Journal of Machine Learning Research 24 (2023) 1-34 Submitted 8/22; Revised 3/23; Published 5/23

A Parameter-Free Conditional Gradient Method for Composite Minimization under Hölder Condition

Masaru Ito ITO.MASARU@NIHON-U.AC.JP
Department of Mathematics
Nihon University
Tokyo 101-8308, JAPAN

Zhaosong Lu ZHAOSONG@UMN.EDU
Department of Industrial and Systems Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Chuan He HE000233@UMN.EDU
Department of Industrial and Systems Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Martin Jaggi

Abstract

In this paper we consider a composite optimization problem that minimizes the sum of a weakly smooth function and a convex function with either a bounded domain or a uniformly convex structure. In particular, we first present a parameter-dependent conditional gradient method for this problem, whose step sizes require prior knowledge of the parameters associated with the Hölder continuity of the gradient of the weakly smooth function, and establish its rate of convergence. Given that these parameters could be unknown or known but possibly conservative, such a method may suffer from implementation issue or slow convergence. We therefore propose a parameter-free conditional gradient method whose step size is determined by using a constructive local quadratic upper approximation and an adaptive line search scheme, without using any problem parameter. We show that this method achieves the same rate of convergence as the parameter-dependent conditional gradient method. Preliminary experiments are also conducted and illustrate the superior performance of the parameter-free conditional gradient method over the methods with some other step size rules.

Keywords: Conditional gradient method, Hölder continuity, uniform convexity, adaptive line search, iteration complexity

1 Introduction

In this paper we consider a composite optimization problem in the form of

$$\varphi^* = \min_{x\in\mathbb{E}} \{\varphi(x) := f(x) + g(x)\}, \tag{1}$$

where $\mathbb{E}$ is a finite dimensional real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$, and the functions $f, g : \mathbb{E} \to \mathbb{R}\cup\{+\infty\}$ are proper and lower-semicontinuous. Assume that φ^* is finite and attainable, and that g is a convex function, while f is possibly nonconvex.

In the recent years conditional gradient methods (e.g., Bach, 2015; Harchaoui et al., 2015; Nesterov, 2018; Ghadimi, 2019) have been developed for solving problem (1), which generate the iterates $\{x_t\}$ by the following scheme:

$$v_t \in \operatorname*{Argmin}_{x\in\mathbb{E}}\{\langle \nabla f(x_t), x\rangle + g(x)\}, \quad x_{t+1} = (1-\tau_t)x_t + \tau_t v_t \tag{2}$$

for each $t \geq 0$, where $\tau_t \in [0,1]$ is a step size chosen by a certain rule. These methods originate from the Frank-Wolfe method (Frank and Wolfe, 1956) that was initially proposed for quadratic programming and later studied for more general or structured problems (e.g., Levitin and Polyak, 1966; Demyanov and Rubinov, 1970; Dunn, 1979; Beck and Teboulle, 2004). Conditional gradient methods have found rich applications in machine learning and statistics, where f and g are typically a loss function and a regularizer, respectively. The advantage of conditional gradient methods in these applications is that the solution v_t to the subproblem in (2) is efficiently computable and also has some desirable property, such as preserving sparsity or low rank (Hazan, 2008; Clarkson, 2010; Jaggi, 2013; Harchaoui et al., 2015; Freund and Grigas, 2016; Nesterov, 2018).

The conditional gradient methods provide a computable quantity δ_t, commonly referred to as the Frank-Wolfe gap, which is given by

$$\delta_t := \langle \nabla f(x_t), x_t - v_t\rangle + g(x_t) - g(v_t) = \max_{x\in\mathbb{E}}\{\langle \nabla f(x_t), x_t - x\rangle + g(x_t) - g(x)\} \geq 0. \tag{3}$$

Notice that $\delta_t = 0$ if and only if x_t is a stationary point of φ. In addition, if f is convex, one can have $\varphi(x_t) - \varphi^* \leq \delta_t$ (see Lemma 3). Consequently, $\delta_t \leq \varepsilon$ is often used as a termination criterion for the conditional gradient methods.

The choice of step sizes is crucial for the performance of the conditional gradient methods, which is typically measured by the iteration complexity, namely, the worst-case number of iterations for reaching $\delta_t \leq \varepsilon$ or $\varphi(x_t) - \varphi^* \leq \varepsilon$ for a prescribed tolerance $\varepsilon > 0$. There are numerous studies (e.g., Frank and Wolfe, 1956; Levitin and Polyak, 1966; Dunn, 1979; Freund and Grigas, 2016) on the step size rule when f is L-smooth, i.e., ∇f is Lipschitz continuous with constant L under a norm $\|\cdot\|$. For example, the choice

$$\tau_t = \min\left\{1, \frac{\delta_t}{L\|x_t - v_t\|^2}\right\} \tag{4}$$

guarantees an iteration complexity of $O(\varepsilon^{-2})$ for reaching $\delta_t \leq \varepsilon$ (Lacoste-Julien, 2016). When f is additionally convex, it can be improved to $O(\varepsilon^{-1})$, which can also be achieved by the step size $\tau_t = \frac{2}{t+2}$, a well-known choice for convex f (Jaggi, 2013; Freund and Grigas, 2016).

More generally, step size rules were studied when f is weakly smooth, that is, ∇f is Hölder continuous with exponent $\nu \in (0,1]$. In particular, when f is additionally convex, the choice $\tau_t = \frac{2}{t+2}$ ensures an iteration complexity of $O(\varepsilon^{-1/\nu})$ for reaching $\varphi(x_t) - \varphi^* \leq \varepsilon$ (Nesterov, 2018), which is known to be nearly optimal from complexity theory perspective (Guzmán and Nemirovski, 2015). When g is strongly convex, Nesterov (2018) proposed the step size $\tau_t = \frac{6(t+1)}{(t+2)(2t+3)}$ for obtaining a better iteration complexity of $O(\varepsilon^{-\frac{1}{2\nu}})$. Recently, Ghadimi (2019) improved this complexity to $O(\varepsilon^{-\frac{1-\nu}{2\nu}}\log\frac{1}{\varepsilon})$ by using a step size τ_t determined by a backtracking line search procedure. Remarkably, his method enjoys a linear

rate of convergence when $\nu = 1$. Ghadimi's method is also applicable to the case where f is nonconvex and achieves an iteration complexity of $O(\varepsilon^{-\frac{1-\nu}{2\nu}-1})$ for reaching $\delta_t \leq \varepsilon$.

Conditional gradient methods were also studied for minimizing an L-smooth convex function over a strongly convex set (e.g., see Levitin and Polyak, 1966; Dunn, 1979; Garber and Hazan, 2015). Under the assumption that the set does not contain a stationary point of the function, it was established that the conditional gradient method with the step size given in (4) has a linear rate of convergence (Levitin and Polyak, 1966; Dunn, 1979). Such a method was also generalized to minimize an L-smooth convex function over a uniformly convex set (Kerdreux et al., 2021a).

In this paper we first study a parameter-dependent conditional gradient method proposed in (Zhao and Freund, 2020, Algorithm 4) with a step size depending explicitly on the problem parameters for solving a broad class of problems in the form of (1), including but not limited to the problems considered in the above references (Levitin and Polyak, 1966; Dunn, 1979; Jaggi, 2013; Garber and Hazan, 2015; Freund and Grigas, 2016; Lacoste-Julien, 2016; Nesterov, 2018; Ghadimi, 2019; Kerdreux et al., 2021a; Zhao and Freund, 2020). Though this method was analyzed in (Zhao and Freund, 2020) for problem (1) with *convex* f and *bounded* $\operatorname{dom} g$, there is a lack of analysis for (1) with nonconvex f. In this paper, we analyze the rate convergence of this method for problem (1) with f being possibly nonconvex under the assumption that ∇f is Hölder continuous and $\operatorname{dom} g$ is bounded or the problem has a uniformly convex structure. As a byproduct, we obtain a new iteration complexity of $O(\varepsilon^{-\frac{1-\nu}{2\nu}})$ [1] for problem (1) with ∇f being Hölder continuous with exponent $\nu \in (0, 1)$ and g being strongly convex, which improves by the factor $\log(1/\varepsilon)$ the previously best known one (Ghadimi, 2019).

Though the aforementioned parameter-dependent method (Algorithm 1) is simple and also enjoys a nice iteration complexity, its step size may suffer from some issues. Indeed, its step size requires prior knowledge of the parameters ν and M_ν associated with the Hölder continuity of ∇f. Since they depend on f, g, and also a particular norm on $\mathbb{E}$, they may be hard to be found if f is sophisticated. On another hand, the parameters ν and M_ν are not unique. The tighter value of them typically leads to a faster convergent algorithm. Yet, it may be challenging to find the tightest possible value for them. Motivated by these, we further propose a parameter-free conditional gradient method (see Algorithm 2) in which the step size is chosen by using a constructive local quadratic upper approximation and an adaptive line search scheme, without using prior knowledge of ν and M_ν. We show that this method achieves the same rate of convergence as the parameter-dependent conditional gradient method, which, however, uses prior knowledge of the problem parameters ν and M_ν.

