Variance Reduced Random Relaxed Projection Method for Constrained Finite-sum Minimization Problems

Zhichun Yang* Fu-quan Xia† Kai Tu‡ Man-Chung Yue§

Abstract. We consider the problem of minimizing a large sum of smooth convex functions subject to a large number of constraints defined by convex functions that are possibly non-smooth. Such a problem finds a wide range of applications in many areas, such as machine learning and signal processing. In this paper, we devise a new random projection method (RPM) for efficiently solving this problem. Compared with existing RPMs, our proposed algorithm features two useful algorithmic ideas. First, at each iteration, instead of projection onto a subset of the feasible region, our algorithm requires only a relaxed projection in the sense that only the projection onto a half-space approximation of the subset is needed. This significantly reduces the per iteration computational cost as the relaxed projection admits a closed-form formula. Second, to exploit the structure that the objective function is a large sum of convex functions, the variance reduction technique is incorporated into our algorithm to improve the empirical convergence behaviour. As our theoretical contributions, under an error bound-type condition and some other standard conditions, we prove that almost surely the proposed algorithm converges to an optimal solution and that both the optimality and feasibility gaps decrease to zero, with rates $\mathcal{O}(1/\sqrt{K})$ and $\mathcal{O}(\log(K)/K)$, respectively. As a side contribution, we also show that the said error bound-type condition holds some mild assumptions, which could be of independent interest. Numerical results on synthetic problem instances are also presented to demonstrate the practical effectiveness of the variance reduction technique and the superiority of our proposed RPM as compared with two existing ones.

Keywords. Constrained Optimization; Finite-Sum Minimization; Random Projection Method; Relaxed Projection; Variance Reduction

*School of Mathematical Science, Sichuan Normal University. E-mail:yangzhichun1994@163.com

†School of Mathematical Science, Sichuan Normal University. E-mail: fuquanxia@163.com

‡Corresponding author. School of Mathematical Science, Laurent Mathematics Center, Sichuan Normal University; Department of Applied Mathematics, The Hong Kong Polytechnic University. E-mail: kaitu_02@163.com

§Musketeers Foundation Institute of Data Science and Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong. Email: mcyue@hku.hk

1. Introduction

This paper considers the following constrained convex optimization problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in C, \end{aligned} \tag{1.1}$$

where

$$C = C_0 \cap \left(\bigcap_{j \in [m]} C_j \right), \quad C_j = \{\boldsymbol{x} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{x}) \leq 0\},$$

and

$$f(\boldsymbol{x}) = \frac{1}{n} \sum_{i \in [n]} f_i(\boldsymbol{x}).$$

Throughout the paper, we assume that $C_0 \subseteq \mathbb{R}^d$ is a non-empty, closed and convex set, that $f_i : \mathbb{R}^d \to \mathbb{R}$ is a differentiable convex function for $i \in [n] = \{1, \cdots, n\}$, and that $\phi_j : \mathbb{R}^d \to \mathbb{R}$ is a convex but possibly non-differentiable function for $j \in [m] = \{1, \cdots, m\}$. Problem (1.1) finds many applications across a wide of areas, including the beamforming problem [31], the constrained LASSO problem [7,9] and the convex regression problem [5,17]. Furthermore, many distributionally robust optimization problems can be equivalently reformulated as a finite-sum minimization problem subject to a large number of constraints, see [15] for example.

A straightforward choice for solving problem (1.1) is the projected gradient method (PGM) whose iteration takes the form

$$\boldsymbol{x}^{k+1} = \Pi_C\big(\boldsymbol{x}^k - \alpha_k \nabla f(\boldsymbol{x}^k)\big),$$

where $\alpha_k > 0$ is the step-size and $\Pi_C(\cdot)$ denotes the projection map onto C. The theory on PGM is rather complete, at least for convex problems. For example, it can be proved that under mild assumptions and a suitable choice of the step-sizes α_k, PGM converges to an optimal solution with a sublinear rate $\mathcal{O}(1/K)$ [3], where K is the total number of iterations. Under stronger assumptions, it is also proved that PGM converges linearly to an optimal solution [24] (*i.e.*, $\mathcal{O}(c^K)$ for some $c \in (0, 1)$).

However, PGM is not a viable algorithm for solving problem (1.1) if

(i) n is large,

(ii) m is large, or

(iii) the projections onto some of the subsets C_j are difficult to compute.

In case of difficulty (i), it is computationally expensive to compute the gradient $\nabla f(\boldsymbol{x}^k)$; and when we have difficulty (ii), difficulty (iii) or both, computing the projection Π_C onto the whole feasible region C is highly computationally demanding, if not impossible.

A standard idea to handle difficulty (i) is to replace the gradient $\nabla f(\boldsymbol{x}^k)$ by the (random) estimator $\nabla f_{i_k}(\boldsymbol{x}^k)$, where i_k is a uniformly random index from $[n]$. The resulting algorithm is known as the stochastic gradient method (SGM), which traces back to the work [27] of Robbins and Monro in the 1951. A natural generalization of SGM is the so-called mini-batch SGM [4], where instead of a single summand function f_i, multiple summand functions are used to form the random gradient estimator. SGM and its variants have become arguably the most popular algorithms for modern, large-scale machine learning. One particular reason comes from its significantly lower per iteration cost, compared to that of PGM. For more details, we refer the readers to the recent survey [25]. The major drawback of SGM and its mini-batch variant is that its gradient estimator tends to introduce a large variance to the algorithm, which necessitates the use of conservative step-size and in turn leads to slow convergence. Indeed, the convergence rate of SGM for minimizing smooth (non-strongly) convex function is only $\mathcal{O}(1/\sqrt{K})$ [23], which is worse than the rate $\mathcal{O}(1/K)$ of the standard PGM based on the exact gradient. Similarly, to minimize a smooth strongly convex function, SGM can only achieve a slower rate of $\mathcal{O}(1/K)$ [4], as compared to the linear convergence rate of the gradient descent.

To remedy this, various variance reduction techniques have been proposed. Notable variants of variance reduced SGM include SAG [16], SAGA [6] and SVRG [13]. Thanks to the variance reduction techniques, the convergence rates of these variants match with that of the gradient descent on both non-strongly and strongly convex problems. Among these variants, the most relevant one to our work is SVRG. SVRG works by dividing iterations into epochs of r iterations, where r is a prescribed positive integer. At the beginning of each epoch, SVRG computes the exact gradient $\nabla f(\tilde{\boldsymbol{x}})$ at the current iterate $\tilde{\boldsymbol{x}}$ and then updates its iterate by using the gradient estimator $\nabla f_{i_k}(\boldsymbol{x}^k) - \nabla f_{i_k}(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})$ for $k \in [r]$, where i_k is again a uniformly random index from $[n]$. The epoch is ended after r iterations (from $k = 1$ to $k = r$), and then $\tilde{\boldsymbol{x}}$ gets updated by setting $\tilde{\boldsymbol{x}} = \boldsymbol{x}^k = \boldsymbol{x}^r$.

To deal with optimization problems with a large number of constraints, *i.e.*, in the presence of difficulty (ii), Wang and Bertsekas [30] developed a random projection method (RPM), which applies to the case where $C_0 = \mathbb{R}^d$ and the objective function is an expectation $f(\boldsymbol{x}) = \mathbb{E}_I[f_I(\boldsymbol{x})]$. In [30], it is assumed that given any $\boldsymbol{x}$, one can access $\nabla f_i(\boldsymbol{x})$ at some realization i of the random variable I, serving as a gradient estimator. The iteration of the RPM in [30] is of the form

$$\boldsymbol{z}^k = \boldsymbol{x}^k - \alpha_k \nabla f_{i_k}(\boldsymbol{x}^k), \quad \boldsymbol{x}^{k+1} = \boldsymbol{z}^k - \beta_k(\boldsymbol{z}^k - \Pi_{C_{j_k}}(\boldsymbol{z}^k)), \tag{1.2}$$

where $\alpha_k > 0$ and $\beta_k \in (0, 2)$ are step-sizes, i_k is a realization of I, and j_k is a uniformly random index from $[m]$. When $\beta_k \equiv 1$, iteration (1.2) simplifies to

$$\boldsymbol{x}^{k+1} = \Pi_{C_{j_k}}(\boldsymbol{x}^k - \alpha_k \nabla f_{i_k}(\boldsymbol{x}^k)),$$

which shows that the new iterate is the projection of the usual stochastic gradient update onto a random subset C_{j_k}. This observation justifies the name of random projection

method. Under certain assumptions, it is proved in [30] that the optimality gap and the feasibility gap both decrease to zero, with rates $\mathcal{O}(1/\sqrt{K})$ and $\mathcal{O}(\log K/K)$, respectively. It should be pointed out that RPMs were first studied in [22]. The work [30] generalized [22] in several aspects.

The purpose of this paper is to develop a new RPM which improves the one in [30] in two regards. First, the optimization problem considered in [30] includes ours as a special case, except that we allow a general C_0. In principle, we could borrow the idea from [30] to handle difficulties (i) and (ii) simultaneously. Nevertheless, when difficulty (iii) is also present, computing the projection $\Pi_{C_{j_k}}$ could be time consuming and hence hinders the efficiency of the algorithm. This motivates us to approximate the subset C_j by a half-space at each iteration. The advantage of such an approximation is that the projection onto a half-space can be computed much more efficiently via an explicit formula. We should emphasize that the idea of approximating complicated feasible region by half-spaces is not new. It has been studied in many other settings and under different names, including the outer approximation method [8] and the cutting-plane method [14].

Second, since the RPM in [30] is based on a stochastic gradient estimator, it similarly suffers from the variance issue discussed above. In view of our previous discussion, it is tempted to replace the gradient estimator $\nabla f_{i_k}(\boldsymbol{x}^k)$ by the SVRG gradient estimator $\nabla f_{i_k}(\boldsymbol{x}^k) - \nabla f_{i_k}(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})$. Unfortunately, this would compromise the theoretical guarantees as the SVRG gradient estimator violates the assumptions required in [30] (see [30, Assumption 1(c)]). Compared with the RPM in [30], our second improvement lies in the incorporation of the SVRG variance reduction technique into the algorithmic framework of [30] while keeping the theoretical convergence guarantees. The nontrivial part is a new analysis which avoids the use of [30, Assumption 1(c)] but still leads the the same convergence rate as the RPM in [30]. Although the use of the SVRG variance reduction technique does not yield a better theoretical convergence rate like in the unconstrained case (which is expected due to the presence of constraints), our experiments showed that it does significantly improve the convergence speed in practice, see Section 4.1.

To summarize our contributions, we developed a new RPM that aims at solving constrained convex optimization problems where the objective is a sum of a large number of differentiable convex functions and the feasible region is defined as the intersection of a large number of complicated convex subsets. The proposed algorithm features two useful algorithmic ideas that can significantly improve the practical performance: variance reduction and half-space approximation of the complicated subsets. To the best of our knowledge, this is the first time these two ideas are simultaneously incorporated into the framework of RPMs. Furthermore, the proposed RPM enjoys rigorous theoretical guarantees. In particular, under standard assumptions (similar to those adopted in [22] and [30]), we proved that the sequence of iterates generated by the proposed RPM converges almost surely to the optimal solution to the constrained convex optimization problem (1.1). Moreover, the convergence rates of the optimality gap and the feasibility gap are $\mathcal{O}(1/\sqrt{K})$

and $\mathcal{O}(\log K/K)$, respectively.

As a side contribution, we formulated and proved an error bound-type condition, see Proposition 3.1. The condition played an instrumental role in the theoretical development of our RPM. We believe that it could be of independent interest in the study of other optimization algorithms, potentially beyond the context of RPMs.

