Stochastic approximation algorithms for DR-submodular maximization with convex functional constraints [*]

Jiachen Ju [†] Xiao Wang [‡] Dachuan Xu [§]

July 11, 2024

Abstract

In this paper, we develop a generic framework of stochastic approximation algorithms for DR-submodular maximization with convex functional constraints, where both objective and constraints take expectation formulations. Solving such type of problems can be challenging due to nonconvexity and submodularity of the objective, feasibility required by the functional constraints, as well as stochastic nature of the problem. To tackle these challenges, we leverage the augmented Lagrangian function associated with the original problem and build the algorithm framework based on stochastic approximations. Theoretical analysis is conducted under mild assumptions and proper update schemes and demonstrates that our proposed algorithm can reach $\left(1 - \min_{x\in X} \|x\|_\infty\right)/4$-approximation for problems with non-monotone objective and $(1 - 1/e)$-approximation when the objective is monotone. And for both cases the expected errors and constraint violations are all bounded by $O(I^{-1/3})$, where I represents the total number of samples. To the best of our knowledge, the study on stochastic approximation algorithms for DR-submodular maximization with convex functional constraints seems new in the literature.

Keywords: DR-submodular maximization, Functional constraints, Augmented Lagrangian, Stochastic approximation, Approximation ratio

1 Introduction

In many real-world applications, decision makers may encounter situations where the objective and constraint functions are not deterministic, but involve uncertainty, either influenced by some random variables or dependent on a large data set. For instance, in financial portfolio [32], decision makers are often tasked to take into account the probability of asset returns and balance the relationship between investment returns and risk, and thus find an optimal investment strategy under uncertain market conditions. In supply chain networks [2], one needs to ascertain inventory levels and order quantities amidst uncertain conditions, such as demand, delivery time, and supply interruptions, in order to maximize the reward function while adhering to service level constraints. Submodularity was originally proposed for optimization over discrete sets [15], and arises in a broad modern applications, such as sensor placement [25], resource allocation [34], and so forth, including [18, 22, 34, 35, 38]. Extending the concept of set submodularity to continuous domains, DR-submodularity [6] portrays

[*] This research was partially supported by the National Natural Science Foundation of China (Nos. 12131003, 12271278) and the Major Key Project of PCL (No. PCL2022A05)

[†] Beijing University of Technology, Beijing, China; Peng Cheng Laboratory, Shenzhen, 518066, China.

[‡] wangx07@pcl.ac.cn , meng Cheng Laboratory, Shenzhen, 518066, China.

[§] Beijing University of Technology, Beijing, China.

a diminishing marginal return property in essence. DR-submodular optimization often models the variables contained in a convex set, whose structure is not specified. And this convex set is always assumed to be simple and easy to handle, that is, easy to project to ensure feasibility of points. This is managable in some easy situations, such as with bound constraints or ball constraints. In general, however, the feasibility to a convex set can be difficult to realize. Under such circumstances, it is acceptable to allow a certain degree of constraint violation, provided that it can be under control.

In this paper, we focus on DR-submodular maximization with convex functional constraints, formulated as

$$\begin{aligned} \max_{x\in X} \quad & f_0(x) \\ \text{s. t.} \quad & f_i(x) \le 0,\ i \in [m] := \{1, \dots, m\}, \end{aligned} \tag{1.1}$$

where function $f_0 : [0,1]^n \to \mathbb{R}_+$, is continuously differentiable and DR-submodular, and $f_i : [0,1]^n \to \mathbb{R},\ i \in [m]$ are convex but possibly nonsmooth. Here, $X \subseteq [0,1]^n$ is a simple convex closed set, and we assume that a strongly convex quadratic minimization over X can be easily solved. Functions are given by $f_0(x) = \mathbb{E}_\xi[F_0(x,\xi)],\ f_i(x) = \mathbb{E}_\xi[F_i(x,\xi)],\ i \in [m]$, where the functions $F_0 : [0,1]^n \times \Xi \to \mathbb{R}_+,\ F_i : [0,1]^n \times \Xi \to \mathbb{R},\ i = 1, \dots, m$, are continuous with respect to x for almost any $\xi \in \Xi$, where $\xi \in \Xi$ is a random vector following probability distribution P over Ξ and $\mathbb{E}_\xi$ represents the expectation taken with respect to ξ. Particularly, when ξ draws a value uniformly from a finite set $\{\xi_1, \dots, \xi_N\}$, associated functions will take finite-sum forms. Hence, the objective and constraint functions of (1.1) can be expressed as

$$f_0(x) = \int_\Xi F_0(x,\xi) d\mathrm{P}(\xi) \quad \text{or} \quad \frac{1}{N}\sum_{j=1}^N F_0(x,\xi_j),$$

$$f_i(x) = \int_\Xi F_i(x,\xi) d\mathrm{P}(\xi) \quad \text{or} \quad \frac{1}{N}\sum_{j=1}^N F_i(x,\xi_j),\ i \in [m].$$

The problem (1.1) arises in a wide range of applications, including data summarization [20], influence maximization [7], and MAP inference of Determinantal Point Processes (DPPs) [21].

In many scenarios, however, the distribution function of random variable may not be explicitly represented or the integral is expensive to compute, or N can be large, which causes that the exact function information of the objective and constraints can be hard to obtain. For any given realization of random vector ξ and $x \in X$, we may only obtain stochastic zeroth- and first-order information

$$F_0(x,\xi),\ F_i(x,\xi),\ \nu_0(x,\xi),\ \nu_i(x,\xi), i \in [m],$$

where $\nu_0(x,\xi)$ is a stochastic gradient of f_0 at x, while $\nu_i(x,\xi)$ is a stochastic subgradient of f_i at x for $i \in [m]$. Addressing problem (1.1) entails overcoming several key challenges: the submodularity of the objective function, feasibility to the functional constraints, and the inherent stochasticity which impedes access to exact function information. To characterize the properties of an approximation algorithm for (1.1), the approximation guarantee is another important issue that we need to address. Thus our goals in this paper are to develop an efficient stochastic approximation algorithm for the case we are interested in, which reaches the highest possible approximation guarantee with lowest possible error and constraint violation. More specifically, let x^* be the optimal solution of (1.1) and $\bar{x}$ be the output of an associated stochastic approximation algorithm for solving (1.1). We will conduct approximation analysis by investigating the approximation error $\gamma f_0(x^*) - \mathbb{E}[f_0(\bar{x})]$ and constraint violation $\mathbb{E}[\|[f(\bar{x})]_+\|]$, where $\gamma > 0$ and $f = (f_1, \dots, f_m)^T$.

1.1 Related Work

DR-submodular maximization

The past decade has witnessed the great development on DR-submodular maximization over a continuous convex domain. Based on differential equations, a continuous time framework, a multiplicative weight updates (MWU) template, is proposed in [10] for handling a class of DR-submodular optimization over a polyhedral. The proposed algorithm can find a solution with objective function reaching a $(1-1/e-\epsilon)$OPT in $\tilde{O}(\epsilon^{-4})$ function value oracles. Here OPT denotes the optimal value of the problem for interest. [6] study a Frank-Wolfe algorithm for maximizing monotone DR-submodular continuous functions under general down-closed convex constraints and guarantee $(1-1/e)$-approximation and sub-linear convergence rate. A non-monotone Frank-Wolfe algorithm with $1/e$-approximation guarantee and sublinear convergence rate is established in [5], aiming for non-monotone DR-submodular maximization under general down-closed convex constraints. [14] extend the deterministic model by discarding the assumption of downclosedness of the constraint set. They apply the core concept of the Frank-Wolfe algorithm and develop a method that achieves a $1/4$-approximation after $O(\epsilon^{-1})$ iteration. [17] propose a gradient algorithm by applying stable point theory and remove the down-closed assumption. A theoretical analysis is presented for both deterministic and stochastic settings with $1/2$-approximation in $O(\epsilon^{-1})$ iterations and $O(\epsilon^{-2})$ iterations, respectively. [26] also study stochastic conditional gradient methods for DR-submodular maximization. Approximation analyses are presented for several scenarios, including non-monotone case and monotone case. Specifically, for the monotone case with general convex set constraint the proposed algorithm can reach $(1-1/e)$-approximation, while for the non-monotone case with downclosed convex set constraint it owns $1/e$-approximation. For both cases the complexity is $\mathcal{O}(\epsilon^{-3})$. For the monotone case [16] develop a variant of stochastic conditional gradient method with $(1-1/e)$-approximation and improve the complexity to $O(\epsilon^{-2})$ in terms of linear optimization oracles. Later, a class of one sample type algorithms [42, see] are designed for both non-oblivious and oblivious setting, and it requires at most $O(\epsilon^{-2})$ stochastic gradient evaluations to achieve a $(1-1/e)$-approximation for monotone DR-submodular maximization. Recently, [30] and [43] introduce and elucidate a unified approach aimed at maximizing continuous DR-submodular functions. This approach spans a broad array of settings and is compatible with various types of oracle access. However, to the best of our knowledge the work on DR-submodular maximization with convex functional constraints is still rare in the literature.

Constrained stochastic optimization

Driven by practical application scenarios, stochastic optimization with functional constraints are attracting much interest. A series of methods such as stochastic SQP methods [4, 13, see], penalty methods [19, 36, see], proximal point methods [9, see], have been proposed and studied for stochastic optimization with deterministic constraints. Stochastic optimization with expectation constraints has also received much attention in recent years. Based on Polayk's subgradient method [31, see] and Nesterov's note [28, see], [23] introduce a cooperative stochastic approximation algorithm for stochastic convex optimization with a single expectation constraint, and establish the iteration complexity of $O(\epsilon^{-2})$ in terms of both optimality gap and constraint violation within $\epsilon > 0$. For convex optimization with multiple expectation constraints, [40] present a stochastic approximation proximal method of multipliers, and the proposed algorithm can achieve an objective regret and a constraint violation regret both of the order $O(T^{-1/2})$ after T iteration. Furthermore, the authors demonstrate that, with a probability of no less than $1-e^{-T^{1/4}}$, the algorithm maintains no more than $O(T^{-1/4})$ objective regret and no more than $O(T^{-1/8})$ constraint violation regret. Based on this work, [41] develop a stochastic augmented Lagrangian-type algorithm based on linearizations of the objective function and constraint functions. The algorithm displays an expected convergence rate of $O(K^{-1/2})$ for both the reduction of the objective function and the constraint violations, where K denotes the number of iterations. Also, from a high probability perspective, the algorithm achieves $O(\log(K)K^{-1/2})$ constraint violation bound and $O(\log^{3/2}(K)K^{-1/2})$ objective bound. Shortly after, [37] propose an adaptive primal-dual stochastic gradient method. At each iteration, an unbiased stochastic estimation of the subgradient of the Lagrangian function is inquired, later the primal variables and dual variables are updated following the adaptive SGM step and the vanilla SGM step respectively. Under the convexity assumption, an ergodic convergence rate of $O(k^{-1/2})$ is also verified in terms of the primal-dual gap and constraint violation, where k is the number of subgradient inquiries. [9] extend the constraint extrapolation method, which is a primal-dual type method for convex functional constrained optimization, within a proximal point algorithm framework to address non-convex functional constrained optimization problems. They establish an oracle complexity of $O(\epsilon^{-6})$, reaching an ϵ-KKT point in expectation. [24] examine the stochastic inexact augmented Lagrangian methods for addressing problems with a non-convex composite objective and non-convex smooth functional constraints, establishing a complexity of $O(\epsilon^{-5})$. [1] introduce single-loop algorithms based on the quadratic penalty method with the complexity guarantee of $\tilde{O}(\epsilon^{-5})$. [12] improves the complexity to $\mathcal{O}(\epsilon^{-5})$ by introducing a feasibility-pursuing phase to locate an approximately feasible initial point. Moreover, the moving average scheme to approximate constraint function values can ensure an $\mathcal{O}(I^{-1/5})$ error in expectation, where I denotes the number of oracle calls.

Online constrained stochastic optimization

In the field of online learning, some researchers have also turned their attention to stochastic constrained optimization problems. [39] propose a framework for online convex optimization with convex stochastic constraints. In their framework, the decision maker performs an action x_t first, and discloses the loss function $f_t(x)$ and the stochastic constraint function realizations $g_t(x;\omega(t))$ at tth round, $t \in [T]$. The authors give an upper bound of the discrete time stochastic process adapted to a filtration. They prove that the orders of both regret and constraint violations in expectation are $O(T^{1/2})$, while in high probability is $O(T^{1/2}\log(T))$. [33] also investigate the online framework under stochastic constraints, where the utility function f_t is assumed to be arbitrary from a class of monotone DR-submodular functions, the constraint function g_t is differentiable and randomly drawn following some unknown underlying distribution over a class of monotone convex functions. Taking inspiration from the Meta-Frank-Wolfe algorithms [11, 27, see], they derive an algorithm that achieves $(1-1/e)$-regret bound of $O(T^{1-\epsilon/2})$ when compared to a benchmark with a window length of $T^{1-\epsilon}$, and establishes a total constraint violation bound of $O(T^{1-\epsilon/2})$.

1.2 Our contributions

In this paper, we study Diminishing Returns (DR)-submodular maximization with convex functional constraints, where both objective function and constraints are in expectation formulations. It poses challenges to solve this type of problems since only stochastic approximations of the objective and constraint functions are available and the feasibility to constraints are hard to maintain. To address these challenges, we propose a generic algorithmic framework that exploits stochastic linear approximations of the objective and constraint functions. Under appropriate update schemes the proposed algorithm framework can be adapted to problems with objective being non-monotone and monotone, respectively. We show that for the non-monotone case a $(1-\min_{x\in X}\|x\|_\infty)/4$-approximation can be achieved with an expected error bounded by $O(I^{-1/3})$, where X represents the problem-dependent convex closed set and I refers to the total number of samples. And the expected constraint violation is in order $O(I^{-1/3})$. For the monotone case and establish the $(1-1/e)$-approximation and $\mathcal{O}(I^{-1/3})$-expected error and constraint violation, when X contains the zero vector. Finally, numerical results are reported on three illustrative examples and showcase the effectiveness and efficiency of the proposed algorithm.