The results of this paper were presented in the SIAM Conference on Optimization in July 2021 (Ito et al., 2021). During the preparation of this paper, a concurrent work (Peña, 2022) proposed a parameter-free conditional gradient method for problem (1) with *convex* f and established similar complexity bounds as the ones obtained in this paper yet in terms of primal-dual optimality gap that is typically weaker than the Frank-Wolfe gap which we

1. For simplicity, the complexity here only emphasizes its dependence on the tolerance parameter ε, while the other parameters such as ν are viewed as a fixed constant and thus omitted. Its detailed expression including the dependence on the other parameters can be found in (22) with $\rho = 2$.

use. It shall be mentioned that our analyses are vastly different from those in (Peña, 2022) and shed new insights into conditional gradients for solving a broader class of problems.

The rest of this paper is organized as follows. In Section 2 we introduce some notation and make some assumptions on the problem studied in this paper. In Section 3 we propose a parameter-dependent conditional gradient method, and show some results on its rate of convergence. In Section 4 we propose a parameter-free conditional gradient method, and establish its rate of convergence and also iteration complexity. Section 5 presents numerical experiments to compare the performance of this method with the conditional gradient methods with some other step size rules. The proofs of main results are given in Section 6. Finally, we make some concluding remarks in Section 7.

2 Notation and Assumptions

Throughout the paper, $\mathbb{E}$ is a finite dimensional real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$. Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{E}$ and $\|\cdot\|_*$ its dual, i.e., $\|z\|_* = \sup_{\|x\|\le 1}\langle z, x\rangle$ for $z \in \mathbb{E}$. We denote by $\mathbb{R}_+$ and $\mathbb{Z}_+$ the set of nonnegative real numbers and the set of nonnegative integers, respectively. For any real number a, we denote by a_+ the nonnegative part of a, that is, $a_+ = \max\{a, 0\}$. For the convex function g, $\operatorname{dom} g$ denotes the domain of g, i.e., $\operatorname{dom} g = \{x \in \mathbb{E} : g(x) \neq +\infty\}$. We denote by D_g the diameter of $\operatorname{dom} g$, that is,

$$D_g = \sup_{x,y\in\operatorname{dom} g} \|x - y\|. \tag{5}$$

Clearly, $D_g = +\infty$ if $\operatorname{dom} g$ is unbounded.

We make the following assumption on the functions f and g throughout this paper.

Assumption 1 *(i) The function $f : \mathbb{E} \to \mathbb{R}\cup\{+\infty\}$ is a proper and lower-semicontinuous function. Moreover, f is differentiable on $\operatorname{dom} g$ and the gradient ∇f is Hölder continuous on $\operatorname{dom} g$, i.e., there exist $\nu \in (0, 1]$ and $M_\nu > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\|_* \le M_\nu \|x - y\|^\nu, \quad \forall x, y \in \operatorname{dom} g. \tag{6}$$

(ii) The function $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower-semicontinuous, and convex function. In addition, for each $x \in \operatorname{dom} g$, the subproblem

$$\min_{v\in\mathbb{E}}\{\langle\nabla f(x), v\rangle + g(v)\}. \tag{7}$$

has at least one optimal solution.

The function f satisfying Assumption 1(i) is commonly referred to as a (ν, M_ν)-*weakly smooth function* on $\operatorname{dom} g$. There are many instances of problem (1) satisfying Assumption 1(i). For example, problem (1) with ∇f being semi-algebraic[2] continuous and $\operatorname{dom} g$ being a compact semi-algebraic set is one of them (Bolte et al., 2020, Proposition C.1). Also, there are some machine learning models satisfying Assumption 1(i) (see, e.g., Bolte

2. A *semi-algebraic set* is a finite union of sets of the form $\{x \in \mathbb{E} \mid p_i(x) = 0,\ i = 1, \ldots, k,\ q_j(x) < 0,\ j = 1, \ldots, l\}$ for some real polynomials p_i, q_j. A map $h : \mathbb{R}^m \to \mathbb{R}^n$ is said to be *semi-algebraic* if its graph $\{(x, y) \mid y = h(x)\}$ is a semi-algebraic set.

et al., 2020; Zwiernik, 2016; Chen et al., 2020). In addition, Assumption 1(ii) plays an important role in a conditional gradient method for solving problem (1). Indeed, all the conditional gradient methods in the literature solve a subproblem in the form of (7) per iteration.

In this paper, we will consider problem (1) satisfying Assumption 1 and one additional assumption that g has a bounded domain or problem (1) has a uniformly convex structure introduced below.

Assumption 2 *Problem* (1) *has a* uniformly convex structure*, that is, there exist* $\kappa > 0$ *and* $\rho \geq 2$ *such that for any* $x \in \operatorname{dom} g$ *and any optimal solution* v^* *to the subproblem* (7)*, we have*

$$\langle \nabla f(x), v\rangle + g(v) - \langle \nabla f(x), v^*\rangle - g(v^*) \geq \frac{\kappa}{\rho}\|v - v^*\|^\rho, \quad \forall v \in \mathbb{E}. \tag{8}$$

There are many instances of problem (1) with a uniformly convex structure. We next provide two examples of (1) for which Assumptions 1 and 2 hold.

Example 1 (Optimization over a uniformly convex set) *Let* $C \subset \mathbb{E}$ *be a nonempty compact* (c,ρ)-uniformly convex set with respect to $\|\cdot\|$ *for some* $c > 0$ *and* $\rho \geq 2$,[3] *that is,* C *is a nonempty compact convex set satisfying that* $(1-\lambda)x + \lambda y + z \in C$ *for any* $x, y \in C$*,* $\lambda \in [0,1]$*, and* z *with* $\|z\| \leq \lambda(1-\lambda)\frac{c}{\rho}\|x-y\|^\rho$*. Consider the problem*

$$\min_{x\in C} f(x), \tag{9}$$

where f *is differentiable on* C*,* ∇f *is Hölder continuous on* C*, and* $\alpha = \min_{x\in C}\|\nabla f(x)\|_* > 0$ *(i.e., no stationary point of* f *belongs to* C*). Let* g *be the indicator function of* C*. One can easily observe that problem* (9) *is a special case of* (1)*, and moreover, Assumption 1 holds for it. We next verify that Assumption 2 also holds for it. Indeed, let* $x \in \operatorname{dom} g$ *and* $v^* \in \operatorname{Argmin}_{v\in\mathbb{E}}\{\langle \nabla f(x), v\rangle + g(v)\}$ *be arbitrarily chosen. Then we have that* $x \in C$ *and* $v^* \in \operatorname{Argmin}_{v\in C}\langle \nabla f(x), v\rangle$*. Since* C *is a* (c,ρ)*-uniformly convex set, one has that* $w = \lambda v^* + (1-\lambda)v + [\lambda(1-\lambda)c\|v-v^*\|^\rho/\rho]z \in C$ *for any* $\lambda \in (0,1)$*,* $v \in C$*, and* z *with* $\|z\| \leq 1$*. This and the optimality of* v^* *lead to*

$$\left\langle \nabla f(x), \lambda v^* + (1-\lambda)v + \lambda(1-\lambda)\frac{c}{\rho}\|v-v^*\|^\rho z - v^* \right\rangle = \langle \nabla f(x), w - v^*\rangle \geq 0,$$

which implies that $\langle \nabla f(x), v - v^*\rangle \geq (\lambda c/\rho)\|v-v^*\|^\rho \langle \nabla f(x), -z\rangle$*. Taking* $\sup_{\|z\|\leq 1}$ *on both sides of this inequality, letting* $\lambda \uparrow 1$*, and using* $\alpha = \min_{x\in C}\|\nabla f(x)\|_*$*, we obtain that*

$$\langle \nabla f(x), v - v^*\rangle \geq \|\nabla f(x)\|_* \cdot \frac{c}{\rho}\|v-v^*\|^\rho \geq \frac{\alpha c}{\rho}\|v-v^*\|^\rho.$$

This, together with g *being the indicator function of* C*, implies that* (8) *is satisfied with* $\kappa = \alpha c$*. Therefore, Assumption 2 holds for problem* (9)*.*

3. For example, the ℓ_p-balls are uniformly convex with $\rho = \max(2,p)$ for $p \in (1,\infty)$. In addition, C is often referred to as a *strongly convex set* if $\rho = 2$. See Vial (1982); Levitin and Polyak (1966); Garber and Hazan (2015); Kerdreux et al. (2021a,b) for the discussion on uniformly or strongly convex sets.

