb{E}[\mathcal{F}(\theta_{t+1})]\Big) + \frac{\eta L\overline{\lambda^2}}{\bar{\lambda}}\sigma_g^2 + \frac{\eta L}{\bar{\lambda}}\frac{dC^2}{b^2}\Psi_{\mathrm{DP}}^2. \tag{39}$$

Finally, by summing this inequality over the complete trajectory from $t = 0$ to $T-1$, the consecutive objective function differences collapse into a telescoping sum, which is upper-bounded by the initial optimality gap $\mathcal{F}(\theta_0) - \mathcal{F}^*$. Dividing this aggregated summation by the total iterations T rigorously extracts the final expected convergence bound in Equation 30. □

A.3 Complexity Analysis

Finally, we analyze the time and space complexity of DP-BiSGD. Let d denote the total number of model parameters, b the expected batch size, and K the discrete grid resolution utilized by the privacy accountant.

Time Complexity. The computational runtime of a single training iteration can be analytically decomposed into four sequential stages:

1. *Gradient Computation and Clipping:* Backpropagating the loss for b individual samples to obtain the per-sample gradients $g_t(x_i, y_i) \in \mathbb{R}^d$ operates in $\mathcal{O}(b \cdot d)$ time. Subsequently, computing the ℓ_2-norm and scaling each vector to satisfy the clipping threshold C requires $\mathcal{O}(d)$ arithmetic operations per instance. Aggregating these clipped gradients yields a total computational cost of $\mathcal{O}(b \cdot d)$.
2. *Bimodal Perturbation:* Constructing the private gradient entails generating a pseudo-random Gaussian noise vector $\xi_t \sim \mathcal{N}(0, \sigma_t^2 C^2 I_{d\times d})$ and computing the normalized aggregation $\widetilde{g}_t = \frac{1}{b}(G_t + \xi_t)$. Because the injected DP noise is strictly isotropic (i.e., its covariance matrix is a scaled identity matrix), the noise generation, addition, and scaling are purely element-wise operations. Consequently, this entire perturbation step executes in $\mathcal{O}(d)$ time.
3. *Bypass-Scaled Update:* The state evolution requires first evaluating the Bernoulli branch Z_t in $\mathcal{O}(1)$ time. Whether the branch activates the standard update (Equation 6) or the momentum-freeze bypass (Equation 7), the ensuing parameter updates (e.g., $m_{t+1} = \mu m_t + d_t$ and $\theta_{t+1} = \theta_t - \eta\Lambda_t\widetilde{g}_t$) are formulated entirely as linear combinations of vectors in $\mathbb{R}^d$. Evaluating these element-wise additions and scalar multiplications demands $\mathcal{O}(d)$ time.
4. *Dynamic Accounting:* Discretizing the single-step distribution onto the grid takes $\mathcal{O}(K)$ time. The frequency-domain composition executes a forward and inverse Fast Fourier Transform, taking $\mathcal{O}(K \log K)$ operations, alongside an $\mathcal{O}(K)$ element-wise exponentiation. Finally, extracting the budget via binary search over the CDF takes $\mathcal{O}(\log K)$. Thus, the tracking module is rigorously bounded by $\mathcal{O}(K \log K)$.

Therefore, the total time complexity per iteration is $\mathcal{O}(b \cdot d + K \log K)$. Crucially, because K is a static hyperparameter (typically $K = 10^4$), the term $\mathcal{O}(K \log K)$ is computationally constant with respect to the model dimension.

Space Complexity. The memory footprint during training is similarly partitioned into two domains:

1. *Parameter Space:* The framework must persistently allocate memory for the model weights θ_t, the isolated momentum buffer m_t, and the active gradient vector $\widetilde{g}_t$. Each of these dynamic states constitutes a dense vector in $\mathbb{R}^d$, bounding the base algorithmic memory requirement strictly to $\mathcal{O}(d)$.
2. *Accountant Space:* The dynamic PLD accountant evaluates the composition entirely in 1D scalar space. It maintains only the discrete probability vectors $\mathbf{u}, \mathbf{v}_t \in \mathbb{R}^K$. Regardless of how massive the model dimension d scales, these vectors require an auxiliary space of exactly $\mathcal{O}(K)$.

The total space complexity is thus $\mathcal{O}(d + K)$.

B Resilience Against Membership Inference Attacks

To verify that DP-BiSGD maintains comparable defensive capabilities against membership inference attacks (MIAs) as the standard DP-SGD, we conduct an empirical evaluation using the strong Offline Likelihood Ratio Attack (LiRA) Carlini et al. (2022).

Table 6: Empirical Resilience of DP-BiSGD Against Membership Inference Attacks (%)

ϵ	Method	AUC	TPR@1% FPR	TPR@0.1% FPR	Attack Advantage
1	DP-SGD	0.493	0.95	0.00	1.20
	DP-BiSGD	0.503	1.25	0.15	1.90
2	DP-SGD	0.494	1.15	0.05	1.40
	DP-BiSGD	0.500	1.40	0.10	3.00
4	DP-SGD	0.495	0.50	0.00	1.65
	DP-BiSGD	0.501	0.45	0.10	1.30

Attack Model. LiRA assumes a realistic black-box threat model where the adversary lacks access to model parameters, gradients, or training status. Instead, the attacker can query the target model for prediction scores on candidate samples. Assuming access to auxiliary data from the same distribution, the attacker trains reference models to estimate the prediction statistic distribution of candidate samples. The more a candidate's output deviates from this distribution, the higher the probability it belongs to the target model's training set.

Evaluation Metrics. We evaluate the empirical privacy of DP-BiSGD and DP-SGD against MIAs using three standard metrics: (1) *Area Under the ROC Curve (AUC)* evaluates the randomness of the attack model's ability to guess whether a sample is in the training set, where a value of 0.5 indicates random guessing; (2) *True Positive Rate at β% False Positive Rate (TPR@β% FPR)* measures the proportion of correctly identified training members when the false positive rate is strictly limited to β%, where a value closer to β indicates lower privacy leakage risk in low-false-positive scenarios; (3) *Attack Advantage* represents the maximum distinguishing gain the attacker can achieve over random guessing, where values approaching 0 indicate stronger privacy preservation.

Experimental Setup. We conduct inference attacks on the CIFAR-10 dataset across privacy budgets $\epsilon \in \{1, 2, 4\}$. For both methods, we set the sampling rate to 0.1 (corresponding to a batch size of 256) and the gradient clipping threshold to 1. For DP-SGD, we set the noise multiplier σ_{base} to 4.18, 2.30, and 1.37 for $\epsilon = 1, 2, 4$, respectively. For DP-BiSGD, we fix $p_{\text{large}} = 0.05$, $\sigma_{\text{large}} = 20$, and $\lambda = 0.1$, while setting σ_{small} to 3.79, 2.11, and 1.28 to align with DP-SGD's noise scales. Finally, following the standard LiRA setup, we train one target model alongside $k = 8$ reference models for attack evaluation.

Results and Analysis. As shown in Table 6, DP-BiSGD exhibits robust resilience virtually identical to DP-SGD. Across all evaluated privacy budgets, the AUC values for both methods consistently hover around 0.50, indicating that the attack performance is degraded to random guessing. Furthermore, the TPRs at strictly limited FPRs (1% and 0.1%) remain extremely low, closely aligning with the theoretical baselines. The attack advantages are also strictly bounded below 3.00%. These empirical results thoroughly confirm that altering the single-noise mechanism to a dual-noise scale mode in DP-BiSGD does not compromise its practical defense capabilities against advanced MIAs.