1.3 Notations, assumptions and organization

We use $\mathbf{1}_d$ and $\mathbf{0}_d$ to denote the unit vector and zero vector in $\mathbb{R}^d$, respectively, and $e_j \in \mathbb{R}^d$ to refer the d-dimensional vector whose jth component is one and other components are zero. We define $[d] = \{1, \dots, d\}$ and $[0] = \emptyset$. For a given vector $x \in \mathbb{R}^d$, we denote $x(i)$ as ith component of x. For a real number a and a vector $x \in \mathbb{R}^d$, we denote $[a]_+ = \max\{a, 0\}$ and $[x]_+ = ([x(1)]_+, \dots, [x(d)]_+)^T$. Without any specification, we refer $\|\cdot\|$ to the standard Euclidean norm in the vector space for interest. We also denote $\sum_1^0 = 0$. The diameter of the convex set X is denoted by $\mathrm{d}(X) = \sup_{x,y\in X} \|x-y\|$. Additionally, for brevity we denote $f(x) := (f_1(x), \dots, f_m(x))^T$, $\nu(x,\xi) := (\nu_1(x,\xi), \dots, \nu_m(x,\xi))^T$ and $\nu(x) := \mathbb{E}_\xi[\nu(x,\xi)]$. Given $x, y \in \mathbb{R}^d$, $x \leq y$ refers to the element-wise inequality. Moreover, $f_0 : [0,1]^n \to \mathbb{R}_+$ is called DR-submodular, if for every $x \leq y \in [0,1]^n$, $i \in [n]$, $a \geq 0$ such that $x + ae_i$, $y + ae_i \in [0,1]^n$, it gives $f_0(x+ae_i) - f_0(x) \geq f_0(y+ae_i) - f_0(y)$.

We now give some standard assumptions that are exploited in the remainder of this paper.

Assumption 1.1. *Function $f_0 : [0,1]^n \to \mathbb{R}_+$ is DR-submodular and continuously differentiable with L_0-Lipschitz continuous gradients. Functions $F_i(\cdot,\zeta) : \mathbb{R}^n \to \mathbb{R}$, $i \in [m]$, are convex over $[0,1]^n$ for almost every $\zeta \in \Xi$.*

Assumption 1.2. *There exist positive constants M_2, M_3 such that for any $x \in X$, $\zeta \in \Xi$ and $i \in [m]$, the stochastic zeroth- and first-order approximations satisfy*

$$\mathbb{E}_\zeta[\nu_0(x,\zeta)] = \nabla f_0(x),\ \mathbb{E}_\zeta[\|\nu_0(x,\zeta)\|^2] \leq M_2^2,\ \mathbb{E}_\zeta[\nu_i(x,\zeta)] \in \partial f_i(x),\ \|\nu_i(x,\zeta)\| \leq M_3.$$

Assumption 1.3. *There exists a positive constant M_1 such that for all $x \in X$, $\mathbb{E}_\zeta[\|F(x,\zeta)\|^2] \leq M_1^2$, where $F(x,\zeta) := (F_1(x,\zeta), \dots, F_m(x,\zeta))^T$.*

Assumption 1.4. *(The Slater's Condition) There exist a positive constant M_4 and $\hat{x} \in X$ such that $f_i(\hat{x}) \leq -M_4$ for any $i \in [m]$.*

Assumption 1.5. *The set $\mathcal{X} := (\arg\min_{x\in X} \|x\|_\infty) \cap \{x \mid f(x) \leq \mathbf{0}_m\}$ is nonempty.*

Remark 1.1. *Under Assumption 1.1, it follows from (7.5) in [17] and Lemma 7 in [10] that for any $x, y \in [0,1]^n$,*

$$\begin{aligned} &\langle \nabla f_0(x), y - x\rangle \geq f_0(x \vee y) + f_0(x \wedge y) - 2f_0(x), \\ &f_0(x \vee y) \geq (1 - \|x\|_\infty)\, f_0(y), \end{aligned} \tag{1.2}$$

where $\vee$ and $\wedge$ are the coordinate-wise maximum and minimum operations, respectively. It is noteworthy that the expected boundedness required in Assumption 1.3 can be realized through the continuity of f_i over $[0,1]^n$ and the variance boundedness of $F_i(x,\zeta), i \in [m]$; that is, $\mathbb{E}_\zeta[\|F_i(x,\zeta) - f_i(x)\|^2]$ is upper bounded, which is commonly adopted in the literature for stochastic optimization.

Our paper is organized as follows. In Section 2, we present a framework of stochastic approximation algorithms for solving (1.1). In Section 3, we exhibit a series of auxiliary lemmas to support the theoretical analysis. In Sections 4 and 5, we delve into the approximation analysis, for the case when objective is non-monotone and monotone, respectively. In Section 6, we report numerical results on three illustrative problems. Finally, we draw a conclusion in Section 7.

2 Algorithm framework

In this section, we propose an algorithm framework of stochastic approximation methods for solving DR-submodular maximization with convex functional constraints, as described in (1.1). To handle the functional constraints, we exploit the augmented Lagrangian (AL) function, which is well-known in continuous optimization community [8, 29]. The AL function associated with (1.1) is defined as

$$\mathcal{L}_\beta(x;\lambda) = -f_0(x) + \frac{1}{2\beta}\left[\|[\lambda+\beta f(x)]_+\|^2 - \|\lambda\|^2\right], \tag{2.1}$$

where $\beta > 0$ is a penalty parameter and $\lambda \in \mathbb{R}^m_+$ is a vector of Lagrange multipliers, also known as dual variables. Keeping the penalty parameter and dual variables fixed, traditional augmented Lagrangian methods aim to (approximately) minimize the AL function at each iteration to update the primal variable. However, in the context of (1.1) only stochastic approximations of the objective function f_0 and constraint functions $f_i, i \in [m]$ are available. We thus need to employ these stochastic approximations to design an algorithm for (1.1).

Our algorithm is organized as a double-loop structure. We set the maximum iteration number in the outer loop and inner loop as T and K, respectively, and simply denote the total number of iterations by $I = TK$. The iteration index in the outer loop is denoted by t located at the subscript of vectors, while the iteration index in the inner loop is denoted by k, placed at the superscript of vectors. For any $t \in [T]$ and $k \in [K]$, let x_t^k and λ_t^k represent the primal iterate and the dual iterate, respectively. We also introduce a sequence of auxiliary variables v_t^k to approximate primal iterates. The basic idea of the algorithm is given as follows. For any $t \in [T]$ and $k \in [K]$, we first select a random sample $\xi_t^k \in \Xi$ and compute stochastic estimates $F_0(x_{t-1}^k, \xi_t^k)$, $\nu_0(x_{t-1}^k, \xi_t^k)$, $F(v_{t-1}^k, \xi_t^k)$ and $\nu(v_{t-1}^k, \xi_t^k)$. Then based on the AL function (2.1), we construct an approximation model to update the primal variable. Note that when x is not far from x_{t-1}^k and v_{t-1}^k, the original objective function and the constraint functions around current iterate can be approximated by

$$\begin{aligned} F_0(x,\xi_t^k) &\approx F_0(x_{t-1}^k,\xi_t^k) + \langle \nu_0(x_{t-1}^k,\xi_t^k), x - x_{t-1}^k\rangle, \\ F_i(x,\xi_t^k) &\approx F_i(v_{t-1}^k,\xi_t^k) + \langle \nu_i(v_{t-1}^k,\xi_t^k), x - v_{t-1}^k\rangle, \quad i \in [m]. \end{aligned}$$

It is noteworthy that we approximate the objective function value around current iterate x_{t-1}^k, while for the constraints we make approximations based on the function information of an approximate point v_{t-1}^k. This plays a crucial role in establishing desired approximation ratios in our paper. We then introduce the approximation function $Q_t^k(x)$ formulated as

$$\begin{aligned} Q_t^k(x) := &- F_0(x_{t-1}^k,\xi_t^k) + \langle -\nu_0(x_{t-1}^k,\xi_t^k), x - x_{t-1}^k\rangle + \frac{1}{2\beta}\sum_{i=1}^m \big[\lambda_t^k(i) + \beta\big(F_i(v_{t-1}^k,\xi_t^k) \\ &+ \langle \nu_i(v_{t-1}^k,\xi_t^k), x - v_{t-1}^k\rangle\big)\big]_+^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2 + \frac{\alpha}{2}\|x - v_{t-1}^k\|^2 \end{aligned} \tag{2.2}$$

and compute v_t^k by solving the minimization problem

$$v_t^k := \operatorname*{argmin}_{x\in X} Q_t^k(x). \tag{2.3}$$

Clearly, $Q_t^k(\cdot)$ is strongly convex over X, thus the subproblem in (2.3) admits a unique solution. We then compute a new primal iterate x_t^{k+1} through a mapping of x_t^k and v_t^k, denoted by

$$x_t^{k+1} = \mathcal{M}(x_t^k, v_t^k). \tag{2.4}$$

We will specify the setting of operator $\mathcal{M}$ in subsequent analysis regarding the non-monotone case and monotone case, respectively.

Finally, with regard to the update of dual iterates, we adopt the following way

$$\lambda_{t+1}^k = \big[\lambda_t^k + \beta\big(F(v_{t-1}^k, \xi_t^k) + \langle \nu(v_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k \rangle\big)\big]_+. \tag{2.5}$$

We formalize above considerations into the main algorithm of our paper, Algorithm 1. Algorithm 1 involves setting up the random process, which comes from the sampling of the random variables $\xi_t^k,\ t \in [T],\ k \in [K]$. During the algorithmic process, there are three kinds of computational expectations to take in our theoretical analysis. The first expectation is the conditional expectation with respect to ξ_t^k given $\mathcal{F}_t^k$, denoted by $\mathbb{E}_{\xi_t^k}[\cdot]$, where

$$\mathcal{F}_t^k := \sigma(\{\xi_i^j \cup \xi_t^z \mid i \in [t-1], j \in [k], z \in [k-1]\}).$$

At this point v_t^k and λ_{t+1}^k depend on the realization of ξ_t^k. The second expectation is taken with respect to the random vectors $\xi_1^1, \xi_1^2, \dots, \xi_T^K$ running through the space Ξ, denoted by $\mathbb{E}_\xi[\cdot]$. The third expectation is taken based on the above randomized process and then takes into account the random subscripts R at the output, meaning that the total expectation with respect to all random variables generated in Algorithm 1. We write the notation $\mathbb{E}[\cdot] := \mathbb{E}_{(R,\xi)}[\cdot]$ for ease.

Algorithm 1:

```
Input: Positive integers $T$ and $K$, parameters $\alpha > 1$, $\beta \in (0,1]$, $x_0^1 = \bar{x}^0 \in \mathcal{X}$, $\{\bar{v}^k\} \subseteq X$ and
    $\{\bar{\lambda}^k\} \subseteq \mathbb{R}_+^m$.
Output: $x_R := x_R^{K+1}$, where $R$ is uniformly randomly chosen from $\{0, 1, \dots, T-1\}$.
for $k = 1$ to $K$ do
    Set $v_0^k = \bar{v}^k$, $\lambda_1^k = \bar{\lambda}^k$
    Compute $x_0^{k+1} = \mathcal{M}(x_0^k, v_0^k)$
end for
for $t = 1$ to $T$ do
    Let $x_t^1 \in \mathcal{X}$
    for $k = 1$ to $K$ do
        Generate an i.i.d. sample $\xi_t^k$ from $\Xi$, and obtain $F_0(x_{t-1}^k, \xi_t^k)$, $\nu_0(x_{t-1}^k, \xi_t^k)$ and $F_i(v_{t-1}^k, \xi_t^k)$,
        $\nu_i(v_{t-1}^k, \xi_t^k)$, $i \in [m]$
        Compute $v_t^k$ through (2.3)
        Compute $x_t^{k+1}$ through (2.4)
        Compute $\lambda_{t+1}^k$ through (2.5)
    end for
end for
```

3 Auxiliary lemmas

In this section, we will present auxiliary lemmas characterizing basic properties of framework Algorithm 1 and prepare for the forthcoming approximation analyses. Proofs of these auxiliary lemmas are presented in Appendix A.

Let $\{x_t^k\}$, $\{\lambda_t^k\}$ and $\{v_t^k\}$ be generated by Algorithm 1. The following lemma explores the behavior of multipliers during updates.

Lemma **3.1.** *Suppose that the Assumption 1.1 is satisfied. Then for any $x \in X$, $t \in [T]$ and $k \in [K]$, it holds that*

$$\frac{1}{2\beta}\|\lambda_{t+1}^k\|^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2 \leq \langle -\nu_0(x_{t-1}^k, \xi_t^k), x - v_{t-1}^k\rangle + \frac{1}{\alpha}\frac{\|\nu_0(x_{t-1}^k, \xi_t^k)\|^2}{2} + \beta\frac{\|F(x, \xi_t^k)\|^2}{2} + \langle \lambda_t^k, F(x, \xi_t^k)\rangle + \frac{\alpha}{2}(\|x - v_{t-1}^k\|^2 - \|x - v_t^k\|^2). \tag{3.1}$$

Our next lemma aims to investigate the upper bound of the difference between two successive auxiliary variables v_t^k and v_{t-1}^k.

Lemma **3.2.** *Suppose that Assumptions 1.1 and 1.2 are satisfied, and $2\alpha - \beta m M_3^2 > 0$. Then for any $t \in [T]$ and $k \in [K]$, it holds that*

$$\|v_{t-1}^k - v_t^k\| \leq \frac{2}{2\alpha - \beta m M_3^2}(\|\nu_0(x_{t-1}^k, \xi_t^k)\| + \sqrt{m}M_3\|\lambda_t^k\| + \beta\sqrt{m}M_3\|[F(v_{t-1}^k, \xi_t^k)]_+\|).$$

The lemma below characterizes an upper bound on $\|\lambda_t^k\|$ in expectation.