Example 2 (Optimization with a uniformly convex function) *Consider a special case of problem* (1)*, where ∇f is Hölder continuous on* $\operatorname{dom} g$*, and g is a proper, lower-semicontinuous and* (κ, ρ)-uniformly convex function with respect to $\|\cdot\|$ *for some $\kappa > 0$ and $\rho \geq 2$, that is, g satisfies that*

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) - \lambda(1-\lambda)\frac{\kappa}{\rho}\|x-y\|^\rho, \quad \forall x, y \in \operatorname{dom} g, \lambda \in [0,1].^4$$

By this, one can observe that (e.g., see Zălinescu, 2002)

$$g(y) \geq g(x) + g'(x; y-x) + \frac{\kappa}{\rho}\|x-y\|^\rho, \quad \forall x, y \in \operatorname{dom} g, \tag{10}$$

where $g'(x;d) = \lim_{s\downarrow 0}(g(x+sd)-g(x))/s$ is the directional derivative of g along the direction d. It then follows that g is coercive, which together with the lower-semicontinuity of g implies that subproblem (7) *has at least one optimal solution. Thus, Assumption 1 holds for this problem. We next verify that Assumption 2 also holds for it. Indeed, let $x \in \operatorname{dom} g$ and $v^* = \operatorname{argmin}_{v\in\mathbb{E}}\{\langle \nabla f(x), v\rangle + g(v)\}$. By these and* (10)*, one has*

$$\langle \nabla f(x), v\rangle + g(v) \geq \langle \nabla f(x), v^*\rangle + \langle \nabla f(x), v - v^*\rangle + g(v^*) + g'(v^*; v-v^*) + \frac{\kappa}{\rho}\|v-v^*\|^\rho,$$

for all $v \in \operatorname{dom} g$. In addition, by the optimality of v^*, we have $\langle \nabla f(x), v-v^*\rangle + g'(v^*; v-v^*) \geq 0$ for all $v \in \operatorname{dom} g$. These two inequalities immediately yield (8)*. Therefore, Assumption 2 also holds for this problem.*

3 A Parameter-Dependent Conditional Gradient Method

In this section we present in Algorithm 1 a parameter-dependent conditional gradient method for solving problem (1), whose step size τ_t depends on the problem parameters ν and M_ν explicitly. This algorithm was proposed in (Zhao and Freund, 2020, Algorithm 4) and analyzed by them for the case where f is convex and $\operatorname{dom} g$ is bounded. However, it was not analyzed for the case where f is nonconvex. In what follows, we will analyze its rate of convergence for solving (1) with f being possibly nonconvex under the assumption that $\operatorname{dom} g$ is bounded or problem (1) has a uniformly convex structure. It shall be mentioned that the convergence results established in this section also hold for a variant of Algorithm 1 with the exact step size $\tau_t \in \operatorname{Argmin}_{\tau\in[0,1]}\varphi((1-\tau)x_t + \tau v_t)$.

Before proceeding, we state a well-known lemma (e.g., see Ghadimi, 2019), which shows that the quantity δ_t, commonly referred to as the *Frank-Wolfe gap*, provides an upper bound on the optimality gap of problem (1) at x_t when f is convex. For the sake of completeness, we include a proof for it.

Lemma 3 *Let the sequences $\{x_t\}$ and $\{\delta_t\}$ be generated in Algorithm 1. Suppose that f is convex. Then it holds that $\delta_t \geq \varphi(x_t) - \varphi^*$ for all $t \geq 0$.*

4. For the case $\rho = 2$, the function g is a usual strongly convex function with modulus κ.

Algorithm 1: A parameter-dependent conditional gradient method

Input: $x_0 \in \operatorname{dom} g$.

1: **for** $t = 0, 1, 2, \ldots,$ **do**
2: $\quad v_t \in \operatorname{Argmin}_{x \in \mathbb{E}} \{\langle \nabla f(x_t), x \rangle + g(x)\}$.
3: $\quad \delta_t = \langle \nabla f(x_t), x_t \rangle + g(x_t) - \langle \nabla f(x_t), v_t \rangle - g(v_t)$.
4: $\quad \tau_t = \min\left\{1, \left(\frac{\delta_t}{M_\nu \|x_t - v_t\|^{1+\nu}}\right)^{\frac{1}{\nu}}\right\}$.
5: $\quad x_{t+1} = (1 - \tau_t) x_t + \tau_t v_t$.
6: **end for**

Proof Let x^* be an arbitrary optimal solution of problem (1). Then $f(x^*) + g(x^*) = \varphi^*$. By this, the expression of δ_t, and the convexity of f, we have that for all $t \geq 0$,

$$\begin{aligned} \delta_t &= \langle \nabla f(x_t), x_t \rangle + g(x_t) - \langle \nabla f(x_t), v_t \rangle - g(v_t) \\ &\geq \langle \nabla f(x_t), x_t \rangle + g(x_t) - \langle \nabla f(x_t), x^* \rangle - g(x^*) \\ &\geq f(x_t) + g(x_t) - f(x^*) - g(x^*) = \varphi(x_t) - \varphi^*. \end{aligned} \tag{11}$$

■

Remark 4 *From the proof of Lemma 3, one can observe that the convexity of f is only used in* (11)*. More generally,* (11) *is also valid if f satisfies the* star-convexity *property (Nesterov and Polyak, 2006): there exists some $x^* \in \operatorname{Argmin}_x \varphi(x)$ such that $f(\lambda x^* + (1-\lambda)x) \leq \lambda f(x^*) + (1-\lambda) f(x)$ for all $\lambda \in [0, 1]$ and $x \in \operatorname{dom} g$. Thus, the conclusion of Lemma 3 also holds if f satisfies the star-convexity property. Moreover, all the results established in this paper for a convex f also hold for a star-convex f.*

In what follows, we state some results regarding the rate of convergence of Algorithm 1 in Theorems 5 and 7, whose proofs are deferred to Section 6.2. In particular, we first present the results under the assumption that g has a bounded domain, namely, $D_g < +\infty$.

Theorem 5 *Let the sequences $\{x_t\}$ and $\{\delta_t\}$ be generated in Algorithm 1. Suppose that Assumption 1 holds, $D_g < +\infty$, and that $\delta_t > 0$ for all $t \geq 0$, where D_g is defined in* (5)*. Let $\delta_t^* = \min_{0 \leq i \leq t} \delta_i$ and*

$$A = M_\nu^{\frac{1}{\nu}} D_g^{\frac{1+\nu}{\nu}}, \quad t_0 = \left\lceil \frac{1+\nu}{\nu} \left(\log \frac{(1+\nu)(\varphi(x_0) - \varphi^*)}{\nu A^\nu} \right)_+ \right\rceil,$$
$$\overline{\gamma}_t = [(\varphi(x_{t_0}) - \varphi^*)^{-\frac{1}{\nu}} + (1+\nu)^{-1} A^{-1} (t - t_0)]^{-\nu}, \quad \forall t \geq t_0.$$

Then the following statements hold.

(i) $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t \to \infty} \varphi(x_t)$ exists. In addition, $\{\delta_t^*\}$ satisfies

$$\delta_t^* \leq \max\left\{ \frac{(1+\nu)(\varphi(x_0) - \varphi_*)}{\nu(t+1)}, \left(\frac{(1+\nu) A (\varphi(x_0) - \varphi_*)}{\nu(t+1)} \right)^{\frac{\nu}{1+\nu}} \right\}, \quad \forall t \geq 0. \tag{12}$$

(ii) Assume additionally that f is convex. Then we have

$$\varphi(x_t) - \varphi^* \le \overline{\gamma}_t, \quad \forall t \ge t_0,$$
$$\delta_t^* \le e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor}, \quad \forall t \ge t_0 + \frac{2(1+\nu)A}{\nu(\varphi(x_{t_0}) - \varphi^*)^{\frac{1}{\nu}}}.$$

Remark 6 *When f is nonconvex, the limit $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ in Theorem 5 can be interpreted as the function value at some stationary point of φ. Indeed, for any convergent subsequence $\{x_t\}_{t\in T}$ with limit x_* such that $\{\delta_t\}_{t\in T} \to 0$, it follows from (3) that*

$$\delta_t \ge \langle \nabla f(x_t), x_t - x \rangle + g(x_t) - g(x), \quad \forall x \in \mathbb{E}.$$

Taking the limit over $t \in T$, we can see that $x_* \in \mathrm{Argmin}_x \{\langle \nabla f(x_*), x\rangle + g(x)\}$. Thus, x_* is a stationary point of φ and moreover $\varphi_* = \varphi(x_*)$.

We next present some results regarding the rate of convergence of Algorithm 1 under the assumption that problem (1) has a uniformly convex structure, namely, Assumption 2 holds.

Theorem 7 *Let the sequences $\{x_t\}$ and $\{\delta_t\}$ be generated in Algorithm 1. Suppose that Assumptions 1 and 2 hold and that $\delta_t > 0$ for all $t \ge 0$. Let $\delta_t^* = \min_{0\le i\le t} \delta_i$ for all $t \ge 0$ and*

$$A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} M_\nu^{\frac{1}{\nu}}, \quad t_0 = \left\lceil \frac{1+\nu}{\nu} \left(\log \frac{(1+\nu)(\varphi(x_0) - \varphi^*)}{\nu A^{\frac{\rho\nu}{\rho-1-\nu}}} \right)_+ \right\rceil,$$
$$\overline{\gamma}_t = \begin{cases} (\varphi(x_0) - \varphi^*) \exp\left(-\frac{1}{2}\min\{1, \frac{\kappa}{2M_1}\} t\right) & \text{if } \nu = 1 \text{ and } \rho = 2, \\ \left[(\varphi(x_{t_0}) - \varphi^*)^{-\frac{\rho-1-\nu}{\rho\nu}} + \frac{\rho-1-\nu}{\rho(1+\nu)} A^{-1}(t - t_0)\right]^{-\frac{\rho\nu}{\rho-1-\nu}} & \text{otherwise.} \end{cases}$$

Then the following statements hold.

(i) $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ exists. In addition, $\{\delta_t^*\}$ satisfies

$$\delta_t^* \le \max\left\{ \frac{(1+\nu)(\varphi(x_0) - \varphi_*)}{\nu(t+1)}, \left(\frac{(1+\nu)A(\varphi(x_0) - \varphi_*)}{\nu(t+1)} \right)^{\frac{\rho\nu}{(\rho-1)(1+\nu)}} \right\}, \tag{13}$$

for all $t \ge 0$.