The rest of this paper is organized as follows. We collect the necessary notations and preliminary results in Section 2. In Section 3, we introduce the assumptions and then present and analyze the proposed random projection method. Numerical results are reported in Section 4.

2. Preliminaries

2.1. Notations

We adopt the following notations. We denote by $\mathbb{R}^d$ the d-dimensional Euclidean space. The Euclidean norm and inner product are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. For two sequences of nonnegative scalars $\{a_k\}$ and $\{b_k\}$, we write $a_k = \mathcal{O}(b_k)$ if there exists a constant $c > 0$ such that $a_k \leq c\, b_k$ for any $k \geq 0$. For a finite set S, we denote by $|S|$ the number of elements in S. For a positive integer m, we write $[m] = \{1, \cdots, m\}$. For any convex function ϕ and any $\boldsymbol{x} \in \mathrm{dom}(\phi)$, we denote by $\partial\phi(\boldsymbol{x}) \equiv \{\boldsymbol{\xi} \in \mathbb{R}^d \mid \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) \geq \langle \boldsymbol{\xi}, \boldsymbol{y} - \boldsymbol{x}\rangle,\ \forall \boldsymbol{y} \in \mathbb{R}^d\}$ the subdifferential of ϕ at $\boldsymbol{x}$. The optimal value and the set of optimal solutions of problem (1.1) are denoted by $f^\star$ and $X^\star$, respectively. In other words, $f^\star = \inf_{\boldsymbol{x}\in C} f(\boldsymbol{x})$ and $X^\star = \{\boldsymbol{x} \in C \mid f(\boldsymbol{x}) = f^\star\}$. We abbreviate "almost surely" as "a.s". For a collection $\mathcal{G}$ of random variables, $\mathbb{E}[\cdot|\mathcal{G}]$ denotes the conditional expected value. Finally, we denote its radius by R_D, *i.e.*, $R_D = \sup\{\|\boldsymbol{x}\| \mid \boldsymbol{x} \in D\}$.

Given any subset $D \subseteq \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$, we define the distance $\mathrm{dist}(\boldsymbol{x}, D)$ from $\boldsymbol{x}$ to D as

$$\mathrm{dist}(\boldsymbol{x}, D) = \inf\{\|\boldsymbol{y} - \boldsymbol{x}\| \mid \boldsymbol{y} \in D\}.$$

If D is closed and convex, then the above minimization is attained at a unique minimizer, which is defined as the projection onto the set D:

$$\Pi_D(\boldsymbol{x}) = \arg\min\{\|\boldsymbol{y} - \boldsymbol{x}\| \mid \boldsymbol{y} \in D\}.$$

2.2. Useful Lemmas

We need the following lemmas, see [26] for example.

Lemma 2.1. Let $D \subseteq \mathbb{R}^d$ be a non-empty, closed and convex set. The following hold.

(i) For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\Pi_D(\boldsymbol{x}) - \Pi_D(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$;

(ii) For any $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in D$, $\|\Pi_D(\boldsymbol{x}) - \boldsymbol{y}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \Pi_D(\boldsymbol{x})\|^2$;

(iii) For any $\boldsymbol{x} \in \mathbb{R}^d$, $\mathrm{dist}(\boldsymbol{x}, D) = \|\boldsymbol{x} - \Pi_D(\boldsymbol{x})\|$.

Lemma 2.2. Let $c \in \mathbb{R}$, $\boldsymbol{\zeta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $H = \{\boldsymbol{x} \mid \boldsymbol{\zeta}^T \boldsymbol{x} \leq c\}$. Then

$$\Pi_H(\boldsymbol{u}) = \boldsymbol{u} - \frac{(\boldsymbol{\zeta}^T \boldsymbol{u} - c)_+}{\|\boldsymbol{\zeta}\|^2} \boldsymbol{\zeta},$$

where $(t)_+ = \max\{t, 0\}$ for any $t \in \mathbb{R}$.

The following lemma will be useful to our development, see [22, 28, 30] for example.

Lemma 2.3. Let $\{a^k\}$, $\{u^k\}$, $\{t^k\}$ and $\{d^k\}$ be sequences of nonnegative random variables satisfying

$$\mathbb{E}[a^{k+1} | \mathcal{G}_k] \leq (1 + t^k) a^k - u^k + d^k \quad \text{for all } k \geq 0 \ a.s.,$$

where $\mathcal{G}_k = \{a^0, \cdots, a^k, u^0, \cdots, u^k, t^0, \cdots, t^k, d^0, \cdots, d^k\}$. Suppose that

$$\sum_{k=0}^{\infty} t^k < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} d^k < \infty \qquad a.s.$$

Then,

$$\sum_{k=0}^{\infty} u^k < \infty \qquad a.s., \quad \text{and} \quad \lim_{k \to 0} a^k = a \qquad a.s.,$$

for some non-negative random variable a.

3. Main Results

3.1. Variance Reduced Random Relaxed Projection Method

Motivated by our discussions in the introduction, we propose a new random projection method for solving problem (1.1), namely the variance reduced random relaxed projection method (VR^{3}PM). From now on, given any convex function ϕ, any vector $\boldsymbol{x} \in \mathbb{R}^d$ and any subgradient $\boldsymbol{\xi} \in \partial\phi(\boldsymbol{x})$, we define the outer approximation

$$H(\phi; \boldsymbol{x}; \boldsymbol{\xi}) = \begin{cases} \{\boldsymbol{y} \in \mathbb{R}^d \mid \phi(\boldsymbol{x}) + \langle \boldsymbol{\xi}, \boldsymbol{y} - \boldsymbol{x} \rangle \leq 0\} & \text{if } \boldsymbol{\xi} \neq \mathbf{0}, \\ \mathbb{R}^d & \text{if } \boldsymbol{\xi} = \mathbf{0}. \end{cases} \tag{3.1}$$

Note that in the case of $\boldsymbol{\xi} \neq \mathbf{0}$, $H(\phi; \boldsymbol{x}; \boldsymbol{\xi})$ is a half-space.

Algorithm 1 Variance Reduced Random Relaxed Projection Method (VR^{3}PM)

Input: Initial point $\boldsymbol{x}^0 \in \mathbb{R}^d$, integers $b \geq 1$ and $r \geq 2$, and a positive sequence $\{\alpha_k\}$ of step-sizes.

for $l = 0, 1, 2, \cdots$ **do**

Set $\tilde{\boldsymbol{x}}^l = \boldsymbol{x}^{lr}$.

for $s = 0, 1, \cdots, r-1$ **do**

Set $k = lr + s$. Generate independently and uniformly distributed indices $I_k = \{i_{k1}, \cdots, i_{kb}\} \subseteq [n]$. Compute

$$\boldsymbol{v}^k = \frac{1}{b} \sum_{i \in I_k} (\nabla f_i(\boldsymbol{x}^k) - \nabla f_i(\tilde{\boldsymbol{x}}^l)) + \nabla f(\tilde{\boldsymbol{x}}^l). \tag{3.2}$$

Generate a random index $j_k \in [m]$ and an arbitrary subgradient $\boldsymbol{\xi}^k \in \partial \phi_{j_k}(\boldsymbol{x}^k)$. Compute

$$\boldsymbol{y}^{k+1} = \Pi_{H_k}\big(\boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k\big),$$

where $H_k = H(\phi_{j_k}; \boldsymbol{x}^k; \boldsymbol{\xi}^k)$ (see (3.1)).

Update the next iterate by

$$\boldsymbol{x}^{k+1} = \Pi_{C_0}(\boldsymbol{y}^{k+1}).$$

end

end

Some remarks about VR^{3}PM are in order. First, by the definition of $\partial \phi_{j_k}(\boldsymbol{x}^k)$, $C_{j_k} \subseteq H_k$. Therefore, H_k is an outer approximation of C_{j_k}. This explains why we call our method a random "relaxed" projection method. Second, because of the relaxation in the projection step, the per iteration cost of our method is arguably lower than that of [30]. Indeed, in case of $\boldsymbol{\xi}^k \neq \mathbf{0}$, by Lemma 2.2, the projection step can be computed via a closed-form formula:

$$\Pi_{H_k}\big(\boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k\big) = \boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k - \frac{\big(\phi_{j_k}(\boldsymbol{x}^k) - \alpha_k \langle \boldsymbol{\xi}^k, \boldsymbol{v}^k \rangle\big)_+}{\|\boldsymbol{\xi}^k\|^2} \boldsymbol{\xi}^k.$$

Third, the SVRG gradient estimator (3.2) is used instead of the standard gradient estimator $\nabla f_{i_k}(\boldsymbol{x}^k)$. This helps reducing the variance of the algorithm and thus improves the convergence speed. Fourth, there is a flexibility in the choice of the distribution of the random index j_k. As long as the index distribution satisfies Assumption 3 below, our theory for the algorithm holds. Finally, since the projection onto the intersection of multiple half-spaces can also be easily calculated (by comparing the projections onto the individual half-spaces), we could divide the m constraints of problem (1.1) into $\bar{m}$ groups of constraints, for some $\bar{m} < m$, and re-write the problem into another constrained optimization problem with only $\bar{m}$ constraints using the maximum function. For example, let $\bar{b} > 0$ be a small integer that divides m and define $\bar{m} = m/\bar{b}$. Then, we could equivalently re-write the

feasible region of problem (1.1) as

$$C = C_0 \cap \left(\bigcap_{t \in [\bar{m}]} \bar{C}_t \right),$$

where $\bar{C}_t = \{\boldsymbol{x} \in \mathbb{R}^d \mid \bar{\phi}_t(\boldsymbol{x}) \leq 0\}$ and

$$\bar{\phi}_t(\boldsymbol{x}) \;=\; \max_{j=(t-1)\bar{b}+1,\,\ldots,\,t\bar{b}} \phi_j(\boldsymbol{x}).$$

This grouping technique can be seen as the constraint analogue of the mini-batch gradient estimator and helps to mitigate the variance issue due to the random sampling of constraints. Therefore, despite the increased computation cost of the relaxed projection step, the overall convergence speed and accuracy could be improved with a suitable grouping. We should point out that the RPM in [22] enjoys a similar flexibility as the algorithm requires computing only a subgradient of the grouped function $\bar{\phi}_t$. However, it might not be worth applying constraint grouping to the RPM in [30], as doing so would require computing the projection onto the grouped subset $\bar{C}_t$ at each iteration.

3.2. Assumptions

To analyze VR^{3}PM, the following blanket assumptions on problem (1.1) are imposed.

Assumption 1. The following hold.

(i) For $i \in [n]$, $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, and there exists $L_i > 0$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L_i(\|\boldsymbol{x} - \boldsymbol{y}\| + 1).$$

(ii) The set $C_0 \subseteq \mathbb{R}^d$ is non-empty, closed and convex. Moreover, for $j \in [m]$, the function ϕ_j is proper, closed and convex.

(iii) The optimal solution set $X^\star$ of problem (1.1) is non-empty.

Assumption 1(i) is a special case of [30, Assumption 1(b)] and is weaker than the requirement on f_i in [22] which assumes that each f_i has a Lipschitz continuous gradient. Indeed, it can be easily checked that Assumption 1(i) holds if for any $i \in [n]$, either the function f_i or its gradient ∇f_i is Lipschitz continuous. Assumptions 1(ii)-(iii) are standard in the literature of constrained convex optimization problem.

