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

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

Abstract

For many applications in signal processing and machine learning, we are tasked with minimizing a large sum of convex functions subject to a large number of convex constraints. In this paper, we devise a new random projection method (RPM) to efficiently solve this problem. Compared with existing RPMs, our proposed algorithm features two useful algorithmic ideas. First, at each iteration, instead of projecting onto the subset defined by one of the constraints, our algorithm only requires projecting onto a half-space approximation of the subset, which significantly reduces the computational cost as it admits a closed-form formula. Second, to exploit the structure that the objective is a sum, variance reduction is incorporated into our algorithm to further improve the performance. As theoretical contributions, under an error bound condition and other standard assumptions, we prove that the proposed RPM converges to an optimal solution and that both optimality and feasibility gaps vanish at a sublinear rate. We also provide sufficient conditions for the error bound condition to hold. Experiments on a beamforming problem and a robust classification problem are also presented to demonstrate the superiority of our RPM over existing ones.

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

1. Introduction

This paper considers the following constrained convex optimization problem

$$\begin{array}{ll} \min & f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in C = C_0 \cap C_1 \cap \cdots \cap C_m, \end{array} \tag{1}$$

[*] School of Mathematical Science, Sichuan Normal University. Email:yangzhichun1994@163.com

[†] School of Mathematical Science, Sichuan Normal University. Email: fuquanxia@163.com

[‡] Corresponding author. College of Mathematics and Statistics, Shenzhen University. Email: 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

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is a differentiable convex function for $i = 1, \ldots, n$, $C_0 \subseteq \mathbb{R}^d$ is a non-empty, closed and convex set, $C_j = \{\boldsymbol{x} \in \mathbb{R}^d \mid \phi_j(\boldsymbol{x}) \leq 0\}$ with $\phi_j : \mathbb{R}^d \to \mathbb{R}$ being a convex but possibly non-differentiable function for $j = 1, \ldots, m$. Problem (1) finds many applications across a wide range of areas, including the beamforming problem [36], the constrained LASSO problem [6, 8] and the convex regression problem [4, 16].

A straightforward choice for solving problem (1) is the projected gradient method (PGM) whose iteration takes the form $\boldsymbol{x}^{k+1} = \Pi_C\left(\boldsymbol{x}^k - \alpha_k \nabla f(\boldsymbol{x}^k)\right)$, 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, PGM converges to an optimal solution with a sublinear rate $\mathcal{O}(1/K)$ [2], where K is the total number of iterations. Under stronger assumptions, it is also proved that PGM converges linearly to an optimal solution [26] (*i.e.*, $\mathcal{O}(c^K)$ for some $c \in (0, 1)$).

However, PGM is not a viable algorithm for solving problem (1) if

(i) the number n of summands f_i is large,

(ii) the number m of constraints 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) or (iii), 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 random index. The resulting algorithm is known as the stochastic gradient method (SGM), which traces back to the work [30] of Robbins and Monro. A natural generalization of SGM is the so-called mini-batch SGM [3], where instead of a single summand f_i, multiple summands 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 [27]. A 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 $\mathcal{O}(1/\sqrt{K})$ [25], 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 $\mathcal{O}(1/K)$ [3], as compared to the linear convergence rate of the gradient descent. Various variance reduction techniques have been proposed to remedy the variance issue. Notable examples of variance reduced SGM include SAG [14], SAGA [5] and SVRG [11]. 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.

To deal with optimization problems with a large number of constraints, *i.e.*, in the presence of difficulty (ii), one could adopt the framework of random projection methods (RPMs). Roughly speaking, at each iteration, an RPM improves feasibility with respect to the intersection $\cap_{j=1}^{m} C_j$ by using only one randomly selected subset C_{j_k} but ignores the others, where j_k is a random index. The computational cost at each iteration can thus be substantially reduced. In the context of optimization, RPMs were first studied by Nedić in [23], which can be viewed as an extension of the algorithm in [29] by Polyak for convex feasibility problems to convex minimization problems. Subsequent works on RPMs include [22, 24, 33, 34]. On the theoretical side, these works typically show that under certain assumptions, both the optimality and feasibility gaps converge to zero at a sublinear rate.

Some of these RPMs, such as [22,33,34], are designed for constrained optimization where the objective is an expectation $f(\boldsymbol{x}) = \mathbb{E}_I[f_I(\boldsymbol{x})]$ and subsumes the finite-sum objective in problem (1) as a special case. In these works, 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 random gradient estimator, which is used in the updating step of the algorithms. Therefore, it similarly suffers from the variance issue. In view of our previous discussion, it is tempted to replace the gradient estimator $\nabla f_i(\boldsymbol{x}^k)$ by its variance reduced counterpart. Unfortunately, it is unclear whether the variance reduced estimators satisfy the assumptions required by the existing RPMs, such as [33, Assumption 1(c)] and [22, Assumption 1]. Hence, incorporating variance reduction into RPMs with provable theoretical guarantees is nontrivial.

When difficulty (iii) is present, computing the projection onto the subset C_j could be time consuming. This hinders the the practicality of projection-based RPMs, such as [33,34], where each iteration requires computing the projection $\Pi_{C_{j_k}}$ onto the randomly selected subset C_{j_k}. In such a case, one could improve the efficiency by approximating the subset C_{j_k} by a half-space at each iteration. The advantage is that the projection onto a half-space can be computed much more efficiently via an explicit formula. 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 [7] and the cutting-plane method [12].

In this paper, we develop a new RPM that aims at solving problem (1) in the face of difficulties (i)-(iii). 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. Under certain assumptions, we prove that the sequence of iterates generated by the proposed RPM converges to an optimal solution to problem (1). Moreover, the convergence rates of the optimality gap and the feasibility gap are $\mathcal{O}(1/\sqrt{K})$ and $\mathcal{O}(1/K)$, respectively. If f satisfies a quadratic growth condition, the rates can be strengthened to $\mathcal{O}(1/K)$ and $\mathcal{O}(1/K^2)$, respectively. As a side contribution, we formulate and prove an error bound-type condition, which plays 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.

The rest of this paper is organized as follows. We present the proposed algorithm and the required assumptions in Section 2 and Section 3, respectively. In Section 4, we study the convergence behaviour of the proposed algorithm. Numerical results are reported in Section 5.

1.1. Notation

The Euclidean norm and inner product are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. For two nonnegative sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k = \mathcal{O}(b_k)$ if there exists a scalar $c > 0$ such that $a_k \leq c\, b_k$ for any $k \geq 0$. We denote by $|S|$ the cardinality of a set S. For a positive integer m, we write $[m] = \{1, \cdots, m\}$. For any convex function ϕ and $\boldsymbol{x} \in \mathrm{dom}(\phi)$, we denote by $\partial\phi(\boldsymbol{x}) = \{\boldsymbol{\xi} \in \mathbb{R}^d \mid \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) \geq \langle \boldsymbol{\xi},\, \boldsymbol{y} - \boldsymbol{x}\rangle,\ \forall \boldsymbol{y} \in \mathbb{R}^d\}$ the subdifferential of ϕ at $\boldsymbol{x}$. The optimal value and the set of optimal solutions of problem (1) are denoted by $f^\star$ and $X^\star$, respectively. We abbreviate "almost surely" as "a.s". For a collection $\mathcal{G}$ of random variables, $\mathbb{E}[\cdot|\mathcal{G}]$ denotes the conditional expectation. Given any subset $D \subseteq \mathbb{R}^d$, we denote the distance from $\boldsymbol{x} \in \mathbb{R}^d$ to D by $\mathrm{dist}(\boldsymbol{x}, D) = \inf\{\|\boldsymbol{y} - \boldsymbol{x}\| \mid \boldsymbol{y} \in D\}$. If D is closed and convex, then the projection is denoted by $\Pi_D(\boldsymbol{x}) = \arg\min\{\|\boldsymbol{y} - \boldsymbol{x}\| \mid \boldsymbol{y} \in D\}$.

2. Variance Reduced Random Relaxed Projection Method

We propose a new RPM for solving problem (1), namely the variance reduced random relaxed projection method (VR^{3}PM) in Algorithm 1. 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

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

Note that $H(\phi;\, \boldsymbol{x};\, \boldsymbol{\xi})$ is a half-space if $\boldsymbol{\xi} \neq \boldsymbol{0}$.