Lemma **3.3.** *Suppose that Assumptions 1.1-1.4 are satisfied. Then for any $t \in [T]$ and $k \in [K]$, it holds that $\mathbb{E}_\xi[\|\lambda_t^k\|] \leq \theta := E_1 + \frac{1}{\alpha}E_2 + \beta E_3 + \alpha\beta E_4$, where*

$$\begin{aligned} &E_1 = \max_{k \in [K]}\|\bar{\lambda}^k\| + 2(M_1 + M_3\mathrm{d}(X))\sqrt{m} + \frac{M_2\mathrm{d}(X)}{M_4}, \ E_2 = \frac{M_2^2}{2M_4}, \\ &E_3 = \sqrt{m}(M_1 + M_3\mathrm{d}(X)) + \frac{M_1^2}{2M_4} \text{ and } E_4 = \frac{\mathrm{d}(X)^2}{2M_4}. \end{aligned} \tag{3.2}$$

4 Non-monotone case

Our aim in this section is to establish an approximation guarantee for Algorithm 1 in general case, where the objective f_0 can be non-monotone. In this case, we define $\mathcal{M}$ as

$$\mathcal{M}(x_t^k, v_t^k) := x_t^k + \left(v_t^k - x_t^k\right)\frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}} \quad \text{with} \quad a^k := \left(1 + \frac{k-1}{K}\right)^2 \tag{4.1}$$

for any $t \in [T] \cup \{0\}$, $k \in [K]$. Let $\{x_t^k\}$ be generated by Algorithm 1 applying (4.1) and $x_R := x_R^{K+1}$ with R being uniformly randomly chosen from $\{0, 1, \dots, T-1\}$.

We first provide an upper bound on $\|x_t^k\|_\infty$ and put the proof in Appendix B.1.

Lemma **4.1.** *For each $t \in [T] \cup \{0\}$ and $k \in [K+1]$, it holds that $\|x_t^k\|_\infty \leq 1 - (1 - \min_{x \in X}\|x\|_\infty)/\sqrt{a^k}$.*

To establish the approximation guarantee, we need to identify the relationships between $f_0(x_t^{k+1})$, $f_0(x_t^k)$ and $f_0(x^*)$. The detailed proof is presented in Appendix B.2.

Lemma **4.2.** *It holds that for any $t \in [T]$ and $k \in [K]$,*

$$\begin{aligned} &a^{k+1}f_0(x_{t-1}^{k+1}) - \frac{k+1}{K}\left(1 - \min_{x \in X}\|x\|_\infty\right)f_0(x^*) - \left(a^k f_0(x_{t-1}^k) - \frac{k}{K}\left(1 - \min_{x \in X}\|x\|_\infty\right)f_0(x^*)\right) \\ &\geq \left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x^*)\frac{\sqrt{a^k}}{K}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2. \end{aligned}$$

We now arrive at the main theorem characterizing the approximation guarantee of Algorithm 1 in non-monotone case.

Theorem **4.3.** *Suppose that Assumptions 1.1-1.3 are satisfied, then*

$$\begin{aligned}\frac{1}{4}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-\mathbb{E}[f_0(x_R)]&\\ \le \frac{1}{K}\frac{L_0\mathrm{d}(X)^2}{8}+\frac{1}{T\beta}\frac{\max_{k\in[K]}\|\bar{\lambda}^k\|^2}{4}&+\frac{1}{\alpha}\frac{M_2^2}{4}+\beta\frac{M_1^2}{4}+\frac{\alpha}{T}\frac{\mathrm{d}(X)^2}{4}.\end{aligned}\tag{4.2}$$

Proof. Proof Recall that for each $t\in[T], k\in[K], \mathcal{F}_t^k=\{\xi_i^j\cup\xi_t^z \mid i\in[t-1], j\in[k], z\in[k-1]\}$. Plugging $x=x^*$ into Lemma 3.1 and then taking expectation with respect to ξ_t^k yield that

$$\begin{aligned}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2]-\frac{1}{2\beta}\|\lambda_t^k\|^2&\le\langle\mathbb{E}_{\xi_t^k}[-\nu_0(x_{t-1}^k,\xi_t^k)],x^*-v_{t-1}^k\rangle+\frac{1}{\alpha}\frac{\mathbb{E}_{\xi_t^k}[\|\nu_0(x_{t-1}^k,\xi_t^k)\|^2]}{2}\\ &+\beta\frac{\mathbb{E}_{\xi_t^k}[\|F(x^*,\xi_t^k)\|^2]}{2}+\langle\lambda_t^k,\mathbb{E}_{\xi_t^k}[F(x^*,\xi_t^k)]\rangle+\frac{\alpha}{2}(\|x^*-v_{t-1}^k\|^2-\mathbb{E}_{\xi_t^k}[\|x^*-v_t^k\|^2]).\end{aligned}$$

From $\mathbb{E}_{\xi_t^k}[\nu_0(x_{t-1}^k,\xi_t^k)]=\nabla f_0(x_{t-1}^k)$ and $\mathbb{E}_{\xi_t^k}[F(x^*,\xi_t^k)]=f(x^*)$, Assumptions 1.3 and 1.2 indicate

$$\begin{aligned}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2]-\frac{1}{2\beta}\|\lambda_t^k\|^2\le&\langle-\nabla f_0(x_{t-1}^k),x^*-v_{t-1}^k\rangle+\frac{1}{\alpha}\frac{M_2^2}{2}+\beta\frac{M_1^2}{2}+\langle\lambda_t^k,f(x^*)\rangle\\ &+\frac{\alpha}{2}(\|x^*-v_{t-1}^k\|^2-\mathbb{E}_{\xi_t^k}[\|x^*-v_t^k\|^2]).\end{aligned}\tag{4.3}$$

For any $t\in[T]$, $k\in[K]$, it follows from $\lambda_t^k\ge\mathbf{0}_m$ and $f(x^*)\le\mathbf{0}_m$ that

$$\begin{aligned}\frac{\sqrt{a^k}}{K}\Big\langle\nabla f_0(x_{t-1}^k),v_{t-1}^k-x^*\Big\rangle&\ge\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2]-\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\|\lambda_t^k\|^2-\frac{\sqrt{a^k}}{K}\Big(\frac{1}{\alpha}\frac{M_2^2}{2}\\ &+\beta\frac{M_1^2}{2}+\langle\lambda_t^k,f(x^*)\rangle+\frac{\alpha}{2}(\|x^*-v_{t-1}^k\|^2-\mathbb{E}_{\xi_t^k}[\|x^*-v_t^k\|^2])\Big)\\ &\ge\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2]-\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\|\lambda_t^k\|^2-\frac{\sqrt{a^k}}{K}\Big(\frac{1}{\alpha}\frac{M_2^2}{2}+\beta\frac{M_1^2}{2}\\ &+\frac{\alpha}{2}(\|x^*-v_{t-1}^k\|^2-\mathbb{E}_{\xi_t^k}[\|x^*-v_t^k\|^2])\Big)\end{aligned}$$

which together with Lemma 4.2 yields

$$\begin{aligned}&a^{k+1}f_0(x_{t-1}^{k+1})-\frac{k+1}{K}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-\Big(a^kf_0(x_{t-1}^k)-\frac{k}{K}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)\Big)\\ &\ge-\frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2+\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2]-\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\|\lambda_t^k\|^2-\frac{\sqrt{a^k}}{K}\Big(\frac{1}{\alpha}\frac{M_2^2}{2}\\ &+\beta\frac{M_1^2}{2}+\frac{\alpha}{2}(\|x^*-v_{t-1}^k\|^2-\mathbb{E}_{\xi_t^k}[\|x^*-v_t^k\|^2])\Big).\end{aligned}$$

Taking expectation with respect to all the random vectors $\xi_1^1,\xi_1^2,\ldots,\xi_T^K$, we derive

$$\begin{aligned}&a^{k+1}\mathbb{E}_\xi[f_0(x_{t-1}^{k+1})]-\frac{k+1}{K}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-\Big(a^k\mathbb{E}_\xi[f_0(x_{t-1}^k)]\\ &-\frac{k}{K}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)\Big)\ge-\frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2+\frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_{t+1}^k\|^2-\|\lambda_t^k\|^2]\\ &-\frac{\sqrt{a^k}}{K}\frac{1}{\alpha}\frac{M_2^2}{2}-\frac{\sqrt{a^k}}{K}\beta\frac{M_1^2}{2}-\frac{\sqrt{a^k}}{K}\frac{\alpha}{2}\mathbb{E}_\xi[\|x^*-v_{t-1}^k\|^2-\|x^*-v_t^k\|^2].\end{aligned}$$

Summing this inequality over all $t \in [T]$ and using $\lambda_1^k = \bar{\lambda}^k$, $1 \leq \sqrt{a^k} \leq 2$, together with $\left\|x^* - v_0^k\right\| \leq \mathrm{d}(X)$, imply that

$$\begin{aligned}
& a^{k+1}\sum_{t=1}^{T}\mathbb{E}_\xi[f_0(x_{t-1}^{k+1})] - \frac{T(k+1)}{K}\Big(1 - \min_{x\in X}\|x\|_\infty\Big)f_0(x^*) - \Big(a^k\sum_{t=1}^{T}\mathbb{E}_\xi[f_0(x_{t-1}^{k})] \\
&\quad - \frac{Tk}{K}\Big(1 - \min_{x\in X}\|x\|_\infty\Big)f_0(x^*)\Big) \\
&\quad \geq -\frac{TL_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2 + \frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\sum_{t=1}^{T}\mathbb{E}_\xi[\|\lambda_{t+1}^k\|^2 - \|\lambda_t^k\|^2] - \frac{\sqrt{a^k}}{K}\frac{T}{\alpha}\frac{M_2^2}{2} - T\frac{\sqrt{a^k}}{K}\beta\frac{M_1^2}{2} \\
&\quad - \frac{\sqrt{a^k}}{K}\frac{\alpha}{2}\sum_{t=1}^{T}\mathbb{E}_\xi[\|x^* - v_{t-1}^k\|^2 - \|x^* - v_t^k\|^2] \\
&\quad \geq -\frac{TL_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2 + \frac{\sqrt{a^k}}{K}\frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_{T+1}^k\|^2 - \|\bar{\lambda}^k\|^2] - \frac{\sqrt{a^k}}{K}\frac{T}{\alpha}\frac{M_2^2}{2} - T\frac{\sqrt{a^k}}{K}\beta\frac{M_1^2}{2} \\
&\quad - \frac{\sqrt{a^k}}{K}\frac{\alpha}{2}\mathbb{E}_\xi[\|x^* - v_0^k\|^2] \\
&\quad \geq -\frac{T}{K^2}\frac{L_0}{2}\mathrm{d}(X)^2 - \frac{1}{K\beta}\|\bar{\lambda}^k\|^2 - \frac{T}{K\alpha}M_2^2 - \frac{T\beta}{K}M_1^2 - \frac{\alpha}{K}\mathrm{d}(X)^2.
\end{aligned}$$

We thus derive the inequality

$$\begin{aligned}
& a^{K+1}\sum_{t=1}^{T}\mathbb{E}_\xi[f_0(x_{t-1}^{K+1})] - \frac{T(K+1)}{K}\Big(1 - \min_{x\in X}\|x\|_\infty\Big)f_0(x^*) - \Big(a^1\sum_{t=1}^{T}\mathbb{E}_\xi[f_0(x_{t-1}^{1})] \\
& - \frac{T}{K}\Big(1 - \min_{x\in X}\|x\|_\infty\Big)f_0(x^*)\Big) \geq -\frac{T}{K}\frac{L_0}{2}\mathrm{d}(X)^2 - \frac{1}{\beta}\max_{k\in[K]}\|\bar{\lambda}^k\|^2 - \frac{T}{\alpha}M_2^2 - T\beta M_1^2 - \alpha\mathrm{d}(X)^2.
\end{aligned}$$

Noting $a^{K+1} = 4$, $a^1 = 1$ and $f_0(\cdot) \geq 0$ and dividing both sides with T leads to (4.2). □

Our next focus is to examine the constraint violation at the output of Algorithm 1.

Theorem **4.4.** *Suppose that Assumptions 1.1-1.5 are satisfied, and $2\alpha - \beta m M_3^2 > 0$, then*

$$\begin{aligned}
\mathbb{E}[\|[f(x_R)]_+\|] \leq\ & \frac{1}{T\beta}\sqrt{m}E_1 + \frac{1}{T\beta\alpha}\sqrt{m}E_2 + \frac{1}{T}\sqrt{m}E_3 + \frac{\alpha}{T}\sqrt{m}E_4 \\
& + \frac{1}{2\alpha - \beta m M_3^2}(2M_3M_2\sqrt{m} + 2M_3^2 m E_1) + \frac{1}{2\alpha^2 - \beta\alpha m M_3^2}2M_3^2 m E_2 \qquad (4.4)\\
& + \frac{\alpha\beta}{2\alpha - \beta m M_3^2}2M_3^2 m E_4 + \frac{\beta}{2\alpha - \beta m M_3^2}(2M_3^2 m E_3 + 2M_3^2 m M_1),
\end{aligned}$$

where E_1, E_2, E_3 and E_4 are introduced in (3.2).

