DERANDOMIZED ONLINE-TO-NON-CONVEX CONVERSION FOR STOCHASTIC WEAKLY CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online-to-non-convex conversion (O2NC) is an online updates learning framework for producing Goldstein (δ, ϵ)-stationary points of non-smooth non-convex functions with optimal oracle complexity $\mathcal{O}(\delta^{-1}\epsilon^{-3})$. Subject to auxiliary *random interpolation or scaling*, O2NC recapitulates the stochastic gradient descent with momentum (SGDM) algorithm popularly used for training neural networks. Randomization, however, introduces deviations from practical SGDM. So a natural question arises: Can we derandomize O2NC to achieve the same optimal guarantees while resembling SGDM? On the negative side, the general answer is *no* due to the impossibility results of Jordan et al. (2023), showing that no dimension-free rate can be achieved by deterministic algorithms. On the positive side, as the primary contribution of the present work, we show that O2NC can be naturally derandomized for *weakly convex* functions. Remarkably, our deterministic algorithm converges at an optimal rate as long as the weak convexity parameter is no larger than $\mathcal{O}(\delta^{-1}\epsilon^{-1})$. In other words, the stronger stationarity is expected, the higher non-convexity can be tolerated by our optimizer. Additionally, we develop a periodically restarted variant of our method to allow for more progressive update when the iterates are far from stationary. The resulting algorithm, which corresponds to a momentum-restarted version of SGDM, has been empirically shown to be effective and efficient for training ResNet and ViT networks.

1 INTRODUCTION

Classic machine learning optimization methods often rely crucially on convexity and/or smoothness assumptions to guarantee the convergence to optima (Nesterov et al., 2018; Bubeck, 2015; Snyman, 2005). However, many modern large-scale machine learning models, such as residual neural networks and transformers (He et al., 2016; Vaswani et al., 2017), involve non-convex and non-smooth objective functions. These models achieve state-of-the-art performance precisely thanks to their capability to learn highly complex, nonlinear hidden representations in data. With such widespread use, efficient non-convex non-smooth optimization algorithms are of fundamental interest.

Specifically, this paper is concerned with stochastic gradient algorithms for solving the following expected risk minimization problem ubiquitous in statistical learning:

$$\min_{w\in\mathbb{R}^d} R(w) := \mathbb{E}_{Z\sim\mathcal{D}}[\ell(w; Z)], \tag{1}$$

where $\ell : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}^+$ is a non-negative loss function whose value $\ell(w; z)$ measures the loss evaluated at $z \in \mathcal{Z}$ with parameter $w \in \mathbb{R}^d$ and $\mathcal{D}$ represents a distribution over the measurable set $\mathcal{Z}$. We consider the setting where the loss ℓ is Lipschitz continuous with respect to its first argument, yet potentially neither convex nor smooth. In contrast to the smooth counterpart, finding an ϵ-stationary point (or even a neighborhood around it) of a non-smooth objective in 1 is generally intractable (Zhang et al., 2020; Kornowski & Shamir, 2022). This intractability motivates the employment of Goldstein (δ, ϵ)-stationarity (see Definition 1) as a notion of approximate convergence for non-convex non-smooth functions (Zhang et al., 2020). The study of efficient algorithms with finite-time complexity guarantees for finding (δ, ϵ)-stationary points has since received ever emerging interests in non-smooth, non-convex optimization (Zhang et al., 2020; Davis et al., 2022; Cutkosky et al., 2023; Jordan et al., 2023; Tian & So, 2024; Kong & Lewis, 2025).

Pioneered by Cutkosky et al. (2023), the online-to-non-convex conversion (O2NC) method identifies (δ, ϵ)-stationary points of 1 using $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ calls to a stochastic gradient oracle, which achieves the optimal first-order complexity. As outlined in Algorithm 1, O2NC essentially converts an online convex learner (in light red) to a stochastic gradient optimizer (in light blue). To be more precise, it recursively updates the increments $\Delta_n := w_n - w_{n-1}$ between two adjacent iterates via invoking an online convex optimization (OCO) algorithm $\mathcal{A}$ to minimize the regret $\sum_{n=1}^{N}\langle \hat{g}_n, \Delta_n - \Delta\rangle$, where the stochastic subgradient $\hat{g}_n$ is evaluated at a random intermediate state $v_n = w_{n-1} + s_n\Delta_n$ with a uniform $s_n \in [0, 1]$. The optimal oracle complexity can be implied by any instantiations of $\mathcal{A}$ with optimal regret bound, such as online gradient descent (OGD) (Zinkevich, 2003).

In addition to theoretical optimality, another attractiveness of the O2NC framework lies in its potential power for recovering stochastic momentum-based optimizers commonly used in training neural networks. Indeed, subject to the random interpolation on the iterates, O2NC equipped with projected OGD turns out to be a clipped variant of SGD with momentum (SGDM) Cutkosky et al. (2023). Alternatively, Zhang & Cutkosky (2024) proposed the Exponentiated O2NC (E-O2NC) framework with exponential random scaling on the updates, which almost exactly recovers the standard SGDM when applied with unconstrained online mirror descent (OMD) (Beck & Teboulle, 2003).

Algorithm 1: Online-to-non-convex Conversion (O2NC) (Cutkosky et al., 2023)

Input : OCO algorithm $\mathcal{A}$, $K, T \in \mathbb{N}$, initial point w_0 and increment Δ_1. Set $N = K \times T$.
for $n = 1, ..., N$ **do**
  `/* Stochastic gradient optimizer */`
  Update $w_n = w_{n-1} + \Delta_n$;
  Compute random interpolation $v_n = w_{n-1} + s_n\Delta_n$, where $s_n \sim \text{Unif}([0, 1])$;
  Randomly sample $z_n \sim \mathcal{D}$ and obtain $\hat{g}_n \in \partial\ell(v_n; z_n)$;
  `/* Online learning of increment */`
  Send the linear loss $\langle \hat{g}_n, \Delta\rangle$ to $\mathcal{A}$ and receive the next increment Δ_{n+1} from $\mathcal{A}$
end
Set $w_t^{(k)} = w_{(k-1)T+t}$, $\forall k \in [K], t \in [T]$, and $\bar{w}^{(k)} = \frac{1}{T}\sum_{t=1}^{T} w_t^{(k)}$.
Output: $\bar{w}_T \sim \text{Unif}(\{\bar{w}^{(k)} : k \in [K]\})$.

Despite the promise of O2NC in justifying the effectiveness/efficiency of SGDM-style optimizers, the recovered algorithmic resemblance will inheritably be subject to some auxiliary randomization operations, say uniform interpolation on iterates or exponential scaling on increments. However, these randomization components are seldom, if not never, employed in the practical implementations of SGDM. Such a fundamental gap motivates us to address the following question:

Can the O2NC technique be derandomized to still achieve optimal dimension-free guarantees and close resemblance to SGDM in the non-smooth and non-convex setting?

The general answer to the above question is unfortunately *negative* as it has been shown by Jordan et al. (2023, Theorem 1) that in the worst case no dimension-free rate can be achieved by deterministic algorithms. Fortunately, on the positive side, we will show in this paper that for a broad class of the so-called weakly convex functions, it is indeed possible to develop deterministic variants of O2NC for identifying Goldstein-style stationary solutions with optimal rates.

1.1 Overview of our results

Our main contribution is a derandomized O2NC framework (Algorithm 2) for solving the stochastic optimization problem 1 with a ρ-weakly convex risk function, i.e., $R(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex. Inspired by the original O2NC, the main development here is using the definition of weak convexity to naturally convert the optimization of iterates w_n to the online learning of increments Δ_n over quadratic losses $\langle \hat{g}_n, \cdot\rangle + \frac{\gamma}{2}\|\cdot\|^2$ for some $\gamma \geq \rho$. Differently, instead of evaluating the gradients at a random intermediate point v_n lying between the two iterates w_n and w_{n-1}, our algorithm exactly evaluates the gradients at each iterate $w_n = w_{n-1} + \Delta_n$, and thus is deterministic (of course, up to the stochastic estimation of gradients). Concretely, we propose two optional online learners for updating the increments Δ_n, which are summarized below:

Derandomized O2NC with clipped OGD (Section 3.2). The first option is a naive projected OGD algorithm under a suitable ball constraint. The resulting algorithm can be interpreted as a clipped version of SGDM but without needing additional random interpolations. Our convergence analysis result (Corollary 1) shows that the proposed deterministic method identifies a (δ, ϵ)-stationary point with $\mathcal{O}\left(\delta^{-1}\epsilon^{-3} + \rho^3\delta^2 + \delta^{-1}\right)$ calls to stochastic gradient oracle, which is dominated by the optimal rate $\delta^{-1}\epsilon^{-3}$. Strikingly, the weak-convexity parameter ρ does not appear in such a dominant component, and it is allowed to scale as large as $\mathcal{O}(\delta^{-1}\epsilon^{-1})$ in its involved component before matching the optimal rate. This phenomenon indicates that the smaller (δ, ϵ) are demanded, the higher non-convexity can be tolerated by our optimizer for achieving optimal complexity.

Derandomized O2NC with periodically restarted OGD (Section 3.3). Like in the original O2NC, our first option of OGD under explicit ball constraint is expected to be over conservative for increments update, and it is also impractical from the perspective of SGDM implementation. To address this issue, as the second option, we further introduce a novel periodically restarted OGD procedure which is characterized by *resetting the increments to zero after a period of iteration*. The resulting method is almost identical to the standard SGDM algorithm, with the only difference that the momentum update is now enforced to start over again periodically. Under a novel notion of (μ, ϵ)-regularized stationarity (see Definition 2), which is equivalent to the Goldstein stationarity, we establish in Corollary 2 that the proposed deterministic and unconstrained O2NC algorithm converges with a composite rate $\mathcal{O}\left(\mu^{1/2}\epsilon^{-7/2} + \rho^{7/3}\mu^{-2/3} + \mu^{1/2}\right)$, in which $\rho = \mathcal{O}(\mu^{1/2}\epsilon^{-3/2})$ is allowable without dominating the optimal component of $\mu^{1/2}\epsilon^{-7/2}$. Coupled with our theoretical findings, we have performed a series of numerical experiments on benchmark tasks to verify that the proposed momentum-restarted SGDM variant achieves performance comparable to or even superior to the standard SGDM in training deep residual networks and vision transformers (Section 4).

1.2 Related work

Our contribution is situated within a broad landscape of non-smooth and non-convex optimization. Below we provide an incomplete review on some prior works most closely related to ours.

Non-smooth optimization. The groundwork for non-smooth optimization date back to the early developments of Clarke (1975); Goldstein (1977). There is a rich history of research on asymptotic analysis for non-smooth optimization problems (Benaïm et al., 2005; Davis et al., 2020; Bolte & Pauwels, 2021). Despite these advances, non-asymptotic guarantees have long been left mysterious for generic non-smooth problems. Recently, Zhang et al. (2020) revolutionized the study on subgradient algorithms with finite-time complexity for finding Goldstein stationary points, which has since attracted much attention (Davis et al., 2022; Kornowski & Shamir, 2022; Cutkosky et al., 2023; Jordan et al., 2023; Kornowski & Shamir, 2024). Particularly, inspired by the idea of online-to-batch conversion Cesa-Bianchi et al. (2004), Cutkosky et al. (2023) introduced the O2NC framework which for the first time established the optimal rate for stochastic non-smooth non-convex optimization. By instantiating different online learners within this framework, it is possible to recover several popular optimizers: SGDM corresponds to choosing online mirror descent (Zhang & Cutkosky, 2024), the Adam optimizer (Kingma & Ba, 2015) corresponds to a variant of follow-the-regularized leader (Ahn & Cutkosky, 2024), and very recently Ahn et al. (2025) showed that a generalized O2NC framework captures the schedule-free SGD (Defazio et al., 2024).

Stochastic weakly convex optimization. The class of weakly convex functions, first introduced in English by Nurminskii (1973), is broad and easy to identify in the sense that it encompasses all composition forms $h \circ c$ of convex functions and smooth maps. For this class of problem, a vast body of asymptotic convergence results have been established for stochastic optimization algorithms (Ermol'ev & Norkin, 1998; Duchi & Ruan, 2018). The finite-time non-asymptotic rates, however, remained largely open until recently a series of breakthrough results were achieved by Davis & Grimmer (2019); Davis & Drusvyatskiy (2019); Mai & Johansson (2020), showing that various SGD/SGDM algorithms can achieve the $\mathcal{O}(\epsilon^{-4})$ optimal rate for producing an ϵ-stationary point of the Moreau envelope of objectives. In terms of the variational analysis, several different notions of approximate subdifferentials were analyzed and compared for weakly convex functions (van Ackooij et al., 2024). In practice, weakly convex optimization has found rich applications in deep learning, signal processing and control theory (see, e.g., Duchi & Ruan, 2018; Davis & Drusvyatskiy, 2019; Drusvyatskiy & Paquette, 2019; Pougkakiotis & Kalogerias, 2023, and the references therein).

