ONE-PROX RANDOM RESHUFFLING FOR NONSMOOTH NONCONVEX OPTIMIZATION: COMPLEXITY AND KL LAST-ITERATE RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Random reshuffling (RR) has been widely applied to large-scale finite-sum optimization. However, existing proximal RR methods for such problems with nonsmooth structure typically either require multiple proximal evaluations per epoch or couple the proximal parameter with the reshuffling stepsize. In this paper, we propose a proximal algorithm for nonsmooth nonconvex random reshuffling (PNRR) that simultaneously addresses these two issues. PNRR requires only one proximal evaluation per epoch while allowing the proximal parameter to be chosen independently of the reshuffling stepsize and kept constant. We establish finite-time complexity guarantees for PNRR, including an incremental first-order oracle complexity of $\mathcal{O}(n\epsilon^{-2})$ and a proximal oracle complexity of $\mathcal{O}(\epsilon^{-2})$. Compared with several representative methods, the complexity advantage of PNRR becomes more pronounced in the high-accuracy regime. Under the Kurdyka–Łojasiewicz (KL) framework, we further prove that the whole sequence generated by PNRR converges to a critical point. Moreover, under polynomially decaying stepsizes, we obtain explicit polynomial convergence rates for the function values, stationarity measure, and iterate error, while exponentially decaying stepsizes yield geometric convergence of all three quantities when the KL exponent is below $1/2$. Numerical experiments on logistic regression and nonconvex binary classification demonstrate the competitive and stable performance of PNRR.

1 INTRODUCTION

In this paper, we consider the following optimization problem:

$$\min_{x\in\mathbb{R}^d} P(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h_1(x) - h_2(x), \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $h_1 : \mathbb{R}^d \to (-\infty, \infty]$ and $h_2 : \mathbb{R}^d \to \mathbb{R}$ are proper, lower semicontinuous (lsc.) and closed. For simplicity, we denote $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. The finite-sum term $f(x)$ naturally models data fitting or empirical loss objectives arising from large-scale datasets, where each component function f_i typically corresponds to the loss associated with one data sample. Such a structure appears broadly in empirical-risk minimization and neural network training (Bottou, 2012; Bottou et al., 2018; Mishchenko et al., 2020; Koloskova et al., 2024), statistical learning (Shamir, 2016; Lok et al., 2025), signal processing and distributed learning (Ying et al., 2019; Huang et al., 2023; Mishchenko et al., 2022; Yun et al., 2022; Sadiev et al., 2024). The additional term $h_1 - h_2$ enriches the finite-sum model by incorporating sparsity, constraints, nonsmooth and nonconvex regularization. Such formulations arise in sparse optimization (Gotoh et al., 2018), nonconvex classification (Qiu et al., 2025), and structured large-scale learning (Duchi & Singer, 2009; Ghadimi et al., 2016). However, solving problem (1) presents two main challenges. First, computing the full gradient ∇f is expensive when n is large. Besides, the nonsmooth and nonconvex structure induced by $h_1 - h_2$ makes the problem more difficult. A common approach to attack such nonsmooth difference structured problems is based on full gradient algorithm, which can be summarized as

$$\begin{cases} w_t \in \partial h_2(x_t), \\ x_{t+1} \in \operatorname{prox}_{\alpha h_1}(x_t - \alpha\nabla f(x_t) + \alpha w_t). \end{cases} \tag{2}$$

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

However, each iteration of (2) computes all n component gradients at the same point x_t before updating the iterate, and therefore cannot exploit the information from already evaluated samples during this process. Motivated by these two challenges, we next focus on stochastic gradient descent (SGD) and random reshuffling (RR) algorithms, which avoid computing the full gradient while making use of each sampled component as soon as its gradient becomes available.

SGD and its proximal variants. Classical SGD avoids the expensive computation of the full gradient by updating the iterate using the gradient of a randomly selected component function f_i. This made it widely used in large-scale machine learning and stochastic optimization (Robbins & Monro, 1951; Ghadimi & Lan, 2013; Bottou et al., 2018). When a nonsmooth term is present, proximal SGD (proxSGD) extends this principle to composite optimization by applying a proximal mapping after each stochastic gradient update (Duchi & Singer, 2009; Ghadimi et al., 2016; Davis & Drusvyatskiy, 2019; Gorbunov et al., 2020). However, SGD and its proximal variants have several limitations. The with-replacement sampling strategy may repeatedly select some component functions while failing to visit others within a given period of iterations. As a result, the available data may not be used evenly, resulting in more oscillatory convergence behavior. Moreover, the proximal parameter in proxSGD is directly coupled with the stochastic gradient stepsize and usually decreases to control stochastic errors and achieve asymptotic convergence. This coupling always leads to a relatively high computational complexity. Davis & Drusvyatskiy (2019) indicated that proxSGD requires $\mathcal{O}\left(\frac{L}{\epsilon^2}\max\left\{1, \frac{B^2}{\epsilon^2}\right\}\right)$ to reach an ϵ-stationary point, where B is a constant. Thus, when $B > 0$ and ϵ is sufficiently small, the complexity is dominated by the $\mathcal{O}(\epsilon^{-4})$ term.

Random reshuffling (RR). A natural direction for addressing these limitations is RR. At the beginning of each epoch, RR generates a random permutation of all component functions and then processes them sequentially according to this permutation. Consequently, every component function is visited exactly once in each epoch, avoiding the repeated or missing samples that may occur in with-replacement SGD. This sampling strategy has motivated extensive theoretical developments for random reshuffling and without-replacement SGD (Ying et al., 2019; Haochen & Sra, 2019; Nagaraj et al., 2019; Rajput et al., 2020; Mishchenko et al., 2020; Gürbüzbalaban et al., 2021; Li et al., 2023), as well as for Shuffle-Once or Single Shuffle schemes (Safran & Shamir, 2020; Ahn et al., 2020; Mishchenko et al., 2020). More general shuffling methods, including cyclic, incremental-gradient, and momentum variants, have also been investigated (Nguyen et al., 2021; Tran et al., 2021). For problems involving nonsmooth structure, the proximity operator needs to be incorporated into the reshuffling procedure. Existing proximal RR methods can be roughly divided into two types. The first type performs only one proximal evaluation per epoch. A representative method is proxRR (Mishchenko et al., 2022), stated in Algorithm 2 placed in Appendix D.1, which reduces the number of proximal evaluations. However, its proximal parameter $n\gamma_t$ is coupled with the reshuffling stepsize γ_t and therefore decreasing γ_t also forces the proximal parameter to decrease, which leads to a relatively high computational complexity. The second type allows the proximal parameter to be chosen independently of the reshuffling stepsize. A representative method is norm-PRR (Qiu et al., 2025), which introduces a separate proximal parameter through a normal-map reformulation. However, norm-PRR evaluates the proximity operator after processing each component function and thus requires n proximal evaluations per epoch, which leads to expensive computation. Recently, several works have considered decoupled stepsizes in RR methods, including Nastya (Malinovsky et al., 2023) and its federated extension under generalized smoothness (Demidovich et al., 2025); however, these studies are restricted to smooth problems.

Contributions. Motivated by the observations above, we would like to propose a RR algorithm in this work to achieve the following targets. First, we develop a method that requires only one proximal evaluation per epoch, while allowing the proximal parameter to remain constant and independent of the reshuffling stepsize. Moreover, we aim to establish complexity and convergence guarantees through a lifted criticality measure that captures the stationarity of the difference structured objective. Furthermore, we seek to establish convergence of the whole sequence under the Kurdyka-Łojasiewicz (KL) framework. Specifically, the main contributions of this work are summarized as follows.

- **Proximal algorithm for nonsmooth nonconvex RR.** We develop a proximal algorithm for nonsmooth nonconvex random reshuffling (PNRR) for the nonsmooth nonconvex composite problem (1). PNRR requires only one proximal evaluation per epoch, while allowing the proximal parameter α to remain fixed and independent of the reshuffling stepsize γ_t.

Table 1: Comparison of IFO and PO complexity and convergence guarantees.*

Algorithm	IFO	PO	Global Conv.	KL Rate
RR[(a)]	$\frac{Ln}{\epsilon^2}\max\left\{1, \frac{\sqrt{A}+B}{\sqrt{n}\,\epsilon}\right\}$	–	$\|\nabla f(x_t)\| \to 0$	$\|x_t - x^*\| = \mathcal{O}(t^{-1})$
proxRR[(b)]	$\frac{Ln}{\epsilon^2}\max\left\{1, \frac{B}{\sqrt{n}\,\epsilon}, \frac{\zeta}{\sqrt{L}\,\epsilon}\right\}$	$\frac{L}{\epsilon^2}\max\left\{1, \frac{B}{\sqrt{n}\,\epsilon}, \frac{\zeta}{\sqrt{L}\,\epsilon}\right\}$	×	×
norm-PRR[(c)]	$\frac{Ln}{\epsilon^2}\max\left\{1, \frac{\sqrt{L}}{\sqrt{n}\,\epsilon}\right\}$	$\frac{Ln}{\epsilon^2}\max\left\{1, \frac{\sqrt{L}}{\sqrt{n}\,\epsilon}\right\}$	$\mathrm{dist}(0, \partial\psi(x_t)) \to 0$	$\|x_t - x^*\| = \mathcal{O}(t^{-1})$ $\|\psi(x_t) - \psi^*\| = \mathcal{O}(t^{-2})$
proxSGD[(d)]	$\frac{L}{\epsilon^2}\max\left\{1, \frac{B^2}{\epsilon^2}\right\}$	$\frac{L}{\epsilon^2}\max\left\{1, \frac{B^2}{\epsilon^2}\right\}$	$\|G_\lambda(x_t)\| \to 0$	×
PNRR[(e)]	$\mathcal{O}(\frac{n}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathrm{dist}(0, \partial P(x_t)) \to 0$	$\|x_t - x^*\| = \mathcal{O}(t^{-q_x})$ $\|P(x_t) - P^*\| = \mathcal{O}(t^{-q_f})$

∗ An IFO call denotes one evaluation of a component gradient ∇f_i, and a PO call denotes one evaluation of a proximity operator. For a fair comparison with existing methods, the PNRR results reported in this table are specialized to the case $h_2 = 0$. The complexities are reported under the criterion that the unsquared stationarity residual is at most ϵ. For results originally stated in terms of a squared stationarity measure, the accuracy parameter is replaced by ϵ^2 accordingly. The best performance is highlighted in blue.
(a) In the smooth setting with $h_1 = h_2 = 0$, L denotes the common Lipschitz constant of the component gradients. The IFO is based on the variance condition $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \le 2A[f(x) - f_{\mathrm{lb}}] + B^2$. The global convergence result and the KL rate are from (Li et al., 2023).
(b) The result in (Mishchenko et al., 2022) holds for the choice $\lambda = \frac{1}{L}\min\left\{\frac{1}{5}, \frac{\varepsilon}{n^{-\frac{1}{2}}B + L^{-\frac{1}{2}}\zeta}\right\}$ if $\|\nabla f(x)\|^2 \le \|\mathcal{G}_\lambda(x)\|^2 + \zeta^2$ for all $x \in \mathrm{dom}(\varphi)$ and if $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \le B^2$ for all $x \in \mathbb{R}^d$.
(c) The result for norm-PRR (Qiu et al., 2025) does not require an explicit bounded-variance assumption. The displayed KL rates correspond to polynomial stepsizes $\alpha_t = \frac{\alpha}{t+\beta}$ and KL exponent $\theta \in [0, \frac{1}{2}]$.
(d) The IFO and PO complexities are from (Davis & Drusvyatskiy, 2019) under $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \le B^2$ and $\lambda \lesssim \frac{1}{\rho}$. The global convergence results are from (Duchi & Ruan, 2018; Majewski et al., 2018; Davis et al., 2020; Li & Milzarek, 2022). Here, L denotes the Lipschitz constant of ∇f.
(e) Our work does not make any explicit bounded variance assumptions. For $\gamma_t = \frac{\gamma_0}{(t+\beta)^p}$ with $p > 1$, $q_x = p - 1$ and $q_f = 2p - 1$ if $\theta \le \frac{p}{2p-1}$; otherwise, $q_x = \frac{1-\theta}{2\theta-1}$ and $q_f = \frac{1}{2\theta-1}$, see Theorem 3.4. For $\gamma_t = \gamma_0\rho^t$ and $\theta < \frac{1}{2}$, the corresponding rates are $\mathcal{O}(\rho^t)$ and $\mathcal{O}(\rho^{2t})$, respectively, see Theorem 3.6.

This design reduces the proximal computational cost while avoiding the shrinkage of the proximal parameter when the reshuffling stepsize decreases.

- **Deterministic full sequence convergence under KL framework.** The difference structure $h_1 - h_2$ introduces additional difficulties into the convergence analysis. Exploiting the convexity of h_2, we introduce the lifted auxiliary function and relate its critical points to those of problem (1). Meanwhile, we use a tail term to absorb the reshuffling errors and recover sufficient descent for the auxiliary function. Under the KL property in Definition A.2, this construction allows us to establish convergence of the whole sequence of iterates. These conclusions differ from existing RR analyses under the KL framework. In particular, we extend the pathwise convergence perspective for RR to problems with a nonsmooth difference structure, showing that convergence and KL rates can be established along each realized reshuffling trajectory without taking expectations over the permutations. The finite-length property also characterizes iterate stabilization, consistent with the reduced oscillations (see Section 4), even though the original objective need not decrease monotonically.
- **Complexity analysis.** We derive finite-time complexity bounds for PNRR in the nonsmooth nonconvex setting, and quantify its computational cost in terms of both the incremental first-order oracle (IFO) and the proximal oracle (PO). We establish complexity and convergence guarantees in terms of a lifted criticality measure tailored to the difference structured objective. In particular, when $h_2 = 0$, the lifted measure reduces to $\mathrm{dist}(0, \partial P(x_t)) \to 0$, which provides a more direct stationarity measure than the natural residual.[1] PNRR attains an IFO complexity of $\mathcal{O}(n\epsilon^{-2})$ and a PO complexity of $\mathcal{O}(\epsilon^{-2})$ for the difference structured model, with the latter benefiting from the single proximal evaluation per epoch. For a fair comparison with existing proximal RR and SGD methods, we

[1] When $h_2 = 0$, problem (1) reduces to $P = f + h_1$, and the lifted criticality measure reduces to $\mathrm{dist}(0, \partial P(x_t)) \to 0$. For the proximal-gradient point $\widehat{x}_t = \mathrm{prox}_{\lambda h_1}(x_t - \lambda\nabla f(x_t))$, the natural residual is defined as $G_\lambda(x_t) := \frac{x_t - \widehat{x}_t}{\lambda} \in \nabla f(x_t) + \partial h_1(\widehat{x}_t)$ and therefore need not belong to $\partial P(x_t) = \nabla f(x_t) + \partial h_1(x_t)$. Thus, $\mathrm{dist}(0, \partial P(x_t)) \to 0$ is a more direct and suitable measure for guaranteeing the convergence than natural residual.