Assumption 2. There exists $\kappa > 0$ such that for any $\boldsymbol{x} \in C_0$,

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \kappa \max_{j \in [m]} \min_{\boldsymbol{\xi}_j \in \partial \phi_j(\boldsymbol{x})} \operatorname{dist}\left(\boldsymbol{x},\, H(\phi_j;\, \boldsymbol{x};\, \boldsymbol{\xi}_j)\right). \tag{3.3}$$

Assumption 2 is a linear regularity-type condition [2], which is instrumental to our convergence analysis. It can also be seen as a generalization of the seminal Hoffman error bound condition [12], which asserts that the distance of any point to a linear system is linearly bounded by its violation of the linear constraints. Indeed, if $C_0 = \mathbb{R}^d$ and all the constraint functions ϕ_j are affine, then inequality (3.3) reduces to the Hoffman error bound condition. Linear regularity-type and error bound-type conditions are themselves important subjects in optimization and have been frequently utilized to study various optimization algorithms. For example, in both [22] and [30], conditions similar to (3.3) are employed to study their RPMs. Going beyond RPMs, linear regularity-type and error bound-type conditions also appeared in the study of first-order methods [20, 21], second-order methods [7, 32, 33], and even manifold optimization algorithms [18, 19].

The next proposition asserts that Assumption 2 holds under mild conditions. We should emphasize that case (i) is not new, see [1, 11, 12] for example. The novelty of the proposition lies in case (ii).

Proposition 3.1. Suppose that Assumption 1 holds. Then, Assumption 2 holds if either of the following is true:

(i) $C_0 = \mathbb{R}^d$ and for all $j \in [m]$, C_j is half-space;

(ii) C_0 is compact and $C_0 \cap \{\boldsymbol{x} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{x}) < 0,\ j \in [m]\}$ is non-empty.

Proof. The proof of case (i) can be founded in [1, 11]. We therefore prove only case (ii). If $\boldsymbol{x} \in C$, then (3.3) trivially holds. It suffices to assume that $\boldsymbol{x} \notin C$. Let $I(\boldsymbol{x})$ be the set defined by

$$I(\boldsymbol{x}) = \{j \in [m] \mid \phi_j(\boldsymbol{x}) > 0\}. \tag{3.4}$$

Since $\boldsymbol{x} \in C_0$, we have that $\boldsymbol{x} \notin \cap_{j\in[m]} C_j$ and hence the index set $I(\boldsymbol{x})$ is non-empty. From Lemma 3.1, it follows that there exists a constant $\gamma > 0$ such that

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \gamma \phi_{j'}(\boldsymbol{x}), \tag{3.5}$$

where $j' \in \arg\max_j \{\phi_j(\boldsymbol{x}) \mid j \in [m]\}$ and γ does not depend on $\boldsymbol{x}$. Clearly, $j' \in I(\boldsymbol{x})$ and hence $\phi_{j'}(\boldsymbol{x}) = \max_{j\in I(\boldsymbol{x})}\{\phi_j(\boldsymbol{x})\}$. By [3, Theorem 3.16] and the assumption that C_0 is compact, there exists a constant $\eta > 0$ such that for any $j \in [m]$, $\boldsymbol{x} \in C_0$ and $\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x})$, we have $\|\boldsymbol{\xi}_j\| \leq \eta$. Fix an arbitrary $\boldsymbol{\xi}_{j'} \in \partial\phi_{j'}(\boldsymbol{x})$. If $\boldsymbol{\xi}_{j'} \neq \boldsymbol{0}$,

$$\phi_{j'}(\boldsymbol{x}) \leq -\langle \boldsymbol{\xi}_{j'}, \Pi_{H(\phi_{j'}; \boldsymbol{x}; \boldsymbol{\xi}_{j'}))}(\boldsymbol{x}) - \boldsymbol{x}\rangle \leq \eta \operatorname{dist}(\boldsymbol{x},\ H(\phi_{j'}; \boldsymbol{x}; \boldsymbol{\xi}_{j'})),$$

where the first inequality follows from the definition (3.1) of $H(\phi_{j'}; \boldsymbol{x}; \boldsymbol{\xi}_{j'})$ and the fact that $\Pi_{H(\phi_{j'}; \boldsymbol{x}; \boldsymbol{\xi}_{j'}))}(\boldsymbol{x}) \in H(\phi_{j'}; \boldsymbol{x}; \boldsymbol{\xi}_{j'})$, and the second inequality follows from the Cauchy-Schwarz inequality and the uniform bound of the subdifferential $\phi_j(\boldsymbol{x})$. If $\boldsymbol{\xi}_{j'} = \boldsymbol{0}$, by convexity, $\boldsymbol{x}$ is a minimizer of $\phi_{j'}$ and

$$\phi_{j'}(\boldsymbol{x}) = \min_{\boldsymbol{y}\in\mathbb{R}^d} \phi_{j'}(\boldsymbol{y}) \leq 0,$$

where the inequality follows from the supposition that $\{\boldsymbol{y} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{y}) < 0,\ j \in [m]\}$ is non-empty. However, by the definition of j', $\phi_{j'}(\boldsymbol{x}) > 0$. Hence, this is a contradiction, and it is impossible to have $\boldsymbol{\xi}_{j'} = \boldsymbol{0}$. Therefore, for any subgradient $\boldsymbol{\xi}_{j'} \in \partial\phi_{j'}(\boldsymbol{x})$,

$$\phi_{j'}(\boldsymbol{x}) \leq \eta \operatorname{dist}(\boldsymbol{x},\ H(\phi_{j'};\, \boldsymbol{x};\, \boldsymbol{\xi}_{j'})).$$

Minimizing the right-hand side over $\boldsymbol{\xi}_{j'}$, we have

$$\phi_{j'}(\boldsymbol{x}) \leq \eta \min_{\boldsymbol{\xi}_{j'} \in \partial\phi_{j'}(\boldsymbol{x})} \operatorname{dist}(\boldsymbol{x}, H(\phi_{j'};\, \boldsymbol{x};\, \boldsymbol{\xi}_{j'})),$$

which, together with (3.5), implies that

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \gamma\eta \min_{\boldsymbol{\xi}_{j'} \in \partial\phi_{j'}(\boldsymbol{x})} \operatorname{dist}(\boldsymbol{x}, H(\phi_{j'};\, \boldsymbol{x};\, \boldsymbol{\xi}_{j'})) \leq \gamma\eta \max_{j \in [m]} \min_{\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x})} \operatorname{dist}(\boldsymbol{x},\ H(\phi_j;\, \boldsymbol{x};\, \boldsymbol{\xi}_j)).$$

Noting that both constants γ and η are independent of $\boldsymbol{x}$, this completes the proof. □

The following lemma is used in the proof of Proposition 3.1.

Lemma 3.1. Suppose that Assumption 1 holds, that the set $C_0 \subseteq \mathbb{R}^d$ is compact and that the set $C_0 \cap \{\boldsymbol{x} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{x}) < 0,\ j \in [m]\}$ is non-empty. Then there exists a constant $\gamma > 0$ such that for any $\boldsymbol{x} \in C_0 \setminus \cap_{j \in [m]} C_j$,

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \gamma\phi_{j'}(\boldsymbol{x}),$$

where $j' \in \arg\max_j \{\phi_j(\boldsymbol{x}) \mid j \in [m]\}$.

Proof. Let $\boldsymbol{x} \in C_0 \setminus \cap_{j \in [m]} C_j$. Then, $I(\boldsymbol{x})$ is non-empty (see (3.4) for the definition of $I(\boldsymbol{x})$). Pick any

$$j' \in \arg\max_{j \in I(\boldsymbol{x})} \phi_j(\boldsymbol{x}).$$

In particular, we have

$$\phi_{j'}(\boldsymbol{x}) = \max_{j \in I(\boldsymbol{x})} \phi_j(\boldsymbol{x}) > 0.$$

Since $C_0 \cap \{\boldsymbol{x} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{x}) < 0,\ j \in [m]\}$ is non-empty, there exists a point $\boldsymbol{z} \in C_0$ such that $\phi_j(\boldsymbol{z}) < 0$ for all $j \in [m]$. Therefore, there exists a constant $\theta > 0$ satisfying

$$\theta \min_{j \in [m]} |\phi_j(\boldsymbol{z})| \geq \max_{j \in [m]} |\phi_j(\boldsymbol{z})|.$$

Clearly, the constant θ does not depend on $\boldsymbol{x}$, and for any $j \in I(\boldsymbol{x})$, $\theta|\phi_j(\boldsymbol{z})| \geq |\phi_{j'}(\boldsymbol{z})|$.

We claim that for any $j \in [m]$,

$$\theta\phi_{j'}(\boldsymbol{x})\phi_j(\boldsymbol{z}) - \phi_{j'}(\boldsymbol{z})\phi_j(\boldsymbol{x}) \leq 0. \tag{3.6}$$

We consider the two cases $j \in I(\boldsymbol{x})$ and $j \in [m] \setminus I(\boldsymbol{x})$ separately. For $j \in I(\boldsymbol{x})$, since $\phi_{j'}(\boldsymbol{x}) \geq \phi_j(\boldsymbol{x})$ and $\theta|\phi_j(\boldsymbol{z})| \geq |\phi_{j'}(\boldsymbol{z})|$, we get $\theta\phi_{j'}(\boldsymbol{x})\phi_j(\boldsymbol{z}) - \phi_{j'}(\boldsymbol{z})\phi_j(\boldsymbol{x}) \leq 0$. For

$j \in [m]\backslash I(\boldsymbol{x})$, it follows from $\phi_{j'}(\boldsymbol{x}) > 0$, $\phi_j(\boldsymbol{z}) < 0$, $\phi_{j'}(\boldsymbol{z}) < 0$ and $\phi_j(\boldsymbol{x}) \leq 0$ that $\theta\phi_{j'}(\boldsymbol{x})\phi_j(\boldsymbol{z}) - \phi_{j'}(\boldsymbol{z})\phi_j(\boldsymbol{x}) \leq 0$. This proves the claim.

We are ready to prove the assertion of the lemma. By the construction of j' and $\boldsymbol{z}$, we know that $\phi_{j'}(\boldsymbol{x}) > 0$, $\phi_{j'}(\boldsymbol{z}) < 0$ and hence $\theta\phi_{j'}(\boldsymbol{x}) - \phi_{j'}(\boldsymbol{z}) > 0$. We can define the point $\boldsymbol{y}$ by

$$\boldsymbol{y} = \frac{\theta\phi_{j'}(\boldsymbol{x})}{\theta\phi_{j'}(\boldsymbol{x}) - \phi_{j'}(\boldsymbol{z})}\boldsymbol{z} + \frac{-\phi_{j'}(\boldsymbol{z})}{\theta\phi_{j'}(\boldsymbol{x}) - \phi_{j'}(\boldsymbol{z})}\boldsymbol{x}.$$

The convexity of C_0 implies that $\boldsymbol{y} \in C_0$. Also, the convexity of $\phi_j(\cdot)$ and inequality (3.6) together yield that $\phi_j(\boldsymbol{y}) \leq 0$ for all $j \in [m]$. Hence, $\boldsymbol{y} \in C$. We thus have

$$\begin{aligned}\operatorname{dist}(\boldsymbol{x}, C) \leq \|\boldsymbol{x} - \boldsymbol{y}\| &= \frac{\theta\phi_{j'}(\boldsymbol{x})}{\theta\phi_{j'}(\boldsymbol{x}) - \phi_{j'}(\boldsymbol{z})}\|\boldsymbol{x} - \boldsymbol{z}\| \\ &\leq \frac{2\theta R_{C_0}}{\min_{j\in[m]}\{-\phi_j(\boldsymbol{z})\}}\phi_{j'}(\boldsymbol{x}),\end{aligned}$$

where the radius R_{C_0} is finite since C_0 is bounded, the first inequality follows from the fact that $\boldsymbol{y} \in C$, the equality follows from the definition of $\boldsymbol{y}$, and the second inequality follows from the fact that $\theta\phi_{j'}(\boldsymbol{x}) - \phi_{j'}(\boldsymbol{z}) > -\phi_{j'}(\boldsymbol{z}) > 0$. Since the constant $\frac{2\theta R_{C_0}}{\min_{j\in[m]}\{-\phi_j(\boldsymbol{z})\}}$ is independent of $\boldsymbol{x}$, the proof is completed. □

We also need an assumption concerning the distribution of the constraint index j_k. A similar assumption is also imposed in [30].

