Journal of Machine Learning Research 26 (2025) 1-46 Submitted 6/24; Revised 3/25; Published 7/25

A New Random Reshuffling Method for Nonsmooth Nonconvex Finite-sum Optimization

Junwen Qiu JWQIU@NUS.EDU.SG
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Guangdong, 518172, P.R. China

Industrial Systems Engineering & Management
National University of Singapore
Singapore, 119077, Singapore

Xiao Li LIXIAO@CUHK.EDU.CN

Andre Milzarek ANDREMILZAREK@CUHK.EDU.CN
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Guangdong, 518172, P.R. China

Editor: Nicolas Le Roux

Abstract

Random reshuffling techniques are prevalent in large-scale applications, such as training neural networks. While the convergence and acceleration effects of random reshuffling-type methods are fairly well understood in the smooth setting, much less studies seem available in the nonsmooth case. In this work, we design a new normal map-based proximal random reshuffling (norm-PRR) method for nonsmooth nonconvex finite-sum problems. We show that norm-PRR achieves the iteration complexity $\mathcal{O}(n^{-1/3}T^{-2/3})$ where n denotes the number of component functions $f(\cdot, i)$ and T counts the total number of iterations. This improves the currently known complexity bounds for this class of problems by a factor of $n^{-1/3}$ in terms of the number of gradient evaluations. Additionally, we prove that norm-PRR converges linearly under the (global) Polyak-Łojasiewicz condition and in the interpolation setting. We further complement these non-asymptotic results and provide an in-depth analysis of the asymptotic properties of norm-PRR. Specifically, under the (local) Kurdyka-Łojasiewicz inequality, the whole sequence of iterates generated by norm-PRR is shown to converge to a single stationary point. Moreover, we derive last-iterate convergence rates that can match those in the smooth, strongly convex setting. Finally, numerical experiments are performed on nonconvex classification tasks to illustrate the efficiency of the proposed approach.

Keywords: Proximal random reshuffling, normal map, complexity, asymptotic convergence, nonconvexity, nonsmoothness

1. Introduction

In this work, we consider the composite optimization problem

$$\min_{w\in\mathbb{R}^d} \ \psi(w) := f(w) + \varphi(w), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable but not necessarily convex and $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ is a weakly convex, lower semicontinuous (lsc.), and proper mapping. Composite models of the form (1) are ubiquitous in structured large-scale applications and stochastic optimization, including, e.g., machine learning (Bottou, 2010; Bottou et al., 2018; LeCun et al., 2015), statistical learning and sparse regression (Tibshirani, 1996; Hastie et al., 2009; Shalev-Shwartz and Tewari, 2011), image and signal processing (Combettes and Pesquet, 2011; Chambolle and Pock, 2011).

We are interested in the case where the smooth part of the objective function has a finite-sum structure, i.e., f can be represented as follows:

$$f(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, i). \tag{2}$$

In machine learning tasks, the number of component functions, n, is typically connected to the number of underlying data points and is often prohibitively large. In this scenario, computing the full gradient ∇f is highly expensive or even impossible. Nonetheless, many classical approaches for solving (1) require exact gradient information in each step, see (Lions and Mercier, 1979; Mine and Fukushima, 1981; Attouch et al., 2010, 2013). In this paper, we adopt a stochastic approximation-based perspective and assume that the gradient of only one single function $f(\cdot, i)$ is available at each iteration. Furthermore, the evaluation of the proximity operator of φ is assumed to be tractable (as is the case in many applications, (Combettes and Wajs, 2005)). Our main objective is to develop a novel proximal random reshuffling algorithm for problem (1) with convincing practical and theoretical properties.

1.1 Related Works and Motivations

Smooth problems. When $\varphi \equiv 0$, (1) reduces to a standard smooth optimization problem. In this case, the stochastic gradient descent method (SGD) proposed in the seminal work by Robbins and Monro (1951) is the prototypical approach for solving (1). The core step of SGD is given by:

$$w^{k+1} = w^k - \alpha_k \nabla f(w^k, i_k), \quad \text{where } i_k \text{ is chosen randomly from } [n]. \tag{SGD}$$

SGD has been studied extensively during the past decades; we refer to (Chung, 1954; Rakhlin et al., 2012; Nguyen et al., 2018; Gower et al., 2019, 2021). In the strongly convex setting, SGD was proven to converge to the optimal solution w^* almost surely with the rate $\|w^k - w^*\| = \mathcal{O}(1/\sqrt{k})$, (Chung, 1954). Moreover, the lower bounds derived in (Nemirovskij and Yudin, 1983; Agarwal et al., 2009; Nguyen et al., 2019) indicate that the rate $\mathcal{O}(1/\sqrt{k})$ is tight for SGD up to a constant. In the nonconvex case, convergence of SGD can be expressed in terms of complexity bounds,

$$\min_{k=1,\dots,T} \mathbb{E}[\|\nabla f(w^k)\|^2] = \mathcal{O}(T^{-1/2}),$$

see, e.g., (Ghadimi and Lan, 2013). Recently, empirical evidence in (Bottou, 2012; Shamir, 2016; Mishchenko et al., 2020) suggests that variants of SGD with *without-replacement sampling* can attain faster convergence. Incorporating such sampling scheme leads to the

so-called random reshuffling method (RR), which is shown below:

$$\left[\begin{array}{l} \text{Set } w_1^k = w^k \text{ and generate a random permutation } \pi^k \text{ of } [n]; \\ \quad \textbf{For } \ i = 1, 2, \ldots, n \ \textbf{ do:} \quad w_{i+1}^k = w_i^k - \alpha_k \nabla f(w_i^k, \pi_i^k); \\ \text{Set } w^{k+1} = w_{n+1}^k. \end{array}\right. \tag{RR}$$

For problems with the finite-sum structure (2), RR exhibits superior theoretical guarantees compared to SGD (Gürbüzbalaban et al., 2021; Mishchenko et al., 2020; Nguyen et al., 2021; Li et al., 2023). In (Gürbüzbalaban et al., 2021), the first theoretical guarantee for faster convergence of RR is provided. In particular, in the strongly convex case, the sequence of q-suffix averaged iterates is shown to converge to the unique optimal solution at a rate of $\mathcal{O}(1/k)$ with high probability. Following this pioneering work, subsequent research has began to explore and understand the theoretical behavior of RR, (Haochen and Sra, 2019; Nagaraj et al., 2019; Mishchenko et al., 2020; Nguyen et al., 2021). In the nonconvex case and under a general variance bound, the authors in (Mishchenko et al., 2020; Nguyen et al., 2021) established the complexity bound

$$\min_{k=1,\ldots,T} \mathbb{E}[\|\nabla f(w^k)\|^2] = \mathcal{O}(n^{-1/3}T^{-2/3}),$$

for RR, when a uniform without-replacement sampling scheme[1] is applied. In addition, Nguyen et al. (2021) show $\liminf_{k\to\infty} \|\nabla f(w^k)\| = 0$ for step sizes $\{\alpha_k\}_k$ satisfying $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^3 < \infty$. In (Li et al., 2023), this was strengthened to full asymptotic convergence, $\nabla f(w^k) \to 0$, and to last-iterate convergence of the form $\|w^k - w^*\| = \mathcal{O}(1/k)$, $k \to \infty$, under the Kurdyka-Łojasiewicz (KL) inequality. Here, w^* generally denotes a stationary point of the problem $\min_w f(w)$, i.e., $\nabla f(w^*) = 0$.

Composite problems. The proximal stochastic gradient method (PSGD) is a standard stochastic approach for solving the composite problem (1). The update of PSGD is given by:

$$w^{k+1} = \operatorname{prox}_{\alpha_k \varphi}(w^k - \alpha_k \nabla f(w^k, i_k)), \quad \text{where } i_k \text{ is chosen randomly from } [n]. \tag{PSGD}$$

In contrast to SGD, the convergence behavior of PSGD is less understood, especially when f is nonconvex. Davis and Drusvyatskiy (2019) present one of the first complexity results for PSGD in the nonconvex setting, i.e., it holds that

$$\min_{k=1,\ldots,T} \mathbb{E}[\|\mathcal{G}_\lambda(w^k)\|^2] = \mathcal{O}(\textstyle\sum_{k=1}^T \alpha_k^2 \,/\, \sum_{k=1}^T \alpha_k), \quad \lambda > 0, \tag{3}$$

where $\mathcal{G}_\lambda(w) := \lambda^{-1}(w - \operatorname{prox}_{\lambda\varphi}(w - \lambda\nabla f(w)))$ is a basic stationarity measure for (1). Earlier studies of PSGD for nonconvex f appeared in (Ghadimi et al., 2016), where convergence is shown if the variance vanishes as $k \to \infty$. Asymptotic convergence guarantees are discussed in (Majewski et al., 2018; Duchi and Ruan, 2018; Davis et al., 2020; Li and Milzarek, 2022) under (almost) sure boundedness of the iterates $\{w^k\}_k$ or global Lipschitz continuity of φ. The easier convex and strongly convex cases have been investigated, e.g., in (Ghadimi et al., 2016; Atchadé et al., 2017; Rosasco et al., 2020; Patrascu and Irofti, 2021). If f

1. Every $\nabla f(\cdot, i)$, $i \in [n]$ has equal probability to be selected without-replacement.

is convex and ψ is strongly convex, convergence in expectation to the unique solution w^*, $\mathbb{E}[\|w^k - w^*\|^2] = \mathcal{O}(1/k)$, can be ensured, (Rosasco et al., 2020; Patrascu and Irofti, 2021).

Naturally and in order to improve the performance of PSGD, we may consider suitable combinations of without-replacement sampling schemes and PSGD. An intuitive and straightforward extension of RR is to perform additional proximal steps "$\mathrm{prox}_{\alpha_k\varphi}(\cdot)$" after each inner iteration $i = 1, \ldots, n$ of RR. However, as highlighted in (Mishchenko et al., 2022, Example 1), such a combination prevents the accurate approximation of the full gradient ∇f after one epoch. Based on this observation, Mishchenko et al. (2022) develop a different proximal-type RR. Unlike PSGD, the proposed method performs a proximal step only once after each epoch. Therefore, we will refer to this approach as epoch-wise proximal random reshuffling method (e-PRR). The main iterative step of e-PRR is shown below:

$$\left[\begin{array}{l} \text{Set } w_1^k = w^k \text{ and generate a random permutation } \pi^k \text{ of } [n]; \\ \quad \textbf{For } \ i = 1, 2, \ldots, n \ \textbf{ do:} \quad w_{i+1}^k = w_i^k - \alpha_k \nabla f(w_i^k, \pi_i^k); \\ \text{Set } w^{k+1} = \mathrm{prox}_{n\alpha_k\varphi}(w_{n+1}^k). \end{array}\right. \tag{e-PRR}$$

In the nonconvex setting and under an additional bound connecting ∇f and the stationarity measure $\mathcal{G}_\lambda$, Mishchenko et al. (2022) derive the complexity bound[2]

$$\min_{k=1,\ldots,T} \mathbb{E}[\|\mathcal{G}_\lambda(w^k)\|^2] = \mathcal{O}(T^{-2/3} + n^{-1}T^{-2/3}), \quad \lambda \sim T^{-1/3}, \tag{4}$$

for e-PRR, where $\lambda > 0$ is a step size parameter. In a recent study of e-PRR, Liu and Zhou (2024) provide additional convergence rates for the objective function values $\{\psi(w^k)\}_k$ in the convex setting (i.e., each $f(\cdot, i)$ and φ are assumed to be convex).

1.2 Contributions

We design a new proximal random reshuffling method (norm-PRR) for nonconvex composite problems. In contrast to existing stochastic proximal methods, our approach is based on the so-called *normal map* which swaps the order of evaluating ∇f and the proximity operator $\mathrm{prox}_{\lambda\varphi}$. We show that this exchanged order generally exhibits a better compatibility with without-replacement sampling schemes. In particular, similar to PSGD but different from e-PRR, norm-PRR performs proximal steps at each inner iteration which allows maintaining feasibility or the structure induced by φ. We now list some of our core contributions:

- We derive finite-time complexity bounds for norm-PRR in the nonconvex, nonsmooth setting. In contrast to previous related works (Davis and Drusvyatskiy, 2019; Mishchenko et al., 2022), our convergence results are formulated in terms of the subdifferential $\partial\psi$ of the original objective function ψ, rather than using the natural stationarity measure $\mathcal{G}_\lambda$. Specifically, under standard assumptions, we establish

$$\begin{aligned} &\min_{k=1,\ldots,T} \mathrm{dist}(0, \partial\psi(w^k))^2 = \mathcal{O}(T^{-2/3}) \quad \text{and} \\ &\min_{k=1,\ldots,T} \mathbb{E}[\mathrm{dist}(0, \partial\psi(w^k))^2] = \mathcal{O}(n^{-1/3}\, T^{-2/3}), \end{aligned} \tag{5}$$

2. The explicit bound in (Mishchenko et al., 2022, Theorem 3) is $\min_{k=1,\ldots,T} \mathbb{E}[\|\mathcal{G}_{n\gamma}(w^k)\|^2] = \mathcal{O}((n\gamma T)^{-1} + n^2\gamma^2 + n\gamma^2)$ with step sizes $\alpha_k = \gamma \lesssim \frac{1}{Ln}$. By replacing $\lambda = n\gamma$, we obtain $\min_{k=1,\ldots,T} \mathbb{E}[\|\mathcal{G}_\lambda(w^k)\|^2] = \mathcal{O}(T^{-2/3} + n^{-1}T^{-2/3})$ for the optimal choice $\lambda \sim T^{-1/3}$.

Alg.	Convergence: Nonconvex Setting			Reference
	complexity *	global conv.	local rate (KL)	
RR (smooth, $\varphi \equiv 0$)	$\frac{\mathsf{L}\sqrt{n}}{\varepsilon^2}\max\{\sqrt{n}, \frac{\sqrt{\mathsf{A}}+\mathsf{B}}{\varepsilon}\}$ (a)	–	–	Mishchenko et al. (2020)
	$\frac{\mathsf{L}\sqrt{n}}{\varepsilon^2}\cdot\frac{\mathsf{B}\sqrt{\mathsf{A}/n+1}}{\varepsilon}$ (b)	$\|\nabla f(w^k)\| \to 0$ (c)	–	Nguyen et al. (2021)
	–	$\|\nabla f(w^k)\| \to 0$	$\|w^k - w^*\| = \mathcal{O}(\frac{1}{k})$	Li et al. (2023)
PSGD	$\frac{\mathsf{L}}{\varepsilon^2}\max\{1, \frac{\mathsf{B}^2}{\varepsilon^2}\}$ (d)	–	✗	Davis and Drusvyatskiy (2019)
	–	$\|\mathcal{G}_\lambda(w^k)\| \to 0$ (e)		Duchi and Ruan (2018); Majewski et al. (2018); Davis et al. (2020); Li and Milzarek (2022)
e-PRR	$\frac{\mathsf{L}\sqrt{n}}{\varepsilon^2}\max\{\sqrt{n}, \frac{\mathsf{B}}{\varepsilon}, \frac{\sqrt{n}\zeta}{\sqrt{\mathsf{L}}\varepsilon}\}$ (f)	✗	✗	Mishchenko et al. (2022)
norm-PRR	$\frac{\mathsf{L}\sqrt{n}}{\varepsilon^2}\max\{\sqrt{n}, \frac{\sqrt{\mathsf{L}}}{\varepsilon}\}$ (g)	$\mathrm{dist}(0, \partial\psi(w^k)) \to 0$	$\|w^k - w^*\| = \mathcal{O}(\frac{1}{k})$ $\|\psi(w^k) - \psi^*\| = \mathcal{O}(\frac{1}{k^2})$	this work

Table 1: Comparison of convergence guarantees for RR and proximal-type methods.

* This column shows the number of individual gradient evaluations $K = T$ (PSGD) and $K = nT$ (RR, e-PRR, norm-PRR) required to reach an ε-accurate solution satisfying $\min_{k=1,\ldots,T} \mathbb{E}[\|\mathcal{G}_\lambda(w^k)\|] \le \varepsilon$ where λ is a (step size) parameter. Note that both PSGD and norm-PRR execute a proximal step after each (stochastic) gradient step, whereas e-PRR only performs a proximal step after each epoch. Hence, e-PRR generally has a better complexity in terms of proximity operator evaluations.

(a) Based on the variance condition $\frac{1}{n}\sum_{i=1}^n \|\nabla f(w,i) - \nabla f(w)\|^2 \le 2\mathsf{A}[f(w) - f_{\mathrm{lb}}] + \mathsf{B}^2$.

(b) Based on the assumption $\frac{1}{n}\sum_{i=1}^n \|\nabla f(w,i) - \nabla f(w)\|^2 \le \mathsf{A}\|\nabla f(w)\|^2 + \mathsf{B}^2$ and if $\sqrt{n} \lesssim \frac{\mathsf{B}}{\varepsilon}$.

(c) Nguyen et al. (2021) provide the weaker result $\liminf_{k\to\infty} \|\nabla f(w^k)\| = 0$.

(d) For PSGD, L denotes the Lipschitz constant of ∇f; For the other RR-based methods, L denotes the common Lipschitz constant of all $\nabla f(\cdot, i)$. Based on the condition $\frac{1}{n}\sum_{i=1}^n \|\nabla f(w,i) - \nabla f(w)\|^2 \le \mathsf{B}^2$ and $\lambda \lesssim \frac{1}{\rho}$, where ρ is the weak convexity parameter of φ.

(e) The results in (Duchi and Ruan, 2018; Majewski et al., 2018; Davis et al., 2020) require almost sure boundedness of $\{w^k\}_k$; Alternatively, the analysis in (Li and Milzarek, 2022) uses Lipschitz continuity of φ.

(f) The result in (Mishchenko et al., 2022) holds for the choice $\lambda = \frac{1}{\mathsf{L}}\min\{\frac{1}{5}, \frac{\varepsilon}{n^{-1/2}\mathsf{B}+\mathsf{L}^{-1/2}\zeta}\}$ if $\|\nabla f(w)\|^2 \le \|\mathcal{G}_\lambda(w)\|^2 + \zeta^2$ for all $w \in \mathrm{dom}(\varphi)$ and if $\frac{1}{n}\sum_{i=1}^n \|\nabla f(w,i) - \nabla f(w)\|^2 \le \mathsf{B}^2$ for all $w \in \mathbb{R}^d$.

(g) Our work does not make any explicit bounded variance assumptions. Instead, in Lemma 6, it is shown that when $\alpha_k = \frac{\eta_k}{n}$, $\sum_{i=1}^\infty \eta_k^3 \lesssim \frac{1}{\mathsf{L}^3}$, $\lambda \lesssim \frac{1}{\mathsf{L}}$, we have $\frac{1}{n}\sum_{i=1}^n \|\nabla f(w^k,i) - \nabla f(w^k)\|^2 \le \mathsf{B}^2$ with $\mathsf{B} \lesssim \sqrt{\mathsf{L}}$; see also Theorem 7 and Remark 10. Note that we provide complexity bounds in terms of $\mathrm{dist}(0, \partial\psi(w^k))^2$, cf. Corollary 8. By (10), these bounds can be readily expressed using the natural residual $\|\mathcal{G}_\lambda(w^k)\|^2$ if $\lambda \lesssim \frac{1}{\rho}$.

where T counts the total number of iterations[3]. The first result in (5) is a worst-case deterministic complexity bound that is applicable to any shuffling scheme. The second

3. The measures $\mathcal{G}_\lambda$ and $\mathrm{dist}(0, \partial\psi(\cdot))$ are closely connected, i.e., we have $(1-\lambda\rho)\|\mathcal{G}_\lambda(w^k)\| \le \mathrm{dist}(0, \partial\psi(w^k))$ for all k and $\lambda < \rho^{-1}$, cf. (10). Hence, our results can also be naturally expressed in terms of $\mathcal{G}_\lambda$.

result shown in (5) holds for a uniform without-replacement sampling strategy. Both bounds match those of RR in the nonconvex smooth setting, (Mishchenko et al., 2020; Nguyen et al., 2021). Compared to the complexity (4) for e-PRR[4], the in-expectation bound for norm-PRR has a better dependence on n. We further prove that the sequence $\{\psi(w^k)\}_k$ can converge linearly to an optimal function value $\psi(w^*)$ under the Polyak-Łojasiewicz condition and in the interpolation setting $\nabla f(w^*, 1) = \cdots = \nabla f(w^*, n)$.

- We provide an in-depth asymptotic analysis of norm-PRR. For diminishing step sizes $\{\alpha_k\}_k$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^3 < \infty$, we derive $\operatorname{dist}(0, \partial\psi(w^k)) \to 0$ and $\psi(w^k) \to \bar{\psi}$. To our knowledge, this is the first asymptotic convergence guarantee for a proximal-type RR method. Moreover, when ψ satisfies the KL inequality, we establish convergence of the whole sequence of iterates, i.e., $w^k \to w^*$, where w^* is a stationary point of ψ. We further quantify the asymptotic rate of convergence of norm-PRR when using polynomial step sizes $\alpha_k \sim k^{-\gamma}$, $\gamma \in (\frac{1}{2}, 1]$. The derived rates depend on the local geometry of ψ around the stationary point w^* which is captured by the KL exponent $\theta \in [0, 1)$. In the case $\gamma = 1$, $\theta \in [0, \frac{1}{2}]$, we obtain

$$\|w^k - w^*\| = \mathcal{O}(k^{-1}), \quad \operatorname{dist}(0, \partial\psi(w^k)) = \mathcal{O}(k^{-1}), \quad |\psi(w^k) - \psi(w^*)| = \mathcal{O}(k^{-2}).$$

  These rates match existing results in the smooth or strongly convex setting, cf. (Gürbüzbalaban et al., 2021; Li et al., 2023). Comparable asymptotic rates for PSGD and e-PRR do not seem to be available in the general nonconvex case.

- Finally, we present experiments that corroborate our theoretical findings on feasibility and potential linear convergence of norm-PRR and we conduct numerical comparisons on a nonconvex binary classification and deep learning image classification problem.

An additional overview and discussion of the obtained results is provided in Table 1.

1.3 Basic Notations

By $\langle \cdot, \cdot \rangle$ and $\|\cdot\| := \|\cdot\|_2$, we denote the standard Euclidean inner product and norm. For $n \in \mathbb{N}$, we define $[n] := \{1, 2, \ldots, n\}$. By convention, we set $\sum_{i=k}^{k-1} \eta_i = 0$ for any $\{\eta_k\}_k$.

2. Preliminaries, the Full Algorithm, and Preparatory Lemmas

We now present technical preliminaries, the main algorithm, and first preparatory results.

2.1 Basic Nonsmooth Concepts and First-Order Optimality

We first recall several useful concepts from nonsmooth and variational analysis. For a function $h : \mathbb{R}^d \to (-\infty, \infty]$, the Fréchet (or regular) subdifferential of h at x is given by

$$\partial h(x) := \{g \in \mathbb{R}^d : h(y) \geq h(x) + \langle g, y - x \rangle + o(\|y - x\|) \text{ as } y \to x\},$$

4. The bounds (4) and (5) capture the complexity of e-PRR and norm-PRR in terms of gradient evaluations. By design, e-PRR has a better per-iteration dependence on proximal evaluations compared to norm-PRR.

see, e.g., (Rockafellar and Wets, 1998, Chapter 8). If h is convex, then the Fréchet subdifferential coincides with the standard (convex) subdifferential. The mapping h is said to be ρ-weakly convex if $h + \frac{\rho}{2}\|\cdot\|^2$, $\rho > 0$, is convex. The ρ-weak convexity of h is equivalent to

$$h(y) \geq h(x) + \langle s, y - x\rangle - \frac{\rho}{2}\|x - y\|^2, \quad \forall\, x, y, \ \ \forall\, s \in \partial h(x), \tag{6}$$

see, e.g., (Vial, 1983; Davis and Drusvyatskiy, 2019).