2 PRELIMINARIES

Let us begin by formally introducing some notation, key assumptions, and preliminary results on non-convex and non-smooth optimization.

2.1 NOTATION AND ASSUMPTIONS

Notation. Throughout this paper, we denote $\|\cdot\|$ as the Euclidean norm, and $\langle\cdot,\cdot\rangle$ as the Euclidean inner product. For a vector set $V \subseteq \mathbb{R}^d$, we denote $\text{dist}(0, V) := \inf_{v\in V}\|v\|$ and $\text{conv}\{V\}$ the convex hull of V. For any positive integer N, we abbreviate $[N] = \{1, ..., N\}$. The symbol $\mathbb{B}_\delta(w)$ denotes the closed ball of radius δ centered on w, and $\texttt{clip}_D(w) := w \min\left\{1, \frac{D}{\|w\|}\right\}$ denotes the Euclidean projection operator associated with the constraint of $\mathbb{B}_D(0)$. For a pair of functions $f, f' \geq 0$, we use $f = \mathcal{O}(f')$ to denote $f \leq cf'$ for some universal constant $c > 0$.

We say that a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is G-Lipschitz continuous if $|f(w) - f(w')| \leq G\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$. The Clarke subdifferential (Clarke, 1990) of a non-smooth function f at $w \in \mathbb{R}^d$ is denoted by $\partial f(w)$. Recall that f is said to be ρ-weakly convex if the quadratically regularized function $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex, or equivalently

$$f(w) \geq f(w') + \langle g', w - w'\rangle - \frac{\rho}{2}\|w - w'\|^2, \quad \forall w, w' \in \mathbb{R}^d, g' \in \partial f(w').$$

A prominent source of weakly convex functions is the composite form $f(x) = h(c(x))$ where h is a convex and G-Lipschitz continuous function, and c is a smooth mapping with a L-Lipschitz Jacobian. These composite functions are neither smooth nor convex, but rather GL-weakly convex (Davis & Drusvyatskiy, 2019). A concrete example, as considered in our experimental study, is neural networks equipped with smooth activation functions (e.g., softplus and GeLU): the loss function is of composite form $h \circ c$ where h is a convex top-layer predictor (e.g., cross-entropy loss) and c is a smooth hierarchical feature mapping. For more examples of weakly convex functions, we refer interested readers to Davis & Drusvyatskiy (2019); Asi & Duchi (2019).

The Moreau envelope (Rockafellar, 1997) of a ρ-weakly convex function f with parameter $\lambda \in (0, \rho^{-1})$ is defined by $f_\lambda(w) := \inf_{u\in\mathbb{R}^d}\left\{f(u) + \frac{1}{2\lambda}\|u - w\|^2\right\}$, and the associated proximal mapping operator is written by $\text{prox}_{\lambda f}(w) := \arg\min_{u\in\mathbb{R}^d}\left\{f(u) + \frac{1}{2\lambda}\|u - w\|^2\right\}$. The following standard result (see, e.g., Böhm & Wright, 2021) summarizes the continuously differential property of the Moreau envelope functions.

Lemma 1. *Let f be a ρ-weakly convex and $\lambda \in (0, \rho^{-1})$ be a scalar. Then the Moreau envelope f_λ is continuously differentiable with gradient $\nabla f_\lambda(w) = \frac{1}{\lambda}\left(w - prox_{\lambda f}(w)\right) \in \partial f(prox_{\lambda f}(w))$, which is L-Lipschitz continuous with parameter $L = \max\left\{\lambda^{-1}, \frac{\rho}{1-\rho\lambda}\right\}$.*

Assumptions. We next impose some basic assumptions on the loss and risk functions in problem 1 for stochastic gradient-based optimization.

Assumption 1. *For any $z \in \mathcal{Z}$, the loss function $\ell(\cdot; z)$ is G-Lipschitz with respect to its first argument. Moreover, the expected risk function R is ρ weakly-convex.*

Assumption 2 (Stochastic oracle)**.** *For each $w \in \mathbb{R}^d$, it holds that $\ell'(w) = \mathbb{E}_{Z\sim\mathcal{D}}[\ell'(w; Z)] \in \partial R(w)$, where $\ell'(w; z) \in \partial\ell(w; z)$ for any $z \in \mathcal{Z}$.*

Also, we assume that $R^* = \min_{w\in\mathbb{R}^d} R(w) > -\infty$ and abbreviate $\Delta R_0 := R(w_0) - R^*$.

2.2 REGULARIZED GOLDSTEIN STATIONARITY CRITERION

For generic non-smooth non-convex functions, the Goldstein (δ, ϵ)-stationarity (Goldstein, 1977) is a standard criterion for convergence analysis, as defined below.

Definition 1 ((δ, ϵ)-Stationarity)**.** *The Goldstein δ-subdifferential of a Lipschitz function f at a point $w \in \mathbb{R}^d$ is the convex hull of all Clarke subgradients at points in a δ-ball around w, i.e.,*

$$\partial_\delta f(w) := conv\left\{\cup_{v\in\mathbb{B}_\delta(w)}\partial f(v)\right\}.$$

A point w is called a (δ, ϵ)-stationary point if dist $(0, \partial_\delta f(w)) \leq \epsilon$.

Despite that the finite-time guarantees on the (δ, ϵ)-stationarity have been well studied in the original O2NC (Cutkosky et al., 2023), the corresponding analysis essentially needs the online increments update to be explicitly constrained inside a tiny ball of radius $\delta\epsilon^2$ which could be over conservative. Inspired by Zhang & Cutkosky (2024), we next introduce a novel regularized version of (δ, ϵ)-stationarity which obviates the need for such explicit constraints, and thus allows for potentially more aggressive update of increments. Given a subset $V \subseteq \mathbb{R}^d$, we denote $\partial_V f := \text{conv}\{\cup_{v\in V}\partial f(v)\}$. Let us define

$$\|\partial f(w)\|_{+\mu} := \inf_{V\subseteq\mathbb{R}^d}\left\{\text{dist}(0, \partial_V f) + \mu \sup_{v\in V}\|v - w\|^2\right\}.$$

Definition 2 ((μ, ϵ)-Regularized Stationarity)**.** *A point w is said to be a (μ, ϵ)-regularized stationary point of a Lipschitz function f if $\|\partial f(w)\|_{+\mu} \le \epsilon$.*

Remark 1. *Intuitively, the (μ, ϵ)-stationarity simultaneously controls the scale of a convex hull of subgradients at points in an underlying subset V and the proximity of V to w. Compared to the relaxed Goldstein stationarity introduced by Zhang & Cutkosky (2024, Definition 2.2), our version uses supreme norm penalty instead of its on-average counterpart, which yields exact equivalence to the original (δ, ϵ)-stationarity, as summarized in the lemma below (see Appendix A.1 for its proof).*

Lemma 2. *Let $\delta, \epsilon, \mu > 0$ be arbitrary positive values. Consider a Lipschitz function f.*

(a) If w is a (δ, ϵ)-stationary point, then it is also a $\left(\frac{\epsilon}{\delta^2}, 2\epsilon\right)$-regularized stationary point.

(b) If w is a (μ, ϵ)-regularized stationary point, then it is also a $\left(\sqrt{\frac{\epsilon}{\mu}}, \epsilon\right)$-stationary point.

We further state the following lemma which shows the monotonicity of $\|\partial f(w)\|_{+\mu}$ with respect to the regularization strength μ. See Appendix A.2 for its proof.

Lemma 3. *Let f be a Lipschitz function. Then for any $w \in \mathbb{R}^d$ and $0 < \mu_1 \le \mu_2$, it holds that $\|\partial f(w)\|_{+\mu_1} \le \|\partial f(w)\|_{+\mu_2}$.*

3 Derandomized O2NC for Weakly Convex Optimization

Building on the O2NC framework, we develop in this section a derandomized stochastic subgradient method for producing Goldstein-style stationary points of weakly convex functions. The overview of algorithm is presented in Section 3.1. There are two optional subroutines for updating the increments in the online learning module of our algorithm: projected OGD and periodically restarted OGD, which are described and analyzed in details respectively in Section 3.2 and Section 3.3.

3.1 Algorithm

The pseudo-code of our Derandomized O2NC (D-O2NC) algorithm is outlined in Algorithm 2. In contrast to the original O2NC (Algorithm 1), the stochastic optimizer module (in light blue) of our algorithm simply eliminates the random interpolation step $v_n = w_{n-1} + s_n\Delta_n$, and directly evaluates the subgradients at each iterate $w_n = w_{n-1} + \Delta_n$. In the online learning module (in light red), we propose two optional variants of OGD for updating the increments Δ_n, both of which are designed for regret minimization over quadratic losses $\langle\hat{g}_n, \cdot\rangle + \frac{\gamma}{2}\|\cdot\|^2$, as described below:

- **Option-I (Clipped OGD):** The online learner $\mathcal{A}$ is instantiated by a standard projected OGD iteration $\Delta_{n+1} = \texttt{clip}_D\left[(1-\eta\gamma)\Delta_n - \eta\hat{g}_n\right]$ with learning rate η over a D-ball constraint.
- **Option-II (Periodically restarted OGD):** We adopt an unconstrained OGD iteration $\Delta_{n+1} = (1-\eta\gamma)\Delta_n - \eta\hat{g}_n$, but reset $\Delta_{n+1} = 0$ whenever $\mod(n+1, T) \equiv 1$. That is, the OGD update of Δ_n is enforced to restart from scratch after every T steps of iteration.

Inspired by the original O2NC, the motivation behind online minimizing a series of quadratic losses in our algorithm is that for a ρ-weakly convex objective and any $\gamma \ge \rho$, we will have $R(w_n) - R(w_{n-1}) \le \mathbb{E}\left[\langle\hat{g}_n, \Delta_n\rangle + \frac{\gamma}{2}\|\Delta_n\|^2\right]$. This suggests that the increments Δ_n might be chosen in a

Algorithm 2: Derandomized O2NC

Input : $\gamma, \eta > 0$, $D > 0$ (optional), $K, T \in \mathbb{N}$, initial w_0 and $\Delta_1 = 0$. Set $N = K \times T$.

for $n = 1, ..., N$ **do**

`/* Stochastic gradient optimizer */`

Update $w_n = w_{n-1} + \Delta_n$;

Randomly sample $z_n \sim \mathcal{D}$ and compute $\hat{g}_n \in \partial \ell(w_n; z_n)$;

`/* Online learning of increments */`

(Option-I) Update $\Delta_{n+1} = \texttt{clip}_D\left[(1-\eta\gamma)\Delta_n - \eta\hat{g}_n\right]$;`/* Clipped OGD */`

(Option-II) `/* Periodically restarted OGD */`

if $\mathrm{mod}(n+1, T) \not\equiv 1$ **then**

  Update $\Delta_{n+1} = (1-\eta\gamma)\Delta_n - \eta\hat{g}_n$;

end

else

  Set $\Delta_{n+1} = 0$;

end

end

Set $w_t^{(k)} = w_{(k-1)T+t}, \forall k \in [K], t \in [T]$, and $\bar{w}^{(k)} = \frac{1}{T}\sum_{t=1}^{T} w_t^{(k)}$.

Output: $\bar{w}_T \sim \mathrm{Unif}(\{\bar{w}^{(k)} : k \in [K]\})$.

sequential manner to make the regret $\sum_{n=1}^{N}\langle \hat{g}_n, \Delta_n\rangle + \frac{\gamma}{2}\|\Delta_n\|^2$ as low as possible, such that the function value gap $R(w_N) - R(w_0)$ can be well upper bounded. See Appendix C for more details on the guarantees of OGD for producing optimal regret over quadratic loss functions.

3.2 Results for D-O2NC under clipped OGD

Recall that in Option-I, the increments are updated with $\Delta_{n+1} = \texttt{clip}_D\left[(1-\eta\gamma)\Delta_n - \eta\hat{g}_n\right]$ which is a projected OGD iteration over the quadratic loss functions $\langle \hat{g}_n, \Delta\rangle + \frac{\gamma}{2}\|\Delta\|^2$ under a D-ball constraint. It is interesting to show a connection of this update to the SGDM method popularly used in training deep learning models (Sutskever et al., 2013; Cutkosky & Orabona, 2019).

Recover SGDM. Let $m_n = -\gamma\Delta_n$ and $\beta = \eta\gamma$, we can reexpress the update with Option-I as

$$w_n = w_{n-1} - \gamma^{-1} m_n;$$
$$m_{n+1} = \texttt{clip}_D\left[(1-\beta)m_n + \beta\hat{g}_n\right].$$

The above procedure can be viewed as a clipped variant of SGDM where m_n is the search direction (which is restricted inside a D-ball), $\hat{g}_n$ is the stochastic subgradient, γ is the learning rate, and β is the momentum parameter. Compared to the clipped SGDM formula implied by the original O2NC (Cutkosky et al., 2023), ours above does not introduce any random perturbation on iterates.

Complexity guarantees. The following theorem is our main result on the convergence of Algorithm 2 for finding (δ, ϵ)-stationary points. See Appendix B.2 for a proof of this result.

