On the Convergence of Constrained Gradient Method

Danqing Zhou*, Hongmei Chen†, Shiqian Ma‡, Junfeng Yang§

November 21, 2025

Abstract

The constrained gradient method (CGM) has recently been proposed to solve convex optimization and monotone variational inequality (VI) problems with general functional constraints. While existing literature has established convergence results for CGM, the assumptions employed therein are quite restrictive; in some cases, certain assumptions are mutually inconsistent, leading to gaps in the underlying analysis. This paper aims to derive rigorous and improved convergence guarantees for CGM under weaker and more reasonable assumptions, specifically in the context of strongly convex optimization and strongly monotone VI problems. Preliminary numerical experiments are provided to verify the validity of CGM and demonstrate its efficacy in addressing such problems.

Mathematical Subject Classifications: 15A18, 15A69, 65F15, 90C33

1 Introduction

The constrained gradient method (CGM) was first proposed in [MJ22] for solving the constrained convex optimization problem:

$$\min_{x\in\mathcal{C}} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a smooth and convex function, and the constraint set $\mathcal{C}\subseteq\mathbb{R}^n$ is defined by m inequality constraints as follows:

$$\mathcal{C}:=\{x\in\mathbb{R}^n \mid g_i(x)\le 0,\, i\in[m]\}, \tag{2}$$

with each g_i being a smooth convex function. Note that throughout this paper, a function is referred to as "smooth" if its gradient operator is Lipschitz continuous, and $[m]:=\{1,\dots,m\}$. This method has been further extended to accelerated variants [MJ25], online and stochastic nonconvex minimization settings [KMM23, STM+22, STMM23], and notably, to solving variational inequality (VI) problems [ZHM25]. A VI problem seeks an $x^*\in\mathcal{C}$ satisfying

$$\langle F(x^*), x^*-x\rangle\le 0,\quad \forall x\in\mathcal{C}, \tag{3}$$

where $F:\mathbb{R}^n\to\mathbb{R}^n$ is a monotone operator. However, as we will discuss in more detail later, the assumptions in [MJ22] and [ZHM25] are quite restrictive, and certain assumptions in [ZHM25] are even mutually inconsistent. The main purpose of this paper is to establish convergence guarantees for CGM under weaker and more reasonable assumptions, specifically when f in (1) is strongly convex and F in (3) is strongly monotone.

Both problems (1) and (3) are of great interest. For instance, problem (1) encompasses a wide variety of applications, including but not limited to operations planning [FJVW86, PEAM21], support vector machines [CV95, HDO+98, GF17], and reinforcement learning [KF11]; and problem (3) offers a powerful, unified framework for modeling and analyzing numerous problems, such as equilibrium problems

*School of Mathematics, Nanjing Audit University, Nanjing, P. R. China. Email: zhoudanqing@nau.edu.cn.

†School of Mathematical Sciences, Sichuan Normal University, Chengdu, P. R. China. Email: hmchen@sicnu.edu.cn.

‡Department of Computational Applied Mathematics and Operations Research, Rice University, Houston, TX, USA. Email: sqma@rice.edu.

§School of Mathematics, Nanjing University, Nanjing, P. R. China. Email: jfyang@nju.edu.cn

[Nas50, Nas51, Nag98], optimization problems (e.g., matrix minimization and generative adversarial networks) [He18, TYH11, GBV+19], and control theory [SB84, NT88]. A more thorough review of algorithms for solving problems (1) and (3) is provided in Appendix A.

CGM for solving problem (1) mirrors the streamlined design of traditional projection-based methods in both the generation of descent directions and the choice of line-search-free step sizes. To avoid the notorious difficulty of projecting onto $\mathcal{C}$, it does not project each iterate directly onto the feasible region; instead, it projects the update direction onto a local and sparse velocity polytope. Built around the iterate x, this polytope takes the form:

$$V_\alpha(x) = \{v \in \mathbb{R}^n \mid \alpha g_i(x) + \nabla g_i(x)^\top v \le 0,\ \forall i \in I_x\} \quad \text{with} \quad I_x := \{i \in [m] \mid g_i(x) \ge 0\}. \tag{4}$$

Here, I_x denotes the set of constraints active at x, and $\alpha > 0$ controls the tradeoff between optimizing the objective function and maintaining feasibility. This design embodies CGM's distinctive hallmark, differentiating it from traditional projection-based methods. Specifically, at the $(t+1)$-th iteration, CGM updates the iterates as follows:

$$\begin{aligned} v^t &= \arg\min_{v \in V_\alpha(x^t)} \|v + \nabla f(x^t)\|^2, \\ x^{t+1} &= x^t + \eta v^t, \end{aligned} \tag{5}$$

where $\eta > 0$ denotes the step size. Regarding the update rule (5), [MJ22] established the following results:

Results adapted from [MJ22]. *Assume the following assumptions:*

(i) The Mangasarian-Fromovitz constraint qualification (MFCQ) holds at every $x \in \mathbb{R}^n$[*]*;*

(ii) f is closed, proper, μ-strongly convex and smooth; all g_i's are smooth;

(iii) The feasible set $\mathcal{C}$ is nonempty, convex and bounded;

(iv) The parameters α and η satisfy $0 < \alpha < \mu$ and $\eta \le 2/(\ell_* + \mu)$, where ℓ_* denotes an upper bound of the smoothness constant of the Lagrangian function $\mathcal{L}(x, \lambda(x)) = f(x) + \sum_{i=1}^m \lambda_i(x) g_i(x)$, and

$$\lambda(x) \in \arg\max \left\{ \mathcal{L}(x,\lambda) - \frac{1}{2\alpha}\|\nabla_x \mathcal{L}(x,\lambda)\|^2 \ \middle|\ \lambda \in \mathbb{R}^m_+,\ \lambda_i = 0,\ \forall i \notin I_x \right\}.$$

Then, the sequence of iterates $\{x^t\}$ generated by (5) from any initial point $x^0 \in \mathbb{R}^n$ converges to the unique minimizer of problem (1). Furthermore, there exist a constant $C > 0$ and a positive integer N such that

$$\min_{t \in \{0,\dots,T\}} \|x^t - x^*\|^2 \le C/T, \quad \forall T \ge N.$$

Note that assumption (iv) is apparently impractical, since ℓ_* is typically unavailable, and therefore it is not clear how to choose η in practice.

Recently, the authors of [ZHM25] extended CGM to solve the VI problem (3), with several modifications made to the update rule (5). Specifically, to ensure the boundedness of the sequences $\{x^t\}$ and $\{v^t\}$, they introduced an auxiliary constraint $g_{m+1}(x) = \|x\|^2 - \widehat{D}^2 \le 0$, where $\widehat{D} > 0$ denotes an upper bound on the norm of elements in $\mathcal{C}$, i.e., $\|x\| \le \widehat{D}$ for all $x \in \mathcal{C}$. As a result, the set of active constraints at x in [ZHM25] is adjusted accordingly to $I_x = \{i \in [m+1] \mid g_i(x) \ge 0\}$. Their CGM for the VIs iterates as follows:

$$\begin{aligned} v^t &= \arg\min_{v \in V_\alpha(x^t)} \|v + F(x^t)\|^2, \\ x^{t+1} &= x^t + \eta_t v^t. \end{aligned} \tag{6}$$

Note that the definition of $V_\alpha(x^t)$ in (6) differs from that in (5) due to the difference in the definition of the active set I_x. Given that $\mathcal{C} \subset \mathcal{B}(0, \widehat{D})$, [ZHM25] analyzed the convergence of (6) under two settings: when F is monotone and when F is strongly monotone. The main results for the strongly monotone case are summarized as follows:

[*] In particular, this condition necessitates that for every $x \in \mathbb{R}^n$ and each $i \in I_x$, there exists a vector $w \in \mathbb{R}^n$ such that $\langle \nabla g_i(x), w \rangle < 0$.

Results adapted from [ZHM25]. *Assume the following assumptions:*

(i) F is continuous, μ-strongly monotone, and L_F-bounded, i.e., $\|F(x)\| \leq L_F$ for all $x \in \mathbb{R}^n$;

(ii) All g_i's are convex, L_g-Lipschitz continuous, and ℓ_g-smooth;

(iii) The feasible set $\mathcal{C}$ is nonempty, closed and bounded, i.e., $\mathcal{C} \subseteq \mathcal{B}(0, \widehat{D})$.

Let $\{x^t\}$ be the sequence generated by (6). Then CGM in (6) achieves the following convergence results:

(a) Let $T \geq 2$ be any integer. When the parameters are chosen as $\eta_t = \frac{1}{\mu(t+1)}$ for $t = 0, \ldots, T$, and $\alpha = \frac{\mu(\gamma-1)}{\gamma+1}$ with $\gamma > 1$, there hold

$$\langle F(x), \bar{x}^T - x \rangle = \mathcal{O}\left(\frac{1}{T}\right), \ \forall x \in \mathcal{C}, \ \text{and} \ g_i(\bar{x}^T) = \mathcal{O}\left(\frac{\gamma^2 + \gamma^2\zeta(1 + 2/(\gamma+1))}{(T+1)^{1-2/(\gamma+1)}}\right), \ \forall i \in [m+1],$$

where $\bar{x}^T = \frac{2}{T(T-1)} \sum_{t=0}^{T-1} t x^t$, and $\zeta(p) = \sum_{s=1}^{\infty} 1/s^p$ represents the Riemann zeta function.

(b) When $F = \nabla f$, with f being a μ-strongly convex and ℓ_f-smooth function, and the parameters are chosen as $\eta_t \equiv \eta = (\log T)/(\mu T)$, $\alpha = \mu$, where T satisfies $T \geq \max\{3, \ell_f/\mu\} \log T$, there hold

$$f(x^T) - f(x^*) = \mathcal{O}\left(\frac{1}{T}\right), \quad \text{and} \ g_i(x^T) = \tilde{\mathcal{O}}\left(\frac{1}{T}\right), \ \forall i \in [m+1].$$

Here, we note that the convergence bound for $g_i(\bar{x}^T)$ in part (a) is weaker than $\mathcal{O}(1/T)$: while taking $\gamma \to +\infty$ makes the denominator approach $T+1$, the numerator also tends to $+\infty$ in this case, which degenerates the bound. Moreover, assumption (i) above is flawed, as discussed below. First, it is not feasible to impose the joint assumptions that the operator F is both μ-strongly monotone and L_F-bounded. To verify this point, suppose for contradiction that both conditions hold. Then, choose $x, y \in \mathbb{R}^n$ satisfying $\|y - x\| > 2L_F/\mu$; we have

$$2L_F\|y - x\| \geq \|F(y) - F(x)\|\|y - x\| \geq \langle F(y) - F(x), y - x \rangle \geq \mu\|y - x\|^2.$$

This would imply that $\|y - x\| \leq 2L_F/\mu$, which cannot be true for arbitrary x and y. Second, for the minimization problem (1), we have $F = \nabla f$, and the requirement that $\|F(x)\| \leq L_F$ for all x simplifies to $\|\nabla f(x)\| \leq L_f$ for all x. This, as clearly noted in [Gri19], contradicts the μ-strong convexity of f.

Motivation and contributions. Based on the above discussions, our primary motivation in this paper is to provide convergence analyses for CGM under more reasonable and justifiable assumptions. Our contributions are twofold.

(i) For strongly convex optimization problems with general functional constraints, we establish the convergence of CGM without relying on the flawed assumption $\|\nabla f\| \leq L_f$ used in [ZHM25], while further relaxing other requirements. Specifically, we neither assume that the constraint set $\mathcal{C}$ is bounded nor that all g_i's are L_g-Lipschitz continuous. Additionally, we establish the same convergence rates as in [ZHM25] for both constant and varying step sizes. In contrast to [MJ22], our analysis does not require dual information, such as ℓ_*, which is not practically available.