Some remarks about VR^{3}PM are in order. First, by definition, $C_{j_k} \subseteq H_k$. 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 lower than that of [33]. Indeed, if $\boldsymbol{\xi}^k \neq \boldsymbol{0}$, the projection can be computed in closed-form by Lemma 2:

$$\boldsymbol{y}^{k+1} = \boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k - \frac{\left(\phi_{j_k}(\boldsymbol{x}^k) - \alpha_k \langle \boldsymbol{\xi}^k, \boldsymbol{v}^k\rangle\right)_+}{\|\boldsymbol{\xi}^k\|^2}\boldsymbol{\xi}^k.$$

Third, the vector $\boldsymbol{v}^k$ is the so-called SVRG gradient estimator [11]. It is used to reduce the variance of the algorithm and thus improves the convergence speed. Fourth, the choice

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

```
Input: Initial point $\boldsymbol{x}^0 \in C_0$, integers $b \geq 1$ and $r \geq 2$, and a positive sequence $\{\alpha_k\}$.
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 i.i.d. uniform indices $I_k = \{i_{k1}, \cdots, i_{kb}\} \subseteq [n]$. Compute
            $\boldsymbol{v}^k = \frac{1}{b} \sum_{i \in I_k} (\nabla f_i(\boldsymbol{x}^k) - \nabla f_i(\tilde{\boldsymbol{x}}^l)) + \nabla f(\tilde{\boldsymbol{x}}^l).$
        Generate a random index $j_k \in [m]$. Compute a subgradient $\boldsymbol{\xi}^k \in \partial \phi_{j_k}(\boldsymbol{x}^k)$ and
            $\boldsymbol{y}^{k+1} = \Pi_{H_k}\big(\boldsymbol{x}^k - \alpha_k \boldsymbol{v}^k\big),$
        where $H_k = H(\phi_{j_k}; \boldsymbol{x}^k; \boldsymbol{\xi}^k)$.
        Update the next iterate by
            $\boldsymbol{x}^{k+1} = \Pi_{C_0}(\boldsymbol{y}^{k+1}).$
    end for
end for
```

of the distribution of the random index j_k is flexible. As long as all the constraints have a positive probability of being chosen (see Assumption 2), our theoretical framework applies. 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) into $\bar{m} < m$ groups and re-write it as another optimization problem with only $\bar{m}$ constraints using the maximum function. For example, let $\bar{b} > 0$ be an integer that divides m and define $\bar{m} = m/\bar{b}$. Then, we could re-write the feasible region of problem (1) as $C = C_0 \cap (\bigcap_{t \in [\bar{m}]} \bar{C}_t)$ where $\bar{C}_t = \{\boldsymbol{x} \in \mathbb{R}^d \mid \max_{j=(t-1)\bar{b}+1,\ldots,t\bar{b}} \phi_j(\boldsymbol{x}) \leq 0\}$. This grouping technique can be seen as the constraint analogue of the mini-batch gradient estimator and mitigates the variance issue due to the random sampling of constraints. Therefore, despite the more difficult projection, the overall speed and accuracy could be improved by a suitable grouping. We should point out that the RPM in [23] is amenable to this technique as it is subgradient-based. However, it might not be worth applying the technique to the RPM in [33], as doing so requires computing the projection onto the grouped subset $\bar{C}_t$.

3. Assumptions

To analyze VR^{3}PM, the following blanket assumptions on problem (1) are imposed.

Assumption 1. The following hold.

(i) 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.

(ii) The optimal solution set $X^\star$ is non-empty.

(iii) For $i \in [n]$, f_i is differentiable and convex, and there exists $L_i > 0$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in C_0$,

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

Assumptions 1(i)-(ii) are standard in the literature of constrained convex optimization. Assumption 1(iii) is weaker than the requirement on f_i in [23], which assumes that each f_i has a Lipschitz continuous gradient. Indeed, it can be easily checked that Assumption 1(iii) holds if for any $i \in [n]$, either the function f_i or its gradient ∇f_i is Lipschitz continuous. Also, [33] requires a similar inequality to hold on $\mathbb{R}^d$, whereas Assumption 1(iii) is assumed to hold on the subset C_0. An immediate consequence of Assumption 1(iii) is that for any $\boldsymbol{x}, \boldsymbol{y} \in C_0$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L(\|\boldsymbol{x} - \boldsymbol{y}\| + 1), \tag{2}$$

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

We also need an assumption concerning the distribution of the constraint index j_k.

Assumption 2. 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 $\mathcal{F}_k = \{j_0, \cdots, j_{k-1}, i_{01}, \cdots, i_{(k-1)b}, \boldsymbol{x}^0\}$ for $k \geq 1$ and $\mathcal{F}_0 = \{\boldsymbol{x}^0\}$.

Assumption 2 ensures that at each iteration, any constraint will be picked with a positive probability.

The next assumption concerns the collection of the subsets $C_1, \dots, C_m$.

Assumption 3. There exists $\kappa > 0$ such that for any $\boldsymbol{x} \in C_0$,

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \kappa \max_{j\in[m]} \min_{\boldsymbol{\xi}_j\in\partial\phi_j(\boldsymbol{x})} \operatorname{dist}\left(\boldsymbol{x},\ H(\phi_j;\ \boldsymbol{x};\ \boldsymbol{\xi}_j)\right). \tag{3}$$

Conditions similar to Assumption 3 are utilized to study RPMs. First, [23] assumes the existence of a scalar $\kappa > 0$ such that for any $\boldsymbol{x} \in C_0$,

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \kappa\, \mathbb{E}[(\phi_j(\boldsymbol{x}))_+ |\, \mathcal{F}_k]. \tag{4}$$

If there exists an upper bound $M > 0$ on the sub-differentials $\partial\phi_j(\boldsymbol{x})$ that is uniform in j and $\boldsymbol{x}$, then condition (4) implies Assumption 3. To see this, we consider the two cases $\mathbf{0} \notin \partial\phi_j(\boldsymbol{x})$ and $\mathbf{0} \in \partial\phi_j(\boldsymbol{x})$ separately. If $\mathbf{0} \notin \partial\phi_j(\boldsymbol{x})$, by Lemma 2,

$$\begin{aligned}(\phi_j(\boldsymbol{x}))_+ \leq & M \min_{\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x})} \frac{(\phi_j(\boldsymbol{x}))_+}{\|\boldsymbol{\xi}_j\|} \\ = & M \min_{\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x})} \operatorname{dist}(\boldsymbol{x},\, H(\phi_j;\, \boldsymbol{x};\, \boldsymbol{\xi}_j)) .\end{aligned}$$

If $\mathbf{0} \in \partial\phi_j(\boldsymbol{x})$, $\boldsymbol{x}$ is a minimzer of ϕ_j. The non-emptiness of C_j then implies that $\phi_j(\boldsymbol{x}) \leq 0$ and hence that $(\phi_j(\boldsymbol{x}))_+ = 0$. Also, since $H(\phi_j;\, \boldsymbol{x};\, \mathbf{0}) = \mathbb{R}^d$, we have $\min_{\boldsymbol{\xi}_j \in \partial\phi_j(\boldsymbol{x})} \operatorname{dist}(\boldsymbol{x},\, H(\phi_j;\, \boldsymbol{x};\, \boldsymbol{\xi}_j)) = 0$. Thus, in both cases, condition (4) implies Assumption 3.

In fact, Assumption 3 is strictly weaker than (4). Consider $\phi_j(\boldsymbol{x}) = \boldsymbol{x}^\top B_j \boldsymbol{x}$ for positive definite matrices B_j. Then, $C = \{\mathbf{0}\}$. If $\boldsymbol{x} = \mathbf{0}$, Assumption 3 holds trivially. If $\boldsymbol{x} \neq \mathbf{0}$, for any $j \in [m]$,

$$\begin{aligned}\frac{(\phi_j(\boldsymbol{x}))_+}{\|\nabla\phi_j(\boldsymbol{x})\|} = & \frac{\boldsymbol{x}^\top B_j \boldsymbol{x}}{2\|B_j \boldsymbol{x}\|} \geq \frac{\lambda_{\min}(B_j)\|\boldsymbol{x}\|^2}{2\lambda_{\max}(B_j)\|\boldsymbol{x}\|} \\ = & \frac{\lambda_{\min}(B_j)}{2\lambda_{\max}(B_j)}\|\boldsymbol{x}\| = \frac{\lambda_{\min}(B_j)}{2\lambda_{\max}(B_j)} \operatorname{dist}(\boldsymbol{x}, C),\end{aligned}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues, respectively. This, together with Lemma 2, shows that Assumption 3 holds with $\kappa = 2\max_{j\in[m]} \lambda_{\max}(B_j)/\lambda_{\min}(B_j)$. Also,

$$\begin{aligned}\frac{\mathbb{E}[(\phi_j(\boldsymbol{x}))_+ | \mathcal{F}_k]}{\operatorname{dist}(\boldsymbol{x}, C)} \leq & \frac{\max_{j\in[m]}(\phi_j(\boldsymbol{x}))_+}{\|\boldsymbol{x}\|} \\ = & \max_{j\in[m]} \frac{\boldsymbol{x}^\top B_j \boldsymbol{x}}{\|\boldsymbol{x}\|} \leq \|\boldsymbol{x}\| \max_{j\in[m]} \lambda_{\max}(B_j),\end{aligned}$$

where the upper bound vanishes as $\boldsymbol{x} \to \mathbf{0}$. This shows that condition (4) does not hold.

Assumption 3 should also be compared to the classical notion of bounded linear regularity [1]:

$$\operatorname{dist}(\boldsymbol{x}, C) \leq \kappa \max_{j\in[m]} \operatorname{dist}(\boldsymbol{x}, C_j),$$

which has been used for analyzing RPMs in [9,33]. Since $C_j \subseteq H(\phi_j;\, \boldsymbol{x};\, \boldsymbol{\xi}_j)$, one readily sees that Assumption 3 is stronger than the bounded linear regularity. That a stronger assumption is required by our algorithm is expected as it relies on a weaker projection.