Theorem 1. *Suppose that Assumption 1 and Assumption 2 hold. Let $\gamma \geq \rho$ be an arbitrary scalar. Suppose that $\eta \leq \frac{1}{8\gamma}$. Let K and T be positive integers and D be an arbitrary positive number. Then for any $\delta \geq TD$, the sequence $\{\bar{w}^{(k)}\}_{k=1}^{K}$ generated by Algorithm 2 with Option-I satisfies*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \mathit{dist}(0, \partial_\delta R(\bar{w}^{(k)}))\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{2}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

As a direct consequence, the following corollary shows the complexity of Algorithm 2 (with Option-I) for producing Goldstein (δ, ϵ)-stationary point. See Appendix B.3 for its proof.

Corollary 1. *Suppose that Assumption 1 and Assumption 2 hold. Let $\delta, \epsilon > 0$ be the desired Goldstein stationarity parameters and N be the total budget of iterates. Set*

$$T = \left\lceil (\delta N)^{2/3} \right\rceil, K = \left\lfloor \frac{N}{T} \right\rfloor, \gamma = \frac{N^{1/3}}{\delta^{2/3}}, \eta = \frac{1}{8N}, D = \frac{\delta^{1/3}}{N^{2/3}}.$$

Suppose that N is sufficiently large such that

$$N \geq \frac{(G^2 + G + 17 + \Delta R_0)^3}{\delta \epsilon^3} + \rho^3 \delta^2 + \frac{1}{\delta}.$$

Then the output $\bar{w}_T$ by Algorithm 2 with Option-I satisfies

$$\mathbb{E}\left[dist\left(0, \partial_\delta R(\bar{w}_T)\right)\right] \leq \epsilon.$$

Remark 2. *We comment that the $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ rate, which dominates the composite complexity bound of Corollary 1, is indeed tight for weakly convex functions. The key insight is that the $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ rate is known to be optimal for all $\epsilon \leq \mathcal{O}(\delta)$ (Cutkosky et al., 2023)—a result that holds even for smooth functions, let alone their superclass of weakly convex functions.*

Remark 3. *It is interesting to note that the weak-convexity parameter ρ does not appear in this dominant rate, but rather in a suboptimal component $\rho^3\delta^2$ which allows it to scale as large as $\mathcal{O}(\delta^{-1}\epsilon^{-1})$ without dominating the optimal rate. In other words, the higher convergence precision is required, the larger weak-convexity can be tolerated by our algorithm to preserve optimality.*

Remark 4. *The hyperparameter choices outlined in Corollary 1 are uniquely determined by the iterate budget N and desired stationarity precisions δ, ϵ, which can typically be specified by users in practical applications. For instance, with $\delta = \epsilon = \mathcal{O}(N^{-1/4})$, Corollary 1 prescribes $T = \mathcal{O}(N^{-1/2})$, $K = \mathcal{O}(N^{1/2})$ $\gamma = \mathcal{O}(N^{1/2})$, $\eta = \mathcal{O}(N^{-1})$ and $D = N^{-3/4}$.*

3.3 Results for D-O2NC under periodically restarted OGD

While Algorithm 2 with Option-I can achieve optimal dimension-free iteration complexity, the clipped OGD iteration enforces the increments Δ_n to stay inside a sufficiently small ball, which could be too conservative. To deal with this issue, we further propose a novel periodically restarted OGD procedure as the Option-II in our algorithm for implementing the OCO module. More precisely, at each time step $n \geq 1$, we update $\Delta_{n+1} = (1 - \eta\gamma)\Delta_n - \eta\hat{g}_n$, and reset $\Delta_{n+1} = 0$ whenever $\mathrm{mod}\,(n+1, T) \equiv 1$. Such an unconstrained OGD procedure allows for more progressive update especially when the iterates are far from stationary.

Remark 5. *In regard with OCO module design, our Algorithm 2 with Option-II shares some spirits with E-O2NC (Zhang & Cutkosky, 2024) where the OCO module is instantiated by an unconstrained OMD. While bearing some similarity, our algorithm has two clear differences from theirs: 1) ours is deterministic without requiring any random scaling on the increments; 2) our algorithm neither needs to exponentially weight the subgradients in constructing losses, nor uses exponential aggregation of iterates for generating output, and thus is perhaps more relevant to practical implementation.*

Remark 6. *It is noteworthy that the proposed periodically restarted OGD can be immediately extended to the original O2NC for generic non-smooth, non-convex optimization. This is true because under the so-called well-behavedness assumption (Cutkosky et al., 2023), similar quadratic losses of the form $\langle \hat{g}_n, \cdot \rangle + \frac{\gamma}{2}\|\cdot\|^2$ can also be constructed in O2NC (or E-O2NC) with arbitrary $\gamma > 0$.*

Recover SGDM. As an interesting consequence of using periodically restarted OGD, we can explicitly write the update of Algorithm 2 with Option-II as

$$w_n = w_{n-1} - \gamma^{-1} m_n;$$
$$m_{n+1} = \left((1-\beta)m_n + \beta\hat{g}_n\right)\mathbf{1}_{\{\mathrm{mod}\,(n+1,T)\neq 1\}},$$

where $m_n = -\gamma\Delta_n$, $\beta = \eta\gamma$, and $\mathbf{1}_{\{\cdot\}}$ represents the indication function. The above update formula is almost identical to the standard SGDM, with the only difference that the update of search direction m_n is now enforced to start over again after every T rounds of iteration. Similar resemblance to SGDM was also revealed for the E-O2NC method (Zhang & Cutkosky, 2024), though under somewhat more sophisticated algorithmic designs as commented in Remark 5.

Complexity guarantees. The following is our main result on the convergence rate of Algorithm 2 with periodically restarted OGD (Option-II). A proof of this result is provided in Appendix B.4.

Theorem 2. *Suppose that Assumption 1 and Assumption 2 hold. Let $\gamma \geq \rho$ be an arbitrary scalar. Suppose that $\eta \leq \frac{1}{8\gamma}$. Let K and T be positive integers and D be an arbitrary positive number. Then for any $\mu \leq \frac{\gamma}{8DT^2}$, the sequence $\{\bar{w}^{(k)}\}_{k=1}^K$ generated by Algorithm 2 with Option-II satisfies*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\partial R(\bar{w}^{(k)})\right\|_{+\mu}\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{1}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

Remark 7. *Unlike in Theorem 1 where D is a hyperparameter in Option-I, the scalar D in Theorem 2 does not actually show up in Option-II: it is introduced for analysis purpose only.*

Based on Theorem 2, we can further establish the following result on the complexity of Algorithm 2 (with Option-II) for producing (μ, ϵ)-regularized stationary points. See Appendix B.5 for its proof.

Corollary 2. *Suppose that Assumption 1 and Assumption 2 hold. Let $\mu, \epsilon > 0$ be the desired regularized-stationarity parameters and N be the total budget of iterates. Set*

$$T = \left\lceil N^{4/7}\mu^{-2/7} \right\rceil, K = \left\lfloor \frac{N}{T} \right\rfloor, \gamma = N^{3/7}\mu^{2/7}, \eta = \frac{1}{8N}.$$

Suppose that

$$N \geq \frac{(4G^2 + 1 + 32\Delta R_0)^{7/2}\mu^{1/2}}{\epsilon^{7/2}} + \frac{\rho^{7/3}}{\mu^{2/3}} + \mu^{1/2}.$$

Then the output $\bar{w}_T$ by Algorithm 2 with Option-II satisfies

$$\mathbb{E}\left[\|\partial R(\bar{w}_T)\|_{+\mu}\right] \leq \epsilon.$$

Remark 8. *In view of Lemma 2, by setting $\mu = \delta^{-2}\epsilon$, the bound in Corollary 2 implies an $\mathcal{O}\left(\delta^{-1}\epsilon^{-3} + \rho^{7/3}\delta^{4/3}\epsilon^{-2/3} + \delta^{-1}\epsilon^{1/2}\right)$ complexity for producing (δ, ϵ)-stationary points, which is dominated by the optimal term $\delta^{-1}\epsilon^{-3}$. Similar to the discussion in Remark 2, the weak-convexity parameter is allowed to scale as $\rho = \mathcal{O}(\mu^{1/2}\epsilon^{-3/2})$ without dominating the optimal component.*

3.4 Comparison with prior results

In Table 1, we summarize the complexity bounds and some important properties of D-O2NC with comparison to several other subgradient-based methods for weakly convex optimization, including SGD (Davis & Drusvyatskiy, 2019), SGDM (Mai & Johansson, 2020) and Interpolated Normalized Gradient Descent (INGD) (Davis et al., 2022). A few comments are in order.

- **Comparison with INGD.** Our D-O2NC is deterministic up to the use of stochastic oracles, with *dimension-free* and optimal complexity in terms of (δ, ϵ)-stationarity. In contrast, INGD is randomized in design and hard to be extended to the stochastic setting; and its corresponding complexity bound is *dimension-dependent*, but with sharper dependence on ρ and (δ, ϵ).
- **Comparison with SGD and SGDM.** The listed optimal complexity of $\mathcal{O}(\epsilon^{-4})$ for SGD and SGDM are about the ϵ-stationarity of Moreau envelope, i.e., $\|\nabla R_{1/\bar{\rho}}\| \leq \epsilon$ with $\bar{\rho} = \mathcal{O}(\rho)$. While our optimal bounds are not directly comparable to this complexity due to the distinct criteria adopted, we still have some observations worth highlighting: 1) Lemma 1 suggests that a ϵ-stationary point of the Moreau envelope implies a $(\epsilon/(2\rho), \epsilon)$-stationary point of the original objective (see Remark 11 in Appendix D for details); thus the bounds of SGD and SGDM imply an $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ complexity for finding(δ, ϵ)-stationary points, albeit under a relatively restrictive choice $\delta = \epsilon/(2\rho)$; 2) As we demonstrate in Theorem 3 (Appendix D) that a (δ, ϵ)-stationary point implies an $\left(\epsilon + \sqrt{\delta}\right)$-stationary point of the Moreau envelope, which however yields the suboptimal complexity $\mathcal{O}(\epsilon^{-5})$ for achieving ϵ-stationarity (Corollary 3); 3) For second-order smooth functions, based on the result of Cutkosky et al. (2023, Proposition 15) it can be readily shown that D-O2NC recovers the optimal $\mathcal{O}(\epsilon^{-3.5})$ complexity for finding ϵ-stationary point, which however cannot be automatically implied by the tabulated results of SGD/SGDM. Last but not least, the weak-convexity parameter ρ is allowed to be as large as $\delta^{-1}\epsilon^{-1}$ in our bound without dominating the optimal rate, which is not applicable to those bounds of SGD and SGDM.

4 Experiments

In this section, we conduct a preliminary experimental study to evaluate the effectiveness of our D-O2NC method when specified with the periodically restarted OGD optimizer (Option-II) for training deep neural networks. Since our algorithm corresponds to a momentum-resetting version of SGDM, we choose to use standard SGDM as a baseline algorithm for comparison. Additional experimental results are provided in the Appendix section E. We emphasize that while our empirical study offers insights into the numerical aspects of D-O2NC, this work is primarily a theoretical contribution.

Method	(δ, ϵ)-stationarity	ϵ-stationarity (Moreau envelope)	DET	SO
SGD (Davis & Drusvyatskiy, 2019)	–	$\mathcal{O}\left(\frac{\rho}{\epsilon^4}\right)$	✓	✓
SGDM (Mai & Johansson, 2020)	–	$\mathcal{O}\left(\frac{\rho^2}{\epsilon^4}\right)$	✓	✓
INGD (Davis et al., 2022)	$\mathcal{O}\left(\frac{d\log(\rho)}{\delta\epsilon}\right)$	–	✗	✗
D-O2NC with Option-I (ours)	$\mathcal{O}\left(\frac{1}{\delta\epsilon^3} + \rho^3\delta^2 + \frac{1}{\delta}\right)$	–	✓	✓
D-O2NC with Option-II (ours)	$\mathcal{O}\left(\frac{1}{\delta\epsilon^3} + \frac{\rho^{7/3}\delta^{4/3}}{\epsilon^{3/2}} + \frac{1}{\delta}\right)$	–	✓	✓

Table 1: Comparison of subgradient-based weakly convex optimization algorithms in terms of complexity bounds, determinism (DET), and applicability with stochastic oracle (SO). The involved quantities: (δ, ϵ): convergence precisions; ρ: weak-convexity parameter; d: dimension of model.

4.1 Experiment Setup

Dataset and backbone. Our experiments are conducted on the CIFAR-10 image classification benchmark dataset Krizhevsky & Hinton (2009) popularly used for evaluating deep learning models and algorithms. It consists of 60,000 color images across 10 classes, with 50,000 allocated for training and 10,000 for testing. We employ ResNet-101 (He et al., 2016) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) as two backbone networks for representation learning, using GeLU (Hendrycks & Gimpel, 2016) as activation functions in both cases. Notably, neural networks with smooth activation functions (e.g., GeLU, softplus) typically match or even outperform their non-smooth ReLU-based counterparts (Clevert et al., 2016; Xu et al., 2015).

