DIFFERENTIALLY PRIVATE NON-CONVEX DISTRIBUTIONALLY ROBUST OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world deployments routinely face distribution shifts, group imbalances, and adversarial perturbations, under which the traditional Empirical Risk Minimization (ERM) framework can degrade severely. Distributionally Robust Optimization (DRO) addresses this issue by optimizing the worst-case expected loss over an uncertainty set of distributions, offering a principled approach to robustness. Meanwhile, as training data in DRO always involves sensitive information, safeguarding it against leakage under Differential Privacy (DP) is essential. In contrast to classical DP-ERM, DP-DRO has received much less attention due to its minimax optimization structure with uncertainty constraints. To bridge the gap, we provide a comprehensive study of DP-(finite-sum)-DRO with ψ-divergence and non-convex loss. First, we study DRO with general ψ-divergence by reformulating it as a minimization problem and developing a novel (ε, δ)-DP optimization method, called DP Double-Spider, tailored to this structure. Under mild assumptions, we show that it achieves a utility bound of $\mathcal{O}(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$ in terms of the gradient norm, where n denotes the data size and d denotes the model dimension. We further improve the utility rate for specific divergences. In particular, for DP-DRO with KL-divergence, by transforming the problem into a compositional finite-sum optimization problem, we develop a DP Recursive-Spider method and show that it achieves a utility bound of $\mathcal{O}((\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$, matching the best-known result for non-convex DP-ERM. Experimentally, we demonstrate that our proposed methods outperform existing approaches for DP minimax optimization.

1 INTRODUCTION

With the rapid expansion of machine learning in real-world applications, the demand for robust model training has become increasingly critical. Conventional Empirical Risk Minimization (ERM) aims to minimize the expected loss under the empirical distribution of the training data, with the hope that the resulting model generalizes well to unseen test data. However, in practice, ERM often struggles when the training and test distributions do not align. Such mismatches are pervasive, including domain shifts in transfer and domain adaptation tasks (Blitzer et al., 2006), imbalanced training datasets where fairness for underrepresented groups is essential (Sagawa et al., 2019), and adversarial perturbations that deliberately mislead deployed models (Goodfellow et al., 2014; Sinha et al., 2017; Madry et al., 2017). In these settings, models trained purely with ERM typically exhibit severe performance degradation, limiting their reliability in safety-critical or fairness-sensitive applications. To overcome these limitations, Distributionally Robust Optimization (DRO) has emerged as a powerful alternative (Ben-Tal et al., 2013; Shapiro, 2017; Rahimian & Mehrotra, 2019). By optimizing performance against the worst-case distribution within a specified uncertainty set, DRO provides models with resilience to distribution shifts, fairness violations, and adversarial attacks. This worst-case perspective has made DRO a central tool in building robust learning systems.

Meanwhile, as the training of DRO often involves private data, safeguarding sensitive information has become a critical priority. Differential privacy (DP) (Dwork et al., 2014) now occupies a central position in data analysis, driven by the growing demand for principled protection guaranties. The foundational work Dwork et al. (2006) established the modern framework for DP, and its incorporation into empirical risk minimization (ERM) has progressed substantially. Early research focused primarily

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.

on convex optimization, delivering strong privacy guarantees alongside efficient learning, for example, (Bassily et al., 2014; 2019; 2021; Choquette-Choo et al., 2024; Feldman et al., 2020; Su et al., 2023; Tao et al., 2022; Wang et al., 2020a;b; 2017). More recently, attention has shifted toward non-convex settings. Notable advances such as (Wang et al., 2019; Zhou et al., 2020; Xiao et al., 2023; Arora et al., 2023) tighten error bounds for non-convex optimization while balancing privacy and utility in stochastic gradient methods. However, relatively few studies have focused on (finite-sum) DRO, which involves minimax optimization over an uncertainty set.

Recently, several works have begun to investigate DP-DRO Zhou & Bassily (2024); Selvi et al. (2025). However, these studies either consider only a special classes of DRO (such as group DRO Zhou & Bassily (2024)) or propose methods that are not scalable to large datasets (Selvi et al., 2025). DRO with ψ-divergence, a widely used framework in machine learning, has not yet been systematically studied under DP. This gap motivates the present work. Although DRO can be formulated as a minimax problem, it differs fundamentally from classical minimax optimization. As a result, recent advances in DP minimax optimization cannot be directly applied to analyze or solve DP-DRO due to their assumptions on the loss function or the absence of uncertainty sets (see Appendix A for details).

To fill this gap, we provide the first systematic study of DP-(finite-sum) DRO with ψ-divergence and non-convex loss. First, by establishing a primal–dual equivalence for distributionally robust optimization, we reduce the original minimax problem to an unconstrained minimization problem over model parameters and a dual variable. Building on this reformulation, we develop a novel private optimization method called DP Double-Spider, which is a differentially private and non-trivial extension of SPIDER (Fang et al., 2018) from ERM to DRO. DP Double-Spider alternates updates of the primal and dual variables using variance-reduced stochastic gradients perturbed with calibrated Gaussian noises, which outperforms classical DP-SGD-based methods such as DP-SGDA. Theoretically, under mild assumptions, we show that the utility of the resulting private model, measured by the gradient norm, is upper bounded by $\mathcal{O}(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$, where n is the data size and d is the model dimension.

We then consider improving the utility bound of the above method. Specifically, we focus on DRO with KL divergence. Specifically, rather than introducing dual variables, the problem can be rewritten as a compositional minimization problem. Based on this, we develop our follow our idea DP-Double Spider and propose DP Recursive-Spider. Theoretically, we show that the utility rate can be improved to $\mathcal{O}\left((\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3}\right)$, which matches the best-known result in DP-ERM with non-convex loss Arora et al. (2023). It is notable that this is also the first result on DP compositional optimization.

Our contribution can be summarized as below:

1. **A generic DP-DRO framework**: For DRO with general ψ-divergence, based on its one primal-dual equivalent formulation, we propose the Differentially Private Double SPIDER algorithm to deal with both the primal and dual variables simultaneously. We claim that the algorithm is (ε, δ)-DP and the utility of our private model is $\mathcal{O}(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$.
2. **A DP method for DRO with KL-divergence**: For DRO with KL-divergence, we further improve its utility. Specifically, we propose DP Recursive-Spider and show its private model achieves an utility bound of $\mathcal{O}((\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$.
3. We implement our algorithms on the constructed imbalanced dataset CIFAR10-ST, MNIST-ST and the inherently imbalanced dataset CelebA and Fashion-MNIST. Our results demonstrate that our algorithms perform better than the previous baselines across distinct datasets and privacy budgets, providing empirical support for our theoretical analysis.

2 Related Works

For further detailed related works on DRO and DP Minimax Optimization, see Appendix A.

DP-DRO As we mentioned, currently there is no work on DP for finite-sum DRO with general ψ-divergence, which has been widely adopted in the deep learning community. Recent work on

DP-DRO only considered a specific class. In detail, Selvi et al. develops a class of mechanisms that enjoy non-asymptotic and unconditional optimality guarantees. (Zhou & Bassily, 2024) investigates worst-group risk minimization (i.e., group DRO) under differential privacy, proposing algorithms that minimize the maximum risk across multiple sub-populations while preserving privacy. It attempts to privatize the given non-private algorithm for convex loss functions. They reach near-optimal guarantees, introducing simple private algorithms. However, in practice, their methods cannot be applied to our continuous uncertainty set.

However, there are two limitations that make these work unsuitable for our DRO problem. First, previous work only considers either convex-(strongly) concave or non-convex-strongly-concave case, while for DRO with ψ-divergence, it is a non-convex-concave problem. Thus, previous methods cannot be used for analyzing our problem. Second, directly using previous methods to our problem requires performing the project onto the simplex in each update, which is prohibitively costly. In our work, we address these limitations and we further improve the utility bound for DRO compared to previous methods.

3 PRELIMINARIES

3.1 DIFFERENTIAL PRIVACY

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** Given a data universe $\mathcal{X}$, we say that two datasets $S, S' \subseteq \mathcal{X}$ are neighbors if they differ by only one entry, which is denoted as $S \sim S'$. A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private (DP) if for all neighboring datasets S, S' and for all events E in the output space of $\mathcal{A}$, the following holds

$$\mathbb{P}(\mathcal{A}(S) \in E) \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(S') \in E) + \delta.$$

If $\delta = 0$, we call algorithm $\mathcal{A}$ ε-DP.

Definition 2. Given a function $q : \mathcal{Z} \to \mathbb{R}^d$, we say q has $\Delta_2(q)$ ℓ_2-sensitivity if for any neighboring datasets D, D' we have $\|q(D) - q(D')\| \leqslant \Delta_2(q)$.

Definition 3. Given any function $q : \mathcal{Z} \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{N}\left(0, \frac{\Delta_2^2(q)\log(1.25/\delta)}{\varepsilon^2}\mathbf{I}_d\right)$. Gaussian mechanism preserves (ε, δ)-DP for $0 < \varepsilon, \delta \leq 1$.

3.2 DISTRIBUTIONALLY ROBUST OPTIMIZATION

As we mentioned earlier, unlike Empirical Risk Minimization, there are different formulations of DRO for various applications. In this paper, we will focus on the fine-sum version of DRO with an uncertainty set, which is the most commonly used model in deep learning to handle distribution shift (Levy et al., 2020; Jin et al., 2021; Qi et al., 2021; Sinha et al., 2017). Given an underlying data distribution P_0 and a loss ℓ, DRO aims to find a model $x \in \mathbb{R}^d$ that minimizes the loss over all data that is close to the distribution P_0:

$$\min_{x\in\mathbb{R}^d} \max_{P\in\mathcal{U}(P_0)} \mathbb{E}_{\xi\sim P}[\ell(x;\xi)],$$

where $\mathcal{U}(P_0)$ is an uncertainty set for P_0. Specifically, when $\mathcal{U}(P_0) = \{P : D_\psi(P, P_0) \leq \rho\}$ for some divergence D_ψ and radius ρ, we refer to the problem as DRO with divergence D_ψ. It includes the χ^2-divergence, KL-divergence, and the Cressie–Read family, defined as

$$D_\psi(Q\|P_0) = \sum_{x\in\mathcal{X}} p_0(x)\psi\left(\frac{q(x)}{p_0(x)}\right), \tag{1}$$

where ψ is a non-negative convex function satisfying $\psi(1) = 0$ and $\psi(t) = +\infty$ for all $t < 0$.

In this paper, we consider the empirical form of DRO. Specifically, given a training data $S = \{\xi_i\}_{i=1}^n$, an equivalent form of DRO with ψ-divergence can be formulated as the following constrained minimax optimization problem:

$$\min_{x\in\mathbb{R}^d} \mathcal{F}(x) := \max_{p\in\Delta_n} \sum_{i=1}^n p_i\ell(x;\xi_i) - \lambda_0 D_\psi(p, \frac{1}{n}\mathbb{I}), \tag{2}$$

where $\frac{1}{n}\mathbb{I}$ is the empirical distribution and $\Delta_n = \{\mathbf{p} \in \mathbb{R}^n : \sum_{i=1}^n p_i = 1, p_i \geqslant 0\}$ denotes a n-dimensional simplex. The hyperparameter λ_0 is pre-specified and remains fixed during training.

Table 1: Common divergences with L-smooth conjugates.

Divergence	$\psi(t)$	$\psi^*(t)$
χ^2	$\frac{1}{2}(t-1)^2$	$-1+\frac{1}{4}(t+2)^2$
KL-regularized CVaR	$\mathbb{I}_{[0,\alpha^{-1})}+t\log(t)-t+1, \alpha\in(0,1)$	$\min(e^t, \alpha^{-1}(1+t+\log(\alpha)))-1$
Cressie-Read	$\frac{t^k-tk+k-1}{k(k-1)}, k\in\mathbb{R}$	$\frac{1}{k}\left(((k-1)t+1)_+^{\frac{k}{k-1}}-1\right)$

Definition 4 (DP-DRO)**.** In the problem of DP-DRO, given a training data S, we aim to develop an (ε,δ)-DP algorithm $\mathcal{A}$ for solving problem equation 2. Moreover, we need to make the utility of the private model $\mathbb{E}[\|\nabla\mathcal{F}(\mathcal{A}(S))\|]$ to be as small as possible, where the expectation takes over the randomness of the algorithm.

Next, we will impose several assumptions on the problem for our theoretical utility analysis.

Definition 5. A function $f:\mathbb{R}^d\to\mathbb{R}$ is G-Lipschitz continuous if $\forall x,y\in\mathbb{R}^d$, $|f(w_1;x)-f(w_2;x)|\leqslant G\|w_1-w_2\|$, where $G>0$ is some finite constant.

Definition 6. A differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth if $\forall x,y\in\mathbb{R}^d$, we have $\|\nabla f(x)-\nabla f(y)\|\leqslant L\|x-y\|$, where $L>0$ is some finite constant.

It is noteworthy that Lipschitz and L-smoothness are commonly adopted in optimization research (Zhang et al., 2025a; Qi et al., 2021; 2022; Zhou & Bassily, 2024; Zhang et al., 2025b). However, as we will mention later, DRO with ψ-divergence may not satisfy smoothness but a more general property called generalized smoothness.

Definition 7 (Generalized (L_0,L_1)-smooth)**.** A differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is genenralized (L_0,L_1)-smooth if for any $x,y\in\mathbb{R}^d$, we have that $\|\nabla_x f(x)-\nabla_x f(y)\|\leqslant(L_0+L_1\|\nabla_x f(x)\|)\|x-y\|$, where $L_0,L_1>0$ are some finite constants.

Definition 8. A function ψ^* called the conjugate function, given ψ, is defined as $\psi^*(t)=\sup_{a\in\mathbb{R}}\{ta-\psi(a)\}$.

Assumptions. Generally, we focus on the non-convex and smooth loss function ℓ.

Assumption 1. For any sample $\xi\in\mathcal{D}$, the loss function $\ell(x,\xi)$ is G-Lipschitz continuous and L-smooth in x.

The smoothness of the primal function alone is insufficient for guarantee the smoothness of its conjugate. To support our analysis, we therefore impose the following additional assumption, which is commonly applied for a wide range of ψ-divergence, where instances can be found in Table 1.

Assumption 2. The conjugate function ψ^* of ψ is M-smooth.

Assumption 3. Assume that the domain of the conjugate function ψ^* is bounded by a constant $S_0>0$. Also, ψ^* remains non-negative on its domain.

4 DP DOUBLE-SPIDER

In this section, we will consider the general ψ-divergence. Due to the divergence regularization and the constrained simplex, it is hard to optimize efficiently. Thus, a popular approach is to investigate its dual formulation. By strong duality (Levy et al., 2020; Shapiro, 2017), we can reformulate the objective function equation 2 as

$$\min_x\Psi(x)=\inf_{\eta\in\mathbb{R}}\hat{\mathcal{L}}(x,\eta):=\frac{\lambda}{n}\sum_{i=1}^n\psi^*\left(\frac{\ell(x;\xi_i)-\eta}{\lambda}\right)+\eta, \tag{3}$$

where $\eta\in\mathbb{R}$ is a dual variable. To simplify the discussion below, we define $\mathcal{L}(x,\eta)=\hat{\mathcal{L}}(x,G\eta)$. Note that there are different forms of the dual problem; each will lead to different methods. The form of duality adopted in this work differs from the classical dual-based approach (Rafique et al., 2022).

Under Assumptions 1 and 2, the function $\Psi(x)$ is differentiable. With this reformulation, Jin et al. (2021) showed that $\|\nabla_{x,\eta}\mathcal{L}(x,\eta)\| \leqslant \alpha/\sqrt{2}$ directly implies $\|\nabla\mathcal{F}(x)\| \leqslant \alpha$. Thus, our goal is to find private x, η to make $\|\nabla_{x,\eta}\mathcal{L}(x,\eta)\|$ as small as possible.

Note that problem equation 3 is a finite-sum minimization problem. Specifically, let $z = (x, \eta)$ and $\mathcal{L}(x, \eta; \mathcal{S}) := \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|}\mathcal{L}(x, \eta; \xi_i)$ with $\mathcal{L}(x, \eta; \xi_i) = \lambda\psi^*(\frac{\ell(x;\xi_i)-\eta}{\lambda}) + \eta$. Thus, one direct way is directly adopting the previous method for DP-ERM, such as DP-SGD (Bassily et al., 2014; Abadi et al., 2016b) on the single composite variable $z = (x, \eta)$. In particular, we may use DP-SPIDER (Arora et al., 2023), which achieves the best-known rate for DP-ERM with non-convex loss.