Proof. Proof It is worthy to note that for any $i \in [m]$, $t \in [T]$ and $k \in [K]$,

$$\begin{aligned}
\lambda_{t+1}^k(i) &\geq \lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) + \beta\langle \nu_i(v_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle \\
&\geq \lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) - \beta M_3\|v_t^k - v_{t-1}^k\| \\
&\geq \lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) - \beta M_3\frac{2}{2\alpha - \beta m M_3^2}(\|\nu_0(x_{t-1}^k, \xi_t^k)\| + \|\lambda_t^k\|\sqrt{m}M_3 \\
&\quad + \beta\sqrt{m}\|[F(v_{t-1}^k, \xi_t^k)]_+\|M_3),
\end{aligned}$$

where the first inequality follows from (2.5), the second inequality follows from Cauchy-Schwarz Inequality and Assumption 1.2, and the third inequality follows from Lemma 3.2. Then it derives

$$\begin{aligned}F_i(v_{t-1}^k,\xi_t^k) \le \frac{1}{\beta}(\lambda_{t+1}^k(i)-\lambda_t^k(i)) + \frac{2M_3}{2\alpha-\beta m M_3^2}(\|\nu_0(x_{t-1}^k,\xi_t^k)\| + \|\lambda_t^k\|\sqrt{m}M_3 \\ + \beta\sqrt{m}\|[F(v_{t-1}^k,\xi_t^k)]_+\|M_3).\end{aligned}$$

Note that it follows from Assumptions 1.3 and 1.2 that

$$\mathbb{E}_{\xi_t^k}[F_i(v_{t-1}^k,\xi_t^k)] = f_i(v_{t-1}^k),\ \mathbb{E}_{\xi_t^k}[\|F(v_{t-1}^k,\xi_t^k)\|] \le M_1 \ \text{ and } \ \mathbb{E}_{\xi_t^k}[\|\nu_0(x_{t-1}^k,\xi_t^k)\|] \le M_2.$$

Then we further take expectation with respect to random vectors ξ_1^1, ξ_1^2, …, ξ_T^K, attaining

$$\mathbb{E}_\xi[f_i(v_{t-1}^k)] \le \frac{1}{\beta}\mathbb{E}_\xi[\lambda_{t+1}^k(i)-\lambda_t^k(i)] + \frac{2M_3}{2\alpha-\beta m M_3^2}(M_2 + \mathbb{E}_\xi[\|\lambda_t^k\|]\sqrt{m}M_3 + \beta\sqrt{m}M_1M_3). \tag{4.5}$$

Let us turn to the iterate update $x_t^{k+1} = x_t^k + (v_t^k - x_t^k)\frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}}$, for any $t \in [T]\cup\{0\}, k \in [K]$. For simplicity, we denote $\eta^k = \frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}}$. We recursively apply the series and subsequently observe that

$$x_t^{k+1} = \prod_{r=1}^{k}(1-\eta^r)\,x_t^1 + \sum_{s=1}^{k-1}\prod_{r=s+1}^{k}(1-\eta^r)\,\eta^s v_t^s + \eta^k v_t^k,$$

and

$$\prod_{r=1}^{k}(1-\eta^r) + \sum_{s=1}^{k-1}\prod_{r=s+1}^{k}(1-\eta^r)\,\eta^s + \eta^k = 1.$$

Recall that f_i is convex, then for any $t \in [T]\cup\{0\}, i \in [m]$, when $k = K$ it implies

$$f_i(x_t^{K+1}) \le \prod_{r=1}^{K}(1-\eta^r)\,f_i(x_t^1) + \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s f_i(v_t^s) + \eta^K f_i(v_t^K).$$

Summing $t \in [T-1]\cup\{0\}$ and applying Assumption 1.5, we obtain

$$\begin{aligned}\sum_{t=0}^{T-1} f_i(x_t^{K+1}) &\le \sum_{t=0}^{T-1}\Big(\prod_{r=1}^{K}(1-\eta^r)\,f_i(x_t^1) + \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s f_i(v_t^s) + \eta^K f_i(v_t^K)\Big) \\ &= \prod_{r=1}^{K}(1-\eta^r)\sum_{t=0}^{T-1} f_i(x_t^1) + \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s\sum_{t=0}^{T-1} f_i(v_t^s) + \eta^K\sum_{t=0}^{T-1} f_i(v_t^K) \\ &\le \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s\sum_{t=0}^{T-1} f_i(v_t^s) + \eta^K\sum_{t=0}^{T-1} f_i(v_t^K).\end{aligned}$$

Taking expectation with the random vector ξ and applying with (4.5), Lemma 3.3 and

$$\sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s + \eta^K = 1 - \prod_{r=1}^{K}(1-\eta^r) \le 1,$$

lead to

$$
\begin{aligned}
&\mathbb{E}_\xi\Big[\sum_{t=0}^{T-1} f_i(x_t^{K+1})\Big] \\
&\quad\leq \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s\bigg(\frac{1}{\beta}\mathbb{E}_\xi[\lambda_{T+1}^s(i)-\lambda_1^s(i)]+\frac{2TM_3M_2}{2\alpha-\beta mM_3^2}+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2} \\
&\quad\quad\cdot\sum_{t=1}^{T}\mathbb{E}_\xi[\|\lambda_t^s\|]+\frac{2T\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2}\bigg)+\eta^K\bigg(\frac{1}{\beta}\mathbb{E}_\xi[\lambda_{T+1}^K(i)-\lambda_1^K(i)] \\
&\quad\quad+\frac{2TM_3M_2}{2\alpha-\beta mM_3^2}+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2}\sum_{t=1}^{T}\mathbb{E}_\xi[\|\lambda_t^K\|]+\frac{2T\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2}\bigg) \\
&\quad\leq \sum_{s=1}^{K-1}\prod_{r=s+1}^{K}(1-\eta^r)\,\eta^s\bigg(\frac{\theta}{\beta}+\frac{2TM_3M_2}{2\alpha-\beta mM_3^2}+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2}T\theta+\frac{2T\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2}\bigg) \\
&\quad\quad+\eta^K\bigg(\frac{\theta}{\beta}+\frac{2TM_3M_2}{2\alpha-\beta mM_3^2}+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2}T\theta+\frac{2T\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2}\bigg) \\
&\quad\leq \frac{\theta}{\beta}+\frac{2TM_3M_2}{2\alpha-\beta mM_3^2}+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2}T\theta+\frac{2T\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2},
\end{aligned}
$$

where θ is defined in Lemma 3.3. We further divide both sides of above inequality with T, attaining

$$
\begin{aligned}
&\mathbb{E}[f_i(x_R)]=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_\xi[f_i(x_t^{K+1})] \\
&\leq \frac{1}{T\beta}(E_1+\frac{1}{\alpha}E_2+\beta E_3+\alpha\beta E_4)+\frac{2M_3M_2}{2\alpha-\beta mM_3^2} \\
&\quad+\frac{2M_3^2\sqrt{m}}{2\alpha-\beta mM_3^2}(E_1+\frac{1}{\alpha}E_2+\beta E_3+\alpha\beta E_4)+\frac{2\beta M_3^2\sqrt{m}M_1}{2\alpha-\beta mM_3^2} \\
&=\frac{1}{T\beta}E_1+\frac{1}{T\beta\alpha}E_2+\frac{1}{T}E_3+\frac{\alpha}{T}E_4+\frac{1}{2\alpha-\beta mM_3^2}(2M_3M_2+2M_3^2\sqrt{m}E_1)+\frac{1}{2\alpha^2-\beta\alpha mM_3^2} \\
&\quad\cdot 2M_3^2\sqrt{m}E_2+\frac{\alpha\beta}{2\alpha-\beta mM_3^2}2M_3^2\sqrt{m}E_4+\frac{\beta}{2\alpha-\beta mM_3^2}(2M_3^2\sqrt{m}E_3+2M_3^2\sqrt{m}M_1).
\end{aligned}
$$

By using $x_R=x_R^{K+1}$ and $\mathbb{E}[\|[f(x_R)]_+\|]\leq\sqrt{m}\mathbb{E}[\|[f(x_R)]_+\|_\infty]$ we obtain (4.4). □

Theorems 4.3 and 4.4 provide estimates on theoretical performances of Algorithm 1. By specifying parameter settings we summarize the approximation ratio and error bounds in the corollary below.

Corollary 4.5. *Suppose that Assumptions 1.1-1.5 are satisfied. Let $T=O(I^{2/3})$, $K=O(I^{1/3})$, $\alpha=I^{1/3}$, and $\beta=I^{-1/3}$. Then it holds that*

$$\frac{1}{4}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-\mathbb{E}[f_0(x_R)]=O(I^{-1/3}),\ \mathbb{E}[\|[f(x_R)]_+\|]=O(I^{-1/3}).$$

Proof. Proof Under the parameter settings of this corollary, it is easy to obtain from Theorem 4.3 and

Theorem 4.4 that

$$\frac{1}{4}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-\mathbb{E}[f_0(x_R)]\le\frac{1}{I^{1/3}}\Big(\frac{L_0\mathrm{d}(X)^2}{8}+\frac{\max\limits_{k\in[K]}\|\bar{\lambda}^k\|^2}{4}+\frac{M_2^2}{4}+\frac{M_1^2}{4}+\frac{\mathrm{d}(X)^2}{4}\Big),$$

$$\mathbb{E}[\|[f(x_R)]_+\|]\le\frac{1}{I^{1/3}}\sqrt{m}(E_1+E_4)+\frac{1}{I^{2/3}}\sqrt{m}(E_2+E_3)+\frac{1}{2I^{1/3}-I^{-1/3}mM_3^2}(2M_3M_2$$
$$\cdot\sqrt{m}+2M_3^2m(E_1+E_4))+\frac{1}{2I^{2/3}-mM_3^2}(2M_3^2m(E_2+E_3+M_1)),$$

which are both in order $\mathcal{O}(I^{-1/3})$. □

Under the same conditions as Corollary 4.5, by applying Markov's inequality we obtain the following high probability bounds at x_R. More specifically, for any $\rho_o,\rho_c\in(0,1)$, it holds that with probability at least $(1-\rho_o)$,

$$\frac{1}{4}\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)-f_0(x_R)$$
$$\le\frac{1}{\rho_o}\Big(\frac{1}{I^{1/3}}\Big(\frac{L_0\mathrm{d}(X)^2}{8}+\frac{\max_{k\in[K]}\|\bar{\lambda}^k\|^2}{4}+\frac{M_2^2}{4}+\frac{M_1^2}{4}+\frac{\mathrm{d}(X)^2}{4}\Big)\Big)$$

and with probability at least $(1-\rho_c)$,

$$\|[f(x_R)]_+\|\le\frac{1}{\rho_c}\Big(\frac{1}{I^{1/3}}\sqrt{m}(E_1+E_4)+\frac{1}{I^{2/3}}\sqrt{m}(E_2+E_3)+\frac{1}{2I^{1/3}-I^{-1/3}mM_3^2}(2M_3M_2$$
$$\cdot\sqrt{m}+2M_3^2m(E_1+E_4))+\frac{1}{2I^{2/3}-mM_3^2}(2M_3^2m(E_2+E_3+M_1))\Big).$$

5 Monotone case

In this section, we study the case when f_0 is monotonously increasing. To proceed, we lay out another assumption that is common for monotone DR-submodular maximization.

Assumption 5.1. *The set $X\subseteq[0,1]^n$ contains the zero vector.*

Under Assumption 5.1, the initial point x_t^1 is $\mathbf{0}_n$. And to accommodate the monotone case, we define the mapping $\mathcal{M}$ as

$$\mathcal{M}(x_t^k,v_t^k)=x_t^k+\frac{1}{K}v_t^k,\quad\forall t\in[T]\cup\{0\},k\in[K]. \tag{5.1}$$

In the following, let $\{x_t^k\}$ be generated by Algorithm 1 with operator $\mathcal{M}$ as defined in (5.1) and $x_R:=x_R^{K+1}$ with R being uniformly randomly chosen from $\{0,1,\ldots,T-1\}$.

The following two theorems demonstrate the expected approximation bound and constraint violation at the output x_R. Detailed proofs are presented in Appendix C.

Theorem **5.1.** *Suppose that Assumptions 1.1-1.3 and Assumption 5.1 are satisfied. Then we have*

$$\Big(1-\frac{1}{e}\Big)f_0(x^*)-\mathbb{E}[f_0(x_R)]$$
$$\le\frac{1}{K}\frac{L_0\mathrm{d}(X)^2}{2}+\frac{1}{\alpha}\frac{M_2^2}{2}+\beta\frac{M_1^2}{2}+\frac{\alpha}{T}\frac{\mathrm{d}(X)^2}{2}+\frac{1}{\beta T}\frac{\max_{k\in[K]}\|\bar{\lambda}^k\|^2}{2}. \tag{5.2}$$

THEOREM **5.2.** *Suppose that Assumptions 1.1-1.4 and Assumption 5.1 are satisfied, and* $2\alpha-\beta m M_3^2 > 0$*. Then it holds that*

$$\begin{aligned}\mathbb{E}[\|[f(x_R)]_+\|] \leq\ & \frac{\sqrt{m}}{T\beta}(E_1+\frac{1}{\alpha}E_2+\beta E_3+\alpha\beta E_4)+\frac{2\sqrt{m}M_3M_2}{2\alpha-\beta m M_3^2}\\ &+\frac{2mM_3^2}{2\alpha-\beta m M_3^2}(E_1+\frac{1}{\alpha}E_2+\beta E_3+\alpha\beta E_4)+\frac{2\beta m M_1 M_3^2}{2\alpha-\beta m M_3^2},\end{aligned}$$

where $E_1,\ E_2,\ E_3,$ *and* E_4 *are introduced in* (3.2).

By specifying the parameters in Theorems 5.1 and 5.2 as

$$T=O(I^{2/3}),\ K=O(I^{1/3}),\ \alpha=I^{1/3},\ \beta=I^{-1/3}, \tag{5.3}$$

we obtain the following relations:

$$\begin{aligned}\Big(1-\frac{1}{e}\Big)f_0(x^*)-\mathbb{E}[f_0(x_R)] &\leq \frac{1}{I^{1/3}}\Big(\frac{L_0\mathrm{d}(X)^2}{2}+\frac{M_2^2}{2}+\frac{M_1^2}{2}+\frac{\mathrm{d}(X)^2}{2}+\frac{\max_{k\in[K]}\|\bar{\lambda}^k\|^2}{2}\Big),\\ \mathbb{E}[\|[f(x_R)]_+\|] &\leq \frac{\sqrt{m}}{I^{1/3}}(E_1+E_4)+\frac{\sqrt{m}}{I^{2/3}}(E_2+E_3)+\frac{2\sqrt{m}M_3M_2}{2I^{1/3}-I^{-1/3}mM_3^2}\\ &\quad+\frac{2mM_3^2}{2I^{1/3}-I^{-1/3}mM_3^2}(E_1+E_4)+\frac{2mM_3^2}{2I^{2/3}-mM_3^2}(E_2+E_3+M_1),\end{aligned}$$

which are both in order $\mathcal{O}(I^{-1/3})$. For completeness, we summarize these results as follows.