The first-order necessary optimality condition for the composite problem (1) is given by

$$0 \in \partial\psi(w) = \nabla f(w) + \partial\varphi(w). \tag{7}$$

A point satisfying this inclusion is called a stationary point and $\mathrm{crit}(\psi) := \{w \in \mathrm{dom}(\varphi) : 0 \in \partial\psi(w)\}$ denotes the set of all stationary points of ψ. It is well-known that the condition (7) can be equivalently represented as a nonsmooth equation, (Rockafellar and Wets, 1998),

$$\mathcal{G}_\lambda(w) := \lambda^{-1}(w - \mathrm{prox}_{\lambda\varphi}(w - \lambda\nabla f(w))) = 0, \quad \lambda > 0,$$

where $\mathcal{G}_\lambda$ is the so-called *natural residual*. The stationarity measure $\mathcal{G}_\lambda$ is widely used in the analysis of proximal methods. If φ is ρ-weakly convex, then the proximity operator $\mathrm{prox}_{\lambda\varphi}(w) := \mathrm{argmin}_{y\in\mathbb{R}^d}\ \varphi(y) + \frac{1}{2\lambda}\|w - y\|^2$, $\lambda \in (0, \rho^{-1})$, is $(1 - \lambda\rho)$-cocoercive, i.e.,

$$\langle w - y, \mathrm{prox}_{\lambda\varphi}(w) - \mathrm{prox}_{\lambda\varphi}(y)\rangle \geq (1 - \lambda\rho)\|\mathrm{prox}_{\lambda\varphi}(w) - \mathrm{prox}_{\lambda\varphi}(y)\|^2, \quad \forall\, w, y, \tag{8}$$

see, e.g., (Hoheisel et al., 2020, Proposition 3.3). In particular, $\mathrm{prox}_{\lambda\varphi}$ is Lipschitz continuous with modulus $(1 - \lambda\rho)^{-1}$. In this work, we use the *normal map* (Robinson, 1992),

$$F^\lambda_{\mathrm{nor}}(z) := \nabla f(\mathrm{prox}_{\lambda\varphi}(z)) + \lambda^{-1}(z - \mathrm{prox}_{\lambda\varphi}(z)) \in \partial\psi(\mathrm{prox}_{\lambda\varphi}(z)), \quad \lambda > 0, \tag{9}$$

as an alternative stationarity measure for (1). Here, the condition $F^\lambda_{\mathrm{nor}}(z) \in \partial\psi(\mathrm{prox}_{\lambda\varphi}(z))$ follows directly from $z - \mathrm{prox}_{\lambda\varphi}(z) \in \lambda\partial\varphi(\mathrm{prox}_{\lambda\varphi}(z))$, see (Hoheisel et al., 2020, Proposition 3.1). The normal map and the natural residual are closely related via

$$(1 - \lambda\rho)\|\mathcal{G}_\lambda(\mathrm{prox}_{\lambda\varphi}(z))\| \leq \mathrm{dist}(0, \partial\psi(\mathrm{prox}_{\lambda\varphi}(z))) \leq \|F^\lambda_{\mathrm{nor}}(z)\| \quad \forall\, z, \tag{10}$$

where the first inequality can be shown by applying (8) and following the proof of (Drusvyatskiy and Lewis, 2018, Theorem 3.5). Throughout this work and motivated by (10), we will typically measure and express convergence in terms of the distance $w \mapsto \mathrm{dist}(0, \partial\psi(w))$.

2.2 Algorithm Design

From proximal gradient to normal map steps. We motivate our approach by introducing an alternative representation of the traditional proximal gradient descent (PGD) method, (Mine and Fukushima, 1981), that separates the gradient and proximal steps. Let us define the auxiliary variable $z^{k+1} = w^k - \lambda\nabla f(w^k)$. The PGD update, $w^{k+1} = \mathrm{prox}_\lambda(w^k - \lambda\nabla f(x^k))$, can then be expressed in the following form:

$$z^{k+1} = z^k - \alpha[\nabla f(w^k) + \lambda^{-1}(z^k - w^k)] \quad \text{and} \quad w^{k+1} = \mathrm{prox}_{\lambda\varphi}(z^{k+1}) \quad \text{where} \quad \alpha = \lambda. \tag{11}$$

Algorithm 1: norm-PRR: **Nor**mal **m**ap-based **p**roximal **r**andom **r**eshuffling

Input: Initial point $z^1 \in \mathbb{R}^d$, $w^1 = \mathrm{prox}_{\lambda\varphi}(z^1)$ and parameters $\{\alpha_k\}_k \subset \mathbb{R}_{++}$, $\lambda > 0$;

for $k = 1, 2, \ldots$ **do**

  Generate a permutation π^k of $[n]$. Set $z_1^k = z^k$ and $w_1^k = w^k$;

  for $i = 1, 2, \ldots, n$ **do**

    Compute $z_{i+1}^k = z_i^k - \alpha_k(\nabla f(w_i^k, \pi_i^k) + \frac{1}{\lambda}(z_i^k - w_i^k))$ and $w_{i+1}^k = \mathrm{prox}_{\lambda\varphi}(z_{i+1}^k)$;

  end

  Set $z^{k+1} = z_{n+1}^k$ and $w^{k+1} = w_{n+1}^k$;

end

This equivalent formulation naturally introduces the *normal map* $F_{\mathrm{nor}}^\lambda(z) = \nabla f(w) + \lambda^{-1}(z - w)$ where $w = \mathrm{prox}_{\lambda\varphi}(z)$. Normal maps have been extensively used in the context of classical variational inequalities for the special case where the proximity operator $\mathrm{prox}_{\lambda\varphi}$ is given as the Euclidean projection onto a closed, convex set. We refer to (Facchinei and Pang, 2003) for more detailed background.

Advantages. A remarkable feature of the normal map is its direct connection to the subdifferential of the objective function ψ. Specifically, based on (9), we have $F_{\mathrm{nor}}^\lambda(z^k) \in \partial\psi(w^k)$, i.e., the normal map $F_{\mathrm{nor}}^\lambda(z^k)$ is a special subgradient of ψ at w^k. By contrast, it holds that $\mathcal{G}_\lambda(w^k) \in \nabla f(w^k) + \partial\varphi(w^{k+1})$, i.e., the natural residual $\mathcal{G}_\lambda(w^k)$ is not necessarily a subgradient of ψ. Our aim is to leverage this new perspective and to study the behavior of the auxiliary iterates $\{z^k\}_k$ using the normal map F_{nor}^λ as the underlying stationarity measure. Indeed, by (9), the stationarity condition $\|F_{\mathrm{nor}}^\lambda(z^k)\| < \varepsilon$ immediately ensures that $\mathrm{dist}(0, \partial\psi(w^k)) < \varepsilon$. Such a direct connection does not seem to exist between the traditional natural residual $\|\mathcal{G}_\lambda(w^k)\|$ and $\mathrm{dist}(0, \partial\psi(w^k))$.

In addition, the formulation (11) facilitates the decoupling of the proximal parameter λ from the step size α. In particular, the constant step size α can be substituted with a sequence of varying step sizes $\{\alpha_k\}_k$ without necessitating adjustments of the proximal parameter λ. This methodological flexibility can be advantageous in stochastic settings where diminishing step sizes are typical to mitigate stochastic errors.

Incorporating reshuffling. We now discuss the full procedures of norm-PRR. Let $\Pi = \{\pi : \pi \text{ is a permutation of } [n]\}$ denote the set of all possible permutations of $[n]$. At each iteration k, a permutation π^k is sampled from Π. The algorithm then updates w^k to w^{k+1} through n consecutive normal map-type steps by using the stochastic gradients $\{\nabla f(\cdot, \pi_1^k), \ldots, \nabla f(\cdot, \pi_n^k)\}$ sequentially:

$$z_{i+1}^k = z_i^k - \alpha_k(\nabla f(w_i^k, \pi_i^k) + \lambda^{-1}(z_i^k - w_i^k)) \quad \text{and} \quad w_i^k = \mathrm{prox}_{\lambda\varphi}(z_i^k), \quad i = 1, \ldots, n. \tag{12}$$

Here, π_i^k represents the i-th element of the permutation π^k and the term $\nabla f(w_i^k, \pi_i^k) + \lambda^{-1}(z_i^k - w_i^k)$ approximates $F_{\mathrm{nor}}^\lambda(z_i^k)$ by replacing the true gradient $\nabla f(w_i^k)$ with the component gradient $\nabla f(w_i^k, \pi_i^k)$. In each step of norm-PRR, only one single gradient component $\nabla f(\cdot, \pi_i^k)$, $i \in [n]$ and one proximal operator is evaluated. The pseudocode of norm-PRR is shown in Algorithm 1. In the case $\varphi \equiv 0$, norm-PRR coincides with the original RR method.

Based on (12), we can express the update of z^k compactly via

$$z^{k+1} = z^k - n\alpha_k F^\lambda_{\mathrm{nor}}(z^k) + e^k, \tag{13}$$

where the error term e^k is given by

$$e^k := -\alpha_k \left[\sum\nolimits_{i=1}^n (F^\lambda_{\mathrm{nor}}(z_i^k) - F^\lambda_{\mathrm{nor}}(z^k)) + \sum\nolimits_{i=1}^n (\nabla f(w_i^k, \pi_i^k) - \nabla f(w_i^k))\right]. \tag{14}$$

We note that, in the deterministic case $n = 1$ and $e^k = 0$, norm-PRR reduces to PGD once setting $\lambda \equiv \alpha_k$. As motivated, the procedure (13) can be interpreted as a special proximal gradient-type method with errors.

2.3 Assumptions and Error Estimates

Assumption 1 (Functions & Sampling) *We consider the following basic conditions:*

- (F.1) *Each mapping $\nabla f(\cdot, i)$, $i \in [n]$, is Lipschitz continuous on $\mathrm{dom}(\varphi)$ with modulus $\mathsf{L} > 0$.*
- (F.2) *The function $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ is ρ-weakly convex, lsc., and proper.*
- (F.3) *There are $f_{\mathrm{lb}}, \varphi_{\mathrm{lb}} \in \mathbb{R}$ such that $f(w, i) \geq f_{\mathrm{lb}}$ and $\varphi(w) \geq \varphi_{\mathrm{lb}}$, for all $w \in \mathrm{dom}(\varphi)$ and $i \in [n]$. This also implies $\psi(w) \geq \psi_{\mathrm{lb}} := f_{\mathrm{lb}} + \varphi_{\mathrm{lb}}$ for all $w \in \mathrm{dom}(\varphi)$.*
- (S.1) *The permutations $\{\pi^k\}_k$ are sampled independently (for each k) and uniformly without replacement from $[n]$.*

The conditions (F.1) and (F.2) are standard in nonconvex and nonsmooth optimization, see, e.g., (Ghadimi et al., 2016; Davis and Drusvyatskiy, 2019; Yang et al., 2021). In (F.2), we only assume φ to be weakly convex. The class of weakly convex functions is rich and allows us to cover important nonconvex regularizations, including, e.g., the student-t loss (Aravkin et al., 2011), the minimax concave penalty (Zhang, 2010), and the smoothly clipped absolute deviation penalty (Fan and Li, 2001). Combining (F.1) and (F.3) and using the descent lemma, we can derive the following bound for the gradients $\nabla f(\cdot, i)$:

$$\|\nabla f(w, i)\|^2 \leq 2\mathsf{L}[f(w, i) - f_{\mathrm{lb}}], \quad \forall\, i \in [n],$$

see, e.g., (Nesterov, 2018) or (Li et al., 2023, eqn. (2.5)). Assumption (S.1) is a standard requirement on the sampling scheme used in random reshuffling methods, (Mishchenko et al., 2020, 2022; Nguyen et al., 2021).

We now establish first estimates for the error term e^k defined in (13). In particular, we provide a link between the errors $\{e^k\}_k$ defined in (14), the step sizes $\{\alpha_k\}_k$, the variance

$$\sigma_k^2 := \frac{1}{n}\sum\nolimits_{i=1}^n \|\nabla f(w^k, i) - \nabla f(w^k)\|^2, \tag{15}$$

and the normal map F^λ_{nor}. To proceed, let us also formally define the filtration $\{\mathcal{F}_k\}_k$ where $\mathcal{F}_k := \sigma(\pi^1, \ldots, \pi^k)$ is the σ-algebra generated by the permutations $\pi^1, \ldots, \pi^k$. (We may also set $\mathcal{F}_0 := \sigma(z^1)$). Then, it follows $z^{k+1}, w^{k+1}, e^k \in \mathcal{F}_k$ for all $k \geq 1$. Let us further introduce $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \,|\, \mathcal{F}_{k-1}]$.

Lemma 2 (Error Estimates) *Let the conditions* (F.1)–(F.2) *be satisfied and let* $\{w^k\}_k$ *and* $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{\rho})$, $0 < \alpha_k \leq \frac{1}{\sqrt{2\mathsf{C}n}}$, *and* $\mathsf{C} := 4[\frac{3\mathsf{L}+2\lambda^{-1}-\rho}{1-\lambda\rho}]^2$.

(a) *It holds that* $\|e^k\|^2 \leq \mathsf{C}n^4\alpha_k^4[\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + \sigma_k^2]$ *for all* $k \geq 1$.

(b) *Additionally, under* (S.1), *it follows*

$$\mathbb{E}_k[\|e^k\|^2] \leq \mathsf{C}n^4\alpha_k^4[\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + n^{-1}\sigma_k^2] \quad \forall\ k \geq 1 \quad (\textit{almost surely}).$$

The proof of Lemma 2 is presented in Appendix B.2. Next, we provide an upper bound for the variance terms $\{\sigma_k^2\}_k$.

Lemma 3 (Variance Bound) *Let* $\{w^k\}_k \subseteq \mathrm{dom}(\varphi)$ *be given and assume that* (F.1) *and* (F.3) *are satisfied. Then, it holds that*

$$\sigma_k^2 \leq 2\mathsf{L}[f(w^k) - f_{\mathrm{lb}}] \leq 2\mathsf{L}[\psi(w^k) - \psi_{\mathrm{lb}}] \quad \forall\ k.$$

Proof Using $\|a-b\|^2 = \|a\|^2 - 2\langle a, b\rangle + \|b\|^2$, we have

$$\begin{aligned}\sigma_k^2 &= \|\nabla f(w^k)\|^2 + \frac{1}{n}\sum\nolimits_{i=1}^n \|\nabla f(w^k, i)\|^2 - \frac{2}{n}\sum\nolimits_{i=1}^n \langle \nabla f(w^k, i), \nabla f(w^k)\rangle \\ &\leq \frac{1}{n}\sum\nolimits_{i=1}^n \|\nabla f(w^k, i)\|^2 \leq 2\mathsf{L}[f(w^k) - f_{\mathrm{lb}}] \leq 2\mathsf{L}[\psi(w^k) - \psi_{\mathrm{lb}}],\end{aligned}$$

where the last line is due to $\sum_{i=1}^n \|\nabla f(w, i)\|^2 \leq 2\mathsf{L}\sum_{i=1}^n [f(w, i) - f_{\mathrm{lb}}] = 2\mathsf{L}n[f(w) - f_{\mathrm{lb}}]$ and $f(w) - f_{\mathrm{lb}} \leq \psi(w) - \psi_{\mathrm{lb}}$ for all $w \in \mathrm{dom}(\varphi)$. ■

2.4 Merit Function and Approximate Descent

Descent-type properties serve as a fundamental cornerstone when establishing iteration complexity and asymptotic convergence of algorithms. In the classical random reshuffling method and in its proximal version e-PRR, descent is measured directly on the objective function f and ψ, respectively (Mishchenko et al., 2020, 2022; Nguyen et al., 2021; Li et al., 2023). By contrast and motivated by the subgradient condition $F_{\mathrm{nor}}^\lambda(z) \in \partial\psi(\mathrm{prox}_{\lambda\varphi}(z))$, we analyze descent of norm-PRR on an auxiliary merit function H_τ that is different from ψ. The merit function H_τ was initially introduced by Ouyang and Milzarek (2024).

Definition 4 (Merit Function) *Let the constants* $\tau, \lambda > 0$ *be given. The merit function* $H_\tau : \mathbb{R}^d \to \mathbb{R}$ *is defined as follows:*

$$H_\tau(z) := \psi(\mathrm{prox}_{\lambda\varphi}(z)) + \frac{\tau\lambda}{2}\|F_{\mathrm{nor}}^\lambda(z)\|^2.$$

Provided that $\lambda \in (0, \frac{1}{4\rho})$, we will typically work with the following fixed choice of τ:

$$\tau := \frac{1 - 4\lambda\rho}{2(1 - 2\lambda\rho + \lambda^2\mathsf{L}^2)}. \tag{16}$$

We now present a first preliminary descent-type property for norm-PRR using H_τ. The detailed proof of Lemma 5 can be found in Appendix B.3.

Lemma 5 *Suppose* (F.1)–(F.2) *are satisfied and let the iterates* $\{w^k\}_k$ *and* $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and* $0 < \alpha_k \le \frac{1}{10\mathsf{L}n}$. *Setting* $\tau = \frac{1-4\lambda\rho}{2(1-2\lambda\rho+\lambda^2\mathsf{L}^2)}$, *it holds that*

$$H_\tau(z^{k+1}) - H_\tau(z^k) \le -\frac{\tau n\alpha_k}{4}\left[1 + \frac{n\alpha_k}{\lambda}\right]\|F^\lambda_{\rm nor}(z^k)\|^2 - \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2 + \frac{1}{n\alpha_k}\|e^k\|^2.$$

Based on Lemmas 2, 3 and 5, we can establish approximate descent of norm-PRR.

Lemma 6 (Approximate Descent) *Let* (F.1)–(F.2) *hold and let* $\{w^k\}_k$, $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and step sizes* $\{\alpha_k\}_k$ *satisfying*

$$0 < \alpha_k \le \frac{1}{n}\cdot\max\left\{\sqrt{2\mathsf{C}},\ 10\mathsf{L},\ 4\mathsf{C}\lambda\tau^{-1}\right\}^{-1} =: \frac{\bar\alpha}{n}.$$

Here, C *and* τ *are defined in Lemma 2 and* (16). *We set* $\Delta(t) := 2\mathsf{LC}\exp(2\mathsf{LC}t)[H_\tau(z^1) - \psi_{\rm lb}]$.

(a) *The following descent-type estimate holds for all* $k \ge 1$*:*

$$H_\tau(z^{k+1}) \le H_\tau(z^k) - \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2 - \frac{\tau n\alpha_k}{4}\|F^\lambda_{\rm nor}(z^k)\|^2 + \mathsf{C}n^3\alpha_k^3\sigma_k^2.$$

Furthermore, if (F.3) *holds and we have* $\sum_{k=1}^\infty \alpha_k^3 < \infty$, *then* $\mathsf{C}\sigma_k^2 \le \Delta(n^3\sum_{i=1}^\infty \alpha_i^3)$.

(b) *Additionally, if the sampling scheme satisfies condition* (S.1), *it holds that*

$$\mathbb{E}_k[H_\tau(z^{k+1})] \le H_\tau(z^k) - \frac{\tau n\alpha_k}{4}\|F^\lambda_{\rm nor}(z^k)\|^2 + \mathsf{C}n^2\alpha_k^3\sigma_k^2 \quad \textit{(almost surely).}$$

Moreover, under (F.3) *and* $\sum_{k=1}^\infty \alpha_k^3 < \infty$, *we have* $\mathsf{C}\mathbb{E}[\sigma_k^2] \le \Delta(n^2\sum_{i=1}^\infty \alpha_i^3)$.

Proof Applying Lemma 5 and Lemma 2 (a), we obtain

$$\begin{aligned} &H_\tau(z^{k+1}) - H_\tau(z^k) + \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2 + \frac{\tau n\alpha_k}{4}\|F^\lambda_{\rm nor}(z^k)\|^2 \\ &\quad\le \left[\mathsf{C}n\alpha_k - \frac{\tau}{4\lambda}\right]n^2\alpha_k^2\|F^\lambda_{\rm nor}(z^k)\|^2 + \mathsf{C}n^3\alpha_k^3\sigma_k^2 \le \mathsf{C}n^3\alpha_k^3\sigma_k^2, \end{aligned} \tag{17}$$

where the last inequality follows from $\alpha_k \le \frac{\tau}{4\mathsf{C}\lambda n}$. Moreover, when (F.3) is satisfied, we can use Lemma 3 in (17). This yields

$$H_\tau(z^{k+1}) - H_\tau(z^k) + \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2 + \frac{\tau n\alpha_k}{4}\|F^\lambda_{\rm nor}(z^k)\|^2 \le 2\mathsf{LC}n^3\alpha_k^3[\psi(w^k) - \psi_{\rm lb}]. \tag{18}$$

Subtracting $\psi_{\rm lb}$ on both sides of (18) and noting $\psi(w^k) \le H_\tau(z^k)$, this further implies

$$H_\tau(z^{k+1}) - \psi_{\rm lb} \le (1 + 2\mathsf{LC}n^3\alpha_k^3)[H_\tau(z^k) - \psi_{\rm lb}] \quad \forall\ k \ge 1.$$

Hence, using $1 + x \le \exp(x)$, $x \ge 0$, we can infer

$$\begin{aligned} H_\tau(z^{k+1}) - \psi_{\rm lb} &\le \textstyle\prod_{i=1}^k \left(1 + 2\mathsf{LC}n^3\alpha_i^3\right)[H_\tau(z^1) - \psi_{\rm lb}] \\ &\le \exp\left(2\mathsf{LC}n^3\textstyle\sum_{i=1}^k \alpha_i^3\right)[H_\tau(z^1) - \psi_{\rm lb}] \le \exp\left(2\mathsf{LC}n^3\textstyle\sum_{i=1}^\infty \alpha_i^3\right)[H_\tau(z^1) - \psi_{\rm lb}]. \end{aligned}$$

Thus, we have $\psi(w^k) - \psi_{\mathrm{lb}} \leq H_\tau(z^k) - \psi_{\mathrm{lb}} \leq \exp(2\mathsf{LC}n^3\sum_{i=1}^\infty \alpha_i^3)[H_\tau(z^1) - \psi_{\mathrm{lb}}]$ for all $k \geq 1$ and it follows $\mathsf{C}\sigma_k^2 \leq \Delta(n^3 \sum_{i=1}^\infty \alpha_i^3)$.