However, due to the specific structure of Ψ in equation 3, directly using these methods may lead to sub-optimal performance. The main reason is that, under Assumptions 1 and 2, (Zhang et al., 2025a) showed that $\mathcal{L}(x, \eta; \xi_i)$ is (L_0, L_1)-smooth in x but L_2-smooth in η, with $L_0 = G + \frac{G^2M}{\lambda}$, $L_1 = L/G$ and $L_2 = \frac{G^2M}{\lambda}$. Due to different smooth properties, training x and η uniformly will lead to worse noise added in each iteration. In particular, for a fixed x, the function $\mathcal{L}(x, \cdot)$ is smooth, which facilitates a more accurate estimation of the gradient $\nabla_\eta\mathcal{L}$. Treating z as a single variable, however, fails to exploit this property. By instead handling the parameter x and the dual variable η separately, we can more precisely characterize and control the behavior of the gradient estimators.

Motivated by this, we proposed our DP Double-SPIDER(see Algorithm 2 for details), a private extension of SPIDER for DRO (Zhang et al., 2025a), which is a variance-reduction optimization method based on SPIDER Fang et al. (2018) that achieves state-of-the-art (SOTA) convergence in the non-private case. Specifically, each variable is updated using a SPIDER-style variance-reduced gradient estimator perturbed with Gaussian noise. Concretely, in Line 4, for η, we perform a large-batch refresh every q iterations with batch size N_1 to form a low-variance estimate, and in Line 6 for the remaining $q - 1$ steps, we use incremental corrections with mini-batch size N_2 via gradient differences, adding an independent Gaussian noise term at each step for privacy. After updating η, we repeat the same refresh–increment pattern for x, using batch sizes N_3 and N_4, again injecting calibrated Gaussian noise. This alternating procedure—refresh once, then correct $q - 1$ times—controls stochastic variance while enforcing privacy, yielding coordinated updates of η and x every q iterations.

Note that there are two differences compared with the non-private one: (1) As we mentioned, due to the different smooth properties of x and η, we need the variance-reduced gradient estimator for $\nabla_\eta\mathcal{L}$ and $\nabla_x\mathcal{L}$ individually, i.e., it is a double version of SPIDER. (2) Under Assumptions 1 and 2, we can show that the sensitivity of the gradient difference in the variance-reduced gradient estimator, such as $\nabla_x\mathcal{L}(x_t, \eta_{t+1}; \mathcal{B}) - \nabla_x\mathcal{L}(x_{t-1}, \eta_t; \mathcal{B})$, is $O(\max\{\|x_t - x_{t-1}\|, \|\eta_{t+1} - \eta_t\|\})$. Thus, the Gaussian noise added to ensure DP also depends on such differences, making the analysis complicated.

The following results provide privacy and utility guarantees of DP Double-SPIDER. For the full versions and their proofs, we refer readers to Appendix D.

Theorem 1. *For any $\varepsilon > 0$ and $\delta \in (0, 1)$, let $\sigma_1 = \mathcal{O}(\frac{C_1\sqrt{T\log(1/\delta)}}{n\sqrt{q}\varepsilon})$, $\sigma_2 = \mathcal{O}(\frac{C_2\sqrt{\log(1/\delta)}}{N_2\varepsilon})$. Similarly, set $\sigma_3 = \mathcal{O}(\frac{C_3\sqrt{T\log(1/\delta)}}{n\sqrt{q}\varepsilon})$ and $\sigma_4 = \mathcal{O}(\frac{C_4\sqrt{\log(1/\delta)}}{\varepsilon}\max\{\frac{1}{N_4}, \frac{\sqrt{T}}{n\sqrt{q}}\})$, Algorithm 2 is (ε, δ)-DP.*

Theorem 2. *Under Assumption 1-3, with the parameter settings in Theorem 1 and some specific values of $\{N_i\}_{i=1}^4$, $\{C_i\}_{i=1}^4$, T, q and step size α, β in Algorithm 2, we have the following guarantee:*

$$\mathbb{E}\|\nabla\mathcal{F}(x)\| \leqslant \mathcal{O}(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{\frac{2}{3}}). \tag{4}$$

Remark 1. For DP-ERM with model size d and data size n, it has been shown that the utility bound given by DP-SGD is $O\big((\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{\frac{1}{2}}\big)$ (Wang et al., 2017). Thus, by adopting the proof in (Zhang et al., 2025a), we can show the same bound when using DP-SGD to equation 3. Comparing the bound of $O(\max\{\frac{1}{\sqrt{n}}, (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3}\})$ in Theorem 2, we can see that Algorithm 2 is better. The main reason is that, in DP-SGD, the noise we add depends on the l_2-norm sensitivity of the loss gradient, which is upper bounded by the Lipschitz constant. Thus, by using the composition theorem, adding the same scale of noise to the gradient in each iteration can guarantee DP.

In comparison, in Algorithm 2, we only add noise to the gradient every q iterations, i.e., T/q times in total, which is significantly smaller than that in DP-SGD. Moreover, when we add noise to the variance-reduced gradient estimator, such as $\nabla_x \mathcal{L}(x_t, \eta_{t+1}; \mathcal{B}) - \nabla_x \mathcal{L}(x_{t-1}, \eta_t; \mathcal{B})$, its sensitivity is $O(\max\{\|x_t - x_{t-1}\|, \|\eta_{t+1} - \eta_t\|\})$. Thus, when t becomes larger, we can show the noise we add becomes smaller than that in DP-SGD.

Remark 2. Note that there is a term $\frac{1}{\sqrt{n}}$ in the utility bound. This term comes from the estimation process of using the large batch size N_1, N_3, where we take all samples to estimate the gradient as the anchor point in preparation for the next $q-1$ updates. After that, accumulating the other $q-1$ gradient variations with the parameter setting we present in the theorem leads to the term with privacy parameters.

5 Improved Rates via DP Recursive-SPIDER

Recall that for DP-ERM with non-convex loss, the best-known result for utility is $\mathcal{O}((\frac{\sqrt{d}}{n\varepsilon})^{2/3})$ (Arora et al., 2023). Thus, from Theorem 2, we can see that Algorithm 2 achieves the same result only if $n \leq O(d^2)$. Thus, a natural question is whether we can further improve the bound for DP-DRO? In this section, we will provide an affirmative answer for DRO with KL-divergence.

For DP-DRO with KL-divergence, besides the dual problem in equation 3, it has been shown that the problem can be reformulated as the following compositional finite-sum optimization problem (Qi et al., 2022) with some ρ, which will be the main focus of this section:

$$\min_x \min_{\lambda \geq \lambda_0} \Psi(x, \lambda) := \lambda \log(\frac{1}{n} \sum_{i=1}^{n} \exp(\frac{\ell(x; \xi_i)}{\lambda})) + \lambda\rho. \tag{5}$$

Specifically, by considering $\mathbf{w} = (x^T, \lambda)^T \in \mathbb{R}^{d+1}$ as a single variable to be optimized, the objective function is a compositional function of $\mathbf{w}$ in the form of $f(g(\mathbf{w}))$, where $g(\mathbf{w}) = [g_1, g_2] = \left[\lambda, \frac{1}{n}\sum_{i=1}^{n} \exp\left(\frac{\ell(x;\xi_i)}{\lambda}\right)\right] \in \mathbb{R}^2$ and $f(g) = g_1 \log(g_2) + g_1\rho$. For convenience, we define $g(\mathbf{w}; \xi) = \exp(\frac{\ell(x;\xi)}{\lambda})$. The gradient of $\Psi(\mathbf{w})$ is given by:

$$\nabla_x \Psi(\mathbf{w}) = \nabla f(g(\mathbf{w})) \nabla_x g(\mathbf{w})$$
$$\nabla_\lambda \Psi(\mathbf{w}) = \nabla f(g(\mathbf{w})) \nabla_\lambda g(\mathbf{w}) + \log(g(\mathbf{w})) + \rho.$$

Note that since f is independent of the data. Thus, it is sufficient to develop a private estimator for $g(\mathbf{w})$, $\nabla_x g(\mathbf{w})$, and $\nabla_\lambda g(\mathbf{w})$. Similar to problem equation 3, we can see that x and λ play different roles. Thus, when considering DP, we need to treat them separately, which motivates our DP Recursive-Spider (see Algorithm 3 for details). DP Recursive-Spider adopts a similar idea as in DP Double-Spider, where we use a private variance-reduced gradient estimator for both $\nabla_x g(\mathbf{w})$ and $\nabla_\lambda g(\mathbf{w})$ in each iteration. The algorithm proceeds in phases of length q, during which it alternately updates x and η in each iteration. At the beginning of each phase (Lines 7–8), the gradient estimators are computed via empirical mean and subsequently privatized with Gaussian noise. In Lines 13–14, the estimators $\mathbf{u}_t$ and v_t are updated via gradient variation for the subsequent $q-1$ iterations. After obtaining the gradient estimate of $g(\mathbf{w})$, the function value $g(\mathbf{w})$ is recursively estimated in Line 16. This is motivated by STORM (Cutkosky & Orabona, 2019), which could provide a better estimate of $g(\mathbf{w})$. It is noteworthy that as $\beta_t \to 0$, s_t approaches the gradient variation, whereas when $\beta_t \to 1$, s_t behaves like the empirical mean.

The following (draft) results provide privacy and utility guarantees of DP Recursive-Spider. For the full versions and their proofs, we refer readers to Appendix D.

Theorem 3. *For any $\varepsilon > 0$ and $\delta \in (0, 1)$, let $\sigma_1 = \sigma_3 = \mathcal{O}(\frac{\sqrt{T\log(1/\delta)}}{nq\varepsilon})$, $\sigma_2 = \mathcal{O}(\|x_t - x_{t-1}\| \frac{\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\})$, $\hat{\sigma}_2 = \mathcal{O}(\frac{\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\})$. Similarly, we set $\sigma_4 = \mathcal{O}(\|\lambda_t - \lambda_{t-1}\| \frac{\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\})$, $\hat{\sigma}_4 = \mathcal{O}(\frac{\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_4}, \frac{\sqrt{T}}{n}\right\})$, and $\sigma_{s_t} = \mathcal{O}(\frac{\sqrt{T\log(1/\delta)}}{n\varepsilon})$. Then Algorithm 3 is (ε, δ)-DP.*

Theorem 4. *Under Assumption 1-3, with the parameters settings in Theorem 3 and some specific values of* $\{b_i\}_{i=1}^4$, $\{C_i\}_{i=1}^5$, T, q *and step size* η *in Algorithm 3, we can have the following guarantee:*

$$\mathbb{E}[\|\nabla \mathcal{F}(x)\|] = \mathcal{O}\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)^{2/3}\right).$$

6 Experiments

6.1 Experimental Setup

Dataset We consider 4 datasets that are widely used for studies in DRO: CIFAR10-ST (Qi et al., 2022), MNIST-ST (Zhang et al., 2025b), CelebA (Sagawa et al., 2019) and Fashion-MNIST. For constructing CIFAR10-ST, we artificially create imbalanced training data, where we only keep the last 100 images of each class for the first half of the classes, while keeping other classes and the test data unchanged. For constructing MNIST-ST, we convert MNIST into a binary task by mapping digits 0-4 $\rightarrow$ to class 0 and 5-9 $\rightarrow$ to class 1, and artificially create imbalanced training data by subsampling the positive class to a target imratio $\in (0, 0.5]$ (fraction of positives); the test split is kept balanced (imratio = 0.5), and training batches are sampled with the same ratio to match the desired imbalance.

Baselines For the comparison of test accuracy, we compare different algorithms for optimizing the minimax problem equation 2, including DP-SGDA (Rafique et al., 2022) and Private Diff (Zhang et al., 2025b), across different privacy budgets ε on the three aforementioned datasets. Besides, we take SCDRO (Zhang et al., 2023) as the non-private baseline, which is state-of-the-art.

Evaluation Metrics The evaluation of the selected four algorithms utilizes four primary metrics to assess utility, privacy, and optimization stability. The gradient norm serves as the central metric for measuring utility in terms of convergence to first-order stationary points. Predictive performance is evaluated via test accuracy across varying privacy budgets. To provide a more nuanced view of classifier performance, the Area Under the Curve (ROC AUC) and the F1 score are employed, with the latter specifically highlighting the balance between precision and recall in imbalanced scenarios. Finally, the Membership Inference Attack (MIA) AUC is used as an empirical measure of privacy robustness, where lower values signify superior protection against training data leakage.

MIA Experiments We evaluate the privacy robustness of different differentially private optimization algorithms using membership inference attacks (MIA) on CIFAR10-ST. Experiments are conducted under privacy budgets $\varepsilon \in \{0.1, 1, 3, 5, 8, 10\}$, covering both high-privacy and low-privacy regimes. For each algorithm–privacy budget configuration, we perform five independent runs with different random seeds and report results as mean $\pm$ standard deviation. We focus on two DP DRO methods: DP Double SPIDER and DP Recursive SPIDER. Privacy leakage is evaluated using a standard black-box MIA, where the attacker infers whether a data point was part of the training set based solely on the model's output confidence. Attack performance is measured by the area under the ROC curve (AUC).

6.2 Experimental Results

We conducted experiments on all four datasets and present representative results for three of them here. Additional experiments on AUC, F1 scores, and hyperparameter sensitivity are provided in the Appendix E.2, E.3, and E.4, respectively.

Performance Analysis. Table 3 illustrates that our proposed algorithms, DP Double-SPIDER (DP DS) and DP Recursive-SPIDER (RS DRO), consistently outperform the baseline methods (DP-SGDA and PrivateDiff) across nearly all datasets and privacy budget settings. The only exception is the MNIST-ST dataset, where all four algorithms achieve high accuracy ($\approx 99\%$), reaching a performance plateau that makes them effectively indistinguishable.

Due to the absence of base model fine tuning, the absolute accuracy on CIFAR10-ST is lower than on other benchmarks. However, the performance gap between our methods and the baselines is most pronounced on this dataset, particularly under stringent privacy constraints (e.g., $\varepsilon = 0.5$).

We draw several key observations from these results:

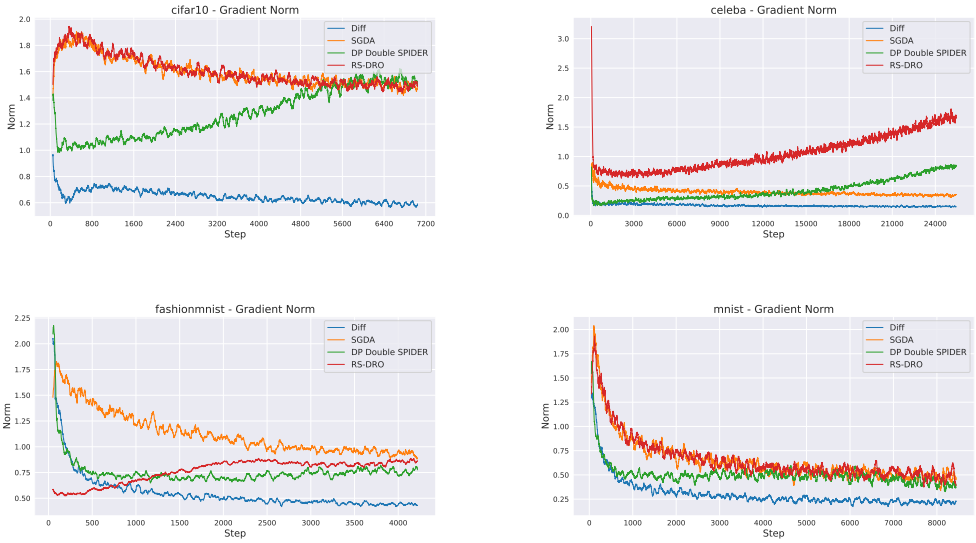


Figure 1: Experimental Results: The performances of four algorithms on CIFAR10-ST, CelebA, Fashion-MNIST, MNIST-ST respectively.

- **Performance Across Datasets:** On CIFAR10-ST, CelebA, and Fashion-MNIST, our proposed algorithms maintain a clear and significant lead over the baselines. This suggests that our variance-reduction approach is more effective for complex data distributions where gradient noise typically hinders convergence.
- **Impact of Privacy Budget ε:** As expected, the test accuracy for all algorithms improves as the privacy budget ε increases. Notably, our algorithms exhibit superior stability; they provide high utility even with very small ε, whereas the baseline methods suffer significantly more from the increased noise required for higher privacy levels.
- **Sensitivity Comparison:** The relative improvement in accuracy as ε increases from 0.5 to 10 is more substantial for DP-SGDA and PrivateDiff, especially on CIFAR10-ST. This high sensitivity indicates that these baseline methods are less robust to the noise injection process, while our methods achieve near-optimal utility even with limited privacy budgets.

Table 2: Membership Inference Attack (MIA) results under different privacy budgets. Lower AUC indicates better privacy. Results are reported as mean $\pm$ standard deviation over 5 runs. Bold values indicate best (lowest) privacy leakage for each privacy budget. DP Recursive SPIDER consistently achieves lower MIA AUC compared to DP Double SPIDER across all epsilon values.