(ii) For strongly monotone VIs with general functional constraints, we conduct a convergence analysis of CGM under more appropriate and relaxed conditions. Specifically, our analysis replaces the flawed assumption $\|F(x)\| \leq L_F$ for all x with a more reasonable one, eliminates the requirement that all g_i's are L_g-Lipschitz continuous, and improves upon the convergence rates in [ZHM25]—which rely on the aforementioned flawed assumptions.

Organization. The rest of this paper is organized as follows. In Section 2, we introduce the notation, optimality measures employed throughout the paper, and present several commonly used lemmas. In Section 3, we provide a refined convergence analysis for the minimization problem. We exclude inappropriate assumptions and the auxiliary constraint, and establish convergence rates for both constant and varying step sizes. In Section 4, we conduct a refined convergence analysis for VIs under relaxed conditions. We derive a convergence rate for the optimality measure that matches the original result, while achieving an improved rate for the feasibility measure compared to the result in [ZHM25]. In Section 5, we carry out numerical experiments to validate the proposed convergence results. Finally, we draw some concluding remarks in Section 6.

2 Preliminaries

In this section, we first introduce the notation and optimality measures employed throughout the paper, followed by the presentation of two useful lemmas.

Notation. Throughout the paper, we use the following notation. The cardinality of a set $\mathcal{W}$ is denoted by $|\mathcal{W}|$. Let $\mathbb{R}^n$ be a finite-dimensional Euclidean space, with the inner products and its induced norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$, respectively. Let $\mathbb{R}^n_+$ denote the set of n-dimensional real vectors with nonnegative entries. The standard basis vector in $\mathbb{R}^n$ is denoted by $\mathbf{e}_i = (\delta_{ij})_{j=1}^n$, where δ_{ij} is the Kronecker delta, i.e., $\delta_{ij} = 1$ if $j = i$, and $\delta_{ij} = 0$ otherwise. Furthermore, $\mathbf{1}_n \in \mathbb{R}^n$ denotes the n-dimensional column vector of all ones, and $\mathbf{0}_n \in \mathbb{R}^n$ the n-dimensional zero vector; I_n stands for the identity matrix of order n. For a symmetric matrix M of appropriate dimensions, $M \succeq 0$ (or $M \succ 0$) denotes that M is positive semidefinite (or positive definite), respectively.

An operator $F : \mathcal{X} \to \mathbb{R}^n$ defined on a convex set $\mathcal{X} \subseteq \mathbb{R}^n$ is called μ-strongly monotone with parameter $\mu > 0$ if $\langle F(x) - F(y), x - y\rangle \geq \mu\|x - y\|^2$ for all $x, y \in \mathcal{X}$. When $\mu = 0$, the operator is simply referred to as a monotone operator. If there exists $\ell_F > 0$ such that $\|F(x) - F(y)\| \leq \ell_F\|x - y\|$ for all $x, y \in \mathcal{X}$, then F is ℓ_F-Lipschitz continuous. Similarly, a function $\theta : \mathcal{X} \to \mathbb{R}$ is said to be L_θ-Lipschitz if $|\theta(x) - \theta(y)| \leq L_\theta\|x - y\|$ for all $x, y \in \mathcal{X}$. When $\theta(\cdot)$ is differentiable, L_θ-Lipschitz is equivalent to $\|\nabla\theta(x)\| \leq L_\theta$ for all $x \in \mathcal{X}$. A function θ is called ℓ_θ-smooth if it is differentiable and its gradient operator is ℓ_θ-Lipschitz, i.e., $\|\nabla\theta(x) - \nabla\theta(y)\| \leq \ell_\theta\|x - y\|$ for all $x, y \in \mathcal{X}$. We denote the set of optimal solutions to problem (3) as $\mathcal{X}^*$, and let x^* be any element in $\mathcal{X}^*$. Finally, $\tilde{\mathcal{O}}$ is the big-$\mathcal{O}$ notation with logarithmic factors ignored.

Optimality measures. The VI problem in (3) is commonly designated as the Stampacchia VI problem, and a solution to (3) is referred to as a strong solution. In this paper, our goal is to find a weak solution $x^* \in \mathcal{C}$, which solves the Minty VI problem:

$$\langle F(x), x^* - x\rangle \leq 0, \quad \forall x \in \mathcal{C}. \tag{7}$$

This Minty VI framework has been widely adopted, see, e.g., [Nem04, Nes07, JNT11, ZHM25]. It is well known (see, e.g., [Min62, KS00]) that if F is continuous and monotone, the strong and weak solutions coincide. For an approximate solution $z \in \mathbb{R}^n$, we use $\max_{x\in\mathcal{C}}\langle F(x), z - x\rangle$ to measure optimality and $g_i(z)$ ($i \in [m]$) to measure feasibility. Specifically, we adopt the notion of *weak ϵ-solution* as in [ZHM25].

Definition 2.1 (Weak ϵ-solution)**.** *Let $\epsilon > 0$, and recall that the feasible set $\mathcal{C}$ is defined in (2). A point $z \in \mathbb{R}^n$ is called a weak ϵ-solution of (3) if $\mathcal{G}(z) := \max_{x\in\mathcal{C}}\langle F(x), z - x\rangle \leq \epsilon$ and $g_i(z) \leq \epsilon$ for all $i \in [m]$.*

Useful lemmas. Next, we present two lemmas that encapsulate key properties of the velocity polytope $V_\alpha(x)$ defined in (4). These properties are implicitly used in [ZHM25] but not explicitly formulated as lemmas. For the sake of completeness, we provide their proofs herein.

Lemma 2.1. *For all $x \in \mathcal{C}$ and $\alpha > 0$, we have $\alpha(x - x^t) \in V_\alpha(x^t)$.*

Proof. Let $x \in \mathcal{C}$, i.e., $g_i(x) \leq 0$ for all $i \in [m]$. Then, for any $i \in I_{x^t}$, the convexity of g_i implies that $g_i(x^t) + \langle\nabla g_i(x^t), x - x^t\rangle \leq g_i(x) \leq 0$. Hence, the conclusion $\alpha(x - x^t) \in V_\alpha(x^t)$ follows from the definition of $V_\alpha(x^t)$ in (4) and $\alpha > 0$. □

Lemma 2.2. *For all $x \in \mathcal{C}$ and $v \in V_\alpha(x^t)$, we have $v + (x - x^t) \in V_\alpha(x^t)$.*

Proof. Let $x \in \mathcal{C}$, $v \in V_\alpha(x^t)$ and $i \in I_{x^t}$ be arbitrarily fixed. Then, $g_i(x) \le 0$ and $g_i(x^t) \ge 0$. Further considering the convexity of g_i, we obtain

$$\langle \nabla g_i(x^t), x - x^t \rangle \le g_i(x^t) + \langle \nabla g_i(x^t), x - x^t \rangle \le g_i(x) \le 0. \tag{8}$$

Since $v \in V_\alpha(x^t)$, we have $\alpha g_i(x^t) + \langle \nabla g_i(x^t), v \rangle \le 0$. Then, by combining this with (8) and (4), we obtain the desired result. □

3 CGM for Strongly Convex Optimization

In this section, we present an enhanced convergence analysis of CGM for solving the optimization problem (1), where f is μ-strongly convex and ℓ_f-smooth. Compared with the results in [ZHM25, Thm. 3], we remove the inappropriate requirement that f is L_f-Lipschitz and further relax other assumptions. Throughout this section, we make the following assumptions.

Assumption 1. *Assume that (i) f is μ-strongly convex and ℓ_f-smooth, (ii) all g_i's are convex and ℓ_g-smooth, and (iii) the feasible set $\mathcal{C}$ is nonempty. Here, $\ell_f \ge \mu > 0$ and $\ell_g > 0$ are some constants.*

Note that under Assumption 1, problem (1) has a unique optimal solution, which we denote by x^*, i.e., $x^* := \arg\min_{x \in \mathcal{C}} f(x)$. We further let $x^\star := \arg\min_{x \in \mathbb{R}^n} f(x)$ (the unique optimal solution of f over $\mathbb{R}^n$) and denote the corresponding function value by $f_\star := f(x^\star)$. The existence of $x^\star$ is also guaranteed by Assumption 1. Note that $f_\star$ is different from $f(x^*)$ and both are finite.

Remark 3.1. *As explained in Section 1, the strong convexity of f is inconsistent with its L_f-Lipschitz continuity. In our analysis, we only assume that f is strongly convex, but not L_f-Lipschitz. Furthermore, we neither impose an L_g-Lipschitz condition on each constraint function g_i nor require the feasible set $\mathcal{C}$ to be bounded. Note that all these conditions were assumed in [ZHM25]. Thus, compared with the results in [ZHM25], our analysis is more rigorous and general. In practice, the feasible set $\mathcal{C}$ can indeed be unbounded. Typical examples include linear inequality constraints $Ax \le b$ and exponential-decay bounds $e^{-x} \le \bar{C}$, which arise in control systems. Even when $\mathcal{C}$ is bounded, it is nontrivial to obtain its diameter in advance—note that this diameter is explicitly required for the algorithmic construction in [ZHM25].*

Our CGM for solving the minimization problem (1) (denoted CGM-Min) is presented in Algorithm 1. Note that in Algorithm 1, we define $I_x := \{i \in [m] \mid g_i(x) > 0\}$, which differs slightly from the definition in (4). Specifically, our I_x denotes the set of constraints violated at x, whereas that in (4) corresponds to active constraints. This change of definition stems from our refined analysis. Despite this minor difference, we retain the notation I_x for simplicity.

Remark 3.2. *We present the following remarks on Algorithm 1.*

1. *Unlike CGM in [ZHM25], we remove the auxiliary constraint $g_{m+1}(x) = \|x\|^2 - \widehat{D}^2 \le 0$. This constraint was originally used to ensure the boundedness of sequences $\{x^t\}$ and $\{v^t\}$ in their theoretical analysis. Additionally, we relax the assumptions on $\mathcal{C}$ by allowing it to be unbounded.*

2. *Leveraging our refined analysis, we relax the active constraint set in (4) to the violated constraint set $I_{x^t} = \{i \in [m] \mid g_i(x^t) > 0\}$. A beneficial byproduct of this relaxation is a reduction in gradient computations $\{\nabla g_i(x^t) : i \in I_{x^t}\}$, owing to the smaller cardinality of our I_{x^t}. Notably, this relaxation does not affect the conclusion of Lemma 2.2, as $\langle \nabla g_i(x^t), x - x^t \rangle \le 0$ remains valid. To see this, for all $i \in I_{x^t}$, since $g_i(x^t) > 0$, (8) becomes $\langle \nabla g_i(x^t), x - x^t \rangle < g_i(x^t) + \langle \nabla g_i(x^t), x - x^t \rangle \le g_i(x) \le 0$.*

3. *When $x^t \in \mathcal{C}$, $I_{x^t} = \varnothing$ holds, in which case $V_\alpha(x^t)$ reduces to the entire space $\mathbb{R}^n$.*

The core idea of our analysis is to leverage function values to bound the magnitude of the projected velocity $\|v^t\|$. The corresponding lemmas are presented as follows.