We continue with the proof of part (b). Taking the conditional expectation in Lemma 5 and using Lemma 2 (b) and $\alpha_k \leq \frac{\tau}{4\mathsf{C}\lambda n}$, we obtain

$$\begin{aligned}&\mathbb{E}_k[H_\tau(z^{k+1})] - H_\tau(z^k) + \frac{\tau n\alpha_k}{4}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 \\ &\quad\leq -\frac{\tau n^2\alpha_k^2}{4\lambda}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + \frac{1}{n\alpha_k}\mathbb{E}_k[\|e^k\|^2] \leq \mathsf{C}n^2\alpha_k^3\sigma_k^2.\end{aligned}$$

Taking the total expectation and applying Lemma 3, this yields

$$\mathbb{E}[H_\tau(z^{k+1})] \leq \mathbb{E}[H_\tau(z^k)] - \frac{\tau n\alpha_k}{4}\mathbb{E}[\|F_{\mathrm{nor}}^\lambda(z^k)\|^2] + 2\mathsf{LC}n^2\alpha_k^3\mathbb{E}[H_\tau(z^k) - \psi_{\mathrm{lb}}]. \tag{19}$$

Mimicking the previous steps, we can infer $\mathbb{E}[H_\tau(z^k) - \psi_{\mathrm{lb}}] \leq \exp(2\mathsf{LC}n^2\sum_{i=1}^\infty \alpha_i^3)[H_\tau(z^1) - \psi_{\mathrm{lb}}]$ and $\mathsf{C}\mathbb{E}[\sigma_k^2] \leq \Delta(n^2 \sum_{i=1}^\infty \alpha_i^3)$ which finishes the proof. ■

3. Iteration Complexity and Global Convergence

Based on the approximate descent properties of the merit function H_τ in Lemma 6, we now establish the iteration complexity of norm-PRR. Applications to constant and polynomial step sizes are presented in Corollaries 8 and 11

Theorem 7 (Complexity Bound for norm-PRR) *Let the conditions* (F.1)–(F.3) *hold and let* $\{w^k\}_k$ *and* $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and step sizes* $\{\alpha_k\}_k$ *satisfying* $\alpha_k = \frac{\eta_k}{n}$ *and* $0 < \eta_k \leq \bar{\alpha}$. *Then, the following statements are valid:*

(a) *If* $\sum_{k=1}^\infty \eta_k^3 \leq \frac{1}{2\mathsf{LC}}$, *then, for all* $k \geq 1$, *it holds that*

$$\min_{k=1,\ldots,T} \operatorname{dist}(0, \partial\psi(w^k))^2 \leq \frac{4 + 24\mathsf{LC}\sum_{k=1}^T \eta_k^3}{\tau\sum_{k=1}^T \eta_k} \cdot [H_\tau(z^1) - \psi_{\mathrm{lb}}].$$

(b) *In addition, under the sampling condition* (S.1) *and if* $\sum_{k=1}^\infty \eta_k^3 \leq \frac{n}{2\mathsf{LC}}$, *it holds that*

$$\min_{k=1,\ldots,T} \mathbb{E}[\operatorname{dist}(0, \partial\psi(w^k))^2] \leq \frac{4n + 24\mathsf{LC}\sum_{k=1}^T \eta_k^3}{\tau n\sum_{k=1}^T \eta_k} \cdot [H_\tau(z^1) - \psi_{\mathrm{lb}}].$$

Here, the constants $\mathsf{C}, \tau, \bar{\alpha} > 0$ *are defined in Lemma 2,* (16), *and Lemma 6, respectively.*

Proof Applying Lemma 6 (a) with $\alpha_k = \frac{\eta_k}{n}$ and dropping the term $\|w^{k+1} - w^k\|^2$, we have

$$\begin{aligned}[H_\tau(z^{k+1}) - \psi_{\mathrm{lb}}] + \frac{\tau\eta_k}{4}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 &\leq [H_\tau(z^k) - \psi_{\mathrm{lb}}] + \Delta(\textstyle\sum_{i=1}^\infty \eta_i^3)\cdot\eta_k^3 \\ &\leq [H_\tau(z^k) - \psi_{\mathrm{lb}}] + 6\mathsf{LC}[H_\tau(z^1) - \psi_{\mathrm{lb}}]\eta_k^3,\end{aligned} \tag{20}$$

where the last line is due to $\Delta(\sum_{i=1}^{\infty}\eta_i^3) \le 2\mathsf{LC}\exp(1)[H_\tau(z^1)-\psi_{\rm lb}] \le 6\mathsf{LC}[H_\tau(z^1)-\psi_{\rm lb}]$. Summing (20) from $k=1$ to T, we obtain

$$\frac{\tau}{4}\sum_{k=1}^{T}\eta_k\|F_{\rm nor}^{\lambda}(z^k)\|^2 \le [H_\tau(z^1)-\psi_{\rm lb}] + 6\mathsf{LC}[H_\tau(z^1)-\psi_{\rm lb}]\sum_{k=1}^{T}\eta_k^3,$$

which further implies $\min_{k=1,\dots,T}\|F_{\rm nor}^{\lambda}(z^k)\|^2 \le \frac{4[H_\tau(z^1)-\psi_{\rm lb}]+24\mathsf{LC}[H_\tau(z^1)-\psi_{\rm lb}]\sum_{k=1}^{T}\eta_k^3}{\tau\sum_{k=1}^{T}\eta_k}$. Noticing $\mathrm{dist}(0,\partial\psi(w^k)) \le \|F_{\rm nor}^{\lambda}(z^k)\|$, this completes the proof of part (a).

To prove (b), we invoke Lemma 6 (b), take total expectation, and set $\alpha_k = \frac{\eta_k}{n}$:

$$\begin{aligned}\mathbb{E}[H_\tau(z^{k+1})-\psi_{\rm lb}] + \frac{\tau\eta_k}{4}\cdot\mathbb{E}[\|F_{\rm nor}^{\lambda}(z^k)\|^2] &\le \mathbb{E}[H_\tau(z^k)-\psi_{\rm lb}] + \Delta(n^{-1}\textstyle\sum_{i=1}^{\infty}\eta_i^3)\cdot n^{-1}\eta_k^3\\ &\le \mathbb{E}[H_\tau(z^k)-\psi_{\rm lb}] + 6\mathsf{LC}[H_\tau(z^1)-\psi_{\rm lb}]n^{-1}\eta_k^3.\end{aligned}$$

The rest of the verification is identical to part (a). ■

We now study specific complexity results for norm-PRR under constant step size schemes.

Corollary 8 *Assume* (F.1)–(F.3) *and let* $\{w^k\}_k$, $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and* $\alpha_k \equiv \frac{\alpha}{n}$ *for all* k.

(a) *If* $\alpha = \frac{\eta}{T^{1/3}}$ *with* $0 < \eta \le \min\{(2\mathsf{LC})^{-\frac{1}{3}}, \bar{\alpha}T^{\frac{1}{3}}\}$, *then we have*

$$\min_{k=1,\dots,T}\mathrm{dist}(0,\partial\psi(w^k))^2 \le 16(\tau\eta T^{2/3})^{-1}[H_\tau(z^1)-\psi_{\rm lb}] = \mathcal{O}(T^{-2/3}).$$

(b) *Moreover, if the sampling condition* (S.1) *is satisfied and if* $\alpha = \frac{\eta n^{1/3}}{T^{1/3}}$ *with* $0 < \eta \le \min\{(2\mathsf{LC})^{-\frac{1}{3}}, \bar{\alpha}n^{-\frac{1}{3}}T^{\frac{1}{3}}\}$, *it holds that*

$$\min_{k=1,\dots,T}\mathbb{E}[\mathrm{dist}(0,\partial\psi(w^k))^2] \le \frac{16[H_\tau(z^1)-\psi_{\rm lb}]}{\tau\eta n^{1/3}T^{2/3}} = \mathcal{O}(n^{-1/3}T^{-2/3}).$$

Proof We substitute $\eta_k = \eta/T^{1/3}$ for all $k \le T$ and $\eta_k = 0$ for all $k > T$ in Theorem 7 (a). Observing $\eta^3 \le \frac{1}{2\mathsf{LC}}$, this yields

$$\min_{k=1,\dots,T}\mathrm{dist}(0,\partial\psi(w^k))^2 \le \frac{4+24\eta^3\mathsf{LC}}{\tau\eta T^{2/3}}\cdot[H_\tau(z^1)-\psi_{\rm lb}] \le \frac{16[H_\tau(z^1)-\psi_{\rm lb}]}{\tau\eta T^{2/3}}.$$

Similarly, setting $\eta_k = \eta n^{1/3}/T^{1/3}$ when $k \le T$ and $\eta_k = 0$ when $k > T$ in Theorem 7 (b) allows completing the proof of part (b). ■

Remark 9 *In the smooth setting, it holds that* $\mathrm{dist}(0,\partial\psi(w^k)) = \|\nabla f(w^k)\|$ *and the results shown in Theorem 7 and Corollary 8 match the existing (nonconvex) complexity bounds presented in (Mishchenko et al., 2020, Theorem 4) and (Nguyen et al., 2021, Theorem 3 & Corollary 2). The complexity result in Corollary 8 (b) improves the best bound for* e-PRR *in (Mishchenko et al., 2022, Theorem 3), in terms of gradient evaluations, by a factor of* $n^{1/3}$. *Moreover, our results do not require a link between the natural residual* $\mathcal{G}_\lambda$ *and the gradient mapping* ∇f *as assumed in (Mishchenko et al., 2022, Assumption 2 & Theorem 3).*

Remark 10 *Let $\varepsilon > 0$ be given and let us set $\rho = 0$ (i.e., φ is convex) and $\lambda = \frac{1}{\mathsf{L}}$. We then obtain $\mathsf{C} = 4[3\mathsf{L} + 2\lambda^{-1}]^2 = 100\mathsf{L}^2$, $\tau = \frac{1}{4}$, and $\bar{\alpha} = \frac{1}{1600\mathsf{L}}$. Let us consider the constant step size $\alpha_k \equiv \alpha = \frac{1}{\mathsf{L}} \min\{\frac{1}{1600n}, \frac{1}{(200T)^{1/3}n^{2/3}}\}$. Thus, following our previous derivations, Theorem 7 (b) is applicable and we can infer*

$$\begin{aligned}\min_{k=1,\dots,T} \mathbb{E}[\operatorname{dist}(0, \partial\psi(w^k))^2] &\leq \left[\frac{1}{Tn\alpha} + 600\mathsf{L}^3 n\alpha^2\right] \cdot 16(H_\tau(z^1) - \psi_{\mathrm{lb}}) \\ &\leq \left[\mathsf{L}\max\left\{\frac{1600}{T}, \frac{200^{1/3}}{T^{2/3}n^{1/3}}\right\} + \frac{600\mathsf{L}}{(200T)^{2/3}n^{1/3}}\right] \cdot 16(H_\tau(z^1) - \psi_{\mathrm{lb}}) = \mathcal{O}(\varepsilon^2),\end{aligned}$$

provided that the total number of gradient evaluations satisfies $Tn \geq \frac{10\mathsf{L}\sqrt{n}}{\varepsilon^2}\max\{160\sqrt{n}, \frac{\sqrt{2\mathsf{L}}}{\varepsilon}\}$. This recovers the result shown in Table 1 and confirms that RR *and* norm-PRR *have similar complexities. Hence, the comparisons between the complexities of* RR *and* SGD *in (Mishchenko et al., 2020; Nguyen et al., 2021) can also be transferred to* norm-PRR *and* PSGD.

Next, we discuss the corresponding complexity results for polynomial step sizes $\alpha_k \sim k^{-\gamma}$, $\gamma \in (0, 1)$. Step sizes of this form are common and popular in stochastic optimization, see, e.g., (Robbins and Monro, 1951; Chung, 1954; Bottou et al., 2018; Nguyen et al., 2021).

Corollary 11 *Assume* (F.1)–(F.3) *and let the sequences $\{w^k\}_k$ and $\{z^k\}_k$ be generated by* norm-PRR *with $\lambda \in (0, \frac{1}{4\rho})$.*

(a) *If $\alpha_k = \frac{\alpha}{nk^\gamma}$ with $\gamma \in (\frac{1}{3}, 1)$ and $0 < \alpha \leq \min\{\bar{\alpha}, (\frac{3\gamma-1}{6\mathsf{LC}})^{\frac{1}{3}}\}$, then*

$$\min_{k=1,\dots,T} \operatorname{dist}(0, \partial\psi(w^k))^2 \leq \frac{16[H_\tau(z^1) - \psi_{\mathrm{lb}}]}{\tau\alpha(T^{1-\gamma} - 1)} = \mathcal{O}(T^{-(1-\gamma)}).$$

(b) *Under the sampling condition* (S.1) *and if $\alpha_k = \frac{\alpha}{n^{2/3}k^\gamma}$ with $\gamma \in (\frac{1}{3}, 1)$ and $0 < \alpha \leq \min\{\bar{\alpha}n^{-\frac{1}{3}}, (\frac{3\gamma-1}{6\mathsf{LC}})^{\frac{1}{3}}\}$, then it holds that*

$$\min_{k=1,\dots,T} \mathbb{E}[\operatorname{dist}(0, \partial\psi(w^k))^2] \leq \frac{16[H_\tau(z^1) - \psi_{\mathrm{lb}}]}{n^{1/3}\tau\alpha(T^{1-\gamma} - 1)} = \mathcal{O}(n^{-1/3}T^{-(1-\gamma)}).$$

As the proof is a routine application of the integral test and Theorem 7, we will omit an explicit derivation here.

The complexity bounds presented in Theorem 7 and Corollaries 8 and 11 do not directly imply convergence of the stationarity measure $\operatorname{dist}(0, \partial\psi(w^k)) \to 0$, i.e., accumulation points of $\{w^k\}_k$ are not automatically stationary points of the problem (1). This subtle technicality is caused by the presence of the min-operation, i.e., the results in Theorem 7 and Corollaries 8 and 11 do not apply to the last iterate w^T. In the following, we close this gap and establish global convergence of norm-PRR under suitable diminishing step size schemes.

Theorem 12 (Global Convergence of norm-PRR) *Let* (F.1)–(F.3) *hold and let $\{w^k\}_k$, $\{z^k\}_k$ be generated by* norm-PRR *with $\lambda \in (0, \frac{1}{4\rho})$ and step sizes $\{\alpha_k\}_k$ satisfying*

$$0 < \alpha_k \leq \frac{\bar{\alpha}}{n}, \quad \sum\nolimits_{k=1}^{\infty} \alpha_k = \infty, \quad \text{and} \quad \sum\nolimits_{k=1}^{\infty} \alpha_k^3 < \infty. \tag{21}$$

Then, $F_{\mathrm{nor}}^\lambda(z^k) \to 0$, $\operatorname{dist}(0, \partial\psi(w^k)) \to 0$, and $\psi(w^k) \to \bar{\psi} \in \mathbb{R}$ as k tends to infinity.

Theorem 12 ensures that accumulation points of the iterates $\{w^k\}_k$ are stationary points of the objective function ψ—independent of the realization of the permutations $\{\pi^k\}_k$.

Proof sketch. By the approximate descent property in Lemma 6, we first show that the merit function values $\{H_\tau(z^k)\}_k$ converge and we have $\sum_{k=1}^\infty \alpha_k \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 < \infty$. In the next step, invoking (F.1) and $\sum_{k=1}^\infty \alpha_k = \infty$, we verify $\lim_{k\to\infty} \|F^\lambda_{\mathrm{nor}}(z^k)\| \to 0$. The convergence of $\{\psi(w^k)\}_k$ then follows from the convergence of the sequences $\{H_\tau(z^k)\}_k$ and $\{\|F^\lambda_{\mathrm{nor}}(z^k)\|\}_k$. A full proof of Theorem 12 can be found in Appendix C.

4. Convergence under the Polyak-Łojasiewicz (PL) Condition

We now study convergence of norm-PRR under the well-known Polyak-Łojasiewicz (PL) condition (Polyak, 1963; Łojasiewicz, 1959, 1963; Karimi et al., 2016).

Definition 13 (PL Condition) *The function ψ is said to satisfy the PL condition if there is $\mu > 0$ such that the following inequality holds:*

$$\operatorname{dist}(0, \partial\psi(w))^2 \geq 2\mu[\psi(w) - \psi^*] \quad \textit{where} \quad \psi^* := \inf_{w\in\mathbb{R}^d} \psi(w). \tag{22}$$

The PL condition is a common tool in stochastic optimization; it provides a measure of the "strong convexity-like" behavior of the objective function without requiring convexity of ψ. Furthermore, under the PL condition, every stationary point $w^* \in \operatorname{crit}(\psi)$ is necessarily an optimal solution to the problem (1). To proceed, we introduce the set of optimal solutions and the variance at optimal points:

$$\mathcal{W} := \{w \in \mathbb{R}^d : \psi(w) = \psi^*\} \quad \text{and} \quad \sigma_*^2 := \sup_{w\in\mathcal{W}} \Big[\frac{1}{n}\sum\nolimits_{j=1}^n \|\nabla f(w, j) - \nabla f(w)\|^2\Big]. \tag{23}$$

In the following, under the PL condition and using constant step sizes $\alpha_k \sim \alpha$, we show that norm-PRR converges linearly to an $\mathcal{O}(\alpha^2\sigma_*^2)$-neighborhood of $\mathcal{W}$.

Theorem 14 *Let* (F.1)–(F.2) *hold and assume that ψ satisfies the PL condition with $\mathcal{W} \neq \emptyset$. Let $\{w^k\}_k$, $\{z^k\}_k$ be generated by* norm-PRR *with $\lambda \in (0, \frac{1}{4\rho})$ and*

$$\alpha_k \equiv \frac{\alpha}{n}, \quad \alpha \leq \min\left\{\bar\alpha,\ \nu^{-1},\ \frac{\mu\sqrt{\tau}}{2\mathsf{L}\sqrt{6\mathsf{C}}}\right\}, \quad \nu := \frac{\mu\tau}{3(1+\mu\tau\lambda)}.$$

(a) *For all $T \geq 1$, it follows $\psi(w^{T+1}) - \psi^* \leq \exp(-T\nu\alpha) \cdot \frac{H_\tau(z^1)-\psi^*}{1+\mu\tau\lambda} + \frac{6\mathsf{C}\alpha^2\sigma_*^2}{\mu\tau}$.*

(b) *In addition, under the sampling condition* (S.1)*, we have*

$$\mathbb{E}[\psi(w^{T+1}) - \psi^*] \leq \exp(-T\nu\alpha) \cdot \frac{H_\tau(z^1) - \psi^*}{1+\mu\tau\lambda} + \frac{6\mathsf{C}\alpha^2\sigma_*^2}{n\mu\tau}, \quad \forall\ T \geq 1.$$

Here, $\psi^* := \inf_w \psi(w)$, $\mathsf{C} > 0$ and $\bar\alpha$ are defined in Lemmas 2 and 6, and τ is given in (16).

Remark 15 *It is possible to express the convergence results shown in Theorem 14 in terms of the distance,* $\operatorname{dist}(w, \mathcal{W})$*, to the set* $\mathcal{W}$*. Indeed, by (Bolte et al., 2017, Theorem 5 and 27) or (Karimi et al., 2016, Theorem 2), the PL condition* (22) *implies the following error bound or quadratic growth condition:*

$$\operatorname{dist}(w, \mathcal{W})^2 \leq 2\mu^{-1}[\psi(w) - \psi^*] \quad \forall\, w \in \mathbb{R}^d. \tag{24}$$

If $\mathcal{W} = \{w^*\}$ *is a singleton (which is the case, e.g., if* ψ *is strongly convex), then Theorem 14 ensures linear convergence of the iterates* $\{w^k\}_k$ *to an* $\mathcal{O}(\alpha^2\sigma_*^2)$*-neighborhood of the optimal solution* w^**. Hence, in the interpolation setting* $\nabla f(w^*, 1) = \cdots = \nabla f(w^*, n)$*, the iterates* $\{w^k\}_k$ *generated by* norm-PRR *converge linearly to* w^**.*

Remark 16 (Comparison with e-PRR) *As shown in (Mishchenko et al., 2022, Theorem 2), applying* e-PRR *with constant step size in the strongly convex case yields linear convergence to an* $\mathcal{O}(\alpha^2\sigma_{rad}^2)$*-neighborhood of the optimal solution* w^**. Here, the shuffling radius* σ_{rad}^2 *is bounded by* $\mathcal{O}(\sigma_*^2) + \mathcal{O}(\|\nabla f(w^*)\|^2)$*. Hence, even if* $f(\cdot, 1) = \cdots = f(\cdot, n)$ *(indicating* $\sigma_*^2 = 0$*), linear convergence of* e-PRR *can not be guaranteed since* $\nabla f(w^*)$ *generally does not vanish in composite problems* (1)*. We illustrate this effect numerically in Section 6.2.*

Proof of Theorem 14 Recalling $F_{\mathrm{nor}}^{\lambda}(z^k) \in \partial\psi(w^k)$, the PL condition implies $\|F_{\mathrm{nor}}^{\lambda}(z^k)\|^2 \geq 2\mu(\psi(w^k) - \psi^*)$. In addition, using the definition of the merit function, $H_\tau(z) = \psi(\operatorname{prox}_{\lambda\varphi}(z)) + \frac{\tau\lambda}{2}\|F_{\mathrm{nor}}^{\lambda}(z)\|^2$, we can infer

$$\psi(w^k) - \psi^* \leq \frac{1}{1+\tau\lambda\mu}[H_\tau(z^k) - \psi^*] \quad \text{and} \quad \|F_{\mathrm{nor}}^{\lambda}(z^k)\|^2 \geq \frac{2\mu}{1+\mu\tau\lambda}[H_\tau(z^k) - \psi^*]. \tag{25}$$

Thus, applying Lemma 6 (a) with (25) and subtracting ψ^*, it follows

$$H_\tau(z^{k+1}) - \psi^* \leq \left[1 - \frac{\mu\tau n\alpha_k}{2(1+\mu\tau\lambda)}\right][H_\tau(z^k) - \psi^*] + \mathsf{C}n^3\alpha_k^3\sigma_k^2. \tag{26}$$

We now bound the variance $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f(w^k, i) - \nabla f(w^k)\|^2$ in terms of σ_*^2, cf. (23). Setting $w_k^* := \operatorname{proj}_{\mathcal{W}}(w^k)$ and using Young's inequality, (F.1), (24), and (25), we have

$$\begin{aligned}
\sigma_k^2 &= \frac{1}{n}\sum\nolimits_{i=1}^n \|\nabla f(w^k, i) - \nabla f(w_k^*, i) + \nabla f(w_k^*, i) - \nabla f(w_k^*) + \nabla f(w_k^*) - \nabla f(w^k)\|^2 \\
&= \frac{1}{n}\sum\nolimits_{i=1}^n \Big[\|\nabla f(w^k, i) - \nabla f(w_k^*, i)\|^2 + 2\langle \nabla f(w^k, i) - \nabla f(w_k^*, i), \nabla f(w_k^*, i) - \nabla f(w_k^*)\rangle \\
&\qquad + 2\langle \nabla f(w^k, i) - \nabla f(w_k^*, i), \nabla f(w_k^*) - \nabla f(w^k)\rangle + \|\nabla f(w_k^*, i) - \nabla f(w_k^*)\|^2 \\
&\qquad + 2\langle \nabla f(w_k^*, i) - \nabla f(w_k^*), \nabla f(w_k^*) - \nabla f(w^k)\rangle + \|\nabla f(w_k^*) - \nabla f(w^k)\|^2\Big] \\
&\leq 2\mathsf{L}^2\|w^k - w_k^*\|^2 + 2\sigma_*^2 \leq 4\mathsf{L}^2\mu^{-1}[\psi(w^k) - \psi^*] + 2\sigma_*^2 \leq \tfrac{4\mathsf{L}^2\mu^{-1}}{1+\tau\lambda\mu}[H_\tau(z^k) - \psi^*] + 2\sigma_*^2.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
H_\tau(z^{k+1}) - \psi^* &\leq \left[1 - \tfrac{\mu\tau n\alpha_k}{2(1+\mu\tau\lambda)} + \tfrac{4\mathsf{C}\mathsf{L}^2n^3\alpha_k^3}{\mu(1+\mu\tau\lambda)}\right][H_\tau(z^k) - \psi^*] + 2\mathsf{C}n^3\alpha_k^3\sigma_*^2 \\
&\leq \left[1 - \tfrac{\mu\tau n\alpha_k}{3(1+\mu\tau\lambda)}\right][H_\tau(z^k) - \psi^*] + 2\mathsf{C}n^3\alpha_k^3\sigma_*^2,
\end{aligned} \tag{27}$$

where we applied $\alpha_k^2 \leq \frac{\mu^2\tau}{24\mathsf{C}\mathsf{L}^2n^2}$. Fixing $\alpha_k = \frac{\alpha}{n}$ and unfolding the recursion (27), it follows

$$\begin{aligned} H_\tau(z^{T+1}) - \psi^* &\leq [1-\nu\alpha]^T(H_\tau(z^1) - \psi^*) + 2\mathsf{C}\alpha^3\sigma_*^2 \cdot \sum\nolimits_{k=0}^{T}[1-\nu\alpha]^k \\ &\leq \exp(-T\nu\alpha)(H_\tau(z^1) - \psi^*) + 2\mathsf{C}\nu^{-1}\alpha^2\sigma_*^2, \end{aligned} \tag{28}$$

where we used $(1-x)^T = \exp(T\log(1-x)) \leq \exp(-Tx)$ and $\sum_{k=0}^{T}(1-x)^k = \frac{1-(1-x)^{T+1}}{x} \leq \frac{1}{x}$ in the last line. Thus, invoking (25), we can infer

$$\psi(w^{T+1}) - \psi^* \leq \exp(-T\nu\alpha) \cdot \frac{H_\tau(z^1) - \psi^*}{1+\mu\tau\lambda} + \frac{6\mathsf{C}\alpha^2\sigma_*^2}{\mu\tau}.$$

Part (b) can be shown in a similar way using Lemma 6 (b). ■

5. Convergence under the Kurdyka-Łojasiewicz (KL) Condition

The PL condition is generally restrictive and the verification of (22) is often not possible in many practical applications. To overcome these limitations, we now study the asymptotic behavior of norm-PRR under a weaker Kurdyka-Łojasiewicz (KL) setting.