Corollary 5.3. *Suppose that Assumptions 1.1-1.4 and Assumption 5.1 hold. Under parameter setting* (5.3) *we have*

$$(1-1/e)f_0(x^*)-\mathbb{E}[f_0(x_R)]=O(I^{-1/3})\ \textit{and}\ \mathbb{E}[\|[f(x_R)]_+\|]=O(I^{-1/3}).$$

By applying Markov's inequality we directly establish high-probability bounds at x_R. That is, for any given $\rho_o,\rho_c\in(0,1)$, with probability $(1-\rho_o)$ such that

$$\Big(1-\frac{1}{e}\Big)f_0(x^*)-f_0(x_R)\leq\frac{1}{\rho_o}\left(\frac{1}{I^{1/3}}\Big(\frac{L_0\mathrm{d}(X)^2}{2}+\frac{M_2^2}{2}+\frac{M_1^2}{2}+\frac{\mathrm{d}(X)^2}{2}+\frac{\max_{k\in[K]}\|\bar{\lambda}^k\|^2}{2}\Big)\right)$$

and with probability $(1-\rho_c)$ such that

$$\begin{aligned}\|[f(x_R)]_+\|\leq &\frac{1}{\rho_c}\left(\frac{\sqrt{m}}{I^{1/3}}(E_1+E_4)+\frac{\sqrt{m}}{I^{2/3}}(E_2+E_3)+\frac{2\sqrt{m}M_3M_2}{2I^{1/3}-I^{-1/3}mM_3^2}\right.\\ &\left.+\frac{2mM_3^2}{2I^{1/3}-I^{-1/3}mM_3^2}(E_1+E_4)+\frac{2mM_3^2}{2I^{2/3}-mM_3^2}(E_2+E_3+M_1)\right).\end{aligned}$$

Remark 5.1. *In comparison to our work, [33] focus on problems in online setting with stochastic constraints and deterministic objective function. Here, the stochasticity of constraints refers to that the constraint function is randomly chosen from a class of functions following an underlining distribution. At each moment, exact gradients of the objective function and constraint at multiple points need to be calculated. However, our work differs in that we consider more general stochastic settings. More specifically, we allow stochastic gradients of the objective function and stochastic subgradients of the constraint can be accessed. Moreover, in our algorithm only one random sample is called at each iteration. In addition, [33] require the assumption of smoothness on the constraint functions. We instead assume the availability of approximate subgradients of constraint functions at a given point. Besides, [33] only study the case with monotone objective, while we consider both non-monotone and monotone cases.*

6 Illustrative examples

To validate our theoretical analysis, we present three illustrative examples and report associated test results. These examples were conducted in Python 3.8 on a server with an Intel® Xeon® Gold 6230 CPU. We apply the V-FISTA variant algorithm ([3]) to solve (2.3).

6.1 Welfare maximization with production cost

In first example, we consider the welfare maximization with production cost [33], formulated as

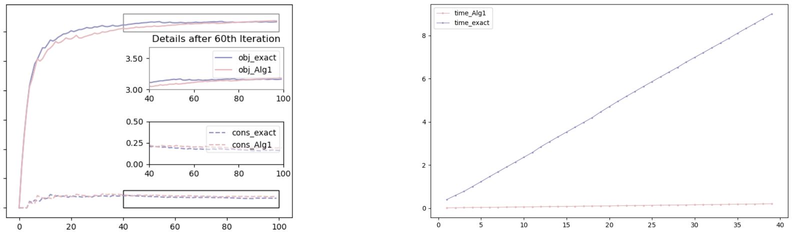
$$\begin{aligned} \max_{\mathbf{0}_n \le x \le \mathbf{1}_n} \quad & f_0(x) := \frac{1}{N}\sum_{i=1}^{N} \log\det\left(\operatorname{diag}(x)\left(L_i - \mathrm{I}\right) + \mathrm{I}\right) \\ \text{s. t.} \quad & f_1(x) := \frac{1}{N}\sum_{i=1}^{N} (x^T P_i x - b) \le 0. \end{aligned} \tag{6.1}$$

We generate $L_i, P_i \in \mathbb{R}^{n\times n}$, $i \in [N]$ as random positive definite matrices whose eigenvalues are uniformly chosen from $[10^{-16}, 3]$ and $[0.3, 6]$, respectively, where $n = 50$, $b = 4$ and I is the identity matrix. In this setting, we can easily see that $f_1(\mathbf{0}_n) < 0$, thus the Slater's condition is satisfied.

For Algorithm 1, we set the maximum number of iterations in the outer loop as $T = 100$, and the number of iterations in the inner loop as $K = 10$. The initial points are $\bar{v}^k = \mathbf{0}_n$ and $\bar{\lambda}^k = 0$, where $k \in [K]$. For any $t \in [T]$ and $k \in [K]$, we randomly and uniformly generate a batch of indices from $[N]$ with size as 10. This is used to compute the mini-batch stochastic function values and stochastic gradients of f_0 and f_1, respectively, aiming for constructing the model function (2.2). Regarding the trend of objective function values along with the outer iteration index t, we compute the averaged values at past iterates to show more stable performance. For instance, we compute $\frac{1}{t}\sum_{s=1}^{t} f_0(x_s)$, where $x_s := x_s^{K+1}$ and $s \in [t]$, and record its trend as t increases to T. Additionally, we compute the averaged constraint violation $\frac{1}{t}\sum_{s=1}^{t}[f_1(x_s)]_+$.

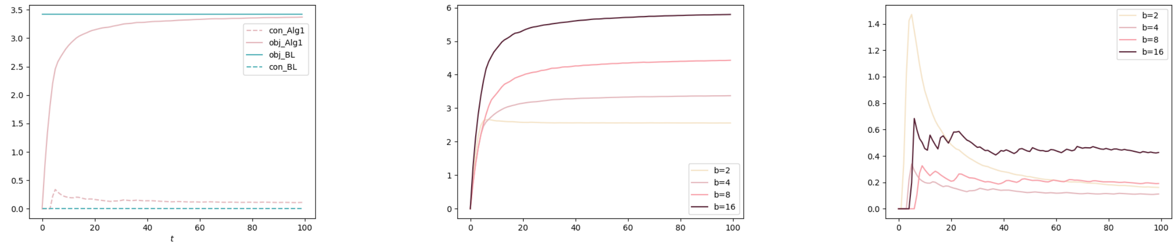
In numerical tests, we first examine the exact version of our algorithm and compare the results with Algorithm 1. The exact algorithm refers to the version that in step 8 of Algorithm 1, we compute exact information of the objective and constraint functions instead of their stochastic approximations. Considering the increased computational time due to the exact information computations, we slightly reduce the scale of N to 500. Figure 1 shows the comparison results of the two algorithms, noting that the accuracy of the solutions from Algorithm 1 is almost indistinguishable from those of exact algorithm. This indicates that our algorithm can achieve similar accuracy to the exact algorithm within much less CPU time, highlighting the efficiency advantage of our algorithm.

To evaluate the performance of Algorithm 1 in solving (6.1), we also compare with the benchmark algorithm, Python's built-in function *minimize*, in Figure 2(a). In this setting, we set $N = 6000$. The pink line represents the results obtained by Algorithm 1, with values reaching as high as 3.4159, while the cyan line corresponds to the results obtained by *minimize*, culminating in a measure of 3.4184 with an slightly constraint violation 10^{-8}. As can be observed from Figure 2(a), the objective function value and constraint violation by both algorithms converge to a similar level. This indicates that both Algorithm 1 and the benchmark algorithm are able to effectively solve the problem. Figures 2(b) and 2(c) show the function value and constraint violation v.s. the outer iteration number for 4 function instances with different values of b. It is evident that despite variations in the parameter b, our proposed algorithm manages to sustain a stable output.



(a) (b)

Figure 1: (a): function value and constraint violation v.s. outer iteration number for Algorithm 1 and exact algorithm. (b): CPU time (minutes) v.s. outer iteration number for Algorithm 1 and exact algorithm at the first 40 iterations out of 100 iterations.



(a) (b) (c)

Figure 2: function value and constraint violation v.s. outer iteration number for different settings of b. (a): $b = 4$. (b) and (c): $b = 2, 4, 8, 16$.

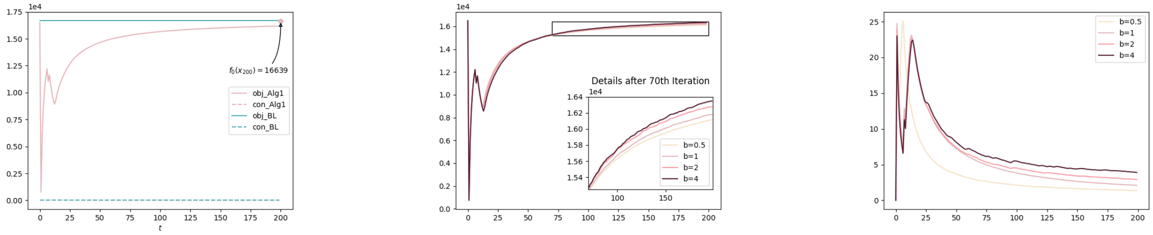
6.2 Finite-sum quadratic programming

In second example, we consider the following quadratic programming problem

$$\begin{aligned} \max_{\mathbf{0}_n \le x \le \mathbf{1}_n} \quad & f_0(x) = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{2}x^T H_i x + h_i^T x + c\Big) \\ \text{s. t.} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N}(A_i x - b_i) \le \mathbf{0}_m. \end{aligned} \tag{6.2}$$

In this example, we define the parameters as follows: $N = 2000$, $n = 100$, and $m = 5$. For each $i \in [N]$, we generate $H_i \in \mathbb{R}^{n\times n}$, a random symmetric matrix with uniformly distributed non-positive entries spanning $[-10, 0]$. Additionally, $A_i \subseteq \mathbb{R}^{m\times n}$ denotes a random matrix with entries uniformly distributed over $[0, 1]$, and $b_i = \mathbf{1}_m$. To preserve the non-monotonic behavior of f_0, we set $h_i = -0.2H_i^T\mathbf{1}_n$. To guarantee that f_0 remains nonnegative, we initially address the optimization problem (6.2) with $c = 0$ using Python's built-in function *minimize*. Denoting the obtained solution by $\tilde{x}$, we subsequently adjust c to be $f(\tilde{x}) + |0.1f(\tilde{x})|$ in (6.2) to maintain non-negativity. It is apparent that the Slater's condition is fulfilled since $f(\mathbf{0}_n) < \mathbf{0}_m$.

For Algorithm 1, the sampling process to compute approximate function information is the same as described in Subsection 6.1, except that the batch size is chosen as 5. We choose $T = 200$ and $K = 44$, respectively. The initial points are set as $\bar{v}^k = \mathbf{0}_n$ and $\bar{\lambda}^k = \mathbf{0}_m$, where $k \in [K]$. To evaluate the performance of Algorithm 1 on problem (6.2), we compare it with the algorithm *minimize*. In Figure 3(a), we plot the objective function values and constraint violations averaged in the way same as first example. The objective function value obtained by *minimize* is 16673. And by the 200th iteration of Algorithm 1, the averaged objective function value $f_0(x_{200})$ reaches 16632. Concerning the violation of constraints, Algorithm 1 reaches a feasible solution at the 200th iteration; however, the value produced by *minimize* slightly exceeds the limits by 0.05. Overall, the results by two algorithms are in the same level, demonstrating the comparable performances of Algorithm 1. In Figure 3(b), we depict the curve of the objective function value as it increases with the number of outer iterations when the parameter b changes. The main figure details the changes in the objective function throughout the entire iteration process, where the value increases from 0 to 16000. We show an enlarged detail of the objective function from the 70th iteration onwards on the secondary axis. Figure 3(c) offers a depiction of the alterations in the constraint violation with the increase in the number of outer iterations. By adjusting the parameter b, it provides additional insights into the stability of our algorithm.



(a)

(b)

(c)

Figure 3: function value and constraint violation v.s. outer iteration number for different settings of b. (a): $b = 1$. (b) and (c): $b = 0.5, 1, 2, 4$.

6.3 Influence maximization

In the influence maximization model, we focus on activating certain nodes in a given social network with the aim of enabling these nodes to influence as many other nodes as possible. The influence maximization model with budget allocation, also known as the source-node bipartite influence maximization model [34], is taken into consideration in the following. Consider a weighted bipartite graph $G(\mathcal{S}, \mathcal{T}, \mathcal{E}, p),\ p : \mathcal{E} \to [0, 1]$ on media channels nodes $\mathcal{S}$ and clients nodes $\mathcal{T}$ with edges $(s, t) \in \mathcal{E}$ implying that media channel $s \in \mathcal{S}$ has the probability $p_{s,t}$ to activate client $t \in \mathcal{T}$. Besides, every media channel $s \in \mathcal{S}$ is limited by a given budget $u(s)$ and has a weight $c(s)$, and we define its neighboring media channel set as $\Gamma(s)$. Allocating budget $x(s)$ to the media channel s allows us to model this problem in the following form:

$$\begin{aligned} \max_{\mathbf{0}_n \le x \le u} \quad & f_0(x) = \frac{1}{N} \sum_{i=1}^{N} \Big(\sum_{t \in \mathcal{T}} \Big(1 - \prod_{s \in \Gamma(t)} \Big(1 - p_{s,t}^{(i)} \Big)^{x(s)} \Big) \Big) \\ \text{s. t.} \quad & f_1(x) = \frac{1}{N} \sum_{i=1}^{N} \Big(\sum_{s \in \mathcal{S}} c(s)^{(i)} x(s) \Big) - b \le 0. \end{aligned}$$

Note that the objective function is monotone. In this model, we employed the MovieLens dataset, with the userId being $\mathcal{S}$ and the movieId being $\mathcal{T}$. We pick a segment of the data where the userId in set $\mathcal{S}$ are below 10 and the movieId in set $\mathcal{T}$ are below 100. Besides, we set $N = 1000, u = \mathbf{1}_n, T = 50, K =$

$20, \bar{v}^k = \mathbf{0}_n, \bar{\lambda}^k = 0, k \in [K]$. And the batch size is 10. It is worth mentioning that the output of our algorithm is always feasible, so we only provide illustrations of changes of objective function values v.s. the maximum outer iteration number in Figure 4, with varying budget b. As can be observed, although b varies, our algorithm performs stable.