Algorithm 1 Constrained Gradient Method for Minimization Problem (CGM-Min)

1: Initialize $x^0 \in \mathcal{C}$, and set α such that $0 < \alpha \leq \mu$.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Construct the set of **violated constraints** at x^t as:

$$I_{x^t} = \{i \in [m] \mid g_i(x^t) > 0\}. \qquad \text{\# removed the auxiliary constraint.}$$

4: Construct the velocity polytope

$$V_\alpha(x^t) = \{v \in \mathbb{R}^n \mid \alpha g_i(x^t) + \nabla g_i(x^t)^\top v \leq 0,\ \forall i \in I_{x^t}\}. \tag{9}$$

5: Solve the quadratic programming problem

$$v^t = \arg\min_{v \in V_\alpha(x^t)} \|v + \nabla f(x^t)\|^2. \tag{10}$$

6: Choose $0 < \eta_t \leq \min\{1/\ell_f, 1/\alpha\}$ and update the iterate $x^{t+1} = x^t + \eta_t v^t$.
7: **end for**

Lemma 3.1 (Bounding $\|v^t\|^2$ via function values)**.** *Under Assumption 1 for Problem (1), Algorithm 1 satisfies, for all $t \geq 0$, the following:*

$$\|v^t\|^2 \leq 4(2\ell_f - \alpha)(f(x^t) - f(x^*)) + 8\ell_f(f(x^*) - f_\star). \tag{11}$$

Proof. Let $t \geq 0$ be arbitrarily fixed. Recall that the optimal function value of f over $\mathbb{R}^n$ is attained at $x^\star$, denoted $f_\star$. Given that f is convex and ℓ_f-smooth, we have

$$f(x^t) - f_\star \geq \frac{1}{2\ell_f}\|\nabla f(x^t)\|^2. \tag{12}$$

From (10), it follows that $\|v^t + \nabla f(x^t)\| \leq \|v + \nabla f(x^t)\|$ for all $v \in V_\alpha(x^t)$. Combining this with Lemma 2.1 and the fact that $x^* \in \mathcal{C}$, we obtain $\alpha(x^* - x^t) \in V_\alpha(x^t)$ and

$$\begin{aligned}
\|v^t + \nabla f(x^t)\|^2 &\leq \|\alpha(x^* - x^t) + \nabla f(x^t)\|^2 = \alpha^2\|x^* - x^t\|^2 + 2\alpha\langle x^* - x^t, \nabla f(x^t)\rangle + \|\nabla f(x^t)\|^2 \\
&\leq \alpha^2\|x^* - x^t\|^2 + 2\alpha\big(f(x^*) - f(x^t) - (\mu/2)\|x^t - x^*\|^2\big) + \|\nabla f(x^t)\|^2 \\
&= \alpha(\alpha - \mu)\|x^t - x^*\|^2 + 2\alpha\big(f(x^*) - f(x^t)\big) + \|\nabla f(x^t)\|^2 \\
&\leq 2\alpha(f(x^*) - f(x^t)) + 2\ell_f(f(x^t) - f_\star) \\
&= 2(\ell_f - \alpha)(f(x^t) - f(x^*)) + 2\ell_f(f(x^*) - f_\star),
\end{aligned} \tag{13}$$

where the second inequality follows from the μ-strong convexity of f, and the third from (12) and $\alpha \leq \mu$. Since $\|v^t\|^2 \leq 2\big(\|v^t + \nabla f(x^t)\|^2 + \|\nabla f(x^t)\|^2\big)$, combining (13) and (12) yields the desired result (11). □

To further bound $\|v^t\|$, the only nontrivial term in (11) is $f(x^t) - f(x^*)$. This can be effectively controlled by setting appropriate step sizes η_t. We now introduce the following lemma, which establishes the boundedness of $\|v^t\|^2$ and $\|x^t - x^*\|$.

Lemma 3.2 (Boundedness of $\|v^t\|$ and $\|x^t - x^*\|$)**.** *Under Assumption 1 for Problem (1), Algorithm 1 satisfies, for all $t \geq 0$, the following:*

$$\|v^t\| \leq C_1 := \sqrt{4(2\ell_f - \alpha)(f(x^0) - f(x^*)) + 8\ell_f(f(x^*) - f_\star)}, \tag{14}$$

$$\|x^t - x^*\| \leq C_2 := \Big(\|\nabla f(x^*)\| + \sqrt{\|\nabla f(x^*)\|^2 + 2\mu\big(f(x^0) - f(x^*)\big)}\Big)\Big/\mu. \tag{15}$$

Proof. Let $t \geq 0$ and $x \in \mathcal{C}$ be arbitrarily fixed. It follows from the update rule $x^{t+1} = x^t + \eta_t v^t$ that

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \eta_t \langle \nabla f(x^t), v^t \rangle + \frac{\ell_f}{2} \eta_t^2 \left\| v^t \right\|^2 \\ &= f(x^t) + \frac{\eta_t}{2} \left\| v^t + \nabla f(x^t) \right\|^2 - \frac{\eta_t}{2} \left\| \nabla f(x^t) \right\|^2 - \frac{\eta_t}{2} \left(1 - \eta_t \ell_f\right) \left\| v^t \right\|^2 \\ &\leq f(x^t) + \frac{\eta_t}{2} \left\| \alpha(x - x^t) + \nabla f(x^t) \right\|^2 - \frac{\eta_t}{2} \left\| \nabla f(x^t) \right\|^2 - \frac{\eta_t}{2} \left(1 - \eta_t \ell_f\right) \left\| v^t \right\|^2 \\ &= f(x^t) + \frac{\eta_t}{2} \alpha^2 \left\| x - x^t \right\|^2 + \alpha \eta_t \langle \nabla f(x^t), x - x^t \rangle - \frac{\eta_t}{2} \left(1 - \eta_t \ell_f\right) \left\| v^t \right\|^2 \\ &\leq f(x^t) - \alpha \eta_t (f(x^t) - f(x)) + \frac{\eta_t \alpha}{2} (\alpha - \mu) \left\| x - x^t \right\|^2 - \frac{\eta_t}{2} \left(1 - \eta_t \ell_f\right) \left\| v^t \right\|^2, \end{aligned} \tag{16}$$

where the first inequality follows from the ℓ_f-smoothness of f, the second from Lemma 2.1, and the third from $\langle \nabla f(x^t), x - x^t \rangle \leq f(x) - f(x^t) - (\mu/2)\|x - x^t\|^2$, which is due to the μ-strong convexity of f. Further considering $\alpha \leq \mu$ and $\eta_t \leq 1/\ell_f$, we obtain from (16) that

$$f(x^{t+1}) - f(x) \leq (1 - \alpha \eta_t)\big(f(x^t) - f(x)\big). \tag{17}$$

Since $0 < \alpha \eta_t \leq 1$, recursive application of (17) yields

$$f(x^{t+1}) - f(x) \leq \big(\Pi_{i=0}^t (1 - \alpha \eta_i)\big)\big(f(x^0) - f(x)\big) \leq f(x^0) - f(x). \tag{18}$$

Note that $\alpha \leq \mu \leq \ell_f$, so $2\ell_f - \alpha \geq \ell_f > 0$. Substituting $x = x^*$ into (18) and using (11), we derive the boundedness result for $\|v^t\|$ given in (14). Moreover, by again utilizing the μ-strong convexity of f and the Cauchy-Schwarz inequality, we have

$$f(x^t) - f(x^*) \geq \langle \nabla f(x^*), x^t - x^* \rangle + \frac{\mu}{2} \|x^t - x^*\|^2 \geq -\|\nabla f(x^*)\| \|x^t - x^*\| + \frac{\mu}{2} \|x^t - x^*\|^2.$$

Further considering (18), this implies $\frac{\mu}{2}\|x^t - x^*\|^2 - \|\nabla f(x^*)\|\|x^t - x^*\| \leq f(x^0) - f(x^*)$, from which the desired result (15) follows. □

Lemma 3.2 implies that the sequence $\{x^t\}$ generated by Algorithm 1 is contained in the bounded region $B(x^*, C_2)$, a closed ball centered at x^* with radius $C_2 > 0$ defined in (15). This boundedness guarantee allows us to remove the assumption that all g_i's are L_g-Lipschitz continuous. With a slight abuse of notation, we define L_g hereafter as

$$L_g := \max\{\|\nabla g_i(x)\| : x \in B(x^*, C_2),\ i \in [m]\}. \tag{19}$$

We are now ready to present the main results in this section.

Theorem 3.1. *Consider Problem (1) under Assumption 1. Let $\kappa := \ell_f/\mu \geq 1$ denote the condition number of f. Recall that the optimality gap $\mathcal{G}(\cdot)$ is defined in Definition 2.1 and the constant C_1 is given in (14). Then, for Algorithm 1 with $\alpha = \mu$, the following convergence results hold:*

1. *Constant step size: Let $T \geq 1$ be any integer satisfying $T \geq \kappa \log T$ and $\eta_t = \eta := \log T/(\mu T)$ for all $t = 0, 1, \ldots, T-1$. Then, Algorithm 1 achieves*

   (Function Value Residual) $\quad f(x^T) - f(x^*) \leq (f(x^0) - f(x^*))/T \sim \mathcal{O}(1/T)$,

   (Optimality Gap) $\quad \mathcal{G}(x^T) \leq (f(x^0) - f(x^*))/T \sim \mathcal{O}(1/T)$,

   (Feasibility) $\quad g_i(x^T) \leq \dfrac{C_1}{\mu} \max\left\{\dfrac{C_1 \ell_g}{2\mu}, L_g\right\} \dfrac{\log T}{T} \sim \tilde{\mathcal{O}}(1/T),\ \forall i \in [m]$.

2. *Varying step size: Let $\eta_t = 1/(\mu(t + \kappa))$ for any $t \geq 0$. Then, for all $t \geq 1$, Algorithm 1 achieves*

   (Function Value Residual) $\quad f(x^t) - f(x^*) \leq \dfrac{\kappa - 1}{t + \kappa - 1}\big(f(x^0) - f(x^*)\big) \sim \mathcal{O}(1/t)$,

   (Optimality Gap) $\quad \mathcal{G}(x^t) \leq \dfrac{\kappa - 1}{t + \kappa - 1}\big(f(x^0) - f(x^*)\big) \sim \mathcal{O}(1/t)$,

   (Feasibility) $\quad g_i(x^{t+1}) \leq \dfrac{2C_1}{\mu(t + \kappa + 1)}\left(L_g + \dfrac{\ell_g C_1}{2\mu}\right) + \dfrac{\ell_g C_1^2 \log t}{\mu^2(t + \kappa + 1)} \sim \tilde{\mathcal{O}}(1/t),\ \forall i \in [m]$.

Proof. First, it is straightforward to verify that $0 < \eta_t \le \min\{1/\ell_f, 1/\alpha\}$ holds for both choices of η_t. Next, we prove the convergence rate results for the function value residual, followed by those for feasibility. Let $T \ge 1$ be any integer satisfying $T \ge \kappa \log T$ and $x \in \mathcal{C}$ be arbitrarily fixed. For the function value residual, using the constant step size $\eta_t = \eta = \log T/(\mu T)$ for all $t = 0, 1, \ldots, T-1$, we have

$$\begin{aligned}\langle \nabla f(x), x^T - x\rangle &\le f(x^T) - f(x) \le f(x^T) - f(x^*) \le (1-\mu\eta)^T \left(f(x^0) - f(x^*)\right)\\ &\le \mathrm{e}^{-\mu\eta T}\left(f(x^0) - f(x^*)\right) = (f(x^0) - f(x^*))/T,\end{aligned}$$

where the first "$\le$" follows from the convexity of f, the second is because $x \in \mathcal{C}$ and $f(x) \ge f(x^*)$, the third follows from $\alpha = \mu$ and (17) with $x = x^*$. Since $x \in \mathcal{C}$ is arbitrarily taken, the above inequality further implies

$$\mathcal{G}(x^T) = \max_{x\in\mathcal{C}} \langle \nabla f(x), x^T - x\rangle \le (f(x^0) - f(x^*))/T.$$

When using the varying step size $\eta_t = 1/(\mu(t+\kappa))$, similarly, let $x \in \mathcal{C}$ and $t \ge 1$ be arbitrarily fixed yields

$$\begin{aligned}\langle \nabla f(x), x^t - x\rangle &\le f(x^t) - f(x) \le \left(\Pi_{i=0}^{t-1}(1-\mu\eta_i)\right)\left(f(x^0) - f(x)\right)\\ &\le \left(\Pi_{i=0}^{t-1}(1-\mu\eta_i)\right)\left(f(x^0) - f(x^*)\right) = \frac{\kappa-1}{t+\kappa-1}\left(f(x^0) - f(x^*)\right),\end{aligned}$$

where the first "$\le$" follows from the convexity of f, the second "$\le$" follows from the first inequality in (18), the third "$\le$" is because $x \in \mathcal{C}$ and $f(x) \ge f(x^*)$, and the "$=$" is because $\eta_i = 1/(\mu(i+\kappa))$ for $i \ge 0$. Letting $x = x^*$ in the above inequality, we obtain

$$f(x^t) - f(x^*) \le \frac{\kappa-1}{t+\kappa-1}\left(f(x^0) - f(x^*)\right), \quad \forall t \ge 1.$$

On the other hand, since $x \in \mathcal{C}$ is arbitrarily taken, we obtain

$$\mathcal{G}(x^t) = \max_{x\in\mathcal{C}} \langle \nabla f(x), x^t - x\rangle \le \frac{\kappa-1}{t+\kappa-1}\left(f(x^0) - f(x^*)\right), \quad \forall t \ge 1.$$

Next, we turn to the convergence rates of the feasibility. When using the constant step size $\eta_t = \eta = \log T/(\mu T)$ for all $t = 0, 1, \ldots, T-1$, we will prove that

$$g_i(x^t) \le \frac{C_1}{\mu}\max\left\{\frac{C_1\ell_g}{2\mu}, L_g\right\}\frac{\log T}{T} \quad \text{for all} \quad t = 0, 1, \ldots, T-1 \quad \text{and} \quad i \in [m], \tag{20}$$

where C_1 is defined in (14). To this end, we consider the following two cases.