5.1 KL Inequality and Accumulation Points

We first present the definition of the KL inequality for nonsmooth functions; see, e.g., (Attouch et al., 2010; Attouch and Bolte, 2009; Attouch et al., 2013; Bolte et al., 2014).

Definition 17 (KL Property) *The function $\psi : \mathbb{R}^d \to (-\infty, \infty]$ is said to have the KL property at a point $\bar{w} \in \mathrm{dom}(\psi)$ if there exist $c > 0$, $\eta \in (0, \infty]$, and a neighborhood U of $\bar{w}$ such that for all $w \in U \cap \{w : 0 < |\psi(w) - \psi(\bar{w})| < \eta\}$, the KL inequality*[5]*,*

$$|\psi(w) - \psi(\bar{w})|^\theta \leq c \cdot \mathrm{dist}(0, \partial\psi(w)), \tag{29}$$

holds. Here, $\theta \in [0, 1)$ is the KL exponent of ψ at $\bar{w}$.

The KL inequality is satisfied for semialgebraic, subanalytic, and log-exp functions, underlining its broad generality; see (Kurdyka, 1998; Bolte et al., 2006a,b, 2007). Particular applications include, e.g., least-squares, logistic and Poisson regression (Li and Pong, 2018), deep learning loss models (Davis et al., 2020; Dereich and Kassing, 2024), and principal component analysis (Liu et al., 2019). Intuitively, the KL inequality implies that ψ can be locally reparameterized as a sharp function near its critical points. It provides a quantitative measure of how quickly the function decreases as it approaches a stationary point which is controlled by the constants $c > 0$ and $\theta \in [0, 1)$. In stark contrast to the PL condition, the KL inequality in Definition 17 is a *local property*. It does not necessitate $\bar{w}\,(\in \mathrm{crit}(\psi))$ to be a global solution of (1). In Table 2, we list several popular and exemplary optimization models together with their corresponding worst-case KL exponent θ.

In (Ouyang and Milzarek, 2024, Lemma 5.3), it was shown that the KL property can be transferred from the objective function ψ to the merit function H_τ. We restate this result in Lemma 18.

5. The specific KL inequality, we use here, is also referred to the (local) Łojasiewicz inequality, see (Bochnak et al., 1998; Hà and Phạm, 2017).

Optimization model	KL exponent	Reference
ℓ_1-regularized least-squares	$\frac{1}{2}$	(Bolte et al., 2017, Lemma 10)
ℓ_1-regularized logistic regression	$\frac{1}{2}$	(Li and Pong, 2018, Remark 5.1)
Quadratic optimization with orthogonality constraints	$\frac{1}{2}$	(Liu et al., 2019, Theorem 1)
Semidefinite programming	$\frac{1}{2}$	(Yu et al., 2022, Theorem 4.1)
Polynomials of degree r	$1-\frac{1}{r(3r-3)^{d-1}}$	(D'Acunto and Kurdyka, 2005, Theorem 4.2)

Table 2: Optimization models and the corresponding KL exponent.

Lemma 18 *Suppose that $\psi:\mathbb{R}^d\to(-\infty,\infty]$ satisfies the KL property at a stationary point $\bar{w}=\mathrm{prox}_{\lambda\varphi}(\bar{z})$ with exponent θ and constant c. The merit function $H_\tau:\mathbb{R}^d\to\mathbb{R}$, $\tau>0$, then satisfies the following KL-type property at $\bar{z}$ with the exponent $\tilde{\theta}:=\max\{\theta,\frac{1}{2}\}$*

$$|H_\tau(z)-H_\tau(\bar{z})|^{\tilde{\theta}}\leq\tilde{c}\cdot\|F^\lambda_{\mathrm{nor}}(z)\|,\quad\forall\, z\in V\cap\{z\in\mathbb{R}^d:|H_\tau(z)-H_\tau(\bar{z})|<\tilde{\eta}\},$$

for $\tilde{c}:=c+\max\{1,\frac{\tau\lambda}{2}\}$, some $\tilde{\eta}\in(0,\infty]$, and some neighborhood V of $\bar{z}$.

Let $\{w^k\}_k$ and $\{z^k\}_k$ be generated by norm-PRR and let us define the associated sets of accumulation points $\mathcal{A}_w$ and $\mathcal{A}_z$:

$$\begin{aligned}\mathcal{A}_w&:=\{w\in\mathbb{R}^d:\exists\text{ a subsequence }\{\ell_k\}_k\subseteq\mathbb{N}\text{ such that }w^{\ell_k}\to w\},\quad\text{and}\\ \mathcal{A}_z&:=\{z\in\mathbb{R}^d:\exists\text{ a subsequence }\{\ell_k\}_k\subseteq\mathbb{N}\text{ such that }z^{\ell_k}\to z\}.\end{aligned}\tag{30}$$

In Lemma 19, we list basic properties of the sets $\mathcal{A}_w$ and $\mathcal{A}_z$, which are direct consequences of Theorem 12. The proof is deferred to Appendix D.1.

Lemma 19 (Accumulation Points) *Let the conditions stated in Theorem 12 be satisfied and let $\{w^k\}_k$ be bounded. Then, the following statements are valid:*

(a) *The sets $\mathcal{A}_w$ and $\mathcal{A}_z$ are nonempty and compact.*

(b) *We have $\mathcal{A}_z\subseteq\{z\in\mathbb{R}^d:F^\lambda_{\mathrm{nor}}(z)=0\}$ and $\mathcal{A}_w=\{w\in\mathbb{R}^d:\exists\, z\in\mathcal{A}_z$ such that $w=\mathrm{prox}_{\lambda\varphi}(z)\}\subseteq\mathrm{crit}(\psi)$.*

(c) *The functions ψ and H_τ, $\tau>0$ are finite and constant on $\mathcal{A}_w$ and $\mathcal{A}_z$, respectively.*

5.2 Strong Convergence

Based on the KL property formulated in Definition 17, our aim is now to show that the whole sequence $\{w^k\}_k$ converges to a stationary point of ψ.

Assumption 20 (KL Conditions) *We consider the following assumptions:*

(K.1) *The sequence $\{w^k\}_k$ generated by* norm-PRR *is bounded.*

(K.2) *The KL property holds on $\mathcal{A}_w$, i.e., (29) holds for all $\bar{w}\in\mathcal{A}_w$.*

Condition (K.1) is a typical and ubiquitous prerequisite appearing in the application of the KL inequality in establishing the convergence of optimization algorithms; see, e.g.,

(Attouch and Bolte, 2009; Attouch et al., 2010, 2013; Bolte et al., 2014; Ochs et al., 2014; Bonettini et al., 2017). Moreover, (K.1) can be ensured if ψ has bounded sub-level sets or $\mathrm{dom}(\varphi)$ is compact. As mentioned, condition (K.2) is naturally satisfied for subanalytic or semialgebraic functions (Kurdyka, 1998, Theorem Ł1).

Theorem 21 (Strong Iterate Convergence) *Assume* (F.1)–(F.3) *and* (K.1)–(K.2). *Let* $\{w^k\}_k$, $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and step sizes* $0 < \alpha_k \leq \frac{\bar{\alpha}}{n}$ *satisfying*

$$\sum\nolimits_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum\nolimits_{k=1}^{\infty} \alpha_k \left[\sum\nolimits_{i=k}^{\infty} \alpha_i^3\right]^{\xi} < \infty, \quad \text{for some } \xi \in (0,1). \tag{31}$$

We then have

$$\sum\nolimits_{k=1}^{\infty} \alpha_k \cdot \mathrm{dist}(0, \partial\psi(w^k)) < \infty \quad \text{and} \quad \sum\nolimits_{k=1}^{\infty} \|w^{k+1} - w^k\| < \infty. \tag{32}$$

In addition, the whole sequence $\{w^k\}_k$ *converges to some stationary point* $w^* \in \mathrm{crit}(\psi)$.

Theorem 21 not only ensures convergence of the iterates $\{w^k\}_k$, but, invoking (32), we also have $\min_{i=1,\dots,k} \mathrm{dist}(0, \partial\psi(w^i))^2 = \mathcal{O}(1/(\sum_{i=1}^k \alpha_i)^2)$. This bound is faster compared to the global complexity $\min_{i=1,\dots,k} \mathrm{dist}(0, \partial\psi(w^i))^2 = \mathcal{O}(1/(\sum_{i=1}^k \alpha_i))$ obtained in Theorem 7.

Proof sketch. By Theorem 12, we conclude that every accumulation point of $\{w^k\}_k$ is a stationary point of ψ. Based on the approximate descent property Lemma 6 (a), the conditions in Lemma 19, and the KL assumptions (K.1)–(K.2), we can show that $\sum_{k=1}^{\infty} \alpha_k \cdot \mathrm{dist}(0, \partial\psi(w^k)) \leq \sum_{k=1}^{\infty} \alpha_k \|F_{\mathrm{nor}}^{\lambda}(z^k)\| < \infty$ and $\sum_{k=1}^{\infty} \|w^{k+1} - w^k\| < \infty$. The latter result indicates that $\{w^k\}_k$ is a Cauchy sequence and thus, $\{w^k\}_k$ converges to some stationary point of the objective function ψ. The full proof and further details are presented in Appendix D.2.

5.3 Convergence Rates under Polynomial Step Sizes

In this section, we establish convergence rates for the sequences $\{w^k\}_k$, $\{\mathrm{dist}(0, \partial\psi(w^k))\}_k$, and $\{\psi(w^k)\}_k$ under more specific step size strategies. We consider polynomial step sizes, (Robbins and Monro, 1951; Chung, 1954; Bottou et al., 2018), of the form:

$$\alpha_k = \frac{\alpha}{(\beta + k)^{\gamma}}, \quad \text{with } \alpha > 0, \quad \beta \geq 0, \quad \gamma \in \left(\frac{1}{2}, 1\right]. \tag{33}$$

Note that the first two step size conditions in Theorem 21 are satisfied when $k \geq (n\alpha/\bar{\alpha})^{\frac{1}{\gamma}} - \beta$. We now present several preparatory bounds to facilitate the derivation of the rates.

Lemma 22 *Let* $\xi \in [0, 1)$ *be given and let us consider step sizes* $\{\alpha_k\}_k$ *of the form* (33).
(a) *For all* $k \geq 1$, *we have*

$$\sum\nolimits_{j=k}^{\infty} \alpha_j^3 \leq \frac{a_\gamma}{(k+\beta)^{3\gamma-1}} \quad \text{and} \quad \alpha_k \left[\sum\nolimits_{j=k}^{\infty} \alpha_j^3\right]^{2\xi} \leq \frac{a_\gamma}{(k+\beta)^{(1+6\xi)\gamma-2\xi}}, \tag{34}$$

where $a_\gamma > 0$ *is a numerical constant depending on* α *and* γ.

(b) *Moreover, if* $\xi > \frac{1-\gamma}{3\gamma-1}$, *then for all* $k \geq 1$, *we have*

$$\sum_{t=k}^{\infty} \alpha_t \left[\sum_{j=t}^{\infty} \alpha_j^3\right]^{\xi} \leq \frac{a_\xi}{(k+\beta)^{(1+3\xi)\gamma-(1+\xi)}}, \tag{35}$$

where $a_\xi > 0$ *is a constant depending on* ξ.

Proof Using the integral test and noting that $3\gamma > 1$, we obtain

$$\sum_{j=k}^{\infty} \alpha_j^3 = \sum_{j=k}^{\infty} \frac{\alpha^3}{(j+\beta)^{3\gamma}} \leq \frac{\alpha^3}{(k+\beta)^{3\gamma}} + \alpha^3 \int_k^{\infty} \frac{1}{(x+\beta)^{3\gamma}}\, \mathrm{d}x \leq \frac{3\gamma\alpha^3}{3\gamma-1} \frac{1}{(k+\beta)^{3\gamma-1}}.$$

In addition, it follows that $\alpha_k[\sum_{j=k}^{\infty} \alpha_j^3]^{2\xi} \leq \alpha^{1+6\xi}(\frac{3\gamma}{3\gamma-1})^2 \cdot \frac{1}{(k+\beta)^{2\xi(3\gamma-1)+\gamma}}$. Noting $\alpha^{1+6\xi} \leq (1+\alpha)^7$, this completes the proof of the statement (a). Part (b) can be shown by applying the results in (a) and the integral test; we refer to (Li et al., 2023, Lemma 3.7). ■

Based on Theorem 21 and Lemma 22, we can ensure the strong-limit convergence of norm-PRR under the polynomial step size rule (33).

Corollary 23 (Strong Convergence: Polynomial Step Sizes) *Assume that the conditions* (F.1)–(F.3) *and* (K.1)–(K.2) *hold. Let the iterates* $\{w^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and polynomial step sizes* $\{\alpha_k\}_k$ *of the form* (33). *Then,* $\{w^k\}_k$ *has finite length and converges to some stationary point* $w^* \in \mathrm{crit}(\psi)$.

Proof We need to verify that $\alpha_k = \frac{\alpha}{(k+\beta)^\gamma}$ satisfies (31) in Theorem 21. Due to $\gamma \leq 1$ and $\alpha_k \to 0$, we have $\alpha_k \leq \frac{\bar{\alpha}}{n}$ (for all k sufficiently large) and $\sum_{k=1}^{\infty} \alpha_k = \infty$. Using Lemma 22 (b), it holds that $\sum_{k=1}^{\infty} \alpha_k \left(\sum_{i=k}^{\infty} \alpha_i^3\right)^{\xi} < \infty$ for all $\xi \in (\frac{1-\gamma}{3\gamma-1}, 1) \subseteq (0,1)$. Therefore, Theorem 21 is applicable and $\{w^k\}_k$ has finite length and converges to some stationary point w^* of ψ. ■

Next, we derive the convergence rates for $\{\psi(w^k)\}_k$, $\{\mathrm{dist}(0, \partial\psi(w^k))^2\}_k$ and $\{w^k\}_k$.

Theorem 24 (Convergence Rates) *Assume* (F.1)–(F.3) *and* (K.1)–(K.2). *Let* $\{w^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{4\rho})$ *and step sizes satisfying* (33). *Then,* $\{w^k\}_k$ *converges to some* $w^* \in \mathrm{crit}(\psi)$ *and for all* $k \geq 1$ *sufficiently large, it holds that*

$$\max\{\mathrm{dist}(0, \partial\psi(w^k))^2, |\psi(w^k) - \psi(w^*)|\} = \begin{cases} \mathcal{O}(k^{-(3\gamma-1)}) & \text{if } 0 \leq \theta < \frac{\gamma}{3\gamma-1}, \\ \mathcal{O}(k^{-\frac{1-\gamma}{2\theta-1}}) & \text{if } \frac{\gamma}{3\gamma-1} \leq \theta < 1, \end{cases} \quad \text{if } \gamma \in (\tfrac{1}{2}, 1)$$

and

$$\|w^k - w^*\| = \begin{cases} \mathcal{O}(k^{-(2\gamma-1)}) & \text{if } 0 \leq \theta < \frac{\gamma}{3\gamma-1}, \\ \mathcal{O}(k^{-\frac{(1-\theta)(1-\gamma)}{2\theta-1}}) & \text{if } \frac{\gamma}{3\gamma-1} \leq \theta < 1, \end{cases} \quad \text{if } \gamma \in (\tfrac{1}{2}, 1).$$

Moreover, if $\theta \in [0, \frac{1}{2}]$, $\gamma = 1$, *and* $\alpha > 16\tilde{c}^2/(\tau n)$, *it follows*

$$\max\{\mathrm{dist}(0, \partial\psi(w^k))^2, |\psi(w^k) - \psi(w^*)|\} = \mathcal{O}(k^{-2}) \quad \text{and} \quad \|w^k - w^*\| = \mathcal{O}(k^{-1}).$$

Here, $\theta \in [0,1)$ *is the KL exponent of* ψ *at* w^* *and* $\tilde{c} > 0$ *is the corresponding KL constant introduced in Lemma 18.*

Remark 25 *Theorem 24 establishes the convergence rates of the function values and the stationarity measure for* norm-PRR. *To the best of our knowledge, this is the first time that such rates have been derived for an* RR *method in the nonconvex setting and under the KL inequality. Moreover, Theorem 21 and Theorem 24 show strong convergence of the iterates* $\{w^k\}_k$*, which is the first among proximal-type* RR *methods. If* $\theta \in [0, \frac{1}{2}]$*, the rate in Theorem 24 aligns with the strongly convex smooth case (cf. Gürbüzbalaban et al. (2021)).*

Remark 26 *As shown in Corollary 11 and using polynomial step sizes, the iteration complexity of* norm-PRR *is given by* $\min_{i=1,\dots,k} \operatorname{dist}(0, \partial\psi(w^i))^2 = \mathcal{O}(k^{-(1-\gamma)})$*. When* $\gamma \in (\frac{1}{2}, 1)$*, we clearly have* $1-\gamma \leq 3\gamma - 1$ *and* $1-\gamma \leq \frac{1-\gamma}{2\theta-1}$ *(if* $\theta > \frac{\gamma}{3\gamma-1}$*). Consequently, the asymptotic rate derived in Theorem 24 is faster than the complexity bound as long as* $\gamma > \frac{1}{2}$*. Furthermore, in stark contrast to Theorem 7, Theorem 24 provides last-iterate convergence guarantees.*

5.4 Proof of Theorem 24

In this section, we present a detailed proof of Theorem 24. Upon first reading of the manuscript, the reader may safely skip to Section 6.

Proof By Corollary 23, $\{w^k\}_k$ converges to some stationary point $w^* \in \operatorname{crit}(\psi)$. Let $\theta \in [0, 1)$ and $c > 0$ denote the KL exponent and constant of ψ at w^*. Applying Theorem 12 and the definition of $F_{\mathrm{nor}}^{\lambda}$, we may infer $z^k = w^k - \lambda\nabla f(w^k) + F_{\mathrm{nor}}^{\lambda}(z^k) \to w^* - \lambda\nabla f(w^*) =: z^* \in \mathcal{A}_z$. Moreover, we notice $\alpha_k \to 0$ and $H_\tau(z^k) \to \bar{\psi} := \psi(w^*)$ (by Lemma 19) as k tends to infinity. Hence, there exists $\tilde{k} \geq 1$ such that $\alpha_k \leq \min\{1, \frac{\bar{\alpha}}{n}\}$, $|H_\tau(z^k) - \bar{\psi}| < 1$, and

$$\tilde{c}\|F_{\mathrm{nor}}^{\lambda}(z^k)\| \geq |H_\tau(z^k) - \bar{\psi}|^{\tilde{\theta}}, \quad \tilde{\theta} = \max\{\tfrac{1}{2}, \theta\}, \quad \text{for all } k \geq \tilde{k}, \tag{36}$$

where (36) follows from Lemma 18. Clearly, due to $|H_\tau(z^k) - \bar{\psi}| \leq 1$, (36) also holds for every exponent $\vartheta \geq \tilde{\theta}$. Thus, we may work with the following Łojasiewicz inequality

$$\tilde{c}\|F_{\mathrm{nor}}^{\lambda}(z^k)\| \geq |H_\tau(z^k) - \bar{\psi}|^{\vartheta}, \quad \vartheta \in [\tilde{\theta}, 1), \quad \text{for all } k \geq \tilde{k}. \tag{37}$$

In the following, we always assume $k \geq \tilde{k}$. Rearranging the terms in Lemma 6 (a), we have

$$r_k - r_{k+1} \geq \frac{\tau n \alpha_k}{4}\|F_{\mathrm{nor}}^{\lambda}(z^k)\|^2 \quad \text{where} \quad r_k := H_\tau(z^k) + u_k - \bar{\psi}, \quad u_k := \mathsf{D}\sum\nolimits_{i=k}^{\infty}\alpha_i^3, \tag{38}$$

and $\mathsf{D} := n^3\Delta(n^3\sum_{i=1}^{\infty}\alpha_i^3) < \infty$. Due to $H_\tau(z^k) \to \bar{\psi}$, $u_k \to 0$, and (38), the sequence $\{r_k\}_k$ monotonically decreases to 0 and it holds that $r_k \geq 0$.

Step 1: Rate for $\{r_k\}_k$. We first establish a rate for $\{r_k\}_k$ through which we can easily derive the rates for $\{|\psi(w^k) - \bar{\psi}|\}_k$ and $\{\|F_{\mathrm{nor}}^{\lambda}(z^k)\|^2\}_k$. Combining (37) and (38), it follows

$$\begin{aligned} r_k - r_{k+1} &\geq \frac{\tau n \alpha_k}{4\tilde{c}^2}|H_\tau(z^k) - \bar{\psi}|^{2\vartheta} = \frac{\tau n \alpha_k}{4\tilde{c}^2}(|H_\tau(z^k) - \bar{\psi}|^{2\vartheta} + u_k^{2\vartheta}) - \frac{\tau n \alpha_k}{4\tilde{c}^2}u_k^{2\vartheta} \\ &\geq \frac{\tau n \alpha_k}{8\tilde{c}^2}|H_\tau(z^k) - \bar{\psi} + u_k|^{2\vartheta} - \frac{\tau n \alpha_k}{4\tilde{c}^2}u_k^{2\vartheta} = \frac{\tau n \alpha_k}{8\tilde{c}^2}r_k^{2\vartheta} - \frac{\tau n \alpha_k}{4\tilde{c}^2}u_k^{2\vartheta}, \end{aligned}$$

where the last line is due to Minkowski's inequality, i.e., $|a|^{2\vartheta} + |b|^{2\vartheta} \geq |a+b|^{2\vartheta}/2$ for all $\vartheta \in [\frac{1}{2}, 1)$, $a, b \in \mathbb{R}$. For simplicity, we limit our discussion to the case $\beta = 0$. Substituting $\alpha_k = \frac{\alpha}{k^\gamma}$ and using Lemma 22 (a) and $\mathsf{D}^{2\vartheta} \leq \max\{1, \mathsf{D}\}^{2\vartheta} \leq \max\{1, \mathsf{D}^2\}$, we obtain

$$r_{k+1} \leq r_k - \frac{\tau n \alpha}{8\tilde{c}^2}\frac{r_k^{2\vartheta}}{k^\gamma} + \frac{\mathsf{E}}{k^{(1+6\vartheta)\gamma - 2\vartheta}} \quad \text{where} \quad \mathsf{E} := \frac{\tau n \alpha a_\gamma}{4\tilde{c}^2} \cdot \max\{1, \mathsf{D}^2\}. \tag{39}$$

In the following, we provide the convergence rates of $\{r_k\}_k$ based on different exponents ϑ.

Our derivations are based on classical convergence results for sequences of numbers shown in (Polyak, 1987, Lemma 4 and 5). For ease of exposition, we state those results in Lemma 27.

Lemma 27 *Let $\{y_k\}_k \subseteq \mathbb{R}_+$ and $b \geq 0$, $d, p, q > 0$, $s \in (0,1)$, $t > s$ be given.*

(a) *Suppose that $\{y_k\}_k$ satisfies*

$$y_{k+1} \leq \left(1 - \frac{q}{k+b}\right) y_k + \frac{d}{(k+b)^{p+1}}, \quad \forall\ k \geq 1.$$

If $q > p$, it holds that $y_k \leq \frac{d}{q-p} \cdot (k+b)^{-p} + o((k+b)^{-p})$ for all sufficiently large k.

(b) *Let the sequence $\{y_k\}_k$ be given with $y_{k+1} \leq (1 - \frac{q}{(k+b)^s})y_k + \frac{d}{(k+b)^t}$ for all $k \geq 1$. Then, it follows $y_k \leq \frac{d}{q} \cdot (k+b)^{s-t} + o((k+b)^{s-t})$.*

Let us now continue with step 1 and with the proof of Theorem 24.

Case 1: $\vartheta = \frac{1}{2}$. In this case, the estimate (39) simplifies to

$$r_{k+1} \leq \left[1 - \frac{\tau n \alpha}{8\tilde{c}^2} \frac{1}{k^\gamma}\right] r_k + \frac{\mathsf{E}}{k^{4\gamma-1}}.$$

If $\gamma < 1$, the rate is $r_k = \mathcal{O}(k^{-(3\gamma-1)})$ by Lemma 27 (b). Moreover, if $\gamma = 1$ and $\alpha > \frac{16\tilde{c}^2}{\tau n}$, then Lemma 27 (a) yields $r_k = \mathcal{O}(k^{-2})$.