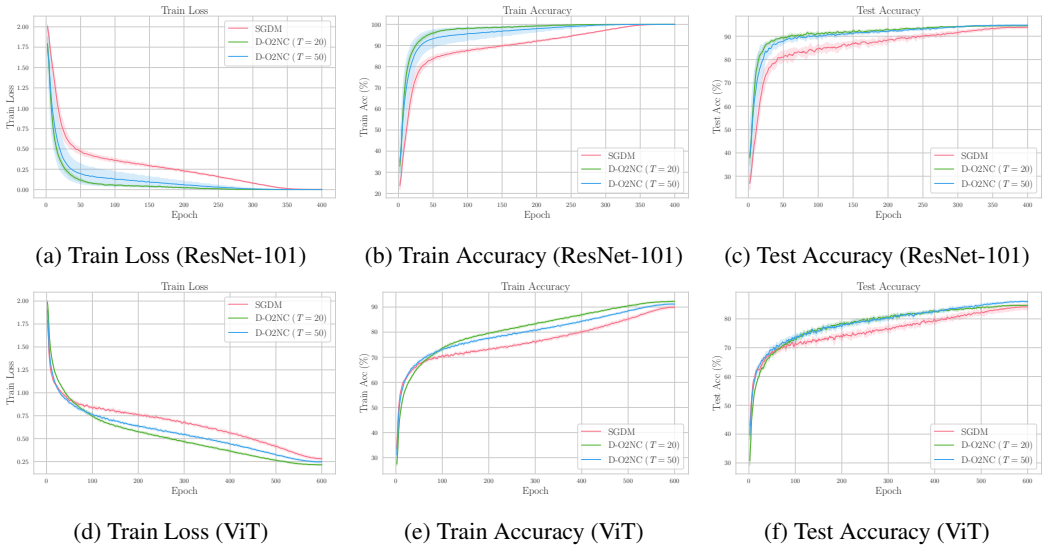
Implementation details and performance metrics. For all considered algorithms, the model parameters are optimized over 400 epochs with a minibatch size of 256 for ResNet-101, and 600 epochs (with the same minibatch size) for ViT, where a patch size of 4 is adopted. The total number of minibatches per epoch is 196. The initial learning rate is 0.01, decayed via cosine annealing to facilitate smoother convergence. The optimizer employs a momentum of 0.99, along with a weight decay of 5×10^{-4}. Our periodically restarted O2NC method is implemented and compared under two different restarting frequency $T \in \{20, 50\}$. For each experiment, we carried out three independent runs with distinct random seeds, recording the empirical loss and training accuracy during training, as well as the prediction accuracy on the test set.

4.2 Results

Figure 1 shows the convergence curves of the considered algorithms. The results of the optimal iterate for each trial are documented in Table 2. From this group of results we can see that D-O2NC converges considerably sharper than SGDM on both models, and averagely the former outperforms the latter in test accuracy by 0.9 percentage points on ResNet-101, and 1.96 on ViT. These obser-

Table 2: Numerical results of the best iterate in each trial on CIFAR-10.

Backbone	Metric	SGDM	D-O2NC ($T = 20$)	D-O2NC ($T = 50$)
ResNet101	Train Loss ($\times 10^{-3}$)	2.55 ± 0.08	0.55 ± 0.02	1.38 ± 0.01
	Train Accuracy (%)	99.98 ± 0.0	100 ± 0.0	100 ± 0.0
	Test Accuracy (%)	93.91 ± 0.55	94.64 ± 0.07	94.81 ± 0.34
ViT	Train Loss ($\times 10^{-1}$)	2.78 ± 0.10	2.12 ± 0.03	2.44 ± 0.10
	Train Accuracy (%)	90.16 ± 0.51	92.38 ± 0.28	91.36 ± 0.22
	Test Accuracy (%)	84.26 ± 0.41	84.92 ± 0.17	86.22 ± 0.29



(a) Train Loss (ResNet-101) (b) Train Accuracy (ResNet-101) (c) Test Accuracy (ResNet-101)

(d) Train Loss (ViT) (e) Train Accuracy (ViT) (f) Test Accuracy (ViT)

Figure 1: Experimental results on CIFAR-10 with ResNet-101 (top) and ViT (bottom) networks.

vations demonstrate that the momentum-resetting mechanism in our method might not only help to improve convergence but also yield superior generalization performance.

To validate the hyperparameter sensitivity of our method, we have conducted additional experiments across diverse configurations of the restarting frequency, learning rate, and momentum coefficient (see Appendix E.1 for detailed results). In Appendix E.2, we further present a set of experiments on robust phase retrieval, a classic weakly convex task, to consolidate the empirical support for our method's effectiveness. The results show that D-O2NC consistently outperforms SGDM under the same configuration of hyperparameters.

Beyond our empirical study, it is noteworthy that the momentum-resetting technique has recently demonstrated experimental effectiveness in training large language models (LLMs) and deep reinforcement learning (deep RL) tasks (Huang et al., 2025; Asadi et al., 2023). Our convergence results on D-O2NC with restarted OGD thus provide a solid theoretical foundation for explaining the empirical success of the momentum-resetting technique.

5 Conclusion

In this paper, we made progress towards resolving a critical issue on the link of O2NC (online-to-non-convex conversion) to SGDM: under auxiliary random interpolation or scaling, O2NC mirrors SGDM but randomization causes deviations from standard SGDM. To this end, for a broad class of weakly convex functions, we presented D-O2NC as a derandomized version of O2NC that maintains optimal oracle complexity $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ while recovering SGDM in a deterministic way. Our method allows the weak-convexity parameter to scale as large as $\mathcal{O}(\delta^{-1}\epsilon^{-1})$ without dominating the optimal rate, meaning that stronger stationarity yields tolerating higher non-convexity. Furthermore, a periodically restarted variant of D-O2NC is developed, enabling more progressive updates when far from stationary. Corresponding to a momentum-restarted SGDM method, this variant has been empirically shown to be effective for training ResNet and ViT models on benchmark datasets. An interesting future work is to extend our periodically restarted O2NC technique to the analysis and improvement of other popular ML optimizers including Adam and schedule-free SGD.

DECLARATION OF LARGE LANGUAGE MODELS (LLMS) USAGE

We acknowledge the use of Doubao 1.6 (ByteDance, 2025) as an auxiliary LLM tool solely for polishing this manuscript, including refining English grammar and improving the readability of non-core descriptive content. All outputs from Doubao were thoroughly reviewed, revised, and validated by the authors to ensure accuracy, consistency with the research context, and alignment with academic standard. Doubao did not participate in idea formalization, theoretical derivation, result analysis, or drafting of core sections (Abstract, Introduction, Method, Experiment). All authors bear full responsibility for the manuscript's content integrity and scientific validity.

REFERENCES

Kwang Jun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free sgd is also effective for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, 2025.

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *arXiv preprint arXiv:2405.18199*, 2024.

Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep rl: An empirical study. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Axel Böhm and Stephen J Wright. Variable smoothing for weakly convex composite functions. *Journal of optimization theory and applications*, 188:628–649, 2021.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

Nicolo Cesa-Bianchi, Alessandro Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004. ISSN 0018-9448. doi: 10.1109/TIT.2004.833339.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*, 2016.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 15236–15245, 2019.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning (ICML)*, pp. 6643–6670. PMLR, 2023.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems*, 37: 9974–10007, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

D Drusvyatskiy and C Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019. doi: 10.1007/s10107-018-1311-3.

John C. Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018.

Yu M Ermol'ev and VI Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, 1998.

Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. Spam: Spike-aware adam with momentum reset for stable llm training. In *International Conference on Learning Representations (ICLR)*, 2025.

Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Conference on Learning Theory*, pp. 4570–4597. PMLR, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

Siyu Kong and Adrian S Lewis. Lipschitz minimization and the goldstein modulus. *Mathematical Programming*, pp. 1–30, 2025.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23:1–43, 2022.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Vien V Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning (ICML)*, 2020.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Evgeni Alekseevich Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, 1973.

Spyridon Pougkakiotis and Dionysios S Kalogerias. A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. *SIAM Journal on Scientific Computing*, 45 (5):2679–2702, 2023. doi: 10.1137/22M1486306.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Jan A Snyman. *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*. Springer, 2005.

Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, volume 28, pp. 1139–1147. JMLR. org, 2013.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

Wim van Ackooij, Felipe Atenas, and Claudia Sagastizábal. Weak convexity and approximate subdifferentials. *Journal of Optimization Theory and Applications*, 203(2):1686–1709, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, Long Beach, CA, 2017.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Forty-first International Conference on Machine Learning (ICML)*, pp. 58780 – 58799. PMLR, 2024.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on International Conference on Machine Learning*, pp. 928–936, 2003.

A PROOFS IN SECTION 2

A.1 PROOF OF LEMMA 2

We prove the following restated lemma which establishes the equivalence between (μ, ϵ)-regularized stationarity and Goldstein (δ, ϵ)-stationarity.

Lemma 2. *Let $\delta, \epsilon, \mu > 0$ be arbitrary positive values. Consider a Lipschitz function f.*

(a) If w is a (δ, ϵ)-stationary point, then it is also a $\left(\frac{\epsilon}{\delta^2}, 2\epsilon\right)$-regularized stationary point.

(b) If w is a (μ, ϵ)-regularized stationary point, then it is also a $\left(\sqrt{\frac{\epsilon}{\mu}}, \epsilon\right)$-stationary point.

Proof. Part(a): Let w be a Goldstein (δ, ϵ)-stationary point of f. Consider $\mu = \frac{\epsilon}{\delta^2}$. Then it follows from Definition 2 that

$$\begin{aligned}\|\partial f(w)\|_{+\mu} \leq & \text{dist}(0, \partial_{\mathbb{B}_\delta(w)} f) + \mu \sup_{v \in \mathbb{B}_\delta(w)} \|v - w\|^2 \\ = & \text{dist}(0, \partial_\delta f(w)) + \mu \sup_{v \in \mathbb{B}_\delta(w)} \|v - w\|^2 \\ \leq & \epsilon + \mu\delta^2 = 2\epsilon,\end{aligned}$$

where in the second inequality we have used the definition of Goldstein (δ, ϵ)-stationarity. Then by definition w must be a $\left(\frac{\epsilon}{\delta^2}, 2\epsilon\right)$-regularized stationary point of f.

Part(b): Let us now consider the case that w is a (μ, ϵ)-regularized stationary point of f. Let $\delta = \sqrt{\frac{\epsilon}{\mu}}$ and $\varepsilon > 0$ be arbitrary. Since $\|\partial f(w)\|_{+\mu} \leq \epsilon$, it follows from Definition 2 that there exists some $V^*(\varepsilon)$ such that

$$\epsilon \geq \|\partial f(w)\|_{+\mu} \geq \text{dist}(0, \partial_{V^*(\varepsilon)} f) + \mu \sup_{v \in V^*(\varepsilon)} \|v - w\|^2 - \varepsilon,$$

which then directly implies

$$\text{dist}(0, \partial_{V^*(\varepsilon)} f) \leq \epsilon + \varepsilon, \quad \sup_{v \in V^*(\varepsilon)} \|v - w\| \leq \sqrt{\frac{\epsilon + \varepsilon}{\mu}} \leq \sqrt{\frac{\epsilon}{\mu}} + \sqrt{\frac{\varepsilon}{\mu}} = \delta + \sqrt{\frac{\varepsilon}{\mu}}.$$

The second inequality in the above implies that $V^*(\varepsilon) \subseteq \mathbb{B}_{\delta + \sqrt{\frac{\varepsilon}{\mu}}}(w)$ and thus $\partial_{V^*(\varepsilon)} f \subseteq \partial_{\delta + \sqrt{\frac{\varepsilon}{\mu}}} f$. Then we have

$$\text{dist}\left(0, \partial_{\delta + \sqrt{\frac{\varepsilon}{\mu}}} f\right) \leq \text{dist}\left(0, \partial_{V^*(\varepsilon)} f\right) \leq \epsilon + \varepsilon.$$

Since ε is allowed to be arbitrarily small and recall that $\delta = \sqrt{\frac{\epsilon}{\mu}}$, the above inequality implies that w deems a Goldstein $\left(\sqrt{\frac{\epsilon}{\mu}}, \epsilon\right)$-stationary point. □

A.2 PROOF OF LEMMA 3

Here we prove the following restated lemma on the monotonicity of $\|\partial F(w)\|_{+\mu}$ with respect to μ.