(ii) Assume additionally that f is convex.

(a) When $\nu = 1$ and $\rho = 2$, we have

$$\varphi(x_t) - \varphi^* \le \overline{\gamma}_t, \quad \forall t \ge 0,$$
$$\delta_t^* \le e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t+2)/2 \rfloor}, \quad \forall t \ge 4\max\left\{1, \frac{2M_1}{\kappa}\right\}.$$

Algorithm 2: A parameter-free conditional gradient method

Input: $x_0 \in \operatorname{dom} g$ and $L_{-1} > 0$.

for $t = 0, 1, 2, \ldots,$ **do**
  $v_t \in \operatorname{Argmin}_{x \in \mathbb{E}}\{\langle \nabla f(x_t), x\rangle + g(x)\}$.
  $\delta_t = \langle \nabla f(x_t), x_t\rangle + g(x_t) - \langle \nabla f(x_t), v_t\rangle - g(v_t)$.
  repeat for $i = 0, 1, 2, \ldots,$
    $L_t^{(i)} = 2^{i-1} L_{t-1}$.
    $\tau_t^{(i)} = \min\left\{1, \frac{\delta_t}{2L_t^{(i)}\|x_t - v_t\|^2}\right\}$.
    $x_{t+1}^{(i)} = (1 - \tau_t^{(i)})x_t + \tau_t^{(i)} v_t$.
  until

$$\varphi(x_{t+1}^{(i)}) \le \varphi(x_t) - \frac{1}{2}\tau_t^{(i)}\delta_t + \frac{1}{2}L_t^{(i)}(\tau_t^{(i)})^2 \|x_t - v_t\|^2. \tag{14}$$

  Set $(x_{t+1}, L_t, \tau_t) \leftarrow (x_{t+1}^{(i)}, L_t^{(i)}, \tau_t^{(i)})$.
end for

(b) When $\nu \neq 1$ or $\rho \neq 2$, we have

$$\varphi(x_t) - \varphi^* \le \overline{\gamma}_t, \quad \forall t \ge t_0,$$
$$\delta_t^* \le e^{\frac{1}{e}}\overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor}, \quad \forall t \ge t_0 + \frac{2(1+\nu)A}{\nu(\varphi(x_{t_0}) - \varphi^*)^{\frac{\rho\nu}{\rho-1-\nu}}}.$$

Remark 8 *It can be observed from Theorem 7 that under Assumptions 1 and 2, Algorithm 1 enjoys a linear rate of convergence when applied to problem* (1) *with f being convex, $\nu = 1$ and $\rho = 2$.*

4 A Parameter-Free Conditional Gradient Method

As seen from above, Algorithm 1 is not only simple but also enjoys a nice rate of convergence. However, its step size τ_t may suffer from some practical issues. Indeed, to evaluate τ_t, one needs to know the problem parameters ν and M_ν in advance. As observed from (6), these parameters depend on f, g, and also a particular norm on $\mathbb{E}$. Thus, it may not be easy to find them if f is a sophisticated function. On another hand, the parameters ν and M_ν are not unique. The tighter value of them typically leads to a faster convergent algorithm. Yet, it may be challenging to find the tightest possible value for them. Motivated by these, we next propose a parameter-free conditional gradient method (Algorithm 2) in which the step size is chosen by using a constructive local quadratic upper approximation and an adaptive line search scheme, without using prior knowledge of ν and M_ν.

We now provide some explanation for Algorithm 2. Observe that the step size τ_t in Algorithm 1 is the minimizer of $h_t(\tau) = \varphi(x_t) - \tau\delta_t + \tau^{1+\nu}\frac{M_\nu}{1+\nu}\|x_t - v_t\|^{1+\nu}$ over $[0, 1]$. Assuming that ν and M_ν are unknown, we can find a quadratic approximation to h_t, without explicitly involving ν and M_ν, and then obtain a step size by minimizing it over $[0, 1]$. Indeed, by the Hölder continuity of ∇f, the following inequality holds (e.g, see

Nesterov, 2015, Lemma 2):

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\varepsilon)}{2} \|x - y\|^2 + \varepsilon, \quad \forall x, y \in \operatorname{dom} g, \ \forall \varepsilon > 0, ^5 \tag{15}$$

where

$$L(\varepsilon) = \left(\frac{1-\nu}{1+\nu} \cdot \frac{1}{2\varepsilon} \right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \quad \forall \varepsilon > 0. \tag{16}$$

By (15) and a suitable choice of ε (see the proof of next theorem for details), one can obtain a quadratic approximation to h_t given by

$$\tilde{h}_t(\tau) = \varphi(x_t) - \frac{1}{2}\tau\delta_t + \frac{1}{2} L_t \tau^2 \|x_t - v_t\|^2$$

for some $L_t > 0$, which is determined by an adaptive line search scheme without explicitly using ν and M_ν. The step size τ_t in Algorithm 2 is then obtained by minimizing $\tilde{h}_t$ in place of h_t over $[0, 1]$. Therefore, Algorithm 2 does not explicitly use the parameters ν and M_ν.

The following theorem shows that Algorithm 2 is well-defined, whose proof is deferred to Section 6.3. In particular, we will establish that in each outer iteration of Algorithm 2 the adaptive line search procedure must terminate after a finite number of trials. We will also establish some other properties for the adaptive line search procedure.

Theorem 9 *Let the sequences $\{L_t\}$ and $\{\delta_t\}$ be generated in Algorithm 2. Suppose that Assumption 1 holds and that $\delta_t > 0$ for all $t \ge 0$.*[6] *Let*

$$\widetilde{L}_t = \max \left\{ L(\delta_t/2), L \left(\frac{\delta_t^2}{4 \|x_t - v_t\|^2} \right)^{\frac{1+\nu}{2\nu}} \right\} \tag{17}$$

for all $t \ge 0$, where $L(\cdot)$ is defined in (16)*. For any $\delta > 0$, let*

$$\overline{L}(\delta) = \begin{cases} \max \left\{ \left(\frac{1-\nu}{1+\nu} \frac{1}{\delta} \right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \quad \left(\frac{2(1-\nu)}{1+\nu} \right)^{\frac{1-\nu}{2\nu}} M_\nu^{\frac{1}{\nu}} \left(\frac{D_g}{\delta} \right)^{\frac{1-\nu}{\nu}} \right\} & \text{if } \operatorname{dom} g \text{ is bounded,} \\ \max \left\{ \left(\frac{1-\nu}{1+\nu} \frac{1}{\delta} \right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \quad \left(\frac{2(1-\nu)}{1+\nu} \right)^{\frac{1-\nu}{2\nu}} M_\nu^{\frac{1}{\nu}} \left(\frac{\rho}{\kappa \delta^{\rho-1}} \right)^{\frac{1-\nu}{\rho\nu}} \right\} & \text{if Assumption 2 holds.} \end{cases} \tag{18}$$

Then the following statements hold.

(i) The inequality (14) *holds whenever $L_t^{(i)} \ge \widetilde{L}_t$.*

(ii) $L_t \le 2 \max_{0 \le i \le t} \widetilde{L}_i$ holds for any $t \ge (\log_2(L_{-1}/\widetilde{L}_0))_+$.

(iii) Suppose further that $\min_{0 \le i \le t} \delta_i \ge \varepsilon$ for some $t \ge 0$ and $\varepsilon > 0$. Then the total number of inner iterations performed by the adaptive line search procedure until the t-th iteration of Algorithm 2 is bounded by $2t + 2 + [\log_2(2\overline{L}(\varepsilon)/L_{-1})]_+$.

5. By convention, we set $0^0 = 1$. One can observe that when $\nu = 1$, $L(\varepsilon)$ becomes M_ν and thus (15) still holds.

6. If $\delta_t = 0$ for some $t \ge 0$, x_t is already a stationary point of problem (1) and Algorithm 2 shall be terminated.

The next theorem establishes some results for the case where g has a bounded domain, namely, $D_g < +\infty$, whose proof is deferred to Section 6.3.

Theorem 10 *Let the sequences $\{x_t\}$ and $\{\delta_t\}$ be generated in Algorithm 2. Suppose that Assumption 1 holds, $D_g < +\infty$, and that $\delta_t > 0$ for all $t \geq 0$, where D_g is defined in (5). Let $\delta_t^* = \min_{0 \leq i \leq t} \delta_i$ for all $t \geq 0$ and*

$$A = (2M_\nu)^{\frac{1}{\nu}} D_g^{\frac{1+\nu}{\nu}}, \quad \tilde{t}_0 = \left\lceil \left(\log_2 \frac{L_{-1}}{\widetilde{L}_0} \right)_+ \right\rceil, \quad t_0 = \left\lceil 4 \left(\log \frac{4(\varphi(x_{\tilde{t}_0}) - \varphi^*)}{A^\nu} \right)_+ \right\rceil,$$
$$\overline{\gamma}_t = \left[(\varphi(x_{t_0+\tilde{t}_0}) - \varphi^*)^{-\frac{1}{\nu}} + (4\nu A)^{-1}(t - t_0) \right]^{-\nu}, \quad \forall t \geq t_0,$$

where $\widetilde{L}_0$ is defined in (17). Then the following statements hold.