Nevertheless, the proposition below asserts that Assumption 3 holds under mild conditions. The proof can be found in Appendix B. We should emphasize that case (i) is not new, see [10] for example. The novelty of the proposition lies in case (ii).

Proposition 1. Suppose that Assumption 1 holds. Then, Assumption 3 holds if

(i) $C_0 = \mathbb{R}^d$ and ϕ_j is affine for all $j \in [m]$, or

(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.

Finally, we remark that Assumption 3 can also be seen as a generalization of the seminal Hoffman error bound [10], which asserts that the distance of any point to a linear system is linearly bounded by its violation of the linear constraints. Indeed, under case (i), inequality (3) reduces to the Hoffman error bound. Error bound conditions are important subjects in optimization and frequently utilized to study optimization algorithms. For example, as discussed above, conditions similar to (but different from) (3) are employed to study RPMs in [23] and [33]. Going beyond RPMs, error bound conditions also appeared in the study of first-order methods [19, 20], second-order methods [6, 38, 39], and even manifold optimization algorithms [17, 18].

4. Convergence Analysis

We now provide a detailed analysis of the convergence behaviour of VR^{3}PM. The proofs can be found in Appendices C-F. The first one shows that VR^{3}PM converges to an optimal solution of problem (1) almost surely under Assumptions 1–3 and a suitable choice for the step-size α_k.

Theorem 1. Suppose that Assumptions 1–3 hold. Let $\{\mu_l\}$ be a positive sequence satisfying $\sum_{l=0}^{\infty} \mu_l = \infty$ and $\sum_{l=0}^{\infty} \mu_l^2 < \infty$. Consider Algorithm 1 with $\alpha_k = \mu_l$ for $k = lr + s - 1$ with $l \geq 0$ and $s \in [r]$. Then, the iterates $\{\boldsymbol{x}^k\}$ converges almost surely to a point in $X^\star$.

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

Theorem 2. Suppose that Assumptions 1–3 hold and that C_0 is compact. Consider Algorithm 1 with $\alpha_k = \frac{\tilde{\alpha}}{\sqrt{k+1}}$, where $\tilde{\alpha} \in (0, \frac{\rho}{16Lm\kappa^2}]$. Then, for any $K \geq 1$, we have that

$$\mathbb{E}\left[\operatorname{dist}^2(\bar{\boldsymbol{x}}^K, C)\right] \leq \mathcal{O}\left(\tfrac{\log(K)}{K}\right)$$
$$\text{and } \mathbb{E}\left[f(\bar{\boldsymbol{x}}^K) - f^\star\right] \leq \mathcal{O}\left(\tfrac{1}{\sqrt{K}}\right),$$

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

With a constant step-size, we can improve the bound on the feasibility gap by a factor of $\log K$.

Theorem 3. Suppose that Assumptions 1–3 hold and that C_0 is compact. Consider Algorithm 1 with $\alpha_k \equiv \alpha = \frac{\tilde{\alpha}}{\sqrt{K+1}}$, where $\tilde{\alpha} \in (0, \frac{\rho}{16Lm\kappa^2}]$. Then, for any $K \geq 1$, we have that

$$\mathbb{E}\left[\operatorname{dist}^2(\bar{\boldsymbol{x}}^K, C)\right] \leq \mathcal{O}\left(\tfrac{1}{K}\right)$$
$$\text{and } \mathbb{E}\left[f(\bar{\boldsymbol{x}}^K) - f^\star\right] \leq \mathcal{O}\left(\tfrac{1}{\sqrt{K}}\right),$$

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

Our last theoretical result relies on the notion of quadratic growth.

Assumption 4. There exists constant $\nu > 0$ such that $f(\boldsymbol{x}) - f^\star \geq \frac{\nu}{2}\operatorname{dist}^2(\boldsymbol{x}, X^\star)$ for any $\boldsymbol{x} \in C$.

Stronger rates can be obtained when the quadratic growth condition is satisfied.

Theorem 4. Suppose that Assumptions 1–4 hold. Consider Algorithm 1 with $\alpha_k = \frac{8}{\nu(l+1)}$ for $k = lr + s - 1$ with $l \geq 0$ and $s \in [r]$. Then, for any $K \geq 1$, we have that

$$\begin{aligned} &\mathbb{E}\left[\operatorname{dist}^2(\bar{\boldsymbol{x}}^K, C)\right] \leq \mathcal{O}\left(\tfrac{1}{K^2}\right) \\ \text{and } &\mathbb{E}[f(\bar{\boldsymbol{x}}^K) - f^\star] \leq \mathcal{O}\left(\tfrac{1}{K}\right), \end{aligned}$$

where $\bar{\boldsymbol{x}}^K = \frac{3}{K(K+1)(K+2)}\sum_{k\in[K]} k(k+1)\boldsymbol{x}^k$.

A similar result has been proved for the RPM in [22]. However, the bounds in [22] hold only if K is sufficiently large, whereas ours hold for any $K \geq 1$. Moreover, [22] requires the quadratic growth condition to hold on the larger set C_0 but not only C.

5. Numerical Experiments

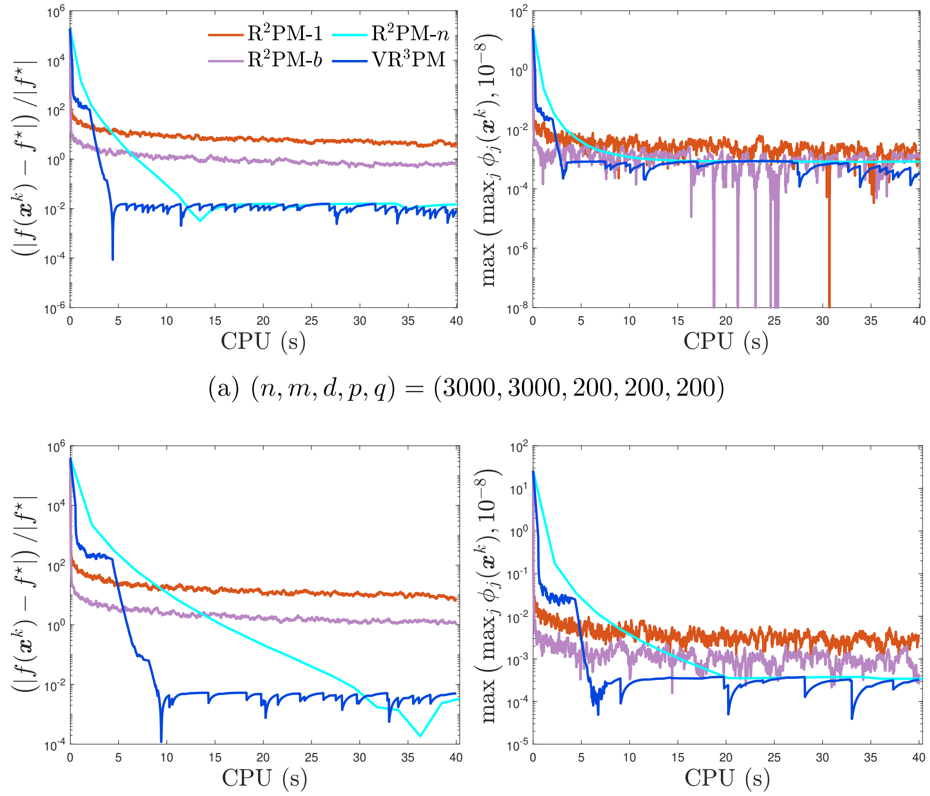
We then study the empirical performance of VR^{3}PM through numerical experiments. All the experiments are performed using MATLAB on a PC with Intel Core i5-1135G7 CPU (2.40 GHz). Because of its high accuracy, the optimal solution and optimal value computed by using YALMIP are taken as the "true" optimal solution $\boldsymbol{x}^\star$.

5.1. Importance of Variance Reduction

In this experiment, we highlight the importance of variance reduction to 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 consider the following quadratically constrained quadratic programming problem (QCQP):

$$\begin{aligned} \min \quad & \frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}^\top A_i^\top A_i \boldsymbol{x} + \boldsymbol{a}_i^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x}^\top B_j^\top B_j \boldsymbol{x} + \boldsymbol{b}_j^\top \boldsymbol{x} \leq w_j, \quad j \in [m], \\ & \boldsymbol{x} \in C_0. \end{aligned} \tag{5}$$

Here, $C_0 = [-10, 10]^d$. The pairs $\{(A_i, \boldsymbol{a}_i)\}_{i\in[n]}$ 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 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 $\tilde{A}_i/\|\tilde{A}_i\|_2 = (A_i^\top \boldsymbol{a}_i)^\top$, where $\|\cdot\|_2$ denotes the operator norm, *i.e.*, the