Assumption 3. There exists a constant $\rho \in (0, 1]$ such that for any $j \in [m]$,

$$\inf_{k\geq 0} P(j_k = j|\mathcal{F}_k) \geq \frac{\rho}{m} \quad a.s.,$$

where $\mathcal{F}_k = \{j_0, \cdots, j_{k-1}, i_{01}, \cdots i_{0b}, \cdots, i_{(k-1)1}, \cdots, i_{(k-1)b}, \boldsymbol{x}^0\}$ for $k \geq 1$ and $\mathcal{F}_0 = \{\boldsymbol{x}^0\}$.

Assumption 3 ensures in particular that at each iteration, any constraint will be picked with a positive probability independent of the iteration counter k.

3.3. Convergence Analysis

We now present the main theoretical results of this paper, the proofs of which are deferred to Section 3.4.

The first one shows that VR^{3}PM converges to an optimal solution of problem (1.1) almost surely under Assumptions 1, 2 and 3 and a suitable choice for the step-size α_k.

Theorem 3.1. Suppose that Assumptions 1, 2 and 3 hold. Let $\{\delta_l\}$ be a positive sequence satisfying

$$\sum_{l=0}^{\infty}\delta_l = \infty \quad \text{and} \quad \sum_{l=0}^{\infty}\delta_l^2 < \infty.$$

Consider Algorithm 1 with $\alpha_k = \delta_l \in (0, 1)$ for $k = lr + s - 1$ with $l \geq 0$ and $s \in [r]$. Then, $\{\boldsymbol{x}^k\}$ converges almost surely to a point in $X^\star$ and $f(\boldsymbol{x}^k)$ converges almost surely to $f^\star$.

Our second main theoretical result characterizes the convergence rates of the optimality gap and feasibility gap of VR^{3}PM under Assumptions 1, 2 and 3.

Theorem 3.2. Suppose that Assumptions 1, 2 and 3 hold and that C_0 is compact. Consider Algorithm 1 with $\boldsymbol{x}^0 \in C_0$, $\alpha_k = \frac{\tilde{\alpha}_0}{\sqrt{k+1}}$, $\tilde{\alpha}_0 \in (0, \frac{\rho}{16Lm\kappa^2}]$, where $L > 0$, $\kappa > 0$ and $\rho > 0$ are constants defined in (3.8), Assumption 2 and Assumption 3, respectively. Then, we have that for integer $K \geq 1$,

$$\mathbb{E}[\mathrm{dist}^2(\bar{\boldsymbol{x}}^K, C)] \leq \mathcal{O}\left(\frac{\log(K)}{K}\right),$$

and that

$$\mathbb{E}[f(\bar{\boldsymbol{x}}^K) - f^\star] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

where $\bar{\boldsymbol{x}}^K = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{x}^k$.

In order for VR^{3}PM to enjoy the convergence rates in Theorem 3.2, the step-size α_k should be set based on the parameters $\kappa > 0$ and $\rho > 0$, which are possibly unknown in practice. In our numerical experiment, we set the step-size according to $\alpha_k = \mathcal{O}(k^{-(0.5+\nu)})$ for some small number $\nu > 0$, see Section 4 for details. We should also point out that such an issue is common in optimization algorithms developed based on linear regularity-type or error bound-type conditions [7, 22, 30, 33].

3.4. Proofs of Theorem 3.1 and Theorem 3.2

Our goal here is prove Theorem 3.1 and Theorem 3.2. Towards that end, we need to prepare several technical results.

The following lemma bounds the distance of an iterate to the feasible region in terms of its distance to the corresponding half-space at that iteration.

Lemma 3.2. Suppose that Assumptions 2 and 3 hold. Then, Algorithm 1 satisfies that for any $k \geq 0$,

$$\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \leq m\kappa^2\rho^{-1}\,\mathbb{E}[\|\boldsymbol{x}^k - \Pi_{H_k}(\boldsymbol{x}^k)\|^2|\mathcal{F}_k] \quad a.s.$$

Proof. We have that for any $j \in [m]$,

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^k - \Pi_{H_k}(\boldsymbol{x}^k)\|^2|\mathcal{F}_k] = \sum_{j'\in[m]} P(j_k = j'|\mathcal{F}_k)\|\boldsymbol{x}^k - \Pi_{H(\phi_{j'};\boldsymbol{x}^k;\boldsymbol{\xi}_{j'})}(\boldsymbol{x}^k)\|^2 \\
\geq\ &\frac{\rho}{m}\|\boldsymbol{x}^k - \Pi_{H(\phi_j;\boldsymbol{x}^k;\boldsymbol{\xi}^k))}(\boldsymbol{x}^k)\|^2 \geq \frac{\rho}{m}\min_{\boldsymbol{\xi}_j\in\partial\phi_j(\boldsymbol{x}^k)}\|\boldsymbol{x}^k - \Pi_{H(\phi_j;\boldsymbol{x}^k;\boldsymbol{\xi}_j))}(\boldsymbol{x}^k)\|^2 \\
=\ &\frac{\rho}{m}\min_{\boldsymbol{\xi}_j\in\partial\phi_j(\boldsymbol{x}^k)}\mathrm{dist}(\boldsymbol{x}^k, H(\phi_j;\,\boldsymbol{x}^k;\,\boldsymbol{\xi}_j))^2,
\end{aligned}$$

where the first equality follows from the definition of H_k, and the first inequality follows from Assumption 3. Therefore, using Assumption 2 and $\boldsymbol{x}^k \in C_0$, we obtain

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}^k - \Pi_{H_k}(\boldsymbol{x}^k)\|^2 | \mathcal{F}_k] &\geq \frac{\rho}{m} \max_{j\in[m]} \min_{\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x}^k)} \operatorname{dist}(\boldsymbol{x}^k, H(\phi_j;\, \boldsymbol{x}^k;\, \boldsymbol{\xi}_j))^2 \\ &\geq \frac{\rho}{m\kappa^2}\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2,\end{aligned}$$

which completes the proof. □

Here, we record a useful consequence of Assumption 1(i) is that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L(\|\boldsymbol{x} - \boldsymbol{y}\| + 1) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \tag{3.7}$$

where

$$L = \sqrt{\frac{1}{n}\sum_{i\in[n]} L_i^2}. \tag{3.8}$$

The next lemma should be compared with [30, Assumption 1(c)] and, to some extent, illustrates why the SVRG gradient estimator cannot be directly used in the algorithmic framework of [30].

Lemma 3.3. Suppose that Assumption 1 holds. Consider Algorithm 1. Then, for any $\boldsymbol{x}^\star \in X^\star$ and $k = lr + s - 1$, where $l \geq 0$, $s \in [r]$, we have

$$\mathbb{E}[\|\boldsymbol{v}^k\|^2 \mid \mathcal{F}_k] \leq 8L^2\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + 16L^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + 12L^2 + 4\|\nabla f(\boldsymbol{x}^\star)\|^2.$$

Proof. We have

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{v}^k\|^2 \mid \mathcal{F}_k] \\ =&\,\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_k}(\nabla f_i(\boldsymbol{x}^k) - \nabla f_i(\tilde{\boldsymbol{x}}^l)) + \nabla f(\tilde{\boldsymbol{x}}^l)\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\ \leq&\, 2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_k}(\nabla f_i(\boldsymbol{x}^k) - \nabla f_i(\tilde{\boldsymbol{x}}^l))\right\|^2 \,\middle|\, \mathcal{F}_k\right] + 2\mathbb{E}[\|\nabla f(\tilde{\boldsymbol{x}}^l))\|^2 \mid \mathcal{F}_k] \\ \leq&\, 2(2L^2\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}^l\|^2 + 2L^2) + 2(2\|\nabla f(\tilde{\boldsymbol{x}}^l) - \nabla f(\boldsymbol{x}^\star)\|^2 + 2\|\nabla f(\boldsymbol{x}^\star)\|^2) \\ \leq&\, 8L^2\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + 16L^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + 12L^2 + 4\|\nabla f(\boldsymbol{x}^\star)\|^2,\end{aligned}$$

where the equality follows from the definition of $\boldsymbol{v}^k$, and the second inequality follows from inequality (3.7). This completes the proof. □

The following technical lemma will also be useful.

Lemma 3.4. Suppose that Assumption 1 holds. Consider Algorithm 1. Then, for any $\boldsymbol{x}^\star \in X^\star$, $k \geq 0$ and $\lambda > 0$, we have

$$\begin{aligned} 2\alpha_k \mathbb{E}[\langle \boldsymbol{v}^k, \boldsymbol{x}^k - \boldsymbol{x}^\star \rangle \mid \mathcal{F}_k] \geq & -(2\alpha_k L + \tfrac{4}{\lambda})\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - \alpha_k^2 L^2 \lambda \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 \\ & - \alpha_k^2 \lambda (2L^2 + \|\nabla f(\boldsymbol{x}^\star)\|^2) + 2\alpha_k \Big(f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star) \Big). \end{aligned}$$

Proof. First, for any $k \geq 0$,

$$\begin{aligned} & 2\alpha_k \mathbb{E}[\langle \boldsymbol{v}^k,\, \boldsymbol{x}^k - \boldsymbol{x}^\star \rangle \mid \mathcal{F}_k] = 2\alpha_k \langle \nabla f(\boldsymbol{x}^k),\, \boldsymbol{x}^k - \boldsymbol{x}^\star \rangle \\ \geq\, & 2\alpha_k (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)) = 2\alpha_k \Big(f(\boldsymbol{x}^k) - f(\Pi_C(\boldsymbol{x}^k)) + f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star) \Big) \\ \geq\, & 2\alpha_k \Big(\langle \nabla f(\Pi_C(\boldsymbol{x}^k)),\, \boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k) \rangle + f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star) \Big), \end{aligned} \tag{3.9}$$

where the first equality follows from the definitions of $\boldsymbol{v}^k$, the random index subset I_k and the collection $\mathcal{F}_k$, the first inequality from the convexity of f, and the second inequality from the convexity of f. Also, we have

$$\begin{aligned} & 2\alpha_k \langle \nabla f(\Pi_C(\boldsymbol{x}^k)),\, \boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k) \rangle \\ \geq\, & -2\alpha_k \Big(\|\nabla f(\Pi_C(\boldsymbol{x}^k)) - \nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| + \|\nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \Big) \\ \geq\, & -2\alpha_k \Big(L(\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| + 1)\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| + \|\nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \Big) \\ \geq\, & -2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\ & - 2\alpha_k \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| - 2\alpha_k \|\nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|, \end{aligned} \tag{3.10}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second from inequality (3.7), and the third from the triangle inequality. We then bound the second, third and fourth terms on the last line of (3.10). We will use multiple times the fact that $2|a_1 a_2| \leq \lambda a_1^2 + \frac{1}{\lambda} a_2^2$ for all $a_1, a_2 \in \mathbb{R}$ and $\lambda > 0$. The second term can be bounded as

$$-2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \geq -\alpha_k^2 L^2 \lambda - \frac{1}{\lambda} \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2. \tag{3.11}$$

The third term can be bounded as

$$\begin{aligned} & -2\alpha_k \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\ \geq\, & -2\alpha_k L(\|\boldsymbol{x}^k - \boldsymbol{x}^\star\| + 1)\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\ \geq\, & -\alpha_k^2 L^2 \lambda \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{1}{\lambda} \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|, \\ \geq\, & -\alpha_k^2 L^2 \lambda \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{2}{\lambda} \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - \alpha_k^2 L^2 \lambda, \end{aligned} \tag{3.12}$$

where the last inequality follows from (3.11). And the fourth term can be bounded as

$$-2\alpha_k \|\nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \geq -\alpha_k^2 \lambda \|\nabla f(\boldsymbol{x}^\star)\|^2 - \frac{1}{\lambda} \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2. \tag{3.13}$$

Substituting inequalities (3.10), (3.11), (3.12) and (3.13) into (3.9) yields the desired result. □

The next proposition establishes a recursion for the distance to optimality $\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|$.