(i) $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ exists. In addition, $\{\delta_t^*\}$ satisfies

$$\delta_t^* \leq \max \left\{ \frac{4(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{t + 1 - \tilde{t}_0}, \left(\frac{4A(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{t + 1 - \tilde{t}_0} \right)^{\frac{\nu}{1+\nu}} \right\}, \quad \forall t \geq \tilde{t}_0. \tag{19}$$

(ii) Assume additionally that f is convex. Then we have

$$\varphi(x_t) - \varphi^* \leq \overline{\gamma}_{t-\tilde{t}_0}, \quad \forall t \geq \tilde{t}_0 + t_0,$$
$$\delta_t^* \leq e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t-\tilde{t}_0+t_0+1)/2 \rfloor}, \quad \forall t \geq \tilde{t}_0 + t_0 + \frac{8A}{(\varphi(x_{\tilde{t}_0+t_0}) - \varphi^*)^{\frac{1}{\nu}}}.$$

As an immediate consequence of Theorem 10, we obtain the following complexity results for Algorithm 2 for finding an approximate solution of problem (1) with an ε-Frank-Wolfe gap, whose proofs are omitted.

Corollary 11 *Under the same settings as in Theorem 10, Algorithm 2 reaches the criterion $\delta_t \leq \varepsilon$ within*

$$\tilde{t}_0 + \frac{4(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{\varepsilon} \max \left\{ 1, \left(\frac{2M_\nu D_g^{1+\nu}}{\varepsilon} \right)^{\frac{1}{\nu}} \right\}$$

iterations. Furthermore, if f is convex, it reaches the criterion $\delta_t \leq \varepsilon$ within

$$\tilde{t}_0 + t_0 + 8 \left(\frac{2M_\nu D_g^{1+\nu}}{\varphi(x_{\tilde{t}_0+t_0}) - \varphi^*} \right)^{\frac{1}{\nu}} \max \left\{ 1, \nu \left[\left(\frac{e^{\frac{1}{e}}(\varphi(x_{\tilde{t}_0+t_0}) - \varphi^*)}{\varepsilon} \right)^{\frac{1}{\nu}} - 1 \right] \right\}$$

iterations.

In what follows, we present some results regarding the rate of convergence of Algorithm 2 for the case where problem (1) has a uniformly convex structure, namely, Assumption 2 holds, whose proof is deferred to Section 6.3.

Theorem 12 *Let the sequences $\{x_t\}$ and $\{\delta_t\}$ be generated by Algorithm 2. Suppose that Assumptions 1 and 2 hold and that $\delta_t > 0$ for all $t \geq 0$. Let $\delta_t^* = \min_{0\leq i\leq t} \delta_i$ for all $t \geq 0$, and*

$$A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} (2M_\nu)^{\frac{1}{\nu}}, \quad \tilde{t}_0 = \left\lceil \left(\log_2 \frac{L_{-1}}{\widetilde{L}_0}\right)_+ \right\rceil, \quad t_0 = \left\lceil 4\left(\log \frac{4(\varphi(x_{\tilde{t}_0}) - \varphi^*)}{A^{\frac{\rho\nu}{\rho-1-\nu}}}\right)_+ \right\rceil,$$

$$\overline{\gamma}_t = \begin{cases} (\varphi(x_{\tilde{t}_0}) - \varphi^*) \exp\left(-\frac{1}{4}\min\{1, \frac{\kappa}{4M_1}\}t\right) & \text{if } \nu = 1 \text{ and } \rho = 2, \\ \left[(\varphi(x_{\tilde{t}_0+t_0}) - \varphi^*)^{-\frac{\rho-1-\nu}{\rho\nu}} + \frac{\rho-1-\nu}{4\rho\nu} A^{-1}(t - t_0)\right]^{-\frac{\rho\nu}{\rho-1-\nu}} & \text{otherwise,} \end{cases}$$

where $\widetilde{L}_0$ is defined in (17)*. Then the following statements hold.*

(i) $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ exists. In addition, $\{\delta_t^*\}$ satisfies

$$\delta_t^* \leq \max\left\{\frac{4(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{t + 1 - \tilde{t}_0}, \left(\frac{4A(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{t + 1 - \tilde{t}_0}\right)^{\frac{\rho\nu}{(\rho-1)(1+\nu)}}\right\}, \tag{20}$$

for all $t \geq \tilde{t}_0$.

(ii) Assume additionally that f is convex.

(a) When $\nu = 1$ and $\rho = 2$, we have

$$\varphi(x_t) - \varphi^* \leq \overline{\gamma}_{t-\tilde{t}_0}, \quad \forall t \geq \tilde{t}_0,$$
$$\delta_t^* \leq e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t-\tilde{t}_0+2)/2 \rfloor}, \quad \forall t \geq \tilde{t}_0 + 4\max\left\{1, \frac{2M_1}{\kappa}\right\}.$$

(b) When $\nu \neq 1$ or $\rho \neq 2$, we have

$$\varphi(x_t) - \varphi^* \leq \overline{\gamma}_{t-\tilde{t}_0}, \quad \forall t \geq \tilde{t}_0 + t_0,$$
$$\delta_t^* \leq e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t-\tilde{t}_0+t_0+1)/2 \rfloor}, \quad \forall t \geq \tilde{t}_0 + t_0 + \frac{8A}{(\varphi(x_{\tilde{t}_0+t_0}) - \varphi^*)^{\frac{\rho-1-\nu}{\rho\nu}}}.$$

As an immediate consequence of Theorem 12, we obtain the following complexity results for Algorithm 2 for finding an approximate solution of problem (1) with an ε-Frank-Wolfe gap, whose proofs are omitted.

Corollary 13 *Under the same settings as in Theorem 12, Algorithm 2 reaches the criterion $\delta_t \leq \varepsilon$ within*

$$\tilde{t}_0 + \frac{4(\varphi(x_{\tilde{t}_0}) - \varphi_*)}{\varepsilon} \max\left\{1, \frac{\rho^{\frac{1+\nu}{\rho\nu}} (2M_\nu)^{\frac{1}{\nu}}}{\kappa^{\frac{1+\nu}{\rho\nu}} \varepsilon^{\frac{\rho-1-\nu}{\rho\nu}}}\right\}$$

iterations. Furthermore, if f is convex, $\nu = 1$ and $\rho = 2$, Algorithm 2 reaches the criterion $\delta_t \leq \varepsilon$ within

$$\tilde{t}_0 + 8\max\left\{1, \frac{4M_1}{\kappa}\right\} \max\left\{1, \log \frac{\varphi(x_{\tilde{t}_0}) - \varphi^*}{\varepsilon}\right\} \tag{21}$$

iterations. In addition, if f is convex, and $\nu \neq 1$ or $\rho \neq 2$, Algorithm 2 reaches the criterion $\delta_t \leq \varepsilon$ within

$$\tilde{t}_0+t_0+\frac{8\rho^{\frac{1+\nu}{\rho\nu}}(2M_\nu)^{\frac{1}{\nu}}}{\kappa^{\frac{1+\nu}{\rho\nu}}\left(\varphi(x_{\tilde{t}_0+t_0})-\varphi^*\right)^{\frac{\rho-1-\nu}{\rho\nu}}}\max\left\{1,\frac{\rho\nu}{\rho-1-\nu}\left[\left(\frac{e^{\frac{1}{e}}(\varphi(x_{\tilde{t}_0+t_0})-\varphi^*)}{\varepsilon}\right)^{\frac{\rho-1-\nu}{\rho\nu}}-1\right]\right\} \tag{22}$$

iterations.

Remark 14 *In view of the identity $\lim_{\alpha\to 0}\frac{1}{\alpha}(x^\alpha-1)=\log x$, one can observe that the limit of (22) as $\nu\to 1$ and $\rho\to 2$ is*

$$O\left(\frac{M_1}{\kappa}\max\left\{1,\log\frac{\varphi(x_{\tilde{t}_0})-\varphi^*}{\varepsilon}\right\}\right),$$

which is consistent with the bound (21) for the case with $\nu=1$ and $\rho=2$.

4.1 Iteration Complexity

As mentioned earlier, the Frank-Wolfe gap δ_t defined in (3) is a computable quantity and can be used to measure whether the associated iterate x_t is an approximate stationary point of problem (1). Therefore, $\delta_t \leq \varepsilon$ can be used as a practical termination criterion for Algorithms 1 and 2 for a prescribed tolerance $\varepsilon > 0$. Besides, one can observe from Theorems 5, 7, 10 and 12 that Algorithms 1 and 2 enjoy the same rate of convergence and thus the same iteration complexity with respect to the termination criterion $\delta_t \leq \varepsilon$. Consequently, it suffices to discuss the iteration complexity of Algorithm 2 with such a termination criterion.

One can observe from Corollaries 11 and 13 that the iteration complexity of Algorithm 2 for reaching the termination criterion $\delta_t \leq \varepsilon$ is:

(i) $O(\varepsilon^{-1-1/\nu})$ if f is nonconvex and $\operatorname{dom} g$ is bounded;

(ii) $O(\varepsilon^{-1-(\rho-1-\nu)/(\rho\nu)})$ if f is nonconvex and problem (1) has a uniformly convex structure;

(iii) $O(\varepsilon^{-1/\nu})$ if f is convex and $\operatorname{dom} g$ is bounded;

(iv) $O(\log(1/\varepsilon))$ if f is convex and problem (1) has a uniformly convex structure with $\nu = 1$ and $\rho = 2$;

(v) $O(\varepsilon^{-(\rho-1-\nu)/(\rho\nu)})$ if f is convex and problem (1) has a uniformly convex structure with $\nu \neq 1$ or $\rho \neq 2$.