(a) If $i \notin I_{x^t}$, then $g_i(x^t) \le 0$. Following from the convexity of g_i, we have

$$g_i(x^{t+1}) \le g_i(x^t) + \eta \nabla g_i(x^{t+1})^\top v^t \le \eta \|\nabla g_i(x^{t+1})\| \|v^t\| \le \frac{C_1 L_g \log T}{\mu T},$$

where the last inequality follows from (14) and (19).

(b) If $i \in I_{x^t}$, then $g_i(x^t) > 0$. From (9), we have $\eta\alpha g_i(x^t) + \nabla g_i(x^t)^\top \left(x^{t+1} - x^t\right) \le 0$. Following from the ℓ_g-smoothness of g_i, we have

$$g_i(x^{t+1}) \le (1-\alpha\eta)\, g_i(x^t) + \frac{\ell_g}{2}\eta^2\|v^t\|^2 \le \left(1 - \frac{\log T}{T}\right) g_i(x^t) + \frac{\ell_g C_1^2 (\log T)^2}{2\mu^2 T^2}. \tag{21}$$

Now we prove (20) by induction. Since $x^0 \in \mathcal{C}$, we have $g_i(x^0) \le 0$ for all $i \in [m]$. Thus, (20) holds for $t = 0$. Assume that (20) holds for some $t \ge 0$. For $(t+1)$, in case (a), we have $g_i(x^{t+1}) \le \frac{C_1 L_g \log T}{\mu T}$, and in case (b), we have

$$\begin{aligned} g_i(x^{t+1}) &\le \left(1 - \frac{\log T}{T}\right) g_i(x^t) + \frac{\ell_g C_1^2 (\log T)^2}{2\mu^2 T^2}\\ &\le \left(1 - \frac{\log T}{T}\right)\frac{C_1}{\mu}\max\left\{\frac{C_1\ell_g}{2\mu}, L_g\right\}\frac{\log T}{T} + \frac{\ell_g C_1^2 (\log T)^2}{2\mu^2 T^2}\\ &\le \frac{C_1}{\mu}\max\left\{\frac{C_1\ell_g}{2\mu}, L_g\right\}\frac{\log T}{T},\end{aligned}$$

where the first "$\leq$" follows from (21), and the second is because the induction hypothesis. This proves (20), and thus the desired result $g_i(x^T) \leq \frac{C_1}{\mu}\max\left\{\frac{C_1\ell_g}{2\mu}, L_g\right\}\frac{\log T}{T} \sim \tilde{\mathcal{O}}(1/T)$, $\forall i \in [m]$.

When using the varying step size $\eta_t = 1/(\mu(t+\kappa))$, we analogously consider the following two cases.

(a) If $i \notin I_{x^t}$, then $g_i(x^t) \leq 0$. Following from the convexity of $g_i(x)$, we have

$$\begin{aligned} g_i(x^{t+1}) &\leq g_i(x^t) + \nabla g_i(x^{t+1})^\top (x^{t+1} - x^t) \leq \eta_t \nabla g_i(x^{t+1})^\top v^t \\ &\leq \eta_t \|\nabla g_i(x^{t+1})\| \|v^t\| \leq \frac{C_1 L_g}{\mu(t+\kappa)}, \quad \forall i \notin I_{x^t}. \end{aligned} \tag{22}$$

(b) If $i \in I_{x^t}$, then $g_i(x^t) > 0$. Since $x^{t+1} = x^t + \eta_t v^t$ and $\alpha = \mu$, it follows from (9) that $\eta_t \mu g_i(x^t) + \langle \nabla g_i(x^t), x^{t+1} - x^t\rangle \leq 0$. Further considering the ℓ_g-smoothness of g_i, we obtain

$$g_i(x^{t+1}) \leq (1 - \mu\eta_t)\, g_i(x^t) + \frac{\ell_g}{2}\eta_t^2 \|v^t\|^2 = \left(1 - \frac{1}{t+\kappa}\right) g_i(x^t) + \frac{\ell_g C_1^2}{2\mu^2(t+\kappa)^2}.$$

Multiplying both sides by $(t+\kappa+1)$ and noting $(t+\kappa+1) \leq 2(t+\kappa)$, we obtain

$$(t+\kappa+1) g_i(x^{t+1}) \leq \left(1 - \frac{1}{t+\kappa}\right)(t+\kappa+1) g_i(x^t) + \frac{\ell_g C_1^2}{\mu^2(t+\kappa)} \leq (t+\kappa) g_i(x^t) + \frac{\ell_g C_1^2}{\mu^2(t+\kappa)}. \tag{23}$$

Define

$$\theta_i(t) = \max\{s \in \{0, 1, 2, \ldots, t-1\} \mid g_i(x^s) \leq 0\}. \tag{24}$$

Since $x^0 \in \mathcal{C}$, it follows that $\theta_i(t) \geq 0$ for every $t \geq 1$. Moreover, since $g_i(x^t) > 0$, we thus have $0 \leq \theta_i(t+1) < t$, $g_i(x^{\theta_i(t+1)}) \leq 0$ and $g_i(x^{\theta_i(t+1)+w}) > 0$ holds for $w = 1, \ldots, t - \theta_i(t+1)$. In other words, we have $i \notin I_{x^{\theta_i(t+1)}}$ and $i \in I_{x^s}$ for any $s = \theta_i(t+1)+1, \ldots, t$. Note that (23) holds for any t as long as $i \in I_{x^t}$. Therefore, we have

$$(s+\kappa+1) g_i(x^{s+1}) - (s+\kappa) g_i(x^s) \leq \frac{\ell_g C_1^2}{\mu^2(s+\kappa)}, \quad s = \theta_i(t+1)+1, \ldots, t.$$

Summing the above inequality from $s = \theta_i(t+1)+1$ to t yields

$$\begin{aligned} (t+\kappa+1) g_i(x^{t+1}) &\leq (\theta_i(t+1)+1+\kappa) g_i(x^{\theta_i(t+1)+1}) + \sum_{s=\theta_i(t+1)+1}^{t} \frac{\ell_g C_1^2}{\mu^2(s+\kappa)} \\ &\leq \frac{C_1 L_g(\theta_i(t+1)+1+\kappa)}{\mu(\theta_i(t+1)+\kappa)} + \frac{\ell_g C_1^2}{\mu^2}\sum_{s=1}^{t}\frac{1}{s} \\ &\leq \frac{2C_1 L_g}{\mu} + \frac{\ell_g C_1^2}{\mu^2}(1+\log t) = \frac{2C_1}{\mu}\left(L_g + \frac{\ell_g C_1}{2\mu}\right) + \frac{\ell_g C_1^2 \log t}{\mu^2}, \end{aligned}$$

where the second inequality uses (22), since we have $i \notin I_{x^{\theta_i(t+1)}}$. Consequently, we obtain

$$g_i(x^{t+1}) \leq \frac{2C_1}{\mu(t+\kappa+1)}\left(L_g + \frac{\ell_g C_1}{2\mu}\right) + \frac{\ell_g C_1^2 \log t}{\mu^2(t+\kappa+1)}, \quad \forall i \in I_{x^t}. \tag{25}$$

Since the term on the far right-hand side of (22) is smaller than that of (25), we conclude that (25) holds for all $i \in [m]$. This completes the proof of the theorem. □

4 CGM for Strongly Monotone VIs

This section focuses on addressing the VI problem (3), for which we introduce the following assumption.

Assumption 2. *The operator $F : \mathbb{R}^n \to \mathbb{R}^n$ in (3) is continuous and μ-strongly monotone over $\mathbb{R}^n$, and it satisfies the relaxed Lipschitz condition:*

$$\|F(x) - F(y)\|^2 \le \ell_F^2 \|x - y\|^2 + B, \quad \forall x, y \in \mathbb{R}^n, \tag{26}$$

where $\ell_F \ge \mu$ and $B \ge 0$ constants. Furthermore, all g_i's are convex and ℓ_g-smooth over $\mathbb{R}^n$. The constraint set $\mathcal{C}$ in (3) is bounded, with its diameter defined as $D := \sup_{x,y\in\mathcal{C}} \|x - y\| < \infty$.

Rather than assuming F is L_F-bounded over the entire space, we instead adopt the more relaxed condition given in (26), for which we provide the following remarks.

(i) The use of weakened conditions traces back to stochastic gradient descent methods, where boundedness conditions such as $\mathbb{E}_\xi\|\nabla f(x;\xi)\|^2 \le B$ or $\mathbb{E}_\xi\|s(x;\xi)\|^2 \le B$ are typically adopted, where $s(x;\xi)$ is an unbiased estimator of a subgradient. For the smooth case, Blum and Gladyshev [Blu54, Gla65] proposed a relaxed condition $\mathbb{E}_\xi\|\nabla f(x;\xi)\|^2 \le L^2\|x\|^2 + B$, which further implies the existence of $\hat{L}, \hat{B} > 0$ such that $\mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla f(x^\star)\|^2 \le \hat{L}^2\|x - x^\star\|^2 + \hat{B}$—a form analogous to that of (26). For the nonsmooth case, [Gri19] assumed $\mathbb{E}_\xi\|s(x;\xi)\|^2 \le B + L\,(f(x) - f_\star)$.

(ii) In our case, condition (26) is different from [BDK24, Eq. (1.3)] and [JNT11, Eq. (4)]: the latter control $\|F(x) - F(y)\|$ (rather than its square) and only requires satisfaction for $x, y \in \mathcal{X}$, where $\mathcal{X}$ is a simple convex compact set admitting easy projection. Additional forms of relaxed conditions are available in [KR23, AMW25] and related works.

(iii) When $B = 0$, condition (26) reduces to the ℓ_F-Lipschitz continuity of F over $\mathbb{R}^n$. Thus, for $B > 0$, this assumption is less restrictive than requiring F to be both ℓ_F-Lipschitz and L_F-bounded. Furthermore, it is consistent with the strong monotonicity of F when $\ell_F \ge \mu$.

We retain the notation $\kappa := \ell_F/\mu \ge 1$ in this section. Our CGM for solving the strongly monotone VI problem (3) is presented in Algorithm 2.