Proposition 3.2. Suppose that Assumptions 1, 2 and 3 hold. Consider Algorithm 1. Then, for any $\lambda > 0$, $\boldsymbol{x}^\star \in X^\star$ and $k = lr + s - 1$, where $l \geq 0$, $s \in [r]$, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k] \\
\leq &(1 + \alpha_k^2(24L^2 + L^2\lambda))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + 48L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 \\
&+ \alpha_k^2\left(2\lambda L^2 + (\lambda + 12)\|\nabla f(\boldsymbol{x}^\star)\|^2 + 36L^2\right) \\
&- (\tfrac{\rho}{2m\kappa^2} - 2\alpha_k L - \tfrac{4}{\lambda})\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star)).
\end{aligned}$$

Proof. Since $\boldsymbol{x}^\star \in X^\star \subseteq C$, we have $\boldsymbol{x}^\star \in C_0$ and $\boldsymbol{x}^\star \in C_{j_k} \subseteq H_k$. Let $\boldsymbol{z}^k = \boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k$. It follows that

$$\begin{aligned}
&\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{y}^{k+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k - \boldsymbol{x}^\star\|^2 - \|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \\
= &\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \alpha_k^2\|\boldsymbol{v}^k\|^2 - 2\alpha_k\langle \boldsymbol{v}^k, \boldsymbol{x}^k - \boldsymbol{x}^\star\rangle - \|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2,
\end{aligned}$$

where the first inequality follows from the definition of $\boldsymbol{y}^{k+1}$, Lemma 2.1(i) and the fact that $\boldsymbol{x}^\star \in C_0$, and the second inequality from the definition of $\boldsymbol{z}^k$, Lemma 2.1(ii) and the fact that $\boldsymbol{x}^\star \in H_k$. Taking conditional expectation on the last inequality and using Lemmas 3.3 and 3.4 yields,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] \\
\leq &\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathbb{E}\left[\alpha_k^2\|\boldsymbol{v}^k\|^2 - 2\alpha_k\langle \boldsymbol{v}^k, \boldsymbol{x}^k - \boldsymbol{x}^\star\rangle - \|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k\right] \\
\leq &(1 + \alpha_k^2(8L^2 + L^2\lambda))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + (2\alpha_k L + \tfrac{4}{\lambda})\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \\
&+ \alpha_k^2(2\lambda L^2 + \lambda\|\nabla f(\boldsymbol{x}^\star)\|^2 + 4\|\nabla f(\boldsymbol{x}^\star)\|^2 + 12L^2) + 16L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 \\
&- 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star)) - \mathbb{E}\left[\|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k\right].
\end{aligned} \tag{3.14}$$

We then bound the term $\mathbb{E}[\|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k]$ in last line of (3.14). First, from the triangle inequality and Lemma 2.1(i), we have that

$$\begin{aligned}
\|\Pi_{H_k}(\boldsymbol{x}^k) - \boldsymbol{x}^k\| &\leq \|\Pi_{H_k}(\boldsymbol{x}^k) - \Pi_{H_k}(\boldsymbol{z}^k)\| + \|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\| + \|\boldsymbol{z}^k - \boldsymbol{x}^k\| \\
&\leq \|\boldsymbol{z}^k - \Pi_{H_k}(\boldsymbol{z}^k)\| + 2\|\boldsymbol{x}^k - \boldsymbol{z}^k\| \\
&\leq \|\boldsymbol{z}^k - \Pi_{H_k}(\boldsymbol{z}^k)\| + 2\alpha_k\|\boldsymbol{v}^k\|,
\end{aligned}$$

which together with Lemma 3.3 yields that

$$\begin{aligned}
&\mathbb{E}[\|\Pi_{H_k}(\boldsymbol{x}^k) - \boldsymbol{x}^k\|^2 \mid \mathcal{F}_k] \\
\leq &2\mathbb{E}[\|\boldsymbol{z}^k - \Pi_{H_k}(\boldsymbol{z}^k)\|^2 \mid \mathcal{F}_k] + 4\alpha_k^2\,\mathbb{E}[\|\boldsymbol{v}^k\|^2 \mid \mathcal{F}_k] \\
\leq &2\mathbb{E}[\|\boldsymbol{z}^k - \Pi_{H_k}(\boldsymbol{z}^k)\|^2 \mid \mathcal{F}_k] + 32L^2\alpha_k^2\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + 64L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 \\
&+ \alpha_k^2(48L^2 + 16\|\nabla f(\boldsymbol{x}^\star)\|^2).
\end{aligned}$$

Rearranging the above inequality gives

$$\begin{aligned}
&-\mathbb{E}[\|\boldsymbol{z}^k-\Pi_{H_k}(\boldsymbol{z}^k)\|^2\mid\mathcal{F}_k]\\
\le&\,16L^2\alpha_k^2\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2+32L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\alpha_k^2(24L^2+8\|\nabla f(\boldsymbol{x}^\star)\|^2)\\
&-\frac{1}{2}\mathbb{E}[\|\Pi_{H_k}(\boldsymbol{x}^k)-\boldsymbol{x}^k\|^2\mid\mathcal{F}_k].
\end{aligned}\tag{3.15}$$

Plugging (3.15) into (3.14) and using Lemma 3.2, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^\star\|^2\mid\mathcal{F}_k]\\
\le&\,(1+\alpha_k^2(24L^2+L^2\lambda))\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2+48L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+(2\alpha_kL+\tfrac{4}{\lambda})\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2\\
&-2\alpha_k(f(\Pi_C(\boldsymbol{x}^k))-f(\boldsymbol{x}^\star))+\alpha_k^2(2\lambda L^2+(\lambda+12)\|\nabla f(\boldsymbol{x}^\star)\|^2+36L^2)\\
&-\frac{1}{2}\mathbb{E}[\|\Pi_{H_k}(\boldsymbol{x}^k)-\boldsymbol{x}^k\|^2\mid\mathcal{F}_k]\\
\le&\,\left(1+\alpha_k^2(24L^2+L^2\lambda)\right)\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2+48L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2\\
&+\alpha_k^2\left(2\lambda L^2+(\lambda+12)\|\nabla f(\boldsymbol{x}^\star)\|^2+36L^2\right)\\
&-(\tfrac{\rho}{2m\kappa^2}-2\alpha_kL-\tfrac{4}{\lambda})\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2-2\alpha_k(f(\Pi_C(\boldsymbol{x}^k))-f(\boldsymbol{x}^\star)).
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.1. We first prove that the sub-sequence $\{\tilde{\boldsymbol{x}}^l\}$ converges. Since $\delta_l\to 0$ as $l\to\infty$, there exists $l_0\ge 0$ such that $2\delta_lL\le\frac{\rho}{8m\kappa^2}$ for any $l\ge l_0$. Take $\lambda=32m\kappa^2\rho^{-1}$. Then, for any for any $l\ge l_0$, we have

$$\frac{\rho}{2m\kappa^2}-2\delta_lL-\frac{4}{\lambda}\ge\frac{\rho}{4m\kappa^2}>0.$$

Fix an arbitrary optimal solution $\boldsymbol{x}^\star\in X^\star$. It follows from Proposition 3.2 that for all $l\ge l_0$,

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^\star\|^2\mid\mathcal{F}_k]\\
\le&\,(1+\mathcal{O}(\alpha_k^2))\,\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2+\mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\mathcal{O}(\alpha_k^2)\\
&-\frac{\rho}{4m\kappa^2}\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2-2\delta_l(f(\Pi_C(\boldsymbol{x}^k))-f(\boldsymbol{x}^\star)).
\end{aligned}\tag{3.16}$$

Note that the constants hidden in the big-O notations $\mathcal{O}(\alpha_k^2)$ are all independent of α_k or k. In particular, for any $k=lr+s-1$, where $l\ge l_0$ and $s\in[r]$, inequality (3.16) implies that

$$\mathbb{E}[\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^\star\|^2\mid\mathcal{F}_k]\le(1+\mathcal{O}(\delta_l^2))\,\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2+\mathcal{O}(\delta_l^2)\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\mathcal{O}(\delta_l^2).\tag{3.17}$$

Using the tower property of conditional expectation (see [29, Theorem 2.3.2(iii)] for example) and inequality (3.17) (with $k=lr+r-1$), we have that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{lr+r}-\boldsymbol{x}^\star\|^2\mid\mathcal{F}_{lr}\right]=\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{x}^{lr+r-1+1}-\boldsymbol{x}^\star\|^2\mid\mathcal{F}_{lr+r-1}\right]\Big|\mathcal{F}_{lr}\right]\\
\le&\,(1+\mathcal{O}(\delta_l^2))\mathbb{E}\left[\|\boldsymbol{x}^{lr+r-1}-\boldsymbol{x}^\star\|^2\Big|\mathcal{F}_{lr}\right]+\mathcal{O}(\delta_l^2)\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\mathcal{O}(\delta_l^2).
\end{aligned}$$

Repeating the same argument, inductively, we get

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\boldsymbol{x}}^{(l+1)} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] = \mathbb{E}[\|\boldsymbol{x}^{(l+1)r} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
\leq &(1 + \mathcal{O}(\delta_l^2))^{r-1}\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
&+ \Big(\mathcal{O}(\delta_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\delta_l^2)\Big)\sum_{s=0}^{r-2}(1 + \mathcal{O}(\delta_l^2))^s.
\end{aligned} \tag{3.18}$$

For the term $\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}]$, we use inequality (3.16) instead of inequality (3.17) and get

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
\leq &(1 + \mathcal{O}(\delta_l^2))\,\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\delta_l^2) \\
&- \frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l - \Pi_C(\tilde{\boldsymbol{x}}^l)\|^2 - 2\delta_l(f(\Pi_C(\tilde{\boldsymbol{x}}^l)) - f(\boldsymbol{x}^\star)).
\end{aligned} \tag{3.19}$$