Since $\rho \geq 2$ and $\nu \in (0, 1]$, one has $\varepsilon^{-(\rho-1-\nu)/(\rho\nu)} < \varepsilon^{-1/\nu}$ and $\varepsilon^{-1-(\rho-1-\nu)/(\rho\nu)} < \varepsilon^{-1-1/\nu}$ when $\nu \neq 1$ or $\rho \neq 2$. In view of this and the above complexity results, we can observe that Algorithm 2 enjoys a lower iteration complexity bound under Assumption 2 than the one under the assumption that $\operatorname{dom} g$ is bounded. Besides, the iteration complexity bound in (iii) matches the ones obtained in (Nesterov, 2015; Zhao and Freund, 2020). It should, however, be noted that the conditional gradient methods in (Nesterov, 2015; Zhao and

Freund, 2020) use the step size $\tau_t = 2/(t+2)$ and the same one as given in Algorithm 1, respectively. These step sizes are not locally adaptive because they use none of local or global problem information, and could be conservative in practice. Moreover, the latter one requires prior knowledge of the parameters ν and M_ν. In contrast with them, the step size in Algorithm 2 is locally adaptive and free of problem parameters. In addition, the iteration complexity bounds in (ii) and (iv) match the ones obtained in (Ghadimi, 2019) for the case with $\rho = 2$. The iteration complexity bound in (v) improves by the factor $\log(1/\varepsilon)$ the one established in (Ghadimi, 2019) for the case with $\nu < 1$ and $\rho = 2$. For a smooth convex f with $\nu = 1$ and g being the indicator function of a uniformly convex set, similar iteration complexity bounds as in (iv) and (v) with $\nu = 1$ were established in (Kerdreux et al., 2021a) for a parameter-dependent conditional gradient method for reaching the criterion $\varphi(x_t) - \varphi^* \leq \varepsilon$.

In addition, as observed from Theorem 9 (iii), the total number of inner iterations of Algorithm 2 for reaching the termination criterion $\delta_t \leq \varepsilon$ is at most $2t + [\log_2(2\overline{L}(\varepsilon)/L_{-1})]_+$. Also, notice from (18) that $\log \overline{L}(\varepsilon) = O(\log(1/\varepsilon))$. In view of these, one can see that the total number of inner iterations of Algorithm 2 enjoys the same complexity bounds as given in (i)-(v) for reaching the termination criterion $\delta_t \leq \varepsilon$.

5 Numerical Experiments

In this section we conduct some numerical experiments to compare the performance of the conditional gradient methods studied in this paper with the ones with some other step size rules. For the comparison, we construct the problems whose Hölder continuity exponent ν and uniform convexity exponent ρ are known in advance. More specifically, we generate the test instances from the problem classes discussed in Examples 1 and 2, respectively. Our experiments are conducted in Matlab on an Apple desktop with the 3.0GHz Intel Xeon E5-1680v2 processor and 64GB of RAM.

5.1 ℓ_p-Norm Minimization over ℓ_q Ball

In our first experiment, we consider the following problem:

$$\begin{array}{ll} \min & \frac{1}{p}\|Ax - b\|_p^p \\ \text{s.t.} & x \in \mathcal{B}_q := \{z \in \mathbb{R}^n : \|z\|_q \leq 1\}, \end{array} \tag{23}$$

where $1 < p \leq 2$, $q > 1$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Note that $\mathcal{B}_q$ is a uniformly convex set with exponent $\rho = \max(2, q)$ (e.g., see Kerdreux et al., 2021a,b). In addition, as shown in Lemma 17 in Section 6.1, the gradient of the objective function of (23) is Hölder continuous with respect to $\|\cdot\|_2$ with exponent $\nu = p - 1$ and modulus $M_\nu = 2^{2-p} m^{\frac{(p-1)(2-p)}{2p}} \|A\|_2^p$. Thus, problem (23) belongs to the class of the problems minimizing a weakly smooth convex function over a uniformly convex set discussed in Example 1. Note that when $A = I$, it reduces to the problem of the ℓ_p-norm projection of a vector onto the ℓ_q unit ball.

We next apply the following three conditional gradient methods to solve problem (23), and compare their performance.

- Algorithm 1 with $\|\cdot\| = \|\cdot\|_2$, $\nu = p - 1$, and $M_\nu = 2^{2-p} m^{\frac{(p-1)(2-p)}{2p}} \|A\|_2^p$.

- Algorithm 2 with $\|\cdot\| = \|\cdot\|_2$.
- The conditional gradient method with the well-known diminishing step size $\tau_t = 2/(t+2)$ (Jaggi, 2013; Freund and Grigas, 2016), which is similar to Algorithm 1 except the choice of τ_t.

As discussed in Section 4.1, for finding an approximate solution of (23) with an ε-Frank-Wolfe gap, the conditional gradient method with step size $\tau_t = 2/(t+2)$ enjoys an iteration complexity of $O(\varepsilon^{-1/\nu})$, while Algorithms 1 and 2 enjoy the following iteration complexity:

$$\begin{cases} O(\log(1/\varepsilon)) & \text{if } p = 2 \text{ and } q \le 2, \\ O(\varepsilon^{-(\rho-1-\nu)/(\rho\nu)}) & \text{otherwise,} \end{cases} \qquad \text{with } \rho = \max(2, q) \text{ and } \nu = p - 1.$$

When applied to (23), the above three methods need to solve the subproblems of the form

$$\min_x \{\langle u, x\rangle : \|x\|_q \le 1\}$$

for $u \in \mathbb{R}^n$. It is not hard to observe that this problem has a closed-form solution given by

$$x_i^* = -\|u\|_q^{-\frac{1}{q-1}} \operatorname{sign}(u_i)|u_i|^{\frac{1}{q-1}}, \quad i = 1, \ldots, n.$$

The instances of problem (23) are generated as follows. In particular, we generate matrix A by letting $A = UDU^T$, where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix, whose diagonal entries are randomly generated according to the uniform distribution over $[1, 100]$ and $U \in \mathbb{R}^{n\times n}$ is a randomly generated orthogonal matrix. We set $b = A\bar{x}$ for some $\bar{x}$ generated from a uniform distribution over $\{x \in \mathbb{R}^n : \|x\|_q = 10\}$.

In this experiment, we consider $p \in \{1.3, 1.6, 2\}$, $q \in \{1.5, 2, 3\}$ and $m = n \in \{1000, 5000\}$. For each choice of (p, q, n), we randomly generate 10 instances of problem (23) by the procedure mentioned above, and apply the aforementioned three conditional gradient methods to solve them, starting with the initial point $x_0 = 0$ and terminating them once the criterion $\delta_t/\delta_0 \le 10^{-6}$ is met, where δ_t and δ_0 are the Frank-Wolfe gap at the iterates x_t and x_0, respectively. Table 1 presents the average CPU time (in seconds) and the average number of iterations of these methods over the 10 random instances. In detail, the values of n, q, p are given in the first three columns, and the average CPU time and the average number of iterations of Algorithms 1, 2 and the conditional gradient method with step size $\tau_t = 2/(t+2)$ are given in the rest of the columns. In addition, Figure 1 illustrates the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0\le i\le t} \delta_i/\delta_0$ and the objective value gap $\varphi(x_t) - \widetilde{\varphi}_*$ with respect to CPU time on a single random instance of problem (23) with $n = 5000$, $q = 3$, and $p = 1.3$, 1.6, 2, respectively, where $\widetilde{\varphi}_*$ is the minimum objective function value of all iterates generated by the three algorithms. One can see that Algorithm 2 generally outperforms the other two methods. This is perhaps because: (i) Algorithm 2 improves the iteration complexity of the conditional gradient method with step size $\tau_t = 2/(t+2)$; (ii) Algorithm 2 uses an adaptive step size determined by using a constructive local quadratic upper approximation of the objective function and an adaptive line search scheme.