Algorithm 1 **P**roximal Algorithm for **N**onsmooth Nonconvex **R**andom **R**eshuffling (PNRR)

Input: Stepsizes $\gamma_t > 0$, $\alpha > 0$, initial vector $x_0 \in \operatorname{dom} h_1$, number of epochs T

for epochs $t = 0, 1, \cdots, T-1$ **do**
  Sample a permutation $\pi = (\pi_0, \pi_1, \cdots, \pi_{n-1})$ of $[n]$
  $w_t \in \partial h_2(x_t)$
  $x_t^0 = x_t$
  for $i = 0, 1, \cdots, n-1$ **do**
    $x_t^{i+1} = x_t^i - \gamma_t \nabla f_{\pi_i}(x_t^i)$
  end for
  $x_{t+1} \in \operatorname{prox}_{\alpha h_1}(x_t + \alpha \frac{x_t^n - x_t}{\gamma_t n} + \alpha w_t)$
end for

specialize PNRR to the subclass $h_2 = 0$ and report the complexity and convergence guarantees together with those of related methods in Table 1. Moreover, under the KL property, we derive explicit last-iterate convergence rates for both polynomial and exponential decreasing reshuffling stepsizes. For polynomial stepsizes $\gamma_t = \frac{\gamma_0}{(t+\beta)^p}$ with $p > 1$, the function gaps decay as $\mathcal{O}(t^{-(2p-1)})$ when the KL exponent $\theta \leq \frac{p}{2p-1}$ and as $\mathcal{O}(t^{-\frac{1}{2\theta-1}})$ otherwise. For exponentially decreasing stepsizes $\gamma_t = \gamma_0 \rho^t$ with $\theta < \frac{1}{2}$, we further obtain geometric rates $\mathcal{O}(\rho^{2t})$ for the function gaps and $\mathcal{O}(\rho^t)$ for the stationarity measure and iterate error. These results characterize the asymptotic behavior and provide explicit polynomial and geometric convergence rates under different stepsize schedules.

The remaining part of this paper is organized as follows. In section 2, we propose the PNRR algorithm and study the convergence properties, while establishing complexity bounds in terms of both IFO and the PO. Section 3 shows the full sequence convergence under the KL framework and derives the corresponding KL convergence rates. Numerical experiments are provided in Section 4. Finally, we conclude the key findings of the paper in Section 5.

2 Proximal Algorithm for Nonsmooth Nonconvex Random Reshuffling

Motivated by the above observations and the limitations of SGD, existing RR methods, and full-gradient methods, we propose a proximal algorithm for nonsmooth nonconvex random reshuffling (PNRR) for solving problem (1), as summarized in Algorithm 1. Note that PNRR differs from proxRR only in the proximal update at the end of each epoch, but this modification yields two important advantages. PNRR decouples the proximal parameter α from the reshuffling stepsize γ_t, allowing α to be chosen as a constant independently. Moreover, PNRR provides a structure that enables us to establish global convergence of the whole iterate sequence under the KL framework.

In the following, we discuss the setting where h_1 is nonsmooth and nonconvex while h_2 is nonsmooth and convex. First, we assume that problem (1) satisfies the following assumptions.

Assumption 2.1. *We consider the following conditions:*

(i) f_i *is* L_i-*smooth, defined* $L := max_{1 \leq i \leq n} L_i$.

(ii) h_1 *is possibly nonsmooth and nonconvex,* h_2 *is convex and nonsmooth.*

(iii) The objective function $P(x) = f(x) + h_1(x) - h_2(x)$ *is bounded from below: there exist* f_{lb}, h_{lb} *and* $P_{lb} := f_{lb} + h_{lb}$ *such that*

$$f_i(x) \geq f_{lb},\ h_1(x) - h_2(x) \geq h_{lb} \ \text{ and } \ P(x) \geq P_{lb}$$

for all $x \in \mathbb{R}^d$, $i = 1, \cdots, n$.

Conditions (i)–(iii) are standard or natural in nonsmooth nonconvex finite-sum optimization; see, e.g., (Ghadimi et al., 2016; Qiu et al., 2025). Condition (i) is used to control the reshuffling error, while condition (ii) captures the nonsmooth difference structure and, in particular, the convexity of

h_2 is crucial for our conjugate-based reformulation. Condition (iii) is essential for the convergence analysis. Combined with Condition (i) and the descent lemma, it can derive the bound for the gradients $\|\nabla f_i\|^2$.

In the following, we first establish a descent property of the objective function and then characterize the critical points of the problem. By further deriving bounds on the variance and reshuffling error, we establish the subsequential convergence of the generated iterates. Specifically, the difference structure $h_1 - h_2$ complicates the analysis of stationarity, we introduce the following **auxiliary function** in our analysis:

$$H(x, w) := f(x) + h_1(x) + h_2^*(w) - \langle x, w\rangle, \tag{3}$$

where $h_2^*(w)$ is the conjugate function of $h_2(x)$ defined as $h_2^*(w) := \sup_x\{\langle x, w\rangle - h_2(x)\}$. For each t, take $w_t \in \partial h_2(x_t)$, then by the Fenchel equality,

$$h_2(x_t) + h_2^*(w_t) = \langle x_t, w_t\rangle. \tag{4}$$

Next we give the definition of critical point of problem (1) corresponding to the auxiliary function H.

Definition 2.2 (**Critical Point**). *We say x^* is a critical point of problem (1) if $x^* \in \mathbb{R}^d$ and the following inclusion holds:*

$$0 \in \nabla f(x^*) + \partial h_1(x^*) - w^*,$$

where $w^* \in \partial h_2(x^*)$.

It is obvious from the generalized Fermat's Rule that a local minimizer of problem (1) must be a critical point of problem (1). The following lemma further connects the accumulation point of the auxiliary function H with critical points of the original problem, and its proof is placed in Appendix B.1.

Lemma 2.3 (**Critical Point**). *Let (x^*, w^*) be an accumulation point of $\{(x_{t+1}, w_t)\}_{t\geq 0}$ and $H(x_{t+1}, w_t) \to H(x^*, w^*)$. If*

$$\operatorname{dist}\big(0, \partial H(x_{t+1}, w_t)\big) \to 0,$$

then x^* is a critical point of problem (1).

Next, we establish the global convergence in the following Theorem by exploiting the auxiliary function H, where its proof is given in Appendix B.2.

Theorem 2.4 (**Subsequential Convergence**). *If Assumption 2.1 is satisfied with*

$$\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty,\ 0 < \alpha < \frac{1}{2L} \text{ and } 0 < \gamma_t \leq \frac{1}{2Ln},$$

then we have

(i) $\|x_{t+1} - x_t\| \to 0$.

(ii) $H(x_{t+1}, w_t)$ *converges to some* $\bar{H} \in \mathbb{R}$ *and* $P(x_{t+1})$ *converges to the same* $\bar{H}$.

(iii) Every accumulation point of $\{x_t\}_{t\geq 0}$ *is a critical point of problem (1).*

Theorem 2.4 guarantees subsequential convergence of PNRR in the nonsmooth nonconvex setting, showing that every accumulation point is a critical point. Based on the approximate descent and subgradient estimates placed in Appendix B, we now establish the iteration complexity of PNRR in the following Theorem, and its proof is detailed in Appendix B.3.

Theorem 2.5 (**Complexity**). *Let Assumption 2.1 is satisfied with*

$$\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty,\ 0 < \alpha < \frac{1}{2L} \text{ and } 0 < \gamma_t \leq \frac{1}{2Ln}.$$

Then, the following statements are valid:

(i) The complexity bound for PNRR is

$$\min_{0\leq t<T} \operatorname{dist}^2(0, \partial H(x_{t+1}, w_t)) \leq \frac{2C_2^2\left[U_0 + 8\alpha L^3 C_1 \sum_{t=0}^{T-1} n^2\gamma_t^2\right]}{T\left(\frac{1}{4\alpha} - \frac{L}{2}\right)} + \frac{2C_3^2}{T}\sum_{t=0}^{T-1} n^2\gamma_t^2.$$

(ii) The IFO for PNRR is $\mathcal{O}(n\epsilon^{-2})$.

(iii) The PO for PNRR is $\mathcal{O}(\epsilon^{-2})$.

Remark 2.6. *We can observe from Theorem 2.5 that PNRR offers clear advantages over representative proximal RR methods. Norm-PRR performs a proximal update at every inner iteration, whereas PNRR requires only one such evaluation per epoch. Compared with proxRR, PNRR retains this low proximal-oracle cost while allowing the proximal parameter to be chosen independently of the reshuffling stepsize. In addition, PNRR incurs fewer proximal-oracle calls than proxSGD. Overall, PNRR achieves both computation efficiency and parameter flexibility.*

3 Convergence under the Kurdyka–Łojasiewicz (KL) Condition

In this section, we study convergence under the Kurdyka-Łojasiewicz (KL) condition. To this end, we further assume H is a KL function with the corresponding φ (see Definition A.2) taking the form $\varphi(z) = \frac{c}{1-\theta} z^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$. Then we can estimate the convergence rate of PNRR for different KL exponents. Moreover, we derive the corresponding rates under polynomial and exponential decaying stepsizes.

3.1 Full Sequence Convergence

In this subsection, we show that the whole sequence $\{x_t\}_t$ converges to a critical point of problem (1). First, we present the definition of the KL inequality in the RR setting in Appendix A, and then assume that problem (1) satisfies the following assumptions.

Assumption 3.1. *We consider the following assumptions:*

(i) The iterate sequence $\{x_t\}$ is bounded.

(ii) H satisfies the KL property (Definition A.2) at every point of the associated set A_H (defined in (30)) with $\varphi(z) = \frac{c}{1-\theta} z^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$, that is,

$$|H(x, w) - \bar{H}|^\theta \leq c \cdot \operatorname{dist}(0, \partial H(x, w)). \tag{5}$$

Conditions (i)–(ii) are standard in KL convergence analysis. Condition (i) is a typical prerequisite for applying the KL framework and guarantees the existence of accumulation points; see, e.g., (Attouch et al., 2013; Bolte et al., 2014; Qiu et al., 2025). Condition (ii) enables an explicit characterization of convergence rates through the KL exponent and is satisfied by a broad class of structured functions, including many semialgebraic and subanalytic functions; see, e.g., (Li et al., 2023).

We present the full sequence convergence in the following Theorem under the KL condition, where its proof is provided in Appendix C.1.

Theorem 3.2 (Full Sequence Convergence). *Suppose conditions in Theorem 2.4 and assumption 3.1 hold, let γ_t satisfy*

$$\sum_{t=0}^{\infty} n\gamma_t < +\infty,\ \sum_{t=0}^{\infty} \left(\sum_{j=t}^{\infty} n^2 \gamma_j^2 \right)^{\theta} < +\infty,\ \theta \in (0, 1).$$

Then

$$\sum_{t=0}^{\infty} \|x_{t+1} - x_t\| < +\infty.$$

Consequently, the whole sequence $\{x_t\}$ converges to a critical point x^* of problem (1).

Remark 3.3. *Theorem 3.2 provides a strong convergence guarantee within the RR framework by establishing full sequence convergence to a critical point under the KL framework. Such a result is particularly valuable, since it provides a deterministic convergence guarantee in the sense that it holds for the iterate sequence itself rather than only in expectation.*

3.2 CONVERGENCE RATES

In this section, we establish convergence rates for the sequences $\{x_t\}_t$, $\{\mathrm{dist}(0, \partial H(x_t, w_{t-1}))\}_t$, $\{P(x_t)\}_t$ and $\{H(x_t, w_{t-1})\}_t$ under two specific stepsize strategies: polynomial decaying stepsize (Robbins & Monro, 1951; Bottou et al., 2018; Nguyen et al., 2021) and exponential decaying stepsize (Davis et al., 2018; 2024). We first present in the following theorem the convergence rate under polynomial stepsizes, whose proof is detailed in Appendix C.2.2.

Theorem 3.4 (**Convergence Rate under Polynomial Stepsizes).** *Suppose conditions in Theorem 2.4 and Assumption 3.1 satisfied. We consider the following polynomial stepsizes,*

$$\gamma_t = \frac{\gamma_0}{(t+\beta)^p},\ \gamma_0 > 0,\ \beta \geq 1,\ p > 1. \tag{6}$$

Then $x_t \to x^*$ for a critical point x^*, and the following rates hold:

(i) If $\theta \in [0, \frac{p}{2p-1}]$, then

$$\begin{aligned} |H(x_t, w_{t-1}) - \bar{H}| &= \mathcal{O}(t^{-(2p-1)}), \\ |P(x_t) - \bar{P}| &= \mathcal{O}(t^{-(2p-1)}), \\ \mathrm{dist}(0, \partial H(x_t, w_{t-1})) &= \mathcal{O}(t^{-(p-\frac{1}{2})}), \\ \|x_t - x^*\| &= \mathcal{O}(t^{-(p-1)}). \end{aligned}$$

(ii) If $\theta \in (\frac{p}{2p-1}, 1)$, then

$$\begin{aligned} |H(x_t, w_{t-1}) - \bar{H}| &= \mathcal{O}\left(t^{-\frac{1}{2\theta-1}}\right), \\ |P(x_t) - \bar{P}| &= \mathcal{O}(t^{-\frac{1}{2\theta-1}}), \\ \mathrm{dist}(0, \partial H(x_t, w_{t-1})) &= \mathcal{O}\left(t^{-\frac{1}{2(2\theta-1)}}\right), \\ \|x_t - x^*\| &= \mathcal{O}\left(t^{-\frac{1-\theta}{2\theta-1}}\right). \end{aligned}$$

Remark 3.5. *Theorem 3.4 characterizes asymptotic convergence rates for the function values and the stationarity measure under polynomially decaying stepsizes, together with an explicit last-iterate convergence rate for PNRR. These rates reveal the joint effect of the polynomial stepsize parameter and the KL exponent on the convergence behavior. Such explicit convergence rates provide a relatively strong asymptotic convergence guarantee for nonsmooth nonconvex RR methods.*

Then, We show the convergence rate under exponential stepsizes as follows, whose proof is placed in Appendix C.2.3.