Lemma 3. *Let f be a Lipschitz function. Then for any $w \in \mathbb{R}^d$ and $0 < \mu_1 \leq \mu_2$, it holds that $\|\partial f(w)\|_{+\mu_1} \leq \|\partial f(w)\|_{+\mu_2}$.*

Proof. Consider a fixed vector w. Let $\varepsilon > 0$ be arbitrary. By definition we know that there exists a subset $V_2^*(\varepsilon) \subseteq \mathbb{R}^d$ such that $\|\partial F(w)\|_{+\mu_2} \geq \text{dist}(0, \partial_{V_2^*(\varepsilon)} F) + \mu_2 \sup_{v \in V_2^*(\varepsilon)} \|v - w\|^2 - \varepsilon$. Again, by definition and the condition $\mu_1 \leq \mu_2$ we can see that

$$\begin{aligned}\|\partial R(w)\|_{+\mu_1} \leq & \text{dist}(0, \partial_{V_2^*(\varepsilon)} R) + \mu_1 \sup_{v \in V_2^*(\varepsilon)} \|v - w\|^2 \\ \leq & \text{dist}(0, \partial_{V_2^*(\varepsilon)} R) + \mu_2 \sup_{v \in V_2^*(\varepsilon)} \|v - w\|^2 \\ \leq & \|\partial R(w)\|_{+\mu_2} + \varepsilon.\end{aligned}$$

By noting that ε can be arbitrarily small, we must have $\|\partial F(w)\|_{+\mu_1} \leq \|\partial F(w)\|_{+\mu_2}$. □

B PROOFS IN SECTION 3

B.1 SOME KEY LEMMAS

The following lemma is key to our analysis of Algorithm 2.

Lemma 4. *Suppose that Assumption 1 and Assumption 2 hold. Let $\gamma \geq \rho$ and $D > 0$ be arbitrary numbers. Suppose that $\eta \leq \frac{1}{8\gamma}$. Then for any $k \in [K]$, the sequence $\{w_t^{(k)}\}_{t=1}^T$ generated by Algorithm 2 satisfies*

$$\begin{aligned}&\mathbb{E}\left[R(w_T^{(k)}) - R(w_0^{(k)}) + \sum_{t=1}^T \frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right]\\ \leq &-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2 + \frac{\|\Delta_1^{(k)}\|^2}{\eta}.\end{aligned}$$

Proof. Let us consider the filtration $\mathcal{F}_t = S\{\Delta_1, \Delta_2, ..., \Delta_{t+1}\}$ where $S\{\cdot\}$ denotes the sigma field. For any $n \geq 1$, by Assumption 2 we have $g_n := \mathbb{E}[\hat{g}_n \mid \mathcal{F}_{n-1}] \in \partial R(w_n)$. For any $\gamma \geq \rho$, from the weak convexity assumption in Assumption 1 we can see that the following holds for all $n \geq 1$,

$$\begin{aligned}R(w_n) - R(w_{n-1}) \leq &\mathbb{E}\left[\langle g_n, \Delta_n\rangle + \frac{\gamma}{2}\|\Delta_n\|^2 \mid \mathcal{F}_{n-1}\right]\\ = &\mathbb{E}\left[\langle \hat{g}_n, \Delta_n\rangle + \frac{\gamma}{2}\|\Delta_n\|^2 \mid \mathcal{F}_{n-1}\right].\end{aligned}$$

It follows from the law of total expectation that

$$\mathbb{E}\left[R(w_n) - R(w_{n-1})\right] \leq \mathbb{E}\left[\langle \hat{g}_n, \Delta_n\rangle + \frac{\gamma}{2}\|\Delta_n\|^2\right].$$

Consider a fixed $k \in [K]$. Recall that $w_t^{(k)} = w_{(k-1)T+t}$, $t \in [T]$ and similarly we use the notations $\Delta_t^{(k)}, \hat{g}_t^{(k)}, g_t^{(k)}$. Then the above inequality implies that for any $t \in [T]$:

$$\mathbb{E}\left[R(w_t^{(k)}) - R(w_{t-1}^{(k)})\right] \leq \mathbb{E}\left[\langle \hat{g}_t^{(k)}, \Delta_t^{(k)}\rangle + \frac{\gamma}{2}\|\Delta_t^{(k)}\|^2\right].$$

By summing the above bound over $t = 1, ..., T$ we obtain

$$\begin{aligned}&\mathbb{E}\left[R(w_T^{(k)}) - R(w_0^{(k)})\right]\\ =&\mathbb{E}\left[\sum_{t=1}^T \left(R(w_t^{(k)}) - R(w_{t-1}^{(k)})\right)\right] \leq \mathbb{E}\left[\sum_{t=1}^T \left(\langle \hat{g}_t^{(k)}, \Delta_t^{(k)}\rangle + \frac{\gamma}{2}\|\Delta_t^{(k)}\|^2\right)\right].\end{aligned} \tag{2}$$

We next upper bound the RHS in the above inequality using the OGD regret bound in Lemma 6. To this end, it can be noted that the sequence $\{\Delta_t^{(k)}\}_{t=1}^T$ generated by Algorithm 2 is the output of OGD, starting from $\Delta_1^{(k)}$ with step-size η, on the following quadratic losses $\{f_t^{(k)}\}_{t\in[T]}$ over the constraint $\mathbb{B}_D(0)$ (for Option-I) (or over the entire space $\mathbb{R}^d$ for Option-II):

$$f_t^{(k)}(\cdot) = \langle \hat{g}_t^{(k)}, \cdot\rangle + \frac{\gamma}{2}\|\cdot\|^2.$$

For arbitrary $D > 0$, let us consider the following comparator:

$$\bar{\Delta}^{(k)} := -D\frac{\sum_{t=1}^T g_t^{(k)}}{\left\|\sum_{t=1}^T g_t^{(k)}\right\|}.$$

Denote $\bar{g}^{(k)} := \frac{1}{T}\sum_{t=1}^{T} g_t^{(k)}$ and $\bar{\hat{g}}^{(k)} := \frac{1}{T}\sum_{t=1}^{T} \hat{g}_t^{(k)}$. Since $\eta \le \frac{1}{8\gamma}$, we can apply Lemma 6 to get

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sum_{t=1}^{T}\langle \hat{g}_t^{(k)}, \Delta_t^{(k)}\rangle + \frac{\gamma}{2}\|\Delta_t^{(k)}\|^2\right)\right] \\
\overset{\text{Lemma 6}}{\le}\ &\mathbb{E}\left[\sum_{t=1}^{T}\left(\left\langle \hat{g}_t^{(k)}, \bar{\Delta}^{(k)}\right\rangle + \frac{\gamma}{2}\|\bar{\Delta}^{(k)}\|^2\right) + \sum_{t=1}^{T}\left(\eta\|\hat{g}_t^{(k)}\|^2 + \frac{\gamma}{2}\|\bar{\Delta}^{(k)}\|^2 - \frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right)\right. \\
&\left. + \frac{1}{\eta}\left(\|\Delta_1^{(k)}\|^2 + \|\bar{\Delta}^{(k)}\|^2\right)\right] \\
=&\mathbb{E}\left[\left\langle \sum_{t=1}^{T}(\hat{g}_t^{(k)} - g_t^{(k)}), \bar{\Delta}^{(k)}\right\rangle + \left\langle \sum_{t=1}^{T} g_t^{(k)}, \bar{\Delta}^{(k)}\right\rangle\right. \\
&\left. + \sum_{t=1}^{T}\left(\eta\|\hat{g}_t^{(k)}\|^2 + \gamma\|\bar{\Delta}^{(k)}\|^2 - \frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right) + \frac{1}{\eta}\left(\|\Delta_1^{(k)}\|^2 + \|\bar{\Delta}^{(k)}\|^2\right)\right] \\
\overset{\zeta_1}{\le}&\mathbb{E}\left[DT\left\|\bar{\hat{g}}^{(k)} - \bar{g}^{(k)}\right\| - DT\left\|\bar{g}^{(k)}\right\| + \sum_{t=1}^{T}\left(\eta\|\hat{g}_t^{(k)}\|^2 + \gamma D^2 - \frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right) + \frac{D^2}{\eta} + \frac{\|\Delta_1^{(k)}\|^2}{\eta}\right] \\
\overset{\zeta_2}{\le}&-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\| + \sum_{t=1}^{T}\frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right] + \eta T G^2 + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2 + \frac{\|\Delta_1^{(k)}\|^2}{\eta},
\end{aligned}
$$

where in "ζ_1" we have used Cauchy–Schwarz inequality and used the fact $\|\Delta_1^{(k)}\| \le D$, and in "ζ_2" we have used $\|\hat{g}_t^{(k)}\| \le G$ implied by Assumption 1 and $\mathbb{E}\left[\left\|\bar{\hat{g}}^{(k)} - \bar{g}^{(k)}\right\|\right] \le \sqrt{\mathbb{E}\left[\left\|\bar{\hat{g}}^{(k)} - \bar{g}^{(k)}\right\|^2\right]} = \frac{1}{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}\|g_t^{(k)} - \hat{g}_t^{(k)}\|^2} \le \frac{1}{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}\|\hat{g}_t^{(k)}\|^2} \le \frac{G}{\sqrt{T}}$. By substituting the previous inequality into equation 2 and rearranging the terms we obtain the desired bound. The proof is completed. □

The following simple lemma is also useful in our analysis. A proof is provided for the sake of completeness.

Lemma 5. *Let $w_1, w_2, ..., w_n$ be a set of vectors and $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$. Then the following holds for all $i \in [n]$:*

$$
\|w_i - \bar{w}\|^2 \le \frac{1}{n}\sum_{i'=1}^{n}\|w_i - w_{i'}\|^2 \le n\sum_{j=1}^{n}\|\Delta_j\|^2,
$$

where $\Delta_j := w_j - w_{j-1}$ and w_0 can be chosen arbitrary for determining Δ_1.

Proof. Fix some $i \in [n]$. It can be shown that

$$
\begin{aligned}
\|w_i - \bar{w}\|^2 =& \left\|w_i - \frac{1}{n}\sum_{i'=1}^{n} w_{i'}\right\|^2 \le \frac{1}{n}\sum_{i'=1}^{n}\|w_i - w_{i'}\|^2 \\
=&\frac{1}{n}\sum_{i'=1}^{n}\left\|\sum_{j=i\wedge i'+1}^{i\vee i'}(w_{j-1} - w_j)\right\|^2 \\
\le&\frac{1}{n}\sum_{i'=1}^{n}\left(\sum_{j=i\wedge i'+1}^{i\vee i'}\|\Delta_j\|\right)^2 \le \left(\sum_{j=2}^{n}\|\Delta_j\|\right)^2 \le n\sum_{j=1}^{n}\|\Delta_j\|^2.
\end{aligned}
$$

The proof is completed. □

B.2 Proof of Theorem 1

Theorem 1. *Suppose that Assumption 1 and Assumption 2 hold. Let $\gamma \geq \rho$ be an arbitrary scalar. Suppose that $\eta \leq \frac{1}{8\gamma}$. Let K and T be positive integers and D be an arbitrary positive number. Then for any $\delta \geq TD$, the sequence $\{\bar{w}^{(k)}\}_{k=1}^{K}$ generated by Algorithm 2 with Option-I satisfies*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} dist(0, \partial_\delta R(\bar{w}^{(k)}))\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{2}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

Proof. Under the given conditions, for any $k \in [K]$, we can invoke Lemma 4 to Algorithm 2 (with Option-I) to get

$$\begin{aligned}
&\mathbb{E}\left[R(w_T^{(k)}) - R(w_0^{(k)}) + \sum_{t=1}^{T}\frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right] \\
\leq& -\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2 + \frac{\|\Delta_1^{(k)}\|^2}{\eta} \\
\leq& -\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)D^2,
\end{aligned}$$

where in the last step we have used the fact $\|\Delta_1^{(k)}\| \leq D$ due to the explicit constraint imposed in Option-I. Note that by definition we have $w_T^{(k)} = w_0^{(k+1)}$. By omitting the non-negative summation term in the LHS of the above inequality we get

$$\mathbb{E}\left[R(w_0^{(k+1)}) - R(w_0^{(k)})\right] \leq -\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)D^2.$$

Rearranging the terms on both sides of the above inequality yields

$$\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] \leq \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)D^2 + \mathbb{E}\left[R(w_0^{(k)}) - R(w_0^{(k+1)})\right].$$

By summing the above inequality of over $k \in [K]$ we get

$$\begin{aligned}
\mathbb{E}\left[DT\sum_{k=1}^{K}\left\|\bar{g}^{(k)}\right\|\right] \leq& \eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)KD^2 + \mathbb{E}\left[\sum_{k=1}^{K}\left(R(w_0^{(k)}) - R(w_0^{(k+1)})\right)\right] \\
=& \eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)KD^2 + \mathbb{E}\left[R(w_0^{(1)}) - R(w_0^{(K+1)})\right] \\
\leq& \eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{2}{\eta}\right)KD^2 + R(w_0) - R^*.
\end{aligned}$$

Dividing the factor DKT on both sides of the above inequality yields

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\bar{g}^{(k)}\right\|\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{2}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}. \tag{3}$$

Since $\|\Delta_t^{(k)}\| \leq D$ almost surely for all $t \in [T]$, by applying Lemma 5 we obtain that,

$$\left\|w_t^{(k)} - \bar{w}^{(k)}\right\| \leq \sqrt{T\sum_{t=1}^{T}\|\Delta_t^{(k)}\|^2} \leq TD \leq \delta, \ \ \forall t \in [T],$$

which implies

$$\text{dist}\left(0, \partial_\delta R(\bar{w}^{(k)})\right) \leq \left\|\frac{1}{T}\sum_{t=1}^{T} g_t^{(k)}\right\| = \left\|\bar{g}^{(k)}\right\|.$$