Algorithm	MIA AUC (Mean $\pm$ Std)					
	ε=0.1	ε=1	ε=3	ε=5	ε=8	ε=10
DP Double SPIDER	0.9719 ± 0.0059	0.9723 ± 0.0045	0.9717 ± 0.0025	0.9723 ± 0.0021	0.9724 ± 0.0033	0.9715 ± 0.0022
DP Recursive SPIDER	$\mathbf{0.8234 \pm 0.0127}$	$\mathbf{0.7932 \pm 0.0072}$	$\mathbf{0.7958 \pm 0.0240}$	$\mathbf{0.8014 \pm 0.0383}$	$\mathbf{0.8001 \pm 0.0302}$	$\mathbf{0.7825 \pm 0.0149}$

Gradient Norm Analysis. Figure 1 illustrates the evolution of gradient norms on CIFAR10-ST, CelebA, Fashion-MNIST, and MNIST-ST, comparing DP-SGD and SGDA with the proposed DP Double SPIDER and DP Recursive SPIDER methods. Across most datasets, the baseline methods exhibit relatively large gradient norms and noticeable oscillations throughout the training process. In particular, DP-SGDA often shows slow convergence with persistent variance, while SGDA displays pronounced fluctuations, especially in later iterations. These behaviors indicate that privacy-induced noise and stochastic saddle-point dynamics hinder stable convergence to stationary points.

In contrast, DP Double SPIDER consistently achieves faster decay and substantially lower variance in gradient norms, reflecting the effectiveness of variance reduction via gradient-difference estimators. DP Recursive SPIDER further enhances stability, maintaining uniformly small gradient norms with minimal oscillations, particularly in the later stages where baseline methods plateau or become

Table 3: Comparison of Test Accuracy (%) Among Two Baseline Methods (DP SGDA and PrivateDiff Minimax) and Our Algorithms (DP Double-SPIDER and DP Recursive-SPIDER). SGDA: DP-SGDA; Private Diff: PrivateDiff Minimax; DP DS: DP Double SPIDER; RS DRO: DP Recursive SPIDER. The results for non-private are give by SCDRO (Zhang et al., 2023).

Dataset	CIFAR10-ST				MNIST-ST				Fashion-MNIST				CelebA			
	SGDA	Private Diff	DP DS	RS DRO	SGDA	Private Diff	DP DS	RS DRO	SGDA	Private Diff	DP DS	RS DRO	SGDA	Private Diff	DP DS	RS DRO
Non-private	-	-	-	67.58	-	-	-	-	-	-	-	-	-	-	-	-
$\varepsilon = 0.5$	41.80	43.06	**56.31**	53.81	99.16	99.18	99.49	**99.66**	88.13	88.82	91.39	**92.27**	88.16	88.18	89.37	**90.37**
$\varepsilon = 1$	45.10	46.13	55.67	**55.93**	99.48	99.30	99.50	**99.56**	89.16	89.51	91.37	**91.87**	89.17	89.01	89.37	**90.32**
$\varepsilon = 5$	51.54	51.29	55.69	**56.58**	**99.49**	99.47	99.47	99.45	90.45	90.70	91.36	**91.59**	89.76	90.14	89.38	**90.41**
$\varepsilon = 10$	52.47	51.68	**57.26**	55.67	**99.55**	99.32	99.46	99.45	90.39	91.29	91.38	**91.58**	89.93	90.35	89.38	**90.45**

Table 4: Clipping-threshold robustness of DP Double-SPIDER on CIFAR10-ST. All four clipping thresholds are tied as $C_1 = C_2 = C_3 = C_4 = C$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

ε	$C = 0.1$	$C = 0.5$	$C = 1.0$	$C = 2.0$	$C = 5.0$	$C = 10.0$	Range
1.0	45.80 ± 0.11	45.87 ± 0.04	45.87 ± 0.05	45.87 ± 0.04	45.87 ± 0.04	45.87 ± 0.04	0.07
5.0	46.02 ± 0.20	46.05 ± 0.17	46.05 ± 0.24	46.07 ± 0.24	46.08 ± 0.21	46.10 ± 0.21	0.08
10.0	45.99 ± 0.12	46.02 ± 0.06	46.05 ± 0.12	46.10 ± 0.15	46.12 ± 0.14	46.12 ± 0.12	0.13

unstable. Overall, these results demonstrate that our algorithms provide significantly more reliable and stable optimization dynamics under differential privacy, leading to improved convergence toward first-order stationary solutions in private nonconvex–strongly-concave optimization.

Hyperparameter Sensitivity. A natural concern is that DP Double-SPIDER introduces four clipping thresholds, C_1, C_2, C_3, C_4, which may appear to increase the tuning burden. We therefore conduct a controlled sensitivity study on CIFAR10-ST, using a fixed 20-epoch ResNet20 pretraining protocol and reporting the mean $\pm$ standard deviation of test accuracy over three random seeds. To test whether the four thresholds need to be tuned independently, we tie them to a single value, $C_1 = C_2 = C_3 = C_4 = C$, and sweep C over two orders of magnitude. As shown in Table 4, the accuracy is nearly unchanged across $C \in \{0.1, 0.5, 1, 2, 5, 10\}$. The range of mean accuracy is only 0.07, 0.08, and 0.13 percentage points for $\varepsilon = 1, 5, 10$, respectively. These results indicate that the four clipping thresholds in DP Double-SPIDER can be tied to a single default value in practice. Thus, although the algorithm contains multiple clipping constants in its theoretical statement, they do not create a practical fine-tuning burden.

Additional sensitivity studies, including pairwise clipping, random independent clipping, comparison with baselines, and DP Recursive-SPIDER hyperparameters, are reported in Appendix E.4.

MIA Experiments. The results of the Membership Inference Attack (MIA) are summarized in Table 2. The empirical results demonstrates that DP Recursive SPIDER consistently outperforms DP Double SPIDER, achieving significantly lower AUC values across all evaluated privacy budgets ε. While DP Double SPIDER exhibits high and nearly static privacy leakage (AUC ≈ 0.97) regardless of the privacy budget, DP Recursive SPIDER provides substantially better protection, with AUC values ranging between 0.78 and 0.82. Furthermore, DP Recursive SPIDER shows greater sensitivity to variations in ε, whereas DP Double SPIDER remains largely unaffected by changes in the privacy budget within the tested range. Overall, the superior privacy-preserving performance of DP Recursive SPIDER is evident across all settings, which aligns with our theoretical analysis.

7 Conclusion

In this paper, we investigate the Distributionally Robust Optimization (DRO) problem with differential privacy and non-convex loss functions. Specifically, we first analyze the Double SPIDER DRO, where the objective function is reformulated as a minimization problem via primal-dual transformation. Then, with the KL-divergence, we recast the original objective as a compositional problem; the inner component is a finite-sum loss function, a structure that facilitates the application of efficient optimization techniques. Inspired by STORM, we introduce the Recursive Double SPIDER, which achieves improved error bounds and matches the best-known results for DP minimax optimization with non-convex loss. Finally, we evaluate our methods against baselines on four datasets, and the empirical results demonstrate superior performance and corroborate our theoretical analysis.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the paper writing. We have not used generative AI tools for proving the theorems, and constructing the paper structure are not applicable to this work. Additionally, we used generative AI tools for the code of the experiments. We have reviewed all AI-assisted work. We have checked the LLM-generated code and review the polished paper by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pp. 308–318, 2016a. URL https://arxiv.org/abs/1607.00133.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016b.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *Advances in neural information processing systems*, 35:22505–22517, 2022.

Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pp. 1060–1092. PMLR, 2023.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pp. 5396–5427. PMLR, 2023.

Christopher A Choquette-Choo, Arun Ganesh, and Abhradeep Thakurta. Optimal rates for dp-sco with a single epoch and large batches. *arXiv e-prints*, pp. arXiv–2406, 2024.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pp. 4337–4348. PMLR, 2021.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34: 2771–2782, 2021.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in neural information processing systems*, 33:8847–8860, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. *Advances in Neural Information Processing Systems*, 34:10067–10080, 2021.

Qi Qi, Jiameng Lyu, Er Wei Bai, Tianbao Yang, et al. Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*, 2022.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

A Selvi, H Liu, and W Wiesemann. Differential privacy via distributionally robust optimization. arxiv 2023. *arXiv preprint arXiv:2304.12681*.

Aras Selvi, Huikang Liu, and Wolfram Wiesemann. Differential privacy via distributionally robust optimization. *Operations Research*, 2025.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Jinyan Su, Changhong Zhao, and Di Wang. Differentially private stochastic convex optimization in (non)-euclidean space revisited. In *Uncertainty in Artificial Intelligence*, pp. 2026–2035. PMLR, 2023.

Youming Tao, Yulian Wu, Xiuzhen Cheng, and Di Wang. Private stochastic convex optimization and sparse learning with heavy-tailed data revisited. In *IJCAI*, pp. 3947–3953, 2022.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pp. 6526–6535. PMLR, 2019.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *Journal of machine learning research*, 21(200):1–39, 2020a.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10081–10091. PMLR, 2020b.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021.

Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. A theory to instruct differentially-private learning via clipping bias reduction. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2170–2189. IEEE, 2023.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Vashney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, pp. 2192–2202. PMLR, 2022.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Liang Zhang, Kiran K Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, 2022.

Lijun Zhang, Peng Zhao, Zhen-Hua Zhuang, Tianbao Yang, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36:52490–52522, 2023.

Qi Zhang, Yi Zhou, Ashley Prater-Bennette, Lixin Shen, and Shaofeng Zou. Large-scale non-convex stochastic constrained distributionally robust optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8217–8225, 2024.

Qi Zhang, Yi Zhou, Simon Khan, Ashley Prater-Bennette, Lixin Shen, and Shaofeng Zou. Revisiting large-scale non-convex distributionally robust optimization. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Ruijia Zhang, Mingxi Lei, Meng Ding, Zihang Xiang, Jinhui Xu, and Di Wang. Improved rates of differentially private nonconvex-strongly-concave minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22524–22532, 2025b.

Canzhe Zhao, Yanjie Ze, Jing Dong, Baoxiang Wang, and Shuai Li. Differentially private temporal difference learning with stochastic nonconvex-strongly-concave optimization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 985–993, 2023.

Xinyu Zhou and Raef Bassily. Differentially private worst-group risk minimization. *arXiv preprint arXiv:2402.19437*, 2024.

Yingxue Zhou, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Arindam Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

A Additional Related Work

DRO Depending on how the uncertain variables are controlled, DRO formulations can be categorized as either constrained, imposing explicit limits on the uncertain variables or regularized, penalizing them through a regularization term in the objective (Levy et al., 2020). Duchi & Namkoong (2019) further showed that a DRO problem with quadratic regularization in constraint form is equivalent to minimizing the empirical loss augmented by a variance regularization term on itself.

Many primal–dual methods for min–max optimization (Duchi & Namkoong, 2021; Levy et al., 2020; Wang et al., 2021; Duchi et al., 2021) assume convexity of the loss function, which restricts their applicability to the broader class of non-convex machine learning problems. For non-convex DRO, existing works (Qi et al., 2021; 2022; Zhang et al., 2024) typically impose boundedness or even stronger assumptions on the loss ℓ, further limiting their practical relevance. The connection between compositional functions and DRO formulations has been recognized and exploited in prior work. Dentcheva et al. (2017) investigated the statistical estimation of compositional functionals with applications to conditional-value-at-risk (CVaR) measures, which are closely related to CVaR-constrained DRO, though without considering stochastic optimization algorithms. To the best of our knowledge, Qi et al. (2021) first applied stochastic compositional optimization to KL-regularized DRO problems, while Qi et al. (2022) extended this framework to the more challenging KL-constrained setting.

DP Minimax Optimization As DRO can be regarded as a constrained minimax problem. Therefore, our work is related to the recent methods for DP minimax optimization Yang et al. (2022); Zhang et al. (2025b; 2022); Zhao et al. (2023). For example, Yang et al. (2022) analyzes DP-SGDA for stochastic minimax learning using an algorithmic-stability framework. They reach the near-optimal bound for convex-concave objectives using SGDA. Zhang et al. (2025b) study differentially private stochastic minimax optimization in the nonconvex-strongly-concave (NC-SC) regime and give the first general results beyond convex/PL settings. However, without tailored algorithmic design, standard minimax optimization methods cannot be directly applied to DRO problems, rendering the two frameworks fundamentally incomparable. Under the DRO constraint, plain Stochastic Gradient Descent–Ascent (SGDA) fails in the non-convex setting, as it cannot guarantee convergence to an optimal point without the convex–concave structure (Hsieh et al., 2021) or additional structural assumptions on the optimization variables.

B Algorithms

Algorithm 1 Clipping (x, C)

Require: x and clipping threshold $C > 0$.

1: $\hat{x} = \min\left\{\frac{C}{\|x\|_2}, 1\right\} x$

2: **return** $\hat{x}$.

C Lemmas

Lemma 1. *(Partially generalized (L_0, L_1, L_2)-smoothness)(Zhang et al., 2025a). Under Assumptions 1 and 2, $\mathcal{L}(x, \eta)$ is (L_0, L_1)-partially smooth in x and L_2-smooth in η such that for any $x, x' \in \mathbb{R}^d$*

Algorithm 2 DP Double-SPIDER

Input: initialization (x_0, η_0), step sizes α_t, β_t, epoch size q, number of iterations T, batch sizes N_1, N_2, N_3, N_4, Dataset S, Clipping thresholds $\{C_i\}_{i=1}^4$
while $t \leq T-1$ **do**
  if $t \bmod q == 0$ **then**
    Draw N_1 samples $\mathcal{B}_1$ from S and compute $g_t \leftarrow \textbf{Clip}(\nabla_\eta \mathcal{L}(x_t, \eta_t; \mathcal{B}_1), C_1) + \omega_t$ where $\omega_t \sim \mathcal{N}(0, \sigma_1^2)$.
  else
    Draw N_2 samples $\mathcal{B}_2$ from S and compute $g_t \leftarrow \textbf{Clip}(\nabla_\eta \mathcal{L}(x_t, \eta_t; \mathcal{B}_2) - \nabla_\eta \mathcal{L}(x_{t-1}, \eta_{t-1}; \mathcal{B}_2), C_2) + g_{t-1} + \xi_t$, where $\xi_t \sim \mathcal{N}(0, \sigma_2^2)$.
  end if
  $\eta_{t+1} \leftarrow \eta_t - \alpha_t g_t$
  if $t \bmod q == 0$ **then**
    Draw N_3 samples $\mathcal{B}_3$ from S and compute $v_t \leftarrow \textbf{Clip}(\nabla_x \mathcal{L}(x_t, \eta_{t+1}; \mathcal{B}_3), C_3) + \tau_t$ where $\tau_t \sim \mathcal{N}(0, \sigma_3^2 \mathbb{I}_d)$.
  else
    Draw N_4 samples $\mathcal{B}_4$ from S and compute $v_t \leftarrow \textbf{Clip}(\nabla_x \mathcal{L}(x_t, \eta_{t+1}; \mathcal{B}_4) - \nabla_x \mathcal{L}(x_{t-1}, \eta_t; \mathcal{B}_4), C_4) + v_{t-1} + \chi_t$ where $\chi_t \sim \mathcal{N}(0, \sigma_4^2 \mathbb{I}_d)$.
  end if
  $x_{t+1} \leftarrow x_t - \beta_t v_t$
  $t \leftarrow t+1$
end while
Return Randomly select x, η from $1, \cdots, T$.