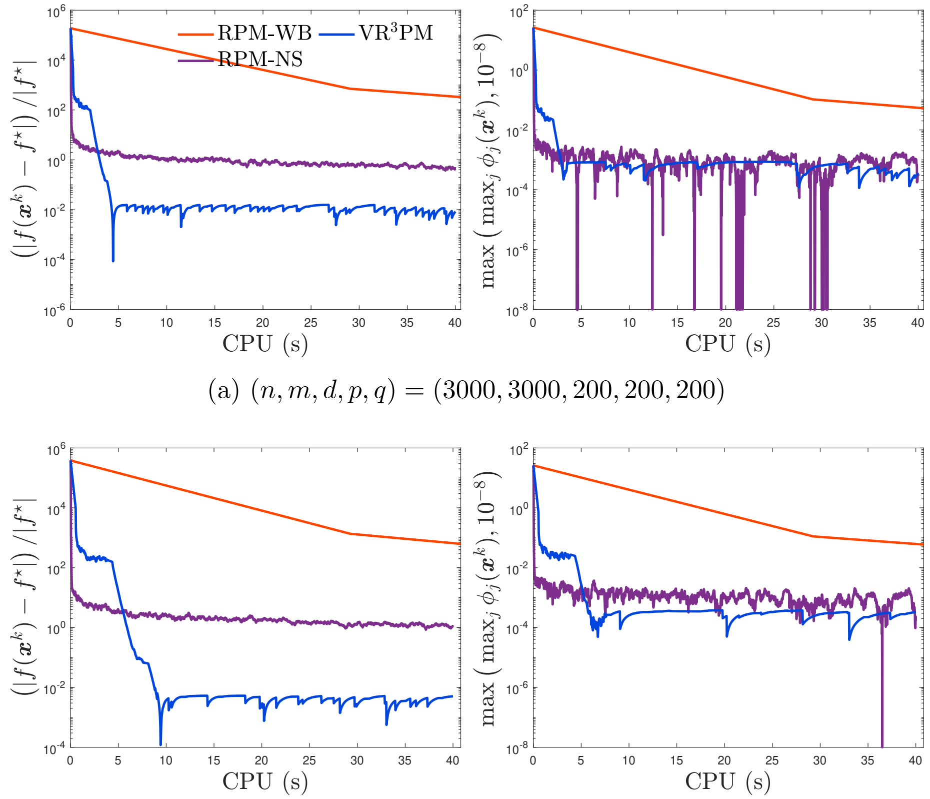
(a) $(n, m, d, p, q) = (3000, 3000, 200, 200, 200)$

(b) $(n, m, d, p, q) = (6000, 6000, 200, 200, 200)$

Figure 1: Comparison of VR^{3}PM and other RPMs using vanilla gradient estimators on problem (5).

largest singular value. The pairs $\{(B_j, b_j)\}_{j\in[m]}$ are generated in the same manner. The constants $w_1, \ldots, w_m$ are i.i.d. uniform random variables on $[0, 0.5]$. We compare VR^{3}PM with three variants of random relaxed projection methods (R^{2}PMs) obtained by replacing the SVRG gradient estimator $\boldsymbol{v}^k$ in Algorithm 1 with the standard gradient estimator using a single summand, the mini-batch gradient estimator using b summands and the full gradient using all n summands. These three R^{2}PMs are denoted, respectively, as R^{2}PM-1, R^{2}PM-b and R^{2}PM-n. The step-sizes for all algorithms are chosen to scale as $\alpha_k = \mathcal{O}(k^{-0.51})$ with optimally tuned constants. The constraint grouping technique is applied to all tested algorithms. Specifically, we group the constraints into $\bar{m} = m/10$ groups of $\bar{b} = 10$ constraints.

The results on problem (5) with parameters $(n, m, d, p, q) = (3000, 3000, 200, 200, 200)$ and $(n, m, d, p, q) = (6000, 6000, 200, 200, 200)$ are shown in Figure 1, where the left and right panels show the optimality gap and constraint violation against the CPU time (in second), respectively. We can see from Figure 1 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.



(a) $(n, m, d, p, q) = (3000, 3000, 200, 200, 200)$

(b) $(n, m, d, p, q) = (6000, 6000, 200, 200, 200)$

Figure 2: Comparison of VR^{3}PM and the RPMs in [22] and [33] on problem (5).

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.

5.2. Comparison with the RPMs in [22] and [33]

Our second experiment aims at comparing the performance of VR^{3}PM with the existing RPMs in [22] by Necoara and Singh and in [33] by Wang and Bertsekas, denoted as RPM-NS, and RPM-WB, respectively. RPM-WB is not amenable to the constraint grouping technique. So, we apply the technique only to our algorithm and RPM-NS. The constraint group size and step-sizes are chosen similarly as in Section 5.1.

The results on problem (5) with the same setting and parameters as in Section 5.1 are shown in Figure 2, where the left and right panels show the optimality gap and constraint violation against the CPU time (in second), respectively. From Figure 2, we can see that our algorithm VR^{3}PM performs substantially better than the competing algorithms RPM-NS and RPM-WB in terms of objective value. As for constraint violation, our algorithm and RPM-NS perform on par, but both better than RPM-WB.

5.3. Applications to Downlink Beamforming

Our next experiment concerns the downlink beamforming problem in wireless communication. Specifically, we consider a single base station equipped d antennas, transmitting data stream to n users. For each user $i \in [n]$, the signal received is given by $y_i = \boldsymbol{h}_i^{\mathsf{H}} \boldsymbol{x}_i + z_i$, where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose, $\boldsymbol{h}_i \in \mathbb{C}^d$ models the downlink channel of user i, $z_i \in \mathbb{C}$ is the error, and $\boldsymbol{x}_i \in \mathbb{C}^d$ represents the beamformer for user i. The quality of the signal received at user i can be measured by the signal-to-interference-plus-noise ratio (SINR):

$$\mathrm{SINR}_i = \frac{|\boldsymbol{h}_i^{\mathsf{H}} \boldsymbol{x}_i|^2}{\sum_{j \neq i} |\boldsymbol{h}_j^{\mathsf{H}} \boldsymbol{x}_j| + \sigma^2},$$

where σ^2 is the noise variance. One formulation of the beamforming problem is to minimize the transmission power subject to SINR constraints [21]:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \\ \text{s.t.} \quad & \mathrm{SINR}_i \geq \gamma_i, \quad i \in [n], \end{aligned} \tag{6}$$

where $\gamma_1, \dots, \gamma_n > 0$ are the desired SINRs. Problem (6) is non-convex since the SINR constraints are non-convex. However, it is shown in [35] that problem (6) is equivalent to a convex, second-order cone program over real numbers:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \|\hat{\boldsymbol{x}}_i\|^2 \\ \text{s.t.} \quad & \hat{\boldsymbol{h}}_i^{\top} \hat{\boldsymbol{x}}_i \geq \sqrt{\gamma_i \textstyle\sum_{j \neq i} (\hat{\boldsymbol{h}}_j^{\top} \hat{\boldsymbol{x}}_j)^2 + \gamma_i \sigma^2}, \; i \in [n], \end{aligned} \tag{7}$$

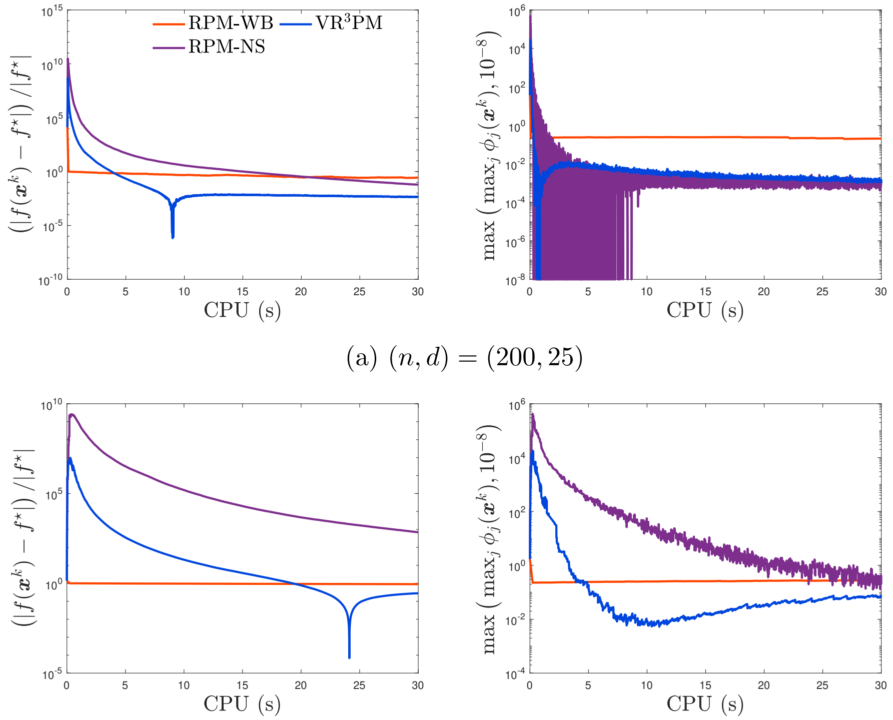
where $\hat{\boldsymbol{h}}_i \in \mathbb{R}^{2d}$ and $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{2d}$ are real vectors obtained by stacking the real and imaginary parts of $\boldsymbol{h}_i$ and $\boldsymbol{x}_i$, respectively, for $i \in [n]$.

In this experiment, we set $\sigma = 1$ and the SINR threshold for all users to be -13 dB (*i.e.*, $\gamma_i \approx 0.0501$). The vectors $\hat{\boldsymbol{h}}_1, \dots, \hat{\boldsymbol{h}}_n$ are i.i.d. $2d$-dimensional standard Gaussian random distribution with zero mean and identity covariance.