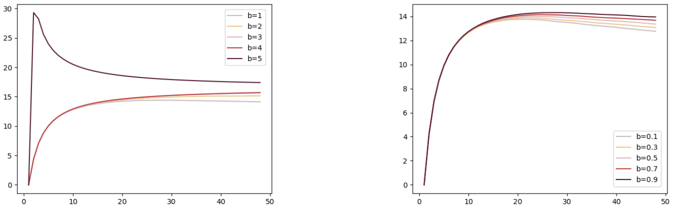


Figure 4: function value v.s. outer iteration number for varying budget b

7 Conclusion

This paper presents an algorithmic framework for stochastic approximation methods to solve DR-submodular maximization with convex functional constraints. Each subproblem is constructed based on zeroth-order and first-order stochastic approximations to the objective function and constraint functions. Under certain update scheme for problems with non-monotone objective and with monotone objective respectively, we present approximation analyses for both cases. For the former case the proposed algorithm achieves $(1 - \min_{x \in X} \|x\|_\infty)/4$-approximation, while for the latter case the approximation ratio is $(1-1/e)$. And the approximation errors and constraint violations for both cases are in order $O(I^{-1/3})$, where I denotes the total number of samples. At last, we provide experimental results on three illustrative examples to showcase the effectiveness of our algorithm.

References

[1] A. Alacaoglu and S. J. Wright. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. *arXiv preprint arXiv:2311.00678*, 2023.

[2] R. Aldrighetti, D. Battini, D. Ivanov, and I. Zennaro. Costs of resilience and disruptions in supply chain network design models: A review and future research directions. *International Journal of Production Economics*, 235:108103, 2021.

[3] A. Beck. *First-order methods in optimization.* SIAM, 2017.

[4] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

[5] A. A. Bian, K. Y. Levy, A. Krause, and J. M. Buhmann. Continuous DR-submodular maximization: Structure and algorithms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 486–496, 2017.

[6] A. A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *Proceedings of the 20th Artificial Intelligence and Statistics (AISTATS)*, pages 111–120, 2017.

[7] S. Bian, Q. Guo, S. Wang, and J. X. Yu. Efficient algorithms for budgeted influence maximization on massive social networks. In *Proceedings of the 46th International Conference on Very Large Data Bases (VLDB)*, pages 1498–1510, 2020.

[8] E. G. Birgin and J. M. Martínez. *Practical augmented Lagrangian methods for constrained optimization*. Society for Industrial and Applied Mathematics, 2014.

[9] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.

[10] C. Chekuri, T. Jayram, and J. Vondrák. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS)*, page 201–210, 2015.

[11] L. Chen, H. Hassani, and A. Karbasi. Online continuous submodular maximization. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTS)*, pages 1896–1905, 2018.

[12] Y. Cui, X. Wang, and X. Xiao. A two-phase stochastic momentum-based algorithm for nonconvex expectation-constrained optimization. *Optimization Online*, 2024.

[13] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, 205(1):431–483, 2024.

[14] D. Du, Z. Liu, C. Wu, D. Xu, and Y. Zhou. An improved approximation algorithm for maximizing a DR-submodular function over a convex set. *arXiv preprint arXiv:2203.14740*, 2022.

[15] J. Edmonds. *Submodular functions, matroids, and certain polyhedra*, pages 69–87. Gordon and Breach, 1970.

[16] H. Hassani, A. Karbasi, A. Mokhtari, and Z. Shen. Stochastic continuous greedy++: when upper and lower bounds match. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS)*, pages 13087–13097, 2019.

[17] H. Hassani, M. Soltanolkotabi, and A. Karbasi. Gradient methods for submodular maximization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 5843–5853, 2017.

[18] R. K. Iyer and J. A. Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pages 2436–2444, 2013.

[19] L. Jin and X. Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Computational Optimization and Applications*, 83(1):143–180, 2022.

[20] S. Kothawade, J. Girdhar, C. Lavania, and R. Iyer. Deep submodular networks for extractive data summarization. *arXiv preprint arXiv:2010.08593*, 2020.

[21] A. Kulesza and B. TaskarLan. Determinantal point processes for machine learning. *Computational Optimization and Applications*, 5(2-3):123–286, 2012.

[22] A. Kulik, H. Shachnai, and T. Tamir. Maximizing submodular set functions subject to multiple linear constraints. In *Proceedings of the 20th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 545–554, 2009.

[23] G. Lan and Z. Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76(2):461–498, 2020.

[24] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Stochastic inexact augmented Lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, 87(1):117–147, 2024.

[25] C. Malings and M. Pozzi. Submodularity issues in value-of-information-based sensor placement. *Reliability Engineering & System Safety*, 183(1):93–103, 2019.

[26] A. Mokhtari, H. Hassani, and A. Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *The Journal of Machine Learning Research*, 21(1):4232–4280, 2020.

[27] M. J. Neely and H. Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:2311.00678*, 2017.

[28] Y. Nesterov. *Introductory lectures on convex optimization: A basic course.* Springer Science & Business Media, 2003.

[29] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, 1999.

[30] M. Pedramfar, C. Quinn, and V. Aggarwal. A unified approach for maximizing continuous DR-submodular functions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS)*, pages 61103–61114, 2023.

[31] B. Polyak. A general method for solving extremum problems. *Soviet Mathematics. Doklady*, 174(1):593–597, 1967.

[32] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2:21–42, 2000.

[33] O. Sadeghi, P. Raut, and M. Fazel. A single recipe for online submodular maximization with adversarial or stochastic constraints. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)*, pages 14712–14723, 2020.

[34] T. Soma, N. Kakimura, K. Inaba, and K.-i. Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 351–359, 2014.

[35] T. Soma and Y. Yoshida. Maximizing monotone submodular functions over the integer lattice. *Mathematical Programming*, 172:539–563, 2018.

[36] X. Wang, S. Ma, and Y. Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of Computation*, 86:1793–1820, 2017.

[37] Y. Yan and Y. Xu. Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation*, 14(2):319–363, 2022.

[38] L. Ye, Z.-W. Liu, M. Chi, and V. Gupta. Maximization of nonsubmodular functions under multiple constraints with applications. *Automatica*, 155:111126, 2023.

[39] H. Yu, M. J. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, page 1427–1437, 2017.

[40] L. Zhang, Y. Zhang, and J. Wu. Stochastic approximation proximal method of multipliers for convex stochastic programming. *arXiv preprint arXiv:1907.12226*, 2019.

[41] L. Zhang, Y. Zhang, J. Wu, and X. Xiao. Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented Lagrangian-type algorithm. *INFORMS Journal on Computing*, 34(6):2989–3006, 2022.

[42] M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi. One sample stochastic Frank-Wolfe. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTS)*, pages 4012–4023, 2020.

[43] Q. Zhang, Z. Wan, Z. Deng, Z. Chen, X. Sun, J. Zhang, and Y. Yang. Boosting gradient ascent for continuous DR-submodular maximization. *arXiv preprint arXiv:2401.08330*, 2024.

A Proofs of auxiliary lemmas in Section 3

A.1 Proof of Lemma 3.1

Proof. Proof Note that Q_t^k is α-strongly convex. Then from the optimality of v_t^k, we have

$$Q_t^k(x) - \frac{\alpha}{2}\|x - v_t^k\|^2 \geq Q_t^k(v_t^k),\ \forall x \in X,\ t \in [T],\ k \in [K],$$

that is,

$$\begin{aligned}
&\langle -\nu_0(x_{t-1}^k, \xi_t^k), x - v_{t-1}^k\rangle + \underbrace{\frac{1}{2\beta}\sum_{i=1}^m [\lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) + \beta\langle \nu_i(v_{t-1}^k, \xi_t^k), x - v_{t-1}^k\rangle]_+^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2}_{A} \\
&+\frac{\alpha}{2}(\|x - v_{t-1}^k\|^2 - \|x - v_t^k\|^2) \geq \underbrace{\frac{1}{2\beta}\sum_{i=1}^m [\lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) + \beta\langle \nu_i(v_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle]_+^2}_{B} \\
&\qquad -\frac{1}{2\beta}\|\lambda_t^k\|^2 + \underbrace{\langle -\nu_0(x_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle + \frac{\alpha}{2}\|v_t^k - v_{t-1}^k\|^2}_{C}. \qquad \text{(A.1)}
\end{aligned}$$

Let us first look at the term "B". Following the computation of λ_{t+1}^k we obtain

$$B = \frac{1}{2\beta}\|\lambda_{t+1}^k\|^2.$$

Regarding the term "A", it follows that

$$\begin{aligned}
A &\leq \frac{1}{2\beta}\sum_{i=1}^m [\lambda_t^k(i) + \beta F_i(x, \xi_t^k)]_+^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2 \qquad \text{(A.2)}\\
&\leq \frac{1}{2\beta}\sum_{i=1}^m (\lambda_t^k(i) + \beta F_i(x, \xi_t^k))^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2 \\
&= \frac{\beta}{2}\|F(x, \xi_t^k)\|^2 + \langle \lambda_t^k, F(x, \xi_t^k)\rangle,
\end{aligned}$$

where the first inequality is due to that $\nu_i(v_{t-1}^k, \xi_t^k)$ is a subgradient of $F_i(x, \xi_t^k)$ at v_{t-1}^k, $\lambda_t^k(i) \geq 0$ and $F_i, i \in [m]$ are convex with respect to x. It also uses the fact that $[a]_+ \leq [b]_+$ for $a \leq b$. To estimate the bound of "C", we can derive that

$$\begin{aligned}
C &= \langle -\nu_0(x_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle + \frac{\alpha}{2}\|v_t^k - v_{t-1}^k\|^2 \\
&= \left\|\sqrt{\frac{\alpha}{2}}(v_t^k - v_{t-1}^k) - \frac{1}{\sqrt{2\alpha}}\nu_0(x_{t-1}^k, \xi_t^k)\right\|^2 - \frac{1}{2\alpha}\|\nu_0(x_{t-1}^k, \xi_t^k)\|^2 \\
&\geq -\frac{1}{2\alpha}\|\nu_0(x_{t-1}^k, \xi_t^k)\|^2.
\end{aligned}$$

Hence, we infer from (A.1) that for any $t \in [T],\ k \in [K]$,

$$\begin{aligned}
&\langle -\nu_0(x_{t-1}^k, \xi_t^k), x - v_{t-1}^k\rangle + \frac{\beta}{2}\|F(x, \xi_t^k)\|^2 + \langle \lambda_t^k, F(x, \xi_t^k)\rangle + \frac{\alpha}{2}(\|x - v_{t-1}^k\|^2 - \|x - v_t^k\|^2) \\
&\geq -\frac{\|\nu_0(x_{t-1}^k, \xi_t^k)\|^2}{2\alpha} + \frac{1}{2\beta}\|\lambda_{t+1}^k\|^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2.
\end{aligned}$$

Rearranging the terms leads to (3.1). □

A.2 Proof of Lemma 3.2

Proof. Proof According to the analysis of term "A" in (A.2), we obtain

$$\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(v_{t-1}^k,\xi_t^k)+\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),x-v_{t-1}^k\rangle]_+^2\le\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(x,\xi_t^k)]_+^2.$$

Hence, it implies from (A.1) that

$$\begin{aligned}&\langle-\nu_0(x_{t-1}^k,\xi_t^k),x-v_t^k\rangle+\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(x,\xi_t^k)]_+^2+\frac{\alpha}{2}\|x-v_{t-1}^k\|^2-\frac{\alpha}{2}\|x-v_t^k\|^2\\&\ge\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(v_{t-1}^k,\xi_t^k)+\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle]_+^2+\frac{\alpha}{2}\|v_t^k-v_{t-1}^k\|^2.\end{aligned}$$

By setting $x=v_{t-1}^k$, we derive

$$\begin{aligned}\alpha\|v_t^k-v_{t-1}^k\|^2&\le\langle-\nu_0(x_{t-1}^k,\xi_t^k),v_{t-1}^k-v_t^k\rangle+\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(v_{t-1}^k,\xi_t^k)]_+^2\\&\quad-\frac{1}{2\beta}\sum_{i=1}^{m}[\lambda_t^k(i)+\beta F_i(v_{t-1}^k,\xi_t^k)+\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle]_+^2\\&\le\langle-\nu_0(x_{t-1}^k,\xi_t^k),v_{t-1}^k-v_t^k\rangle+\frac{1}{2\beta}\sum_{i=1}^{m}[2[\lambda_t^k(i)+\beta F_i(v_{t-1}^k,\xi_t^k)]_+\\&\quad\cdot|\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle|+(\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle)^2]\\&\le\langle-\nu_0(x_{t-1}^k,\xi_t^k),v_{t-1}^k-v_t^k\rangle+\frac{1}{2\beta}\sum_{i=1}^{m}[2(|\lambda_t^k(i)|+\beta[F_i(v_{t-1}^k,\xi_t^k)]_+)\\&\quad\cdot|\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle|+(\beta\langle\nu_i(v_{t-1}^k,\xi_t^k),v_t^k-v_{t-1}^k\rangle)^2]\\&\le\|\nu_0(x_{t-1}^k,\xi_t^k)\|\|v_{t-1}^k-v_t^k\|+\sum_{i=1}^{m}\Big[(|\lambda_t^k(i)|+\beta[F_i(v_{t-1}^k,\xi_t^k)]_+)\\&\quad\cdot M_3\|v_t^k-v_{t-1}^k\|+\beta\frac{M_3^2}{2}\|v_t^k-v_{t-1}^k\|^2\Big]\\&\le\|\nu_0(x_{t-1}^k,\xi_t^k)\|\|v_{t-1}^k-v_t^k\|+(\sqrt{m}\|\lambda_t^k\|+\beta\sqrt{m}\|[F(v_{t-1}^k,\xi_t^k)]_+\|)\\&\quad\cdot M_3\|v_t^k-v_{t-1}^k\|+\beta\frac{mM_3^2}{2}\|v_t^k-v_{t-1}^k\|^2\end{aligned}$$

where the second and third inequalities are due to $[a]_+^2-[b]_+^2\le 2[a]_+|a-b|+(a-b)^2$ and $[a+b]_+\le|a|+[b]_+$ for any $a,b\in\mathbb{R}$, the fourth inequality is due to Cauchy-Schwarz Inequality together with the Assumption 1.2. Dividing both sides by $\|v_t^k-v_{t-1}^k\|$, we attain