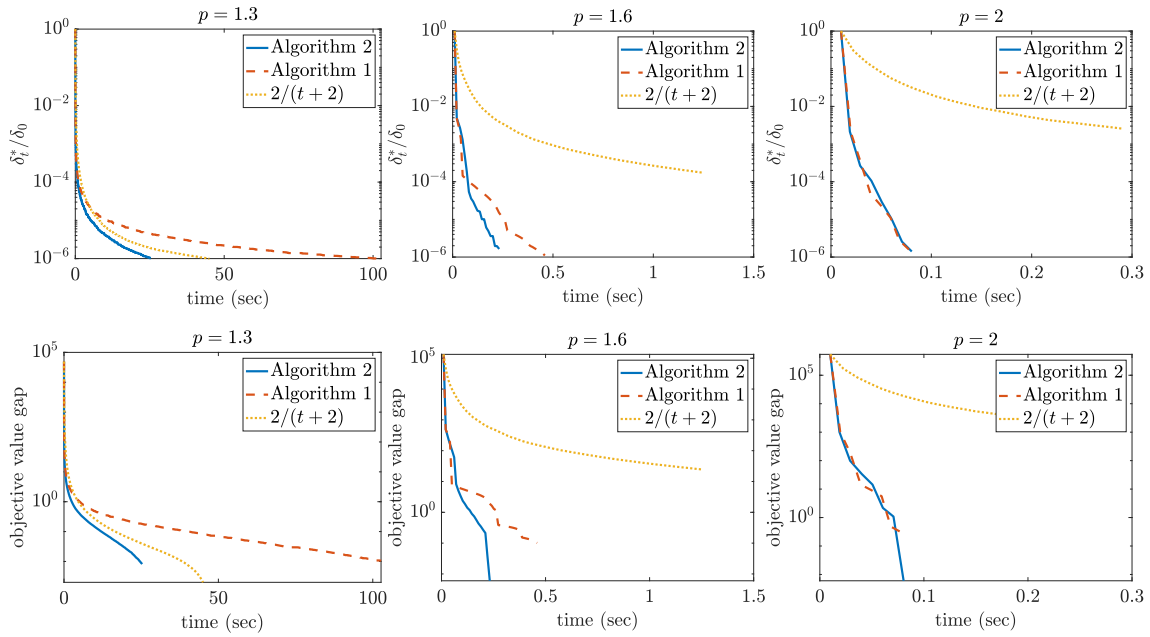

Figure 1: Numerical results on a single random instance of problem (23) with $n = 5000$, $q = 3$, and $p = 1.3$, 1.6, 2, respectively. These sub-figures illustrate the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0\le i\le t} \delta_i/\delta_0$ and the objective value gap $\varphi(x_t) - \widetilde{\varphi}_*$ with respect to CPU time in seconds, where $\widetilde{\varphi}_*$ is the minimum objective function value of all iterates generated by the three algorithms for solving one problem instance.

			Average CPU time (sec)			Average number of iterations		
n	q	p	Algorithm 1	Algorithm 2	$\frac{2}{t+2}$	Algorithm 1	Algorithm 2	$\frac{2}{t+2}$
1000	1.5	1.3	3.51	0.054	0.56	8881.4	84.9	1404.5
		1.6	0.0065	0.0055	0.52	13.5	6.2	1333.5
		2.0	0.0028	0.0037	0.44	5.0	6.2	1287.1
	2.0	1.3	1.63	0.13	0.50	4901.2	252.5	1544.4
		1.6	0.0054	0.0060	0.44	13.8	6.9	1335.1
		2.0	0.0020	0.0022	0.37	4.0	4.0	1299.4
	3.0	1.3	10.9	1.18	1.76	27442.3	2038.1	4449.2
		1.6	0.028	0.012	0.52	68.4	18.5	1323.1
		2.0	0.0036	0.0038	0.45	7.7	7.4	1289.8
5000	1.5	1.3	20.6	1.34	13.3	2223.8	132.5	1424.4
		1.6	0.063	0.078	12.5	5.7	6.2	1334.6
		2.0	0.056	0.067	12.0	5.0	6.2	1288.0
	2.0	1.3	7.03	3.23	13.2	809.5	341.3	1506.4
		1.6	0.10	0.083	11.6	10.5	7.2	1335.8
		2.0	0.044	0.043	11.2	4.0	4.0	1300.1
	3.0	1.3	161.2	28.4	37.3	17364.6	2827.7	3972.8
		1.6	0.41	0.20	12.4	43.7	18.7	1323.4
		2.0	0.084	0.083	12.0	7.8	7.8	1289.9

Table 1: Numerical results for problem (23)

5.2 Entropy Regularized ℓ_p-Norm Minimization

In our second experiment, we consider the following problem:

$$\begin{array}{ll} \min & \frac{1}{p}\|Ax-b\|_p^p + \lambda \sum_{i=1}^n x_i \log x_i \\ \text{s.t.} & x \in \Delta_n := \{z \in \mathbb{R}^n : \sum_{i=1}^n z_i = 1,\ z_i \geq 0,\ i = 1, \ldots, n\}, \end{array} \tag{24}$$

where $p > 1$, $\lambda > 0$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Note that $f(x) = \frac{1}{p}\|Ax-b\|_p^p$ is weakly smooth and $g(x) = \lambda \sum_{i=1}^n x_i \log x_i + \iota_{\Delta_n}(x)$ is λ-strongly convex with respect to the ℓ_1-norm (e.g., see Beck and Teboulle, 2003), where ι_{Δ_n} denotes the indicator function of Δ_n. Thus, problem (24) is a special case of Example 2 with $\nu = p-1$ and $\rho = 2$.

We next apply Algorithms 1 and 2 with the same settings as in Subsection 5.1 and the conditional gradient method with the diminishing step size

$$\tau_t = \frac{6(t+1)}{(t+2)(2t+3)} \tag{25}$$

proposed by Nesterov (2018) to solve problem (24), and compare their performance. The latter method is similar to Algorithm 1 except the choice of τ_t and enjoys an iteration complexity of $O(\varepsilon^{-1/(2(p-1))})$ for finding an approximate solution of (24) with an ε-Frank-Wolfe gap (Nesterov, 2018). In addition, as seen from Section 4.1, Algorithms 1 and 2 enjoy the following iteration complexity for finding an approximate solution of (24) with

an ε-Frank-Wolfe gap:

$$\begin{cases} O\left(\frac{\|A\|_2^2}{\lambda}\log\frac{1}{\varepsilon}\right) & \text{if } p=2, \\ O(\varepsilon^{-(2-p)/(2(p-1))}) & \text{otherwise.} \end{cases}$$

When applied to (24), the aforementioned three methods need to solve the subproblems of the form

$$\min_{x\in\Delta_n} \langle u, x\rangle + \lambda\sum_{i=1}^n x_i\log x_i$$

for some $u\in\mathbb{R}^n$. It is well-known that this problem has a closed-form solution given by

$$x_i^* = \frac{e^{-u_i/\lambda}}{\sum_{j=1}^n e^{-u_j/\lambda}}, \quad i=1,\ldots,n.$$

The instances of problem (24) are generated as follows. In particular, we generate matrix A with $\|A\|_2\le 100$ by letting $A=VDU^T$, where D is a $m\times m$ diagonal matrix, whose diagonal entries are randomly generated according to the uniform distribution over $[0,100]$, and $U\in\mathbb{R}^{n\times m}$ and $V\in\mathbb{R}^{m\times m}$ are randomly generated orthonormal matrices. Each entry of $b\in\mathbb{R}^m$ is generated from the uniform distribution on $[0,1]$.

In this experiment, we consider $m=n/2\in\{1000,5000\}$, $p\in\{1.5,1.75,2\}$, and $\lambda\in\{1,10,50\}$. For each choice of (m,n,p,λ), we randomly generate 10 instances of problem (24) by the procedure mentioned above, and apply the aforementioned three conditional gradient methods to solve them, starting with the initial point $x_0=(1/n,\ldots,1/n)^T$ and terminating them once the criterion $\delta_t/\delta_0\le 10^{-8}$ is met, where δ_t and δ_0 are the Frank-Wolfe gap at the iterates x_t and x_0, respectively. Table 2 presents the average CPU time (in seconds) and the average number of iterations of these methods over the 10 random instances. In detail, the values of m,p,λ are given in the first three columns, and the average CPU time and the average number of iterations of Algorithms 1, 2 and the conditional gradient method with step size $\tau_t=6(t+1)/((t+2)(2t+3))$ are given in the rest of the columns. In addition, Figure 2 illustrates the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0\le i\le t}\delta_i/\delta_0$ and the objective value gap $\varphi(x_t)-\widetilde{\varphi}_*$ with respect to CPU time on a single random instance of problem (24) with $m=5000$, $n=10{,}000$, $\lambda=10$, and $p=1.5,1.75,2$, respectively, where $\widetilde{\varphi}_*$ is the minimum objective function value of all iterates generated by the three algorithms. One can see that Algorithm 2 generally outperforms the other two methods, which is perhaps for the similar reasons as explained at the end of Subsection 5.1.