Theorem 3.6 (**Convergence Rate under Exponential Stepsizes).** *Supposing conditions in Theorem 2.4 and Assumption 3.1 is satisfied. We consider the following exponential stepsizes with $\theta \in [0, \frac{1}{2})$,*

$$\gamma_t = \gamma_0 \rho^t,\ \gamma_0 \leq \frac{1}{2Ln},\ \rho \in (0, 1). \tag{7}$$

Then $x_t \to x^*$ for a critical point x^*, and the following rates hold:

$$\begin{aligned} |H(x_t, w_{t-1}) - \bar{H}| &= \mathcal{O}(\rho^{2t}), \\ |P(x_t) - \bar{P}| &= \mathcal{O}(\rho^{2t}), \\ \mathrm{dist}(0, \partial H(x_t, w_{t-1})) &= \mathcal{O}(\rho^t), \\ \|x_t - x^*\| &= \mathcal{O}(\rho^t). \end{aligned}$$

Remark 3.7. *Theorem 3.6 provides asymptotic convergence rates for the function values, the stationarity measure, and the iterates of PNRR under exponentially decaying stepsizes. In particular, all these quantities converge at geometric rates. This demonstrates the advantage of exponential stepsizes in obtaining faster asymptotic rates for nonsmooth nonconvex RR methods.*

4 NUMERICAL EXPERIMENTS

In this section, we conduct experiments on logistic regression and nonconvex binary classification to compare PNRR with several representative algorithms, including the standard proximal stochastic gradient (proxSGD, Duchi & Singer (2009)), proximal random reshuffling (proxRR, Mishchenko et al. (2022)), normal map-based proximal random reshuffling (norm-PRR, Qiu et al. (2025)), SGD-based method, and full-gradient method. In our experiments[2], we evaluate the performance of the tested algorithms by the following criteria: (i) objective value $P(x)$, and (ii) relative error, defined as $(P(x_t) - P_{\min})/\max(1, P_{\min})$.

4.1 LOGISTIC REGRESSION WITH CONVEX REGULARIZATION

We consider the logistic regression with elastic net regularization

$$\min_{x\in\mathbb{R}^d} f(x) + h_1(x) - h_2(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|_2^2,$$

where $f_i(x) := -(b_i \log(h(a_i^\top x)) + (1-b_i)\log(1-h(a_i^\top x)))$, $(a_i, b_i) \in \mathbb{R}^d \times \{0,1\}, i = 1, \ldots, n$ are the data samples, $h : t \to 1/(1+e^{-t})$ is the sigmoid function, and $\lambda_1, \lambda_2 \geq 0$ are the elastic net parameters. In this experiment, $h_2 = 0$ is set for a fair comparison with existing proximal RR and SGD methods reported in Table 1. The implementation details are shown in Appendix D.2.

Experimental results are shown in Figure 1. Across all three datasets, PNRR clearly outperforms proxSGD and proxRR, exhibiting faster and smoother decreases in both the objective value and relative error, together with smaller final errors. Compared with norm-PRR, which performs n proximal computation per epoch, PNRR achieves comparably strong performance in terms of convergence speed and final accuracy, while requiring only one proximal evaluation per epoch. These results suggest that PNRR is competitive with the existing stochastic and RR methods.

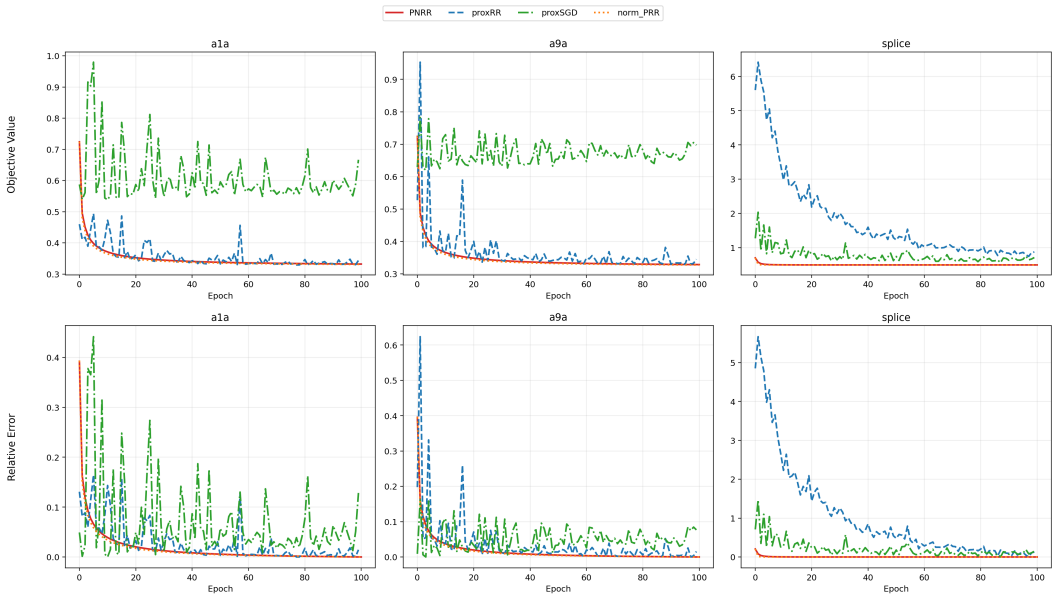

Figure 1: Logistic regression with convex regularization.

4.2 BINARY CLASSIFICATION WITH NONCONVEX REGULARIZATION

In this subsection, we conduct experiments with stochastic gradient descent algorithm in the proximal difference of convex setting (SGD-DC) and full gradient algorithm (summarized in (2)). For

[2]Code and datasets are available at `https://anonymous.4open.science/status/PNRR-2750`

completeness, we give the pseudocode for SGD-DC in Algorithm 3 placed in Appendix D.1. Next, we consider binary classification problem with the nonconvex $\ell_1 - \ell_2$ regularization

$$\min_{x\in\mathbb{R}^d} f(x) + h_1(x) - h_2(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \lambda_1\|x\|_1 - \lambda_2\|x\|_2,$$

where $f_i(x) = 1 - \tanh(b_i \cdot a_i^\top x)$, where $\tanh$ denotes the hyperbolic tangent function. The implementation details are shown in Appendix D.2.

Experimental results are shown in Figure 2. We observe that PNRR generally shows faster convergence than SGD-DC and the full-gradient algorithm across the three datasets. In the initial epochs, a rapid decrease in the objective value can be observed for PNRR. Moreover, PNRR typically achieves smaller objective values and relative errors than the compared methods. In the later epochs, PNRR maintains stable convergence behavior and exhibits less fluctuation than the compared methods.

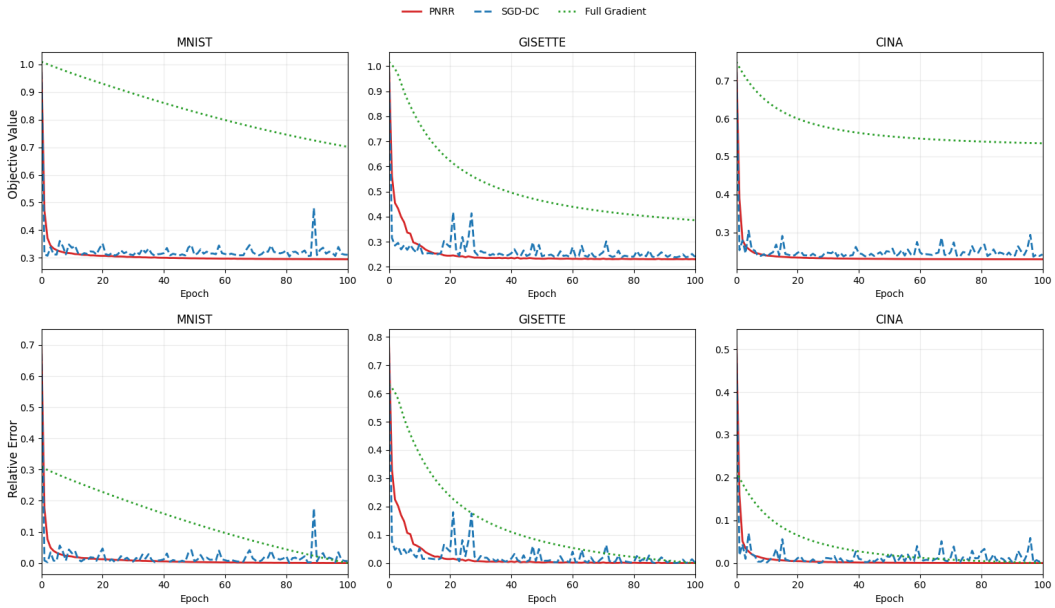

Figure 2: Binary classification with nonconvex regularization.

5 Conclusion

In this paper, we proposed a proximal algorithm for nonsmooth nonconvex random reshuffling (PNRR) for nonsmooth nonconvex finite-sum optimization problem. PNRR requires only one evaluation of the proximity operator in each epoch, while the proximal parameter can be chosen independently of the reshuffling stepsize and kept constant. We established convergence and complexity results for the proposed PNRR. In particular, PNRR achieves an IFO complexity of $\mathcal{O}(n\epsilon^{-2})$ and a PO complexity of $\mathcal{O}(\epsilon^{-2})$. Moreover, we further investigated the convergence of PNRR under the KL condition. By introducing the auxiliary function H and controlling the accumulated reshuffling errors, we proved the finite-length property of the iterates and the convergence of the whole sequence to a critical point. We also characterized the asymptotic convergence behavior of the function values, stationarity measure, and iterates. The subsequent rate analysis further characterizes the asymptotic behavior of PNRR under polynomial and exponential decaying stepsizes. More importantly, it goes beyond finite-time best-iterate complexity by providing convergence guarantees for the actual iterate sequence, together with the corresponding function values and stationarity measure. Finally, numerical experiments illustrate that PNRR effectively performs on nonsmooth nonconvex problems and learning tasks.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, including improving grammar, clarity, and presentation, and to assist with literature retrieval and discovery of potentially relevant related work. We did not use generative AI tools to fabricate datasets, construct unsupported or invalid mathematical proofs, or invent factual claims. All AI-assisted content was carefully reviewed and verified by the authors, and all retrieved references and bibliographic information were checked against the original sources. We take full responsibility for the final content of this work, including all text, mathematical claims, results, and references.

REPRODUCIBILITY STATEMENT

All assumptions and algorithmic details are stated explicitly in the main text. Complete proofs of the theoretical results are provided in the appendix. The experimental settings, parameter choices, and implementation details are reported in the experimental section and appendix to facilitate reproduction of the reported results. Anonymous source code is provided in the paper.

REFERENCES

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Advances in Neural Information Processing Systems 33*, pp. 17526–17535, 2020.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, pp. 421–436. Springer, 2012.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. SIAM, 2021.

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

Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179 (3):962–982, 2018.

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, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *Mathematical Programming*, 207(1): 145–190, 2024.

Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtarik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. In *International Conference on Learning Representations*, volume 2025, pp. 80916–80983, 2025.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(99):2899–2934, 2009.

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

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 680–690, 2020.

Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1):141–176, 2018.

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2624–2633, 2019.

Kun Huang, Xiao Li, Andre Milzarek, Shi Pu, and Junwen Qiu. Distributed random reshuffling over networks. *IEEE Transactions on Signal Processing*, 71:1143–1158, 2023.

Anastasia Koloskova, Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 25058–25086, 2024.

Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. In *Advances in Neural Information Processing Systems 35*, pp. 33107–33119, 2022.

Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the kurdyka–łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

Yuqing Liang and Dongpo Xu. Last-iterate convergence of shuffling momentum gradient method under the kurdyka-lojasiewicz inequality. *Journal of Machine Learning Research*, 26(123):1–51, 2025.

Jackie Lok, Rishi Sonthalia, and Elizaveta Rebrova. Error dynamics of mini-batch gradient descent with random reshuffling for least squares regression. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272, pp. 736–770, 2025.

Szymon Majewski, Blazej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: The differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. In *Proceedings of the 4th International Workshop on Distributed Machine Learning*, pp. 85–104. Association for Computing Machinery, 2023.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems 33*, pp. 17309–17320, 2020.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15718–15749, 2022.

Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of *Fundamental Principles of Mathematical Sciences*. Springer, 2006.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4703–4711, 2019.

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

Junwen Qiu, Xiao Li, and Andre Milzarek. A new random reshuffling method for nonsmooth nonconvex finite-sum optimization. *Journal of Machine Learning Research*, 26(191):1–46, 2025.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7964–7973. PMLR, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, 1970.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Springer, 1998.

Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Burlachenko, and Peter Richtárik. Don't compress gradients in random reshuffling: Compress gradient differences. In *Advances in Neural Information Processing Systems 37*, pp. 84523–84605, 2024.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125, pp. 3250–3284. PMLR, 2020.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pp. 46–54, 2016.

Trang H. Tran, Lam M. Nguyen, and Quoc Tran-Dinh. SMG: A shuffling gradient-based method with momentum. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10379–10389. PMLR, 2021.

Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H. Sayed. Stochastic learning under random reshuffling with constant step-sizes. *IEEE Transactions on Signal Processing*, 67(2):474–489, 2019.

Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022.

APPENDIX

A Notations and Auxiliary Results

This appendix supplies the notation and background used in the main text.

A.1 Variational Notations and Auxiliary Convergence Result

We begin with our preferred notations. An extended-real-valued function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be proper if its domain $\operatorname{dom} \varphi := \{x \in \mathbb{R}^n : \varphi(x) < +\infty\}$ is nonempty. A proper function φ is said to be closed if φ is lower semicontinuous on $\mathbb{R}^n$. A function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be Lipschitz continuous around x with a modulus $L > 0$, if there exists $\delta > 0$ such that

$$|\varphi(u) - \varphi(v)| \leq L\|u - v\|$$

holds for all $u, v \in \mathcal{B}(x, \delta)$.

We next review some preliminaries on subdifferentials of nonconvex functions (Cui & Pang, 2021; Mordukhovich, 2006; Rockafellar & Wets, 1998). For a proper function φ, the Fréchet subdifferential at $x \in \operatorname{dom}\varphi$ is defined by

$$\widehat{\partial}\varphi(x) := \left\{ y \in \mathbb{R}^n : \liminf_{\substack{z \to x \\ z \neq x}} \frac{\varphi(z) - \varphi(x) - \langle y, z - x\rangle}{\|z - x\|_2} \geq 0 \right\},$$

and the limiting subdifferential at $x \in \operatorname{dom}\varphi$ is defined by

$$\partial\varphi(x) := \left\{ y \in \mathbb{R}^n : \exists x^k \to x, \varphi(x^k) \to \varphi(x), y^k \to y \text{ with } y^k \in \widehat{\partial}\varphi(x^k) \text{ for each } k \right\}.$$

We set $\widehat{\partial}\varphi(x) = \partial\varphi(x) = \varnothing$ by convention when $x \notin \operatorname{dom}\varphi$, and define $\operatorname{dom}\partial\varphi := \{x : \partial\varphi(x) \neq \varnothing\}$. For a convex φ, the Fréchet and limiting subdifferential both consist with the classical subdifferential of a convex function at each $x \in \operatorname{dom}\varphi$ (Rockafellar & Wets (1998), Proposition 8.12), that is,

$$\widehat{\partial}\varphi(x) = \partial\varphi(x) = \{y \in \mathbb{R}^n : \varphi(z) - \varphi(x) - \langle y, z - x\rangle \geq 0, \text{ for all } z \in \mathbb{R}^n\}.$$

It is clear that $\partial(\alpha\varphi)(x) = \alpha\partial\varphi(x)$ for all $\alpha > 0$ and $x \in \operatorname{dom}\varphi$. For proper lower semicontinuous functions $\varphi_1, \varphi_2 : \mathbb{R}^n \to (-\infty, +\infty]$, there holds $\partial(\varphi_1 + \varphi_2)(x) \subseteq \partial\varphi_1(x) + \partial\varphi_2(x)$ if φ_1 is locally Lipschitz continuous around x. If φ_1 and φ_2 are further convex, we have $\partial(\varphi_1 - \varphi_2)(x) \subseteq \partial\varphi_1(x) - \partial\varphi_2(x)$ for $x \in \operatorname{dom}\partial\varphi_2$. For a function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$, the generalized *Fermat's Rule* states that $0 \in \partial\varphi(x^*)$ holds if $x^* \in \operatorname{dom}(\varphi)$ is a local minimizer of φ (Rockafellar & Wets (1998), Theorem 10.1).

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be proper, lsc., and convex. The convex conjugate function $\varphi^* : \mathbb{R}^n \to (-\infty, +\infty]$ is defined at $y \in \mathbb{R}^n$ as $\varphi^*(y) := \sup\{\langle x, y\rangle - \varphi(x) : x \in \mathbb{R}^n\}$. Then conjugate φ^* is a proper closed convex function (Rockafellar (1970), Theorem 12.2). Furthermore, for any $x, y \in \mathbb{R}^n$, the following three statements are equivalent (Rockafellar & Wets (1998), Proposition 11.3):

$$\langle x, y\rangle = \varphi(x) + \varphi^*(y) \Leftrightarrow y \in \partial\varphi(x) \Leftrightarrow x \in \partial\varphi^*(y).$$

Therefore, it is clear that the *Fenchel-Young Inequality*, i.e., $\varphi(x) + \varphi^*(y) \geq \langle x, y\rangle$, holds for all $x, y \in \mathbb{R}^n$, and it becomes an equality when $y \in \partial\varphi(x)$. Besides, from the *Fenchel-Young Inequality*, we know that $y \in \operatorname{dom}\varphi^*$ if $y \in \partial\varphi(x)$ for some $x \in \mathbb{R}^n$.

Having introduced the basic variational tools above, we next recall a standard supermartingale convergence result that will be useful for analyzing the perturbed descent sequences arising in our convergence analysis.

Lemma A.1 (**Supermartingale Convergence Theorem**, Proposition A.31, Bertsekas (2016))**.** *Let* $\{y_t\}_{t\geq 0}, \{z_t\}_{t\geq 0}, \{w_t\}_{t\geq 0}$, *and* $\{v_t\}_{t\geq 0}$ *be sequences such that*

$$y_{t+1} \leq (1 + v_t)y_t - z_t + w_t, \ t = 0, 1, \ldots,$$

$\{z_t\}_{t\geq 0}, \{w_t\}_{t\geq 0}$, *and* $\{v_t\}_{t\geq 0}$ *are nonnegative, and*

$$\sum_{t=0}^{\infty} w_t < \infty, \ \sum_{t=0}^{\infty} v_t < \infty.$$

Then, either $y_t \to -\infty$ *or* $\{y_t\}_{t\geq 0}$ *converges to a finite value and* $\sum_{t=0}^{\infty} z_t < \infty$.

A.2 KL Properties

We recall the KL property in this subsection. The KL property and the associated notion of KL exponent have been extensively used in the convergence analysis of first-order methods.

For the RR setting considered in this paper, its KL property is measured by the function-value gap through $|h(x) - h(\bar{x})|$ rather than $h(x) - h(\bar{x})$, since the function values may approach the limiting value from either side (Li et al., 2023; Qiu et al., 2025; Liang & Xu, 2025). The corresponding definition is given below.

Definition A.2 (Kurdyka-Łojasiewicz (KL) property). *A proper function $h : \mathbb{R}^d \to (-\infty, \infty]$ is said to satisfy the KL property at a point $\bar{x} \in dom(\partial h)$ if there exist $\eta \in (0, \infty]$, a neighborhood U of $\bar{x}$, and a concave continuous function $\varphi : [0, \eta) \to [0, +\infty)$ such that*

(i) $\varphi(0) = 0$;

(ii) $\varphi \in C^1((0, \eta))$ *with* $\varphi'(x) > 0$;

(iii) for any $x \in U \cap \{x : 0 < |h(x) - h(\bar{x})| < \eta\}$, *there holds the KL inequality,*

$$\varphi'(|h(x) - h(\bar{x})|) \operatorname{dist}(0, \partial h(x)) \geq 1.$$

Furthermore, we recall the following proposition that establishes a uniformized KL property on the set of accumulation points.

Proposition A.3 (Uniformized KL Property, Lemma 3.5, Li et al. (2023)**).** *Suppose that h is a proper closed function, consider the following conditions satisfied:*

(i) A_h *be a compact set;*

(ii) h *is a constant on* A_h *and satisfies the KL property at each point of* A_h.

Then there exist $\delta, \eta > 0$ and φ such that

$$\varphi'(|h(x) - h(\bar{x})|) \operatorname{dist}(0, \partial h(x)) \geq 1. \tag{8}$$

for all $\bar{x} \in A_h$ and $x \in \{\operatorname{dist}(x, A_h) < \delta\} \cap \{x : 0 < |h(x) - h(\bar{x})| < \eta\}$.

B Proof of Subsequential Convergence

B.1 Proof of Lemma 2.3

Proof. Since (x^*, w^*) is an accumulation point of $\{(x_{t+1}, w_t)\}$, there exists a subsequence $\{t_j\}$ such that $(x_{t_j+1}, w_{t_j}) \to (x^*, w^*)$. Together with $\operatorname{dist}\big(0, \partial H(x_{t_j+1}, w_{t_j})\big) \to 0$ and $H(x_{t_j+1}, w_{t_j}) \to H(x^*, w^*)$, the closedness property of the limiting subdifferential yields $0 \in \partial H(x^*, w^*)$. Moreover, the definition of H gives

$$\partial H(x^*, w^*) = \big(\nabla f(x^*) + \partial h_1(x^*) - w^*\big) \times \big(\partial h_2^*(w^*) - x^*\big).$$

Thus, we can get

$$0 \in \nabla f(x^*) + \partial h_1(x^*) - w^* \text{ and } x^* \in \partial h_2^*(w^*). \tag{9}$$

Since h_2 is proper closed convex, the conjugate subdifferential relation gives $x^* \in \partial h_2^*(w^*)$, that is $w^* \in \partial h_2(x^*)$. Then, (9) indicates

$$0 \in \nabla f(x^*) + \partial h_1(x^*) - w^*,$$

which shows that x^* is a critical point of problem (1). □

B.2 Proof of Theorem 2.4

We first give in the following lemmas to control the gradient variance and the stochastic error.

Lemma B.1 (variance bound). *If Assumption 2.1 is satisfied, then the gradient variance*

$$\sigma_t^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_t) - \nabla f(x_t)\|^2$$

can be bounded by the objective function, that is,

$$\sigma_t^2 \leq 2L(P(x_t) - P_{lb}).$$

Proof. From the definition of variance, we have

$$\begin{aligned}\sigma_t^2 =&\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_t)\|^2 - \|\nabla f(x_t)\|^2\\ \le&\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_t)\|^2\\ \le&\frac{1}{n}\sum_{i=1}^{n}2L(f_i(x_t)-f_{lb})\\ \le&2L(f(x_t)-f_{lb}) \le 2L(P(x_t)-P_{lb}),\end{aligned}$$

where the second inequality is applying the descent lemma, and the last inequality is due to $P(x_t) - P_{lb} = f(x_t) - f_{lb} + h_1(x_t) - h_2(x_t) - h_{lb} \ge f(x_t) - f_{lb}$ for all $x \in \mathbb{R}^d$.

□

Lemma B.2 (error bound). *If Assumption 2.1 is satisfied with $\gamma_t \le \frac{1}{2Ln}$, then the stochastic error*

$$\|g_t - \nabla f(x_t)\|^2 \le 8n^2\gamma_t^2L^3\big(P(x_t)-P_{lb}\big),$$

where $g_t = \frac{1}{\gamma_t n}(x_t - x_t^n) = \frac{1}{n}\sum_{i=0}^{n-1}\nabla f_{\pi_i}(x_t^i)$.

Proof. From the definition of g_t, we have

$$\begin{aligned}\|g_t - \nabla f(x_t)\|^2 =&\|\frac{1}{n}\sum_{i=0}^{n-1}\big(\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)\big)\|^2\\ \le&\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)\|^2\\ \le&\frac{L^2}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2,\end{aligned} \tag{10}$$

where the first inequality is applying the Jensen's inequality and the last inequality since each f_i is L_i-smooth.

Let $v_i := x_t^i - x_t = -\gamma_t\sum_{j=0}^{i-1}\nabla f_{\pi_j}(x_t^j)$, $s_i := \sum_{j=0}^{i-1}\nabla f_{\pi_j}(x_t)$, then

$$\begin{aligned}v_i =&-\gamma_t\big(s_i + \sum_{j=0}^{i-1}\nabla f_{\pi_j}(x_t^j) - s_i\big)\\ =&-\gamma_t s_i - \gamma_t\sum_{j=0}^{i-1}\big(\nabla f_{\pi_j}(x_t^j) - \nabla f_{\pi_j}(x_t)\big).\end{aligned}$$

Therefore,

$$\begin{aligned}\|v_i\|^2 \le&2\gamma_t^2\|s_i\|^2 + 2\gamma_t^2\left\|\sum_{j=0}^{i-1}\big(\nabla f_{\pi_j}(x_t^j) - \nabla f_{\pi_j}(x_t)\big)\right\|^2\\ \le&2\gamma_t^2\|s_i\|^2 + 2\gamma_t^2 i\sum_{j=0}^{i-1}\|\nabla f_{\pi_j}(x_t^j) - \nabla f_{\pi_j}(x_t)\|^2\\ \le&2\gamma_t^2\|s_i\|^2 + 2\gamma_t^2L^2 i\sum_{j=0}^{i-1}\|v_j\|^2,\end{aligned} \tag{11}$$

where the second inequality is using the Cauchy–Schwarz inequality and the last inequality since each f_i is L-smooth.

Next we solve the $\|s_i\|^2$ term, we can derive from the definition that

$$\|s_i\|^2 \leq i\sum_{j=0}^{i-1}\|\nabla f_{\pi_j}(x_t)\|^2 \leq n\sum_{i=1}^{n}\|\nabla f_i(x_t)\|^2 \leq 2Ln^2(P(x_t)-P_{lb}).$$

Let $A_t := \max_{1\leq i\leq n}\|v_i\|^2$, then (11) leads to

$$A_t \leq 4Ln^2\gamma_t^2(P(x_t)-P_{lb}) + 2\gamma_t^2L^2n^2A_t. \tag{12}$$

Combining the stepsize condition $\gamma_t \leq \frac{1}{2Ln}$ and (12), then (12) indicates

$$A_t \leq 8Ln^2\gamma_t^2(P(x_t)-P_{lb}).$$

Therefore, (10) leads to $\|g_t - \nabla f(x_t)\|^2 \leq L^2A_t \leq 8n^2\gamma_t^2L^3(P(x_t)-P_{lb})$. □

Then, we establish an approximate descent property for the objective function and auxiliary function in the following lemma by exploiting the auxiliary function H.