We thus have

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\boldsymbol{x}}^{(l+1)} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
\leq &(1 + \mathcal{O}(\delta_l^2))^r\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
&+ \Big(\mathcal{O}(\delta_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\delta_l^2)\Big)\sum_{s=0}^{r-1}(1 + \mathcal{O}(\delta_l^2))^s \\
&- \frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l - \Pi_C(\tilde{\boldsymbol{x}}^l)\|^2 - 2\delta_l(f(\Pi_C(\tilde{\boldsymbol{x}}^l)) - f(\boldsymbol{x}^\star)) \\
\leq &(1 + \mathcal{O}(\delta_l^2))\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] + \mathcal{O}(\delta_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\delta_l^2) \\
&- \frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l - \Pi_C(\tilde{\boldsymbol{x}}^l)\|^2 - 2\delta_l(f(\Pi_C(\tilde{\boldsymbol{x}}^l)) - f(\boldsymbol{x}^\star)) \\
\leq &(1 + \mathcal{O}(\delta_l^2))\,\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 - \frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l - \Pi_C(\tilde{\boldsymbol{x}}^l)\|^2 + \mathcal{O}(\delta_l^2),
\end{aligned} \tag{3.20}$$

where the first inequality follows by substituting inequality (3.19) into inequality (3.18), the second inequality follows from the fact that $(1 + \mathcal{O}(\delta_l^2))^s \leq 1 + \mathcal{O}(\delta_l^2)$ for any $s \in [r]$, and the third inequality from the fact that $f(\Pi_C(\tilde{\boldsymbol{x}}^l)) - f(\boldsymbol{x}^\star) \geq 0$. Note that now the constants hidden in the big-O notation could possibly depend on r, but they are independent of δ_l or l. Applying Lemma 2.3 to the recursion (3.20), we have that the sequence $\{\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2\}$ converges almost surely, that

$$\sum_{l=0}^{\infty}\delta_l[f(\Pi_C(\tilde{\boldsymbol{x}}^l)) - f(\boldsymbol{x}^\star)] < \infty \quad a.s., \tag{3.21}$$

and that

$$\sum_{l=0}^{\infty}\|\tilde{\boldsymbol{x}}^l - \Pi_C(\tilde{\boldsymbol{x}}^l)\|^2 < \infty \quad a.s. \tag{3.22}$$

Inequality (3.21) and the fact that $\sum_{l=0}^{\infty}\delta_l = \infty$ imply that

$$\liminf_{l\to\infty} f(\Pi_C(\tilde{\boldsymbol{x}}^l)) = f(\boldsymbol{x}^\star) \quad a.s. \tag{3.23}$$

Also, inequality (3.22) implies that

$$\lim_{l\to\infty} \|\Pi_C(\tilde{\boldsymbol{x}}^l) - \tilde{\boldsymbol{x}}^l\| = 0 \quad a.s. \tag{3.24}$$

Since the sequence $\{\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|\}$ converges almost surely, the sequence $\{\tilde{\boldsymbol{x}}^l\}$ is bounded and has an accumulation point $\tilde{\boldsymbol{x}}^\star$ almost surely. Therefore, there exists a sub-sequence $\{\tilde{\boldsymbol{x}}^{l_t}\}$ of $\{\tilde{\boldsymbol{x}}^l\}$ such that $\tilde{\boldsymbol{x}}^{l_t} \to \tilde{\boldsymbol{x}}^\star$ as $t \to \infty$. By relation (3.24) and continuity of $\Pi_C(\cdot)$, the sequence $\Pi_C(\tilde{\boldsymbol{x}}^{l_t})$ converges almost surely to $\Pi_C(\tilde{\boldsymbol{x}}^\star) = \tilde{\boldsymbol{x}}^\star$. Clearly, $\tilde{\boldsymbol{x}}^\star \in C$. It follows from (3.23) and the continuity of f that $f(\tilde{\boldsymbol{x}}^\star) = f(\boldsymbol{x}^\star)$. Hence, $\tilde{\boldsymbol{x}}^\star \in X^\star$. Since $\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|$ converges almost surely for every $\boldsymbol{x}^\star \in X^\star$, we have that $\|\tilde{\boldsymbol{x}}^l - \tilde{\boldsymbol{x}}^\star\|$ converges almost surely. Since $\|\tilde{\boldsymbol{x}}^{l_t} - \tilde{\boldsymbol{x}}^\star\| \to 0$ as $t \to \infty$ almost surely, we have that $\|\tilde{\boldsymbol{x}}^l - \tilde{\boldsymbol{x}}^\star\| \to 0$ as $l \to \infty$ almost surely. Thus, almost surely, we have

$$\lim_{l\to\infty} \tilde{\boldsymbol{x}}^l = \tilde{\boldsymbol{x}}^\star \quad \text{and} \quad \lim_{l\to\infty} f(\tilde{\boldsymbol{x}}^l) = f^\star.$$

We then prove the convergence in $\{\boldsymbol{x}^k\}$. First, the boundedness of the sequence $\{\|\tilde{\boldsymbol{x}}^l - \tilde{\boldsymbol{x}}^\star\|^2\}$ and Proposition 3.2 imply that

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}^{k+1} - \tilde{\boldsymbol{x}}^\star\|^2 \mid \mathcal{F}_k] \\
&\leq (1 + \mathcal{O}(\alpha_k^2))\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}^\star\|^2 + \mathcal{O}(\alpha_k^2) - \frac{\rho}{4m\kappa^2}\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f(\tilde{\boldsymbol{x}}^\star)),
\end{aligned}$$

where the constants hidden in the big-O notations do not depend on α_k or k. Using the last inequality, Lemma 2.3 and the fact that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we get that the sequence $\{\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}^\star\|^2\}$ converges almost surely. Since the sub-sequence $\{\|\tilde{\boldsymbol{x}}^l - \tilde{\boldsymbol{x}}^\star\|^2\}$ converges almost surely to 0, we have that $\{\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}^\star\|^2\}$ converges almost surely to 0 as well, which shows that

$$\lim_{k\to\infty} \boldsymbol{x}^k = \tilde{\boldsymbol{x}}^\star \quad \text{and} \quad \lim_{k\to\infty} f(\boldsymbol{x}^k) = f^\star.$$

This completes the proof. □

Proof of Theorem 3.2. We first prove the convergence rate of the feasibility gap. Fix an arbitrary optimal solution $\boldsymbol{x}^\star \in X^\star$. By using Proposition 3.2 with $\lambda = 16m\kappa^2\rho^{-1}$ and the definition of α_k, we have that for all $k \geq 0$,

$$\frac{\rho}{2m\kappa^2} - 2\alpha_k L - \frac{4}{\lambda} \geq \frac{\rho}{8m\kappa^2},$$

and hence that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k] &\leq (1 + \mathcal{O}(\alpha_k^2))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2) \\
&\quad - \frac{\rho}{8m\kappa^2}\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f(\boldsymbol{x}^\star)) \qquad (3.25) \\
&\leq (1 + \mathcal{O}(\alpha_k^2))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2).
\end{aligned}$$

Similarly, the constants hidden in $\mathcal{O}(\cdot)$ do not depend on α_k or k. Since C_0 is compact, $\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2$ and $\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2$ are, respectively, uniformly bounded in k and l, which together with (3.25), imply that for any $k \geq 0$,

$$\frac{\rho}{8m\kappa^2}\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] + \mathcal{O}(\alpha_k^2).$$

Summing the last inequality over $k = 0, \cdots, K-1$, we have

$$\begin{aligned}\frac{\rho}{8m\kappa^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2\right] &\leq \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 - \mathbb{E}\left[\|\boldsymbol{x}^K - \boldsymbol{x}^\star\|^2\right] + \mathcal{O}(1)\sum_{k=0}^{K-1}\alpha_k^2 \\ &\leq \mathcal{O}(1) + \mathcal{O}(1)\sum_{k=1}^{K}\frac{1}{k} \leq \mathcal{O}(1)\log(K).\end{aligned}$$

By the convexity of function $\mathrm{dist}^2(\cdot, C)$, it follows that

$$\mathbb{E}\left[\mathrm{dist}^2(\bar{\boldsymbol{x}}^K, C)\right] \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2\right] \leq \mathcal{O}\left(\frac{\log(K)}{K}\right).$$

Next, we prove the convergence rate of the optimality gap. By the definition of $\boldsymbol{v}^k$ in Algorithm 1, we have

$$\mathbb{E}\left[\langle \boldsymbol{v}^k, \boldsymbol{x}^k - \boldsymbol{x}^\star\rangle \mid \mathcal{F}_k\right] \geq f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star),$$

which together with Lemma 3.3 implies that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] \\ &\leq (1 + 8L^2\alpha_k^2)\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + 16L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 - 2\alpha_k\left(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)\right) \\ &\quad - \mathbb{E}\left[\|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k\right] + 4\|\nabla f(\boldsymbol{x}^\star)\|^2\alpha_k^2 + 12L^2\alpha_k^2 \\ &\leq \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2\alpha_k\left(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)\right) + \mathcal{O}(1)\alpha_k,\end{aligned}$$

where the last inequality follows from the boundedness of C_0 and the fact that $\tilde{\boldsymbol{x}}^l, \boldsymbol{x}^k, \boldsymbol{x}^\star \in C_0$. Taking expectation on both sides, we obtain

$$\mathbb{E}\left[f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)\right] \leq \frac{1}{2\alpha_k}\left(\mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2\right]\right) + \mathcal{O}(1)\alpha_k.$$

Summing of the last inequality over $k = 0, 1, \cdots, K-1$, it follows that

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[f(\boldsymbol{x}^k) - f^\star\right] \\
\leq &\sum_{k=0}^{K-1} \frac{1}{2\alpha_k}\left(\mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2\right]\right) + \mathcal{O}(1)\sum_{k=0}^{K-1}\alpha_k \\
\leq &\frac{1}{2\tilde{\alpha}}\mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2\right] - \frac{1}{2\tilde{\alpha}}\sqrt{K}\mathbb{E}\left[\|\boldsymbol{x}^K - \boldsymbol{x}^\star\|^2\right] \\
&+ \frac{1}{2\tilde{\alpha}}\sum_{k\in[K-1]}\left(\sqrt{k+1} - \sqrt{k}\right)\mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] + \mathcal{O}\left(\sum_{k\in[K]}\frac{1}{\sqrt{k}}\right) \qquad (3.26) \\
\leq &\frac{1}{2\tilde{\alpha}}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 + \frac{1}{2\tilde{\alpha}}\sum_{k\in[K-1]}\left(\sqrt{k+1} - \sqrt{k}\right)\mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] + \mathcal{O}\left(\sum_{k\in[K]}\frac{1}{\sqrt{k}}\right) \\
\leq &\mathcal{O}(1) + \mathcal{O}(1)\sum_{k\in[K-1]}\left(\sqrt{k+1} - \sqrt{k}\right) + \mathcal{O}\left(\sum_{k\in[K]}\frac{1}{\sqrt{k}}\right) \\
\leq &\mathcal{O}(1) + \mathcal{O}(\sqrt{K-1}) + \mathcal{O}(\sqrt{K}) = \mathcal{O}(\sqrt{K}),
\end{aligned}$$

where the fourth inequality follows from the boundedness of C_0, and the fifth inequality follows from the fact that

$$\sum_{k\in[K]}\frac{1}{\sqrt{k}} \leq \int_1^{K+1}\frac{dt}{\sqrt{t}} = 2(\sqrt{K+1} - 1).$$

By the convexity of f and (3.26), we have

$$\mathbb{E}[f(\bar{\boldsymbol{x}}^K)] \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[f(\boldsymbol{x}^k)] \leq f^\star + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

This completes the proof. □

4. Numerical Experiments

In this section, we conduct numerical experiments to study the empirical performance of VR^{3}PM. All codes are implemented using MATLAB, and all the experiments are performed on a PC with Intel Core i7-6700HQ CPU (2.60 GHz). Because of their high accuracy, the optimal solution and optimal value computed by using CVX [10] are taken as the "true" optimal solution $\boldsymbol{x}^\star$ and optimal value $f^\star$, respectively.