The results on problem (7) with parameters $(n, d) = (200, 25)$ and $(n, d) = (600, 20)$ are shown in Figure 3, where we can see that VR^{3}PM performs better than RPM-NS and RPM-WB in terms of both optimality and feasibility gaps.

5.4. Applications to Robust Classification

Finally, we compare our algorithm against the RPM-NS and RPM-WB on a distributionally robust classification problem with using real data-sets. To begin, we define a distance



(a) $(n, d) = (200, 25)$

(b) $(n, d) = (600, 20)$

Figure 3: Comparison of VR^{3}PM and the RPMs in [22] and [33] on problem (7).

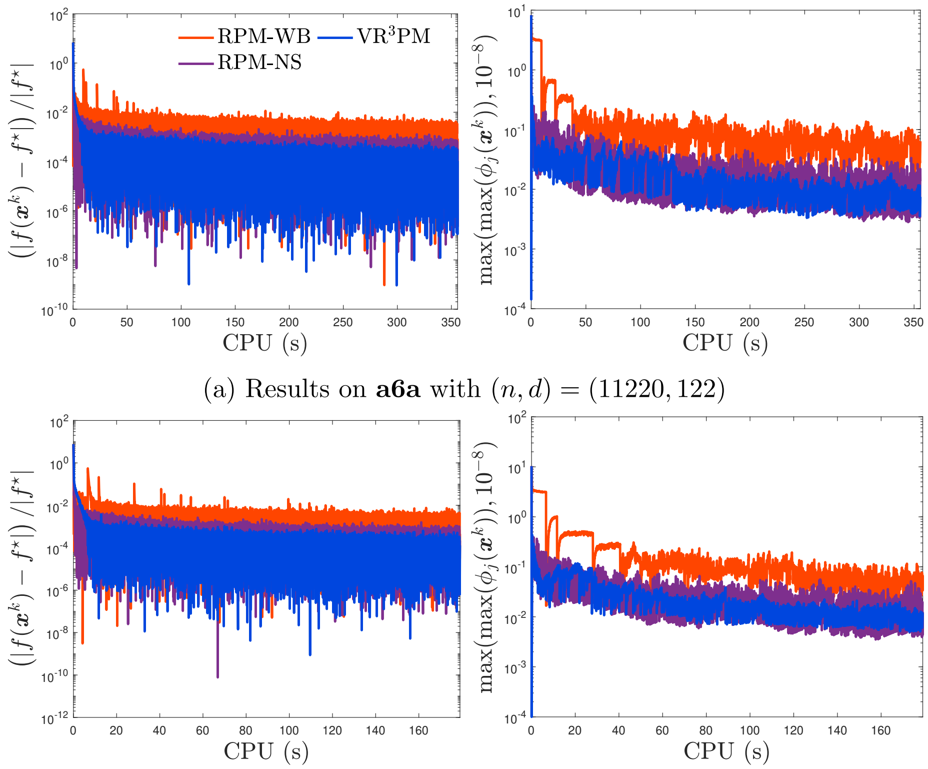
between two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$:

$$\begin{aligned} & W(\mathbb{P}_1, \mathbb{P}_2) \\ & = \min_{\mathbb{Q} \in \Gamma(\mathbb{P}_1, \mathbb{P}_2)} \int_{\Xi \times \Xi} \left(\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_\infty + |y_1 - y_2| \right) \mathrm{d}\mathbb{Q}((\boldsymbol{w}_1, y_1), (\boldsymbol{w}_2, y_2)), \end{aligned}$$

where $\Xi = \mathbb{R}^d \times \{+1, -1\}$ and $\Gamma(\mathbb{P}_1, \mathbb{P}_2)$ is the set of joint distributions on $\Xi \times \Xi$ with first marginal $\mathbb{P}_1$ and second marginal $\mathbb{P}_2$. The distance W is an instance of Wasserstein distance [37], specialized to the context of classification problems. The ball centered at $\widehat{\mathbb{P}}$ of radius $\epsilon > 0$ in the space of probability distributions is denoted by $B_\epsilon(\widehat{\mathbb{P}})$. Based on ideas from distributionally robust optimization [13], the following model for binary classification is considered in [31]:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \sup_{\mathbb{P} \in B_\epsilon(\widehat{\mathbb{P}})} \mathbb{E}_{(\boldsymbol{w}, y) \sim \mathbb{P}} \left[\log \left(1 + \exp \left(-y \, \boldsymbol{x}^\top \boldsymbol{w} \right) \right) \right],$$

which is intractable in general, due to the infinite-dimensionality of the maximization. However, if the center $\widehat{\mathbb{P}}$ is an empirical distribution associated with a sample $\{(\boldsymbol{w}_i, y_i)\}_{i \in [n]}$,



(a) Results on **a6a** with $(n, d) = (11220, 122)$

(b) Results on **a7a** with $(n, d) = (16100, 122)$

Figure 4: Comparison of VR^{3}PM and the RPMs in [22] and [33] on problem (8).

then by [31] it is equivalent to the problem

$$\begin{aligned} \min \quad & \lambda\epsilon + \frac{1}{n}\sum_{i\in[n]} s_i + \log\left(1 + \exp\left(-y_i\, \boldsymbol{x}^\top \boldsymbol{w}_i\right)\right) \\ \text{s.t.} \quad & \max\left(y_i\, \boldsymbol{x}^\top \boldsymbol{w}_i - \lambda,\, 0\right) \le s_i,\ \ i \in [n] \\ & \|\boldsymbol{x}\|_2 \le \lambda,\ \ \boldsymbol{x} \in \mathbb{R}^d,\ \boldsymbol{s} \in \mathbb{R}^n,\ \lambda \in \mathbb{R}. \end{aligned} \tag{8}$$

In this experiment, we use real data-sets[*] **a6a** with $(n, d) = (11220, 122)$ and **a7a** with $(n, d) = (16100, 122)$. In both data-sets, the sample size n is very large. This makes problem (8) extremely computationally challenging as the number of decision variables, the number of summands in the objective and the number of constraints all increase linearly in n. We compare our algorithm VR^{3}PM against RPM-NS and RPM-WB with all the algorithmic parameters similarly chosen.

The results are presented in Figure 4, where we can see that VR^{3}PM performs better than the two competing RPMs in terms of both the optimality gap and constraint violation.

[*] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

A. Lemmas

The following are standard results on projections.

Lemma 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 $\boldsymbol{y} \in D$,

$$\|\Pi_D(\boldsymbol{x}) - \boldsymbol{y}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \Pi_D(\boldsymbol{x})\|^2.$$

Lemma 2. Let $c \in \mathbb{R}$, $\boldsymbol{\zeta} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ 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 are useful to our development, see [23, 33] for example.

Lemma 3. Let $\{a^k\}$, $\{u^k\}$, $\{t^k\}$ and $\{d^k\}$ be sequences of nonnegative random variables satisfying

$$\begin{aligned} &\mathbb{E}[a^{k+1} | a^0, \cdots, a^k, u^0, \cdots, u^k, t^0, \cdots, t^k, d^0, \cdots, d^k] \\ \leq &(1 + t^k)a^k - u^k + d^k \quad \text{for all} \quad k \geq 0 \; a.s. \end{aligned}$$

Suppose that

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

Then,

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

for some non-negative random variable a.

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 4. Suppose that Assumptions 2–3 hold. Then, Algorithm 1 satisfies that for any $k \geq 0$,

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

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

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

where the inequality follows from Assumption 2. Therefore, using Assumption 3, we obtain

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

which completes the proof. □

The next lemma should be compared with [33, Assumption 1(c)] and, to some extent, illustrates why the SVRG gradient estimator cannot be directly used in the algorithmic framework of [33].

Lemma 5. 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

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

Proof. We have

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

where the equality follows from the definition of $\boldsymbol{v}^k$, the second inequality follows from Assumption 1(iii), $\boldsymbol{x}^k \in C_0$ and $\tilde{\boldsymbol{x}}^l \in C_0$, and the last inequality follows from inequality (2). This completes the proof. □

The following technical lemma is also used in the proof of our main results.