$$\alpha\|v_t^k-v_{t-1}^k\|\le\|\nu_0(x_{t-1}^k,\xi_t^k)\|+(\sqrt{m}\|\lambda_t^k\|+\beta\sqrt{m}\|[F(v_{t-1}^k,\xi_t^k)]_+\|)M_3+\beta\frac{mM_3^2}{2}\|v_t^k-v_{t-1}^k\|,$$

which further yields the final result due to $2\alpha-\beta mM_3^2>0$. □

A.3 Proof of Lemma 3.3

Proof. Proof For any $t \in [T],\ k \in [K],\ i \in [m]$, we have

$$\begin{aligned}\mathbb{E}_{\xi_t^k}[\lambda_{t+1}^k(i)] &= \mathbb{E}_{\xi_t^k}[[\lambda_t^k(i) + \beta F_i(v_{t-1}^k, \xi_t^k) + \beta\langle \nu_i(v_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle]_+] \\ &\leq [\lambda_t^k(i)]_+ + \beta\mathbb{E}_{\xi_t^k}[[F_i(v_{t-1}^k, \xi_t^k)]_+] + \beta\mathbb{E}_{\xi_t^k}[[\langle \nu_i(v_{t-1}^k, \xi_t^k), v_t^k - v_{t-1}^k\rangle]_+] \\ &\leq \lambda_t^k(i) + \beta M_1 + \beta\mathbb{E}_{\xi_t^k}[\|\nu_i(v_{t-1}^k, \xi_t^k)\|\|v_t^k - v_{t-1}^k\|] \\ &\leq \lambda_t^k(i) + \beta M_1 + \beta M_3 \mathrm{d}(X),\end{aligned}$$

where the first equality follows from (2.5), the first inequality is by the fact $[a+b]_+ \leq [a]_+ + [b]_+$, for any $a, b \in \mathbb{R}$, the second inequality is because of $\lambda_t^k(i) \geq 0$, Assumption 1.3 and Cauchy-Schwarz Inequality, and the last inequality is from Assumption 1.2. Then we take expectation with ξ, proving that

$$\mathbb{E}_\xi[\lambda_{t+1}^k(i)] \leq \mathbb{E}_\xi[\lambda_t^k(i)] + \beta M_1 + \beta M_3 \mathrm{d}(X),$$

In other words, $\mathbb{E}_\xi[\lambda_{t+1}^k(i)]$ increases at most $\beta(M_1 + M_3\mathrm{d}(X))$ based on $\mathbb{E}_\xi[\lambda_t^k(i)]$ for any $k \in [K]$. Using these recursive relations, and recalling that $\lambda_1^k(i) = \bar{\lambda}^k(i)$, we can bound $\mathbb{E}_\xi[\lambda_t^k(i)]$ and $\mathbb{E}_\xi[\|\lambda_{t+1}^k\|]$ by

$$\mathbb{E}_\xi[\lambda_t^k(i)] \leq \bar{\lambda}^k(i) + (t-1)\beta(M_1 + M_3\mathrm{d}(X)), \tag{A.3}$$

$$\mathbb{E}_\xi[\|\lambda_{t+1}^k\|] \leq \mathbb{E}_\xi[\|\lambda_t^k + \beta(M_1 + M_3\mathrm{d}(X))\mathbf{1}_m\|] \leq \mathbb{E}_\xi[\|\lambda_t^k\|] + \beta(M_1 + M_3\mathrm{d}(X))\sqrt{m}. \tag{A.4}$$

We now introduce $\gamma := \lceil \frac{1}{\beta} \rceil$. Our subsequent analysis splits into two cases for $t \in [T]$.

Case 1. $t \in \{1, \ldots, \gamma + 1\}$. It follows from (A.3) that

$$\begin{aligned}\mathbb{E}_\xi[\|\lambda_t^k\|] &\leq \|\bar{\lambda}^k + (t-1)\beta(M_1 + M_3\mathrm{d}(X))\mathbf{1}_m\| \\ &\leq \|\bar{\lambda}^k\| + (t-1)\beta(M_1 + M_3\mathrm{d}(X))\sqrt{m} \\ &\leq \|\bar{\lambda}^k\| + 2(M_1 + M_3\mathrm{d}(X))\sqrt{m} \leq \theta.\end{aligned}$$

Hence, the conclusion holds.

Case 2. $t \in \{\gamma+1, \gamma+2, \ldots, T\}$. We will prove the conclusion by induction. Note that we have proved the case when $t = \gamma + 1$ in Case 1. Now suppose $\mathbb{E}_\xi[\|\lambda_t^k\|] \leq \theta$ holds for any $\gamma + 1 < t \leq T - 1$. We next show that $\mathbb{E}_\xi[\|\lambda_T^k\|] \leq \theta$. It suffices to prove the conclusion when $\mathbb{E}_\xi[\|\lambda_T^k\|] > \mathbb{E}_\xi[\|\lambda_{T-\gamma}^k\|]$. From Lemma 3.1, for any $x \in X,\ t \in \{T-\gamma,\ T-\gamma+1, \ldots, T-1\}$ and $k \in [K]$, it holds that

$$\begin{aligned}\frac{1}{2\beta}\|\lambda_{t+1}^k\|^2 - \frac{1}{2\beta}\|\lambda_t^k\|^2 \leq \|\nu_0(x_{t-1}^k, \xi_t^k)\|\mathrm{d}(X) + \frac{1}{\alpha}\frac{\|\nu_0(x_{t-1}^k, \xi_t^k)\|^2}{2} + \beta\frac{\|F(x, \xi_t^k)\|^2}{2} \\ + \langle \lambda_t^k, F(x, \xi_t^k)\rangle + \frac{\alpha}{2}(\|x - v_{t-1}^k\|^2 - \|x - v_t^k\|^2).\end{aligned}$$

Plugging $x = \hat{x}$ which satisfies the Slater's condition as in Assumption 1.4, taking conditional expectation given $\mathcal{F}_t^k$ with respect to ξ_t^k and utilizing Assumptions 1.3 and 1.2, we obtain

$$\begin{aligned}&\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2] - \frac{1}{2\beta}\|\lambda_t^k\|^2 \\ &\quad\leq M_2\mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{\mathbb{E}_{\xi_t^k}[\|F(\hat{x}, \xi_t^k)\|^2]}{2} + \langle \lambda_t^k, \mathbb{E}_{\xi_t^k}[F(\hat{x}, \xi_t^k)]\rangle + \frac{\alpha}{2}(\|\hat{x} - v_{t-1}^k\|^2 - \mathbb{E}_{\xi_t^k}[\|\hat{x} - v_t^k\|^2]) \\ &\quad= M_2\mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2} + \langle \lambda_t^k, f(\hat{x})\rangle + \frac{\alpha}{2}\left(\|\hat{x} - v_{t-1}^k\|^2 - \mathbb{E}_{\xi_t^k}[\|\hat{x} - v_t^k\|^2]\right).\end{aligned}$$

From Assumption 1.4, $\lambda_t^k(i) \geq 0$ and $\|\lambda_t^k\| \leq \sum_{i=1}^m \lambda_t^k(i)$ it indicates

$$\langle \lambda_t^k, f(\hat{x}) \rangle = \sum_{i=1}^m \lambda_t^k(i) f_i(\hat{x}) \leq \sum_{i=1}^m \lambda_t^k(i)(-M_4) \leq -M_4 \|\lambda_t^k\|,$$

which further implies that

$$\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2] - \frac{1}{2\beta}\|\lambda_t^k\|^2 \leq M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2} - M_4\|\lambda_t^k\| + \frac{\alpha}{2}[\|\hat{x} - v_{t-1}^k\|^2 - \mathbb{E}_{\xi_t^k}[\|\hat{x} - v_t^k\|^2]].$$

Taking expectation with respect to random vectors $\xi_1^1, \xi_1^2, \ldots, \xi_T^K$, we derive

$$\frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_{t+1}^k\|^2] - \frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_t^k\|^2] \leq M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2} - M_4\mathbb{E}_\xi[\|\lambda_t^k\|] + \frac{\alpha}{2}\mathbb{E}_\xi[\|\hat{x} - v_{t-1}^k\|^2 - \|\hat{x} - v_t^k\|^2].$$

Then summing over all $\{T-\gamma,\ T-\gamma+1, \ldots, T-1\}$ yields

$$\begin{aligned}
&\frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_T^k\|^2 - \|\lambda_{T-\gamma}^k\|^2] \\
&\quad \leq \Big(M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2}\Big)\gamma - M_4 \sum_{t=T-\gamma}^{T-1} \mathbb{E}_\xi[\|\lambda_t^k\|] + \frac{\alpha}{2}\mathbb{E}_\xi[\|\hat{x} - v_{T-\gamma-1}^k\|^2 - \|\hat{x} - v_{T-1}^k\|^2].
\end{aligned}$$

Recall that $\mathbb{E}_\xi[\|\lambda_T^k\|] > \mathbb{E}_\xi[\|\lambda_{T-\gamma}^k\|]$, which implies

$$0 < \Big(M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2}\Big)\gamma - M_4 \sum_{t=T-\gamma}^{T-1} \mathbb{E}_\xi[\|\lambda_t^k\|] + \frac{\alpha}{2}\mathrm{d}(X)^2.$$

By rearranging terms of above inequality the following relation holds:

$$M_4 \sum_{t=T-\gamma}^{T-1} \mathbb{E}_\xi[\|\lambda_t^k\|] < \Big(M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2}\Big)\gamma + \frac{\alpha}{2}\mathrm{d}(X)^2. \tag{A.5}$$

Assume, for a contradiction, that $\mathbb{E}_\xi[\|\lambda_T^k\|] > \theta$, which joint with (A.4) indicates

$$\mathbb{E}_\xi[\|\lambda_t^k\|] > \theta - (T-t)\beta(M_1 + M_3\mathrm{d}(X))\sqrt{m},$$

for any $t \in \{T-\gamma,\ T-\gamma+1, \ldots, T-1\}$. Thus, the left hand side of (A.5) is lower bounded by

$$\begin{aligned}
M_4 \sum_{t=T-\gamma}^{T-1} \mathbb{E}_\xi[\|\lambda_t^k\|] &> M_4 \sum_{t=T-\gamma}^{T-1} [\theta - (T-t)\beta(M_1 + M_3\mathrm{d}(X))\sqrt{m}] \\
&= \theta\gamma M_4 - \beta\frac{(1+\gamma)\gamma}{2} M_4\sqrt{m}(M_1 + M_3\mathrm{d}(X)) \\
&\geq \theta\gamma M_4 - (1+\gamma)M_4\sqrt{m}(M_1 + M_3\mathrm{d}(X)),
\end{aligned}$$

where the last inequality is due to $\beta\gamma = \beta \cdot \lceil \frac{1}{\beta} \rceil \leq 1 + \beta \leq 2$. Then combing with (A.5), we obtain

$$\theta\gamma M_4 - (1+\gamma)M_4\sqrt{m}(M_1 + M_3\mathrm{d}(X)) < \Big(M_2 \mathrm{d}(X) + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2}\Big)\gamma + \frac{\alpha}{2}\mathrm{d}(X)^2.$$

It leads to

$$\begin{aligned}\theta &< \Big(1+\frac{1}{\gamma}\Big)\sqrt{m}(M_1+M_3\mathrm{d}(X)) + \Big(M_2\mathrm{d}(X)+\frac{1}{\alpha}\frac{M_2^2}{2}+\beta\frac{M_1^2}{2}\Big)\frac{1}{M_4}+\alpha\frac{1}{\gamma}\frac{\mathrm{d}(X)^2}{2M_4}\\ &\le (1+\beta)\sqrt{m}(M_1+M_3\mathrm{d}(X)) + \Big(M_2\mathrm{d}(X)+\frac{1}{\alpha}\frac{M_2^2}{2}+\beta\frac{M_1^2}{2}\Big)\frac{1}{M_4}+\alpha\beta\frac{\mathrm{d}(X)^2}{2M_4}\\ &= \sqrt{m}(M_1+M_3\mathrm{d}(X)) + \frac{M_2\mathrm{d}(X)}{M_4}+\frac{1}{\alpha}\frac{M_2^2}{2M_4}+\beta\sqrt{m}(M_1+M_3\mathrm{d}(X))+\beta\frac{M_1^2}{2M_4}+\alpha\beta\frac{\mathrm{d}(X)^2}{2M_4}.\end{aligned}$$

This however contradicts the setting of θ. Hence, conclusion of the lemma is derived. □

B Proofs of lemmas in Section 4

B.1 Proof of Lemma 4.1

Proof. Proof Observe that this is certainly true for $k=1$ from $x_t^1 \in \arg\min_{x\in X}\|x\|_\infty$ for any $t\in[T]\cup\{0\}$. When $k\ge 2$, according to Step 10 in Algorithm 1, it yields from $\frac{1}{K}=\sqrt{a^k}-\sqrt{a^{k-1}}$, $\frac{a^{k-1}}{a^k}\le 1$ and $v_t^{k-1}\in X\subseteq[0,1]^n$ that

$$\begin{aligned}\mathbf{1}_n - x_t^k &= \Big(\mathbf{1}_n - x_t^{k-1}\Big)-\Big(x_t^k-x_t^{k-1}\Big) = \Big(\mathbf{1}_n - x_t^{k-1}\Big)-\Big(v_t^{k-1}-x_t^{k-1}\Big)\frac{1}{K}\frac{\sqrt{a^{k-1}}}{a^k}\\ &\ge \Big(\mathbf{1}_n - x_t^{k-1}\Big)-\Big(\mathbf{1}_n - x_t^{k-1}\Big)\left(\frac{\sqrt{a^{k-1}}}{\sqrt{a^k}}-\frac{a^{k-1}}{a^k}\right)\ge\Big(\mathbf{1}_n - x_t^{k-1}\Big)\frac{\sqrt{a^{k-1}}}{\sqrt{a^k}}.\end{aligned}$$

Continuing this procedure and considering that $a^1=1$, we reach the conclusion. □