Algorithm 2 Constrained Gradient Method for strongly monotone VIs (CGM-VI)

1: Initialize $x^0 \in \mathcal{C}$, set parameters: $\alpha = \mu$, $\eta_t = \frac{1}{\mu(t+16\kappa^2)}$ for all $t \ge 0$ and $\Delta \ge \max\left\{1, \frac{D^2\ell_F^2}{\|F(x^0)\|^2+B}\right\}$.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Construct the set of **violated constraints** as $I_{x^t} = \{i \in [m+1] \mid g_i(x^t) > 0\}$, where for $i = m+1$, the auxiliary constraint function is defined as $g_{m+1}(x) = \|x - x^0\|^2 - \frac{\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B)$.
4: Construct the velocity polytope

$$V_\alpha(x^t) = \{v \in \mathbb{R}^n \mid \alpha g_i(x^t) + \nabla g_i(x^t)^\top v \le 0,\ \forall i \in I_{x^t}\}.$$

5: Solve the quadratic program

$$v^t = \arg\min_{v \in V_\alpha(x^t)} \|v + F(x^t)\|^2. \tag{27}$$

6: Update the iterate $x^{t+1} = x^t + \eta_t v^t$.
7: **end for**

Remark 4.1. *Given the choice of $\Delta \ge \max\left\{1, \frac{D^2\ell_F^2}{\|F(x^0)\|^2+B}\right\}$ and the definition of the auxiliary constraint function g_{m+1}, we can deduce that $g_{m+1}(x) \le 0$ holds for all $x \in \mathcal{C}$. In the extreme case where $\|F(x^0)\| = 0$ and $B = 0$, a zero denominator can be avoided by selecting a different initial point.*

Next, we present a key lemma that underpins our main convergence results.

Lemma 4.1 (Boundedness of $\|x^t - x^0\|$ and $\|v^t\|$)**.** *Under Assumption 2 for Problem (3), Algorithm 2 satisfies, for all $t \ge 0$, the following:*

$$\|x^t - x^0\| \le C_3 := \sqrt{\big(2\Delta + 5/4\big)(\|F(x^0)\|^2 + B)/\ell_F^2}, \tag{28a}$$

$$\|v^t\| \le C_4 := \sqrt{(16\Delta + 20)(\|F(x^0)\|^2 + B)}. \tag{28b}$$

Proof. For any $x \in \mathcal{C}$, by Lemma 2.1 and the definition of v^t, we have

$$\begin{aligned}\big\|v^t + F(x^t)\big\|^2 &\le \big\|\alpha(x - x^t) + F(x^t)\big\|^2 = \alpha^2\|x - x^t\|^2 + 2\alpha\langle x - x^t, F(x^t)\rangle + \|F(x^t)\|^2\\ &= \alpha^2\|x - x^t\|^2 + 2\alpha\langle x - x^t, F(x^t) - F(x)\rangle + 2\alpha\langle x - x^t, F(x)\rangle + \|F(x^t)\|^2\\ &\le (\alpha^2 - 2\mu\alpha)\|x - x^t\|^2 + 2\alpha\langle x - x^t, F(x)\rangle + \|F(x^t)\|^2\\ &\le -\mu^2\|x - x^t\|^2 + 2\mu\Big(\frac{1}{2\mu}\|F(x)\|^2 + \frac{\mu}{2}\|x - x^t\|^2\Big) + \|F(x^t)\|^2\\ &= \|F(x)\|^2 + \|F(x^t)\|^2, \end{aligned} \tag{29}$$

where the second inequality follows from the μ-strong monotonicity of F, and the third inequality follows from Young's inequality and the fact that $\alpha = \mu$. Using the inequality $\|a\|^2 \le 2(\|a - b\|^2 + \|b\|^2)$ (which holds for any vectors a and b of the same dimension), we further derive

$$\begin{aligned}\big\|v^t\big\|^2 &\le 2\|v^t + F(x^t)\|^2 + 2\|F(x^t)\|^2 \overset{(29)}{\le} 2\|F(x)\|^2 + 4\|F(x^t)\|^2\\ &\le 10\|F(x)\|^2 + 8\|F(x^t) - F(x)\|^2 \overset{(26)}{\le} 10\|F(x)\|^2 + 8\ell_F^2\|x^t - x\|^2 + 8B\\ &\le 8\ell_F^2\|x^t - x\|^2 + 10(\|F(x)\|^2 + B).\end{aligned}$$

Since the above inequality holds for any $x \in \mathcal{C}$ and $x^0 \in \mathcal{C}$, by letting $x = x^0$ we derive

$$\big\|v^t\big\|^2 \le 8\ell_F^2\|x^t - x^0\|^2 + 10(\|F(x^0)\|^2 + B). \tag{30}$$

Since $\eta_t = \frac{1}{\mu(t+16\kappa^2)}$ for all $t \ge 0$ and $\kappa = \ell_F/\mu$, we have

$$\eta_t \le \mu/(16\ell_F^2) \quad \text{for all} \quad t \ge 0. \tag{31}$$

We now prove (28a) by induction. First, (28a) clearly holds for $t = 0$. Assuming it holds for some $t = k \ge 0$, we will then show that it also holds for $t = k + 1$. Consider the following two cases:

(i) $\|x^k - x^0\|^2 \le \frac{\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B)$, i.e., $g_{m+1}(x^k) \le 0$;

(ii) $\frac{\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B) < \|x^k - x^0\|^2 \le \big(2\Delta + 5/4\big)(\|F(x^0)\|^2 + B)/\ell_F^2$.

Using the inequality $\|a + b\|^2 \le \frac{4}{3}\|a\|^2 + 4\|b\|^2$ and the update rule $x^{k+1} = x^k + \eta_k v^k$, in case (i), we have

$$\begin{aligned}\|x^{k+1} - x^0\|^2 &\le \frac{4}{3}\|x^k - x^0\|^2 + 4\eta_k^2\|v^k\|^2 \overset{(30)}{\le} \Big(\frac{4}{3} + 32\ell_F^2\eta_k^2\Big)\|x^k - x^0\|^2 + 40\eta_k^2(\|F(x^0)\|^2 + B)\\ &\overset{(31),\,(i)}{\le} \Big(\frac{4}{3} + \frac{\mu^2}{8\ell_F^2}\Big)\frac{\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B) + \frac{5\mu^2}{32\ell_F^4}(\|F(x^0)\|^2 + B)\\ &\le \Big(\frac{4}{3} + \frac{1}{8} + \frac{5}{32}\Big)\frac{\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B) \le \frac{2\Delta}{\ell_F^2}(\|F(x^0)\|^2 + B),\end{aligned} \tag{32}$$

where the second-to-last inequality follows from $\kappa \ge 1$ and $\Delta \ge 1$. On the other hand, in case (ii), we have $g_{m+1}(x^k) = \|x^k - x^0\|^2 - \Delta(\|F(x^0)\|^2 + B)/\ell_F^2 > 0$, i.e., $(m+1) \in I_{x^k}$. By the definitions of g_{m+1}, $V_\alpha(x^k)$ and $v^k \in V_\alpha(x^k)$, we have $\mu\big(\|x^k - x^0\|^2 - \Delta(\|F(x^0)\|^2 + B)/\ell_F^2\big) + 2(x^k - x^0)^\top v^k \le 0$. Using the updating rule $x^{k+1} = x^k + \eta_k v^k$ again, we derive

$$\begin{aligned}\|x^{k+1} - x^0\|^2 &= \eta_k^2\|v^k\|^2 + \|x^k - x^0\|^2 + 2\eta_k(x^k - x^0)^\top v^k\\ &\le \eta_k^2\|v^k\|^2 + \|x^k - x^0\|^2 - \mu\eta_k\big(\|x^k - x^0\|^2 - \Delta(\|F(x^0)\|^2 + B)/\ell_F^2\big)\\ &\overset{(30)}{\le} \eta_k^2\big(8\ell_F^2\|x^k - x^0\|^2 + 10(\|F(x^0)\|^2 + B)\big) + (1 - \mu\eta_k)\|x^k - x^0\|^2 + \mu\eta_k\Delta(\|F(x^0)\|^2 + B)/\ell_F^2\\ &\overset{(31)}{\le} \big(1 - \mu\eta_k/2\big)\|x^k - x^0\|^2 + 10\eta_k^2(\|F(x^0)\|^2 + B) + \mu\eta_k\Delta(\|F(x^0)\|^2 + B)/\ell_F^2\\ &\overset{(ii)}{\le} \Big(\big(1 - \mu\eta_k/2\big)\big(2\Delta + 5/4\big) + \mu\eta_k\Delta\Big)(\|F(x^0)\|^2 + B)/\ell_F^2 + 10\eta_k^2(\|F(x^0)\|^2 + B)\\ &\le \big(2\Delta + 5/4\big)(\|F(x^0)\|^2 + B)/\ell_F^2,\end{aligned} \tag{33}$$

where the last inequality uses $10\eta_k^2 \le 5\mu\eta_k/(8\ell_F^2)$. Combining (32) and (33), we thus conclude that (28a) holds in all cases. Finally, (28b) follows immediately from combining (28a) and (30). □

Now we are ready to present our main result.

Theorem 4.1. *Consider Problem (3) under Assumption 2. Let $\{x^t\}$ be the sequence generated by Algorithm 2. Let $T \ge 1$ be any integer and define $\bar{x}^T = \sum_{t=0}^{T-1}(t+16\kappa^2-1)x^t / \sum_{t=0}^{T-1}(t+16\kappa^2-1)$. We then have the following convergence results:*

$$\mathcal{G}(\bar{x}^T) \le \frac{2\mu D^2(8\kappa^2-1)(16\kappa^2-1)}{T(T+32\kappa^2-3)} + \frac{(16\Delta+20)(\|F(x^0)\|^2+B)}{\mu(T+32\kappa^2-3)} \sim \mathcal{O}(1/T), \tag{34}$$

$$g_i(\bar{x}^T) \le \frac{4C_4}{\mu(T+32\kappa^2-3)}\Big(L_g + \frac{\ell_g C_4}{2\mu}\Big) + \frac{2\ell_g C_4^2 \log T}{\mu^2(T+32\kappa^2-3)} \sim \tilde{\mathcal{O}}(1/T), \quad \forall i \in [m+1]. \tag{35}$$

Here, the optimality gap $\mathcal{G}(\cdot)$ is defined in Definition 2.1, $L_g := \max\{\|\nabla g_i(x)\| : x \in B(x^0, C_3), i \in [m+1]\}$, C_3 and C_4 are defined in (28), and $\kappa = \ell_F/\mu$. Additionally, for the constraint violation, we also have the following non-ergodic convergence rate for all $t \ge 1$:

$$g_i(x^{t+1}) \le \frac{2C_4}{\mu(t+16\kappa^2+1)}\left(L_g + \frac{\ell_g C_4}{2\mu}\right) + \frac{\ell_g C_4^2 \log t}{\mu^2(t+16\kappa^2+1)} \sim \tilde{\mathcal{O}}(1/t), \quad \forall i \in [m+1]. \tag{36}$$

Proof. Using the optimality conditions of (27), we obtain $\langle v^t + F(x^t), v^t - v\rangle \le 0$ for any $v \in V_\alpha(x^t)$. Let $x \in \mathcal{C}$ be arbitrarily fixed. Then, we have $v = v^t + x - x^t \in V_\alpha(x^t)$ from Lemma 2.2. Hence, we have $\langle v^t + F(x^t), x^t - x\rangle \le 0$, which further implies

$$\langle F(x^t), x^t - x\rangle \le \langle v^t, x - x^t\rangle = \frac{1}{\eta_t}\langle x^{t+1} - x^t, x - x^t\rangle = \frac{1}{2\eta_t}\big(\|x - x^t\|^2 - \|x - x^{t+1}\|^2\big) + \frac{\eta_t}{2}\|v^t\|^2. \tag{37}$$

Moreover, since $F(x)$ is μ-strongly convex, we have

$$\begin{aligned}
\langle F(x), x^t - x\rangle &\le \langle F(x^t), x^t - x\rangle - \mu\left\|x^t - x\right\|^2 \\
&\overset{(37)}{\le} \left(\frac{1}{2\eta_t} - \mu\right)\left\|x - x^t\right\|^2 - \frac{1}{2\eta_t}\left\|x - x^{t+1}\right\|^2 + \frac{\eta_t}{2}\left\|v^t\right\|^2 \\
&\overset{(28b)}{\le} \frac{\mu(t+16\kappa^2-2)}{2}\left\|x - x^t\right\|^2 - \frac{\mu(t+16\kappa^2)}{2}\left\|x - x^{t+1}\right\|^2 + \frac{C_4^2}{2\mu(t+16\kappa^2)},
\end{aligned}$$

where the last inequality also uses $\eta_t = \frac{1}{\mu(t+16\kappa^2)}$ for $t \ge 0$. For simplicity, we let $C_t' := t + 16\kappa^2 - 1$ for $t \ge 0$. Multiplying both sides of the above inequality by C_t', we obtain

$$C_t'\langle F(x), x^t - x\rangle \le \frac{\mu}{2}\Big((C_t'-1)C_t'\left\|x - x^t\right\|^2 - C_t'(C_t'+1)\left\|x - x^{t+1}\right\|^2\Big) + \frac{C_4^2}{2\mu}.$$

Summing this inequality for $t = 0, \dots, T-1$, and using the definition of $\bar{x}^T$, we derive

$$\begin{aligned}
\langle F(x), \bar{x}^T - x\rangle &= \frac{2}{T(T+32\kappa^2-3)}\sum_{t=0}^{T-1} C_t'\langle F(x), x^t - x\rangle \\
&\le \frac{2}{T(T+32\kappa^2-3)}\Big(\frac{\mu}{2}(C_0'-1)C_0'\left\|x - x^0\right\|^2 + \frac{TC_4^2}{2\mu}\Big) \\
&= \frac{2\mu(8\kappa^2-1)(16\kappa^2-1)}{T(T+32\kappa^2-3)}\|x - x^0\|^2 + \frac{(16\Delta+20)(\|F(x^0)\|^2+B)}{\mu(T+32\kappa^2-3)},
\end{aligned}$$

where the last equality follows from direct computation and the definition of C_4 in (28b). Then, the desired result (34) follows immediately from the definition of D in Assumption 2 and the arbitrariness of $x \in \mathcal{C}$.