Case 2: $\vartheta \in (\frac{1}{2}, 1)$, $\gamma \neq 1$. In this case, the mapping $x \mapsto h_\vartheta(x) := x^{2\vartheta}$ is convex for all $x > 0$ and we have

$$h_\vartheta(y) \geq h_\vartheta(x) + h'_\vartheta(x)(y-x) = 2\vartheta x^{2\vartheta-1} y + (1-2\vartheta) h_\vartheta(x) \quad \forall\ x, y > 0. \tag{40}$$

Our next step is to reformulate the recursion (39) into a suitable form so that Lemma 27 is applicable. To that end, in (40), we set $x = \bar{c}k^{-\sigma}$ and $y = r_k$, where $\bar{c} := (\frac{8\tilde{c}^2\sigma}{\tau\alpha n\vartheta})^{1/(2\vartheta-1)}$ and $\sigma := \min\{\frac{1-\gamma}{2\vartheta-1}, 3\gamma - 1\}$. Then, it follows $r_k^{2\vartheta} \geq \frac{16\tilde{c}^2\sigma}{\tau\alpha n} \frac{r_k}{k^{(2\vartheta-1)\sigma}} + \frac{(1-2\vartheta)\bar{c}^{2\vartheta}}{k^{2\vartheta\sigma}}$. Using this bound in (39), we obtain

$$r_{k+1} \leq \left[1 - \frac{2\sigma}{k^{\gamma+(2\vartheta-1)\sigma}}\right] r_k + \frac{\mathsf{E}}{k^{(1+6\vartheta)\gamma-2\vartheta}} + \frac{(2\vartheta-1)\tau n\alpha\bar{c}^{2\vartheta}}{8\tilde{c}^2} \frac{1}{k^{\gamma+2\vartheta\sigma}}.$$

Noticing $\gamma + 2\vartheta\sigma \leq \gamma + 2\vartheta(3\gamma-1) = (1+6\vartheta)\gamma - 2\vartheta$ (by definition of σ), there exists $\hat{c} > 0$ such that

$$r_{k+1} \leq \left[1 - \frac{2\sigma}{k^{\gamma+(2\vartheta-1)\sigma}}\right] r_k + \frac{\hat{c}}{k^{\gamma+2\vartheta\sigma}}.$$

Lemma 27 then yields $r_k = \mathcal{O}(k^{-\sigma})$. Since the parameter σ is determined by the adjusted KL exponent $\vartheta \in [\max\{\frac{1}{2}, \theta\}, 1)$, to maximize σ, we shall always choose $\vartheta = \theta$ when $\theta > \frac{1}{2}$. On the other hand, if $\theta \in [0, \frac{1}{2}]$, we set $\vartheta = \frac{1}{2}$ and the results in **Case 1** apply.

Therefore, we can express the rate of $\{r_k\}_k$ in terms of the original KL exponent θ and the step size parameter γ:

$$r_k = \mathcal{O}(k^{-R(\theta,\gamma)}) \quad \text{where} \quad R(\theta,\gamma) := \begin{cases} 3\gamma - 1 & \text{if } \theta \in [0, \frac{\gamma}{3\gamma-1}] \\ \frac{1-\gamma}{2\theta-1} & \text{if } \theta \in (\frac{\gamma}{3\gamma-1}, 1) \end{cases} \quad \text{when} \quad \gamma \in (\tfrac{1}{2}, 1),$$

and $R(\theta,\gamma) := 2$ when $\gamma = 1, \theta \in [0, \frac{1}{2}]$ and $\alpha > \frac{16\tilde{c}^2}{\tau n}$.

Step 2: Rate for $\{\|F^\lambda_{\rm nor}(z^k)\|^2\}_k$. Based on our discussion of $\{r_k\}_k$, we now compute the rate for $\{\|F^\lambda_{\rm nor}(z^k)\|^2\}_k$ using the sufficient descent property (38) and Lemma 28:

Lemma 28 *Assume* (F.1)–(F.3) *and let* $\{w^k\}_k$ *and* $\{z^k\}_k$ *be generated by* norm-PRR *with* $\lambda \in (0, \frac{1}{\rho})$. *Let* $\varsigma > 0$ *be given such that*

$$\mathsf{Q}\varsigma \exp(\mathsf{Q}\varsigma) \le \tfrac{1}{2} \quad \textit{where} \quad \mathsf{Q} := (1-\lambda\rho)^{-1}(\mathsf{L} + 2\lambda^{-1} - \rho)n \max\{1, \sqrt{\mathsf{C}(2\mathsf{L}+1)}n\}$$

and $\mathsf{C} > 0$ *is introduced in Lemma 2. Consider the following additional assumptions:*

- *There exists* $\mathsf{P} > 0$ *such that we have* $\max\{\|F^\lambda_{\rm nor}(z^k)\|^2, \psi(w^k) - \psi_{\rm lb}\} \le \mathsf{P}$ *for all* $k \ge 1$.
- *For* $k \ge 1$, *there exists* $i = i(k) \ge 1$ *such that* $\sum_{j=0}^{i-1} \alpha_{k+j} \le \varsigma$.

Then, the relation $\|F^\lambda_{\rm nor}(z^{k+i})\|^2 \ge \frac{1}{8}\|F^\lambda_{\rm nor}(z^k)\|^2 - \frac{\mathsf{P}}{4\varsigma^2}(\sum_{j=0}^{i-1} \alpha_{k+j}^2)^2$ *holds.*

A proof of the auxiliary results in Lemma 28 is presented in Appendix D.3. The first condition, $\max\{\|F^\lambda_{\rm nor}(z^k)\|^2, \psi(w^k) - \psi_{\rm lb}\} \le \mathsf{P}$, in Lemma 28 is guaranteed by the convergence of $\{\psi(w^k)\}_k$ and $\{F^\lambda_{\rm nor}(z^k)\}_k$ (cf. Theorem 12). The conditions $\alpha_k \to 0$ and $\sum_{k=0}^\infty \alpha_k = \infty$ imply that for every k sufficiently large, there exists an integer $t = t(k) \ge 1$ such that $\frac{\varsigma}{2} \le \sum_{j=0}^{t-1} \alpha_{k+j} \le \varsigma$. Hence, the requirements in Lemma 28 are satisfied for all k sufficiently large and all $i = 1, \ldots, t(k)$, i.e., we have $\|F^\lambda_{\rm nor}(z^{k+i})\|^2 \ge \frac{1}{8}\|F^\lambda_{\rm nor}(z^k)\|^2 - \frac{\mathsf{P}}{4\varsigma^2}(\sum_{j=0}^{i-1} \alpha_{k+j}^2)^2$ for all $1 \le i \le t(k)$. Noting $t = t(k)$ and summing (38) for $k, k+1, \ldots, k+t$, this yields

$$\begin{aligned} r_k \ge r_k - r_{k+t} &\ge \frac{\tau n}{4} \sum\nolimits_{i=0}^{t-1} \alpha_{k+i} \|F^\lambda_{\rm nor}(z^{k+i})\|^2 \\ &\ge \frac{\tau n}{32} \big[\sum\nolimits_{i=0}^{t-1} \alpha_{k+i}\big] \|F^\lambda_{\rm nor}(z^k)\|^2 - \frac{\mathsf{P}\tau n}{16\varsigma^2} \sum\nolimits_{i=1}^{t-1} \alpha_{k+i} \big[\sum\nolimits_{j=0}^{i-1} \alpha_{k+j}^2\big]^2. \end{aligned}$$

Since $\frac{\varsigma}{2} \le \sum_{j=0}^{t-1} \alpha_{k+j} \le \varsigma$ and $\{\alpha_k\}_k$ is monotonically decreasing, we further obtain

$$\begin{aligned} \frac{\tau n \varsigma}{64} \|F^\lambda_{\rm nor}(z^k)\|^2 \le \frac{\tau n}{32} \big[\sum\nolimits_{i=0}^{t-1} \alpha_{k+i}\big] \|F^\lambda_{\rm nor}(z^k)\|^2 &\le r_k + \frac{\mathsf{P}\tau n}{16\varsigma^2} \sum\nolimits_{i=1}^{t-1} \alpha_{k+i} \big[\sum\nolimits_{j=0}^{i-1} \alpha_{k+j}^2\big]^2 \\ &\le r_k + \frac{\mathsf{P}\tau n}{16\varsigma^2} \sum\nolimits_{i=1}^{t-1} \alpha_{k+i} \big[\alpha_k \sum\nolimits_{j=0}^{t-1} \alpha_{k+j}\big]^2 \le r_k + \frac{\mathsf{P}\tau n \varsigma}{16} \alpha_k^2 = \mathcal{O}(k^{-R(\theta,\gamma)}). \end{aligned}$$

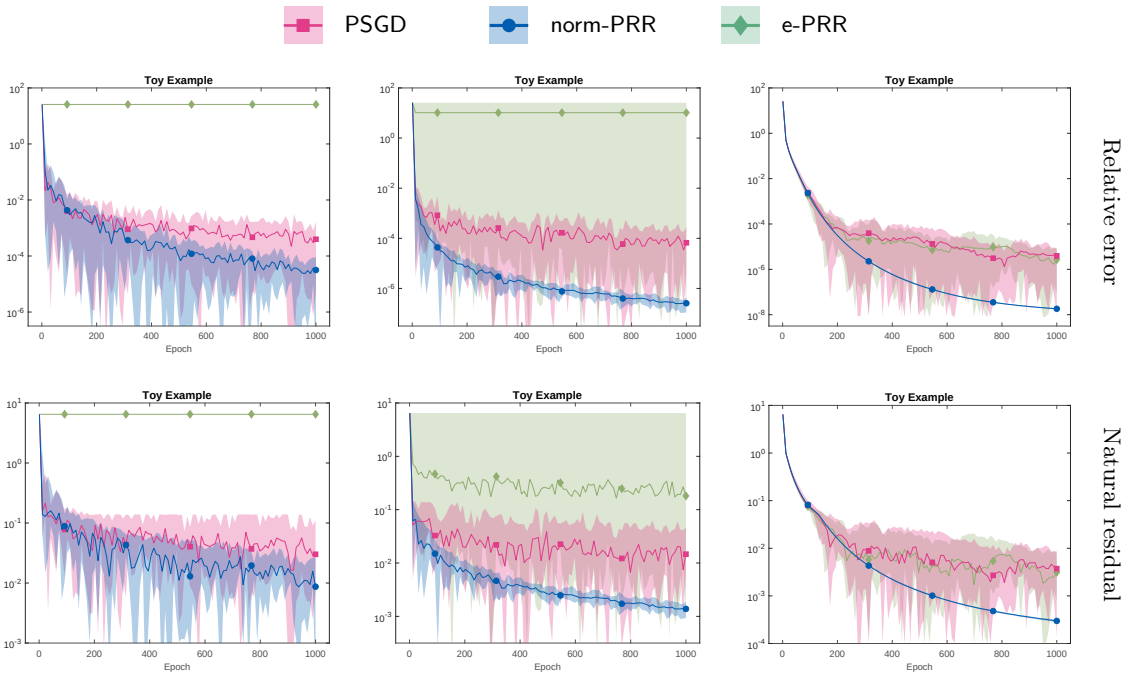
The last line is due to $\alpha_k^2 = \mathcal{O}(k^{-2\gamma})$ and $2\gamma \ge R(\theta,\gamma)$.

Step 3: Rate for $\{\psi(w^k)\}_k$. Recalling $r_k = H_\tau(z^k) + u_k - \bar{\psi} = \psi(w^k) - \bar{\psi} + \frac{\tau\lambda}{2}\|F^\lambda_{\rm nor}(z^k)\|^2 + u_k$ and invoking the triangle inequality, it follows

$$|\psi(w^k) - \bar{\psi}| = \Big|r_k - \tfrac{\tau\lambda}{2}\|F^\lambda_{\rm nor}(z^k)\|^2 - u_k\Big| \le |r_k| + \tfrac{\tau\lambda}{2}\|F^\lambda_{\rm nor}(z^k)\|^2 + u_k = \mathcal{O}(k^{-R(\theta,\gamma)}),$$

where the last equality holds due to $u_k = \mathcal{O}(\sum_{j=k}^\infty \alpha_j^3) = \mathcal{O}(k^{-3\gamma+1})$ (cf. Lemma 22 (a)) and $3\gamma - 1 \ge R(\theta,\gamma)$.

The last step of the proof, i.e., the derivation of the rate of convergence of the iterates $\{w^k\}_k$, can be found in Appendix D.4. ■



(a) $\alpha = 1$ (b) $\alpha = 0.1$ (c) $\alpha = 0.01$

Figure 1: Performance of PSGD, e-PRR, and norm-PRR on the toy example (41). We report 10 independent runs; the average performance is shown using a thicker line style.

6. Numerical Experiments

In this section, we compare norm-PRR with two prevalent proximal stochastic algorithms; the standard proximal stochastic gradient (PSGD, Duchi and Singer (2009)) and the epoch-wise proximal random reshuffling method (e-PRR, Mishchenko et al. (2022)). In the experiments, we typically evaluate the performance of the tested algorithms based on the following two criteria: (i) The *relative error* is defined as $(\psi(w^k) - \psi_{\min})/\max\{1, \psi_{\min}\}$, where $\psi_{\min}$ is the smallest function value among all generated iterations of the algorithms; (ii) the *natural residual* $\mathcal{G}_1(w) := w - \mathrm{prox}_{\varphi}(w - \nabla f(w))$. Similar to norm-PRR and e-PRR, we will keep the step size α_k fixed in each epoch when applying PSGD.

6.1 A Toy Example: Testing feasibility

We consider a one-dimensional toy example with smooth component functions $\mathbb{R} \ni w \mapsto f(w, i)$ that are not well-defined for $w \notin \mathrm{dom}(\varphi)$:

$$\min_{w\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} f(w,i) + \varphi(w) := \frac{1}{100}\sum_{i=1}^{100} \frac{\sin(\frac{i\pi}{100})w^2 + \log^2(w + \frac{i}{10})}{2} + \iota_{\mathbb{R}_+}(w), \tag{41}$$

where $\iota_{\mathbb{R}_+}$ denotes the indicator function of $\mathbb{R}_+$. In this case, the proximity operator reduces to the projection onto the half space $\mathbb{R}_+$. Notice that the function f is not well-defined for

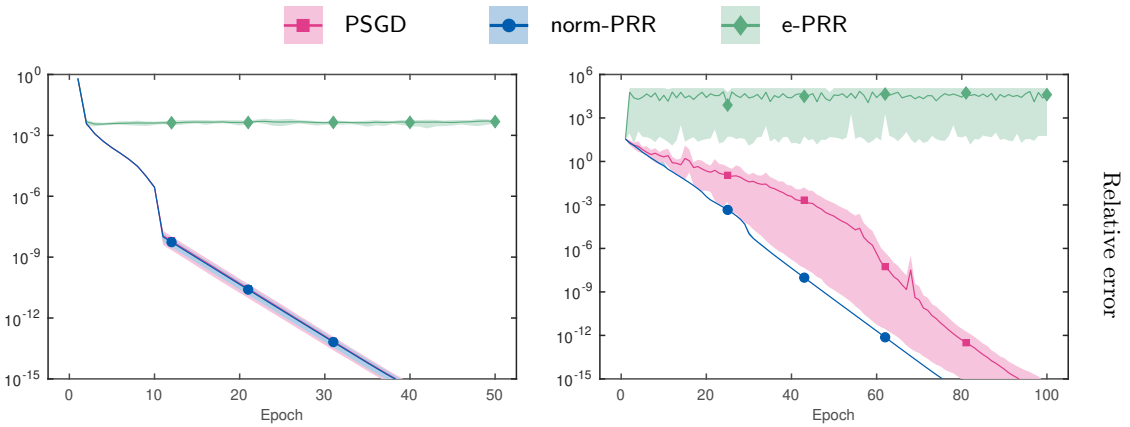


Figure 2: This plot illustrates that norm-PRR (and PSGD) can achieve linear convergence when $\sigma_*^2 = 0$, whereas e-PRR converges to a neighborhood of w^*. Left: A in (42) is generated using a uniform distribution and all methods use $\alpha_k = \frac{4}{Ln}$. Right: A follows a Student's t distribution and we apply the step size $\alpha_k = \frac{0.04}{Ln}$. We depict the relative error $(\psi(w^k) - \psi(w^*))/\max\{1, \psi(w^*)\} = \psi(w^k)$ of 10 independent runs. The average performance is shown using a thicker line style.

$w \leq -\frac{1}{10}$. Hence, once we detect iterates w^k with $w^k \leq -\frac{1}{10}$, we stop the tested algorithm and mark such run as "failed". (Due to the projection steps in PSGD and norm-PRR, failed runs can only occur when testing e-PRR).

Implementation details. We set $w^0 = 10$, $\lambda = 1$, and use diminishing step sizes of the form $\alpha_k = \alpha/k$. Here, α takes values from the set $\{1, 0.1, 0.01\}$. We conduct 10 independent run of each algorithm and depict their overall performance in Figure 1.

In Figure 1 (a), when $\alpha = 1$, the success rate of e-PRR is 0%, i.e., every run of e-PRR returns invalid values. Both PSGD and norm-PRR show larger fluctuations. In Figure 1 (b), when $\alpha = 0.1$, the success rate of e-PRR increases to 40%. norm-PRR converges faster with less oscillations compared to the other two algorithms. In Figure 1 (c), when $\alpha = 0.01$, the success rate of e-PRR is 100%; its performance is similar to PSGD. As expected, both PSGD and norm-PRR are not affected by the potential infeasibility $\mathrm{dom}(\varphi) \subseteq \mathrm{dom}(f) \neq \mathbb{R}$.

6.2 Linear Convergence and Interpolation

We now numerically illustrate the convergence results obtained in Section 4 and study convergence of norm-PRR in the interpolation setting as discussed in Remarks 15 and 16. We consider the problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f(w,i) + \varphi(w) := \frac{1}{n}\sum_{i=1}^n \left[0.5(a_i^\top w - b_i)^2 + c^\top w\right] + \iota_{\Delta^d}(w), \tag{42}$$

where $\Delta^d := \{w : w_i \in [0,1] \text{ and } \mathbb{1}^\top w = 1\}$ is the d-simplex. We choose $n = 5000$, $d = 250$, and select $A = [a_1, \ldots, a_n]^\top \in \mathbb{R}^{n \times d}$ randomly following a uniform distribution, $a_{ij} \sim \mathcal{U}[0,1]$,

or a Student's t distribution with degree of freedom 1.5. We generate $w^* \in \Delta^d$, b, and c, via

$$w_i^* = 0.2,\ \forall\ i \in \mathcal{I}, \quad w_i^* = 0,\ \forall\ i \notin \mathcal{I}, \quad b = Aw^*, \quad c_i = 0,\ \forall\ i \in \mathcal{I}, \quad c_i \sim \mathcal{U}[0,1],\ \forall\ i \notin \mathcal{I},$$

where $\mathcal{I} \subseteq [d]$ is a random index set with $|\mathcal{I}| = 5$. We have $\nabla f(w^*, 1) = \cdots = \nabla f(w^*, n) = \nabla f(w^*) = c$ and by construction, it holds that $-\langle c, y - w^* \rangle = -\sum_{i \notin \mathcal{I}} c_i y_i \leq 0$ for all $y \in \Delta^d$, i.e., $-\nabla f(w^*) \in N_{\Delta^d}(w^*)$. Hence, w^* is a solution to (42). In the tests, we generate A such that $A^\top A$ is invertible, i.e., problem (42) is strongly convex with $\sigma_*^2 = 0$. We run PSGD, norm-PRR, and e-PRR with $w^0 = e_n$, $\lambda = \frac{1}{\mathsf{L}}$, and constant step sizes $\alpha_k = \frac{4}{\mathsf{L}n}$ and $\alpha_k = \frac{0.04}{\mathsf{L}n}$ where $\mathsf{L} = \frac{1}{n}\|A\|_2$. The results are presented in Figure 2.

6.3 Nonconvex Binary Classification

Next, we consider a nonconvex binary classification problem with ℓ_1-regularization (Mason et al., 1999; Wang et al., 2017; Milzarek et al., 2019):

$$\min_{w \in \mathbb{R}^d}\ \frac{1}{n}\sum\nolimits_{i=1}^{n} f(w, i) + \varphi(w) := \frac{1}{n}\sum\nolimits_{i=1}^{n} [1 - \tanh(b_i \cdot a_i^\top w)] + \nu\|w\|_1. \tag{43}$$

Here, tanh denotes the hyperbolic tangent function and the parameter ν is set to $\nu = 0.01$. We conduct the binary classification task on real datasets $(A, b) \in \mathbb{R}^{n \times d} \times \{0, 1\}^n$. In our tests, we use the datasets `CINA`, `MNIST`, and `GISETTE`[6]. (In `MNIST`, we only keep two features).

Implementation details. For all algorithms, we use polynomial step sizes of the form $\alpha_k = \alpha/(\mathsf{L} + k)$ with $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1\}$; the Lipschitz constant is $\mathsf{L} = 0.8 \cdot \lambda_{\max}(AA^\top)/n$; the index k represents the k-th epoch. We run each algorithm for 200 epochs with $w^0 = 0$; this process is repeated 10 times for each dataset. The parameter λ is set to $\lambda = 1$. The results are reported in Figure 3.

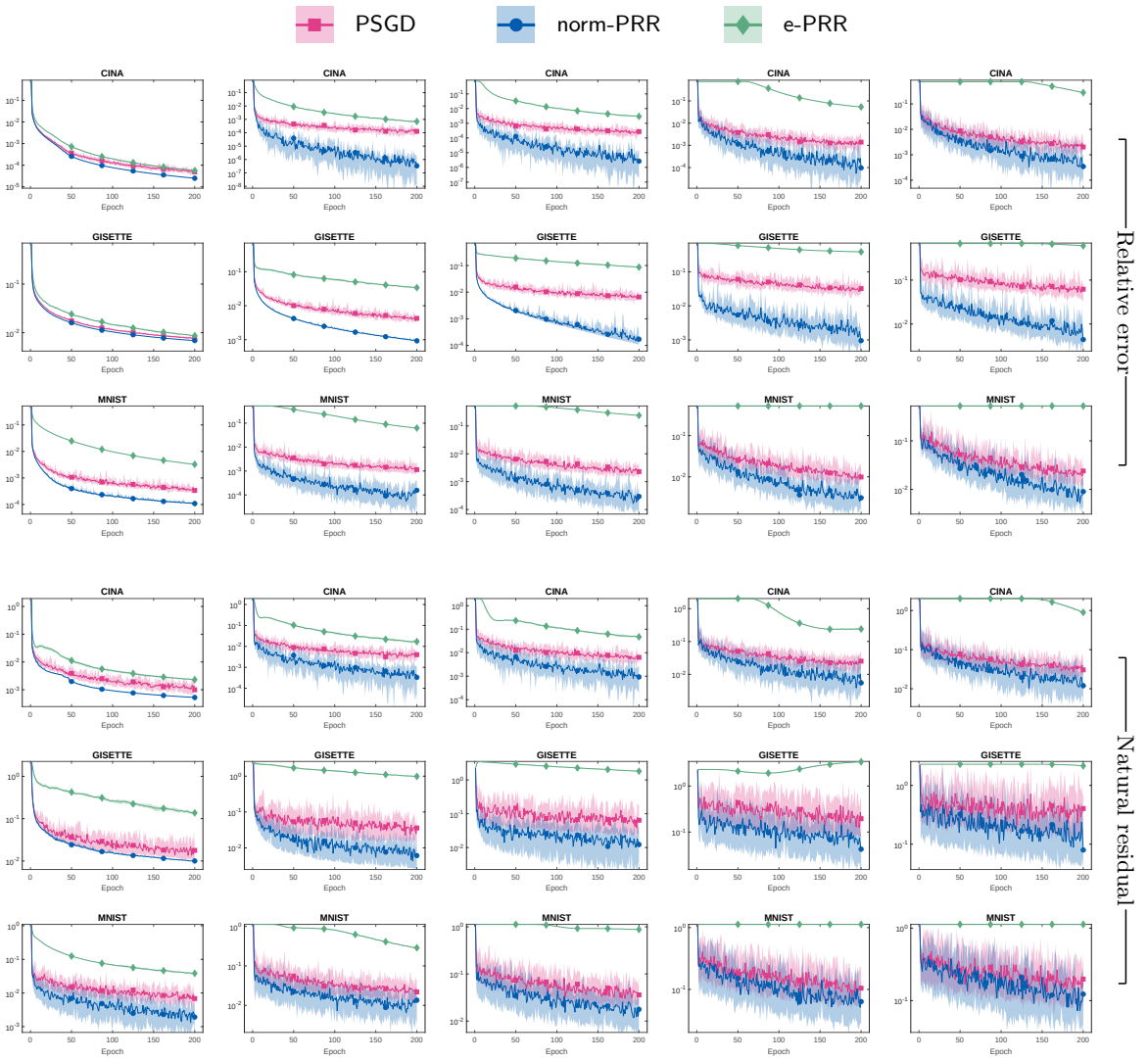
Across all datasets, norm-PRR appears to converge faster than PSGD and e-PRR and is relatively robust w.r.t. the choice of α. In the initial phases of the training, rapid convergence can be observed for all methods. However, norm-PRR typically achieves a smaller relative error and natural residual than PSGD and e-PRR. This improved performance might originate from the design of norm-PRR: it incorporates without-replacement sampling and applies the ℓ_1-proximity operator at each iteration to maintain sparsity.