B.2 Proof of Lemma 4.2

Proof. Proof For any $t\in[T]$, $k\in[K+1]$, by (1.2) and Lemma 4.1, we attain that

$$\begin{aligned}\langle\nabla f_0(x_{t-1}^k), x^*-x_{t-1}^k\rangle &\ge f_0(x_{t-1}^k\vee x^*)+f_0(x_{t-1}^k\wedge x^*)-2f_0(x_{t-1}^k)\\ &\ge f_0(x_{t-1}^k\vee x^*)+0-2f_0(x_{t-1}^k)\\ &\ge(1-\|x_{t-1}^k\|_\infty)f_0(x^*)-2f_0(x_{t-1}^k)\\ &\ge\frac{1-\min_{x\in X}\|x\|_\infty}{\sqrt{a^k}}f_0(x^*)-2f_0(x_{t-1}^k).\end{aligned}$$

Rearranging these terms derives

$$\Big(1-\min_{x\in X}\|x\|_\infty\Big)f_0(x^*)\le\sqrt{a^k}\big(\langle\nabla f_0(x_{t-1}^k),x^*-x_{t-1}^k\rangle+2f_0(x_{t-1}^k)\big). \tag{B.1}$$

Then it yields that

$$
\begin{aligned}
& a^{k+1} f_0(x_{t-1}^{k+1}) - \frac{1-\min_{x\in X}\|x\|_\infty}{K} f_0(x^*) - a^k f_0(x_{t-1}^k) \\
&= a^{k+1}\Big(f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k)\Big) + (a^{k+1}-a^k) f_0(x_{t-1}^k) - \frac{1-\min_{x\in X}\|x\|_\infty}{K} f_0(x^*) \\
&= a^{k+1}\left(f_0\left(x_{t-1}^k + (v_{t-1}^k - x_{t-1}^k)\frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}}\right) - f_0(x_{t-1}^k)\right) + (a^{k+1}-a^k) f_0(x_{t-1}^k) \\
&\quad - \frac{1-\min_{x\in X}\|x\|_\infty}{K} f_0(x^*) \\
&\geq a^{k+1}\left(\left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x_{t-1}^k)\frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\frac{a^k}{(a^{k+1})^2}\|v_{t-1}^k - x_{t-1}^k\|^2\right) \\
&\quad + (a^{k+1}-a^k) f_0(x_{t-1}^k) - \frac{1-\min_{x\in X}\|x\|_\infty}{K} f_0(x^*) \\
&\geq a^{k+1}\left(\left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x_{t-1}^k)\frac{1}{K}\frac{\sqrt{a^k}}{a^{k+1}}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\frac{a^k}{(a^{k+1})^2}\mathrm{d}(X)^2\right) \\
&\quad + (a^{k+1}-a^k) f_0(x_{t-1}^k) - \frac{\sqrt{a^k}}{K}\left(\left\langle \nabla f_0(x_{t-1}^k), x^* - x_{t-1}^k\right\rangle + 2 f_0(x_{t-1}^k)\right) \\
&\geq f_0(x_{t-1}^k)\left((a^{k+1}-a^k) - \frac{2\sqrt{a^k}}{K}\right) + \left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x^*)\frac{\sqrt{a^k}}{K}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2 \\
&= f_0(x_{t-1}^k)\frac{1}{K^2} + \left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x^*)\frac{\sqrt{a^k}}{K}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2 \\
&\geq \left\langle \nabla f_0(x_{t-1}^k), (v_{t-1}^k - x^*)\frac{\sqrt{a^k}}{K}\right\rangle - \frac{L_0}{2}\frac{1}{K^2}\mathrm{d}(X)^2
\end{aligned}
$$

for any $t \in [T]$, $k \in [K]$, where the second equality follows from the update of x_{t-1}^{k+1} in Algorithm 1, the first inequality is because of the L_0-smoothness of f_0, the second inequality is due to (B.1) and the definition of $\mathrm{d}(X)$, and the last equality follows from the setting of $a^k, k \in [K+1]$. □

C Proofs of theorems in Section 5

C.1 Proof of Theorem 5.1

Proof. Proof The monotonicity of f_0, along with the DR-submodularity of f_0 as given in (1.2), ensures that

$$
\begin{aligned}
\langle -\nabla f_0(x_{t-1}^k), x^*\rangle \leq \langle -\nabla f_0(x_{t-1}^k), (x^* - x_{t-1}^k)\vee \mathbf{0}_n\rangle &= \langle -\nabla f_0(x_{t-1}^k), x^*\vee x_{t-1}^k - x_{t-1}^k\rangle \\
&\leq -(f_0(x^*\vee x_{t-1}^k) - f_0(x_{t-1}^k)) \\
&\leq -(f_0(x^*) - f_0(x_{t-1}^k)).
\end{aligned}
\tag{C.1}
$$

Besides, the L_0-smoothness of f_0 joint with the update (5.1) leads to

$$
\begin{aligned}
f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k) &= f_0\left(x_{t-1}^k + \frac{1}{K}v_{t-1}^k\right) - f_0(x_{t-1}^k) \\
&\geq \left\langle \nabla f_0(x_{t-1}^k), \frac{1}{K}v_{t-1}^k\right\rangle - \frac{L_0}{2K^2}\|v_{t-1}^k\|^2 \\
&\geq \left\langle \nabla f_0(x_{t-1}^k), \frac{1}{K}v_{t-1}^k\right\rangle - \frac{L_0\mathrm{d}(X)^2}{2K^2}
\end{aligned}
$$

implying that

$$\langle \nabla f_0(x_{t-1}^k), v_{t-1}^k \rangle \leq K(f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k)) + \frac{L_0 \mathrm{d}(X)^2}{2K}. \tag{C.2}$$

Adding (C.1) and (C.2) enables us gain the following bound

$$\langle \nabla f_0(x_{t-1}^k), -x^* + v_{t-1}^k \rangle \leq -(f_0(x^*) - f_0(x_{t-1}^k)) + K(f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k)) + \frac{L_0 \mathrm{d}(X)^2}{2K}.$$

which, together with (4.3), $f(x^*) \leq \mathbf{0}_m$ and $\lambda_t^k \geq \mathbf{0}_m$, indicates that

$$\begin{aligned}\frac{1}{2\beta}\mathbb{E}_{\xi_t^k}[\|\lambda_{t+1}^k\|^2] - \frac{1}{2\beta}\|\lambda_t^k\|^2 \leq & -(f_0(x^*) - f_0(x_{t-1}^k)) + K(f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k)) + \frac{L_0 \mathrm{d}(X)^2}{2K} \\ & + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2} + 0 + \frac{\alpha}{2}(\|x^* - v_{t-1}^k\|^2 - \mathbb{E}_{\xi_t^k}[\|x^* - v_t^k\|^2]).\end{aligned}$$

By taking expectation with respect to all the random vectors $\xi_1^1, \xi_1^2, \ldots, \xi_T^K$ we obtain

$$\begin{aligned}\frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_{t+1}^k\|^2] - \frac{1}{2\beta}\mathbb{E}_\xi[\|\lambda_t^k\|^2] \leq & -\mathbb{E}_\xi[f_0(x^*) - f_0(x_{t-1}^k)] + K\mathbb{E}_\xi[f_0(x_{t-1}^{k+1}) - f_0(x_{t-1}^k)] \\ & + \frac{L_0 \mathrm{d}(X)^2}{2K} + \frac{1}{\alpha}\frac{M_2^2}{2} + \beta\frac{M_1^2}{2} + \frac{\alpha}{2}(\mathbb{E}_\xi[\|x^* - v_{t-1}^k\|^2] - \mathbb{E}_\xi[\|x^* - v_t^k\|^2]).\end{aligned}$$

After rearrangement, dividing both sides by K shows that

$$\begin{aligned}\mathbb{E}_\xi[f_0(x^*) - f_0(x_{t-1}^{k+1})] \leq & \Big(1 - \frac{1}{K}\Big)\mathbb{E}_\xi[f_0(x^*) - f_0(x_{t-1}^k)] + \frac{1}{K^2}\frac{L_0 \mathrm{d}(X)^2}{2} + \frac{1}{K\alpha}\frac{M_2^2}{2} + \frac{\beta}{K}\frac{M_1^2}{2} \\ & + \frac{1}{2\beta K}\mathbb{E}_\xi[\|\lambda_t^k\|^2 - \|\lambda_{t+1}^k\|^2] + \frac{\alpha}{2K}\mathbb{E}_\xi[\|x^* - v_{t-1}^k\|^2 - \|x^* - v_t^k\|^2].\end{aligned}$$

By summing up the above inequality over $t = 1, \ldots, T$, we obtain from $\lambda_1^k = \bar{\lambda}^k$ and $\|x^* - v_0^k\| \leq \mathrm{d}(X)$ that

$$\begin{aligned}\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^T f_0(x_{t-1}^{k+1})] \leq & \Big(1 - \frac{1}{K}\Big)\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^T f_0(x_{t-1}^k)] + \frac{T}{K^2}\frac{L_0 \mathrm{d}(X)^2}{2} + \frac{T}{K\alpha}\frac{M_2^2}{2} + \frac{T\beta}{K}\frac{M_1^2}{2} \\ & + \frac{1}{2\beta K}\|\bar{\lambda}^k\|^2 + \frac{\alpha}{K}\frac{\mathrm{d}(X)^2}{2}, \quad \forall k \in [K].\end{aligned}$$

From the above recursive relations and denoting $C_1 = \frac{T}{K^2}\frac{L_0 \mathrm{d}(X)^2}{2} + \frac{T}{K\alpha}\frac{M_2^2}{2} + \frac{T\beta}{K}\frac{M_1^2}{2} + \frac{\alpha}{K}\frac{\mathrm{d}(X)^2}{2}$, we further derive

$$\begin{aligned}
&\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^{T} f_0(x_{t-1}^{K+1})] \\
&\quad\le \Big(1-\frac{1}{K}\Big)\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^{T} f_0(x_{t-1}^{K})] + C_1 + \frac{1}{2\beta K}\|\bar\lambda^K\|^2 \\
&\quad\le \Big(1-\frac{1}{K}\Big)\Big[\Big(1-\frac{1}{K}\Big)\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^{T} f_0(x_{t-1}^{K-1})] + C_1 + \frac{1}{2\beta K}\|\bar\lambda^{K-1}\|^2\Big] + C_1 + \frac{\|\bar\lambda^K\|^2}{2\beta K} \\
&\quad\le \Big(1-\frac{1}{K}\Big)^2\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^{T} f_0(x_{t-1}^{K-1})] + 2C_1 + \frac{1}{2\beta K}(\|\bar\lambda^{K-1}\|^2 + \|\bar\lambda^K\|^2) \le \cdots \le \\
&\quad\le \Big(1-\frac{1}{K}\Big)^K\mathbb{E}_\xi[Tf_0(x^*) - \sum_{t=1}^{T} f_0(x_{t-1}^{1})] + KC_1 + \frac{1}{2\beta K}\sum_{k=1}^{K}\|\bar\lambda^k\|^2 \\
&\quad\le \frac{1}{e}\mathbb{E}_\xi[Tf_0(x^*)] + KC_1 + \frac{1}{2\beta}\max_{k\in[K]}\|\bar\lambda^k\|^2,
\end{aligned}$$

where the last inequality follows from $(1-\frac{1}{K})^K \le \frac{1}{e}$ and $f_0(\cdot) \ge 0$. Rearranging the terms, dividing by T and applying the randomness of R imply that

$$\Big(1-\frac{1}{e}\Big)f_0(x^*) - \mathbb{E}[f_0(x_R)] \le \frac{K}{T}C_1 + \frac{1}{2\beta T}\max_{k\in[K]}\|\bar\lambda^k\|^2,$$

which indicates (5.2). □

C.2 Proof of Theorem 5.2

Proof. Proof For any $t \in [T]$, according to the update scheme defined through (5.1), we have $x_{t-1}^{K+1} = \mathbf{0}_n + \frac{1}{K}\sum_{k=1}^{K} v_{t-1}^k$. It then indicates from the convexity of f_i that

$$\mathbb{E}_\xi[f_i(x_{t-1}^{K+1})] = \mathbb{E}_\xi\Big[f_i\Big(\frac{1}{K}\sum_{k=1}^{K} v_{t-1}^k\Big)\Big] \le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_\xi[f_i(v_{t-1}^k)].$$

Therefore, we conclude from (4.5) and $I = TK$ that for any $i \in [m]$,

$$\begin{aligned}
\mathbb{E}[f_i(x_R)] =& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\xi[f_i(x_{t-1}^{K+1})] \le \frac{1}{I}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}_\xi[f_i(v_{t-1}^k)] \\
\le& \frac{1}{I}\sum_{k=1}^{K}\sum_{t=1}^{T}\Big\{\frac{1}{\beta}\mathbb{E}_\xi[\lambda_{t+1}^k(i) - \lambda_t^k(i)] + \frac{2M_3}{2\alpha - \beta m M_3^2}[M_2 + \mathbb{E}_\xi[\|\lambda_t^k\|]\sqrt{m}M_3 + \beta\sqrt{m}M_1M_3]\Big\} \\
\le& \frac{1}{I\beta}\sum_{k=1}^{K}\mathbb{E}_\xi[\lambda_{T+1}^k(i) - \lambda_1^k(i)] + \frac{2M_3M_2}{2\alpha - \beta m M_3^2} + \frac{1}{I}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}_\xi[\|\lambda_t^k\|]\frac{2\sqrt{m}M_3^2}{2\alpha - \beta m M_3^2} + \frac{2\beta\sqrt{m}M_1M_3^2}{2\alpha - \beta m M_3^2} \\
\le& \frac{1}{T\beta}\theta + \frac{2M_3M_2}{2\alpha - \beta m M_3^2} + \theta\frac{2\sqrt{m}M_3^2}{2\alpha - \beta m M_3^2} + \frac{2\beta\sqrt{m}M_1M_3^2}{2\alpha - \beta m M_3^2},
\end{aligned}$$

where the fourth inequality is due to Lemma 3.3. Therefore, the conclusion is valid relying on the definition of θ and the relationship between $\|\cdot\|$ and $\|\cdot\|_\infty$. □