Combining the above with equation 3 yields

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\text{dist}\left(0, \partial_\delta R(\bar{w}^{(k)})\right)\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{2}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

The proof is completed. □

B.3 Proof of Corollary 1

Corollary 1. *Suppose that Assumption 1 and Assumption 2 hold. Let $\delta, \epsilon > 0$ be the desired Goldstein stationarity parameters and N be the total budget of iterates. Set*

$$T = \left\lceil (\delta N)^{2/3} \right\rceil, K = \left\lfloor \frac{N}{T} \right\rfloor, \gamma = \frac{N^{1/3}}{\delta^{2/3}}, \eta = \frac{1}{8N}, D = \frac{\delta^{1/3}}{N^{2/3}}.$$

Suppose that N is sufficiently large such that

$$N \geq \frac{(G^2 + G + 17 + \Delta R_0)^3}{\delta \epsilon^3} + \rho^3 \delta^2 + \frac{1}{\delta}.$$

Then the output $\bar{w}_T$ by Algorithm 2 with Option-I satisfies

$$\mathbb{E}\left[dist\left(0, \partial_\delta R(\bar{w}_T)\right)\right] \leq \epsilon.$$

Proof. The given choice of the hyperparameters ensures that $TD \leq \delta$. Under the condition on N we can verify that

$$\gamma \geq \rho, \quad \gamma\eta = \frac{1}{8(\delta N)^{2/3}} \leq \frac{1}{8}.$$

Then all the conditions of Theorem 1 are fulfilled in our setting, and the theorem can be applied to obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \text{dist}\left(0, \partial_\delta R(\bar{w}^{(k)})\right)\right] \\
\leq &\frac{\eta G^2}{D} + \left(\gamma T + \frac{2}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT} \\
\leq &\left(\frac{G^2}{8} + 1 + 16 + G + \Delta R_0\right)\frac{1}{(\delta N)^{1/3}} \\
\leq &\left(G^2 + G + 17 + \Delta R_0\right)\frac{1}{(\delta N)^{1/3}} \leq \epsilon,
\end{aligned}$$

where the last inequality is due to the condition on N. The desired bound follows by noting that $\bar{w}_T \sim \text{Unif}(\{\bar{w}^{(k)} : k \in [K]\})$. The proof is completed. □

B.4 Proof of Theorem 2

Theorem 2. *Suppose that Assumption 1 and Assumption 2 hold. Let $\gamma \geq \rho$ be an arbitrary scalar. Suppose that $\eta \leq \frac{1}{8\gamma}$. Let K and T be positive integers and D be an arbitrary positive number. Then for any $\mu \leq \frac{\gamma}{8DT^2}$, the sequence $\{\bar{w}^{(k)}\}_{k=1}^{K}$ generated by Algorithm 2 with Option-II satisfies*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\partial R(\bar{w}^{(k)})\right\|_{+\mu}\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{1}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

Proof. Under the given conditions, for any $k \in [K]$, we can invoke Lemma 4 to Algorithm 2 (with Option-II) to get

$$\begin{aligned}
&\mathbb{E}\left[R(w_T^{(k)}) - R(w_0^{(k)}) + \sum_{t=1}^{T}\frac{\gamma}{8}\|\Delta_t^{(k)}\|^2\right] \\
\leq &-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2 + \frac{\|\Delta_1^{(k)}\|^2}{\eta} \\
\leq &-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2,
\end{aligned}$$

where in the last inequality we have used the fact $\|\Delta_1^{(k)}\| = 0$ according to the periodic restarting step in Option-II of Algorithm 2. Note that by definition we have $w_T^{(k)} = w_0^{(k+1)}$. Then the above implies that

$$\begin{aligned}&\mathbb{E}\left[R(w_0^{(k+1)}) - R(w_0^{(k)}) + \underbrace{\frac{\gamma}{8}\sum_{t=1}^{T}\|\Delta_t^{(k)}\|^2}_{A}\right]\\ \leq &-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)D^2.\end{aligned}$$

By applying Lemma 5 we can lower bound the term $A^{(k)}$ on the LHS of the above inequality as

$$A \geq \frac{\gamma}{8T}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2.$$

It follows that

$$\begin{aligned}&\mathbb{E}\left[R(w_0^{(k+1)}) - R(w_0^{(k)}) + \frac{\gamma}{8T}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2\right]\\ \leq &-\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\|\right] + \eta G^2 T + DG\sqrt{T} + \left(\frac{\gamma T}{2} + \frac{1}{\eta}\right)D^2.\end{aligned}$$

Rearranging the terms on both sides of the above inequality yields

$$\begin{aligned}&\mathbb{E}\left[DT\left\|\bar{g}^{(k)}\right\| + \frac{\gamma}{8T}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2\right]\\ \leq &\eta G^2 T + DG\sqrt{T} + \left(\frac{\gamma T}{2} + \frac{1}{\eta}\right)D^2 + \mathbb{E}\left[R(w_0^{(k)}) - R(w_0^{(k+1)})\right].\end{aligned}$$

By summing the above inequality of over $k \in [K]$ we get

$$\begin{aligned}&\mathbb{E}\left[DT\sum_{k=1}^{K}\left\|\bar{g}^{(k)}\right\| + \frac{\gamma}{8T}\sum_{k=1}^{K}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2\right]\\ \leq &\eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)KD^2 + \mathbb{E}\left[\sum_{k=1}^{K}\left(R(w_0^{(k)}) - R(w_0^{(k+1)})\right)\right]\\ = &\eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)KD^2 + \mathbb{E}\left[R(w_0^{(1)}) - R(w_0^{(K+1)})\right]\\ \leq &\eta G^2 KT + DGK\sqrt{T} + \left(\gamma T + \frac{1}{\eta}\right)KD^2 + R(w_0) - R^*.\end{aligned}$$

Finally, dividing the factor DKT on both sides of the above inequality yields

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(\left\|\bar{g}^{(k)}\right\| + \frac{\gamma}{8DT^2}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2\right)\right] \leq \frac{\eta G^2}{D} + \left(\gamma T + \frac{1}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT}.$$

Since $\mu \leq \mu' = \frac{\gamma}{8DT^2}$, in view of Lemma 3 we get

$$\left\|\partial R(\bar{w}^{(k)})\right\|_{+\mu} \leq \left\|\partial R(\bar{w}^{(k)})\right\|_{+\mu'} \leq \left\|\bar{g}^{(k)}\right\| + \frac{\gamma}{8DT^2}\max_{t\in[T]}\left\|w_t^{(k)} - \bar{w}^{(k)}\right\|^2.$$

Combining the preceding two inequalities leads to the desired result. The proof is completed. □

B.5 Proof of Corollary 2

Corollary 2. *Suppose that Assumption 1 and Assumption 2 hold. Let $\mu, \epsilon > 0$ be the desired regularized-stationarity parameters and N be the total budget of iterates. Set*

$$T = \left\lceil N^{4/7}\mu^{-2/7}\right\rceil, K = \left\lfloor\frac{N}{T}\right\rfloor,\ \gamma = N^{3/7}\mu^{2/7},\ \eta = \frac{1}{8N}.$$

Suppose that

$$N \geq \frac{(4G^2 + 1 + 32\Delta R_0)^{7/2}\mu^{1/2}}{\epsilon^{7/2}} + \frac{\rho^{7/3}}{\mu^{2/3}} + \mu^{1/2}.$$

Then the output $\bar{w}_T$ by Algorithm 2 with Option-II satisfies

$$\mathbb{E}\left[\|\partial R(\bar{w}_T)\|_{+\mu}\right] \leq \epsilon.$$

Proof. Under the conditions on N we can verify that

$$\gamma \geq \rho, \quad \eta\gamma = \frac{\mu^{2/7}}{8N^{4/7}} \leq \frac{1}{8}.$$

Let us now consider the number $D = \frac{1}{32}\mu^{-1/7}N^{-5/7}$. Again the condition on N implies that

$$T' := N^{4/7}\mu^{-2/7} \geq 1.$$

With the given choice of T, γ, D, it can be readily shown that

$$\frac{\gamma}{8DT^2} = \frac{\gamma}{8D\lceil T' \rceil^2} \geq \frac{\gamma}{8D(T'+1)^2} \geq \frac{\gamma}{32DT'^2} = \mu.$$

In view of the above arguments, the conditions of Theorem 2 are fulfilled in our setting, and thus we can apply it to obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\partial R(\bar{w}^{(k)})\right\|_{+\mu}\right] \leq & \frac{\eta G^2}{D} + \left(\gamma T + \frac{1}{\eta}\right)\frac{D}{T} + \frac{G}{\sqrt{T}} + \frac{\Delta R_0}{DKT} \\
\leq & \left(4G^2 + \frac{1}{32} + \frac{1}{4} + 32\Delta R_0\right)\frac{\mu^{1/7}}{N^{2/7}} \\
\leq & (4G^2 + 1 + 32\Delta R_0)\frac{\mu^{1/7}}{N^{2/7}} \leq \epsilon,
\end{aligned}$$

where in the last step we have used the condition on N. The desired bound follows by noting that $\bar{w}_T \sim \text{Unif}(\{\bar{w}^{(k)} : k \in [K]\})$. This proves the desired bound. □

C Analysis of online gradient descent for quadratic losses

Consider the quadratic loss functions of the form $f_t(x) = \langle u_t, x\rangle + \frac{\gamma}{2}\|x\|^2, t \geq 1$ over a convex constraint $\mathcal{C}$. We will analyze the following standard online gradient descent (OGD) method starting from an initial iterate x_1 with step-sizes $\eta > 0$:

$$x_{t+1} := \Pi_{\mathcal{C}}\left[x_t - \eta\nabla f_t(x_t)\right] = \Pi_{\mathcal{C}}\left[(1-\eta\gamma)x_t - \eta u_t\right], \tag{4}$$

where $\Pi_{\mathcal{C}}$ denotes the Euclidian projection operator associated with $\mathcal{C}$. Let $\text{Regret}_T(\bar{x})$ be the regret of algorithm w.r.t. some comparator $\bar{x} \in \mathcal{C}$ after T iterations, as defined below:

$$\text{Regret}_T(\bar{x}) := \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(\bar{x}).$$

Based on standard analysis, we can show the following result on the regret bound of the above OGD algorithm.

Lemma 6. *Suppose that $\eta\gamma \leq \frac{1}{8}$. Then the OGD procedure 4 applied on $\{f_t\}_{t=1}^T$ over a convex constraint $\mathcal{C}$ guarantees that for all $T \geq 1$ and $\bar{x}$:*

$$\textit{Regret}_T(\bar{x}) \leq \sum_{t=1}^{T}\left(\eta\|u_t\|^2 + \frac{\gamma}{2}\|\bar{x}\|^2 - \frac{\gamma}{8}\|x_t\|^2\right) + \frac{1}{\eta}\left(\|x_1\|^2 + \|\bar{x}\|^2\right).$$

Proof. First, it can be verified that

$$\begin{aligned}\|x_{t+1}-\bar{x}\|^2 &= \|\Pi_{\mathcal{C}}\left(x_t-\eta\nabla f_t(x_t)\right)-\bar{x}\|^2\\ &\le \|x_t-\eta\nabla f_t(x_t)-\bar{x}\|^2\\ &= \|x_t-\bar{x}\|^2+\eta^2\|\nabla f_t(x_t)\|^2-2\eta\langle\nabla f_t(x_t),x_t-\bar{x}\rangle,\end{aligned}$$

which implies

$$\langle\nabla f_t(x_t),x_t-\bar{x}\rangle=\frac{\|x_t-\bar{x}\|^2-\|x_{t+1}-\bar{x}\|^2}{2\eta}+\frac{\eta\|\nabla f_t(x_t)\|^2}{2}.$$