Algorithm 3 DP Recursive-SPIDER

Require: Dataset: $S \in \mathcal{X}^n$, Function: $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, Learning Rate: η, Phase Size: q, Batch Sizes $\{b_i\}_{i=1}^4$, Privacy Parameters: (ε, δ), Iterations: T, Clipping threashold: $\{C_i\}_{i=1}^5$ and $\mathbf{w} = (x, \lambda)$.
$w_0 = 0$
for $t = 1, \ldots, T$ **do**
  Update $s_{t+1}, \mathbf{v}_{t+1}, u_{t+1}$:
  if $\mathrm{mod}(t, q) = 0$ **then**
    Sample batch $\mathcal{B}_1$, $\mathcal{B}_3$ of size b_1, b_3 respectively.
    Sample $\nu_t \sim \mathcal{N}(0, \mathbf{I}_d \sigma_1^2)$, Sample $\omega_t \sim \mathcal{N}(0, \sigma_3^2)$
    $\mathbf{v}_t = \frac{1}{b_1} \sum_{\xi_i \in \mathcal{B}_1} \textbf{Clip}(\nabla_x g(\mathbf{w}_t; \xi_i), C_1) + \nu_t$
    $u_t = \frac{1}{b_3} \sum_{\xi_i \in \mathcal{B}_3} \textbf{Clip}(\nabla_\lambda g(\mathbf{w}_t; \xi_i), C_3) + \omega_t$
  else
    Sample batch $\mathcal{B}_2$, $\mathcal{B}_4$ of size b_2, b_4 respectively.
    Sample $\kappa_t \sim \mathcal{N}\left(0, \mathbf{I}_d \min\left\{\sigma_2^2, \hat{\sigma}_4^2\right\}\right)$
    Sample $\zeta_t \sim \mathcal{N}\left(0, \min\left\{\sigma_4^2, \hat{\sigma}_2^2\right\}\right)$
    $\mathbf{v}_t = \frac{1}{b_2} \sum_{\xi_i \in \mathcal{B}_2} \textbf{Clip}([\nabla_x g(\mathbf{w}_t; \xi_i) - \nabla_x g(\mathbf{w}_{t-1}; \xi_i)], C_2) + \kappa_t$
    $u_t = \frac{1}{b_4} \sum_{\xi_i \in \mathcal{B}_4} \textbf{Clip}([\nabla_\lambda g(\mathbf{w}_t; \xi_i) - \nabla_\lambda g(\mathbf{w}_{t-1}; \xi_i)], C_4) + \zeta_t$
  end if
  Take one sample ξ_0 from the dataset $\mathcal{D}$. And $s_t = \textbf{Clip}(g(\mathbf{w}_t; \xi_0), C_5) + (1-\beta_t)(s_{t-1} - \textbf{Clip}(g(\mathbf{w}_t; \xi_0), C_5)) + \mathcal{N}(0, \sigma_{s_t}^2)$
  Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t$, where $\mathbf{z}_t$ is given below:
  $\mathbf{z}_t = (\nabla f(s_t) \mathbf{v}_t^\top, \nabla f(s_t) u_t + \log(s_t) + \rho)^\top$
end for
return $(\mathbf{w}_\tau, \mathbf{v}_\tau, u_\tau, s_\tau)$, where $\tau \sim [T]$

and $\eta, \eta' \in \mathbb{R}$ *we have that*

$$\|\nabla_x \mathcal{L}(x, \eta) - \nabla_x \mathcal{L}(x', \eta)\| \leq \big(L_0 + L_1 |\nabla_\eta \mathcal{L}(x, \eta)|\big) \|x - x'\|, \tag{6}$$

$$|\nabla_\eta \mathcal{L}(x, \eta) - \nabla_\eta \mathcal{L}(x, \eta')| \leq L_2 |\eta - \eta'|, \tag{7}$$

where $L_0 = G + \frac{G^2 M}{\lambda}$, $L_1 = \frac{L}{G}$, *and* $L_2 = \frac{G^2 M}{\lambda}$.

It is worth noting that (L_0, L_1)-smoothness can be defined in two ways: in one formulation, the inequality is required only when $|x - y| \leqslant \frac{1}{L_0}$ Zhang et al. (2019), whereas in the other, no such restriction is imposed (Chen et al., 2023). As we will treat the dual variable and parameter variable separately, we will introduce partially generalized smoothness as shown in Lemma 1 (Zhang et al., 2025a).

Lemma 2. *(Partially affine variance noise)(Zhang et al., 2025a). Under Assumptions 1, 2 and 3, for any* $x \in \mathbb{R}^d$ *and* $\eta \in \mathbb{R}$, *we have that*

$$\mathbb{V}_{S\sim P_0}\big[\nabla_x \mathcal{L}(x, \eta, S)\big] \leq D_0 + D_1(\nabla_\eta \mathcal{L}(x, \eta))^2, \tag{8}$$

$$\mathbb{V}_{S\sim P_0}\big[\nabla_\eta \mathcal{L}(x, \eta, S)\big] \leq D_2, \tag{9}$$

where $D_0 = 8G^2 + 10G^2M^2\lambda^{-2}\sigma^2$, $D_1 = 8$, *and* $D_2 = G^2M^2\lambda^{-2}\sigma^2$.

To analyze Algorithm 2, we further develop the following property of gradients $\nabla_x \mathcal{L}(x, \eta, s)$ and $\nabla_\eta \mathcal{L}(x, \eta, s)$.

Lemma 3. *(Zhang et al., 2025a) For any* $x, x' \in \mathbb{R}^d$, $\eta, \eta' \in \mathbb{R}$ *and* $s \in \mathcal{S}$, $\nabla_\eta \mathcal{L}(x, \eta, s)$ *is* L_2*-continuous in* x:

$$\big|\nabla_\eta \mathcal{L}(x, \eta) - \nabla_\eta \mathcal{L}(x', \eta, s)\big| \leq L_2\|x - x'\|, \tag{10}$$

and $\nabla_x \mathcal{L}(x, \eta)$ *is* L_2*-continuous in* η:

$$\|\nabla_x \mathcal{L}(x, \eta, s) - \nabla_x \mathcal{L}(x, \eta', s)\| \leq L_2|\eta - \eta'|. \tag{11}$$

Theorem 5 (Abadi et al. (2016a))**.** *Let* $\varepsilon, \delta \in (0, 1]$ *and* c *be a universal constant. Let* $D \in \mathcal{Y}^n$ *be a dataset over some domain* $\mathcal{Y}$, *and let* $h_1, \ldots, h_T : \mathcal{Y} \to \mathbb{R}^d$ *be a series of (possibly adaptive) queries such that for any* $y \in \mathcal{Y}$, $t \in [T]$, $\|h_t(y)\|_2 \leq \lambda_t$. *Let* $\sigma_t = c\lambda_t\sqrt{\frac{\log(1/\delta)}{\varepsilon}}\max\left\{\frac{1}{b}, \frac{\sqrt{T}}{n}\right\}$. *Then the algorithm which samples batches of size* $B_1, \ldots, B_t$ *of size* b *uniformly at random and outputs* $\frac{1}{b}\sum_{y\in B_t} h_t(y) + g_t$ *for all* $t \in [T]$ *where* $g_t \sim \mathcal{N}(0, I\sigma_t^2)$, *is* (ε, δ)*-DP.*

D Omitted Proof

Besides Lemma 1, we have the following depiction of $\nabla_\eta \mathcal{L}(x, \eta)$:

Lemma 4. *For the function* $\nabla_\eta \mathcal{L}(x, \eta)$, *we have*

$$\|\nabla_\eta \mathcal{L}(x, \eta) - \nabla_\eta \mathcal{L}(x', \eta)\| \leqslant \frac{MG^2\|x - x'\|}{\lambda} \tag{12}$$

Proof. Recall the definition of $\mathcal{L}(x, \eta)$,

$$\mathcal{L}(x, \eta) = \lambda\sum_{i=1}^n p_i\psi^*(\frac{\ell_i(x) - G\eta}{\lambda}) + \eta.$$

Recall by Assumption 2, ψ^* is M-smooth, and $\ell(x)$ is G-Lipschitz. Therefore we have

$$\begin{aligned}
&\|\nabla_\eta \mathcal{L}(x, \eta) - \nabla_\eta \mathcal{L}(x', \eta)\| \\
=&\left\|\sum_{i=1}^n [p_i(\psi^*)'(\frac{\ell_i(x) - \eta}{\lambda}) - p_i(\psi^*)'(\frac{\ell_i(x') - \eta}{\lambda})]\right\| \\
\leqslant&\sum_{i=1}^n p_i\frac{M(\ell_i(x) - \ell_i(x'))}{\lambda} = \frac{MG^2\|x - x'\|}{\lambda}
\end{aligned}$$

□

Similarly, we have

Lemma 5. *For function $\nabla_x \mathcal{L}(x,\eta)$,we have*

$$\|\nabla_x \mathcal{L}(x,\eta) - \nabla_x \mathcal{L}(x,\eta')\| \leqslant \frac{ML\|\eta-\eta'\|}{\lambda}.$$

Proof.

$$\begin{aligned}
&\|\nabla_x \mathcal{L}(x,\eta) - \nabla_x \mathcal{L}(x,\eta)\| \\
=&\left\|\lambda\sum_{i=1}^n p_i \ell_i'(x)(\psi^*)'(\frac{\ell_i(x)-G\eta}{\lambda}) - \lambda\sum_{i=1}^n p_i \ell_i'(x)(\psi^*)'(\frac{\ell_i(x)-G\eta'}{\lambda})\right\| \\
\leqslant&\left\|\sum_{i=1}^n p_i \ell_i'(x)\frac{M\|\eta-\eta'\|}{\lambda}\right\| \leqslant LM\|\eta-\eta'\|/\lambda
\end{aligned}$$

□

Under Assumption 3, by Arora et al. (2022) we have the following guarantee:

Proposition 1. *To quantify the gradient $\nabla_\eta \mathcal{L}$, we have*

$$\|\nabla_\eta \mathcal{L}(x_t,\eta_{t+1})\| \leqslant H_0. \tag{13}$$

Proof. By Assumption 3, we have $\forall x,\eta, \max\{\|x\|,\|\eta\|\} \leqslant S_0$. Then, by Lemma 2 in Arora et al. (2022), we have $\|(\psi^*)'\| \leqslant 2S_0^2\sqrt{M} + 2MS_0^3$.

Therefore, by the definition of $\mathcal{L}$, we can derive the final guarantee:

$$\|\nabla_\eta \mathcal{L}\| \leqslant \frac{\lambda}{n}\sum(\psi^*)'(-\frac{\eta}{\lambda}) + 1 \leqslant 2S_0^3\sqrt{M} + 2MS_0^4 + 1 := H_0. \tag{14}$$

□

Theorem 6. *By setting $\sigma_1 = \frac{cC_1\sqrt{T\log(1/\delta)}}{n\sqrt{q}\varepsilon}$, $\sigma_2 = \frac{cC_2\sqrt{\log(1/\delta)}}{N_2\varepsilon}$, $C_2 = 2\max\{L_2\|\eta_t - \eta_{t-1}\|, GM\|x_t - x_{t-1}\|/\lambda\}$, $\sigma_3 = \frac{cC_3\sqrt{\log(1/\delta)}}{\varepsilon}\max\{\frac{1}{N_3}, \frac{\sqrt{T}}{n\sqrt{q}}\}$, $(L_0 + L_1\sqrt{H})$, $\sigma_4 = \frac{cC_4\sqrt{\log(1/\delta)}}{\varepsilon}\max\{\frac{1}{N_4}, \frac{\sqrt{T}}{n\sqrt{q}}\}$ where $C_4 = 2\max\{ML\|\eta_t - \eta_{t-1}\|/\lambda, (L_0 + L_1\sqrt{H})\|x_t - x_{t-1}\|\}$ based on the Gaussian mechanism, we claim that the Algorithm 2 is (ε,δ)-DP.*

Proof. **Privacy Proof:** Based on Theorem 5, we note that each estimate computed in line 4 has elements with ℓ_2 norm at most L_2, and this is involved in computation at most $\frac{T}{q}$ times. Similarly, for gradient variation at step t in line 6, we have norm bound C_2 and have that at most T such estimates are computed. Similar Results hold for Line 10 and 12. Therefore, by Gaussian mechanism 5, we know that Algorithm 2 is (ε,δ)-DP. □

Formal statement of Theorem 2.

Theorem 7. *Let $c_0 = \max(32L_2, 8L_0)$, $c_1 = \left(4 + \frac{8L_1^2D_2}{N_1L_0^2} + \frac{32L_1^2D_2}{N_1L_0^2} + \frac{16L_1^2L_2}{5D_1L_0^3}\right)$, $c_2 = \max\left(\frac{1}{8L_2} + \frac{L_1}{L_0^3}, 1\right)$, $c_3 = 1 + \frac{L_2}{10L_0} + \frac{L_0D_1+L_0+2L_0L_2D_2}{L_2} + \frac{33L_2^2}{5L_0L_2} + \frac{L_1^2}{15L_2^3} + \frac{L_1^2}{2L_0L_2^2}$, $c_4 = \frac{17}{4} + \sqrt{c_3} + \sqrt{\frac{1}{60L_2}}$. For $N_1 \geq \frac{6D_2c_0c_2}{\varepsilon^2}$, $N_2 \geq \max\left(\frac{20qD_1L_2}{L_0}, 20qc_2L_2, \frac{12qL_1^2c_0c_2}{L_0^2}, q\right)$, $N_3 \geq \max\left(\frac{200D_1L_2}{L_0}, \frac{3c_0(D_0+4D_1D_2)n}{2L_0}\right)$, $N_4 \geq \max\left(\frac{5qL_2}{L_0}, \frac{6qc_1c_0}{L_0}\right)$ and $q = O((\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})^{2/3})$ Given the above parameter setting and $\alpha = \frac{1}{4L_2}, \beta_t = \min\{\frac{1}{2L_0+L_1\sqrt{H}}, \frac{1}{L_0\sqrt{n}\|v_t\|}\}$ for Algorithm 2, we can have the following guarantee:*

$$\mathbb{E}\|\nabla\Psi\| \leqslant O(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3}). \tag{15}$$

Proof. **Utility Proof:** Since $\mathcal{L}(x,\eta)$ is L_2-smooth in η, we have that

$$\begin{aligned}\mathcal{L}(x_t,\eta_{t+1}) \leq& \mathcal{L}(x_t,\eta_t) - \langle \nabla_\eta \mathcal{L}(x_t,\eta_t), \alpha_t g_t\rangle + \frac{L_2}{2}(\alpha_t g_t)^2 \\ \leq& \mathcal{L}(x_t,\eta_t) - \langle \nabla_\eta \mathcal{L}(x_t,\eta_t), \alpha_t g_t\rangle + \alpha_t^2 L_2\Big((\nabla_\eta \mathcal{L}(x_t,\eta_t))^2 + (g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t))^2\Big).\end{aligned} \tag{16}$$

Taking the expectation on both sides of the above inequality, we can further show that, for the update of η, we have that

$$\begin{aligned}\mathbb{E}[\mathcal{L}(x_t,\eta_{t+1})] &\leq \mathbb{E}\left[\mathcal{L}(x_t,\eta_t) - \langle \nabla_\eta \mathcal{L}(x_t,\eta_t), \alpha_t g_t\rangle + \frac{L_2}{2}(\alpha_t g_t)^2\right] \\ &\leq \mathbb{E}[\mathcal{L}(x_t,\eta_t)] - \frac{1}{2}\mathbb{E}[\alpha_t (g_t)^2] + \frac{1}{2}\mathbb{E}[\alpha_t (g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t))^2] + \frac{L_2}{2}\mathbb{E}[\alpha_t^2 (g_t)^2].\end{aligned} \tag{17}$$

Similar to the previous one, for the update of x and any $\gamma > 0$, we have that

$$\begin{aligned}\mathcal{L}(x_{t+1},\eta_{t+1}) \leq& \mathcal{L}(x_t,\eta_{t+1}) - \langle \nabla_x \mathcal{L}(x_t,\eta_{t+1}), \beta_t v_t\rangle \\ &+ \frac{L_0 + L_1\|\nabla_\eta \mathcal{L}(x_t,\eta_t)\| + L_1 L_2 |\alpha_t g_t|}{2}\|\beta_t v_t\|^2 \\ \leq& \mathcal{L}(x_t,\eta_{t+1}) - \frac{\beta_t}{2}\left(\|v_t\|^2 + \|\nabla_x \mathcal{L}(x_t,\eta_{t+1})\|^2 - \|v_t - \nabla_x \mathcal{L}(x_t,\eta_{t+1})\|^2\right) \\ &+ \frac{L_0 + L_1\|\nabla_\eta \mathcal{L}(x_t,\eta_t) - g_t\| + (L_1 + L_1 L_2\alpha_t)|g_t|}{2}\|\beta_t v_t\|^2 \\ \leq& \mathcal{L}(x_t,\eta_{t+1}) - \frac{\beta_t}{2}\|v_t\|^2 + \frac{\beta_t}{2}\|v_t - \nabla_x \mathcal{L}(x_t,\eta_{t+1})\|^2 + \frac{L_0}{2}\beta_t^2\|v_t\|^2 \\ &+ \gamma(L_1 + L_1 L_2\alpha_t)^2 (g_t)^2 + \frac{\beta_t^4}{16\gamma}\|v_t\|^4 + \gamma L_1^2\|g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t)\|^2 + \frac{\beta_t^4}{16\gamma}\|v_t\|^4.\end{aligned} \tag{18}$$

Combine 17 and 18 and it follows that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{2}\alpha_t - \frac{L_2\alpha_t^2}{2} - \gamma(L_1 + L_1 L_2\alpha_t)^2\right)(g_t)^2 + \left(\frac{\beta_t}{2} - \frac{L_0\beta_t^2}{2}\right)\|v_t\|^2\right] \\ \leq& \mathbb{E}[\mathcal{L}(x_t,\eta_t) - \mathcal{L}(x_{t+1},\eta_{t+1})] + \frac{\beta_t^4}{8\gamma}\|v_t\|^4 + \frac{\beta_t}{2}\|v_t - \nabla_x \mathcal{L}(x_t,\eta_{t+1})\|^2 + \frac{\alpha_t}{2}(g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t))^2 + \gamma L_1^2(g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t)\end{aligned}$$