The following two optimization problems will be used in our experiments. The first

one is a linearly constrained quadratic programming problem (LCQP):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}^\top A_i^\top A_i \boldsymbol{x} + \boldsymbol{a}_i^\top \boldsymbol{x} \\ \text{subject to} \quad & Q\boldsymbol{x} \leq \boldsymbol{w}. \end{aligned} \tag{4.1}$$

where $A_1, \ldots, A_n \in \mathbb{R}^{p\times d}, Q \in \mathbb{R}^{m\times d}$ are matrices and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^d, \boldsymbol{w} \in \mathbb{R}^m$ are vectors. In our experiment, these vectors and matrices are generated as follows. For each $i \in [n]$, we first generate a random matrix $\tilde{A}_i \in \mathbb{R}^{(p+1)\times d}$ with independently and identically distributed (i.i.d.) standard Gaussian entries. Then, the matrix $A_i \in \mathbb{R}^{p\times d}$ and the vector $\boldsymbol{a}_i \in \mathbb{R}^d$ are defined as submatrices of the normalized matrix $\tilde{A}_i/\|\tilde{A}_i\|_2$:

$$\tilde{A}_i/\|\tilde{A}_i\|_2 = \begin{pmatrix} A_i \\ \boldsymbol{a}_i^\top \end{pmatrix},$$

where $\|\cdot\|_2$ denotes the operator norm, *i.e.*, the largest singular value. For the matrix $Q \in \mathbb{R}^{m\times d}$, we take

$$Q = \begin{pmatrix} \boldsymbol{q}_1^\top \\ \vdots \\ \boldsymbol{q}_m^\top \end{pmatrix},$$

where each $\boldsymbol{q}_j = \tilde{\boldsymbol{q}}_j/\|\tilde{\boldsymbol{q}}_j\|$ and $\tilde{\boldsymbol{q}}_j \in \mathbb{R}^d$ is a random vector with i.i.d. standard Gaussian entries. Such summand-wise and constraint-wise normalizations for, respectively, the objective and constraints are to ensure uniformity and make the problem instances better conditioned. The vector $\boldsymbol{w} \in \mathbb{R}^m$ is a random vector with entries being i.i.d. uniform random variables on $[0, 0.5]$.

The second tested problem is a quadratically constrained quadratic programming problem (QCQP):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}^\top A_i^\top A_i \boldsymbol{x} + \boldsymbol{a}_i^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x}^\top B_j^\top B_j \boldsymbol{x} + \boldsymbol{b}_j^\top \boldsymbol{x} \leq \boldsymbol{w}, \quad j \in [m], \\ & \boldsymbol{x} \in C_0, \end{aligned} \tag{4.2}$$

where $A_1, \ldots, A_n \in \mathbb{R}^{p\times d}, B_1, \ldots, B_m \in \mathbb{R}^{q\times d}$ are matrices, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \in \mathbb{R}^d, \boldsymbol{w} \in \mathbb{R}^m$ are vectors, and $C_0 \subseteq \mathbb{R}^d$ is a convex subset. In our experiments, we take $C_0 = [-10, 10]^d$. The pairs $\{(A_i, a_i)\}_{i\in[n]}$ and $\{(B_j, b_j)\}_{j\in[m]}$ are generated in the same manner as the pairs $\{(A_i, a_i)\}_{i\in[n]}$ for problem (4.1). The vector $\boldsymbol{w}$ is a random vector with entries being i.i.d. uniform random variables on $[0, 0.5]$.

In all the experiments, the initial point $\boldsymbol{x}^0$ is a random vector with entries being i.i.d. uniform random variables on $[0, 1]$, the batch size for the SVRG gradient estimator is $b = 5$, the number of inner loop iterations is $r = n/b$, and the step-size α_k is given by

$$\alpha_k = \frac{0.01}{k^{0.55}}.$$

To improve the convergence behaviour, we also adopt the grouping technique as discussed in the final remark at the end of Section 3.1. In particular, we group the constraints into $\bar{m} = m/5$ groups of $\bar{b} = 5$ constraints.

4.1. Usefulness of the SVRG Gradient Estimator

Unlike the unconstrained case, the use of variance reduction technique does not improve the theoretical convergence rate in terms of the order of the iteration counter K, see Theorem 3.2 and [30, Theorem 2]. Nevertheless, in this experiment, we highlight the importance of the incorporation of the variance into our algorithm VR^{3}PM by empirically showing that the SVRG gradient estimator does substantially improve the practical convergence behaviour upon the vanilla gradient estimators. Specifically, we compare VR^{3}PM with three variants of random relaxed projection methods (R^{2}PMs) that are obtained by replacing the SVRG gradient estimator $\boldsymbol{v}^k$ in Algorithm 1 with the standard gradient estimator using a single summand function, the mini-batch gradient estimator using b summand functions and the full gradient using all the n summand functions. These three R^{2}PMs are denoted, respectively, as R^{2}PM-1, R^{2}PM-b and R^{2}PM-n. The constraint grouping technique is applied to both our algorithm VR^{3}PM and all these three variants of R^{2}PMs. However, for these three variants, we use the step-size $\alpha_k = 1/k^{0.55}$ as we found it performs empirically better than the choice $\alpha_k = 0.01/k^{0.55}$.

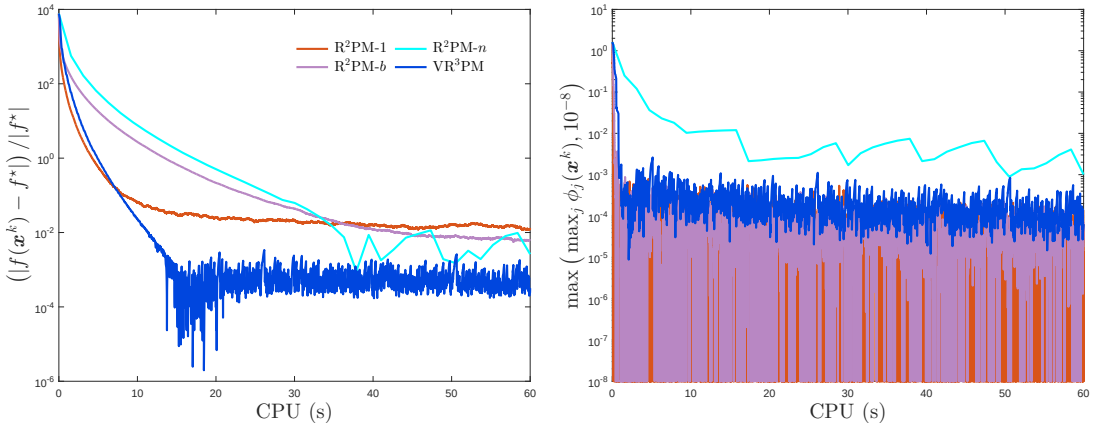
The results on problem (4.1) with parameters $(n, m, d, p) = (2000, 500, 200, 30)$ and $(n, m, d, p) = (4000, 1000, 400, 30)$ are shown in Figure 1, whereas the results on problem (4.2) with parameters $(n, m, d, p, q) = (2000, 500, 200, 30, 50)$ and $(n, m, d, p, q) = (4000, 1000, 400, 30, 15)$ are shown in Figure 2. In both figures, the left and right columns show the sub-optimality gap and constraint violation against the computation time (in second), respectively.

We can see from Figure 1 and Figure 2 that in terms of objective value, our algorithm VR^{3}PM is faster and reaches a higher accuracy than the other three R^{2}PMs. As for the constraint violation, our algorithm VR^{3}PM performs on par with R^{2}PM-1 and R^{2}PM-b and outperforms the full gradient variant R^{2}PM-n.

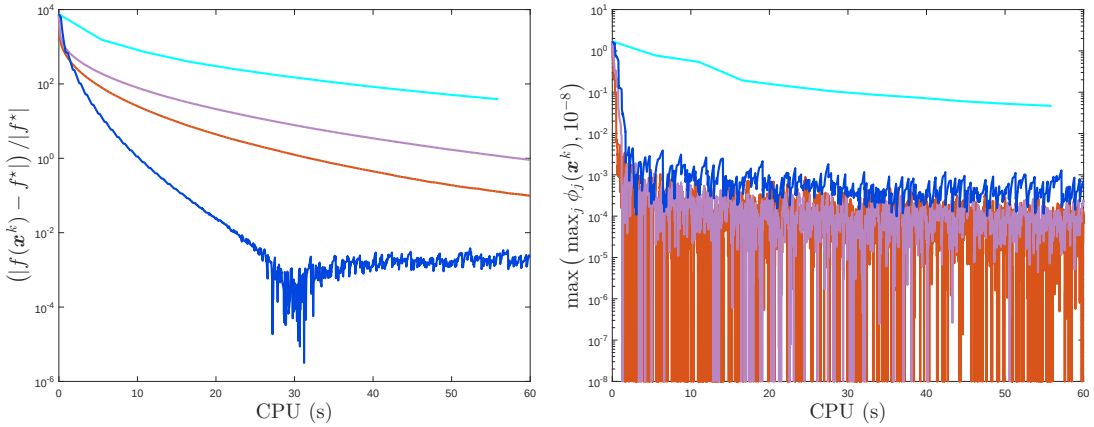
4.2. Comparison with the RPMs in [22] and [30]

Our second experiment aims at comparing the performance of VR^{3}PM with two existing RPMs, namely, the RPM by Nedić [22] and the RPM by Wang and Bertsekas [30]. They are denoted as RPM-N and RPM-WB, respectively. As we mentioned at the end of Section 3.1, RPM-N is amenable to the constraint grouping technique. So, we apply the technique to RPM-N in our experiments to improve its performance.

The results on problem (4.1) with parameters $(n, m, d, p) = (2000, 500, 200, 30)$ and $(n, m, d, p) = (4000, 1000, 200, 30)$ are shown in Figure 3, whereas the results on problem (4.2) with parameters $(n, m, d, p, q) = (2000, 500, 200, 30, 15)$ and $(n, m, d, p, q) =$



(a) $(n, m, d, p) = (2000, 500, 200, 30)$



(b) $(n, m, d, p) = (4000, 1000, 400, 30)$

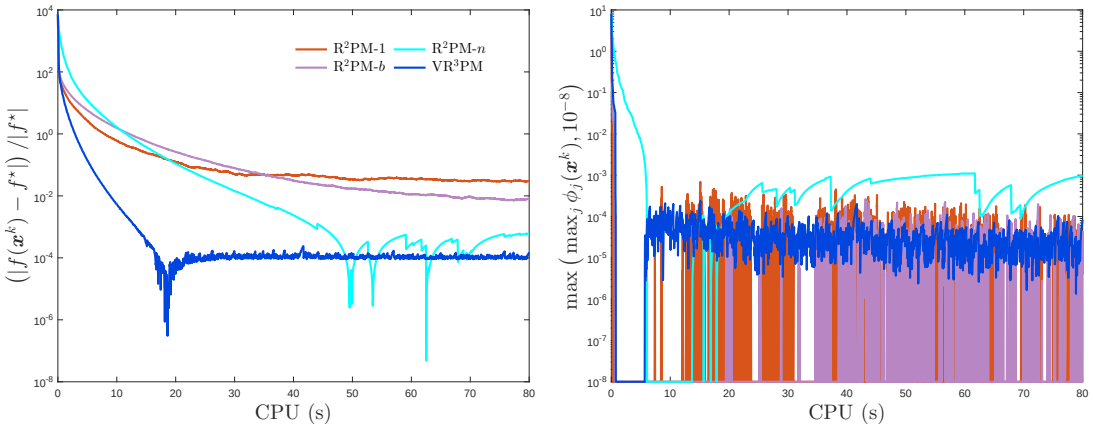
Figure 1: Comparison of VR^{3}PM and other R^{2}PMs using vanilla gradient estimators.

$(4000, 1000, 400, 30, 15)$ are shown in Figure 4. In both figures, the left and right columns show the sub-optimality gap and constraint violation against the computation time (in second), respectively.