Varying the parameter λ. In Figure 4, we additionally evaluate the performance of norm-PRR using different values of the hyperparameter $\lambda \in \{0.1, 1, 10\}$. As illustrated in Figure 4, larger values of λ generally lead to an improved performance when solving the problem (43) across the tested datasets. In this experiment, the overall convergence behavior norm-PRR seems to show similar trends when using different λ. Note that in the previous tests presented in Figure 3, we select $\lambda = 1$. The results in Figure 4 indicate that better performance can be potentially achieved if the parameter λ is tuned more carefully.

6.4 Deep Learning for Image Classification

In this subsection, we study multiclass image classification for the dataset `CIFAR-10` (Krizhevsky, 2009). `CIFAR-10` contains 10 classes of images and each class contains 6,000

6. Datasets are available at `http://www.causality.inf.ethz.ch/data` and `www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`.



(a) $\alpha = 0.01$ (b) $\alpha = 0.05$ (c) $\alpha = 0.1$ (d) $\alpha = 0.5$ (e) $\alpha = 1$

Figure 3: Performance of PSGD, e-PRR, and norm-PRR on the nonconvex binary classification problem (43) using different step size parameters α.

images. In this experiment, we split the dataset into $n_{\text{train}} = 50,000$ training samples and $n_{\text{test}} = 10,000$ test samples. We consider standard ResNet-18 (He et al., 2016) and VGG-16 (Simonyan and Zisserman, 2015) architectures with the cross-entropy loss function (Yang et al., 2021; Tang et al., 2021) and elastic net regularizer (Zou and Hastie, 2005):

$$\min_{w\in\mathbb{R}^d} \; -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(\mathcal{T}(w,a_i)^\top e^{b_i})}{\sum_{j=1}^{10}\exp(\mathcal{T}(w,a_i)^\top e^{j})}\right) + \nu_1\|w\|_1 + \nu_2\|w\|^2. \tag{44}$$

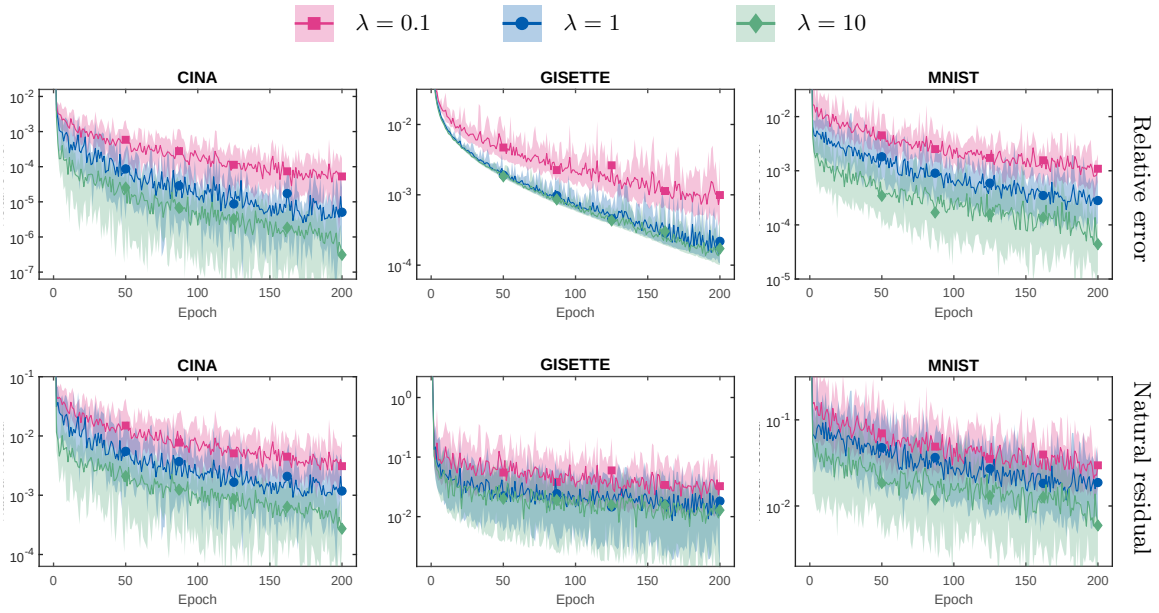


Figure 4: Performance of norm-PRR on the nonconvex binary classification problem (43) for different choices of $\lambda > 0$.

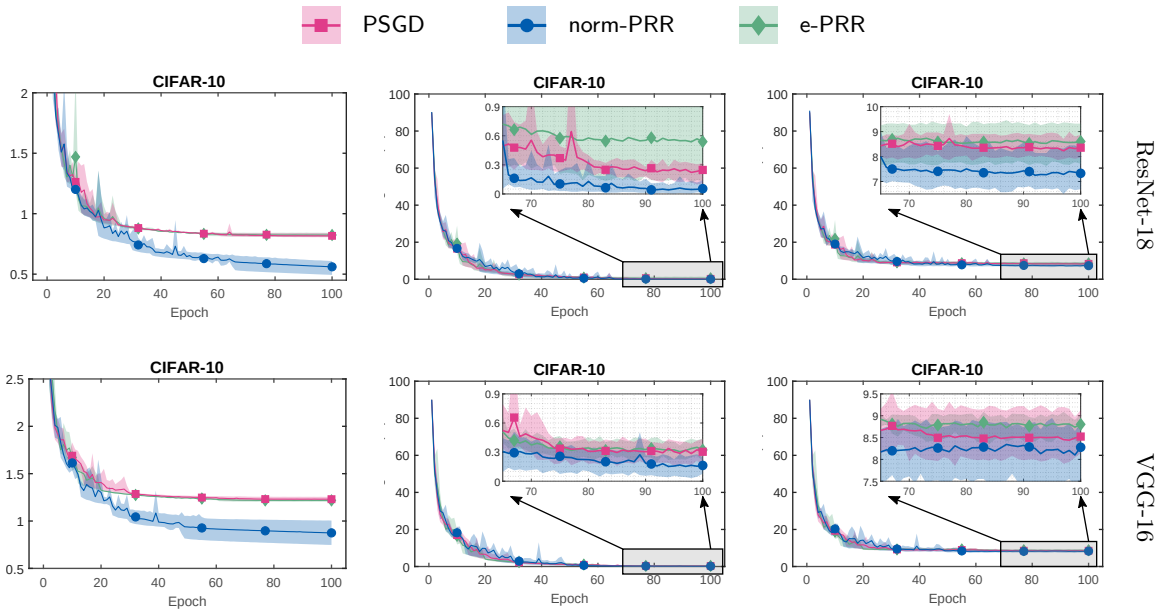
Here, the tuple (a_i, b_i) is a training sample with the color image $a_i \in \mathbb{R}^{32\times32\times3}$ and the corresponding label $b_i \in [10]$. The operator $\mathcal{T}(w, \cdot) : \mathbb{R}^{32\times32\times3} \mapsto \mathbb{R}^{10}$, which maps an image to a ten dimensional vector, represents the neural network architecture with the weights w. The vector $e^j \in \mathbb{R}^{10}$ is a unit vector with its j-th element equal to 1. To avoid overfitting while maintaining the sparsity of the parameters of the model, the elastic net regularizer is applied with parameters set to $\nu_1 = 10^{-6}$ and $\nu_2 = 10^{-4}$.

Implementation details. We use adaptive step sizes (ReduceLROnPlateau with initial step size $\alpha = 0.1$ in PyTorch (Paszke et al., 2019)) and set $\lambda = 10^{-2}$ for norm-PRR (in both architectures). We train ResNet-18 and VGG-16 for 100 epochs with batch size 128 and run each algorithm 5 times independently.

The results in Figure 5 show that norm-PRR achieves the lowest training loss. While all tested methods reach a low training error at the end, norm-PRR seems to slightly outperform PSGD and e-PRR. Similar trends can also be observed when considering the test error.

7. Conclusion

In this paper, we propose a new proximal random reshuffling method (norm-PRR) for large-scale composite problems. Our approach adopts a normal map-based perspective and uses stochastic gradient information generated through without-replacement sampling. The obtained complexity results match the existing bounds for RR in the smooth case and improve other known complexity bounds for this class of problems in terms of gradient evaluations. Under the (global) Polyak-Łojasiewicz condition and in the interpolation setting, norm-PRR



(a) Function value. (b) Training error. (c) Test error.

Figure 5: Performance of PSGD, e-PRR, and norm-PRR on the deep learning problem (44) applying the same learning rate policy. The performance is evaluated on `CIFAR-10` using the neural network architectures ResNet-18 and VGG-16.

provably converges to an optimal function value at a linear rate. In addition, accumulation points of a sequence of iterates $\{w^k\}_k$ generated by norm-PRR are shown to correspond to stationary points of the problem. Finally, under the (local) Kurdyka-Łojasiewicz inequality, we establish last-iterate convergence for norm-PRR and derive asymptotic rates of convergence. Numerical experiments illustrate that the proposed method effectively maintains feasibility and performs favorably on nonconvex, nonsmooth problems and learning tasks.

Acknowledgments

The authors would like to thank the Action Editor and two anonymous reviewers for their detailed and constructive comments, which have helped greatly to improve the quality and presentation of the manuscript.

Xiao Li was partly supported by the National Natural Science Foundation of China under grant No. 12201534 and by the Shenzhen Science and Technology Program under Grant No. RCYX20221008093033010. Andre Milzarek was partly supported by the National Natural Science Foundation of China (Foreign Young Scholar Research Fund Project) under Grant No. 12150410304, by the Shenzhen Science and Technology Program under Grant No. RCYX20210609103124047 and RCYX20221008093033010, by the Shenzhen Stability Science

Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing, and by the Internal Project Fund from the Shenzhen Research Institute of Big Data under Grant T00120230001.

Appendix A. Preparatory Results

A.1 Basic Mathematical Tools

In the following, we present several fundamental results concerning sequences $\{y_k\}_k$ of real numbers and their convergence. We start with a discrete variant of Gronwall's inequality (Borkar, 2009, Appendix B).

Lemma 29 (Gronwall's Inequality) *Let $\{a_k\}_k$, $\{y_k\}_k \subseteq \mathbb{R}_+$ and $p, q \geq 0$ be given. Assume $y_{k+1} \leq p + q\sum_{j=1}^{k} a_j y_j$ for all $k \geq 1$. Then, $y_{k+1} \leq p \cdot \exp(q\sum_{j=1}^{k} a_j)$ for all $k \geq 1$.*

In order to establish global convergence of norm-PRR, we will use the well-known supermartingale convergence theorem, see, e.g., (Bertsekas, 2016, Proposition A.31).

Theorem 30 *Let $\{y_k\}_k$, $\{p_k\}_k$, $\{q_k\}_k$, and $\{\gamma_k\}_k \subseteq \mathbb{R}_+$ be non-negative sequences. Assume that $\sum_{k=1}^{\infty} \gamma_k < \infty$, $\sum_{k=1}^{\infty} q_k < \infty$, and $y_{k+1} \leq (1+\gamma_k)y_k - p_k + q_k$ for all $k \geq 1$. Then, $\{y_k\}_k$ converges to some $y \geq 0$ and it follows $\sum_{k=1}^{\infty} p_k < \infty$.*

A.2 Sampling Without Replacement

In the following lemma, we restate results for the variance of sampling a collection of vectors from a finite set of vectors without replacement, (Mishchenko et al., 2020, Lemma 1).

Lemma 31 *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be given and let $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\sigma^2 := \frac{1}{n}\sum_{i=1}^{n} \|X_i - \bar{X}\|^2$ denote the associated average and population variance. Let $t \in [n]$ be fixed and let $X_{\pi_1}, \ldots, X_{\pi_t}$ be sampled uniformly without replacement from $\{X_1, \ldots, X_n\}$. Then, we have*

$$\mathbb{E}[\bar{X}_\pi] = \bar{X} \quad \text{and} \quad \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-t}{t(n-1)} \cdot \sigma^2, \quad \text{where} \quad \bar{X}_\pi = \frac{1}{t}\sum_{i=1}^{t} X_{\pi_i}.$$

Appendix B. Approximate Descent of the Merit Function

B.1 Basic Estimates

In this subsection, we provide two basic estimates that will be used in the derivation of the approximate descent property of H_τ stated in Lemma 6.

Lemma 32 *Let (F.1)–(F.2) hold and let $\{w^k\}_k$ and $\{z^k\}_k$ be generated by norm-PRR with $\lambda \in (0, \frac{1}{\rho})$ and step sizes $\{\alpha_k\}_k \subseteq \mathbb{R}_{++}$. Then, for all $k \geq 1$, we have*

$$\psi(w^{k+1}) - \psi(w^k) \leq \langle F_{\mathrm{nor}}^{\lambda}(z^k) + \lambda^{-1}(z^{k+1} - z^k), w^{k+1} - w^k \rangle + \left[\frac{\mathsf{L}+\rho}{2} - \frac{1}{\lambda}\right] \|w^{k+1} - w^k\|^2.$$

Proof First, invoking L-smoothness of f, (6), and $\lambda^{-1}(z^{k+1}-w^{k+1}) \in \partial\varphi(w^{k+1})$, we obtain

$$\begin{aligned}
&\psi(w^{k+1}) - \psi(w^k) \\
&\quad \le \langle \nabla f(w^k), w^{k+1}-w^k\rangle + \frac{\mathsf{L}}{2}\|w^{k+1}-w^k\|^2 + \varphi(\mathrm{prox}_{\lambda\varphi}(z^{k+1})) - \varphi(\mathrm{prox}_{\lambda\varphi}(z^k)) \\
&\quad \le \langle \nabla f(w^k), w^{k+1}-w^k\rangle + \frac{\mathsf{L}+\rho}{2}\|w^{k+1}-w^k\|^2 + \frac{1}{\lambda}\langle z^{k+1}-w^{k+1}, w^{k+1}-w^k\rangle,
\end{aligned}$$

Using $F^\lambda_{\mathrm{nor}}(z^k) = \nabla f(w^k) + \lambda^{-1}(z^k - w^k)$, this yields

$$\begin{aligned}
&\psi(w^{k+1}) - \psi(w^k) \\
&\quad \le \langle \nabla f(w^k), w^{k+1}-w^k\rangle + \frac{\mathsf{L}+\rho}{2}\|w^{k+1}-w^k\|^2 + \langle F^\lambda_{\mathrm{nor}}(z^k) - \nabla f(w^k), w^{k+1}-w^k\rangle \\
&\qquad + \lambda^{-1}\langle z^{k+1}-w^{k+1}, w^{k+1}-w^k\rangle - \lambda^{-1}\langle z^k - w^k, w^{k+1}-w^k\rangle \\
&\quad = \left[\frac{\mathsf{L}+\rho}{2} - \frac{1}{\lambda}\right]\|w^{k+1}-w^k\|^2 + \langle F^\lambda_{\mathrm{nor}}(z^k) + \lambda^{-1}(z^{k+1}-z^k), w^{k+1}-w^k\rangle,
\end{aligned}$$

which completes the proof. ■

Lemma 33 *Under the conditions stated in Lemma 32, we have*

$$\begin{aligned}
\|F^\lambda_{\mathrm{nor}}(z^{k+1})\|^2 \le{}& \left[1 - \frac{n\alpha_k}{\lambda}\right]^2 \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + [\mathsf{L}+\lambda^{-1}]^2\|w^{k+1}-w^k\|^2 + \frac{1}{\lambda^2}\|e^k\|^2 \\
&+ \frac{2}{\lambda}\left[1 - \frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), \lambda(\nabla f(w^{k+1}) - \nabla f(w^k)) - (w^{k+1}-w^k) + e^k\rangle \\
&+ 2\lambda^{-1}\langle \nabla f(w^{k+1}) - \nabla f(w^k), e^k\rangle - 2\lambda^{-2}\langle e^k, w^{k+1}-w^k\rangle.
\end{aligned}$$

Proof Applying the definition of F^λ_{nor} and (13), we can expand $F^\lambda_{\mathrm{nor}}(z^{k+1})$ as follows

$$\begin{aligned}
F^\lambda_{\mathrm{nor}}(z^{k+1}) &= \nabla f(w^{k+1}) + \lambda^{-1}(z^{k+1}-w^{k+1}) \\
&= F^\lambda_{\mathrm{nor}}(z^k) + (\nabla f(w^{k+1}) - \nabla f(w^k)) + \lambda^{-1}(z^{k+1} - z^k + w^k - w^{k+1}) \\
&= \left[1 - \frac{n\alpha_k}{\lambda}\right] F^\lambda_{\mathrm{nor}}(z^k) + (\nabla f(w^{k+1}) - \nabla f(w^k)) - \lambda^{-1}(w^{k+1}-w^k) + \lambda^{-1}e^k.
\end{aligned}$$

Using the L-smoothness of f, we can infer

$$\begin{aligned}
&\|F^\lambda_{\mathrm{nor}}(z^{k+1})\|^2 \\
&\quad = \left[1 - \frac{n\alpha_k}{\lambda}\right]^2 \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \|\nabla f(w^{k+1}) - \nabla f(w^k)\|^2 - \frac{2}{\lambda^2}\langle e^k, w^{k+1}-w^k\rangle \\
&\qquad + 2\left[1 - \frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), (\nabla f(w^{k+1}) - \nabla f(w^k)) - \lambda^{-1}(w^{k+1}-w^k) + \lambda^{-1}e^k\rangle \\
&\qquad + 2\langle \nabla f(w^{k+1}) - \nabla f(w^k), \lambda^{-1}e^k - \lambda^{-1}(w^{k+1}-w^k)\rangle + \frac{1}{\lambda^2}\|w^{k+1}-w^k\|^2 + \frac{1}{\lambda^2}\|e^k\|^2 \\
&\quad \le \left[1 - \frac{n\alpha_k}{\lambda}\right]^2 \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \left[\mathsf{L} + \frac{1}{\lambda}\right]^2 \|w^{k+1}-w^k\|^2 \\
&\qquad + 2\lambda^{-1}\langle \nabla f(w^{k+1}) - \nabla f(w^k), e^k\rangle + \lambda^{-2}\|e^k\|^2 - 2\lambda^{-2}\langle e^k, w^{k+1}-w^k\rangle
\end{aligned}$$

$$+\frac{2}{\lambda}\left[1-\frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k),\lambda(\nabla f(w^{k+1})-\nabla f(w^k))-(w^{k+1}-w^k)+e^k\rangle,$$

as desired. ■

B.2 Proof of Lemma 2

Proof We start with the proof of part (a). Using the definition of the error term e^k in (14) and the triangle inequality, we have

$$\begin{aligned}\|e^k\| &\leq \alpha_k\sum\nolimits_{i=1}^n\|F^\lambda_{\mathrm{nor}}(z_i^k)-F^\lambda_{\mathrm{nor}}(z^k)\|+\alpha_k\left\|\sum\nolimits_{i=1}^n\nabla f(w_i^k,\pi_i^k)-\nabla f(w^k,\pi_i^k)\right\|\\ &\quad+\alpha_k\left\|\sum\nolimits_{i=1}^n\nabla f(w^k,\pi_i^k)-\nabla f(w_i^k)\right\|\\ &\leq \alpha_k\sum\nolimits_{i=1}^n\left[\|F^\lambda_{\mathrm{nor}}(z_i^k)-F^\lambda_{\mathrm{nor}}(z^k)\|+\mathsf{L}\|w_i^k-w^k\|\right]+\alpha_k\left\|\sum\nolimits_{i=1}^n\nabla f(w^k)-\nabla f(w_i^k)\right\|\\ &\leq \alpha_k\sum\nolimits_{i=1}^n\|F^\lambda_{\mathrm{nor}}(z_i^k)-F^\lambda_{\mathrm{nor}}(z^k)\|+\frac{2\mathsf{L}}{1-\lambda\rho}\cdot\alpha_k\sum\nolimits_{i=1}^n\|z_i^k-z^k\|,\end{aligned}\tag{45}$$

where we applied the Lipschitz continuity of the proximity operator, $\|w_i^k-w^k\|\leq\|z_i^k-z^k\|/(1-\lambda\rho)$, in the last line, cf. (8). Similarly, it holds that

$$\begin{aligned}\|F^\lambda_{\mathrm{nor}}(z_i^k)-F^\lambda_{\mathrm{nor}}(z^k)\| &= \|\nabla f(w_i^k)-\nabla f(w^k)+\lambda^{-1}[(z_i^k-z^k)-(w_i^k-w^k)]\|\\ &\leq(\mathsf{L}+\lambda^{-1})\|w_i^k-w^k\|+\lambda^{-1}\|z_i^k-z^k\|\leq\frac{\mathsf{L}+2\lambda^{-1}-\rho}{1-\lambda\rho}\|z_i^k-z^k\|.\end{aligned}\tag{46}$$

Combining (46) and (45), we readily obtain

$$\|e^k\|\leq\mathsf{C}_r\alpha_k\sum\nolimits_{i=1}^n\|z_i^k-z^k\|\quad\text{where}\quad\mathsf{C}_r:=\frac{3\mathsf{L}+2\lambda^{-1}-\rho}{1-\lambda\rho}=\frac{\sqrt{\mathsf{C}}}{2}.\tag{47}$$

Based on (12), (13) and applying (46), it follows

$$\begin{aligned}\|z_{i+1}^k-z^k\| &= \alpha_k\left\|iF^\lambda_{\mathrm{nor}}(z^k)+\sum\nolimits_{j=1}^i\left[F^\lambda_{\mathrm{nor}}(z_j^k)-F^\lambda_{\mathrm{nor}}(z^k)+\nabla f(w_j^k,\pi_j^k)-\nabla f(w_j^k)\right]\right\|\\ &\leq i\alpha_k\|F^\lambda_{\mathrm{nor}}(z^k)\|+(\mathsf{L}+2\lambda^{-1}-\rho)(1-\lambda\rho)^{-1}\alpha_k\sum\nolimits_{j=1}^i\|z^k-z_j^k\|\\ &\quad+\alpha_k\left\|\sum\nolimits_{j=1}^i\left[\nabla f(w_j^k,\pi_j^k)-\nabla f(w_j^k)\right]\right\|\end{aligned}\tag{48}$$

for all $i=0,\ldots,n-1$. We now continue with the last term in the estimate (48). Setting $\Upsilon_i:=\|\sum_{j=1}^i[\nabla f(w^k,\pi_j^k)-\nabla f(w^k)]\|$ and using the triangle inequality, the L-Lipschitz continuity of the component gradients $\nabla f(\cdot,i)$, $i\in[n]$, and the $(1-\lambda\rho)^{-1}$-Lipschitz continuity of the proximity operator $\mathrm{prox}_{\lambda\varphi}$, it holds that

$$\begin{aligned}&\left\|\sum\nolimits_{j=1}^i\left[\nabla f(w_j^k,\pi_j^k)-\nabla f(w_j^k)\right]\right\|\\ &\quad\leq\sum\nolimits_{j=1}^i\left[\|\nabla f(w_j^k,\pi_j^k)-\nabla f(w^k,\pi_j^k)\|+\|\nabla f(w^k)-\nabla f(w_j^k)\|\right]+\Upsilon_i\\ &\quad\leq 2\mathsf{L}\sum\nolimits_{j=1}^i\|w_j^k-w^k\|+\Upsilon_i\leq\frac{2\mathsf{L}}{1-\lambda\rho}\sum\nolimits_{j=1}^i\|z_j^k-z^k\|+\Upsilon_i.\end{aligned}\tag{49}$$