To prove (35) and (36), we consider the following two cases:

(a) If $i \notin I_{x^t}$, then $g_i(x^t) \le 0$. Using the convexity of g_i, the definition of L_g, and (28b), we obtain

$$\begin{aligned} g_i(x^{t+1}) &\le g_i(x^t) + \nabla g_i(x^{t+1})^\top (x^{t+1} - x^t) \\ &\le \eta_t \nabla g_i(x^{t+1})^\top v^t \le \eta_t \|\nabla g_i(x^{t+1})\| \|v^t\| \le \frac{C_4 L_g}{\mu(t+16\kappa^2)}, \quad \forall i \notin I_{x^t}. \end{aligned} \tag{38}$$

(b) If $i \in I_{x^t}$, then we have $g_i(x^t) > 0$, and $\eta_t \mu g_i(x^t) + \langle \nabla g_i(x^t), x^{t+1} - x^t \rangle \le 0$. Since g_i is ℓ_g-smooth, we have

$$\begin{aligned} g_i(x^{t+1}) &\le g_i(x^t) + \langle \nabla g_i(x^t), x^{t+1} - x^t \rangle + \frac{\ell_g}{2} \|x^{t+1} - x^t\|^2 \\ &\le (1 - \mu\eta_t)\, g_i(x^t) + \frac{\ell_g}{2} \eta_t^2 \|v^t\|^2 \le \left(1 - \frac{1}{t+16\kappa^2}\right) g_i(x^t) + \frac{\ell_g C_4^2}{2\mu^2 (t+16\kappa^2)^2}. \end{aligned}$$

Multiplying both sides by $(t+16\kappa^2+1)$, we derive

$$\begin{aligned} (t+16\kappa^2+1) g_i(x^{t+1}) &\le \left(1 - \frac{1}{t+16\kappa^2}\right)(t+16\kappa^2+1) g_i(x^t) + \frac{\ell_g C_4^2}{\mu^2 (t+16\kappa^2)} \\ &\le (t+16\kappa^2) g_i(x^t) + \frac{\ell_g C_4^2}{\mu^2 (t+16\kappa^2)}, \end{aligned}$$

which further implies

$$(t+16\kappa^2+1) g_i(x^{t+1}) - (t+16\kappa^2) g_i(x^t) \le \frac{\ell_g C_4^2}{\mu^2 (t+16\kappa^2)}. \tag{39}$$

Recall the definition of $\theta_i(t)$ in (24). Since $g_i(x^t) > 0$, we have $0 \le \theta_i(t+1) < t$, $g_i(\theta_i(t+1)) \le 0$ and $g_i(\theta_i(t+1) + w) > 0$ holds for $w = 1, \ldots, t - \theta_i(t+1)$. In other words, we have $i \notin I_{x^{\theta_i(t+1)}}$ and $i \in I_{x^s}$ for any $s = \theta_i(t+1)+1, \ldots, t$. Note that (39) holds for any t such that $i \in I_{x^t}$. Therefore, the following inequalities hold:

$$(s+16\kappa^2+1) g_i(x^{s+1}) - (s+16\kappa^2) g_i(x^s) \le \frac{\ell_g C_4^2}{\mu^2 (s+16\kappa^2)}, \quad s = \theta_i(t+1)+1, \ldots, t.$$

Summing the above inequality for $s = \theta_i(t+1)+1, \ldots, t$ yields

$$\begin{aligned} (t+16\kappa^2+1) g_i(x^{t+1}) &\le (\theta_i(t+1)+1+16\kappa^2) g_i(x^{\theta_i(t+1)+1}) + \sum_{s=\theta_i(t+1)+1}^{t} \frac{\ell_g C_4^2}{\mu^2 (s+16\kappa^2)} \\ &\le \frac{C_4 L_g (\theta_i(t+1)+1+16\kappa^2)}{\mu(\theta_i(t+1)+16\kappa^2)} + \frac{\ell_g C_4^2}{\mu^2} \sum_{s=1}^{t} \frac{1}{s} \\ &\le \frac{2C_4 L_g}{\mu} + \frac{\ell_g C_4^2}{\mu^2}(1 + \log t) \le \frac{2C_4}{\mu}\left(L_g + \frac{\ell_g C_4}{2\mu}\right) + \frac{\ell_g C_4^2 \log t}{\mu^2}, \end{aligned}$$

where the second inequality follows from (38), since $i \notin I_{x^{\theta_i(t+1)}}$. Consequently, we obtain

$$g_i(x^{t+1}) \le \frac{2C_4}{\mu(t+16\kappa^2+1)}\left(L_g + \frac{\ell_g C_4}{2\mu}\right) + \frac{\ell_g C_4^2 \log t}{\mu^2 (t+16\kappa^2+1)}, \quad \forall i \in I_{x^t}. \tag{40}$$

Since the term on the far right-hand side of (38) is smaller than that of (40), we conclude that (36) holds for all $i \in [m+1]$. Finally, by the definition of $\bar{x}^T$, leveraging the convexity of g_i, and using (36), we can derive the following via simple calculations and appropriate inequalities:

$$\begin{aligned} g_i(\bar{x}^T) &\le \frac{2}{T(T+32\kappa^2-3)} \sum\nolimits_{t=0}^{T-1} (t+16\kappa^2-1) g_i(x^t) \\ &\le \frac{2}{T(T+32\kappa^2-3)} \sum\nolimits_{t=0}^{T-1} \left(\frac{2C_4}{\mu}\left(L_g + \frac{\ell_g C_4}{2\mu}\right) + \frac{\ell_g C_4^2 \log(t+1)}{\mu^2}\right) \\ &\le \frac{4C_4}{\mu(T+32\kappa^2-3)}\left(L_g + \frac{\ell_g C_4}{2\mu}\right) + \frac{2\ell_g C_4^2 \log T}{\mu^2 (T+32\kappa^2-3)}, \end{aligned}$$

which establishes (35). This completes the proof. □

5 Numerical Experiments

In this section, we present numerical experiments to validate the convergence results of CGM with general functional constraints, applied to both minimization and VI problems. All implementations were conducted in Python 3.11.0, with quadratic programming (QP) subproblems (10) and (27) solved using `cvxpy`. A fixed random seed (`np.random.seed(42)`) was used across all experiments to ensure reproducibility.

5.1 Resource Allocation Problem

In this section, we consider the following resource allocation problem (RAP):

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(x) := \frac{1}{2}x^\top \Sigma x + a^\top x \\ \text{s.t.} \quad & x \in \mathcal{C} := \{x \mid x \geq 0,\ \mathbf{1}^\top x = 1,\ r^\top x \leq R_{\max},\ x^\top E x \leq E_{\max}\}, \end{aligned} \tag{41}$$

where $a, r \in \mathbb{R}^d_+$, $\Sigma \succ 0$, $E \succeq 0$, and $E_{\max}, R_{\max} > 0$ are positive constants. Here, the linear constraint $r^\top x \leq R_{\max}$ defines the resource budget limit, the quadratic constraint $x^\top E x \leq E_{\max}$ regulates the allowable risk threshold, and the objective function quantifies the allocation cost. Notably, the objective function f is $\lambda_{\max}(\Sigma)$-smooth and $\lambda_{\min}(\Sigma)$-strongly convex. To cast the set $\mathcal{C}$ in (41) into the form (2), we introduce constraint functions g_i for $i = 1, 2, \ldots, d+4$ as follows: $g_i(x) = -\mathbf{e}_i^\top x$ for $i \in [d]$, $g_{d+1}(x) = \mathbf{1}^\top x - 1$, $g_{d+2}(x) = 1 - \mathbf{1}^\top x$, $g_{d+3}(x) = r^\top x - R_{\max}$, and $g_{d+4}(x) = x^\top E x - E_{\max}$. Let $\mathcal{N}(0,1)$ denote the normal distribution with mean 0 and variance 1, and $\mathcal{U}(0,1)$ denote the uniform distribution over $[0,1]$. In our numerical setup, we generate two independent $d \times 10$ Gaussian matrices G_1 and G_2 with entries sampled from $\mathcal{N}(0,1)$, and set $\Sigma = G_1 G_1^\top + 5I_d$ and $E = G_2 G_2^\top + 10 I_d$. We define $\bar{\sigma} = \frac{1}{d}\sum_{i=1}^d \sqrt{\Sigma_{ii}}$, sample $u \sim \mathcal{U}(0,1)^d$, and set $a = \bar{\sigma} u$. For generating r, we sample each component $r_i \sim |\mathcal{N}(0,1)| + 0.1$ for $i = 1, \ldots, d$. Additionally, $R_{\max}$ and $E_{\max}$ are computed as $R_{\max} = \frac{1}{d}\sum_{i=1}^d r_i$ and $E_{\max} = \frac{1}{d^2}\mathbf{1}^\top E \mathbf{1}$, respectively. The initial point is set to $x^0 = \frac{1}{d}\mathbf{1}$ with $d = 50$, while the optimal solution x^* and optimal value $f(x^*)$ are obtained using `cvxpy`.

We first evaluate the performance of Algorithm 1 for solving (41) with a constant step size $\eta_t = \eta = \log T/(\mu T)$ for all $t = 0, 1, \ldots, T-1$, considering both small iteration counts $T = \{100, 150, 200, 250\}$ and large iteration counts $T = \{1500, 2000, 2500, 3000\}$. The results are reported in Figure 1.

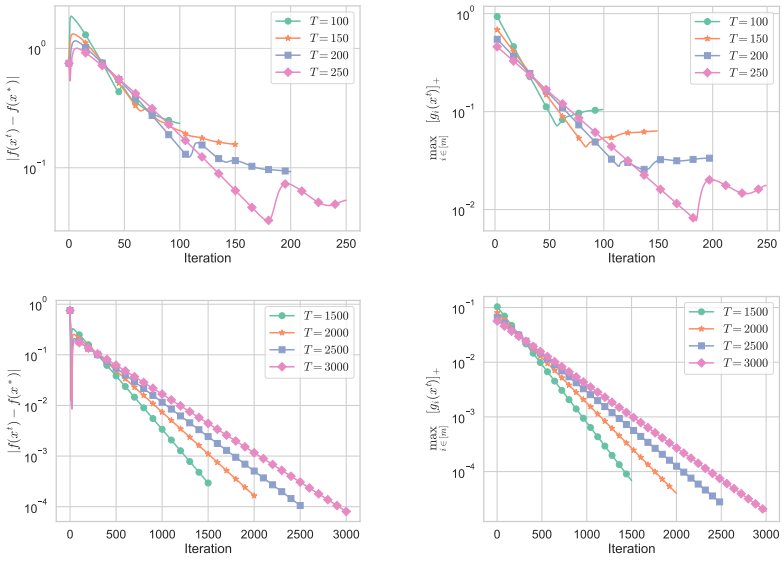

Figure 1: Convergence results of CGM for the RAP with varying iteration counts: small $T = \{100, 150, 200, 250\}$ and large $T = \{1500, 2000, 2500, 3000\}$. Left: absolute function value residual; Right: constraint violation with $m = d + 4$.