Lemma B.3 (Approximate Descent). *For all $w_t \in \partial h_2(x_t)$, let parameters satisfy $\alpha < \frac{1}{2L}$ and $\gamma_t \leq \frac{1}{2Ln}$, then*

$$\begin{aligned}
P(x_{t+1}) \leq& H(x_{t+1}, w_t)\\
\leq& P(x_t) - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1}-x_t\|^2 + 8\alpha n^2\gamma_t^2L^3\big(P(x_t)-P_{lb}\big)\\
\leq& H(x_t, w_{t-1}) - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1}-x_t\|^2 + 8\alpha n^2\gamma_t^2L^3\big(H(x_t,w_{t-1})-P_{lb}\big).
\end{aligned}$$

Proof. By the definition of $x_{t+1} \in \operatorname{prox}_{\alpha h_1}(x_t - \alpha g_t + \alpha w_t)$, we have

$$h_1(x_{t+1}) + \frac{1}{2\alpha}\|x_{t+1} - x_t + \alpha(g_t - w_t)\|^2 \leq h_1(x_t) + \frac{\alpha}{2}\|g_t - w_t\|^2,$$

Rearranging the terms, this further yields

$$h_1(x_{t+1}) \leq h_1(x_t) - \frac{1}{2\alpha}\|x_{t+1}-x_t\|^2 + \langle w_t - g_t, x_{t+1} - x_t\rangle. \tag{13}$$

Besides, f is L-smooth leading to

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1}-x_t\rangle + \frac{L}{2}\|x_{t+1}-x_t\|^2. \tag{14}$$

Combining the Fenchel equality (4), (13) and (14), then

$$H(x_{t+1}, w_t) \leq P(x_t) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|x_{t+1}-x_t\|^2 + \langle \nabla f(x_t) - g_t, x_{t+1} - x_t\rangle.$$

By Young's inequality,

$$\langle \nabla f(x_t) - g_t, x_{t+1} - x_t\rangle \leq \alpha\|\nabla f(x_t) - g_t\|^2 + \frac{1}{4\alpha}\|x_{t+1}-x_t\|^2.$$

Thus,

$$H(x_{t+1}, w_t) \leq P(x_t) - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1}-x_t\|^2 + \alpha\|\nabla f(x_t) - g_t\|^2. \tag{15}$$

We can find that $\alpha < \frac{1}{2L}$ leads to $\frac{1}{4\alpha} - \frac{L}{2} > 0$, thus $\|x_{t+1}-x_t\|^2$ is the descent term. Moreover, combining the definition of $H(x,w)$, that is, for any $t = 0, 1, \cdots, T-1$,

$$H(x_t, w_{t-1}) = f(x_t) + h_1(x_t) + h_2^*(w_{t-1}) - \langle x_t, w_{t-1}\rangle \geq f(x_t) + h_1(x_t) - h_2(x_t) = P(x_t),$$

and using the Lemma B.2 to upper bound $\|\nabla f(x_t) - g_t\|^2$ in (15), then we can finish the proof. □

Based on the approximate descent property, we next establish the boundedness of the objective gap and derive a useful estimate for the stationarity residual of the auxiliary function H in the following proposition.

Proposition B.4 (Boundedness). *If Assumption 2.1 is satisfied with*

$$\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty,\ 0 < \alpha < \frac{1}{2L} \text{ and } 0 < \gamma_t \leq \frac{1}{2Ln},$$

then we can upper bound the difference of objective function value and its lower bound, while control the subgradient of H, that is

(i) There exists a constant $C_1 > 0$ such that

$$P(x_t) - P_{lb} \leq H(x_t, w_{t-1}) - P_{lb} \leq C_1.$$

(ii) There exist constants $C_2 > 0$ and $C_3 > 0$ such that

$$\operatorname{dist}(0, \partial H(x_{t+1}, w_t)) \leq C_2\|x_{t+1} - x_t\| + C_3 n\gamma_t.$$

Proof. For part (i). First, the definition of $H(x_t, w_{t-1})$ leads to $P(x_t) - P_{lb} \leq H(x_t, w_{t-1}) - P_{lb}$. For notational consistency at $t = 0$, choose any $w_{-1} \in \partial h_2(x_0)$, $x_0 \in \operatorname{dom} h_1$, and set $U_t := H(x_t, w_{t-1}) - P_{lb}$ for $t \geq 0$. From Lemma B.3, we have

$$\begin{aligned} U_{t+1} \leq & (1 + 8\alpha n^2\gamma_t^2 L^3)U_t - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 \\ \leq & (1 + 8\alpha n^2\gamma_t^2 L^3)U_t \\ \leq & \prod_{j=0}^{t}(1 + 8\alpha n^2\gamma_j^2 L^3)U_0 \leq \prod_{t=0}^{\infty}(1 + 8\alpha n^2\gamma_t^2 L^3)U_0. \end{aligned} \tag{16}$$

Using $\ln(1 + x) \leq x$, we have $\prod_{t=0}^{\infty}(1 + 8\alpha n^2\gamma_t^2 L^3)U_0 \leq (e^{\sum_{t=0}^{\infty} n^2\gamma_t^2})^{8\alpha L^3}U_0$. Moreover, since $\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty$, there exists a constant $C_0 > 0$ such that $e^{\sum_{t=0}^{\infty} n^2\gamma_t^2} \leq C_0$, that is, there exists a constant $C_1 \geq {C_0}^{8\alpha L^3}U_0 > 0$ such that for all $t \geq 0$,

$$U_{t+1} \leq C_0^{8\alpha L^3}U_0 \leq C_1. \tag{17}$$

For part (ii). It is obvious that

$$\partial H(x_{t+1}, w_t) = (\nabla f(x_{t+1}) + \partial h_1(x_{t+1}) - w_t) \times (\partial h_2^*(w_t) - x_{t+1}).$$

In order to show the conclusion, it suffices to select subgradients from $\nabla f(x_{t+1}) + \partial h_1(x_{t+1}) - w_t$ and $\partial h_2^*(w_t) - x_{t+1}$, respectively. First, since $w_t \in \partial h_2(x_t)$ leading to $x_t \in \partial h_2^*(w_t)$, then we have

$$x_t - x_{t+1} \in \partial h_2^*(w_t) - x_{t+1}.$$

Besides, from the first order optimal condition of the proximity operator of x_{t+1}, there has

$$\nabla f(x_{t+1}) - g_t - \frac{1}{\alpha}(x_{t+1} - x_t) \in \nabla f(x_{t+1}) + \partial h_1(x_{t+1}) - w_t$$

Then, we can get $(\nabla f(x_{t+1}) - g_t - \frac{1}{\alpha}(x_{t+1} - x_t), x_t - x_{t+1}) \in \partial H(x_{t+1}, w_t)$. Thus,

$$\begin{aligned} \operatorname{dist}(0, \partial H(x_{t+1}, w_t)) \leq & \|\nabla f(x_{t+1}) - g_t - \frac{1}{\alpha}(x_{t+1} - x_t)\| + \|x_{t+1} - x_t\| \\ \leq & \|\nabla f(x_{t+1}) - \nabla f(x_t)\| + \|\nabla f(x_t) - g_t\| + (1 + \frac{1}{\alpha})\|x_{t+1} - x_t\| \\ \leq & \left(L + \frac{1}{\alpha} + 1\right)\|x_{t+1} - x_t\| + \sqrt{8L^3C_1}n\gamma_t, \end{aligned}$$

where the last inequality is combining the L-smoothness of f, Lemma B.2, and the conclusion of part (i). Let $C_2 = L + \frac{1}{\alpha} + 1$ and $C_3 = \sqrt{8L^3C_1}$, then we finish the proof.

□

Lemma B.5 (Function value consistency). *Suppose $\bar{x}$ is an accumulation point of $\{x_t\}$, and $\{t_j\}$ is a subsequence such that*

$$x_{t_j} \to \bar{x},\ x_{t_j+1} \to \bar{x},\ w_{t_j} \to \bar{w},\ \text{and}\ g_{t_j} \to \nabla f(\bar{x}). \tag{18}$$

If $\{P(x_{t_j})\}$ are bounded from above, then

$$H(x_{t_j+1}, w_{t_j}) \to H(\bar{x}, \bar{w}).$$

Proof. Since h_2 is convex and real-valued, it is continuous on $\mathbb{R}^d$ and the graph of ∂h_2 is closed. Then, together with (18), there implies $\bar{w} \in \partial h_2(\bar{x})$. Since $\{P(x_{t_j})\}$ is bounded from above, the continuity of f and h_2 yields that $\{h_1(x_{t_j})\}$ is bounded from above. By the lsc. of h_1, we have $h_1(\bar{x}) < +\infty$.

Let $z_{t_j} := x_{t_j} - \alpha g_{t_j} + \alpha w_{t_j}$ and $\bar{z} := \bar{x} - \alpha\nabla f(\bar{x}) + \alpha\bar{w}$, then we have $z_{t_j} \to \bar{z}$. Since $x_{t_j+1} \in \operatorname{prox}_{\alpha h_1}(z_{t_j})$,

$$h_1(x_{t_j+1}) + \frac{1}{2\alpha}\|x_{t_j+1} - z_{t_j}\|^2 \le h_1(\bar{x}) + \frac{1}{2\alpha}\|\bar{x} - z_{t_j}\|^2.$$

Taking the limit superior gives $\limsup_{j\to\infty} h_1(x_{t_j+1}) \le h_1(\bar{x})$. Together with the lsc. of h_1 and $x_{t_j+1} \to \bar{x}$, we obtain

$$h_1(x_{t_j+1}) \to h_1(\bar{x}). \tag{19}$$

Furthermore, by the Fenchel equality (4),

$$h_2^*(w_{t_j}) = \langle x_{t_j}, w_{t_j}\rangle - h_2(x_{t_j}) \to \langle \bar{x}, \bar{w}\rangle - h_2(\bar{x}) = h_2^*(\bar{w}). \tag{20}$$

Therefore, combining (19), (20) and the continuity of f, we have $H(x_{t_j+1}, w_{t_j}) \to H(\bar{x}, \bar{w})$. □

Now, we can prove the Theorem 2.4 in the following.

Proof. For part (i). Combining Proposition B.4 (i) in (16), then we have

$$\|x_{t+1} - x_t\|^2 \le \frac{4\alpha}{1 - 2\alpha L}\left(U_t - U_{t+1} + 8\alpha n^2\gamma_t^2 L^3 C_1\right), \tag{21}$$

Summing (21) from 0 to $+\infty$, we can get

$$\sum_{t=0}^{\infty}\|x_{t+1} - x_t\|^2 \le \frac{4\alpha}{1 - 2\alpha L}\left(U_0 + 8\alpha L^3 C_1 \sum_{t=0}^{\infty} n^2\gamma_t^2\right), \tag{22}$$

where the inequality is due to $U_{t+1} = H(x_{t+1}, w_t) - P_{lb} \ge 0$ for all $t \ge 0$. Thus, combining $\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty$, the right side of the inequality is a finite value, that is

$$\sum_{t=0}^{\infty}\|x_{t+1} - x_t\|^2 < +\infty,$$

this also leads to

$$\|x_{t+1} - x_t\| \to 0. \tag{23}$$

For part (ii). Lemma B.3 is equal to

$$H(x_{t+1}, w_t) - P_{lb} \le (1 + 8\alpha n^2\gamma_t^2 L^3)\left[H(x_t, w_{t-1}) - P_{lb}\right] - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2,$$

and

$$P(x_{t+1}) - P_{lb} \le (1 + 8\alpha n^2\gamma_t^2 L^3)\left[P(x_t) - P_{lb}\right] - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2,$$

which satisfies the supermartingale-type recursion present in Lemma A.1. This implies that $H(x_{t+1}, w_t)$ converges to some $\bar{H} \in \mathbb{R}$ and $P(x_{t+1})$ converges to some $\bar{P} \in \mathbb{R}$.

Next, we show $\bar{P} = \bar{H}$. On one hand, from the definition of H, we have $P(x_{t+1}) \le H(x_{t+1}, w_t)$. Letting $t \to \infty$ gives $\bar{P} \le \bar{H}$. On the other hand, Lemma B.3 indicates $H(x_{t+1}, w_t) \le P(x_t) +$

$8\alpha n^2\gamma_t^2 L^3 C_1$. Since $\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty$, we have $n^2\gamma_t^2 \to 0$. Letting $t \to +\infty$ for the both side of inequality gives $\bar{H} \leq \bar{P}$. Thus, we can get $\bar{P} = \bar{H}$.

For part (iii). Since $\sum_{t=0}^{\infty} n^2\gamma_t^2 < +\infty$, we have

$$n\gamma_t \to 0. \tag{24}$$

Combining Proposition B.4 (ii), (23), and (24), we obtain

$$\operatorname{dist}(0, \partial H(x_{t+1}, w_t)) \to 0. \tag{25}$$

Let x^* be an arbitrary accumulation point of $\{x_t\}_{t\geq 0}$. Then, there exists a subsequence $\{t_j\}$ such that $x_{t_j} \to x^*$. By (23),

$$x_{t_j+1} \to x^*. \tag{26}$$

Since h_2 is convex and finite-valued on $\mathbb{R}^d$, $\cup_{x\in S}\partial h_2(x)$ is bounded, where S is a bounded set. Since $x_{t_j} \to x^*$, the sequence $\{w_{t_j}\}$, with $w_{t_j} \in \partial h_2(x_{t_j})$, is bounded. Passing to a further subsequence if necessary, we may assume that $w_{t_j} \to w^*$. Thus, the closedness graph property of h_2 leads to

$$w^* \in \partial h_2(x^*). \tag{27}$$

Moreover, Lemma B.2 and Proposition B.4 (i) give $\|g_t - \nabla f(x_t)\|^2 \leq 8L^3C_1n^2\gamma_t^2$. Hence, by (24), $x_{t_j} \to x^*$ and the continuity of ∇f,

$$g_{t_j} \to \nabla f(x^*). \tag{28}$$

Proposition B.4 (i) also implies that $\{P(x_{t_j})\}$ is bounded from above. Therefore, Lemma B.5 together with (26), (27) and (28) yield

$$H(x_{t_j+1}, w_{t_j}) \to H(x^*, w^*). \tag{29}$$

Combining (25), (29), and Lemma 2.3, we conclude that x^* is a critical point of problem (1). Since x^* is arbitrary, every accumulation point of $\{x_t\}_{t\geq 0}$ is a critical point of problem (1). □

B.3 Proof of Theorem 2.5

Proof. For part (i). From Proposition B.4 (ii), we have

$$\operatorname{dist}^2(0, \partial H(x_{t+1}, w_t)) \leq 2C_2^2\|x_{t+1} - x_t\|^2 + 2C_3^2n^2\gamma_t^2.$$

Thus, it holds that

$$\min_{0\leq t\leq T-1} \operatorname{dist}^2(0, \partial H(x_{t+1}, w_t)) \leq \frac{2C_2^2}{T}\sum_{t=0}^{T-1}\|x_{t+1} - x_t\|^2 + \frac{2C_3^2}{T}\sum_{t=0}^{T-1} n^2\gamma_t^2.$$

Combining (22), we can finish the proof for part (i).