Setting $\alpha = \frac{1}{4L_2}$ and $\beta_t = \min\left(\frac{1}{2L_0}, \frac{1}{L_0\sqrt{n}\|v_t\|}\right)$, we have that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{3}{32L_2} - \gamma\frac{25L_1^2}{16}\right)(g_t)^2 + \frac{\beta_t}{4}\|v_t\|^2\right] \\ \leq& \mathbb{E}[\mathcal{L}(x_t,\eta_t) - \mathcal{L}(x_{t+1},\eta_{t+1})] + \frac{1}{8\gamma n^2 L_0^4} + \frac{1}{4L_0}\mathbb{E}\left[\|v_t - \nabla_x \mathcal{L}(x_t,\eta_{t+1})\|^2\right] + \left(\frac{1}{8L_2} + \gamma L_1^2\right)\mathbb{E}\left[(g_t - \nabla_\eta \mathcal{L}(x_t,\eta_t))^2\right].\end{aligned}$$

Let $\gamma = \frac{1}{50L_1}$ and take the sum of preceding inequality from $t = 0$ to $T-1$. According to Lemma 6 we then have that

$$\begin{aligned}
&\sum_{t=0}^{T-1} \frac{1}{16L_2}\mathbb{E}[(g_t)^2] + \mathbb{E}\left[\frac{\beta_t}{4}\|v_t\|^2\right] \\
\leq &\mathbb{E}[\mathcal{L}(x_0,\eta_0) - \mathcal{L}(x_T,\eta_T)] + \frac{7L_1T}{n^2L_0^4} \\
&+ \frac{1}{4L_0}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{D_0}{N_3} + \frac{4D_1D_2}{N_3N_1} + \frac{q}{nN_4}\left(4 + \frac{8L_1^2D_2}{L_0^2} + \frac{32L_1^2D_2}{N_1L_0^2} + \frac{64qL_1^2L_2^2}{nN_2L_0^4}\right)\right] \\
&+ \frac{1}{4L_0}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{4D_1}{N_3} + \frac{D_1}{8N_3} + \frac{33qL_1^2}{nN_4L_0^2} + \frac{q}{8N_4} + \frac{qD_1}{2N_2N_3} + \frac{4q^2L_1^2}{nN_2N_4L_0^2}\right](g_t)^2 \\
&+ \frac{1}{8L_2} + \frac{L_1}{50}\sum_{t=0}^{T-1}\left[\left(\frac{D_2}{N_1} + \frac{2qL_2^2}{nN_2L_0^2} + \frac{q}{8N_2}(g_t)^2\right)\right] + \frac{T}{q}\sigma_0^2.
\end{aligned}$$

Let $c_0 = \max(32L_2, 8L_0)$, $c_1 = \left(4 + \frac{8L_1^2D_2}{L_0^2} + \frac{32L_1^2D_2}{N_1L_0^2} + \frac{16L_1^2L_2}{5D_1L_0^3}\right)$, and $c_2 = \max\left(\frac{1}{8L_2} + \frac{L_1}{50}, 1\right)$. For $N_3 \geq \max\left(\frac{200D_1L_2}{L_0}, \frac{3c_0n(D_0+4D_1D_2)}{2L_0}\right)$, $N_4 \geq \max\left(\frac{5qL_2}{L_0}, \frac{6qc_1c_0}{L_0}\right)$, $N_2 \geq \max\left(\frac{20qD_1L_2^2}{L_0}, 20qc_2L_2, \frac{12qL_2^2c_0c_2}{L_0^2}, q\right)$, and $N_1 \geq n6D_2c_0c_2$, we then have that

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{32L_2}(g_t)^2 + \frac{\beta_t}{4}\|v_t\|^2\right] \\
\leq &\mathbb{E}[\mathcal{L}(x_0,\eta_0) - \mathcal{L}(x_T,\eta_T)] + \frac{T}{6nc_0} + \frac{D_0 + 4D_1D_2}{4L_0N_3}T \\
&+ T\frac{q}{nN_4L_0}\left(4 + \frac{8L_1^2D_2}{L_0^2} + \frac{32L_1^2D_2}{N_1L_0^2} + \frac{16L_1^2L_2}{5D_1L_0^3}\right) + \left(\frac{1}{8L_2} + \frac{L_1}{50}\right)\frac{D_2T}{N_1} + \left(\frac{1}{8L_2} + \frac{L_1}{50}\right)\frac{T\,2qL_2^2}{nN_2L_0^2} \\
\leq &\mathbb{E}[\mathcal{L}(x_0,\eta_0) - \mathcal{L}(x_T,\eta_T)] + \frac{5T}{6nc_0} + \frac{T}{q}\sigma_0^2.
\end{aligned}$$

For $T = n_0q \geq 6n(c_0(\mathcal{L}(x_0,\eta_0) - \inf_{x,\eta}\mathcal{L}(x,\eta)))$, we can find some t_0 with $\bmod q = 0$ such that

$$\sum_{t'=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{1}{32L_2}(g_{t'})^2 + \frac{\beta_{t'}}{4}\|v_{t'}\|^2\right] \leq \frac{q}{32nL_2} + \sigma_0^2. \tag{19}$$

Moreover we can find some $t \in [t_0, t_0+q-1]$ such that

$$\frac{1}{32L_2}\mathbb{E}[(g_t)^2] + \mathbb{E}\left[\frac{\beta_t}{4}\|v_t\|^2\right] \leq \frac{1}{c_0n} + \frac{\sigma_0^2}{q}. \tag{20}$$

Based on 20 and $c_0 = \max(32L_2, 8L_0)$, we have that $\mathbb{E}[(g_t)^2] \leq \frac{1}{n} + \sigma_1 + q\sigma_2^2$. Based on 22 and 23, we can further show that

$$\begin{aligned}
\mathbb{E}[(g_t - \nabla_\eta\mathcal{L}(x_t,\eta_t))^2] \leq &\frac{D_2}{N_1} + \sum_{t'=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{2L_2^2}{nN_2L_0^2} + \frac{2L_2^2}{N_2}\alpha_{t'-1}^2(g_{t'})^2\right] + \sigma_1^2 + (q-1)\sigma_2^2, \\
\leq &\frac{1}{6nc_0} + \frac{1}{6nc_0} + \frac{1}{160L_2n} + \sigma_1^2 + (q-1)\sigma_2^2 \leq \frac{1}{60L_2n} + \sigma_1^2 + q\sigma_2^2.
\end{aligned}$$

Thus we have that

$$\mathbb{E}[\|\nabla_\eta\mathcal{L}(x_t,\eta_{t+1})\|] \leq \mathbb{E}[\|g_t - \nabla_\eta\mathcal{L}(x_t,\eta_t)\|] + |g_t| + L_2\alpha_t|g_t| \leq \left(\frac{5}{4} + \sqrt{\frac{1}{60L_2}}\right)\frac{1}{\sqrt{n}} + 3\sigma_1^2 + 3q\sigma_2^2.$$

Moreover, we can show that

$$\frac{1}{nL_0} + \sigma_3^2 + q\sigma_4^2 \geq \mathbb{E}[\beta_t \|v_t\|^2]$$

As a result, we have that $\mathbb{E}[\|v_t\|] \leq O\left(\frac{3}{\sqrt{n}} + L_0(\sigma_3 + \sqrt{q}(\sigma_4))\right)$.

Based on 24 and 25, we can further show that

$$\begin{aligned}
&\mathbb{E}[\|v_t - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\|^2] \\
\leq &\mathbb{E}\left[\frac{D_0 + 4D_1(g_{t_0})^2 + 2D_1 L_2^2 \alpha_{t_0}^2 (g_{t_0})^2}{N_3} + \frac{4D_1(g_{t_0} - \nabla_\eta \mathcal{L}(x_{t_0}, \eta_{t_0}))^2}{N_3}\right] \\
&+ \sum_{t'=t_0+1}^{t_0+q-1} \mathbb{E}\left[\frac{2}{nN_4}\left(2 + \frac{8L_1^2 L_2^2}{L_0^2}\alpha_{t'-1}^2 (g_{t'-1})^2 + \frac{L_1^2 L_2}{L_0^2} D_2\right) + \frac{2}{N_4} L_2^2 \alpha_{t'}^2 (g_{t'})^2\right] \\
&+ \frac{2}{nN_4} \sum_{t'=t_0+1}^{t_0+q-1} \mathbb{E}\left[\frac{16L_2^2}{L_0^2}(g_{t'-1})^2 + \frac{16L_2^2}{L_0^2}(\nabla_x \mathcal{L}(x_{t'-1}, \eta_{t'-1}) - g_{t'-1})^2\right] + \frac{\sigma_0^2}{q} \\
\leq &\frac{D_0}{N_3} + \left(\frac{4D_1 + 2D_1 L_2^2 \alpha_{t_0}^2}{N_3} + \frac{L_1^2}{nN_4 L_0^2} + \frac{L_2^2}{8N_4 L_2} + \frac{32L_1^2}{nN_4 L_0^2}\right) \sum_{t'=t_0+1}^{t_0+q-1} \mathbb{E}[(g_{t'})^2] \\
&+ \frac{4q}{nN_4} + \frac{8qL_1^2 D_2}{nN_4 L_0^2} + \left(\frac{4D_1}{N_3} + \frac{32L_1^2}{nN_4 L_0^2}\right) \sum_{t'=t_0+1}^{t_0+q-1} \mathbb{E}[(\nabla_\eta \mathcal{L}(x_{t'}, \eta_{t'}) - g_{t'})^2] + \frac{\sigma_0^2}{q} \\
\leq &\frac{1}{n} + \left(\frac{D_1 L_0}{L_2} + \frac{33L_1^2}{5L_0 L_2} + \frac{L_2}{40L_0}\right)(\frac{1}{n} + \sigma_1^2 + q\sigma_2^2) + \frac{4L_0}{5L_2 n} + \frac{8L_0 L_2 D_2}{5L_2 n} \\
&+ \left(\frac{4D_1 L_0}{5L_2} + \frac{32L_1^2}{5L_0 L_2}\right)\left(\frac{1}{15L_2 n} + \sigma_1^2 + q\sigma_2^2\right) \\
\leqslant &\frac{c}{n} + c'(\sigma_1^2 + \sigma_3^2 + q\sigma_2^2 + q\sigma_4^2)
\end{aligned}$$

By taking $q = \left(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}}\right)^{2/3}$, and select t randomly from $[T]$, we have

$$\mathbb{E}[\|\nabla_x \mathcal{L}(x_t, \eta_{t+1})\|] \leqslant \mathbb{E}[\|v_t - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\| + \|v_t\|] \leqslant O\left(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3}\right)$$

By Jin et al. (2021), we know that

$$\mathbb{E}\|\nabla \Psi\| \leqslant O(\frac{1}{\sqrt{n}} + (\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3}). \tag{21}$$

□

In the following section, we are going to prove some key lemmas used in the proof of the mian theorem. First, in Lemma 6, we give the control of two critical terms: $\mathbb{E}[(g_t - \nabla_\eta \mathcal{L}(x_t, \eta_t))^2]$ and $\mathbb{E}[\|v_t - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\|^2]$

Lemma 6. *With the parameters selected in Theorem 2, for each $t < T$ that can be divided by q, we have that*

$$\sum_{t=t_0}^{t_0+q-1} \mathbb{E}[(g_t - \nabla_\eta \mathcal{L}(x_t, \eta_t))^2] \leq \sum_{t=t_0}^{t_0+q-1} \mathbb{E}\left[\frac{D_2}{N_1} + \frac{2qL_2^2}{nN_2 L_0^2} + \frac{2qL_2^2}{N_2}\alpha_t (g_t)^2\right] + q\sigma_1^2 + \frac{q(q-1)}{2}\sigma_2^2$$

and

$$\sum_{t=t_0}^{t_0+q-1} \mathbb{E}[\|v_t - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\|^2] \leq \sum_{t=t_0}^{t_0+q-1} \mathbb{E}\left[\frac{D_0}{N_3} + \frac{4D_1 D_2}{N_3 N_1} + \frac{q}{nN_4}\left(4 + \frac{8L_1^2 D_2}{L_0^2} + \frac{32L_1^2 D_2}{N_1 L_0^2} + \frac{64qL_1^2 L_2^2}{nN_2 L_0^4}\right)\right.$$

$$+\frac{1}{4L_0}\sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{4D_1}{N_3}+\frac{D_1}{8N_3}+\frac{33qL_1^2}{nN_4L_0^2}+\frac{q}{8N_4}+\frac{qD_1}{2N_2N_3}+\frac{4q^2L_1^2}{nN_2N_4L_0^2}\right](g_t)^2\Bigg]+\sigma_0^2,$$

where $\sigma_0 = q(\sigma_2^2+\sigma_3^2)+\frac{q(q-1)}{2}(\sigma_2^2+\sigma_4^2)$.

Proof. If $t \bmod q = 0$, g_t is an unbiased estimate of $\nabla_\eta \mathcal{L}(x_t,\eta_t)$ and according to Lemma 2, we have that

$$\mathbb{E}[(g_t-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2] \le \frac{D_2}{N_1}+\sigma_1^2.$$

Otherwise, we have that

$$\begin{aligned}&\mathbb{E}[(g_t-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2] = \mathbb{E}[(\nabla_\eta\mathcal{L}(x_t,\eta_t,B_2)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},B_2)+g_{t-1}-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2]\\ =&\mathbb{E}[(\nabla_\eta\mathcal{L}(x_t,\eta_t,B_2)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},B_2)+\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1})-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2]+\mathbb{E}[(g_{t-1}-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1}))^2]+\sigma_2^2,\end{aligned} \tag{22}$$

where the last inequality is due to the fact that $\nabla_\eta\mathcal{L}(x_t,\eta_t,B_2)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},B_2)$ is an unbiased estimate of $\nabla_\eta\mathcal{L}(x_t,\eta_t)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1})$.

We now focus on the first term of equation, which can be further bounded as follows:

$$\begin{aligned}&\mathbb{E}[(\nabla_\eta\mathcal{L}(x_t,\eta_t,B_2)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},B_2)+\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1})-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2]\\ \le&\frac{1}{N_2}\mathbb{E}[(\nabla_\eta\mathcal{L}(x_t,\eta_t,S))^2-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},S))^2]\\ \le&\frac{2}{N_2}\mathbb{E}\left[(\nabla_\eta\mathcal{L}(x_t,\eta_t,S)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},S))^2+(\nabla_\eta\mathcal{L}(x_t,\eta_t,S)-\nabla_\eta\mathcal{L}(x_{t-1},\eta_{t-1},S))^2\right]\\ \le&\mathbb{E}\left[\frac{2L_2^2}{N_2}(\alpha_{t-1}^2(g_{t-1})^2+\beta_{t-1}^2\|v_{t-1}\|^2)\right]\\ \le&\frac{2L_2^2}{nN_2L_0^2}+\frac{2L_2^2}{N_2}\alpha_{t-1}^2(g_{t-1})^2,\end{aligned} \tag{23}$$

where the first inequality is due to the fact that the square of expectation is not larger than the expectation of square, the third inequality is due to the continuous properties shown in Lemma 4, and the last inequality is due to the fact that $\beta_t = \min\left(\frac{1}{2L_0},\frac{1}{L_0\sqrt{n}\|v_t\|}\right)$.

Combining the above equations, for $t_0 \bmod q = 0$, we have that

$$\sum_{t=t_0}^{t_0+q-1}\mathbb{E}[(g_t-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2] \le \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{D_2}{N_1}+\frac{2qL_2^2}{N_2L_0^2}\varepsilon^2+\frac{q}{8N_2}(g_t)^2\right]+\frac{q(q-1)}{2}\sigma_2^2.$$

We then focus on the estimate of the gradient to x. If $t \bmod q = 0$, we have that

$$\mathbb{E}[\|v_t-\nabla_x\mathcal{L}(x_t,\eta_{t+1})\|^2] \le \frac{D_0+D_1(\nabla_\eta\mathcal{L}(x_t,\eta_t))^2}{N_3}+\sigma_3^2.$$

$$\begin{aligned}\mathbb{E}[\|v_t-\nabla_x\mathcal{L}(x_t,\eta_{t+1})\|^2] &\le \frac{D_0+2D_1(\nabla_\eta\mathcal{L}(x_t,\eta_t))^2+2D_1L_2^2\alpha_t^2(g_t)^2}{N_3}+\sigma_3^2\\ &\le \frac{D_0+4D_1(g_t)^2+2D_1L_2^2\alpha_t^2(g_t)^2+4D_1(g_t-\nabla_\eta\mathcal{L}(x_t,\eta_t))^2}{N_3}+\sigma_3^2,\end{aligned}$$

where the second inequality is due to the L_2-smoothness on η and the update of η. Otherwise, we have that

$$\begin{aligned}&\mathbb{E}[\|v_t-\nabla_x\mathcal{L}(x_t,\eta_{t+1})\|^2] = \mathbb{E}[\|\nabla_x\mathcal{L}(x_t,\eta_{t+1},B_4)-\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1},B_4)+v_{t-1}-\nabla_x\mathcal{L}(x_t,\eta_{t+1})\|^2]\\ =&\mathbb{E}[\|\nabla_x\mathcal{L}(x_t,\eta_{t+1},B_4)-\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1},B_4)+\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1})-\nabla_x\mathcal{L}(x_t,\eta_{t+1})\|^2]+\mathbb{E}[\|v_{t-1}-\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1})\|^2],\end{aligned} \tag{24}$$

since $\nabla_x\mathcal{L}(x_t,\eta_{t+1},B_4)-\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1},B_4)$ is an unbiased estimate of $\nabla_x\mathcal{L}(x_t,\eta_{t+1})-\nabla_x\mathcal{L}(x_{t-1},\eta_{t-1})$.