Lemma 6. 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

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

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

$$\begin{aligned}
&2\alpha_k \mathbb{E}\left[\langle \boldsymbol{v}^k,\, \boldsymbol{x}^k - \boldsymbol{x}^\star \rangle \mid \mathcal{F}_k\right] \\
=& 2\alpha_k \langle \nabla f(\boldsymbol{x}^k),\, \boldsymbol{x}^k - \boldsymbol{x}^\star \rangle \geq 2\alpha_k (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)) \\
=& 2\alpha_k \left(f(\boldsymbol{x}^k) - f(\Pi_C(\boldsymbol{x}^k)) + f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right) \\
\geq& 2\alpha_k \langle \nabla f(\Pi_C(\boldsymbol{x}^k)),\, \boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k) \rangle \\
&+ 2\alpha_k \left(f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right),
\end{aligned} \tag{9}$$

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

$$\begin{aligned}
&2\alpha_k \langle \nabla f(\Pi_C(\boldsymbol{x}^k)),\, \boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k) \rangle \\
\geq& - 2\alpha_k \|\nabla f(\Pi_C(\boldsymbol{x}^k)) - \nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\
&- 2\alpha_k \|\nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\
\geq& - 2\alpha_k \left(L(\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| + 1)\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|\right) \\
&- 2\alpha_k \|\nabla f(\boldsymbol{x}^k)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\
\geq& - 2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\alpha_k L \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\
&- 2\alpha_k \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\| \\
&- 2\alpha_k \|\nabla f(\boldsymbol{x}^\star)\| \|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|,
\end{aligned} \tag{10}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second from inequality (2), $\boldsymbol{x}^k \in C_0$ and $\Pi_C(\boldsymbol{x}^k) \in C \subseteq C_0$, and the third from the triangle inequality. We then bound the second, third and fourth terms on the last line of (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

$$\begin{aligned}&-2\alpha_k L\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|\\ \geq\ &-\alpha_k^2L^2\lambda-\frac{1}{\lambda}\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2.\end{aligned}\tag{11}$$

The third term can be bounded as

$$\begin{aligned}&-2\alpha_k\|\nabla f(\boldsymbol{x}^k)-\nabla f(\boldsymbol{x}^\star)\|\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|\\ \geq&-2\alpha_k L(\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|+1)\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|\\ \geq&-\alpha_k^2L^2\lambda\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2-\frac{1}{\lambda}\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2\\ &-2\alpha_k L\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|,\\ \geq&-\alpha_k^2L^2\lambda\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2-\alpha_k^2L^2\lambda\\ &-\frac{2}{\lambda}\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2,\end{aligned}\tag{12}$$

where the last inequality follows from (11). And the fourth term can be bounded as

$$\begin{aligned}&-2\alpha_k\|\nabla f(\boldsymbol{x}^\star)\|\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|\\ \geq&-\alpha_k^2\lambda\|\nabla f(\boldsymbol{x}^\star)\|^2-\frac{1}{\lambda}\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2.\end{aligned}\tag{13}$$

Substituting inequalities (10), (11), (12) and (13) into (9) yields the desired result. □

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

Lemma 7. Suppose that Assumptions 1–3 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

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

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

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

where the first inequality follows from the definition of $\boldsymbol{y}^{k+1}$, Lemma 1(i) and the fact that $\boldsymbol{x}^\star \in C_0$, and the second inequality from the definition of $\boldsymbol{z}^k$, Lemma 1(ii) and the fact that $\boldsymbol{x}^\star \in H_k$. Taking conditional expectation on the last inequality and using Lemmas 5 and 6 yields,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] \\
\leq\, & \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2\alpha_k \mathbb{E}\left[\langle \boldsymbol{v}^k, \boldsymbol{x}^k - \boldsymbol{x}^\star\rangle \mid \mathcal{F}_k\right] \\
& + \mathbb{E}\left[\alpha_k^2 \|\boldsymbol{v}^k\|^2 - \|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k\right] \\
\leq\, & (1 + \alpha_k^2(8L^2 + L^2\lambda))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 \\
& + (2\alpha_k L + \tfrac{4}{\lambda})\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \\
& + \alpha_k^2(2\lambda L^2 + (\lambda + 4)\|\nabla f(\boldsymbol{x}^\star)\|^2 + 12L^2) \\
& + 16L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 - 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f^\star) \\
& - \mathbb{E}\left[\|\Pi_{H_k}(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \mid \mathcal{F}_k\right].
\end{aligned} \tag{14}$$

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 (14). First, from the triangle inequality and Lemma 1(i), we have that

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

which together with Lemma 5 yields that

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

Rearranging the above inequality gives

$$\begin{aligned}
& - \mathbb{E}\left[\|\boldsymbol{z}^k - \Pi_{H_k}(\boldsymbol{z}^k)\|^2 \mid \mathcal{F}_k\right] \\
\leq\, & 16L^2\alpha_k^2\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{1}{2}\mathbb{E}\left[\|\Pi_{H_k}(\boldsymbol{x}^k) - \boldsymbol{x}^k\|^2 \mid \mathcal{F}_k\right] \\
& + 32L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \alpha_k^2(24L^2 + 8\|\nabla f(\boldsymbol{x}^\star)\|^2).
\end{aligned} \tag{15}$$

Plugging (15) into (14) and using Lemma 4, we obtain

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

This completes the proof. □

B. Proof of Proposition 1

The proof of case (i) can be founded in [10]. We therefore prove only case (ii). If $\boldsymbol{x}\in C$, then (3) trivially holds. It suffices to assume that $\boldsymbol{x}\notin C$. Let $I(\boldsymbol{x})$ be the set defined by

$$I(\boldsymbol{x})=\{j\in[m]\mid \phi_j(\boldsymbol{x})>0\}. \tag{16}$$

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. By [15, Lemma 2.3] and [28, Theorem 10], there exists a constant $\gamma>0$ such that

$$\operatorname{dist}(\boldsymbol{x},C)\leq\gamma\phi_{j'}(\boldsymbol{x}), \tag{17}$$

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 [2, 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 $\boldsymbol{\xi}_{j'}\neq\boldsymbol{0}$,

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

where the first inequality follows from the definition (2) 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 $\boldsymbol{\xi}_{j'}\in\partial\phi_{j'}(\boldsymbol{x})$, $\phi_{j'}(\boldsymbol{x})\leq\eta\operatorname{dist}(\boldsymbol{x},\,H(\phi_{j'};\,\boldsymbol{x};\,\boldsymbol{\xi}_{j'}))$. We therefore 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 (17), implies that

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

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

C. Proof of Theorem 1

We first prove that the sub-sequence $\{\tilde{\boldsymbol{x}}^l\}$ converges. Since $\mu_l \to 0$ as $l \to \infty$, there exists $l_0 \geq 0$ such that $2\mu_l L \leq \frac{\rho}{8m\kappa^2}$ for any $l \geq l_0$. Take $\lambda = 32m\kappa^2\rho^{-1}$. Then, for any for any $l \geq l_0$,

$$\frac{\rho}{2m\kappa^2} - 2\mu_l L - \frac{4}{\lambda} \geq \frac{\rho}{4m\kappa^2} > 0.$$

Fix any optimal solution $\boldsymbol{x}^\star \in X^\star$. It follows from Lemma 7 and the definition of α_k that for all $l \geq l_0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] \\
\leq\, &(1 + \mathcal{O}(\alpha_k^2))\, \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2) + \mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 \\
&- \frac{\rho}{4m\kappa^2}\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 - 2\mu_l(f(\Pi_C(\boldsymbol{x}^k)) - f^\star) \\
\leq\, &(1 + \mathcal{O}(\mu_l^2))\, \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\mu_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\mu_l^2).
\end{aligned} \tag{18}$$

Using the tower property of conditional expectation [32, Theorem 2.3.2(iii)] and inequality (18),

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{lr+r} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}\right] \\
=\, &\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{x}^{lr+r-1+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr+r-1}\right] \middle| \mathcal{F}_{lr}\right] \\
\leq\, &(1 + \mathcal{O}(\mu_l^2))\mathbb{E}\left[\|\boldsymbol{x}^{lr+r-1} - \boldsymbol{x}^\star\|^2 \middle| \mathcal{F}_{lr}\right] \\
&\quad + \mathcal{O}(\mu_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\mu_l^2) \\
&\ \vdots \\
\leq\, &(1 + \mathcal{O}(\mu_l^2))^{r-1}\mathbb{E}[\|\boldsymbol{x}^{lr+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}] \\
&\quad + \left(\mathcal{O}(\mu_l^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2\right)\sum_{s=0}^{r-2}(1 + \mathcal{O}(\mu_l^2))^s \\
&\quad + \mathcal{O}(\mu_l^2)\sum_{s=0}^{r-2}(1 + \mathcal{O}(\mu_l^2))^s.
\end{aligned} \tag{19}$$

Similarly, we get

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}^{lr+1}-\boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}\right]\\ \le\,&(1+\mathcal{O}(\mu_l^2))\,\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2-\frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l-\Pi_C(\tilde{\boldsymbol{x}}^l)\|^2\\ &+\mathcal{O}(\mu_l^2)-2\mu_l(f(\Pi_C(\tilde{\boldsymbol{x}}^l))-f^\star).\end{aligned}\tag{20}$$