For part (ii) and (iii). Let $S_T := \sum_{t=0}^{T-1} n^2\gamma_t^2$ and $S_\infty := \sum_{t=0}^{\infty} n^2\gamma_t^2$. By assumption, $S_\infty < +\infty$ and $S_T \leq S_\infty$. Hence, part (i) implies

$$\begin{aligned}\min_{0\leq t<T} \operatorname{dist}^2(0, \partial H(x_{t+1}, w_t)) \leq& \frac{2C_2^2\left(U_0 + 8\alpha L^3C_1S_\infty\right)}{T\left(\frac{1}{4\alpha} - \frac{L}{2}\right)} + \frac{2C_3^2S_\infty}{T} \\ =& \mathcal{O}(T^{-1}).\end{aligned}$$

Therefore, to obtain an ϵ-stationary point in the sense that $\operatorname{dist}^2(0, \partial H(x_{t+1}, w_t)) \leq \epsilon^2$, it suffices to take $T = \mathcal{O}(\epsilon^{-2})$ epochs. Each epoch of PNRR evaluates exactly one pass over the n component gradients, and hence costs n IFO calls. Thus the total IFO complexity is $nT = \mathcal{O}(n\epsilon^{-2})$. Since the number of proximal oracle calls is still exactly one per epoch, the PO complexity for reaching an unsquared ϵ-stationary point is $\mathcal{O}(\epsilon^{-2})$.

□

C Proof of Full sequence Convergence

C.1 Proof of Theorem 3.2

Let $\{x_t\}$ be generated by PNRR, $\{w_t\}$ be taken from $\partial h_2(x_t)$, we define the **associated sets** of accumulation points in the following:

$$A_H := \left\{(x,w) : \exists \text{ a subsequence } \{t_j\} \text{ such that } (x_{t_j}, w_{t_j}) \to (x,w) \text{ when } \mathrm{t_j} \to \infty\right\}. \quad (30)$$

The following Lemma indicates the uniform KL property on A_H.

Lemma C.1 (Accumulation points). *Supposing the conditions of Theorem 2.4 and Assumption 3.1 are satisfied, then we have*

(i) A_H is nonempty and compact.

(ii) H satisfies the uniform KL property on A_H. Consequently, there exists a $\tilde{t} \in \mathbb{N}$ such that for all $t > \tilde{t}$ with $0 < |H(x,w) - \bar{H}| < \eta$, the uniformized KL inequality (8) holds at (x_{t+1}, w_t).

Proof. For part (i). Since h_2 is finite convex, ∂h_2 is locally bounded. Thus, Assumption 3.1 (i) implies that $\{(x_t, w_t)\}$ is bounded. Therefore, it has at least one accumulation point and its accumulation set A_H is nonempty and compact.

For part (ii), let $(\bar{x}, \bar{w}) \in A_H$. From the definition of A_H, there exists a subsequence $\{t_j\}$ such that $(x_{t_j}, w_{t_j}) \to (\bar{x}, \bar{w})$. By Theorem 2.4 (i), $x_{t_j+1} \to \bar{x}$. Moreover, Lemma B.2, Proposition B.4 (i) and (24) imply $g_{t_j} \to \nabla f(\bar{x})$, while Proposition B.4 (i) implies that $\{P(x_{t_j})\}$ is bounded from above. Therefore, Lemma B.5 gives

$$H(x_{t_j+1}, w_{t_j}) \to H(\bar{x}, \bar{w}). \quad (31)$$

On the other hand, Theorem 2.4 (ii) gives $H(x_{t_j+1}, w_{t_j}) \to \bar{H}$. Hence, $H(\bar{x}, \bar{w}) = \bar{H}$. Since $(\bar{x}, \bar{w})$ is arbitrary, H is constant on A_H.

Combining Assumption 3.1 (ii) and Proposition A.3, there exist $\delta > 0$, $\eta > 0$, and a desingularizing function φ such that

$$\varphi'(|H(x,w) - \bar{H}|)\,\mathrm{dist}(0, \partial H(x,w)) \geq 1 \quad (32)$$

whenever $\mathrm{dist}((x,w), A_H) < \delta$ and $0 < |H(x,w) - \bar{H}| < \eta$.

Since $\{(x_t, w_t)\}$ is bounded, $\mathrm{dist}((x_t, w_t), A_H) \to 0$. Together with Theorem 2.4 (i),

$$\mathrm{dist}((x_{t+1}, w_t), A_H) \leq \|x_{t+1} - x_t\| + \mathrm{dist}((x_t, w_t), A_H) \to 0. \quad (33)$$

Moreover, Theorem 2.4 (ii) gives $H(x_{t+1}, w_t) \to \bar{H}$. Therefore, there exists a sufficiently large $\tilde{t} \in \mathbb{N}$ such that for all $t > \tilde{t}$ with $0 < |H(x,w) - \bar{H}| < \eta$, the uniformized KL inequality (8) holds at (x_{t+1}, w_t). □

Then, we can give the detail proof of Theorem 3.2 as follows.

Proof. From the approximate descent inequality (16) and the boundedness of U_t (17), we have

$$H(x_{t+1}, w_t) \leq H(x_t, w_{t-1}) - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 + 8\alpha n^2 \gamma_t^2 L^3 C_1. \quad (34)$$

Let $D := 8\alpha n^2 L^3 C_1,\ u_t := D\sum_{j=t}^{\infty} \gamma_j^2$, then $u_t - u_{t+1} = D\gamma_t^2$. Hence, (34) leads to

$$H(x_{t+1}, w_t) + u_{t+1} \leq H(x_t, w_{t-1}) + u_t - \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2. \quad (35)$$

Since $u_t \to 0$ and Theorem 2.4 indicating $H(x_t, w_{t-1}) \to \bar{H}$, the sequence $\{H(x_t, w_{t-1}) + u_t\}$ is sufficiently decreasing and converges to $\bar{H}$. Thus $H(x_t, w_{t-1}) + u_t - \bar{H}$ is positive. Let

$$S_t := \varphi\big(H(x_t, w_{t-1}) + u_t - \bar{H}\big),$$

where $\varphi(z) = \frac{c}{1-\theta} z^{1-\theta},\ c > 0,\ \theta \in (0,1)$, and thus $\varphi'(z) = cz^{-\theta},\ \frac{1}{\varphi'(z)} = \frac{z^\theta}{c}$. For completeness, we let $\frac{1}{\varphi'(0)} = \frac{0^\theta}{c} = 0$.

Next we bound $\|x_{t+1} - x_t\|^2$ by constructing the decreasing of S_t. We can get

$$\begin{aligned}
S_t - S_{t+1} \geq & \varphi'\big(H(x_t, w_{t-1}) - \bar{H} + u_t\big)\big[H(x_t, w_{t-1}) + u_t - H(x_{t+1}, w_t) - u_{t+1}\big] \\
\geq & \varphi'\big(|H(x_t, w_{t-1}) - \bar{H}| + u_t\big)\big[H(x_t, w_{t-1}) + u_t - H(x_{t+1}, w_t) - u_{t+1}\big] \\
\geq & \varphi'\big(|H(x_t, w_{t-1}) - \bar{H}| + u_t\big)\left(\frac{1}{4\alpha} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 \\
\geq & \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\frac{\|x_{t+1} - x_t\|^2}{[\varphi'(|H(x_t, w_{t-1}) - \bar{H}|)]^{-1} + [\varphi'(u_t)]^{-1}} \\
\geq & \left(\frac{1}{4\alpha} - \frac{L}{2}\right)\frac{\|x_{t+1} - x_t\|^2}{\operatorname{dist}(0, \partial H(x_t, w_{t-1})) + [\varphi'(u_t)]^{-1}},
\end{aligned} \tag{36}$$

where the first inequality is using the concavity of φ, the second inequality is due to monotonicity of φ', the third inequality follows the sufficient decrease of $H + u_t$, the fourth inequality since $(\varphi')^{-1}$ is subadditive, and the last inequality applies the KL inequality (8). Rearrange the terms in (36), this further yields

$$\begin{aligned}
\|x_{t+1} - x_t\|^2 \leq & \frac{S_t - S_{t+1}}{\frac{1}{4\alpha} - \frac{L}{2}}\left(\operatorname{dist}(0, \partial H(x_t, w_{t-1})) + [\varphi'(u_t)]^{-1}\right) \\
\leq & \frac{S_t - S_{t+1}}{\frac{1}{4\alpha} - \frac{L}{2}}\left(\left(L + \frac{1}{\alpha} + 1\right)\|x_t - x_{t-1}\| + \|\nabla f(x_{t-1}) - g_{t-1}\| + [\varphi'(u_t)]^{-1}\right),
\end{aligned}$$

where the second inequality is due to Proposition B.4 (ii).

Thus, taking square roots and using Young's inequality, we can get for all $t > \tilde{t}$

$$\begin{aligned}
\|x_{t+1} - x_t\| \leq & \frac{2(\alpha L + \alpha + 1)(S_t - S_{t+1})}{1 - 2\alpha L} + \frac{1}{2}\|x_t - x_{t-1}\| + \\
& \frac{\alpha}{2(\alpha L + \alpha + 1)}\|\nabla f(x_{t-1}) - g_{t-1}\| + \frac{\alpha}{2(\alpha L + \alpha + 1)}[\varphi'(u_t)]^{-1}.
\end{aligned} \tag{37}$$

Summing from $t = \tilde{t}$ to $T - 1$, we obtain

$$\begin{aligned}
\sum_{t=\tilde{t}}^{T-1}\|x_{t+1} - x_t\| \leq & \frac{2(\alpha L + \alpha + 1)(S_{\tilde{t}} - S_T)}{1 - 2\alpha L} + \frac{1}{2}\sum_{t=\tilde{t}}^{T-1}\|x_t - x_{t-1}\| + \\
& \frac{\alpha}{2(\alpha L + \alpha + 1)}\sum_{t=\tilde{t}}^{T-1}\|\nabla f(x_{t-1}) - g_{t-1}\| + \frac{\alpha}{2(\alpha L + \alpha + 1)}\sum_{t=\tilde{t}}^{T-1}[\varphi'(u_t)]^{-1}.
\end{aligned}$$

Since $S_t \geq 0$ for all $t \geq 0$, and

$$\frac{1}{2}\sum_{t=\tilde{t}}^{T-1}\|x_t - x_{t-1}\| = \frac{1}{2}\sum_{t=\tilde{t}-1}^{T-2}\|x_{t+1} - x_t\| \leq \frac{1}{2}\sum_{t=\tilde{t}}^{T-1}\|x_{t+1} - x_t\| + \frac{1}{2}\|x_{\tilde{t}} - x_{\tilde{t}-1}\|,$$

then we can get

$$\begin{aligned}
\sum_{t=\tilde{t}}^{T-1}\|x_{t+1} - x_t\| \leq & \frac{4(\alpha L + \alpha + 1)S_{\tilde{t}}}{1 - 2\alpha L} + \|x_{\tilde{t}} - x_{\tilde{t}-1}\| + \\
& \frac{\alpha}{\alpha L + \alpha + 1}\sum_{t=\tilde{t}}^{T-1}\|\nabla f(x_{t-1}) - g_{t-1}\| + \frac{\alpha}{\alpha L + \alpha + 1}\sum_{t=\tilde{t}}^{T-1}[\varphi'(u_t)]^{-1}.
\end{aligned}$$

Let $T \to +\infty$, thus

$$\sum_{t=\tilde{t}}^{\infty} \|x_{t+1} - x_t\| \leq \frac{4(\alpha L + \alpha + 1)S_{\tilde{t}}}{1 - 2\alpha L} + \|x_{\tilde{t}} - x_{\tilde{t}-1}\| + \frac{\alpha}{\alpha L + \alpha + 1} \sum_{t=\tilde{t}}^{\infty} \|\nabla f(x_{t-1}) - g_{t-1}\| + \frac{\alpha}{\alpha L + \alpha + 1} \frac{(8\alpha L^3 C_1)^\theta}{c} \sum_{t=\tilde{t}}^{\infty} \left(\sum_{j=t}^{\infty} n^2 \gamma_j^2 \right)^\theta. \tag{38}$$

We next attack the third term in (38) by applying Lemma B.2, that is,

$$\sum_{t=\tilde{t}}^{\infty} \|\nabla f(x_{t-1}) - g_{t-1}\| \leq \sqrt{8L^3 C_1} \sum_{t=\tilde{t}}^{\infty} n\gamma_{t-1} < +\infty. \tag{39}$$

Moreover, the last term in (38) is a finite value due to

$$\sum_{t=\tilde{t}}^{\infty} \left(\sum_{j=t}^{\infty} n^2 \gamma_j^2 \right)^\theta < +\infty. \tag{40}$$

Combining (38), (39) and (40), we can get

$$\sum_{t=0}^{\infty} \|x_{t+1} - x_t\| < +\infty \text{ and } \|x_{t+1} - x_t\| \to 0.$$

This implies that $\{x_t\}$ has finite length, and hence converges to some x^*. Finally, by Theorem 2.4, the limit x^* is a critical point of problem (1). □

C.2 Proof of Convergence Rate

C.2.1 Important lemmas

We first establish several important lemmas that provide the key estimates for the subsequent convergence rate analysis. The following lemma gives key estimates for the corrected function value residual and the tail length of the iterates, which form the basis of the subsequent convergence rate analysis.