We now focus on the first term of equation , which can be further bounded as follows:

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_x \mathcal{L}(x_t, \eta_{t+1}, B_4) - \nabla_x \mathcal{L}(x_{t-1}, \eta_{t-1}, B_4) + \nabla_x \mathcal{L}(x_{t-1}, \eta_{t-1}) - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\|^2] \\
\leq &\frac{1}{N_4}\mathbb{E}[\|\nabla_x \mathcal{L}(x_t, \eta_{t+1}, S) - \nabla_x \mathcal{L}(x_{t-1}, \eta_{t-1}, S)\|^2] \\
\leq &\frac{2}{N_4}\mathbb{E}[\|\nabla_x \mathcal{L}(x_t, \eta_{t+1}, S) - \nabla_x \mathcal{L}(x_t, \eta_t, S)\|^2 + \|\nabla_x \mathcal{L}(x_t, \eta_t, S) - \nabla_x \mathcal{L}(x_{t-1}, \eta_{t-1}, S)\|^2] \\
\leq &\frac{2}{N_4}\mathbb{E}\left[L_2^2\alpha_t^2(g_t)^2 + (2L_2^2 + 2L_2^2)\|\nabla_x \mathcal{L}(x_t, \eta_t, S)\|^2\beta_t^2\|v_t - u_t\|^2\right] \\
\leq &\frac{2}{N_4}\mathbb{E}\left[\frac{4}{n}\left(2 + \frac{8L_1^2L_2^2}{L_0^2}(\nabla_x \mathcal{L}(x_t, \eta_t))^2 + \frac{8L_1^2L_2^2}{L_0^2}\alpha_{t-1}^2(g_{t-1})^2 + \frac{4L_1^2L_2}{L_0^2}D_2\right) + \frac{2}{N_4}L_2^2\alpha_t^2(g_t)^2\right] \\
&+ \frac{2}{nsN_4}\sum_{t'=t_0+1}^{t_0+q-1}\mathbb{E}\left[16\frac{L_2^2}{L_0^2}(g_{t'-1})^2 + 16\frac{L_2^2}{L_0^2}(\nabla_x \mathcal{L}(x_{t'-1}, \eta_{t'-1}) - g_{t'-1})^2\right],
\end{aligned} \tag{25}
$$

where the third inequality is due to Lemma 3, the fourth inequality is due to $\beta_t = \min\left(\frac{1}{2L_0}, \frac{1}{L_0\sqrt{n}\|v_t\|}\right)$, and Lemma 2.

Combining the above inequality, and for $t_0 \bmod q = 0$, we have that

$$
\begin{aligned}
&\sum_{t=t_0}^{t_0+q-1}\mathbb{E}[\|v_t - \nabla_x \mathcal{L}(x_t, \eta_{t+1})\|^2] \leq \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{D_0}{N_3} + \frac{q}{N_4}\left(4 + \frac{8L_1^2D_2}{L_0^2}\right)\right] \\
&+ \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{4D_1}{N_3} + \frac{D_1}{8N_3} + \frac{33qL_1^2}{nN_4L_0^2} + \frac{q}{8N_4}\right](g_t)^2 + \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{4D_1}{N_3} + \frac{32qL_1^2}{nN_4L_0^2}\right](g_t - \nabla_\eta \mathcal{L}(x_t, \eta_t))^2 \\
\leq &\sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{D_0}{N_3} + \frac{q}{nN_4}\left(4 + \frac{8L_1^2D_2}{L_0^2}\right)\right] + \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\left(\frac{4D_1}{N_3} + \frac{D_1}{8N_3} + \frac{33qL_1^2}{nN_4L_0^2} + \frac{q}{8N_4}\right)(g_t)^2\right] \\
&+ \left(\frac{4D_1}{N_3} + \frac{32qL_1^2}{nN_4L_0^2}\right)\sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\left(\frac{D_2}{N_1} + \frac{2qL_2^2}{nN_2L_0^2} + \frac{2qL_2^2}{N_2}\alpha_t^2(g_t)^2\right)\right] + q\sigma_1^2 + \frac{q(q-1)}{2}\sigma_2^2 + q\sigma_3^2 + \frac{q(q-1)}{2}\sigma_4^2 \\
\leq &\sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\frac{D_0}{N_3} + \frac{4D_1D_2}{N_3N_1} + \frac{q}{nN_4}\left(4 + \frac{8L_1^2D_2}{L_0^2} + \frac{32L_1^2D_2}{N_1L_0^2} + \frac{64qL_1^2L_2^2}{nN_2L_0^4}\right)\right] \\
&+ \sum_{t=t_0}^{t_0+q-1}\mathbb{E}\left[\left(\frac{4D_1}{N_3} + \frac{D_1}{8N_3} + \frac{33qL_1^2}{nN_4L_0^2} + \frac{q}{8N_4} + \frac{qD_1}{2N_2N_3} + \frac{4q^2L_1^2}{nN_2N_4L_0^2}\right)(g_t)^2\right] + \sigma_0^2.
\end{aligned}
$$

□

Next, here is the full statement of Theorem 4:

Theorem 8. *Let* $\sigma_1 = \frac{cG\sqrt{T\log(1/\delta)}}{nq\varepsilon}$, $\sigma_2 = \frac{cL\sqrt{\log(1/\delta)}}{\varepsilon}\max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$, $\hat{\sigma}_2 = \frac{2cG\sqrt{\log(1/\delta)}}{\varepsilon}\max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$, $\sigma_{s_t} = \frac{cG\sqrt{T\log(1/\delta)}}{n\varepsilon}$ *where* c *is a universal constant. Let* $n \geq \max\left\{\frac{(G\varepsilon)^2}{\Psi_0 Ld\log(1/\delta)}, \frac{\sqrt{d}\max\{1, \sqrt{L\Psi_0/G}\}}{\varepsilon}\right\}$. *Algorithm 3 run with parameter settings* $\eta = \min\{\frac{1}{2L}, c\eta_t^2\}$, $b_1 = n$, $b_2 = \left\lfloor\max\left\{\left(\frac{Gn\varepsilon}{\sqrt{\Psi_0 Ld\log(1/\delta)}}\right)^{2/3}, \frac{(Gnd\log(1/\delta))^{1/3}}{(L\Psi_0)^{1/6}\varepsilon^{2/3}}\right\}\right\rfloor$, $T = \left\lfloor\max\left\{\left(\frac{(\Psi_0 L)^{1/4}n\varepsilon}{\sqrt{Gd\log(1/\delta)}}\right)^{4/3}, \frac{n\varepsilon}{\sqrt{d\log(1/\delta)}}\right\}\right\rfloor$, *and* $q = \left\lfloor\frac{n^2\varepsilon^2}{L^2Td\log(1/\delta)}\right\rfloor$ *and* $\Psi_0 = \Psi(\mathbf{0}) - \min_{w\in\mathbb{R}^d}\{\Psi(\mathbf{w})\}$. *Furthermore, let* $\varepsilon, \delta \in [0, 1]$ *and* $n \geq$

$\max\left\{\frac{(L_0)^2}{\Psi_0 L d\log(1/\delta)}, \frac{\sqrt{d}\max\{1,\sqrt{L\Psi_0/G}\}}{\varepsilon}\right\}$. *Algorithm 3 is* (ε,δ)*-DP and Algorithm 3 satisfies Let* $\sigma_1 = \frac{cG\sqrt{T\log(1/\delta)}}{nq\varepsilon}$, $\sigma_2 = \frac{cL\sqrt{\log(1/\delta)}}{\varepsilon}\max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$, $\hat{\sigma}_2 = \frac{2cG\sqrt{\log(1/\delta)}}{\varepsilon}\max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$, $\sigma_{s_t} = \frac{cG\sqrt{T\log(1/\delta)}}{n\varepsilon}$ *where* c *is a universal constant.*

$$\mathbb{E}[\|\nabla\Psi(x;S)\|] = O\left(\left(\frac{\sqrt{\Psi_0 LG}\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)^{2/3}\right).$$

Proof. **Privacy Proof:** We rely on the moment accountant analysis of the Gaussian mechanism per iteration. Note that each gradient estimate computed in Line 7–8 of Algorithm 3 has elements with ℓ_2-norm at most G, and this estimator is computed at most $\frac{T}{q}$ times. Similarly, in Line 13-14 we have the norm bound $L\|w_t - w_{t-1}\|$ and have at most T such estimates are computed. Therefore, according to the Gaussian mechanism, we can claim that Algorithm 3 is (ε,δ)-DP.

Utility Proof: First we start with the first technical lemma:

Lemma 7. *Let* $\mathbf{z}_t = \nabla f_{\lambda_t}(s_t)\mathbf{q}_t + \mathbf{q}_{\lambda_t}$, *where* $\mathbf{q}_t = (\mathbf{v}_t^\top, u_t)^\top$, $\mathbf{q}_{\lambda_t} = (\boldsymbol{\theta}^\top, \log(s_t)+\rho)^\top$ *and* $\boldsymbol{\theta}\in\mathbb{R}^d$. *Let* $\|\mathscr{X}_t\|^2 = \|s_t - g(\mathbf{w}_t)\|^2$. *Run Algorithm 3, and then for every* $t\in\{1,\dots,T\}$ *we have*

$$\|\mathbf{z}_t - \nabla F(\mathbf{w}_t)\|^2 \le 4L_\Psi^2\|\mathscr{X}_t\|^2 + \|\mathbf{v}_t - \nabla_w g(\mathbf{w}_t)\|^2 + \|u_t - \nabla_\lambda g(\mathbf{w}_t)\|^2.$$

Proof. By simple expansion, it holds that

$$\begin{aligned}
\|\mathbf{z}_t - \nabla F(\mathbf{w}_t)\|^2 &= \|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla_w g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{v}_t\|^2 \\
&\quad + \|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla_\lambda g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{v}_t + \log(g(\mathbf{w}_t)) - \log(s_t)\|^2 \\
&\overset{(a)}{\le} 2\|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla_w g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{v}_t\|^2 + 2\|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla_\lambda g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)u_t\|^2 \\
&\quad + 2\|g(\mathbf{w}_t) - s_t\|^2 \\
&= 2\|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{q}_t\|^2 + 2\|g(\mathbf{w}_t) - s_t\|^2, \qquad (26)
\end{aligned}$$

where the inequality (a) is because $\|\mathbf{a}+\mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, and $|\log(x) - \log(y)| \le |x-y|$ for all $x, y \ge 1$.

Applying the smoothness and Lipschitz continuity of f_λ and g, we obtain

$$\begin{aligned}
&\|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{q}_t\|^2 \\
&= \|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\nabla g(\mathbf{w}_t) + \nabla f_{\lambda_t}(s_t)\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{q}_t\|^2 \\
&\le 2\|\nabla f_{\lambda_t}(g(\mathbf{w}_t))\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\nabla g(\mathbf{w}_t)\|^2 + 2\|\nabla f_{\lambda_t}(s_t)\nabla g(\mathbf{w}_t) - \nabla f_{\lambda_t}(s_t)\mathbf{q}_t\|^2 \\
&\le 2L_g^2 L^2_{\nabla f_{\lambda_t}}\|s_t - g(\mathbf{w}_t)\|^2 + 2L_{f_{\lambda_t}}\|\mathbf{q}_t - \nabla g(\mathbf{w}_t)\|^2 + 2\|g(\mathbf{w}_t) - s_t\|^2. \qquad (27)
\end{aligned}$$

Noting $\|\mathbf{q}_t - \nabla g(\mathbf{w}_t)\|^2 = \|\mathbf{v}_t - \nabla_w g(\mathbf{w}_t)\|^2 + \|u_t - \nabla_\lambda g(\mathbf{w}_t)\|^2$ and combining Eqs. Eqs. equation 26 and equation 27, we have

$$\begin{aligned}
\|\mathbf{z}_t - \nabla F(\mathbf{w}_t)\|^2 &\le (4L_g^2 L^2_{\nabla f_{\lambda_t}} + 2)\|s_t - g(\mathbf{w}_t)\|^2 + 4L^2_{f_{\lambda_t}}\|\mathbf{q}_t - \nabla g(\mathbf{w}_t)\|^2 \\
&\le 4L_\Psi^2\|s_t - g(\mathbf{w}_t)\|^2 + 4L_\Psi^2\|\mathbf{q}_t - \nabla g(\mathbf{w}_t)\|^2 \\
&= 4L_\Psi^2(\|s_t - g(\mathbf{w}_t)\|^2 + \|\mathbf{v}_t - \nabla_w g(\mathbf{w}_t)\|^2 + \|u_t - \nabla_\lambda g(\mathbf{w}_t)\|^2),
\end{aligned}$$

where we define L_Ψ, L_g as the Lipschitz constant of function $F(\mathbf{w})$ and $g(\cdot)$ respectively. This completes the proof. □

Lemma 8. *Run Algorithm 3, by defining* K_b *as the general notation for minibatch size* b_1 *and* b_2, *then for iteration* T *we have*

$$\mathbb{E}\,\|\mathscr{X}_{t+1}\|^2 \leqslant (1-beta_t)^2\mathbb{E}\,\|\mathscr{X}_t\|^2 + 8(1-\beta_t)^2L_\Psi^2\mathbb{E}\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + 6\beta_t^2\sigma^2 + \sigma_{s_t}^2.$$

Proof. We make a convenient assumption here

Assumption 4. Let σ_g, $\sigma_{\nabla g}$ be positive constants and $\sigma^2 = \max\{\sigma_g, \sigma_{\nabla g}\}$. For $i \in \mathcal{D}$, assume that

$$\mathbb{E}[||g(\mathbf{w};\xi_i) - g(\mathbf{w})||^2] \le \sigma_g^2, \quad \mathbb{E}[||\nabla g(\mathbf{w};\xi_i) - \nabla g(\mathbf{w})||^2] \le \sigma_{\nabla g}^2.$$

Since $s_{t+1} = g(\mathbf{w}_{t+1};\xi_i) + (1-\beta_t)(s_t - g(\mathbf{w}_t;\xi_i))) + \zeta_t$, it holds that

$$\begin{aligned}
\mathbb{E}[\|s_{t+1} - g(\mathbf{w}_{t+1})\|^2] &= \mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) + (1-\beta_t)(s_t - g(\mathbf{w}_t;\xi_i))) - g(\mathbf{w}_{t+1}) + \zeta_t\|^2] \\
&\le \mathbb{E}[\|(1-\beta_t)(s_t - g(\mathbf{w}_t)) + \beta_t(g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_{t+1})) \\
&\quad + (1-\beta_t)(g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i)) - (g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t)))\|^2] + \mathbb{E}\zeta_t^2 \\
&= \mathbb{E}[(1-\beta_t)^2\|s_t - g(\mathbf{w}_t)\|^2] + \mathbb{E}[\|\beta_t(g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_{t+1})) \\
&\quad + (1-\beta_t)(g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i)) - (g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t)))\|^2] + \mathbb{E}\zeta_t^2,
\end{aligned}$$

where the last inequality is due to $\mathbb{E}[g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_{t+1})] = 0$.