Hence, combining (48) and (49), we can infer

$$\|z_{i+1}^k - z^k\| \le \alpha_k\Big[i\|F_{\rm nor}^\lambda(z^k)\| + \mathsf{C}_r\sum\nolimits_{j=1}^{i}\|z_j^k - z^k\| + \Upsilon_i\Big] \quad \forall\, i \in [n-1]. \tag{50}$$

Consequently, summing the term $\|z_i^k - z^k\|$ from $i = 1$ to n and using (50), we obtain

$$\begin{aligned}\sum\nolimits_{i=1}^{n}\|z_i^k - z^k\| &= \sum\nolimits_{i=1}^{n-1}\|z_{i+1}^k - z^k\|\\ &\le \mathsf{C}_r\alpha_k\sum\nolimits_{i=1}^{n-1}\sum\nolimits_{j=1}^{i}\|z^k - z_j^k\| + \alpha_k\sum\nolimits_{i=1}^{n-1}[i\|F_{\rm nor}^\lambda(z^k)\| + \Upsilon_i]\\ &\le \mathsf{C}_r n\alpha_k\sum\nolimits_{i=1}^{n}\|z_i^k - z^k\| + \frac{n(n-1)\alpha_k}{2}\|F_{\rm nor}^\lambda(z^k)\| + \alpha_k\sum\nolimits_{i=1}^{n-1}\Upsilon_i.\end{aligned}$$

Rearranging the terms in the previous inequality, it follows

$$\sum\nolimits_{i=1}^{n}\|z_i^k - z^k\| \le \frac{\alpha_k[\frac{n^2-n}{2}\|F_{\rm nor}^\lambda(z^k)\| + \sum_{i=1}^{n-1}\Upsilon_i]}{1 - \mathsf{C}_r n\alpha_k} \le \alpha_k\Big[n^2\|F_{\rm nor}^\lambda(z^k)\| + 2\sum\nolimits_{i=1}^{n-1}\Upsilon_i\Big],$$

where the last line is due to $\mathsf{C}_r n\alpha_k \le \frac{1}{2}$. Inserting this estimate into (47), we obtain

$$\|e^k\| \le \mathsf{C}_r\alpha_k^2\Big[n^2\|F_{\rm nor}^\lambda(z^k)\| + 2\sum\nolimits_{i=1}^{n-1}\Upsilon_i\Big].$$

Taking squares on both sides of this inequality and using $(\sum_{i=1}^{j} a_i)^2 \le j\sum_{i=1}^{j} a_i^2$ with $j = 2$ and $j = n-1$, this yields

$$\|e^k\|^2 \le 2\mathsf{C}_r^2\alpha_k^4\Big[n^4\|F_{\rm nor}^\lambda(z^k)\|^2 + 4(n-1)\sum\nolimits_{i=1}^{n-1}\Upsilon_i^2\Big]. \tag{51}$$

Thus, noticing that $\Upsilon_i^2 \le i\sum_{j=1}^{n}\|\nabla f(w^k, j) - \nabla f(w^k)\|^2 = in\sigma_k^2$ and $\sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}$, we finally obtain

$$\|e^k\|^2 \le 2\mathsf{C}_r^2 n^4\alpha_k^4[\|F_{\rm nor}^\lambda(z^k)\|^2 + 2\sigma_k^2].$$

We continue with the proof of part (b). Taking the conditional expectation $\mathbb{E}_k[\cdot]$ in (51), we have

$$\mathbb{E}_k[\|e^k\|^2] \le 2\mathsf{C}_r^2\alpha_k^4\Big[n^4\|F_{\rm nor}^\lambda(z^k)\|^2 + 4(n-1)\sum\nolimits_{i=1}^{n-1}\mathbb{E}_k[\Upsilon_i^2]\Big]. \tag{52}$$

According to Lemma 31, it holds that

$$\begin{aligned}\mathbb{E}_k[\Upsilon_i^2] &= \mathbb{E}_k\Big[\Big\|\sum\nolimits_{j=1}^{i}(\nabla f(w^k, \pi_j^k) - \nabla f(w^k))\Big\|^2\Big]\\ &= \frac{i(n-i)}{n(n-1)}\sum\nolimits_{j=1}^{n}\|\nabla f(w^k, j) - \nabla f(w^k)\|^2 = \frac{i(n-i)\sigma_k^2}{n-1} \le \frac{n^2\sigma_k^2}{4(n-1)}.\end{aligned}$$

Using this in (52), it follows $\mathbb{E}_k[\|e^k\|^2] \le 2\mathsf{C}_r^2 n^4\alpha_k^4[\|F_{\rm nor}^\lambda(z^k)\|^2 + \frac{\sigma_k^2}{n}]$. ■

B.3 Proof of Lemma 5

Proof Applying Lemmas 32 and 33, it follows

$$
\begin{aligned}
&H_\tau(z^{k+1}) - H_\tau(z^k) \\
&\le \left[\frac{\mathsf{L}+\rho}{2} - \frac{1}{\lambda}\right]\|w^{k+1}-w^k\|^2 + \langle F^\lambda_{\mathrm{nor}}(z^k) + \lambda^{-1}(z^{k+1}-z^k), w^{k+1}-w^k\rangle \\
&\quad + \frac{\tau\lambda}{2}[\|F^\lambda_{\mathrm{nor}}(z^{k+1})\|^2 - \|F^\lambda_{\mathrm{nor}}(z^k)\|^2] \\
&\le \left[\frac{\mathsf{L}+\rho}{2} - \frac{1}{\lambda}\right]\|w^{k+1}-w^k\|^2 + \langle F^\lambda_{\mathrm{nor}}(z^k) + \lambda^{-1}(z^{k+1}-z^k) - \lambda^{-1}\tau e^k, w^{k+1}-w^k\rangle \\
&\quad + \frac{\tau\lambda}{2}\left[\left(1-\frac{n\alpha_k}{\lambda}\right)^2 - 1\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 - \tau\left[1-\frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), w^{k+1}-w^k\rangle \\
&\quad + \tau\left[1-\frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), \lambda(\nabla f(w^{k+1}) - \nabla f(w^k)) + e^k\rangle \\
&\quad + \frac{\tau\lambda}{2}[\mathsf{L}+\lambda^{-1}]^2\|w^{k+1}-w^k\|^2 + \tau\langle \nabla f(w^{k+1}) - \nabla f(w^k), e^k\rangle + \frac{\tau}{2\lambda}\|e^k\|^2 \\
&= -\tau n\alpha_k\left[1-\frac{n\alpha_k}{2\lambda}\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \langle h^k, w^{k+1}-w^k\rangle \\
&\quad + \tau\left[1-\frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), \lambda(\nabla f(w^{k+1}) - \nabla f(w^k)) + e^k\rangle + \tau\langle \nabla f(w^{k+1}) - \nabla f(w^k), e^k\rangle \\
&\quad + \frac{\tau}{2\lambda}\|e^k\|^2 + \left[\frac{\mathsf{L}\tau(\mathsf{L}\lambda+2)+(\mathsf{L}+\rho)}{2} - \frac{2-\tau}{2\lambda}\right]\|w^{k+1}-w^k\|^2, \qquad (53)
\end{aligned}
$$

where $h^k := F^\lambda_{\mathrm{nor}}(z^k) + \lambda^{-1}(z^{k+1}-z^k) - \tau[1-\frac{n\alpha_k}{\lambda}]F^\lambda_{\mathrm{nor}}(z^k) - \lambda^{-1}\tau e^k$. Furthermore, according to the definition of F^λ_{nor} and e^k in (13) and (14), we have

$$
\begin{aligned}
h^k &= (1-\tau)\,F^\lambda_{\mathrm{nor}}(z^k) + \frac{1}{\lambda}(z^{k+1}-z^k) - \frac{\tau}{\lambda}[e^k - n\alpha_k F^\lambda_{\mathrm{nor}}(z^k)] \\
&= (1-\tau)\,F^\lambda_{\mathrm{nor}}(z^k) + \frac{1-\tau}{\lambda}(z^{k+1}-z^k) = (1-\tau)\left[\frac{1}{\lambda} - \frac{1}{n\alpha_k}\right](z^{k+1}-z^k) + \frac{1-\tau}{n\alpha_k}e^k.
\end{aligned}
$$

Inserting this expression into the estimate (53), we obtain

$$
\begin{aligned}
&H_\tau(z^{k+1}) - H_\tau(z^k) \qquad (54)\\
&\le -\tau n\alpha_k\left[1-\frac{n\alpha_k}{2\lambda}\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 - (1-\tau)\left[\frac{1}{n\alpha_k} - \frac{1}{\lambda}\right]\langle z^{k+1}-z^k, w^{k+1}-w^k\rangle + \frac{\tau}{2\lambda}\|e^k\|^2 \\
&\quad + \left[\frac{\mathsf{L}\tau(\mathsf{L}\lambda+2)+(\mathsf{L}+\rho)}{2} - \frac{2-\tau}{2\lambda}\right]\|w^{k+1}-w^k\|^2 + \frac{1-\tau}{n\alpha_k}\langle e^k, w^{k+1}-w^k\rangle \\
&\quad + \tau\langle \nabla f(w^{k+1}) - \nabla f(w^k), e^k\rangle + \tau\left[1-\frac{n\alpha_k}{\lambda}\right]\langle F^\lambda_{\mathrm{nor}}(z^k), \lambda(\nabla f(w^{k+1}) - \nabla f(w^k)) + e^k\rangle.
\end{aligned}
$$

We now estimate and bound the different terms appearing in (54). First, due to $\tau \in (0,1)$ and $n\alpha_k \in (0,\lambda)$ for all k, the coefficient in front of the inner product $\langle z^{k+1}-z^k, w^{k+1}-w^k\rangle$ is negative. Hence, using (8), we have

$$
-\langle z^{k+1}-z^k, w^{k+1}-w^k\rangle \le -(1-\lambda\rho)\|w^{k+1}-w^k\|^2.
$$

Applying the Cauchy-Schwartz inequality, the Lipschitz continuity of ∇f, Young's inequality—$\langle a,b\rangle \le \frac{\varepsilon}{2}\|a\|^2 + \frac{1}{2\varepsilon}\|b\|^2,\ a,b\in\mathbb{R}^d$ and $\varepsilon>0$—(with $a=w^{k+1}-w^k$, $b=e^k$, and $\varepsilon=1$), we further obtain

$$\langle \nabla f(w^{k+1})-\nabla f(w^k), e^k\rangle \le \mathsf{L}\|w^{k+1}-w^k\|\|e^k\| \le \frac{\mathsf{L}}{2}\|w^{k+1}-w^k\|^2 + \frac{\mathsf{L}}{2}\|e^k\|^2$$

and, similarly, $\langle e^k, w^{k+1}-w^k\rangle \le \frac{1}{2}\|w^{k+1}-w^k\|^2 + \frac{1}{2}\|e^k\|^2$. Moreover, setting $a=F^\lambda_{\mathrm{nor}}(z^k)$, $b=\lambda(\nabla f(w^{k+1})-\nabla f(w^k))$ and $\varepsilon = n\alpha_k$ in Young's inequality and applying the Lipschitz continuity of ∇f, it holds that

$$\begin{aligned}\langle F^\lambda_{\mathrm{nor}}(z^k), \lambda(\nabla f(w^{k+1})-\nabla f(w^k))\rangle &\le \frac{n\alpha_k}{2}\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \frac{\lambda^2}{2n\alpha_k}\|\nabla f(w^{k+1})-\nabla f(w^k)\|^2\\ &\le \frac{n\alpha_k}{2}\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \frac{\mathsf{L}^2\lambda^2}{2n\alpha_k}\|w^{k+1}-w^k\|^2.\end{aligned}$$

Repeating this step once more with $a=F^\lambda_{\mathrm{nor}}(z^k)$, $b=e^k$ and $\varepsilon=\frac{n\alpha_k}{2}$, we have $\langle F^\lambda_{\mathrm{nor}}(z^k), e^k\rangle \le \frac{n\alpha_k}{4}\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \frac{1}{n\alpha_k}\|e^k\|^2$. Plugging these different estimates into (54), we can conclude

$$\begin{aligned}&H_\tau(z^{k+1}) - H_\tau(z^k)\\ &\le -\tau n\alpha_k\left[1-\frac{n\alpha_k}{2\lambda}-\frac{3}{4}\left(1-\frac{n\alpha_k}{\lambda}\right)\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2\\ &\quad+\left[\frac{\tau}{2\lambda}+\frac{1-\tau}{2n\alpha_k}+\frac{\mathsf{L}\tau}{2}+\frac{\tau}{n\alpha_k}\left(1-\frac{n\alpha_k}{\lambda}\right)\right]\|e^k\|^2+\left[\frac{\mathsf{L}\tau(\mathsf{L}\lambda+2)+(\mathsf{L}+\rho)}{2}-\frac{2-\tau}{2\lambda}\quad\cdots\right.\\ &\quad\left.\cdots-(1-\tau)(1-\lambda\rho)\left(\frac{1}{n\alpha_k}-\frac{1}{\lambda}\right)+\frac{1-\tau}{2n\alpha_k}+\frac{\mathsf{L}\tau}{2}+\frac{\mathsf{L}^2\lambda^2\tau}{2n\alpha_k}\left(1-\frac{n\alpha_k}{\lambda}\right)\right]\|w^{k+1}-w^k\|^2\\ &=-\frac{\tau n\alpha_k}{4}\left[1+\frac{n\alpha_k}{\lambda}\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2+\left[\frac{1+\tau}{2n\alpha_k}+\frac{\mathsf{L}\tau}{2}-\frac{\tau}{2\lambda}\right]\|e^k\|^2\\ &\quad+\left[\frac{3\mathsf{L}\tau}{2}+\frac{\mathsf{L}+\rho}{2}-\frac{\tau+2\lambda\rho(1-\tau)}{2\lambda}-\frac{(1-\tau)(1-2\lambda\rho)-\mathsf{L}^2\lambda^2\tau}{2n\alpha_k}\right]\|w^{k+1}-w^k\|^2.\end{aligned}$$

By assumption, it holds that $\lambda\rho<\frac{1}{4}$ and we have $\tau=\frac{1-4\lambda\rho}{2(1-2\lambda\rho+\mathsf{L}^2\lambda^2)}\le\frac{1}{2}$. We then may infer

$$\frac{3\mathsf{L}\tau}{2}+\frac{\mathsf{L}+\rho}{2}-\frac{(1-2\lambda\rho)\tau+2\lambda\rho}{2\lambda}-\frac{(1-\tau)(1-2\lambda\rho)-\mathsf{L}^2\lambda^2\tau}{2n\alpha_k}\le\frac{5\mathsf{L}}{4}-\frac{1}{4n\alpha_k}$$

and $\frac{1+\tau}{2n\alpha_k}+\frac{\mathsf{L}\tau}{2}-\frac{\tau}{2\lambda}\le\frac{3}{4n\alpha_k}+\frac{\mathsf{L}}{4}$. The choice of the step sizes $\{\alpha_k\}_k$ implies $\frac{5\mathsf{L}}{4}\le\frac{1}{8n\alpha_k}$ and hence, it follows $\frac{5\mathsf{L}}{4}-\frac{1}{4n\alpha_k}\le-\frac{1}{8n\alpha_k}$ and $\frac{3}{4n\alpha_k}+\frac{\mathsf{L}}{4}\le\frac{1}{n\alpha_k}$. Using these bounds in the previous estimate, we finally obtain

$$H_\tau(z^{k+1})-H_\tau(z^k)+\frac{1}{8n\alpha_k}\|w^{k+1}-w^k\|^2\le-\frac{\tau n\alpha_k}{4}\left[1+\frac{n\alpha_k}{\lambda}\right]\|F^\lambda_{\mathrm{nor}}(z^k)\|^2+\frac{1}{n\alpha_k}\|e^k\|^2,$$

as desired. ■

Appendix C. Global Convergence: Proof of Theorem 12

Proof Invoking Lemma 6 (a), we have

$$H_\tau(z^{k+1}) \leq H_\tau(z^k) - \frac{\tau n \alpha_k}{4}\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \mathsf{D}\alpha_k^3,$$

where $\mathsf{D} := n^3\Delta(n^3\sum_{i=1}^\infty \alpha_i^3) < \infty$. Thus, the sequence $\{H_\tau(z^k)\}_k$ satisfies a supermartingale-type recursion. Consequently, applying Theorem 30 and using the lower bound $H_\tau(z^k) \geq \psi(w^k) \geq \psi_{\mathrm{lb}}$ (as stated in assumption (F.3)) and the condition $\sum_{k=1}^\infty \alpha_k^3 < \infty$ (as stated in (21)), we can infer $H_\tau(z^k) \to \bar{\psi}$, $k \to \infty$ for some $\bar{\psi} \in \mathbb{R}$ and

$$\sum\nolimits_{k=1}^\infty \alpha_k \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 < \infty. \tag{55}$$

Due to $\sum_{k=1}^\infty \alpha_k = \infty$, this immediately implies $\liminf_{k\to\infty} \|F^\lambda_{\mathrm{nor}}(z^k)\| = 0$. In order to show $\lim_{k\to\infty} \|F^\lambda_{\mathrm{nor}}(z^k)\| = 0$, let us, on the contrary, assume that $\{\|F^\lambda_{\mathrm{nor}}(z^k)\|\}_k$ does not converge to zero. Then, there exist $\varepsilon > 0$ and two infinite subsequences $\{t_j\}_j$ and $\{\ell_j\}_j$ such that $t_j < \ell_j \leq t_{j+1}$,

$$\|F^\lambda_{\mathrm{nor}}(z^{t_j})\| \geq 2\varepsilon, \quad \|F^\lambda_{\mathrm{nor}}(z^{\ell_j})\| < \varepsilon, \quad \text{and} \quad \|F^\lambda_{\mathrm{nor}}(z^k)\| \geq \varepsilon \tag{56}$$

for all $k = t_j + 1, \ldots, \ell_j - 1$. Combining this observation with (55), this yields

$$\infty > \sum\nolimits_{k=1}^\infty \alpha_k \|F^\lambda_{\mathrm{nor}}(z^k)\|^2 \geq \varepsilon^2 \sum\nolimits_{j=1}^\infty \sum\nolimits_{k=t_j}^{\ell_j-1} \alpha_k,$$

which implies $\lim_{j\to\infty}\ \beta_j := \sum_{k=t_j}^{\ell_j-1}\alpha_k = 0$. Next, applying the triangle and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\|z^{\ell_j} - z^{t_j}\| &\leq \sum\nolimits_{k=t_j}^{\ell_j-1} \sqrt{\alpha_k}\left[\frac{\|z^{k+1}-z^k\|}{\sqrt{\alpha_k}}\right] \\ &\leq \left[\sum\nolimits_{k=t_j}^{\ell_j-1}\alpha_k \cdot \sum\nolimits_{k=t_j}^{\ell_j-1}\frac{\|z^{k+1}-z^k\|^2}{\alpha_k}\right]^{\frac{1}{2}} \leq \sqrt{\beta_j}\cdot\left[\sum\nolimits_{k=1}^\infty \frac{\|z^{k+1}-z^k\|^2}{\alpha_k}\right]^{\frac{1}{2}}. \end{aligned} \tag{57}$$

Using the recursion (13) and Lemmas 2 and 3, we further have

$$\begin{aligned}\sum\nolimits_{k=1}^\infty \alpha_k^{-1}\|z^{k+1}-z^k\|^2 &\leq 2\sum\nolimits_{k=1}^\infty [n^2\alpha_k\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \alpha_k^{-1}\|e^k\|^2] \\ &\leq 2\sum\nolimits_{k=1}^\infty [(n^2 + \mathsf{C}n^4\alpha_k^2)\alpha_k\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + \mathsf{C}n^4\alpha_k^3\sigma_k^2] \\ &\leq 2(n^2 + \mathsf{C}n^2\bar{\alpha}^2)\sum\nolimits_{k=1}^\infty \alpha_k\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 + 2\mathsf{D}n\sum\nolimits_{k=1}^\infty \alpha_k^3 < \infty,\end{aligned}$$

where the last line is by (21), (55) and $\mathsf{C}\sigma_k^2 \leq \Delta(n^3\sum_{i=1}^\infty \alpha_i^3) = \frac{\mathsf{D}}{n^3}$ (cf. Lemma 6 (a)). Hence, taking the limit $j \to \infty$ in (57), it follows

$$\lim\nolimits_{j\to\infty} \|z^{\ell_j} - z^{t_j}\| = 0.$$

Moreover, applying (56), the inverse triangle inequality, and (46), it holds that

$$\varepsilon \leq |\|F^\lambda_{\mathrm{nor}}(z^{\ell_j})\| - \|F^\lambda_{\mathrm{nor}}(z^{t_j})\|| \leq \|F^\lambda_{\mathrm{nor}}(z^{\ell_j}) - F^\lambda_{\mathrm{nor}}(z^{t_j})\| \leq \frac{\mathsf{L} + 2\lambda^{-1} - \rho}{1-\lambda\rho}\|z^{\ell_j} - z^{t_j}\|.$$

Taking the limit $j \to \infty$, we reach a contradiction and thus, we conclude $\lim_{k\to\infty} \|F^\lambda_{\mathrm{nor}}(z^k)\| = 0$. Finally, recalling $H_\tau(z^k) := \psi(z^k) + \frac{\lambda\tau}{2}\|F^\lambda_{\mathrm{nor}}(z^k)\|$ and $H_\tau(z^k) \to \bar{\psi}$, we have

$$\bar{\psi} = \lim_{k\to\infty} H_\tau(z^k) = \lim_{k\to\infty} \psi(w^k) + \lim_{k\to\infty} \frac{\lambda\tau}{2}\|F^\lambda_{\mathrm{nor}}(z^k)\| = \lim_{k\to\infty} \psi(w^k).$$

This completes the proof. ■

Appendix D. Strong Convergence

D.1 Proof of Lemma 19

Proof By the definition of the normal map, we have $z^k = \lambda F^\lambda_{\mathrm{nor}}(z^k) + w^k - \lambda\nabla f(w^k)$. Since $F^\lambda_{\mathrm{nor}}(z^k) \to 0$ (cf. Theorem 12) and $\{w^k\}_k$ is bounded, we can conclude that the sequence $\{z^k\}_k$ is bounded. This verifies (a).

The inclusions $\mathcal{A}_z \subseteq \{z : F^\lambda_{\mathrm{nor}}(z) = 0\}$ and $\mathcal{A}_w \subseteq \mathrm{crit}(\psi)$ follow directly from Theorem 12. Next, let $\bar{w} \in \mathcal{A}_w$ with $w^{\ell_k} \to \bar{w}$ be arbitrary. Then, we have $z^{\ell_k} \to \bar{w} - \lambda\nabla f(\bar{w}) =: \bar{z}$ and $\bar{w} = \lim_{k\to\infty} \mathrm{prox}_{\lambda\varphi}(z^{\ell_k}) = \mathrm{prox}_{\lambda\varphi}(\bar{z})$. Conversely, let $\bar{w} = \mathrm{prox}_{\lambda\varphi}(\bar{z})$ with $\bar{z} \in \mathcal{A}_z$ be given. Thus, by definition, there is $z^{\ell_k} \to \bar{z}$ and using the Lipschitz continuity of the proximity operator, it holds that $\|w^{\ell_k} - \bar{w}\| = \|\mathrm{prox}_{\lambda\varphi}(z^{\ell_k}) - \mathrm{prox}_{\lambda\varphi}(\bar{z})\| \le \|z^{\ell_k} - \bar{z}\|/(1-\lambda\rho) \to 0$. This verifies $\bar{w} \in \mathcal{A}_w$ and finishes the proof of part (b).