Since $\mu_l<1$ for sufficiently large l, $(1+\mathcal{O}(\mu_l^2))^s\le 1+\mathcal{O}(\mu_l^2)$ for any l and $s\in[r]$. We thus have

$$\begin{aligned}&\mathbb{E}\left[\|\tilde{\boldsymbol{x}}^{(l+1)}-\boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}\right]\\ \le\,&(1+\mathcal{O}(\mu_l^2))^r\mathbb{E}\left[\|\boldsymbol{x}^{lr+1}-\boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{lr}\right]\\ &+\Big(\mathcal{O}(\mu_l^2)\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\mathcal{O}(\mu_l^2)\Big)\sum_{s=0}^{r-1}(1+\mathcal{O}(\mu_l^2))^s\\ &-\frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l-\Pi_C(\tilde{\boldsymbol{x}}^l)\|^2-2\mu_l(f(\Pi_C(\tilde{\boldsymbol{x}}^l))-f^\star)\\ \le\,&(1+\mathcal{O}(\mu_l^2))\,\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2+\mathcal{O}(\mu_l^2)\\ &-\frac{\rho}{4m\kappa^2}\|\tilde{\boldsymbol{x}}^l-\Pi_C(\tilde{\boldsymbol{x}}^l)\|^2,\end{aligned}\tag{21}$$

where the first inequality follows by substituting (20) into (19). Note that now the constants hidden in the big-O notation could possibly depend on r, but they are independent of μ_l or l. Applying Lemma 3 to the recursion (21), we have that the sequence $\{\|\tilde{\boldsymbol{x}}^l-\boldsymbol{x}^\star\|^2\}$ converges almost surely, that

$$\sum_{l=0}^{\infty}\mu_l[f(\Pi_C(\tilde{\boldsymbol{x}}^l))-f^\star]<\infty \quad a.s.,\tag{22}$$

and that

$$\sum_{l=0}^{\infty}\|\tilde{\boldsymbol{x}}^l-\Pi_C(\tilde{\boldsymbol{x}}^l)\|^2<\infty \quad a.s.\tag{23}$$

By inequality (22) and the fact that $\sum_{l=0}^\infty \mu_l=\infty$,

$$\liminf_{l\to\infty} f(\Pi_C(\tilde{\boldsymbol{x}}^l))=f^\star \quad a.s.\tag{24}$$

Also, inequality (23) implies that

$$\lim_{l\to\infty}\|\Pi_C(\tilde{\boldsymbol{x}}^l)-\tilde{\boldsymbol{x}}^l\|=0 \quad a.s.\tag{25}$$

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 $\{\tilde{\boldsymbol{x}}^{l_t}\}$ such that $\tilde{\boldsymbol{x}}^{l_t}\to\tilde{\boldsymbol{x}}^\star$ as $t\to\infty$. By relation (25) and continuity of $\Pi_C(\cdot)$, the sequence

$\Pi_C(\tilde{\boldsymbol{x}}^{l_t})$ converges almost surely to $\Pi_C(\tilde{\boldsymbol{x}}^\star) = \tilde{\boldsymbol{x}}^\star \in C$. It follows from (24) and the continuity of f that $f(\tilde{\boldsymbol{x}}^\star) = f^\star$. 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 $\lim_{l\to\infty} \tilde{\boldsymbol{x}}^l = \tilde{\boldsymbol{x}}^\star$. To prove the convergence in $\{\boldsymbol{x}^k\}$, by the boundedness of the sequence $\{\|\tilde{\boldsymbol{x}}^l - \tilde{\boldsymbol{x}}^\star\|^2\}$ and Lemma 7,

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

which, together with Lemma 3 and the fact that $\sum_{k=0}^\infty \alpha_k^2 < \infty$, implies 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 $\lim_{k\to\infty} \boldsymbol{x}^k = \tilde{\boldsymbol{x}}^\star$. This completes the proof.

D. Proof of Theorem 2

We first prove the convergence rate of the feasibility gap. Fix an arbitrary optimal solution $\boldsymbol{x}^\star \in X^\star$. By using Lemma 7 with $\lambda = 16m\kappa^2\rho^{-1}$ and the definition of α_k, we have that for all $k \ge 0$,

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

and hence that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right] \\
\le& (1 + \mathcal{O}(\alpha_k^2))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 \\
&+ \mathcal{O}(\alpha_k^2) - \frac{\rho}{8m\kappa^2}\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \\
&- 2\alpha_k(f(\Pi_C(\boldsymbol{x}^k)) - f^\star) \\
\le& (1 + \mathcal{O}(\alpha_k^2))\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2)\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2 + \mathcal{O}(\alpha_k^2).
\end{aligned} \tag{26}$$

Since C_0 is compact, the sequences $\{\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\}_k$ and $\{\|\tilde{\boldsymbol{x}}^l - \boldsymbol{x}^\star\|^2\}_l$ are bounded. Inequality (26) then implies that for any $k \ge 0$,

$$\begin{aligned}
&\frac{\rho}{8m\kappa^2}\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2\right] \\
\le& \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] + \mathcal{O}(\alpha_k^2) \\
&- \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_k\right]\right] \\
=& \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2\right] + \mathcal{O}(\alpha_k^2).
\end{aligned} \tag{27}$$

Summing the last inequality over k, we get

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

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

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

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

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

which, together with Lemma 5, implies that

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

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

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

Summing of the last inequality over k, we get

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[f(\boldsymbol{x}^k) - f^\star\right] \\
\leq &\frac{1}{2\tilde{\alpha}}\mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2\right] - \frac{\sqrt{K}}{2\tilde{\alpha}}\mathbb{E}\left[\|\boldsymbol{x}^K - \boldsymbol{x}^\star\|^2\right] \\
&+ \frac{1}{2\tilde{\alpha}} \sum_{k\in[K-1]} \left(\sqrt{k+1} - \sqrt{k}\right) \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right] \\
&+ \mathcal{O}\left(\sum_{k\in[K]} \frac{1}{\sqrt{k}}\right) \\
\leq &\mathcal{O}(1) + \mathcal{O}(1) \sum_{k\in[K-1]} \left(\sqrt{k+1} - \sqrt{k}\right) + \mathcal{O}\left(\sum_{k\in[K]} \frac{1}{\sqrt{k}}\right) \\
\leq &\mathcal{O}(1) + \mathcal{O}(\sqrt{K-1}) + \mathcal{O}(\sqrt{K}) = \mathcal{O}(\sqrt{K}),
\end{aligned} \tag{28}$$

where the second inequality follows from the boundedness of C_0, and the last 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 (28), 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.

E. Proof of Theorem 3

Similarly to (27), we get for any $k \geq 0$,

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

Summing the above inequality over k, we have

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

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

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

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

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

which, together with Lemma 5, implies that

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

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

$$\begin{aligned}
&\mathbb{E}\left[f(\boldsymbol{x}^k)-f^\star\right]\\
\leq\,&\frac{1}{2\alpha}\left(\mathbb{E}\left[\|\boldsymbol{x}^k-\boldsymbol{x}^\star\|^2\right]-\mathbb{E}\left[\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^\star\|^2\right]\right)+\mathcal{O}\left(\frac{1}{\sqrt{K}}\right).
\end{aligned}$$

Summing of the last inequality over k, we get

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

which, together with the convexity of f, yields

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

This completes the proof.

F. Proof of Theorem 4

By Theorem 1, $\{\boldsymbol{x}^k\}$ converges almost surely to a point in $X^\star$. Therefore, the sequence $\{\mathbb{E}[\|\boldsymbol{x}^k\|]\}_k$ is bounded, and by Lemma 1(i), inequality (2) and triangle inequality, so are the sequences $\{\mathbb{E}[\|\tilde{\boldsymbol{x}}^l\|]\}_l$, $\{\mathbb{E}[\|\Pi_{X^\star}(\boldsymbol{x}^k)\|]\}_k$, $\{\mathbb{E}[\|\boldsymbol{x}^k-\Pi_{X^\star}(\boldsymbol{x}^k)\|]\}_k$, $\{\mathbb{E}[\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|]\}_k$ and $\{\mathbb{E}\left[\|\nabla f(\Pi_{X^\star}(\boldsymbol{x}^k))\|^2\right]\}_k$. Using Lemma 1(i),

$$\begin{aligned}&\|\boldsymbol{x}^k-\Pi_{X^\star}(\boldsymbol{x}^k)\|\\ \le&\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|+\|\Pi_C(\boldsymbol{x}^k)-\Pi_{X^\star}(\Pi_C(\boldsymbol{x}^k))\|\\ &+\|\Pi_{X^\star}(\Pi_C(\boldsymbol{x}^k))-\Pi_{X^\star}(\boldsymbol{x}^k)\|\\ \le&2\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|+\|\Pi_C(\boldsymbol{x}^k)-\Pi_{X^\star}(\Pi_C(\boldsymbol{x}^k))\|,\end{aligned}$$

which implies that

$$\begin{aligned}&-\|\Pi_C(\boldsymbol{x}^k)-\Pi_{X^\star}(\Pi_C(\boldsymbol{x}^k))\|^2\\ \le&4\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2-\frac{1}{2}\|\boldsymbol{x}^k-\Pi_{X^\star}(\boldsymbol{x}^k)\|^2.\end{aligned}$$

Using this inequality, Assumption 4 and the fact that $\alpha_k\ge\frac{8}{\nu(k+1)}$, we have

$$\begin{aligned}&-\alpha_k(f(\Pi_C(\boldsymbol{x}^k))-f^\star)\\ \le&-\frac{\alpha_k\nu}{2}\|\Pi_C(\boldsymbol{x}^k)-\Pi_{X^\star}\left(\Pi_C(\boldsymbol{x}^k)\right)\|^2\\ \le&2\nu\alpha_k\|\boldsymbol{x}^k-\Pi_C(\boldsymbol{x}^k)\|^2-\frac{2}{k+1}\|\boldsymbol{x}^k-\Pi_{X^\star}(\boldsymbol{x}^k)\|^2.\end{aligned}$$