Then based on the strong convexity of f_t we can show that

$$\begin{aligned}\text{Regret}_T(x) &= \sum_{t=1}^T f_t(x_t)-\sum_{t=1}^T f_t(\bar{x})\\ &\le \sum_{t=1}^T\langle\nabla f_t(x_t),x_t-\bar{x}\rangle-\frac{\gamma}{2}\|x_t-\bar{x}\|^2\\ &\le \sum_{t=1}^T\left(\frac{\|x_t-\bar{x}\|^2-\|x_{t+1}-\bar{x}\|^2}{2\eta}-\frac{\gamma}{2}\|x_t-\bar{x}\|^2\right)+\sum_{t=1}^T\frac{\eta\|\nabla f_t(x_t)\|^2}{2}\\ &= -\sum_{t=1}^T\frac{\gamma}{2}\|x_t-\bar{x}\|^2+\frac{1}{2\eta}\|x_1-\bar{x}\|^2-\frac{1}{2\eta}\|x_{T+1}-\bar{x}\|^2+\sum_{t=1}^T\frac{\eta\|u_t+\gamma x_t\|^2}{2}\\ &\le -\sum_{t=1}^T\frac{\gamma}{2}\|x_t-\bar{x}\|^2+\frac{1}{\eta}\left(\|x_1\|^2+\|\bar{x}\|^2\right)+\sum_{t=1}^T\eta(\|u_t\|^2+\gamma^2\|x_t\|^2)\\ &\overset{\zeta_1}{\le} -\sum_{t=1}^T\frac{\gamma}{2}\left(\frac{\|x_t\|^2}{2}-\|\bar{x}\|^2\right)+\frac{1}{\eta}\left(\|x_1\|^2+\|\bar{x}\|^2\right)+\sum_{t=1}^T\eta(\|u_t\|^2+\gamma^2\|x_t\|^2)\\ &= \sum_{t=1}^T\left(\eta\|u_t\|^2+\frac{\gamma}{2}\|\bar{x}\|^2-\gamma\left(\frac{1}{4}-\eta\gamma\right)\|x_t\|^2\right)+\frac{1}{\eta}\left(\|x_1\|^2+\|\bar{x}\|^2\right)\\ &\le \sum_{t=1}^T\left(\eta\|u_t\|^2+\frac{\gamma}{2}\|\bar{x}\|^2-\frac{\gamma}{8}\|x_t\|^2\right)+\frac{1}{\eta}\left(\|x_1\|^2+\|\bar{x}\|^2\right),\end{aligned}$$

where in "ζ_1" we have used the fact $\|a-b\|^2\ge\frac{\|a\|^2}{2}-\|b\|^2$, and in the last inequality we have used the condition $\eta\gamma\le\frac{1}{8}$. This proves the desired bound. □

Remark 9. *The main message conveysed by Lemma 6 is that it is beneficial to control the scales of the competitor $\bar{x}$ and the initial x_1 to make the regret small, even the domain of interest is allowed to be unbounded. This result inspires us to explicitly control the scale of the initial iterate.*

D From Goldstein to Clarke stationarity

As a side contribution of our work, we have established in the following theorem a set of results on the connection between the Goldstein stationarity of a weakly convex function and the Clarke stationarity of its Moreau envelope, which are believed to be of independent interests.

Theorem 3. *Let f be a G-Lipschitz and ρ-weakly convex function.*

(a) If w is a (δ,ϵ)-stationary point of f, then it holds that

$$\left\|\nabla f_{1/(3\rho)}(w)\right\|\le 3\sqrt{\frac{\epsilon^2}{2}+4G\rho\delta+2\rho^2\delta^2}.$$

(b) If w is a (μ,ϵ)-regularized stationary point of f, then it holds that

$$\left\|\nabla f_{1/(3\rho)}(w)\right\|\le 3\sqrt{\frac{\epsilon^2}{2}+4G\rho\sqrt{\frac{\epsilon}{\mu}}+2\rho^2\frac{\epsilon}{\mu}}.$$

Proof. Part (a): Let w be a (δ,ϵ)-stationary point of f. Then by definition there exists a subset $V \subseteq \mathbb{B}_\delta(w)$ and $\{\alpha_v\}_{v\in V}$ such that $\alpha_v \geq 0, \sum_{v\in V}\alpha_v = 1$ and

$$\left\|\sum_{v\in V}\alpha_v g_v\right\| \leq \epsilon, \tag{5}$$

where $g_v \in \partial f(v)$. For any w', let us consider a subgradient $g' \in \partial f(w')$. Since f is ρ-weakly convex, we can show that

$$\begin{aligned}
f(w') =& \sum_{v\in V}\alpha_v f(w') \\
\geq& \sum_{v\in V}\alpha_v\left(f(v) + \langle g_v, w'-v\rangle - \frac{\rho}{2}\|w'-v\|^2\right) \\
=& f(w) + \left\langle \sum_{v\in V}\alpha_v g_v, w'-w\right\rangle + \sum_{v\in V}\alpha_v\left(f(v)-f(w)+\langle g_v, w-v\rangle - \frac{\rho}{2}\|w'-w+w-v\|^2\right) \\
\overset{\zeta_1}{\geq}& f(w) + \left\langle \sum_{v\in V}\alpha_v g_v, w'-w\right\rangle - \rho\|w'-w\|^2 + \sum_{v\in V}\alpha_v\left(f(v)-f(w)+\langle g_v, w-v\rangle - \rho\|w-v\|^2\right) \\
\overset{\zeta_2}{\geq}& f(w) - \frac{1}{4\rho}\left\|\sum_{v\in V}\alpha_v g_v\right\|^2 - 2\rho\|w'-w\|^2 - \sum_{v\in V}\alpha_v\left(2G\|v-w\| + \rho\|w-v\|^2\right) \\
\overset{\zeta_3}{\geq}& f(w) - 2\rho\|w'-w\|^2 - \frac{\epsilon^2}{4\rho} - 2G\delta - \rho\delta^2 \\
\geq& f(w') + \langle g', w-w'\rangle - \frac{\rho}{2}\|w-w'\|^2 - 2\rho\|w'-w\|^2 - \frac{\epsilon^2}{4\rho} - 2G\delta - \rho\delta^2 \\
=& f(w') + \langle g', w-w'\rangle - \frac{5\rho}{2}\|w-w'\|^2 - \frac{\epsilon^2}{4\rho} - 2G\delta - \rho\delta^2,
\end{aligned}$$

where we have used in "ζ_1" the Cauchy–Schwarz inequality, in "ζ_2" the Cauchy–Schwarz inequality and the G-Lipschitzness of R, in "ζ_3" $V \subseteq \mathbb{B}_\delta(w)$ and equation 5, and in the last inequality the ρ-weak-convexity of f. Now let us consider $\hat{w} = \text{prox}_{f/\bar{\rho}}(w)$ for some $\bar{\rho} > \frac{5\rho}{2}$, which by Lemma 1 satisfies that

$$\nabla f_{1/\bar{\rho}}(w) = \bar{\rho}(w-\hat{w}) \in \partial f(\hat{w}).$$

Subsisting $w' = \hat{w}$ into the preceding inequality and rearranging the terms yields

$$\|w-\hat{w}\|^2 \leq \left(\bar{\rho} - \frac{5\rho}{2}\right)^{-1}\left(\frac{\epsilon^2}{4\rho} + 2G\delta + \rho\delta^2\right).$$

It follows from the above inequality that

$$\left\|\nabla f_{1/\bar{\rho}}(w)\right\| = \|\bar{\rho}(w-\hat{w})\| \leq \bar{\rho}\left(\bar{\rho} - \frac{5\rho}{2}\right)^{-1/2}\left(\frac{\epsilon^2}{4\rho} + 2G\delta + \rho\delta^2\right)^{1/2}.$$

Finally, setting $\bar{\rho} = 3\rho$ in the above and applying some slight algebraic manipulation yields the desired bound in Part (a). The bound in Part (b) follows directly from Part(a) and Lemma 2. The proof is completed. □

Remark 10. *Theorem 3 essentially shows that the (δ,ϵ)-stationarity of a weakly convex function implies the $(\epsilon+\sqrt{\delta})$-stationarity of its Moreau envelope, and correspondingly the (μ,ϵ)-regularized stationary implies the $(\epsilon+\sqrt{\epsilon/\mu})$-stationarity.*

Remark 11. *Conversely, for a ρ-weakly convex function f, the translate from the Clarke stationarity of its Moreau envelop to the Goldstein stationarity of the original objective is relatively straightforward. Indeed, suppose that w is an ϵ-stationary point of the Moreau envelope $f_{1/(2\rho)}$ such that $\|\nabla f_{1/(2\rho)}(w)\| \leq \epsilon$. Consider $\hat{w} := prox_{f/(2\rho)}(w)$. Then according to Lemma 1 we must have*

$$\nabla f_{1/(2\rho)}(w) \in \partial f(\hat{w}),\ \|w-\hat{w}\| \leq \frac{\|\nabla f_{1/(2\rho)}(w)\|}{2\rho} \leq \frac{\epsilon}{2\rho},$$

which implies that dist $\left(0, \partial_{\frac{\epsilon}{2\rho}} f(w)\right) \leq \|\nabla f_{1/(2\rho)}(w)\| \leq \epsilon$, *and thus* w *is a* (δ, ϵ)*-stationary point of* f *with* $\delta = \frac{\epsilon}{2\rho}$. *However, a limitation of translating rates via the setting* $\delta = \epsilon/(2\rho)$ *is that it excludes the range of relatively large* δ *(e.g.,* $\delta = \sqrt{\epsilon}$*, a choice critical for sharper rates in second-order smooth functions), as* ρ *is typically lower bounded by a constant.*

The following corollary is a direct consequence of Theorem 3 when applied to Algorithm 2 with Option-I.

Corollary 3. *Suppose that Assumption 1 and Assumption 2 hold. Let* $\epsilon > 0$ *be the desired stationarity precision and* N *be the total budget of iterates. Set*

$$T = \left\lceil (\epsilon^2 N)^{2/3} \right\rceil, K = \left\lfloor \frac{N}{T} \right\rfloor, \gamma = \frac{N^{1/3}}{\epsilon^{4/3}}, \eta = \frac{1}{8N}, D = \frac{\epsilon^{2/3}}{N^{2/3}}.$$

Suppose that N *is sufficiently large such that*

$$N \geq \frac{(G^2 + G + 17 + \Delta R_0)^3}{\epsilon^5} + \rho^3\epsilon^4 + \frac{1}{\epsilon^2}.$$

Then the output $\bar{w}_T$ *by Algorithm 2 with Option-I satisfies*

$$\mathbb{E}\left[\left\|\nabla f_{1/(3\rho)}(\bar{w}_T)\right\|\right] \leq \mathcal{O}\left(\sqrt{G\rho}\epsilon + \rho\epsilon^2\right).$$

Proof. Let $\delta = \epsilon^2$ and $\varepsilon(\delta, \bar{w}_T) := \text{dist}\,(0, \partial_\delta R(\bar{w}_T))$. Under the given conditions, it follows from Corollary 1 that

$$\mathbb{E}\left[\varepsilon(\delta, \bar{w}_T)\right] = \mathbb{E}\left[\text{dist}\,(0, \partial_\delta R(\bar{w}_T))\right] \leq \epsilon. \tag{6}$$

Conditioned on $\bar{w}_T$, it is natural that $\bar{w}_T$ is a $(\delta, \varepsilon(\delta, \bar{w}_T))$-stationary point of R. Therefore from the Part (a) of Theorem 3 we have

$$\left\|\nabla R_{1/(3\rho)}(\bar{w}_T)\right\| \leq 3\sqrt{\frac{\varepsilon^2(\delta, \bar{w}_T)}{2} + 4G\rho\delta + 2\rho^2\delta^2} \leq \frac{3\sqrt{2}}{2}\varepsilon(\delta, \bar{w}_T) + 6\sqrt{G\rho\delta} + 3\sqrt{2}\rho\delta.$$

Taking expectation on both sides of the above yields

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla R_{1/(3\rho)}(\bar{w}_T)\right\|\right] \leq & \mathbb{E}\left[\frac{3\sqrt{2}}{2}\varepsilon(\delta, \bar{w}_T) + 6\sqrt{G\rho\delta} + 3\sqrt{2}\rho\delta\right] \\ \leq & \frac{3\sqrt{2}}{2}\epsilon + 6\sqrt{G\rho}\epsilon + 3\sqrt{2}\rho\epsilon^2,\end{aligned}$$

where in the last step we have used 6 and $\delta = \epsilon^2$. This proves the desired bound. □

Remark 12. *The* $\mathcal{O}(\epsilon^{-5})$ *complexity established in Corollary 3 is suboptimal compared to the* $\mathcal{O}(\epsilon^{-4})$ *optimal complexity of SGD (Davis & Grimmer, 2019) and SGDM (Mai & Johansson, 2020) for achieving the* ϵ*-stationarity of the Moreau envelope. Such a slower rate is mainly due to the* $\sqrt{\delta}$ *component appeared in the bound of Theorem 3 (Part a), which is open for improvement in future.*

Similarly, we have the following corollary as a direct consequence of Theorem 3 when applied to Algorithm 2 with Option-II.