Noting $\mathbb{E}[\langle g(\mathbf{w}_{t+1};\xi_i) - g(\mathbf{w}_t;\xi_i)), g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t)\rangle] = \mathbb{E}[\|(g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t))\|^2]$ and applying the Lipschitz continuity of $g(\mathbf{w};\xi_i)$, we have

$$\begin{aligned}
&\mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i)) - (g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t))\|^2] \\
&= \mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i))\|^2 + \|(g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t))\|^2 \\
&\quad - 2\langle g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i)), g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t)\rangle] \\
&= \mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i))\|^2 - \|(g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t))\|^2] \\
&\le \mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t;\xi_i))\|^2] \\
&\le L_g^2\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2].
\end{aligned}$$

Combining the above inequality and invoking the Lipschitz continuity of $g(\mathbf{w};\xi_i)$, we have

$$\begin{aligned}
\mathbb{E}[\|s_{t+1} - g(\mathbf{w}_{t+1})\|^2] &\le (1-\beta_t)^2\mathbb{E}[\|s_t - g(\mathbf{w}_t)\|^2] \\
&\quad + 2\beta_t^2\mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_t)\|^2] \\
&\quad + 2(1-\beta_t)^2\mathbb{E}[\|g(\mathbf{w}_{t+1};\xi_i)) - g(\mathbf{w}_{t+1};\xi_i)) - (g(\mathbf{w}_{t+1}) - g(\mathbf{w}_t))\|^2] \\
&\le (1-\beta_t)^2\mathbb{E}[\|s_t - g(\mathbf{w}_t)\|^2] + 2\beta_t^2\sigma^2 \\
&\quad + 2(1-\beta_t)^2L_g^2\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] + \sigma_{\zeta_t}^2.
\end{aligned}$$

□

Lemma 9. *By setting $\beta_t = c\eta_t^2$ and $\eta_t = \frac{k}{(m+t\sigma^2)^{1/3}}$, $m = \max\{2\sigma^2, (16L_\Psi^2 k)^2\}$, $k = \frac{\sigma^{2/3}}{L_\Psi}$ and $c = \frac{\sigma^2}{14L_\psi k^3} + 130L_\Psi^2$. Then running algorithm 3 satisfies*

$$4L_\Psi^4\sum_{t=1}^{T}\eta_t\mathbb{E}[||\mathscr{X}_t||^2] \le \frac{\mathbb{E}[||\mathscr{X}_1||^2]}{\eta_0} - \frac{\mathbb{E}[||\mathscr{X}_{T+1}||^2]}{\eta_T} + \sum_{t=1}^{T}6c^2\eta_t^3\sigma^{-2} + 64L_F^2\Psi_0 + c_{\sigma_{s_t}}\sigma_{s_t}^2.$$

Proof. With $\eta_t = \frac{k}{(w+t\sigma^2)^{1/3}}$, we obtain

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{(m+t\sigma^2)^{1/3} - (m+(t-1)\sigma^2)^{1/3}}{k} \overset{(a)}{\le} \frac{\sigma^2}{3k(m+(t-1)\sigma^2)^{2/3}} \tag{28}$$

$$\overset{(b)}{\le} \frac{\sigma^2}{3k(m/2+t\sigma^2)^{2/3}} \le \frac{\sigma^2}{3k(m/2+t\sigma^2/2)^{2/3}} = \frac{2^{2/3}\sigma^2}{3k(m+t\sigma^2)^{2/3}} \tag{29}$$

$$= \frac{2^{2/3}\sigma^2}{3k^3}\eta_t^2 \overset{(c)}{\le} \frac{2^{2/3}}{12L_\Psi k^3}\eta_t \le \frac{\sigma^2}{7L_\Psi k^3}\eta_t, \tag{30}$$

where the inequality (a) uses the inequality $(x+y)^{1/3} - x^{1/3} \le \frac{yx^{-2/3}}{3}$, the inequality (b) is due to $m \ge 2\sigma^2$, and the inequality (c) is due to $\eta_t \le \frac{1}{4L_\Psi}$.

Noting $\beta_t = c\eta_t^2$ and $0 \leq (1-\beta_t) \leq 1$, by Lemma 8 we have

$$\frac{\mathbb{E}[\|\mathscr{X}_{t+1}\|^2]}{\eta_t} - \frac{\mathbb{E}[\|\mathscr{X}_t\|^2]}{\eta_{t-1}} \leq \left(\frac{(1-\beta_t)^2}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\mathbb{E}[\|\mathbf{z}_t\|^2] + 6c^2\eta_t^3\sigma^2 + \frac{8(1-\beta_t)^2L_\Psi^2}{\eta_t}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2] \tag{31}$$

$$\leq (\eta_t^{-1} - \eta_{t-1}^{-1} - 2c\eta_t)\mathbb{E}[\|\mathbf{z}_t\|^2] + 6c^2\eta_t^3\sigma^2 + \frac{8(1-\beta_t)^2L_\Psi^2}{\eta_t}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2] \tag{32}$$

$$\leq -260L_\Psi^4\eta_t\mathbb{E}[\|\mathbf{z}_t\|^2] + 6c^2\eta_t^3\sigma^2 + \frac{8(1-\beta_t)^2L_\Psi^2}{\eta_t}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2], \tag{33}$$

where the last inequality is due to $\eta_t^{-1} - \eta_{t-1}^{-1} - 2c\eta_t \leq \frac{\sigma^2}{7L_\Psi k^3}\eta_t - 2\left(\frac{1}{14L_\Psi k^3} + 130L_\Psi\right)\eta_t \leq -260L_\Psi^4\eta_t$.

Taking summation of the above equation from 1 to T, we have

$$260L_\Psi^4\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\mathscr{X}_t\|^2] \leq \frac{\mathbb{E}[\|\mathscr{X}_1\|^2]}{\eta_0} - \frac{\mathbb{E}[\|\mathscr{X}_{T+1}\|^2]}{\eta_T} + \sum_{t=1}^{T}6c^2\eta_t^3\sigma^2 + 8L_\Psi^2\sum_{t=1}^{T}\frac{1}{\eta_t}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2]. \tag{34}$$

Using the descent argument leading to Eq. equation 44, with iteration-dependent step sizes, and $\eta_t \leq \eta_1, \forall t \geq 1$, we could also have

$$\frac{1-2\eta_1L_\Psi}{4}\sum_{t=1}^{T}\frac{1}{\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 \leq \sum_{t=1}^{T}\frac{1-2\eta_tL_\Psi}{4\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 \leq \Psi_0 + \sum_{t=1}^{T}\eta_t\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2. \tag{35}$$

Noting $\eta_1L_\Psi \leq \frac{1}{4}$ and invoking Lemma 7, we obtain

$$\sum_{t=1}^{T}\frac{1}{\eta_t}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2] \leq \frac{4}{1-2\eta_1L_\Psi}\left(\Psi_0 + \sum_{t=1}^{T}\eta_t\mathbb{E}[\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2]\right) \tag{36}$$

$$\leq 8\Psi_0 + 8\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2] \tag{37}$$

$$\leq 8\Psi_0 + 32L_\Psi^2\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\mathscr{X}_t\|^2]. \tag{38}$$

Combining the above two inequality, we have

$$4L_\Psi^4\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\mathscr{X}_t\|^2] \leq \frac{\mathbb{E}[\|\mathscr{X}_1\|^2]}{\eta_0} - \frac{\mathbb{E}[\|\mathscr{X}_{T+1}\|^2]}{\eta_T} + \sum_{t=1}^{T}6c^2\eta_t^3\sigma^2 + 64L_\Psi^2\Psi_0 + c_{\sigma_{s_t}}\sigma_{s_t}^2. \tag{39}$$

This complete the proof.

□

By Theorem 1 in Arora et al. (2023), we have the following guarantee:

$$\mathbb{E}[\|\mathbf{v}_{t+1} - \nabla_w g(\mathbf{w}_{t+1})\|^2] \leqslant \frac{\tau_2\eta^2q\Psi_0}{T} + \frac{\eta^3q\tau_2^2\tau_1^2}{2A} + \tau_1^2$$

where we define $A = \frac{\eta}{2} - \frac{L_1\eta^2}{2} - \frac{\eta^3\tau_2^2q}{2}$, $\tau_1^2 = O(\frac{G^2TA_d^2}{q})$ and $\tau_2^2 = O(\frac{L^2}{b_2} + L^2TA_d^2)$, $A_d = \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}$

By the same method as Theorem 2 in Xu et al. (2019), we have the following inequality,

$$\begin{aligned}&\left\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_t - \mathbf{w}_{t+1})\right\|^2 \\ \leqslant& 2\left\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\right\|^2 + \frac{2(\Psi(\mathbf{w}_{t+1}) - \Psi(\mathbf{w}_t))}{\eta_t} + (2L_\Psi^2 + \frac{3L_\Psi}{\eta_t})\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2.\end{aligned}$$

Therefore, all we need to do is to bound the three terms. Actually, the first term can be bounded by Lemma 7 and then, then by the Lipschitz property of function $\Psi(\mathbf{w})$, then it can be reduced to bound the term with $\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2$

Since $F(\mathbf{w})$ is smooth with parameter L_F, then

$$\Psi(\mathbf{w}_{t+1}) \leq \Psi(\mathbf{w}_t) + \langle\nabla\Psi(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L_\Psi}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2. \tag{40}$$

Combining the above two inequalities, we get

$$\langle\mathbf{z}_t - \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{1}{2}(1/\eta - L_\Psi)\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}). \tag{41}$$

That is

$$\begin{aligned}\frac{1}{2}(1/\eta - L_\Psi)\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 &\leq \Psi(\mathbf{w}_t) - \Psi(\mathbf{w}_{t+1}) - \langle\mathbf{z}_t - \nabla\psi(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle &(42)\\ &\leq \Psi(\mathbf{w}_t) - \Psi(\mathbf{w}_{t+1}) + \eta\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2 + \frac{1}{4\eta}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2, &(43)\end{aligned}$$

where the last inequality uses Young's inequality $\langle\mathbf{a}, \mathbf{b}\rangle \leq \|\mathbf{a}\|^2 + \frac{\|\mathbf{b}\|^2}{4}$. Then by rearranging the above inequality and summing it across $t = 1, \cdots, T$, we have

$$\begin{aligned}\sum_{t=1}^{T}\frac{1 - 2\eta L_\Psi}{4\eta}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 &\leq \Psi(\mathbf{w}_1) - \Psi(\mathbf{w}_{T+1}) + \sum_{t=1}^{T}\eta\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2 \\ &\leq \Psi(\mathbf{w}_1) - \inf_{\mathbf{w}\in\mathcal{W}}\Psi(\mathbf{w}) + \sum_{t=1}^{T}\eta\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2 \\ &\leq \Psi_0 + \sum_{t=1}^{T}\eta\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2. &(44)\end{aligned}$$

Thus,

$$\frac{1 - 2\eta_1 L_\Psi}{4}\sum_{t=1}^{T}\frac{1}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq \sum_{t=1}^{T}\frac{1 - 2\eta_t L_\Psi}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq \Psi_0 + \sum_{t=1}^{T}\eta_t\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2.$$

Dividing $\frac{\eta_1}{1/4 - \eta_1 L_\Psi/2}$ on both sides for the above inequality, we have:

$$\sum_{t=1}^{T}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq \frac{1}{1/4 - \eta_1 L_\Psi/2}\left(\eta_1\Psi_0 + \eta_1\sum_{t=1}^{T}\eta_t\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_t)\|^2\right).$$

Therefore, combining the above inequality and choosing the parameter we set before, we can derive that

$$\mathbb{E}\left\|\nabla\Psi(\mathbf{w}_\tau)\right\|^2 \leqslant \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{z}_t - \nabla\Psi(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|^2\right] \leqslant O((\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon})^{2/3})$$

□

E EXPERIMENTAL RESULTS

E.1 COMPUTER RESOURCES

We use AMD Ryzen Threadripper 3970X and 2 NVIDIA GeForce RTX 3090 (24 GB VRAM each, 48 GB total) with RAM: 125 GiB total.

Parameter Setup We will use ResNet20 for our model. To make a fair comparison, we will consider DRO with KL-divergence. The constrained parameter ρ is set to be 0.5 and λ_0 is set to $1e-3$. For all experiments, the batch size is 128. In addition, we set privacy budget $\varepsilon = \{0.5, 1, 5, 10\}$ with $\delta = \frac{1}{n^{1.1}}$.

E.2 TEST AUC ANALYSIS

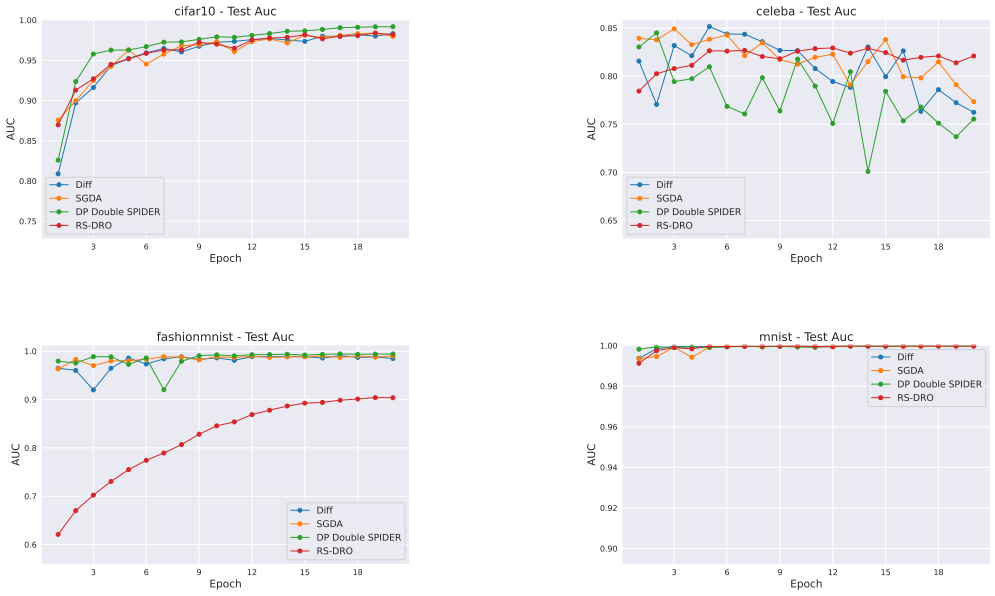


Figure 2: Test AUC Results: The performances of four algorithms on CIFAR10-ST, CelebA, Fashion-MNIST, MNIST-ST respectively

Figure 2 presents the Area Under the Curve (AUC) performance of four algorithms across different datasets. The AUC metric is particularly valuable as it measures the model's ability to distinguish between classes regardless of the classification threshold, providing a comprehensive view of classifier performance.

Across the four datasets (CIFAR10-ST, CelebA, Fashion-MNIST, and MNIST-ST), we observe distinct performance patterns. The MNIST-ST dataset generally exhibits the highest AUC scores, which is expected given its relatively simple binary classification task with well-separated digit classes. Fashion-MNIST presents a more challenging scenario due to the visual similarity between certain clothing categories, resulting in comparatively lower AUC values. The CelebA dataset, dealing with facial attribute classification, demonstrates intermediate performance, while CIFAR10-ST shows varying convergence patterns depending on the algorithm employed.

The comparative analysis reveals that different algorithms exhibit varying degrees of robustness across datasets. Some algorithms demonstrate consistent performance across all datasets, suggesting good generalization capabilities, while others show dataset-specific strengths. The convergence behavior also varies, with some algorithms achieving stable AUC values earlier in training, while others require more iterations to reach optimal performance.

E.3 Test F1 Score Analysis

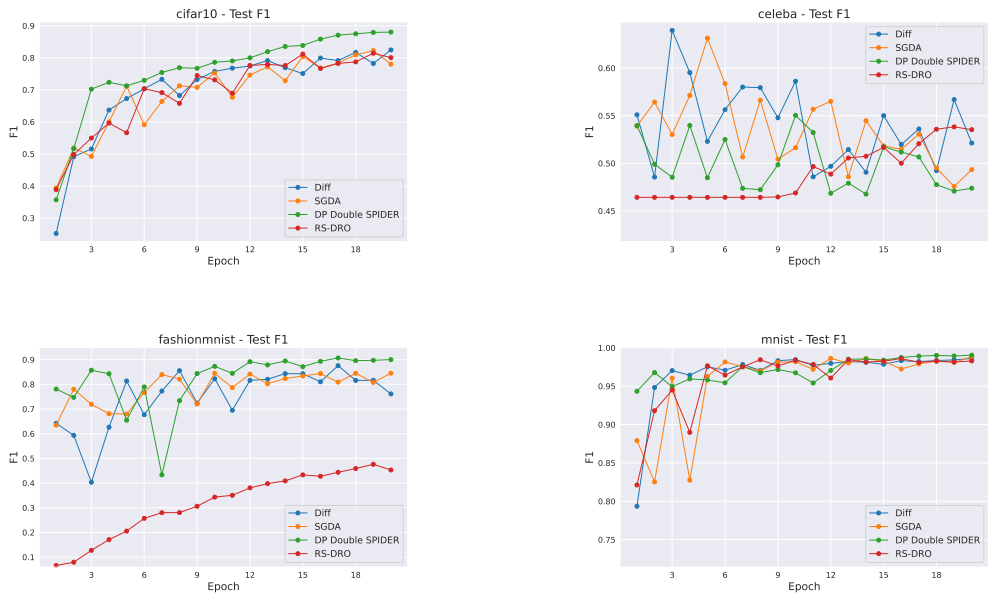


Figure 3: Test F1 Score Results: The performances of four algorithms on CIFAR10-ST, CelebA, Fashion-MNIST, MNIST-ST respectively

Figure 3 presents the F1 score results, which provide a harmonic mean of precision and recall. This metric is particularly important when dealing with imbalanced datasets or when both false positives and false negatives carry significant costs.