Lemma C.2 (Corrected Residual and Tail Length). *Supposing conditions in Theorem 3.2 and Assumption 3.1 are satisfied. Then, there exist constants $C > 0$, the following statements hold for all sufficiently large t.*

(i) $\left(H(x_{t+1}, w_t) - \bar{H} + u_{t+1}\right)^\theta \leq C\|x_{t+1} - x_t\| + Cn\gamma_t + Cu_{t+1}^\theta$.

(ii) $\|x_t - x^*\| \leq CS_t + C\|x_t - x_{t-1}\| + C\sum_{j=t}^{\infty} n\gamma_{j-1} + C\sum_{j=t}^{\infty} \left(\sum_{\ell=j}^{\infty} n^2\gamma_\ell^2\right)^\theta$.

Proof. We first prove part (i). The KL inequality (5) and Proposition B.4 (ii) give

$$|H(x_{t+1}, w_t) - \bar{H}|^\theta \leq C\|x_{t+1} - x_t\| + Cn\gamma_t.$$

Thus, we have

$$\begin{aligned}\left(H(x_{t+1}, w_t) - \bar{H} + u_{t+1}\right)^\theta &\leq |H(x_{t+1}, w_t) - \bar{H}|^\theta + u_{t+1}^\theta \\ &\leq C\|x_{t+1} - x_t\| + Cn\gamma_t + Cu_{t+1}^\theta,\end{aligned}$$

where the first inequality since $|H(x_{t+1}, w_t) - \bar{H}|$ and u_t is nonnegative and uses $(a+b)^\theta \leq a^\theta + b^\theta$ for $a, b > 0$ and $\theta \in (0, 1)$.

For part (ii). Replacing t by j in (37) summing from $j = t$ to N yield

$$\sum_{j=t}^{N} \|x_{j+1} - x_j\| \leq C\sum_{j=t}^{N}(S_j - S_{j+1}) + \frac{1}{2}\sum_{j=t}^{N}\|x_j - x_{j-1}\| + C\sum_{j=t}^{N}\|\nabla f(x_{j-1}) - g_{j-1}\| + C\sum_{j=t}^{N}[\varphi'(u_j)]^{-1}. \tag{41}$$

Then we attack the second term by

$$\sum_{j=t}^{N}\|x_j - x_{j-1}\| \le \|x_t - x_{t-1}\| + \sum_{j=t}^{N}\|x_{j+1} - x_j\|.$$

Thus, rearrange (41) and it gives

$$\sum_{j=t}^{N}\|x_{j+1} - x_j\| \le CS_t + C\|x_t - x_{t-1}\| + C\sum_{j=t}^{N}\|\nabla f(x_{j-1}) - g_{j-1}\| + C\sum_{j=t}^{N}[\varphi'(u_j)]^{-1}. \tag{42}$$

Letting $N \to \infty$, we have

$$\begin{aligned}\|x_t - x^*\| = \left\|\sum_{j=t}^{\infty}(x_j - x_{j+1})\right\| \le& \sum_{j=t}^{\infty}\|x_{j+1} - x_j\| \\ \le& CS_t + C\|x_t - x_{t-1}\| + C\sum_{j=t}^{\infty}\|\nabla f(x_{j-1}) - g_{j-1}\| + C\sum_{j=t}^{\infty}[\varphi'(u_j)]^{-1} \\ \le& CS_t + C\|x_t - x_{t-1}\| + \sqrt{8L^3C_1}\sum_{j=t}^{\infty}n\gamma_{j-1} + C\sum_{j=t}^{\infty}\left(\sum_{\ell=j}^{\infty}n^2\gamma_\ell^2\right)^{\theta},\end{aligned}$$

where the equality is due to Theorem 3.2, the second inequality since equation 42, and the last inequality applies lemma B.2 and the definition of φ and u_t. □

In order to derive explicit rates from the recursive inequalities, we introduce the following lemma as a preparation for the convergence rate analysis.

Lemma C.3 (**Sequence Decay Rates**). *Let $\{a_t\}$ be a nonnegative sequence. For constants $c_0, C_0 > 0$, and all sufficiently large t, the following statements hold:*

(i) If $p > 1$, $\beta \ge 1$, $\vartheta \in (\frac{1}{2}, 1)$, and $\{a_t\}$ satisfies

$$a_t - a_{t+1} \ge c_0 a_{t+1}^{2\vartheta} - C_0(t+\beta)^{-2p}, \tag{43}$$

then $a_t = O\left(t^{-r}\right)$, where $r = \min\left\{2p - 1, \frac{1}{2\vartheta - 1}\right\}$.

(ii) If $\rho \in (0,1)$, $\vartheta \in (0, \frac{1}{2})$, and $\{a_t\}$ satisfies

$$a_t - a_{t+1} \ge c_0 a_{t+1}^{2\vartheta} - C_0\rho^{4\vartheta t}, \tag{44}$$

then $a_t = O\left(\rho^{2t}\right)$.

Proof. For part (i). Let $b_t := M(t+\beta)^{-r}$, where $M > 0$ is a constant. The mean value theorem gives that there exists a $\xi \in (t+\beta, t+\beta+1)$ such that $b_{t+1} - b_t = -Mr\xi^{-(r+1)}$, meanwhile $\{b_t\}$ is a decrease sequence, then we have

$$\begin{aligned}b_t - b_{t+1} \le& rM(t+\beta)^{-(r+1)} \\ \le& rM(t+\beta)^{-2\vartheta r} \\ \le& (c_0\, 2^{-2\vartheta r}M^{2\vartheta} - C_0)(t+\beta)^{-2\vartheta r} \\ \le& c_0\, 2^{-2\vartheta r}M^{2\vartheta}\,(t+\beta)^{-2\vartheta r} - C_0(t+\beta)^{-2p} \\ \le& c_0 b_{t+1}^{2\vartheta} - C_0(t+\beta)^{-2p},\end{aligned}$$

where the first inequality is due to the mean value theorem, the second inequality since the definition of r gives $2\vartheta r \le r+1$ and $2\vartheta r \le 2p$, the third inequality applies that a sufficiently large M can

lead to $c_0\, 2^{-2\vartheta r} M^{2\vartheta} \geq rM + C_0$, and the last inequality since $t+1+\beta \leq 2(t+\beta)$ leads to $c_0 b_{t+1}^{2\vartheta} \geq c_0\, 2^{-2\vartheta r} M^{2\vartheta} (t+\beta)^{-2\vartheta r}$.

Thus, we can get

$$b_{t+1} + c_0 b_{t+1}^{2\vartheta} \geq b_t + C_0 (t+\beta)^{-2p}. \tag{45}$$

Next we apply mathematical induction to complete the proof. Let $a_t \leq b_t$, then (43) and (45) lead to

$$\begin{aligned} a_{t+1} + c_0 a_{t+1}^{2\vartheta} \leq & a_t + C_0 (t+\beta)^{-2p} \leq b_t + C_0 (t+\beta)^{-2p} \\ \leq & b_{t+1} + c_0 b_{t+1}^{2\vartheta}. \end{aligned} \tag{46}$$

Since the map $s \mapsto s + c_0 s^{2\vartheta}$ is strictly increasing on $[0, \infty)$, (46) yields $a_{t+1} \leq b_{t+1}$. Thus, the induction $a_t \leq b_t$ yields $a_t \leq M(t+\beta)^{-r} = \mathcal{O}(t^{-r})$.

For part (ii). First, we show that $a_t \to 0$. (44) yields that for all sufficiently large t,

$$c_0 a_{t+1}^{2\vartheta} \leq a_t - a_{t+1} + C_0 \rho^{4\vartheta t}.$$

Summing the above inequality from $t = t_0$ to T, where t_0 is sufficiently large such that (44) holds, then we have

$$c_0 \sum_{t=t_0}^{T} a_{t+1}^{2\vartheta} \leq a_{t_0} - a_{T+1} + C_0 \sum_{t=t_0}^{T} \rho^{4\vartheta t} \leq a_{t_0} + C_0 \sum_{t=t_0}^{\infty} \rho^{4\vartheta t} < +\infty,$$

where the last inequality follows from $\rho \in (0,1)$ and $\vartheta > 0$. Letting $T \to \infty$, we obtain

$$\sum_{t=t_0}^{\infty} a_{t+1}^{2\vartheta} < +\infty$$

and hence $a_t \to 0$. Moreover, (44) gives

$$\begin{aligned} a_{t+1} \leq & c_0^{-\frac{1}{2\vartheta}} (a_t + C_0 \rho^{4\vartheta t})^{\frac{1}{2\vartheta}} \\ \leq & C a_t^{\frac{1}{2\vartheta}} + C\rho^{2t} \\ = & C a_t^{\frac{1}{2\vartheta}-1} a_t + C\rho^{2t} \\ \leq & \frac{\rho^2}{2} a_t + C\rho^{2t}, \end{aligned}$$

where the second inequality applies the inequality $(a+b)^\lambda \leq 2\lambda^{-1}(a^\lambda + b^\lambda)$ with $a, b \geq 0$ and $\lambda > 1$, and the last inequality since a sufficiently large t leads to $a_t \to 0$, then $a_t^{\frac{1}{2\vartheta}-1} \to 0$ with $\frac{1}{2\vartheta} - 1 > 0$, thus $C a_t^{\frac{1}{2\vartheta}-1} \leq \frac{\rho^2}{2}$ is satisfied for all sufficiently large t.

Similarly to the proof of part (i), let $b_t := M\rho^{2t}$, where $M > 0$ is a constant. Next we apply mathematical induction to complete the proof. Let $a_t \leq b_t$, then we have

$$a_{t+1} \leq \frac{\rho^2}{2} a_t + C\rho^{2t} \leq \frac{\rho^2}{2} b_t + C\rho^{2t} = (\frac{M\rho^2}{2} + C)\rho^{2t} \leq M\rho^2 \rho^{2t} = b_{t+1},$$

where the equality follows from the definition of b_t, and the last inequality is due to $\frac{M\rho^2}{2} > C$ when M is sufficiently large. Thus, the induction $a_t \leq b_t$ yields $a_t \leq M\rho^{2t} = \mathcal{O}(\rho^{2t})$.

□

C.2.2 Proof of Theorem 3.4

Proof. From the definition of γ_t, we have

$$n\gamma_t = \frac{n\gamma_0}{(t+\beta)^p} = \mathcal{O}(t^{-p}). \tag{47}$$

Then, applying the integral test $\sum_{j=t}^{\infty}(j+\beta)^{-q} \leq (t+\beta)^{-q} + \int_t^{\infty}(x+\beta)^{-q}\,dx$ to (6), it gives

$$u_t = 8\alpha L^3 C_1 \sum_{j=t}^{\infty} n^2\gamma_j^2 = 8\alpha L^3 C_1 n^2\gamma_0^2 \sum_{j=t}^{\infty} \frac{1}{(j+\beta)^{2p}} = \mathcal{O}(t^{-(2p-1)}), \tag{48}$$

and

$$\sum_{j=t}^{\infty} n\gamma_{j-1} = n\gamma_0 \sum_{j=t-1}^{\infty} \frac{1}{(j+\beta)^{p}} = \mathcal{O}(t^{-(p-1)}). \tag{49}$$

Define

$$\widehat{\theta} := \max\left\{\theta, \frac{p}{2p-1}\right\}. \tag{50}$$

Since $\widehat{\theta} \geq \theta$, then the KL inequality (5) implies

$$|H(x,w) - \bar{H}|^{\widehat{\theta}} \leq |H(x,w) - \bar{H}|^{\theta} \leq c \cdot \operatorname{dist}(0, \partial H(x,w)),$$

where the first inequality follows from $|H(x,w) - \bar{H}|^{\widehat{\theta}} < 1$ for a sufficiently large t. Hence, $\widehat{\theta}$ is also an admissible KL exponent. Moreover, (48) yields

$$\sum_{t=0}^{\infty}\left(\sum_{j=t}^{\infty} n^2\gamma_j^2\right)^{\widehat{\theta}} \leq C\sum_{t=0}^{\infty}(t+\beta)^{-(2p-1)\widehat{\theta}} < +\infty, \tag{51}$$

where the last inequality follows from $(2p-1)\widehat{\theta} \geq p > 1$. Therefore, Theorem 3.2 is applicable with the KL exponent $\widehat{\theta}$, and we can discuss in the following for the convergence rate of $\theta \in [0, \frac{p}{2p-1}]$ and $\theta \in (\frac{p}{2p-1}, 1)$.

For part (i), let $\theta \in [0, \frac{p}{2p-1}]$. Then (50) gives $\widehat{\theta} = \frac{p}{2p-1}$. Therefore, Lemma C.2 can be applied with the KL exponent $\frac{p}{2p-1}$. First, we can deduce in the following

$$\begin{aligned} H(x_t, w_{t-1}) + u_t - \bar{H} - \big(H(x_{t+1}, w_t) + u_{t+1} - \bar{H}\big) & \\ \geq & C\|x_{t+1} - x_t\|^2 \qquad (52)\\ \geq & c_0\big(H(x_{t+1}, w_t) + u_{t+1} - \bar{H}\big)^{\frac{2p}{2p-1}} - C_0(t+\beta)^{-2p}, \qquad (53) \end{aligned}$$

where the first inequality is due to (35), the last inequality is combining Lemma C.2 (i) with KL exponent $\frac{p}{2p-1}$, (47) and (48).