Corollary 4. *Suppose that Assumption 1 and Assumption 2 hold. Let* $\epsilon > 0$ *be the desired stationarity precision and* N *be the total budget of iterates. Set*

$$T = \left\lceil N^{4/7}\epsilon^{6/7} \right\rceil, K = \left\lfloor \frac{N}{T} \right\rfloor, \gamma = N^{3/7}\epsilon^{-6/7}, \eta = \frac{1}{8N}.$$

Suppose that

$$N \geq \frac{(4G^2 + 1 + 32\Delta R_0)^{7/2}}{\epsilon^5} + \rho^{7/3}\epsilon^2 + \frac{1}{\epsilon^{3/2}}.$$

Then the output $\bar{w}_T$ *by Algorithm 2 with Option-II satisfies*

$$\mathbb{E}\left[\left\|\nabla f_{1/(3\rho)}(\bar{w}_T)\right\|\right] \leq \mathcal{O}\left(\sqrt{G\rho}\epsilon + \rho\epsilon^2\right).$$

Proof. The proof basically mimics that of Corollary 3 and is restated for the sake of completeness. Let $\mu = \epsilon^{-3}$ and $\varepsilon(\mu, \bar{w}_T) := \|\partial R(\bar{w}_T)\|_{+\mu}$. Under the given conditions, it follows from Corollary 2 that

$$\mathbb{E}\left[\varepsilon(\mu, \bar{w}_T)\right] = \mathbb{E}\left[\|\partial R(\bar{w}_T)\|_{+\mu}\right] \leq \epsilon. \tag{7}$$

Conditioned on $\bar{w}_T$, it is natural that $\bar{w}_T$ is a $(\delta, \varepsilon(\delta, \bar{w}_T))$-stationary point of R. Therefore from the Part (b) of Theorem 3 we have

$$\left\|\nabla R_{1/(3\rho)}(\bar{w}_T)\right\| \leq 3\sqrt{\frac{\varepsilon^2(\mu, \bar{w}_T)}{2} + 4G\rho\sqrt{\frac{\epsilon}{\mu}} + 2\rho^2\frac{\epsilon}{\mu}} \leq \frac{3\sqrt{2}}{2}\varepsilon(\mu, \bar{w}_T) + 6\sqrt{G\rho\sqrt{\frac{\epsilon}{\mu}}} + 3\sqrt{2}\rho\sqrt{\frac{\epsilon}{\mu}}.$$

Taking expectation on both sides of the above yields

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla R_{1/(3\rho)}(\bar{w}_T)\right\|\right] \leq & \mathbb{E}\left[\frac{3\sqrt{2}}{2}\varepsilon(\mu, \bar{w}_T) + 6\sqrt{G\rho\sqrt{\frac{\epsilon}{\mu}}} + 3\sqrt{2}\rho\sqrt{\frac{\epsilon}{\mu}}\right] \\ \leq & \frac{3\sqrt{2}}{2}\epsilon + 6\sqrt{G\rho}\epsilon + 3\sqrt{2}\rho\epsilon^2,\end{aligned}$$

where in the last step we have used 7 and $\mu = \epsilon^{-3}$. This proves the desired bound. □

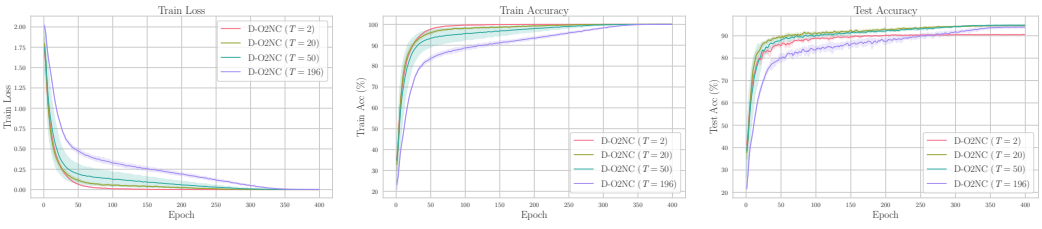
E Some additional details and results of experiment

In this appendix section, we present additional experimental results on neural networks (Appendix E.1) and robust phase retrieval (Appendix E.2) to further validate the effectiveness and efficiency of our D-O2NC method when applied with periodically restarted OGD.

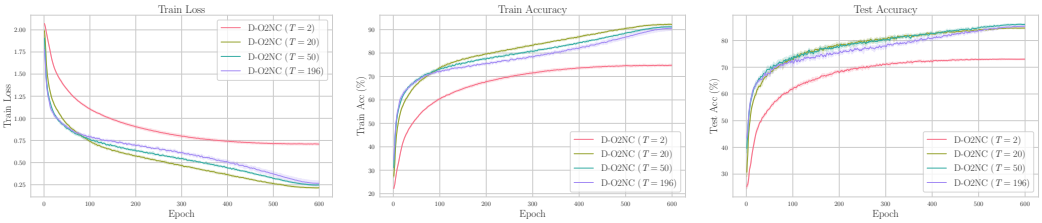
E.1 Experiments on Neural Networks

Descriptions of backbones. We employ the ResNet-101 and ViT models to evaluate our method. ResNet-101 stands as a hallmark architecture in the ResNet family, featuring 101 layers formed by stacking residual blocks, each composed of 1×1, 3×3, and 1×1 convolutional layers. This model is commonly adopted as a backbone in downstream computer vision applications, including object detection and image segmentation. In our empirical study, the ViT model incorporates 6 Transformer encoder layers, each equipped with 8 multi-head self-attention heads and a 512-dimensional multilayer perceptron (MLP); the dropout rate is configured at 0.1, with the input segmented into 4 patches. Both models were trained from scratch.

Results under various restarting frequency. In our experiments on the CIFAR-10 dataset, we configured the restarting frequency T to a broad value range of $\{2, 20, 50, 196\}$, and the total number of minibatches in one epoch is 196. Consistent with the parameter settings in the main text, we adopt a learning rate of 0.01 and a momentum of 0.99. As illustrated in Figure 2, the experimental results reveal that in most cases, as T increases, the model's performance exhibits an initial gradual improvement followed by a subsequent decline. Notably, extreme values of T (e.g., $T = 2$) exert a detrimental impact on performance. Therefore, selecting an appropriate T is crucial for optimizing the final model efficacy.

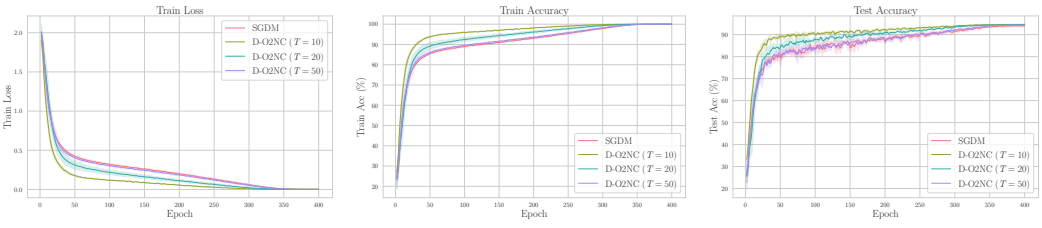


(a) Train Loss (ResNet-101) (b) Train Accuracy (ResNet-101) (c) Test Accuracy (ResNet-101)

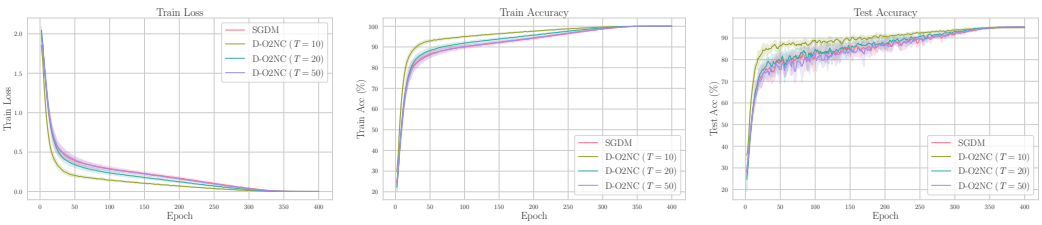


(d) Train Loss (ViT) (e) Train Accuracy (ViT) (f) Test Accuracy (ViT)

Figure 2: Experimental results on CIFAR-10 with ResNet-101 (top) and ViT (bottom) networks under various values of T.



(a) Train Loss (ResNet-101) (b) Train Accuracy (ResNet-101) (c) Test Accuracy (ResNet-101)



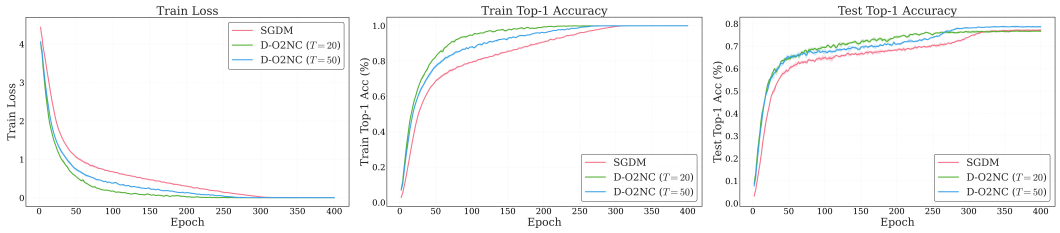
(d) Train Loss (ResNet-101) (e) Train Accuracy (ResNet-101) (f) Test Accuracy (ResNet-101)

Figure 3: Experimental results of with relaxed momentum and learning rate parameters (The top two rows correspond to momentum coefficient 0.95 and learning rate 0.05, and the bottom row to momentum coefficient 0.9 and learning rate 0.1).

Results under various configurations of momentum coefficient and learning rate. For the experimental results reported in the main text, we have adopted a tight momentum coefficient of value 0.99, along with a learning rate of 0.01. In this section, we additionally report the experimental results obtained using two other configurations of the momentum coefficient and learning rate: $(0.95, 0.05)$ and $(0.9, 0.1)$. The results are presented in Figure 3, from which we can observe that our method consistently demonstrates considerable superiority in both convergence speed and prediction accuracy. Additionally, it is evident that for these two parameter configurations, the superiority of the method decreases as T increases.

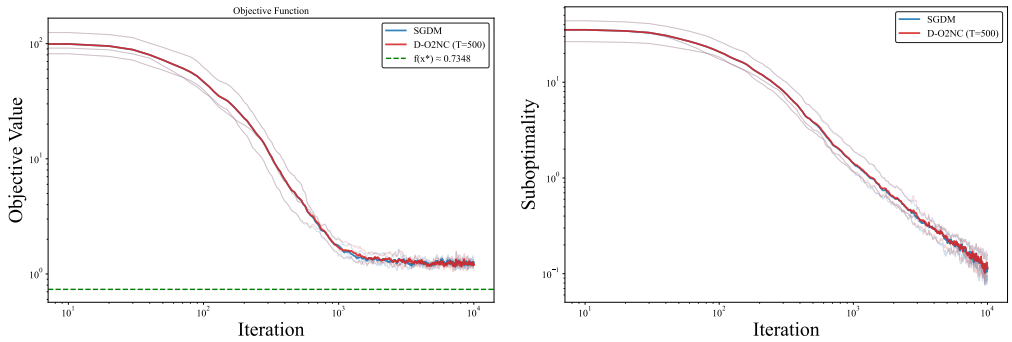
Results on CIFAR-100. Finally, in addition to CIFAR-10, we have also conducted the algorithm evaluation on the CIFAR-100 dataset. CIFAR-100 is an advanced counterpart of CIFAR-10, com-

prising 60,000 32×32 color images. While CIFAR-10 contains 10 coarse categories, CIFAR-100 extends this to 100 fine-grained classes. For this more fine-grained dataset, we resort ResNet-152 as the backbone network. The optimizer employs a learning rate of 0.01, a momentum of 0.99, and a weight decay of 5×10^{-4}. The experimental results are demonstrated in Figure 4. It can be observed from this set of results that 1) our D-O2NC method converges faster than the standard SGDM in terms of training loss and accuracy; and 2) our D-O2NC method achieves higher test accuracy than SGDM, which further demonstrates the superiority of our algorithm for generalization.



(a) Train Loss (b) Train Accuracy (c) Test Accuracy

Figure 4: Experimental results on CIFAR-100 with ResNet-152 network.



(a) $\kappa = 1$ (b) $\kappa = 10$

Figure 5: Experimental results of phase retrieval with various value of κ.

E.2 Experiments on Robust Phase Retrieval

To further validate the effectiveness of our D-O2NC method for weakly convex problems, we have conducted an additional set of experiments on the robust phase retrieval task (Duchi & Ruan, 2018). Given a set of m measurement-amplitude pairs $\{x_i, y_i\}_{i=1}^m$, robust phase retrieval is conventionally formulated as the following composite loss minimization problem:

$$\mathcal{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \left| (w^\top x_i)^2 - y_i \right|,$$

where w denotes the unknown signal to be recovered, x_i represent the measurement vectors, y_i are the observed squared amplitude values. The considered loss function quantifies the discrepancy between the predicted squared amplitude of inner products and the actual observed data. Clearly, it is of the composition form $h \circ c$ where $h(u) = |u|$ is Lipschitz continuous and $c(w) = (w^\top x_i)^2 - y_i$ is a smooth mapping. Therefore, the loss function is weakly convex.

Following the related experimental setup of Mai & Johansson (2020), we generate a measurement matrix of size 300×100 using $X = QP$, where P is diagonal matrix with condition number $\kappa \in \{1, 10\}$, and Q is a matrix whose entries are *i.i.d.* The observation noise follows $\mathcal{N}(0, 25)$

and affects 20% of the data points. The considered SGDM and D-O2NC methods employ shared momentum coefficient of 0.92 and learning rate of 2×10^{-4}, and the total number of iteration is set as 10,000. Figure 5 plots the sub-optimality gap as a function of the iteration count for two distinct values of κ. It can be observed from this group of results that, for both values of κ, our method achieves performance comparable to that of SGDM under the same learning rate and momentum coefficient.