The F1 scores reveal nuanced performance differences that may not be immediately apparent from accuracy alone. For MNIST-ST, the F1 scores closely track accuracy values, confirming robust performance across both metrics. However, for potentially imbalanced scenarios in CelebA (where certain facial attributes may be rare), the F1 score provides critical insights into the model's ability to correctly identify minority class instances without sacrificing precision.

On Fashion-MNIST F1 scores indicate how well algorithms balance precision and recall when distinguishing between similar clothing categories. Lower F1 scores compared to accuracy might suggest that while overall accuracy is reasonable, the model struggles with specific class pairs, leading to either high false positive or false negative rates for certain categories.

CIFAR10-ST F1 scores demonstrate the algorithms' effectiveness in handling the complexity of natural images. The relationship between F1 scores and accuracy across different algorithms reveals which approaches better balance the trade-off between precision and recall. Algorithms showing F1 scores close to their accuracy values indicate balanced performance, while larger gaps suggest potential issues with either over-prediction or under-prediction of certain classes.

Overall, the F1 score analysis complements the AUC and accuracy metrics by highlighting the algorithms' ability to maintain balanced performance across different evaluation criteria, which is essential for practical deployment scenarios.

E.4 Hyperparameter Sensitivity

We conduct additional hyperparameter sensitivity experiments on CIFAR10-ST. CIFAR10-ST is used because it is the most challenging benchmark in our experiments and because the performance differences between methods are most visible on this dataset. Unless otherwise stated, each configuration is averaged over three random seeds, and we report mean $\pm$ standard deviation of test accuracy.

The goal of this appendix is to examine whether the additional algorithm-specific parameters introduced by DP Double-SPIDER and DP Recursive-SPIDER create a practical tuning burden. We

Table 5: Pairwise clipping sensitivity of DP Double-SPIDER on CIFAR10-ST. Top: vary the dual-side clipping thresholds $C_1 = C_2$ while fixing $C_3 = C_4 = 1$. Bottom: vary the primal-side clipping thresholds $C_3 = C_4$ while fixing $C_1 = C_2 = 1$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

Setting		Clip value				
Group	ε	0.25	0.5	1.0	2.0	4.0
Vary $C_1 = C_2$	1.0	45.86 ± 0.04	45.86 ± 0.05	45.87 ± 0.05	45.87 ± 0.04	45.88 ± 0.04
	5.0	46.03 ± 0.24	46.05 ± 0.19	46.05 ± 0.24	46.04 ± 0.25	46.07 ± 0.20
	10.0	46.07 ± 0.13	46.04 ± 0.13	46.05 ± 0.12	46.05 ± 0.16	46.06 ± 0.14
Vary $C_3 = C_4$	1.0	45.87 ± 0.06	45.87 ± 0.05	45.87 ± 0.05	45.87 ± 0.05	45.87 ± 0.04
	5.0	46.05 ± 0.20	46.05 ± 0.24	46.05 ± 0.24	46.06 ± 0.23	46.06 ± 0.22
	10.0	46.03 ± 0.02	46.10 ± 0.07	46.05 ± 0.12	46.09 ± 0.14	46.09 ± 0.15

focus on the parameters that are specific to the proposed variance-reduced private algorithms and that could reasonably raise sensitivity concerns: clipping thresholds, the SPIDER phase size, and the learning-rate multiplier. We do not perform exhaustive ablations over all implementation constants. Some quantities, such as the privacy budget (ε, δ), define the privacy requirement rather than a tuning choice. Other quantities, such as the model architecture, batch size, DRO radius ρ, and lower-bound parameter λ_0, are fixed for fair comparison across algorithms or follow common settings in related DP/DRO experiments. The Gaussian noise scales are also not independent hyperparameters, since they are determined by the privacy accounting once the clipping thresholds, sampling scheme, and privacy parameters are fixed.

All experiments use KL-DRO with radius $\rho = 0.5$, lower-bound parameter $\lambda_0 = 10^{-3}$, batch size 128, and privacy failure probability $\delta = n^{-1.1}$. Non-swept algorithmic parameters are fixed to their default values.

The sensitivity study requires many repeated runs, so the experiments below use controlled protocols. For DP Double-SPIDER and the clipping comparison with DP-SGDA and PrivateDiff, we use a 20-epoch ResNet20 pretraining protocol on CIFAR10-ST, whose test accuracy is approximately 46%. For DP Recursive-SPIDER, we use the training framework of the main paper: full CIFAR10-ST with $n = 25{,}500$, ResNet20 with width factor 0.5, a 12-epoch class-balanced pretrain with test accuracy around 55.5%, a Ψ_0 warm start, and theory-derived values of T, b_1, b_2, q with the same implementation caps used in the main experiments. Therefore, absolute accuracies should not be compared across the DP Double-SPIDER and DP Recursive-SPIDER sensitivity blocks. The relevant comparison is the variation across hyperparameter values within each controlled protocol.

The main paper reports the uniform clipping sweep for DP Double-SPIDER, where all four clipping thresholds are tied as $C_1 = C_2 = C_3 = C_4 = C$. That result directly addresses the main concern that multiple clipping thresholds may require careful tuning. To avoid overlap, this appendix does not repeat that table. Instead, it reports the remaining sensitivity studies: pairwise clipping, random independent clipping, comparison with baselines, and DP Recursive-SPIDER sensitivity to clipping scale, phase size, and learning-rate multiplier.

DP Double-SPIDER Clipping Sensitivity DP Double-SPIDER contains four clipping thresholds. The thresholds C_1 and C_2 are used for the refresh and recursive update of the dual-variable gradient estimator, while C_3 and C_4 are used for the corresponding primal-variable gradient estimator. The main paper shows that these four thresholds can be tied to a single value without meaningful performance loss. Here, we further test whether the dual-side and primal-side clipping thresholds need to be tuned separately, whether heterogeneous random clipping values affect performance, and whether DP Double-SPIDER is more sensitive to clipping than the baselines.

Pairwise clipping independence. We first test whether the dual-side and primal-side clipping thresholds need to be tuned separately. In the first group, we vary $C_1 = C_2$ while fixing $C_3 = C_4 = 1$. In the second group, we vary $C_3 = C_4$ while fixing $C_1 = C_2 = 1$. The results are reported in Table 5.

Table 6: Random independent clipping stress test for DP Double-SPIDER on CIFAR10-ST with $\varepsilon = 5$. Each configuration samples C_1, C_2, C_3, C_4 independently from $[0.5, 2.0]$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

Config	C_1	C_2	C_3	C_4	Accuracy (%)
1	1.46	0.54	0.91	0.83	46.05 ± 0.22
2	1.60	1.52	1.84	0.63	46.22 ± 0.25
3	1.13	0.54	0.83	1.26	46.03 ± 0.24
4	0.54	0.80	1.47	1.32	46.06 ± 0.21
5	0.83	1.38	1.71	0.51	46.16 ± 0.25
6	1.71	1.55	1.01	0.73	46.11 ± 0.23
7	1.94	1.00	0.64	0.65	46.08 ± 0.22
8	1.77	1.41	1.71	1.59	46.08 ± 0.24
9	1.30	1.96	1.07	1.33	46.03 ± 0.21
10	1.74	1.43	1.79	1.37	46.09 ± 0.24
		Across configurations			mean $= 46.09$, std $= 0.06$, range $= 0.19$

The pairwise sweep shows that the dual-side and primal-side clipping thresholds do not require separate tuning. Across all rows in Table 5, the largest range of mean accuracy is only 0.07 percentage points. Varying $C_1 = C_2$ while fixing $C_3 = C_4 = 1$ changes accuracy by at most 0.04 percentage points for $\varepsilon = 5$, and varying $C_3 = C_4$ while fixing $C_1 = C_2 = 1$ changes accuracy by at most 0.07 percentage points for $\varepsilon = 10$. These variations are smaller than the seed-level standard deviations. Thus, the two sides of the estimator do not behave like independent fragile tuning knobs. This supports the practical implementation rule of using a common clipping value for all four thresholds.

Random independent clipping stress test. The pairwise sweep still varies the clipping thresholds in a structured way. To test a more heterogeneous setting, we sample each clipping threshold independently from $[0.5, 2.0]$ and evaluate ten random configurations at $\varepsilon = 5$. This setting allows the four clipping values to differ from each other rather than being tied.

The random stress test further confirms clipping robustness. Across ten heterogeneous clipping configurations, the mean accuracy is 46.09%, the cross-configuration standard deviation of the reported means is only 0.06%, and the range is 0.19 percentage points. Therefore, even when the four clipping thresholds are independently sampled rather than tied, the final accuracy remains stable. This result provides a stronger test than the tied clipping sweep: it shows that the method is not only robust to the scale of a common clipping value, but also robust to moderate heterogeneity among the four clipping thresholds.

Comparison with baselines. We next compare clipping sensitivity with DP-SGDA and PrivateDiff. For each method, all clipping thresholds are tied to a common value $C \in \{0.5, 1, 2, 4\}$. This comparison directly tests whether having more clipping thresholds in the proposed method makes it harder to tune than the baselines.

Table 7 shows that DP Double-SPIDER is less sensitive to clipping than the baselines, despite having more clipping variables in its theoretical statement. For DP Double-SPIDER, the range across clipping values is 0.00, 0.03, and 0.09 percentage points for $\varepsilon = 1, 5, 10$, respectively. In contrast, the corresponding ranges are 3.91, 2.23, and 2.09 percentage points for DP-SGDA, and 5.66, 1.70, and 1.09 percentage points for PrivateDiff. Therefore, the number of clipping constants in the algorithmic description does not translate into practical tuning difficulty. Empirically, DP Double-SPIDER is more robust to the clipping scale than both baseline methods.

DP Recursive-SPIDER Sensitivity We next evaluate the sensitivity of DP Recursive-SPIDER. Algorithm 3 contains five clipping thresholds: C_1–C_4 are used for the recursive gradient estimators, and C_5 is used for the value tracker s_t. In addition, the method has a phase size q, which controls the frequency of full-batch refreshes, and a learning-rate parameter. All experiments in this subsection are conducted on CIFAR10-ST with $\varepsilon = 5$.

Table 7: Clipping sensitivity comparison on CIFAR10-ST. For each method, all clipping thresholds are tied to a common value C. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

ε	Method	$C = 0.5$	$C = 1.0$	$C = 2.0$	$C = 4.0$
1.0	DP Double-SPIDER	45.87 ± 0.04	45.87 ± 0.05	45.87 ± 0.04	45.87 ± 0.04
1.0	DP-SGDA	29.09 ± 2.20	28.39 ± 1.81	29.51 ± 1.28	25.60 ± 1.71
1.0	PrivateDiff	27.49 ± 0.91	25.80 ± 1.51	25.37 ± 1.16	21.83 ± 0.98
5.0	DP Double-SPIDER	46.05 ± 0.17	46.05 ± 0.24	46.07 ± 0.24	46.08 ± 0.22
5.0	DP-SGDA	30.51 ± 2.59	29.44 ± 2.46	30.75 ± 2.03	28.52 ± 1.15
5.0	PrivateDiff	28.24 ± 1.19	27.41 ± 2.41	26.54 ± 0.87	27.83 ± 1.91
10.0	DP Double-SPIDER	46.02 ± 0.06	46.05 ± 0.12	46.10 ± 0.15	46.11 ± 0.16
10.0	DP-SGDA	31.33 ± 0.76	29.24 ± 1.85	29.35 ± 3.76	29.32 ± 3.38
10.0	PrivateDiff	29.09 ± 2.11	28.00 ± 0.64	28.44 ± 1.67	28.52 ± 1.71

Table 8: DP Recursive-SPIDER under uniform clipping scale $G = L = C$ on CIFAR10-ST with $\varepsilon = 5$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

C	0.5	1.0	2.0	4.0
Accuracy (%)	55.97 ± 1.65	54.81 ± 2.63	53.62 ± 3.43	53.53 ± 3.29

Clipping-scale sensitivity. In the implementation of DP Recursive-SPIDER, the clipping thresholds are controlled through the gradient-bound and smoothness constants. We therefore tie them by setting $G = L = C$ and sweep $C \in \{0.5, 1, 2, 4\}$. The results are reported in Table 8.

The mean accuracy ranges from 53.53% to 55.97%, giving a range of 2.44 percentage points and a cross-C standard deviation of 1.153% over the reported means. The results show moderate sensitivity to the clipping scale, especially as C increases. However, the method remains stable over an eight-fold sweep from $C = 0.5$ to $C = 4.0$: no tested value causes training collapse, and the variation is comparable to the seed-level standard deviations. Thus, DP Recursive-SPIDER benefits from a reasonable default clipping scale, but it does not require precise fine tuning of the clipping threshold.

Phase-size sensitivity. The phase size q controls how often DP Recursive-SPIDER performs a refresh step rather than a recursive variance-reduced update. The theory-derived default in this setting is $q_0 = 5$. We test $q \in \{2, 5, 10, 20\}$, corresponding to approximately half, one, two, and four times the default value.

The phase-size sweep is highly stable. The range of mean accuracy is 0.65 percentage points, and the cross-q standard deviation of the reported means is 0.275%. The default value $q_0 = 5$ achieves the highest mean accuracy, but the differences between $q = 2, 5, 10, 20$ are small relative to the seed-level standard deviations. This indicates that the refresh period is not a delicate hyperparameter. In practice, the theory-derived value can be used without extensive tuning.

Learning-rate multiplier sensitivity. Finally, we vary the learning-rate multiplier. Let η_0 denote the default step size. We test the effective step size $s\eta_0$ with $s \in \{0.5, 1, 2\}$.

The learning rate has a larger effect than the phase size. Halving the default step size gives the highest mean accuracy, while doubling the default step size lowers the mean accuracy. The range of the reported means is 2.72 percentage points, and the cross-s standard deviation is 1.392%. Nevertheless, all tested settings produce valid training behavior, and the default setting $s = 1$ remains close to the best tested configuration within the observed seed-level variability. Thus, the learning rate should be set to a reasonable default, but the method does not require delicate fine tuning.

The sensitivity results support the claim that the additional parameters introduced by the proposed algorithms do not create a practical fine-tuning burden. For DP Double-SPIDER, the evidence is particularly strong. The main paper reports that tying all four clipping thresholds to one common value leads to almost unchanged accuracy over a two-order-of-magnitude sweep. The additional

Table 9: DP Recursive-SPIDER under varying phase size on CIFAR10-ST with $\varepsilon = 5$. The default value is $q_0 = 5$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

Phase-size setting	Actual q	Accuracy (%)
$0.5q_0$	2	55.20 ± 2.68
q_0	5	55.38 ± 2.15
$2q_0$	10	55.03 ± 2.30
$4q_0$	20	54.73 ± 2.57

Table 10: DP Recursive-SPIDER under learning-rate multiplier s on CIFAR10-ST with $\varepsilon = 5$, where the effective step size is $s\eta_0$. We report test accuracy (%) as mean $\pm$ standard deviation over three random seeds.

s	0.5	1.0	2.0
Accuracy (%)	56.24 ± 1.41	55.38 ± 2.15	53.52 ± 3.53

experiments in this appendix further show that the dual-side and primal-side clipping thresholds do not need to be tuned separately, that independently sampled heterogeneous clipping thresholds produce stable accuracy, and that DP Double-SPIDER is less sensitive to the clipping scale than DP-SGDA and PrivateDiff.

These results directly address the concern that DP Double-SPIDER has "too many parameters." Although the theoretical algorithm contains four clipping thresholds, the experiments show that these thresholds can be treated as one practical parameter. Moreover, compared with the baselines, the proposed method is more robust to the clipping value. Therefore, the number of clipping constants in the theoretical description does not imply a larger tuning burden in practice.

For DP Recursive-SPIDER, the phase size q is highly stable around the theory-derived default. The clipping scale and learning-rate multiplier have larger effects than q, but the method remains stable over multi-fold sweeps. These results support a simple practical implementation strategy: use tied clipping thresholds, use the theory-derived phase size, and fix standard training choices consistently across methods.

We also emphasize that not every parameter appearing in the algorithms requires a separate ablation. Some quantities are part of the privacy requirement, such as (ε, δ). Some are fixed for fair comparison, such as the model architecture, batch size, DRO radius ρ, and lower-bound parameter λ_0. Others, especially the Gaussian noise scales, are determined by the privacy analysis once the clipping thresholds, sampling scheme, and privacy parameters are specified. The sensitivity study therefore focuses on the algorithm-specific choices that could reasonably raise concerns about tuning. Under this targeted evaluation, the proposed methods show stable performance and do not require exhaustive hyperparameter search.