Applying the Lemma C.3 (i) to (53) with $\vartheta = \frac{p}{2p-1}$, it yields

$$H(x_t, w_{t-1}) + u_t - \bar{H} = \mathcal{O}(t^{-(2p-1)}). \tag{54}$$

Since (48), (54) further gives

$$|H(x_t, w_{t-1}) - \bar{H}| = \mathcal{O}(t^{-(2p-1)}). \tag{55}$$

Moreover, Lemma B.3 and Proposition B.4 indicate $H(x_{t+1}, w_t) \leq P(x_t) + C\gamma_t^2$, then we have

$$\bar{P} - P(x_t) \leq \bar{H} - H(x_{t+1}, w_t) + C\gamma_t^2 \leq |\bar{H} - H(x_{t+1}, w_t)| + C\gamma_t^2. \tag{56}$$

From the definition of $P(x_t)$ and $H(x_t, w_{t-1})$, we can get

$$P(x_t) - \bar{P} \leq H(x_t, w_{t-1}) - \bar{H} \leq |H(x_t, w_{t-1}) - \bar{H}|. \tag{57}$$

Combining (56) and (57), it follows that

$$|P(x_t) - \bar{P}| \leq |H(x_t, w_{t-1}) - \bar{H}| + |\bar{H} - H(x_{t+1}, w_t)| + C\gamma_t^2. \tag{58}$$

Since $\theta \leq \frac{p}{2p-1}$, (6), (55) and (58) give $|P(x_t) - \bar{P}| = \mathcal{O}(t^{-(2p-1)})$.

Next, (52) and (54) indicate

$$\|x_t - x_{t-1}\| \leq C\big(H(x_{t-1}, w_{t-2}) + u_{t-1} - \bar{H}\big)^{\frac{1}{2}} = \mathcal{O}(t^{-(p-\frac{1}{2})}). \tag{59}$$

Then, combining Proposition B.4 (ii) and (59), we have

$$\operatorname{dist}(0, \partial H(x_t, w_{t-1})) \leq C_2\|x_t - x_{t-1}\| + C_3 n\gamma_{t-1} = \mathcal{O}(t^{-(p-\frac{1}{2})}).$$

In the following, we estimate the last iterate sequence directly from Lemma C.2 (ii). First we consider the order of S_t by

$$S_t = C(H(x_t, w_{t-1}) + u_t - \bar{H}\big)^{1-\frac{p}{2p-1}} = \mathcal{O}(t^{-(p-1)}), \tag{60}$$

where the first equality is from the definition of S_t, and the last equation is due to (54). Then, we attack the last term of Lemma C.2 (ii):

$$\sum_{j=t}^{\infty}\left(\sum_{\ell=j}^{\infty} n^2\gamma_\ell^2\right)^{\frac{p}{2p-1}} = \sum_{j=t}^{\infty}\mathcal{O}(j^{-p}) = \mathcal{O}(t^{-(p-1)}), \tag{61}$$

where the first equality is due to (48). Combining (60), (59), (61) and (49) in Lemma C.2 (ii), then we have

$$\|x_t - x^*\| = \mathcal{O}(t^{-(p-1)}).$$

For part (ii). In this case, $\theta \in (\frac{p}{2p-1}, 1)$, thus the KL exponent is denoted by θ. The same argument as the former gives

$$\begin{aligned} H(x_t, w_{t-1}) + u_t - \bar{H} - \big(H(x_{t+1}, w_t) + u_{t+1} - \bar{H}\big) \\ \geq c_0\big(H(x_{t+1}, w_t) + u_{t+1} - \bar{H}\big)^{2\theta} - C_0(t+\beta)^{-2p}, \end{aligned} \tag{62}$$

where the inequality is due to $\theta > \frac{p}{2p-1}$ leading to the decay of $t^{-\theta(2p-1)}$ faster than t^{-p}. Then, apply Lemma C.3 (i) to (62) with $\vartheta = \theta$. Since $\theta > \frac{p}{2p-1}$, it gives

$$H(x_t, w_{t-1}) + u_t - \bar{H} = \mathcal{O}(t^{-\frac{1}{2\theta-1}}). \tag{63}$$

Since $u_t = \mathcal{O}(t^{-(2p-1)})$, which decays faster than $\mathcal{O}(t^{-\frac{1}{2\theta-1}})$. Then, (63) indicates

$$|H(x_t, w_{t-1}) - \bar{H}| = \mathcal{O}(t^{-\frac{1}{2\theta-1}}). \tag{64}$$

Moreover, since $\theta > \frac{p}{2p-1}$, (6), (58) and (64) indicate $|P(x_t) - \bar{P}| = \mathcal{O}(t^{-\frac{1}{2\theta-1}})$.

Next, (52) and (63) indicates

$$\|x_t - x_{t-1}\| \leq C\big(H(x_{t-1}, w_{t-2}) + u_{t-1} - \bar{H}\big)^{\frac{1}{2}} = \mathcal{O}(t^{-\frac{1}{2(2\theta-1)}}). \tag{65}$$

Then, combining Proposition B.4 (ii) and (65), we have

$$\operatorname{dist}(0, \partial H(x_t, w_{t-1})) \leq C_2\|x_t - x_{t-1}\| + C_3 n\gamma_{t-1} = \mathcal{O}(t^{-\frac{1}{2(2\theta-1)}}).$$

In the following, we estimate the last iterate sequence directly from Lemma C.2 (ii). First we consider the order of S_t by

$$S_t = C(H(x_t, w_{t-1}) + u_t - \bar{H}\big)^{1-\theta} = \mathcal{O}(t^{-\frac{1-\theta}{2\theta-1}}) \tag{66}$$

Then, we attack the last term of Lemma C.2 (ii):

$$\sum_{j=t}^{\infty}\left(\sum_{\ell=j}^{\infty} n^2\gamma_\ell^2\right)^{\theta} = \sum_{j=t}^{\infty}\mathcal{O}(j^{-\theta(2p-1)}) = \mathcal{O}(t^{-(\theta(2p-1)-1)}), \tag{67}$$

where the first equality is due to (48). Combining (66), (65), (49) and (67) in Lemma C.2 (ii), then we have

$$\|x_t - x^*\| = \mathcal{O}(t^{-\frac{1-\theta}{2\theta-1}}) + \mathcal{O}(t^{-\frac{1}{2(2\theta-1)}}) + \mathcal{O}(t^{-(p-1)}) + \mathcal{O}(t^{-(\theta(2p-1)-1)}) = \mathcal{O}(t^{-\frac{1-\theta}{2\theta-1}}),$$

where the last equality is due to $p - 1 > \frac{1-\theta}{2\theta-1}$, $\theta(2p-1) - 1 > \frac{1-\theta}{2\theta-1}$ and $\frac{1}{2(2\theta-1)} \geq \frac{1-\theta}{2\theta-1}$.

□

C.2.3 Proof of Theorem 3.6

Proof. First, from the definition of γ_t, we can get

$$n\gamma_t = \mathcal{O}(\rho^t) \text{ and } u_t = C\sum_{j=t}^{\infty}\gamma_j^2 = \mathcal{O}(\rho^{2t}). \tag{68}$$

We set $\widehat{\theta} = \theta$ if $\theta \in (0, \frac{1}{2})$, and choose any $\widehat{\theta} \in (0, \frac{1}{2})$ if $\theta = 0$. In the latter case, (5) gives

$$|H(x,w) - \bar{H}|^{\widehat{\theta}} \leq 1 \leq c \cdot \operatorname{dist}(0, \partial H(x,w)).$$

Hence, H satisfies the KL inequality with exponent $\widehat{\theta} \in (0, \frac{1}{2})$. Moreover, the definition of γ_t follows $\sum_{t=0}^{\infty} n\gamma_t = n\gamma_0 \sum_{t=0}^{\infty} \rho^t < +\infty$ and $\sum_{t=0}^{\infty}\left(\sum_{j=t}^{\infty} n^2\gamma_j^2\right)^{\widehat{\theta}} = \left(\frac{n^2\gamma_0^2}{1-\rho^2}\right)^{\widehat{\theta}} \sum_{t=0}^{\infty}\rho^{2\widehat{\theta}t} < +\infty$. Therefore, Theorem 3.2 is applicable with the KL exponent $\widehat{\theta}$.

Besides, Lemma C.2 (i) points out that

$$\begin{aligned}\big(H(x_{t+1},w_t) - \bar{H} + u_{t+1}\big)^{\widehat{\theta}} \leq& C\|x_{t+1} - x_t\| + Cn\gamma_t + Cu_{t+1}^{\widehat{\theta}} \\ \leq& C\|x_{t+1} - x_t\| + C\rho^{2\widehat{\theta}t},\end{aligned} \tag{69}$$

where the last inequality since $\widehat{\theta} \in [0, \frac{1}{2})$ and (68) lead to $Cn\gamma_t + Cu_{t+1}^{\widehat{\theta}} = \mathcal{O}(\rho^{2\widehat{\theta}t})$. Then, we have

$$\begin{aligned}H(x_t,w_{t-1}) + u_t - \bar{H} - \big(H(x_{t+1},w_t) + u_{t+1} - \bar{H}\big)& \\ \geq C\|x_{t+1} - x_t\|^2& \qquad (70)\\ \geq c_0\big(H(x_{t+1},w_t) + u_{t+1} - \bar{H}\big)^{2\widehat{\theta}} - C_0\rho^{4\widehat{\theta}t},& \qquad (71)\end{aligned}$$

where the first inequality follows from (35), and the second inequality follows from (69). Applying Lemma C.3 (ii) to (71) with $\vartheta = \widehat{\theta}$ yields

$$H(x_t,w_{t-1}) + u_t - \bar{H} = \mathcal{O}(\rho^{2t}). \tag{72}$$

Combining with (68), then we can attain the convergence $|H(x_t,w_{t-1}) - \bar{H}| = \mathcal{O}(\rho^{2t})$. Simultaneously, (58) leads to $|P(x_t) - \bar{P}| = \mathcal{O}(\rho^{2t})$.

Next, (70) and (72) indicates

$$\|x_t - x_{t-1}\| \leq C\big(H(x_{t-1},w_{t-2}) + u_{t-1} - \bar{H}\big)^{\frac{1}{2}} = \mathcal{O}(\rho^t). \tag{73}$$

Then, combining Proposition B.4 (ii), (68) and (73), we have

$$\operatorname{dist}(0, \partial H(x_t,w_{t-1})) \leq C_2\|x_t - x_{t-1}\| + C_3 n\gamma_{t-1} = \mathcal{O}(\rho^t).$$

Finally, Theorem 3.2 gives $x_t \to x^*$, then, combining (73) further yields

$$\|x_t - x^*\| \leq \sum_{j=t}^{\infty}\|x_{j+1} - x_j\| = \mathcal{O}\left(\sum_{j=t}^{\infty}\rho^j\right) = \mathcal{O}(\rho^t).$$

□

D Experimental Details

D.1 Related Works

For completeness, we present the pseudocode of proximal random reshuffling (proxRR) and stochastic gradient descent algorithm in the proximal difference of convex setting (SGD-DC), which are used as representative comparison methods in this paper.

Algorithm 2 Proximal Random Reshuffling (proxRR)

Input: Stepsizes $\gamma_t > 0$, initial vector $x_0 \in \mathbb{R}^d$, number of epochs T

for epochs $t = 0, 1, \ldots, T-1$ **do**
  Sample a permutation $\pi = (\pi_0, \pi_1, \ldots, \pi_{n-1})$ of $[n]$
  $x_t^0 = x_t$
  for $i = 0, 1, \ldots, n-1$ **do**
    $x_t^{i+1} = x_t^i - \gamma_t \nabla f_{\pi_i}(x_t^i)$
  end for
  $x_{t+1} = \text{prox}_{\gamma_t n \psi}(x_t^n)$
end for

ProxRR processes all component functions once in a randomly reshuffled order through incremental gradient updates and applies a single proximal update at the end of each epoch. As shown in Algorithm 2, the proximal parameter is $n\gamma_t$ and is therefore directly determined by the reshuffling stepsize γ_t.

We next present SGD-DC, which is used as a comparison method for the nonsmooth nonconvex difference-structured problem. Different from proxRR, SGD-DC incorporates the difference structure into every inner iteration. Specifically, for each sampled component, it selects $w_t^i \in \partial h_2(x_t^i)$ and then performs a proximal update with respect to h_1, as summarized in Algorithm 3.

Algorithm 3 Stochastic Gradient Descent algorithm in the Proximal Difference of Convex setting (SGD-DC)

Input: Stepsizes $\alpha_t > 0$, initial vector $x_0 \in \mathbb{R}^d$, number of epochs T

for epochs $t = 0, 1, \cdots, T-1$ **do**
  $x_t^0 = x_t$
  for $i = 0, 1, \cdots, n-1$ **do**
    Sample an index j_t^i uniformly at random from $[n]$
    $w_t^i \in \partial h_2(x_t^i)$
    $x_t^{i+1} = \text{prox}_{\alpha_t h_1}(x_t^i - \alpha_t \nabla f_{j_t^i}(x_t^i) + \alpha_t w_t^i)$
  end for
  $x_{t+1} = x_t^n$
end for

D.2 Implementation Details

In this subsection, we provide the implementation details for the two sets of experiments presented in Subsections 4.1 and 4.2.

Implementation details for 4.1. We conduct the experiment on a1a, a9a and splice datasets and compare PNRR with proxSGD, proxRR and norm-PRR. Parameters are set to $\lambda_1 = 10^{-5}$, $\lambda_2 = L/n$, and the proximity operator is equal to the combination of the proximity operator of $\lambda_1\|x\|_1$ and then the proximity operator of $\frac{\lambda_2}{2}\|x\|^2$. Besides, we take the stepsizes $\gamma_t = \frac{2}{n(t+1)}$ and $\alpha = 2$.

Implementation details for 4.2. We conduct the binary classification task on MNIST, GISETTE and CINA datasets $(A, b) \in \mathbb{R}^{n\times d} \times \{-1, 1\}^n$ with PNRR, SGD-DC and full gradient algorithm. For MNIST, we only retain samples from two classes (digits 3 and 8). The parameters are set to $\lambda_1 = 0.01$ and $\lambda_2 = 0.0001$. Besides, we take the stepsizes $\gamma_t = \frac{2}{n(t+1)^{0.51}}$ and $\alpha = 1.5$ in MNIST and CINA, and $\gamma_t = \frac{0.8}{n(t+1)^{0.8}}$ and $\alpha = 1.1$ in GISETTE for PNRR.