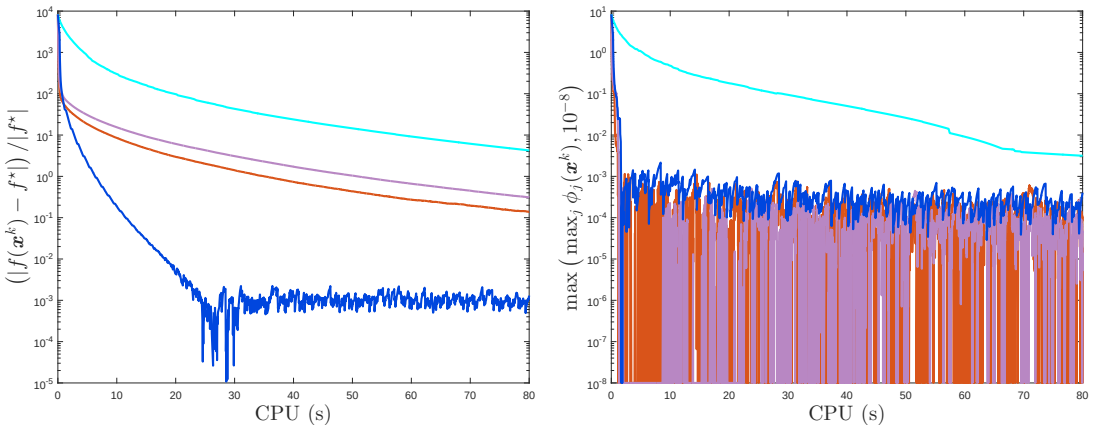
From Figure 3 and Figure 4, we can see that our algorithm VR^{3}PM performs substantially better than the competing algorithms RPM-N and RPM-WB in terms of both objective value and constraint violation.

Acknowledgements

Kai Tu is supported by National Natural Science Foundation of China (No. 12101436). Man-Chung Yue is supported by the Hong Kong Research Grants Council under the General Research Fund project 15305321.

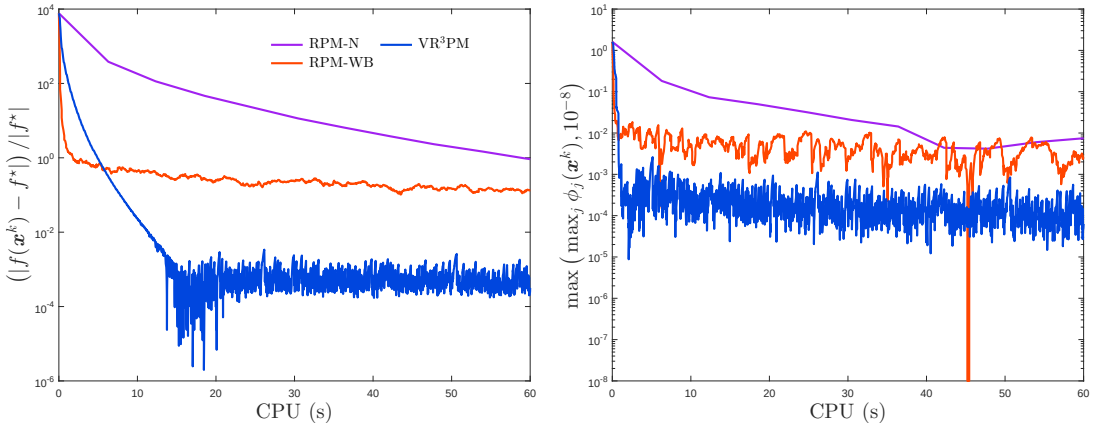


(a) $(n, m, d, p) = (2000, 500, 200, 30, 15)$

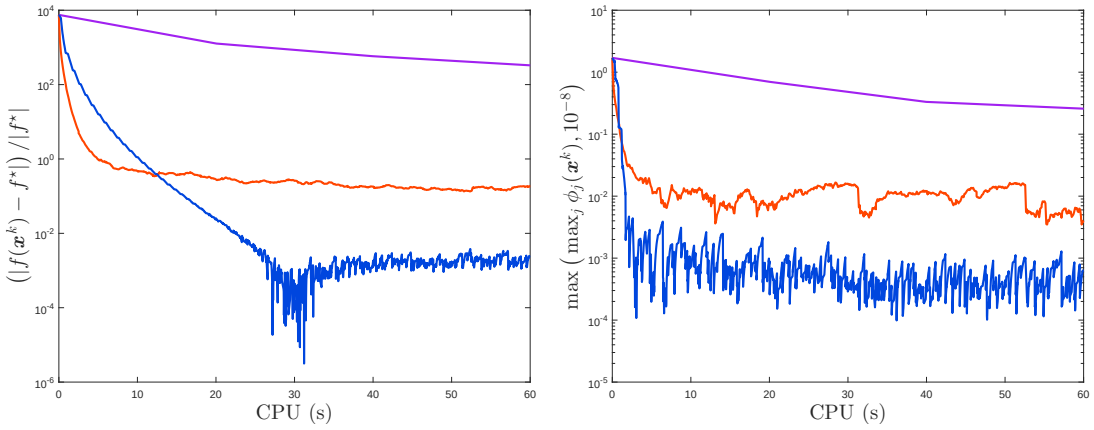


(b) $(n, m, d, p, q) = (4000, 1000, 400, 30, 15)$

Figure 2: Comparison of VR^{3}PM and other R^{2}PMs using vanilla gradient estimators.

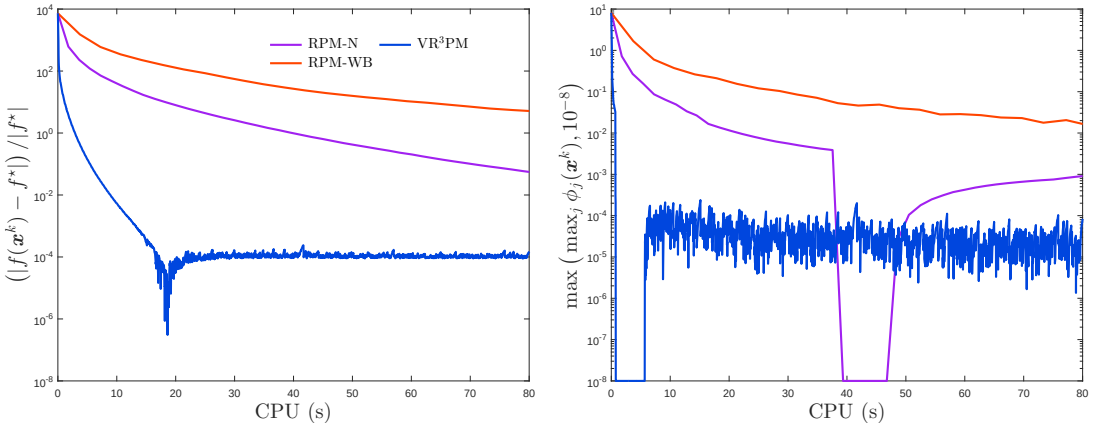


(a) $(n, m, d, p) = (2000, 500, 200, 30)$

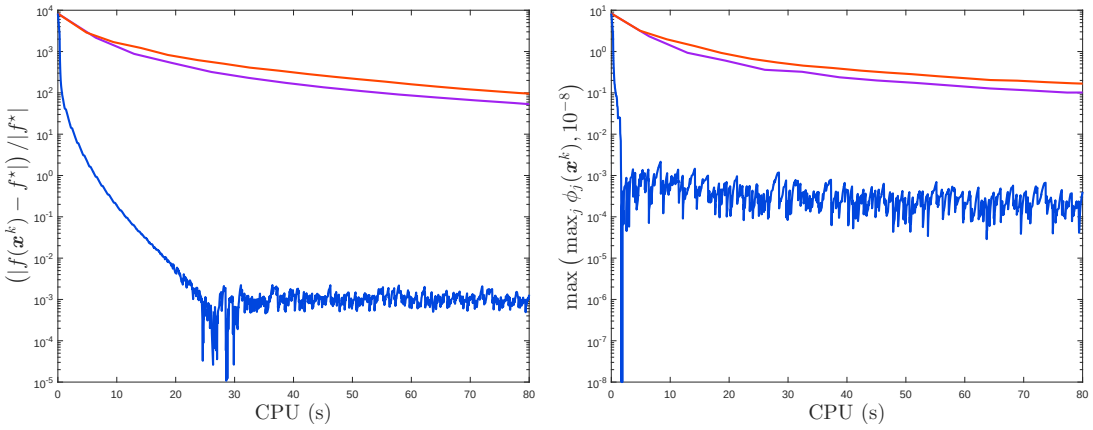


(b) $(n, m, d, p) = (4000, 1000, 400, 30)$

Figure 3: Comparison of VR^{3}PM and the RPMs by Nedić [22] and by Wang and Bertsekas [30].



(a) $(n, m, d, p) = (2000, 500, 200, 30, 15)$



(b) $(n, m, d, p, q) = (4000, 1000, 400, 30, 15)$

Figure 4: Comparison of VR^{3}PM and the RPMs by Nedić [22] and by Wang and Bertsekas [30].

References

1. S. Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954.

2. H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993.

3. A. Beck. *First-order Methods in Optimization*. SIAM, 2017.

4. L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

5. W. Chen and R. Mazumder. Multivariate convex regression at scale. *arXiv preprint arXiv:2005.11588*, 2020.

6. A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27:1646–1654, 2014.

7. Z. Deng, M.-C. Yue, and A. M.-C. So. An efficient augmented Lagrangian-based method for linear equality-constrained Lasso. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5760–5764. IEEE, 2020.

8. M. Fukushima. On the convergence of a class of outer approximation algorithms for convex programs. *Journal of Computational and Applied Mathematics*, 10(2):147–156, 1984.

9. B. R. Gaines, J. Kim, and H. Zhou. Algorithms for fitting the constrained Lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018.

10. M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming, 2008.

11. L. G. Gubin, B. Polyak, and É. V. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7:1–24, 1967.

12. A. J. Hoffman. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 174–176. World Scientific, 2003.

13. R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26:315–323, 2013.

14. J. E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

15. D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

16. N. Le Roux, M. W. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, 4:2663–2671, 2013.

17. M. Lin, D. Sun, and K.-C. Toh. An augmented Lagrangian method with constraint generation for shape-constrained convex regression problems. *Mathematical Programming Computation*, 14(2):223–270, 2022.

18. H. Liu, M.-C. Yue, and A. Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

19. H. Liu, M.-C. Yue, and A. M.-C. So. A unified approach to synchronization problems over subgroups of the orthogonal group. *arXiv preprint arXiv:2009.07514*, 2020.

20. H. Liu, M.-C. Yue, A. M.-C. So, and W.-K. Ma. A discrete first-order method for large-scale MIMO detection with provable guarantees. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2017.

21. Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

22. A. Nedić. Random algorithms for convex minimization problems. *Mathematical Programming*, 129(2):225–253, 2011.

23. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

24. Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2018.

25. P. Netrapalli. Stochastic gradient descent and its variants in machine learning. *Journal of the Indian Institute of Science*, 99(2):201–213, 2019.

26. B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

27. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

28. H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

29. S. E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer, 2004.

30. M. Wang and D. P. Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016.

31. S. X. Wu, M.-C. Yue, A. M.-C. So, and W.-K. Ma. SDR approximation bounds for the robust multicast beamforming problem with interference temperature constraints. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4054–4058. IEEE, 2017.

32. M.-C. Yue, Z. Zhou, and A. Man-Cho So. On the quadratic convergence of the cubic regularization method under a local error bound condition. *SIAM Journal on Optimization*, 29(1):904–932, 2019.

33. M.-C. Yue, Z. Zhou, and A. M.-C. So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. *Mathematical Programming*, 174(1):327–358, 2019.