The last inequality and Lemma 7 with $\lambda = 32m\kappa^2\rho^{-1}$ yield that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \Pi_{X^\star}(\boldsymbol{x}^{k+1})\|^2 \mid \mathcal{F}_k\right] \\
\leq&\,\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \Pi_{X^\star}(\boldsymbol{x}^{k})\|^2 \mid \mathcal{F}_k\right] \\
\leq&\,(1 + \alpha_k^2(24L^2 + L^2\lambda))\|\boldsymbol{x}^k - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2 \\
&+ 48L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2 + 36L^2\alpha_k^2 \\
&+ \alpha_k^2\left(2\lambda L^2 + (\lambda+12)\|\nabla f(\Pi_{X^\star}(\boldsymbol{x}^k))\|^2\right) \\
&- (\tfrac{3\rho}{8m\kappa^2} - 2\alpha_k L)\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \\
&- \left(\alpha_k + \frac{8}{\nu(k+1)}\right)\left(f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right) \qquad (29) \\
\leq&\left(1 + \alpha_k^2(24L^2 + L^2\lambda) - \frac{2}{k+1}\right)\|\boldsymbol{x}^k - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2 \\
&- \frac{8}{\nu(k+1)}\left(f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right) \\
&- \left(\frac{3\rho}{8m\kappa^2} - 2\alpha_k(L+\nu)\right)\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2 \\
&+ \alpha_k^2\left(2\lambda L^2 + (\lambda+12)\|\nabla f(\Pi_{X^\star}(\boldsymbol{x}^k))\|^2 + 36L^2\right) \\
&+ 48L^2\alpha_k^2\|\tilde{\boldsymbol{x}}^l - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2.
\end{aligned}$$

Noting that

$$\begin{aligned}
&\frac{3\rho}{8m\kappa^2} - 2\alpha_k(L+\nu) + 4\alpha_k^2(L+\nu)^2 m\kappa^2\rho^{-1} \\
=&\,\frac{\rho}{8m\kappa^2} + \left(\sqrt{\frac{\rho}{4m\kappa^2}} - \sqrt{4m\kappa^2\rho^{-1}}\alpha_k(L+\nu)\right)^2 \\
\geq&\,\frac{\rho}{8m\kappa^2},
\end{aligned}$$

and taking total expectation on (29), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \Pi_{X^\star}(\boldsymbol{x}^{k+1})\|^2\right] \\
\leq &\left(1 - \frac{2}{k+1}\right)\mathbb{E}[\|\boldsymbol{x}^k - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2] \\
&- \frac{8}{\nu(k+1)}\mathbb{E}\left[f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right] + \mathcal{O}(\alpha_k^2) \\
&- \left(\frac{3\rho}{8m\kappa^2} - 2\alpha_k(L+\nu)\right)\mathbb{E}[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2] \\
\leq &\left(1 - \frac{2}{k+1}\right)\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_{X^\star}(\boldsymbol{x}^k)\|^2\right] \\
&- \frac{8}{\nu(k+1)}\mathbb{E}\left[f(\Pi_C(\boldsymbol{x}^k)) - f^\star\right] \\
&- \frac{\rho}{8m\kappa^2}\mathbb{E}\left[\|\boldsymbol{x}^k - \Pi_C(\boldsymbol{x}^k)\|^2\right] + \frac{\mathcal{O}(1)}{(k+1)^2},
\end{aligned}$$

where we used the boundednes of the sequences mentioned at the beginning of the proof. Multiplying both sides of the last inequality by $(k+1)k$, we have that for all $K \geq k$,

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

Summing the above inequality over k, we get

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

which, together with the convexity of $\|\cdot\|^2$ and f, implies that

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\boldsymbol{x}}^K - \hat{\boldsymbol{x}}^K\|^2\right] \leq \frac{K}{S_K}\mathcal{O}(1) \quad \text{and} \\
&\mathbb{E}\left[f(\hat{\boldsymbol{x}}^K) - f^\star\right] \leq \frac{K(K+1)}{S_K}\mathcal{O}(1),
\end{aligned} \tag{30}$$

where $\bar{\boldsymbol{x}}^K = \frac{1}{S_K}\sum_{k\in[K]} k(k+1)\boldsymbol{x}^k$, $\hat{\boldsymbol{x}}^K = \frac{1}{S_K}\sum_{k\in[K]} k(k+1)\Pi_C(\boldsymbol{x}^k)$ and

$$S_K = \sum_{k\in[K]}(k^2+k) \geq \frac{1}{6}K^3,$$

which, together with (30), implies that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2(\bar{\boldsymbol{x}}^K, C)\right] \leq \mathbb{E}\left[\|\bar{\boldsymbol{x}}^K - \hat{\boldsymbol{x}}^K\|^2\right] \leq \mathcal{O}\left(\frac{1}{K^2}\right) \\
&\text{and } \mathbb{E}\left[f(\hat{\boldsymbol{x}}^K) - f^\star\right] \leq \mathcal{O}\left(\frac{1}{K}\right).
\end{aligned}$$

These two inequalities and the mean value theorem imply the existence of $\theta \in [0,1]$ such that

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\boldsymbol{x}}^K) - f^\star\right] \leq \mathbb{E}\left[|f(\bar{\boldsymbol{x}}^K) - f^\star|\right] \\
\leq &\mathbb{E}\left[f(\hat{\boldsymbol{x}}^K) - f^\star\right] + \mathbb{E}\left[|f(\bar{\boldsymbol{x}}^K) - f(\hat{\boldsymbol{x}}^K)|\right] \\
\leq &\mathcal{O}\left(\frac{1}{K}\right) + \mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K))\|\|\bar{\boldsymbol{x}}^K - \hat{\boldsymbol{x}}^K\|\right] \\
\leq &\mathcal{O}\left(\frac{1}{K}\right) + \frac{K}{2}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^K - \hat{\boldsymbol{x}}^K\|^2\right] \\
&+ \frac{1}{2K}\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K))\|^2\right] \\
\leq &\mathcal{O}\left(\frac{1}{K}\right) + \mathcal{O}\left(\frac{1}{K}\right)\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K))\|^2\right].
\end{aligned}$$

We get the desired result if the sequence $\{\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K))\|^2\right]\}_K$ is bounded. To prove the boundedness,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K))\|^2\right] \\
\leq &2\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K)) - \nabla f(\boldsymbol{x}^0)\|^2\right] + 2\|\nabla f(\boldsymbol{x}^0)\|^2 \\
\leq &2L^2\mathbb{E}\left[(\|\bar{\boldsymbol{x}}^K + \theta(\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K) - \boldsymbol{x}^0\| + 1)^2\right] + 2\|\nabla f(\boldsymbol{x}^0)\|^2 \\
\leq &\mathcal{O}(1) + 8L^2\mathbb{E}\left[(\|\bar{\boldsymbol{x}}^K - \boldsymbol{x}^0\|^2 + \|\hat{\boldsymbol{x}}^K - \bar{\boldsymbol{x}}^K\|^2)\right] \\
\leq &\mathcal{O}(1) + 16L^2\|\boldsymbol{x}^0\|^2 + \mathcal{O}\left(\frac{1}{K}\right) \\
&+ \frac{16L^2}{S_K}\sum_{k\in[K]}(k+1)k\mathbb{E}\left[\|\boldsymbol{x}^k\|^2\right] \leq \mathcal{O}(1),
\end{aligned}$$

where the second inequality follows from (2), the third from $\theta \in [0,1]$, the fourth from the convexity of $\|\cdot\|^2$, and the last from the boundedness of the sequence $\left\{\mathbb{E}\left[\|\boldsymbol{x}^k\|^2\right]\right\}_k$. This completes the proof.

References

1. H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

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

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

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

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

6. 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.

7. 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.

8. 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.

9. 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.

10. A. J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49:263–265, 1952.

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

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

13. 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.

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

15. W. Li. Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities. *SIAM Journal on Optimization*, 7(4):966–978, 1997.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Z.-Q. Luo and W. Yu. An introduction to convex optimization for communications and signal processing. *IEEE Journal on Selected Areas in Communications*, 24:1426–1438, 2006.

22. I. Necoara and N. K. Singh. Stochastic subgradient projection methods for composite optimization with functional constraints. *Journal of Machine Learning Research*, 23:1–35, 2022.

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

24. A. Nedić and I. Necoara. Random minibatch subgradient algorithms for convex problems with functional constraints. *Applied Mathematics and Optimization*, 80:801–833, 2019.

25. 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.

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

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

28. J.-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1-3):299–332, 1997.

29. B. T. Polyak. Random algorithms for solving convex inequalities. *Studies in Computational Mathematics*, 8:409–422, 2001.

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

31. S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

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

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

34. M. Wang, Y. Chen, J. Liu, and Y. Gu. Random multi-constraint projection: Stochastic gradient methods for convex optimization with many constraints. *arXiv preprint arXiv:1511.03760*, 2015.

35. A. Wiesel, Y. C. Eldar, and S. Shamai. Linear precoding via conic optimization for fixed MIMO receivers. *IEEE Transactions on Signal Processing*, 54:161–176, 2006.

36. 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.

37. M.-C. Yue, D. Kuhn, and W. Wiesemann. On linear optimization over Wasserstein balls. *Mathematical Programming*, 195(1):1107–1122, 2022.

38. 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.

39. 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.