From Figure 1, we observe that as T increases, the step size decreases and convergence slows (this trend is more pronounced in the second row), while the resulting solution becomes more accurate. Furthermore, by comparing the first row (corresponding to smaller T) and the second row (corresponding to larger T) of Figure 1, we find that setting T too small may lead to unsatisfactory solution accuracy—whether measured by function value residual or the feasibility.

We also compare the efficacy of Algorithm 1 with two step size strategies: a constant step size $\eta_t = \eta = \log T/(\mu T)$ with $T = 2000$ and a diminishing step size $\eta_t = 1/\mu(t+\kappa)$. The comparison results are presented in Figure 2. From the second subfigure in Figure 2, the initial step sizes of the varying strategy may be too large, causing the iterates to exit the constraint set $\mathcal{C}$ and resulting in objective values significantly smaller than $f(x^*)$. Both approaches for setting the step sizes eventually converge to the optimal value. Additionally, while the constant step size yields better performance than the varying step size in Figure 2, it requires predefining an appropriate T. As illustrated in Figure 1 (particularly in the first row), an improperly chosen T will limit the precision of both the function value residual and constraint violation.

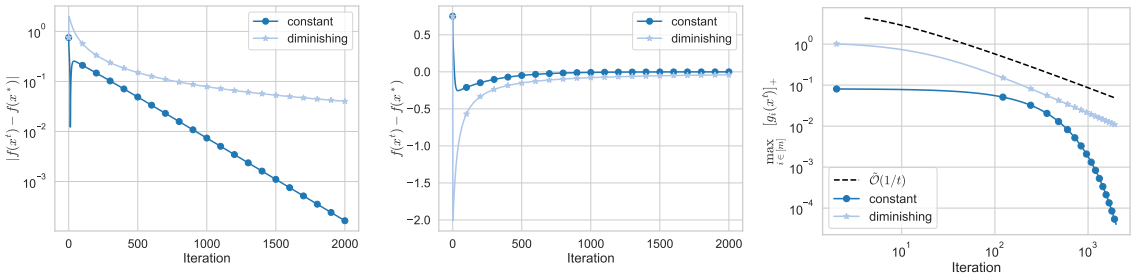

Figure 2: Convergence results of CGM for the RAP: constant step size $\eta_t = \eta = \log T/(\mu T)$ (with $T = 2000$) versus diminishing step size $\eta_t = 1/\mu(t+\kappa)$. Left: absolute function value residual; Middle: function value residual; Right: constraint violation with $m = d + 4$.

5.2 High-Dimensional Bilinear Game

In this section, we consider a high-dimensional bilinear game (HBG) problem formulated as the following min-max problem [CYPJ24]:

$$\begin{aligned} \min_{x_1\in\mathbb{R}^d}\max_{x_2\in\mathbb{R}^d} \quad & \beta x_1^\top x_1 + (1-\beta)x_1^\top x_2 - \beta x_2^\top x_2, \\ \text{s.t.} \quad & x \in \mathcal{C} := \{x = [x_1^\top, x_2^\top]^\top \mid x \geq \mathbf{0}_{2d},\ \mathbf{1}_d^\top x_1 = 1,\ \mathbf{1}_d^\top x_2 = 1\}, \end{aligned} \tag{42}$$

where $x = [x_1^\top, x_2^\top]^\top$ encompasses the strategies of two players (i.e., x_1 and x_2), and the parameter $\beta \in (0,1)$ modulates the rotational influence within the game. Problem (42) can be reformulated as the VI problem (3), with F and constraint functions g_i $(i = 1,\ldots,2d+4)$ defined as follows:

$$F(x) = \begin{bmatrix} 2\beta I_d & (1-\beta)I_d \\ -(1-\beta)I_d & 2\beta I_d \end{bmatrix} x, \quad g_i(x) = -\mathbf{e}_i^\top x \leq 0,\ \forall i \in [2d],$$

along with $g_{2d+1}(x) = \mathbf{1}_d^\top x_1 - 1 \leq 0$, $g_{2d+2}(x) = 1 - \mathbf{1}_d^\top x_1 \leq 0$, $g_{2d+3}(x) = \mathbf{1}_d^\top x_2 - 1 \leq 0$ and $g_{2d+4}(x) = 1 - \mathbf{1}_d^\top x_2 \leq 0$. In this problem, F satisfies (26) with $\ell_F = \sqrt{5\beta^2 - 2\beta + 1}$ and $B = 0$. Furthermore, F is μ-strong monotone with $\mu = 2\beta$. The initial point x^0 is constructed as follows: first generate a $2d$-dimensional vector sampled from $\mathcal{U}(0,1)$, then normalize its first d components and last d components to sum to 1 respectively—this ensures both x_1^0 and x_2^0 lie in the probability simplex. In this case, x^* is given by $\frac{1}{d}[\mathbf{1}_d^\top, \mathbf{1}_d^\top]^\top$, and we set $d = 500$ in our experiment. For this problem, we have $D = 2$.

First, we validate Algorithm 2 with parameters $\alpha = \mu$, $\Delta = \max\left\{1, D^2\ell_F^2/(\|F(x^0)\|^2 + B)\right\}$, and the step size $\eta_t = 1/\mu(t + 16\kappa^2)$ for $t \geq 0$. For computational efficiency, we adopt the (strong) gap function

$\max_{x\in\mathcal{C}}\langle F(x^t), x^t - x\rangle$ as the optimality measure. This is because the constraint set $\mathcal{C}$ is the Cartesian product of two simplices, which allows this gap function to be computed in closed form as:

$$\max_{x\in\mathcal{C}}\langle F(x^t), x^t - x\rangle = \langle F(x^t), x^t\rangle - \min_{j\in[d]}\left(2\beta[x_1^t]_j + (1-\beta)[x_2^t]_j\right) - \min_{j\in[d]}\left(-(1-\beta)[x_1^t]_j + 2\beta[x_2^t]_j\right),$$

where $x^t = [(x_1^t)^\top, (x_2^t)^\top]^\top$, and $[x]_j$ denotes the j-th component of the vector x. The numerical results are presented in Figure 3. As shown in Figure 3, the results align with Theorem 4.1, confirming the validity of our theoretical analysis.

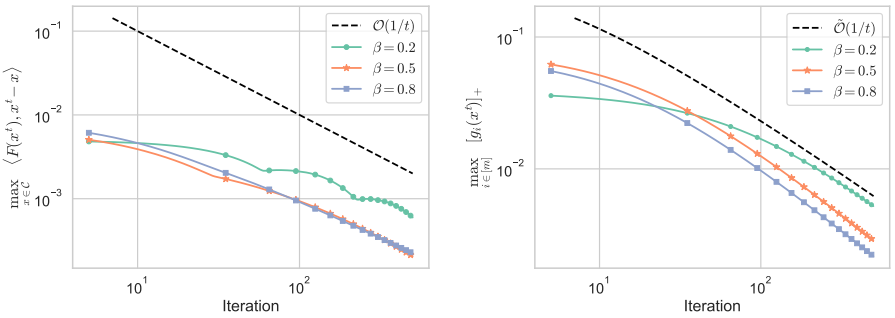

Figure 3: Convergence results of CGM for the HBG problem with different β values. Left: optimality residual; Right: constraint violation with $m = 2d + 5$ (including the auxiliary constraint).

Furthermore, we conducted numerical experiments to compare CGM (with the same parameter settings as above in this subsection) with other projection-based methods under fixed iteration counts and CPU time constraints, with β set to 0.8. The compared methods are as follows:

1. Projected gradient descent ascent (GDA) method: $x^{t+1} = \mathrm{Proj}_{\mathcal{C}}(x^t - \eta F(x^t))$ with a conservative step size $\eta = 0.005$;
2. Projected extra gradient (EG) method [Kor76]: $x^{t+1} = \mathrm{Proj}_{\mathcal{C}}\left(x^t - \eta F\left(\mathrm{Proj}_{\mathcal{C}}\left(x^t - \eta F\left(x^t\right)\right)\right)\right)$ with convergence-guaranteed step size $\eta = 1/\ell_F$.

Here, $\mathrm{Proj}_{\mathcal{C}}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{C}$. The comparison results are presented in Figure 4, which indicate that the CGM outperforms GDA and EG methods in reducing the relative error $\|x^t - x^*\|/\|x^*\|$, both in terms of iteration counts and CPU time.

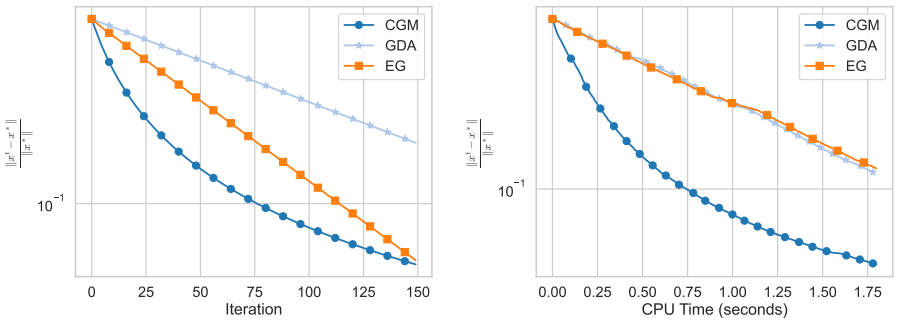

Figure 4: Comparison results of CGM versus GDA and EG methods on the HBG problem: relative error in distance to the optimum under fixed iteration counts and fixed CPU time constraints. Left: relative error v.s. iteration; Right: relative error v.s. CPU time.

6 Conclusions

In this paper, we investigated the convergence analysis of CGM for solving strongly convex optimization problems and strongly monotone VI problems with general functional constraints. We find that the assumptions made in the literature to establish the convergence of CGM are highly restrictive, and some are even inconsistent. To address this gap, we provide a new convergence rate analysis for CGM with appropriately chosen step sizes, under weaker and more reasonable assumptions, for the aforementioned problem classes. Preliminary numerical results further demonstrate the effectiveness and potential of CGM in solving these problems.

References

[AAS+25] Mohammad Alkousa, Belal A. Alashqar, Fefor S. Stonyakin, Tarek NABHANI, and Seydamet Ablaev. Mirror descent methods with weighting scheme for outputs for constrained variational inequality problems. In *First Conference of Mathematics of AI*, 2025.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[AMW25] Ahmet Alacaoglu, Yura Malitsky, and Stephen J. Wright. Towards weaker variance assumptions for stochastic optimization. arXiv:2504.09951, 2025.

[BDK24] Digvijay Boob, Qi Deng, and Mohammad Khalafi. First-order methods for stochastic variational inequality problems with function constraints. arXiv:2304.04778, 2024.

[BDL23] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, 197(1):215–279, 2023.

[Blu54] Julius R. Blum. Approximation methods which converge with probability one. *Ann. Math. Statistics*, 25:382–386, 1954.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[CYPJ24] Tatjana Chavdarova, Tong Yang, Matteo Pagliardini, and Michael Jordan. A primal-dual approach to solving variational inequalities with general constraints. In *ICLR*, 2024.

[FJVW86] Marshall L. Fisher, R. Jaikumar, and Luk N. Van Wassenhove. A multiplier adjustment method for the generalized assignment problem. *Manage. Sci.*, 32(9):1095–1103, 1986.

[FP03] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, New York, 2003.