For part (c) and due to the structure of $\mathcal{A}_w$, we may pick some arbitrary $\bar{z} \in \mathcal{A}_z$ and $\bar{w} \in \mathcal{A}_w$ with $\bar{w} = \mathrm{prox}_{\lambda\varphi}(\bar{z})$. Let $\{\ell_k\}_k$ be a subsequence such that $z^{\ell_k} \to \bar{z}$ and $w^{\ell_k} \to \bar{w}$. Applying Theorem 12, we conclude $\psi(w^k) \to \bar{\psi} \in \mathbb{R}$. Since $z \mapsto \psi(\mathrm{prox}_{\lambda\varphi}(z))$ is continuous, we further have $\bar{\psi} = \lim_{k\to\infty} \psi(\mathrm{prox}_{\lambda\varphi}(z^{\ell_k})) = \psi(\mathrm{prox}_{\lambda\varphi}(\bar{z}))$. This implies $\lim_{k\to\infty} \psi(w^k) = \psi(\bar{w}) = \bar{\psi}$ for all $\bar{w} \in \mathcal{A}_w$. Finally, we have $H_\tau(\bar{z}) = \psi(\bar{w}) + \frac{\tau\lambda}{2}\|F^\lambda_{\mathrm{nor}}(\bar{z})\|^2 = \psi(\bar{w}) = \bar{\psi}$ for all $\tau > 0$, i.e., H_τ is constant on $\mathcal{A}_z$. ■

D.2 Proof of Theorem 21

Proof The properties in Lemma 19 allow us to apply the uniformization technique in (Bolte et al., 2014, Lemma 6) to the KL-type inequality derived in Lemma 18. Specifically, there exist $\hat{\zeta}, \hat{c} > 0$, $\hat{\eta} \in (0, 1]$, and $\hat{\theta} \in [\frac{1}{2}, 1)$ such that for all $\bar{z} \in \mathcal{A}_z$ and $z \in V_{\hat{\zeta},\hat{\eta}} := \{z \in \mathbb{R}^d : \mathrm{dist}(z, \mathcal{A}_z) < \hat{\zeta}\} \cap \{z \in \mathbb{R}^d : 0 < |H_\tau(z) - \bar{\psi}| < \hat{\eta}\}$, we have

$$\hat{c} \cdot \|F^\lambda_{\mathrm{nor}}(z)\| \ge |H_\tau(z) - \bar{\psi}|^{\hat{\theta}} \ge |H_\tau(z) - \bar{\psi}|^{\vartheta},$$

where $\vartheta := \max\{\hat{\theta}, \xi\}$ and $\bar{\psi} = \psi(\bar{w}) = H_\tau(\bar{z})$ (for all $\bar{z} \in \mathcal{A}_z$). Setting $\varrho_\vartheta(s) := \frac{\hat{c}}{1-\vartheta} \cdot s^{1-\vartheta}$, this can be written as

$$\varrho'_\vartheta(|H_\tau(z) - \bar{\psi}|) \cdot \|F^\lambda_{\mathrm{nor}}(z)\| \ge 1. \tag{58}$$

Since $\mathrm{dist}(z^k, \mathcal{A}_z) \to 0$ and $H_\tau(z^k) \to \bar{\psi}$, there exists $\bar{k} \ge 1$ such that $z^k \in V_{\hat{\zeta},\hat{\eta}}$ for all $k \ge \bar{k}$. Applying Lemma 6 (a), it holds that

$$H_\tau(z^{k+1}) \le H_\tau(z^k) - \frac{\tau n\alpha_k}{4}\|F^\lambda_{\mathrm{nor}}(z^k)\|^2 - \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2 + \mathsf{D}\alpha_k^3,$$

where $\mathsf{D} = n^3\Delta(n^3\sum_{i=1}^{\infty}\alpha_i^3) < \infty$. Defining $u_k := \mathsf{D}\sum_{j=k}^{\infty}\alpha_j^3$ and adding u_{k+1} on both sides of this inequality, it follows

$$H_\tau(z^{k+1}) + u_{k+1} \leq H_\tau(z^k) + u_k - \frac{\tau n\alpha_k}{4}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 - \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2. \tag{59}$$

In the following, without loss of generality, let us assume $z^k \notin \mathcal{A}_z$ or $u_k \neq 0$ for all $k \geq \bar{k}$. Let us set $\delta_k := \varrho_\vartheta(H_\tau(z^k) - \bar{\psi} + u_k)$. Due to the monotonicity of the sequence $\{H_\tau(z^k) + u_k\}_k$ and $H_\tau(z^k) + u_k \to \bar{\psi}$, δ_k is well defined as $H_\tau(z^k) - \bar{\psi} + u_k \geq 0$ for all $k \geq 1$. Hence, for all $k \geq \bar{k}$, we obtain

$$\begin{aligned}
\delta_k - \delta_{k+1} &\geq \varrho_\vartheta'(H_\tau(z^k) - \bar{\psi} + u_k)[H_\tau(z^k) + u_k - H_\tau(z^{k+1}) - u_{k+1}] \\
&\geq \varrho_\vartheta'(|H_\tau(z^k) - \bar{\psi}| + u_k)[H_\tau(z^k) + u_k - H_\tau(z^{k+1}) - u_{k+1}] \\
&\geq \varrho_\vartheta'(|H_\tau(z^k) - \bar{\psi}| + u_k)\left[\frac{\tau n\alpha_k}{4}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2\right] \\
&\geq \frac{\frac{\tau n\alpha_k}{4}\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + \frac{1}{8n\alpha_k}\|w^{k+1} - w^k\|^2}{[\varrho_\vartheta'(|H_\tau(z^k) - \bar{\psi}|)]^{-1} + [\varrho_\vartheta'(u_k)]^{-1}} \geq \frac{\tau n}{4}\frac{\alpha_k^2\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 + \frac{1}{2\tau n^2}\|w^{k+1} - w^k\|^2}{\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| + \alpha_k[\varrho_\vartheta'(u_k)]^{-1}},
\end{aligned} \tag{60}$$

where the first inequality uses the concavity of ϱ_ϑ, the second inequality is due to monotonicity of $s \mapsto \varrho_\vartheta'(s) = \hat{c}s^{-\vartheta}$, the third inequality is from (59), the fourth inequality follows from subadditivity of $[\varrho_\vartheta'(s)]^{-1}$—i.e., $[\varrho_\vartheta'(s_1 + s_2)]^{-1} \leq [\varrho_\vartheta'(s_1)]^{-1} + [\varrho_\vartheta'(s_2)]^{-1}$ for all $s_1, s_2 \geq 0$—and the last inequality applies the KL inequality (58). Rearranging the terms in (60), this further yields

$$\begin{aligned}
&\frac{4(\delta_k - \delta_{k+1})}{\tau n} \cdot [\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| + \alpha_k[\varrho_\vartheta'(u_k)]^{-1}] \\
&\quad\geq \frac{1}{2\tau n^2}\|w^{k+1} - w^k\|^2 + \alpha_k^2\|F_{\mathrm{nor}}^\lambda(z^k)\|^2 \geq \frac{1}{2}\Big[\frac{1}{\sqrt{2\tau}n}\|w^{k+1} - w^k\| + \alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\|\Big]^2
\end{aligned} \tag{61}$$

for all $k \geq \bar{k}$, where we used $a^2 + b^2 \geq (a+b)^2/2$. Multiplying both sides of (61) with 8 and taking square root, we obtain

$$\begin{aligned}
\frac{\sqrt{2}}{\sqrt{\tau}n}\|w^{k+1} - w^k\| + 2\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| &\leq \sqrt{\frac{16(\delta_k - \delta_{k+1})}{\tau n} \cdot 2[\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| + \alpha_k[\varrho_\vartheta'(u_k)]^{-1}]} \\
&\leq \frac{8(\delta_k - \delta_{k+1})}{\tau n} + \alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| + \alpha_k[\varrho_\vartheta'(u_k)]^{-1},
\end{aligned} \tag{62}$$

where the last inequality is due to the AM-GM inequality, $\sqrt{ab} \leq (a+b)/2$. Summing the inequality (62) from $k = \bar{k}$ to T, we obtain

$$\frac{\sqrt{2}}{\sqrt{\tau}n}\sum_{k=\bar{k}}^{T}\|w^{k+1} - w^k\| + \sum_{k=\bar{k}}^{T}\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| \leq \frac{8(\delta_{\bar{k}} - \delta_{T+1})}{\tau n} + \sum_{k=\bar{k}}^{T}\alpha_k[\varrho_\vartheta'(u_k)]^{-1}. \tag{63}$$

Since ϱ_ϑ is continuous with $\varrho_\vartheta(0) = 0$, we have $\delta_{T+1} \to 0$ as $T \to \infty$. In addition, by (31) and using $\vartheta \geq \xi$, it follows $\sum_{k=\bar{k}}^{\infty}\alpha_k[\varrho_\vartheta'(u_k)]^{-1} = \hat{c}\sum_{k=\bar{k}}^{\infty}\alpha_k u_k^\vartheta < \infty$. Thus, taking the limit $T \to \infty$ in (63), it holds that

$$\sum_{k=1}^{\infty}\alpha_k\|F_{\mathrm{nor}}^\lambda(z^k)\| < \infty \quad \text{and} \quad \sum_{k=1}^{\infty}\|w^{k+1} - w^k\| < \infty.$$

The second estimate implies that $\{w^k\}_k$ has finite length, and hence it is convergent to some w^*. Finally, by Theorem 12, the limit w^* is a stationary point of ψ. ■

D.3 Proof of Lemma 28

Proof Similar to (46), we have

$$\|F^\lambda_{\rm nor}(z^{k+i}) - F^\lambda_{\rm nor}(z^k)\| \le \frac{\mathsf{L} + 2\lambda^{-1} - \rho}{1-\lambda\rho}\|z^{k+i} - z^k\| \quad \forall\, i \ge 1 \text{ and } k \ge 1. \tag{64}$$

Applying Lemmas 2 and 3 and $\max\{\|F^\lambda_{\rm nor}(z^k)\|^2, \psi(w^k) - \psi_{\rm lb}\} \le \mathsf{P}$, it further follows

$$\|e^k\| \le \sqrt{\mathsf{C}} n^2 \alpha_k^2 (\|F^\lambda_{\rm nor}(z^k)\|^2 + 2\mathsf{L}[\psi(w^k) - \psi_{\rm lb}])^{\frac{1}{2}} \le \sqrt{\mathsf{CP}(2\mathsf{L}+1)} n^2 \alpha_k^2. \tag{65}$$

Setting $\mathsf{C}_1 := \sqrt{\mathsf{C}(2\mathsf{L}+1)}n$, the update rule (13) and the estimate (65) imply

$$\begin{aligned}\|z^{k+i} - z^k\| &\le n\sum\nolimits_{j=0}^{i-1} \alpha_{k+j}\|F^\lambda_{\rm nor}(z^{k+j})\| + \sum\nolimits_{j=0}^{i-1}\|e^{k+j}\| \\ &\le n\left[\varsigma\|F^\lambda_{\rm nor}(z^k)\| + \sum\nolimits_{j=0}^{i-1}\alpha_{k+j}\|F^\lambda_{\rm nor}(z^{k+j}) - F^\lambda_{\rm nor}(z^k)\|\right] + \mathsf{C}_1 n\sqrt{\mathsf{P}}\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}^2,\end{aligned}$$

where the last line is due to $\sum_{j=0}^{i-1}\alpha_{k+j} \le \varsigma$. Defining $\mathsf{C}_f := (1-\lambda\rho)^{-1}(\mathsf{L} + 2\lambda^{-1} - \rho)n$ and combining the previous estimates, we obtain

$$\begin{aligned}&\|F^\lambda_{\rm nor}(z^{k+i}) - F^\lambda_{\rm nor}(z^k)\| \le \mathsf{C}_f n^{-1}\|z^{k+i} - z^k\| \\ &\quad\le \mathsf{C}_f\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}\|F^\lambda_{\rm nor}(z^{k+j}) - F^\lambda_{\rm nor}(z^k)\| + \mathsf{C}_f\left[\varsigma\|F^\lambda_{\rm nor}(z^k)\| + \mathsf{C}_1\sqrt{\mathsf{P}}\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}^2\right].\end{aligned}$$

We now apply Gronwall's inequality (Lemma 29) upon setting $\mathsf{Q} := \mathsf{C}_f\max\{1, \mathsf{C}_1\}$,

$$p := \mathsf{Q}\varsigma\|F^\lambda_{\rm nor}(z^k)\| + \mathsf{Q}\sqrt{\mathsf{P}}\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}^2, \quad q := \mathsf{Q}, \quad a_j := \alpha_{k+j},$$

$y_j := \|F^\lambda_{\rm nor}(z^{k+j}) - F^\lambda_{\rm nor}(z^k)\|$, and $t := i-1$. This establishes the following upper bound:

$$\|F^\lambda_{\rm nor}(z^{k+i}) - F^\lambda_{\rm nor}(z^k)\| \le \mathsf{Q}\exp(\mathsf{Q}\varsigma)\left[\varsigma\|F^\lambda_{\rm nor}(z^k)\| + \sqrt{\mathsf{P}}\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}^2\right].$$

Noticing $\varsigma > 0$ and $\mathsf{Q}\varsigma\exp(\mathsf{Q}\varsigma) \le \frac{1}{2}$ (per assumption), we can infer

$$\begin{aligned}\|F^\lambda_{\rm nor}(z^k)\| &\le \|F^\lambda_{\rm nor}(z^{k+i}) - F^\lambda_{\rm nor}(z^k)\| + \|F^\lambda_{\rm nor}(z^{k+i})\| \\ &\le \frac{1}{2}\|F^\lambda_{\rm nor}(z^k)\| + \|F^\lambda_{\rm nor}(z^{k+i})\| + \frac{\sqrt{\mathsf{P}}}{2\varsigma}\sum\nolimits_{j=0}^{i-1}\alpha_{k+j}^2.\end{aligned}$$

Rearranging the terms yields $\|F^\lambda_{\rm nor}(z^{k+i})\| + \frac{\sqrt{\mathsf{P}}}{2\varsigma}\sum_{j=0}^{i-1}\alpha_{k+j}^2 \ge \frac{1}{2}\|F^\lambda_{\rm nor}(z^k)\|$. Taking square and using $a^2 + b^2 \ge \frac{1}{2}(a+b)^2$, we have $\|F^\lambda_{\rm nor}(z^{k+i})\|^2 + \frac{\mathsf{P}}{4\varsigma^2}(\sum_{j=0}^{i-1}\alpha_{k+j}^2)^2 \ge \frac{1}{8}\|F^\lambda_{\rm nor}(z^k)\|^2$. ■

D.4 Proof of Theorem 24: Rate of $\{w^k\}_k$

Proof In this section, we complete the proof of Theorem 24 and provide rates for the iterates $\{w^k\}_k$. Applying (63) with $T \to \infty$, we have

$$\frac{\sqrt{2}}{\sqrt{\tau}n}\|w^k - w^*\| \leq \frac{\sqrt{2}}{\sqrt{\tau}n}\sum\nolimits_{i=k}^{\infty}\|w^{i+1} - w^i\| \leq \frac{8\delta_k}{\tau n} + \sum\nolimits_{i=k}^{\infty}\alpha_i[\varrho_\vartheta'(u_i)]^{-1}, \tag{66}$$

for all $k \geq \bar{k}$, where $\delta_k := \varrho_\vartheta(r_k)$, $\varrho_\vartheta(s) := \frac{\tilde{c}}{1-\vartheta}s^{1-\vartheta}$, and $\vartheta \in [\max\{\frac{1}{2}, \theta\}, 1)$. (As argued in the first parts of the proof, due to $w^k \to w^*$ and $z^k \to z^*$, we can directly work with the KL exponent θ and adjusted constant $\tilde{c}$ instead of using the uniformized versions $\hat{\theta}$ and $\hat{c}$). Hence, based on the rate for $\{r_k\}_k$ and using $R(\vartheta, \gamma) \leq R(\theta, \gamma)$ and Lemma 22 (b), we have

$$\delta_k = \mathcal{O}(1/k^{(1-\vartheta)R(\vartheta,\gamma)}) \text{ and } \sum\nolimits_{i=k}^{\infty}\alpha_i[\varrho_\vartheta'(u_i)]^{-1} = \mathcal{O}\Big(\sum\nolimits_{i=k}^{\infty}\alpha_i u_i^{\vartheta}\Big) = \mathcal{O}(1/k^{(3\gamma-1)\vartheta-(1-\gamma)}).$$

Note that the adjusted KL exponent ϑ can be selected freely. Thus, we may increase it to ensure $\vartheta > \frac{1-\gamma}{3\gamma-1}$ such that $\sum_{i=k}^{\infty}\alpha_i[\varrho_\vartheta'(u_i)]^{-1} \to 0$ as $k \to \infty$. Hence, provided $\vartheta > \frac{1-\gamma}{3\gamma-1}$, the convergence rate for $\{w^k\}_k$ is

$$\|w^k - w^*\| = \mathcal{O}(k^{-Q(\vartheta,\gamma)}) \quad \text{where} \quad Q(\vartheta, \gamma) := \min\{(1-\vartheta)R(\vartheta, \gamma), (3\gamma-1)\vartheta - (1-\gamma)\}.$$

We observe that $Q(\vartheta, \gamma) = (3\gamma-1)\vartheta - (1-\gamma)$ if $\vartheta \in (\frac{1-\gamma}{3\gamma-1}, \frac{\gamma}{3\gamma-1}]$ and $Q(\vartheta, \gamma) = \frac{(1-\vartheta)(1-\gamma)}{2\vartheta-1}$ if $\vartheta \in (\frac{\gamma}{3\gamma-1}, 1)$. This means that the mapping $Q(\cdot, \gamma)$ is increasing over the interval $(\frac{1-\gamma}{3\gamma-1}, \frac{\gamma}{3\gamma-1}]$ and decreasing over $(\frac{\gamma}{3\gamma-1}, 1)$. Therefore, we can choose ϑ in the following way to maximize $Q(\vartheta, \gamma)$: when $\theta \leq \frac{\gamma}{3\gamma-1}$, set $\vartheta = \frac{\gamma}{3\gamma-1}$; otherwise, set $\vartheta = \theta$. In summary, we obtain

$$\|w^k - w^*\| = \mathcal{O}(k^{-R_w(\theta,\gamma)}) \quad \text{where} \quad R_w(\theta, \gamma) := \begin{cases} 2\gamma - 1 & \text{if } \theta \in [0, \frac{\gamma}{3\gamma-1}] \\ \frac{(1-\gamma)(1-\theta)}{2\theta-1} & \text{if } \theta \in (\frac{\gamma}{3\gamma-1}, 1) \end{cases} \quad \gamma \in (\tfrac{1}{2}, 1),$$

and $R_w(\theta, \gamma) := 1$ when $\gamma = 1, \theta \in [0, \frac{1}{2}]$ and $\alpha > \frac{16\tilde{c}^2}{\tau n}$. ■

References

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Aleksandr Aravkin, Tristan Van Leeuwen, and Felix Herrmann. Robust full-waveform inversion using the student's t-distribution. In *SEG Technical Program Expanded Abstracts 2011*, pages 2669–2673. Society of Exploration Geophysicists, 2011.

Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.*, 18(10):1–33, 2017.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2, Ser. B):5–16, 2009.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

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. *Math. Program.*, 137(1-2, Ser. A):91–129, 2013.

Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, third edition, 2016.

Jacek Bochnak, Michel Coste, and Marie-Fran¸coise Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1998. Translated from the 1987 French original, Revised by the authors.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. A nonsmooth Morse-Sard theorem for subanalytic functions. *J. Math. Anal. Appl.*, 321(2):729–740, 2006a.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2006b.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, 18(2):556–572, 2007.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2, Ser. A):459–494, 2014.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2):471–507, 2017.

S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi. On the convergence of a linesearch based proximal-gradient method for nonconvex optimization. *Inverse Problems*, 33(5):055005, 30, 2017.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

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

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

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

K. L. Chung. On a stochastic approximation method. *Ann. Math. Statistics*, 25:463–483, 1954.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

Didier D'Acunto and Krzysztof Kurdyka. Explicit bounds for the Łojasiewicz exponent in the gradient inequality for polynomials. *Ann. Polon. Math.*, 87:51–61, 2005.

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

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

Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes. *Journal of Machine Learning*, 3(3):245–281, 2024.

Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

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

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems. Vol. II.* Springer-Verlag, New York, 2003. ISBN 0-387-95581-X.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

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

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2, Ser. A):267–305, 2016.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

Mert Gürbüzbalaban, Asu Ozdaglar, and PA Parrilo. Why random reshuffling beats stochastic gradient descent. *Math. Program.*, 186(1-2):49–84, 2021.

Huy-Vui Hà and Tiê'n-So'n Phạm. *Genericity in polynomial optimization*, volume 3 of *Series on Optimization and its Applications*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017. With a foreword by Jean Bernard Lasserre.

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633. PMLR, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Tim Hoheisel, Maxime Laborde, and Adam Oberman. A regularization interpretation of the proximal point method for weakly convex functions. *J. Dyn. Games*, 7(1):79–96, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.*, 18(5):1199–1232, 2018.

Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. In *Adv. Neural Inf. Process. Syst.*, volume 35, pages 33107–33119, 2022.

Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality. *SIAM J. Optim.*, 33(2):1092–1120, 2023.

P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

Huikang Liu, Anthony Man-Cho So, and Weijie Wu. Quadratic optimization with orthogonality constraint: explicit Łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Math. Program.*, 178(1): 215–262, 2019.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. *arXiv preprint arXiv:2403.07723*, 2024.

Stanislaw Łojasiewicz. Sur le problème de la division. *Studia Mathematica*, 18:87–136, 1959.

Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Szymon Majewski, Blazej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. arXiv preprint, arXiv:1805.01916v1, 2018.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 12, 1999.

Andre Milzarek, Xiantao Xiao, Shicong Cen, Zaiwen Wen, and Michael Ulbrich. A stochastic semismooth newton method for nonsmooth nonconvex optimization. *SIAM J. Optim.*, 29 (4):2916–2948, 2019.

H. Mine and M. Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *J. Optim. Theory Appl.*, 33(1):9–23, 1981.

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

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pages 15718–15749. PMLR, 2022.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience Series in Discrete Mathematics. Wiley, Chichester, 1983.

Yurii Nesterov. *Lectures on Convex Optimization.* Springer, Berlin, Heidelberg, 2018.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. SGD and Hogwild! Convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

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. *J. Mach. Learn. Res.*, 22:1–44, 2021.

Phuong_ha Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: inertial proximal algorithm for nonconvex optimization. *SIAM J. Imaging Sci.*, 7(2):1388–1419, 2014.

Wenqing Ouyang and Andre Milzarek. A trust region-type normal map-based semismooth Newton method for nonsmooth nonconvex composite optimization. *Math. Program.*, pages 1–47, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Andrei Patrascu and Paul Irofti. Stochastic proximal splitting algorithm for composite minimization. *Optim. Lett.*, 15(6):2255–2273, 2021.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Boris T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. ISBN 0-911575-14-6. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

Stephen M Robinson. Normal maps induced by linear transformations. *Math. Oper. Res.*, 17 (3):691–714, 1992.

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

Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *Appl. Math. Optim.*, 82(3):891–917, 2020.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *J. Mach. Learn. Res.*, 12:1865–1892, 2011.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Jialiang Tang, Mingjin Liu, Ning Jiang, Huan Cai, Wenxin Yu, and Jinjia Zhou. Data-free network pruning for model compression. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2021.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Math. Oper. Res.*, 8(2): 231–259, 1983.

Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM J. Optim.*, 27(2):927–956, 2017.

Minghan Yang, Andre Milzarek, Zaiwen Wen, and Tong Zhang. A stochastic extra-step quasi-newton method for nonsmooth nonconvex optimization. *Math. Program.*, pages 1–47, 2021.

Peiran Yu, Guoyin Li, and Ting Kei Pong. Kurdyka-Łojasiewicz exponent via inf-projection. *Found. Comput. Math.*, 22(4):1171–1217, 2022.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67(2):301–320, 2005.