[GBV+19] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

[GF17] Raoof Gholami and Nikoo Fakhari. Chapter 27 - support vector machine: Principles, parameters, and applications. In Pijush Samui, Sanjiban Sekhar, and Valentina E. Balas, editors, *Handbook of Neural Computation*, pages 515–535. Academic Press, 2017.

[Gla65] E. G. Gladyvsev. On stochastic approximation. *Teor. Verojatnost. i Primenen.*, 10:297–300, 1965.

[Gri19] Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM J. Optim.*, 29(2):1350–1365, 2019.

[HDO+98] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intell. Syst.*, 13(4):18–28, 1998.

[He18] Bing-sheng He. A uniform framework of contraction methods for convex optimization and monotone variational inequality. *Sci. Sin. Math.*, 48(2):255–272, 2018.

[Inc25] The MathWorks Inc. Constrained nonlinear optimization algorithms, 2025.

[JNT11] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.*, 1(1):17–58, 2011.

[KF11] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *Int. J. Rob. Res.*, 30(7):846–894, June 2011.

[KMM23] Pavel Kolev, Georg Martius, and Michael Muehlebach. Online learning under adversarial nonlinear constraints. In *NeurIPS*, volume 36, pages 53227–53238, 2023.

[Kor76] G. M. Korpelevìvc. An extragradient method for finding saddle points and for other problems. *Èkonom. i Mat. Metody*, 12(4):747–756, 1976.

[KR23] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *TMLR*, 2023.

[KS00] David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. SIAM, Philadelphia, 2000.

[Liu05] Xin-Wei Liu. Global convergence on an active set SQP for inequality constrained optimization. *J. Comput. Appl. Math.*, 180(1):201–211, 2005.

[Min62] George J Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29(1):341–346, 1962.

[MJ22] Michael Muehlebach and Michael I. Jordan. On constraints in first-order optimization: A view from non-smooth dynamical systems. *JMLR*, 23(256):1–47, 2022.

[MJ25] Michael Muehlebach and Michael I. Jordan. Accelerated first-order optimization under nonlinear constraints. *Math. Program.*, 2025.

[Nag98] Anna Nagurney. *Network Economics: A Variational Inequality Approach*. Springer Science & Business Media, Dordrecht, 1998.

[Nas50] John Nash. Equilibrium points in n-person games. *PNAS*, 36(1):48–49, 1950.

[Nas51] John Nash. Non-cooperative games. *Ann. Math.*, pages 286–295, 1951.

[Nem04] Arkadi Nemirovski. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2004.

[Nes07] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2-3):319–344, 2007.

[NT88] P. Neittaanmäki and D. Tiba. A variational inequality approach to constrained control problems for parabolic equations. *Appl. Math. Optim.*, 17(3):185–201, 1988.

[PEAM21] Jungho Park, Hadi El-Amine, and Nevin Mutlu. An exact algorithm for large-scale continuous nonlinear resource allocation problems with minimax regret objectives. *INFORMS J. Comput.*, 33(3):1213–1228, 2021.

[Pow78a] M. J. D. Powell. The convergence of variable metric methods for nonlinearly constrained optimization calculations. In *Nonlinear programming, 3 (Proc. Sympos., Special Interest Group Math. Programming, Univ. Wisconsin, Madison, Wis., 1977)*, pages 27–63. Academic Press, New York-London, 1978.

[Pow78b] M. J. D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis (Proc. 7th Biennial Conf., Univ. Dundee, Dundee, 1977)*, volume Vol. 630 of *Lecture Notes in Math.*, pages 144–157. Springer, Berlin-New York, 1978.

[SB84] C. Saguez and A. Bermudez. Optimal control of variational inequalities. In *CDC*, pages 249–251, 1984.

[Sch92] K. Schittkowski. Solving nonlinear programming problems with very many constraints. *Optimization*, 25(2-3):179–196, 1992.

[Sch09] K. Schittkowski. An active set strategy for solving optimization problems with up to 200,000,000 nonlinear constraints. *Appl. Numer. Math.*, 59(12):2999–3007, 2009.

[STM+22] Sholom Schechtman, Daniil Tiapkin, Eric Moulines, Michael I. Jordan, and Michael Muehlebach. First-order constrained optimization: Non-smooth dynamical system viewpoint. *IFAC-PapersOnLine*, 55(16):236–241, 2022.

[STMM23] Sholom Schechtman, Daniil Tiapkin, Eric Moulines, and Michael Muehlebach. Orthogonal directions constrained gradient method: from non-linear equality constraints to Stiefel manifold. In *COLT*, volume 195, pages 1228–1258. PMLR, 2023.

[TJNO20] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method. In *NeurIPS*, volume 33, pages 12211–12224, 2020.

[TYH11] Min Tao, Xiao-ming Yuan, and Bing-sheng He. Solving a class of matrix minimization problems by linear variational inequality approaches. *Linear Algebra Appl.*, 434(11):2343–2352, 2011.

[YJC23] Tong Yang, Michael I. Jordan, and Tatjana Chavdarova. Solving constrained variational inequalities via a first-order interior point-based method. In *ICLR*, 2023.

[ZHM25] Liang Zhang, Niao He, and Michael Muehlebach. Primal methods for variational inequality problems with functional constraints. *Math. Program.*, 2025.

[ZZZ25] Lei Zhao, Daoli Zhu, and Shuzhong Zhang. An augmented lagrangian approach to conically constrained nonmonotone variational inequality problems. *Math. Oper. Res.*, 50(3):1868–1900, 2025.

A Related Work

A.1 Methods related to CGM for solving constrained minimization problems

For the constrained minimization problem (1), the algorithm closest to CGM is probably the sequential quadratic programming method (SQP) [Pow78b, Pow78a, Inc25]. Recall that one key feature of CGM is its reduction of a complex problem to a sequence of simpler quadratic programming (QP) subproblems with linearized constraints. Likewise, SQP bases its iterations on lightweight QP subproblems. With this in mind, we will first focus on highlighting the differences between CGM and SQP.

Under the additional assumption that f and g are twice continuously differentiable, to solve problem (1), at iterate t, SQP requires solving the following QP subproblem to obtain the descent direction:

$$\begin{aligned} \min_{v\in\mathbb{R}^n} \quad & \nabla f(x^t)^\top v + \frac{1}{2} v^\top B^t v \\ \text{s.t.} \quad & g_i\big(x^t\big) + \nabla g_i\big(x^t\big)^\top v \le 0, \quad i \in [m], \end{aligned} \tag{43}$$

where $B^t \in \mathbb{R}^{n\times n}$ is typically required to be positive definite, and it serves as an approximation of the Hessian of the Lagrangian function

$$\mathcal{L}(x,\lambda) = f(x) + \lambda^\top g(x), \text{ with } g(x) = (g_1(x), g_2(x), \dots, g_m(x)), \ \lambda \in \mathbb{R}^m_+$$

evaluated at the current iterate (x^t, λ^t). The iterate is updated by $x^{t+1} = x^t + \alpha_t v^t$, where v^t is the optimal solution of (43), and α_t is used determined by line search. The matrix B^t in (43) is often updated by quasi-Newton methods. SQP is effective for solving medium and small-sized nonlinear programs, but its scalability is limited by the high memory cost of QP subproblems in large-scale settings. This limitation has, in turn, motivated the development of various SQP methods incorporating the active set technique. Both Schittkowski [Sch92, Sch09] and Liu [Liu05] constructed (working) active sets to reduce the size of the QP subproblems in SQP methods by utilizing information from local constraint violations and multipliers. Specifically, Schittkowski's method presupposes that at most m_w constraints are active. Using a fixed user-defined tolerance ε, it defines indices

$$J_t^* = \{i \mid g_i(x^t) \geq -\varepsilon \text{ or } \lambda_i^t > 0\} \quad \text{and} \quad K_t^* = [m] \setminus J_t^*.$$

A heuristic then selects a subset $\bar{K}_t^* \subset K_t^*$ such that the working active set $W_t = J_t^* \cup \bar{K}_t^*$ has exactly m_w elements. Liu's method, on the other hand, requires extra theoretical conditions, including a strengthened MFCQ and uniformly bounded B^t. Its active set, $I(x^t, \lambda^t, \varepsilon_t) = \{i \mid g_i(x^t) \geq -(\varepsilon_t + \lambda_i^t)\}$, relies on a varying tolerance parameter ε_t that must converge to zero to ensure the algorithm's convergence. In contrast, the QP subproblem in CGM is much simpler, as it does not require estimating second-order information of the Lagrangian function and typically employs a simple active-set technique that relies solely on local constraint violations $I_{x^t} := \{i \in [m] \mid g_i(x^t) \geq 0\}$. This feature makes the QP subproblem in CGM easier to solve. Moreover, the step size strategy in CGM can adopt either a constant or a variable step size that satisfies certain conditions, without requiring a complex line search.

Next, consider that CGM computes a descent direction by projecting the negative gradient onto the linearized, sparse active constraints in (10), and this projection-based approach for generating descent directions shares conceptual similarities with the Riemannian gradient descent method (RGD) [AMS08]. We then discuss the similarities and differences between RGD and CGM. In RGD, the descent direction is obtained by projecting the negative Euclidean gradient onto the tangent space of the constraint manifold—a local linear approximation of the feasible set. This yields the Riemannian gradient $\mathrm{grad} f(x)$, specifically, for an equality-constrained manifold $\mathcal{M} = \{x \in \mathbb{R}^n : g_i(x) = 0, \forall i \in [m]\}$ with regular constraints, i.e., the gradient vectors $\{\nabla g_1(x), \ldots, \nabla g_m(x)\}$ are linearly independent at $x \in \mathcal{M}$, the tangent space at x is given by $T_x\mathcal{M} = \{v \in \mathbb{R}^n : \nabla g_i(x)^\top v = 0, \forall i \in [m]\}$, and the Riemannian gradient admits the variational characterization:

$$\mathrm{grad} f(x) = \arg\min_{v \in T_x\mathcal{M}} \frac{1}{2}\|v + \nabla f(x)\|^2,$$

which exhibits strong similarity with the update direction in (10), in fact, the two are identical when $x^t \in \mathcal{M}$. The key difference lies in how the next iterate is produced: RGD requires an additional retraction step to map points from the tangent space back onto the manifold, thereby maintaining feasibility. In contrast, CGM imposes no such feasibility requirement on x^{t+1} lying in the feasible set.

A.2 Recent progress for solving constrained VIs

For solving (3), one can adopt the traditional projected gradient method [FP03] and the Frank-Wolfe method [TJNO20], which necessitates explicit projection or linear minimization oracles to address the feasibility issue. However, for general constrained sets, performing such oracles is prohibitively expensive, so an alternative line of work turns to primal-dual algorithms that exploit (augmented) Lagrangian reformulations [YJC23, CYPJ24, BDK24, BDL23, ZZZ25]. Nevertheless, both these algorithms and their theoretical analysis heavily depend on the existence or the magnitude of the optimal Lagrange multipliers. Specifically, [YJC23] proposed an ADMM-based interior point method, which was further improved by [CYPJ24] through a warm-starting technique. [BDK24] extended the constraint extrapolation method [BDL23] for constrained minimization problems to constrained VIs, relying on bounded optimal Lagrange multipliers for step size selection and convergence rates, and requiring one projection per iteration onto a "simple" convex compact set to ensure the boundedness of the iterates theoretically. [ZZZ25] developed a primal-dual system and designed the augmented Lagrangian-based ALAVI method. This algorithm ensures global convergence under a condition weaker than monotonicity—termed primal-dual variational coherence, but it requires two projections each iteration onto the dual cone $\mathcal{C}^*$. More recently, [AAS+25] proposed a mirror descent type method for solving (3). This method dynamically switches between objective-descent and constraint-descent steps, depending

on whether the functional constraints satisfy $\max_{i\in[m]} g_i(x) \leq \varepsilon$ at each iteration. With time-varying step sizes, and assuming F is bounded and monotone, the method achieves a $\mathcal{O}(1/\sqrt{T})$ convergence rate.