5.3 Simplex-Constrained Nonnegative Matrix Factorization

In our third experiment, we consider the following simplex-constrained nonnegative matrix factorization problem (see Thanh et al., 2022):

$$\begin{array}{ll} \min & \frac{1}{2}\|X-UV\|_F^2 + \lambda(\|U\|_F^2+\|V\|_F^2) \\ \text{s.t.} & U\in B_{n,k} := \{U\in\mathbb{R}^{n\times k} \mid 0\le U_{ij}\le\alpha,\ 1\le i\le n,\ 1\le j\le k\}, \\ & V\in\Delta_{k,m} := \{V\in\mathbb{R}_+^{k\times m} \mid V^T1_k = 1_m\}, \end{array} \tag{26}$$

			Average CPU time (sec)			Average number of iterations		
m	p	λ	Algorithm 1	Algorithm 2	Rule (25)	Algorithm 1	Algorithm 2	Rule (25)
1000	1.5	1	314.3	0.47	0.49	640908.6	607.5	982.4
		10	33.4	0.030	0.046	67184.7	34.1	87.4
		50	2.60	0.010	0.016	5186.6	8.1	28.1
	1.75	1	0.77	0.17	0.39	1446.4	210.8	731.3
		10	0.21	0.020	0.048	404.3	21.4	85.0
		50	0.015	0.0056	0.016	27.6	5.0	28.0
	2	1	0.12	0.14	0.32	236.8	237.1	647.0
		10	0.021	0.015	0.050	40.2	24.0	94.7
		50	0.0035	0.0047	0.015	5.0	5.1	27.4
5000	1.5	1	8300.7	5.23	8.99	459440.4	272.6	518.4
		10	1943.9	0.33	1.77	107117.5	16.3	100.4
		50	16.5	0.10	0.55	952.0	4.3	31.0
	1.75	1	30.9	2.77	8.99	1716.2	143.0	513.6
		10	6.28	0.30	1.86	353.9	14.7	106.3
		50	0.16	0.071	0.56	8.3	3.0	30.6
	2	1	4.00	2.67	8.99	230.7	150.7	513.7
		10	0.78	0.33	2.11	44.8	17.5	119.2
		50	0.071	0.070	0.55	3.0	3.0	31.0

Table 2: Numerical results for problem (24)

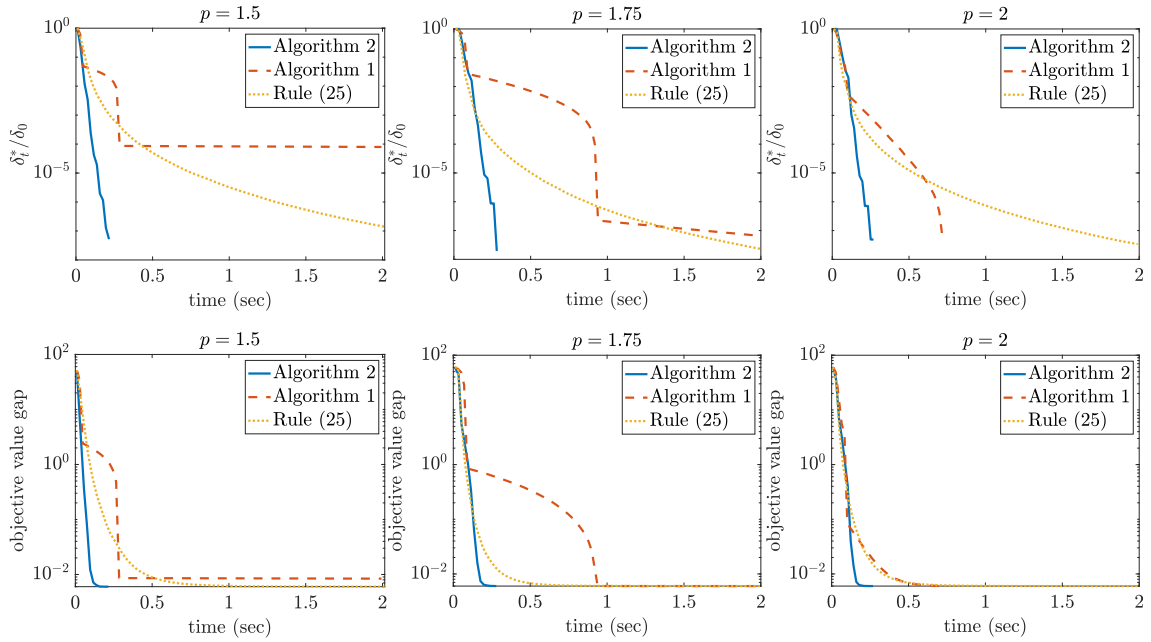

Figure 2: Numerical results on a single random instance of problem (24) with $m = 5000$, $n = 10{,}000$, $\lambda = 10$, and $p = 1.5, 1.75, 2$, respectively. These sub-figures illustrate the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0\le i\le t} \delta_i/\delta_0$ and the objective function value gap $\varphi(x_t) - \widetilde{\varphi}_*$ with respect to CPU time in seconds. Here, $\widetilde{\varphi}_*$ denotes the minimum objective function value of all iterates generated by the three algorithms for solving one problem instance.

where $X \in \mathbb{R}^{n\times m}$, $\alpha, \lambda > 0$, $\|\cdot\|_F$ is the Frobenius norm, and $1_d \in \mathbb{R}^d$ is the all-ones vector for any $d \geq 1$. Problem (26) can be viewed as $\min_{U,V}\{\varphi(U,V) := f(U,V) + g(U,V)\}$ with

$$f(U,V) = \frac{1}{2}\|X - UV\|_F^2, \quad g(U,V) = \lambda(\|U\|_F^2 + \|V\|_F^2) + \iota_{B_{n,k}}(U) + \iota_{\Delta_{k,m}}(V),$$

where $\iota_{B_{n,k}}$ and $\iota_{\Delta_{k,m}}$ denote the indicator function of $B_{n,k}$ and $\Delta_{k,m}$, respectively. Notice that f is nonconvex and smooth, $\operatorname{dom} g$ is compact, and g is strongly convex. Clearly, problem (26) is a special case of problem (1) satisfying Assumptions 1 and 2 with $\nu = 1$ and $\rho = 2$.

We next apply the following two conditional gradient methods to solve problem (26), and compare their performance.

- Algorithm 2 with $\|\cdot\| = \|\cdot\|_F$.
- The conditional gradient method with line search (Ghadimi, 2019, Algorithm 2), abbreviated as CGM-LS. We set the parameters $\gamma = 0.5$ and $\delta = \varepsilon\delta_0/4$ for this method as suggested in (Ghadimi, 2019), where δ_0 is the Frank-Wolfe gap at the initial point and ε is the targeted tolerance for the final relative Frank-Wolfe gap.

It shall be mentioned that these two methods enjoy an iteration complexity of $O(1/\varepsilon)$ for finding an approximate solution of (26) with an ε-Frank-Wolfe gap (see Section 4.1 and equation (1.8) in (Ghadimi, 2019)).

The data matrix $X \in \mathbb{R}^{n\times m}$ for problem (26) is generated as follows. In particular, we first randomly generate $U^* \in \mathbb{R}^{n\times k}$ with all entries following the uniform distribution over $[0, \alpha]$. We next randomly generate $\widetilde{V} \in \mathbb{R}^{k\times m}$ with all entries following the standard normal distribution and set $V^* = \widetilde{V}D$, where $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix such that $(V^*)^T 1_k = 1_m$. Finally, we set $X = U^*V^* + E$, where the entries of $E \in \mathbb{R}^{n\times m}$ follow the normal distribution with mean zero and standard deviation 0.01.

In this experiment, we set $\lambda = 0.01$, $\alpha = 2$, and consider $m = n \in \{100, 200, 300, 400, 500\}$ and $k \in \{5, 10\}$. For each choice of (m, n, k), we randomly generate 10 instances of problem (26) by the procedure mentioned above. Then we apply the aforementioned two conditional gradient methods to solve them with the initial point U^0 and V^0 being the matrices of all entries equal to 1 and $1/k$, respectively, and terminate the methods once the criterion $\delta_t/\delta_0 \leq 10^{-5}$ is met, where δ_t is the Frank-Wolfe gap at the t-th iteration (U^t, V^t). The computational results are presented in Table 3. In particular, the values of m and k are given in the first two columns, and the average CPU time (in seconds) and the average number of iterations over each set of 10 random instances for these methods are given in the rest of the columns. Besides, in Figure 3 we illustrate the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0\leq i\leq t}\delta_i/\delta_0$ and the relative objective function value $\varphi(U^t, V^t)/\varphi(U^0, V^0)$ with respect to CPU time on a single random instance of problem (24) with $m = n = 300$, $\lambda = 0.01$, $\alpha = 2$, and $k = 5, 10$, respectively.

One can observe that Algorithm 2 significantly outperforms the conditional gradient method with line search proposed in Ghadimi (2019). This is perhaps because: (i) the line search criterion of the conditional gradient method in Ghadimi (2019) explicitly depends on the targeted accuracy ε, while the line search criterion of Algorithm 2 does not; (ii) at each iteration, the initial trial step size in Algorithm 2 is determined by using a constructive local

quadratic upper approximation of the objective function, while the line search procedure in Ghadimi (2019) does not use such a novel scheme.

6 Proof of the Main Results

In this section, we provide a proof of our main results presented in Sections 3 and 4.

6.1 Auxiliary Lemmas

In this subsection we establish some technical lemmas that will be used subsequently.

Lemma 15 *Suppose that $\{\beta_t\}$ and $\{\gamma_t\}$ are sequences of nonnegative real numbers such that*

$$\gamma_{t+1} \leq \gamma_t - c\beta_t \min\{1, \beta_t^\alpha / A\}, \quad \forall t \geq 0 \tag{27}$$

for some constants $c \in (0,1)$, $\alpha \geq 0$ and $A > 0$. Then, $\underline{\gamma} = \lim_{t\to\infty} \gamma_t$ exists and the sequence $\beta_t^* = \min_{0\leq i\leq t} \beta_i$ satisfies

$$\beta_t^* \leq \max\left\{ \frac{\gamma_0 - \underline{\gamma}}{c(t+1)}, \ \left(\frac{A(\gamma_0 - \underline{\gamma})}{c(t+1)} \right)^{\frac{1}{1+\alpha}} \right\}, \quad \forall t \geq 0.$$

In particular, we have $\beta_t^* \leq \varepsilon$ whenever

$$t \geq \frac{\gamma_0 - \underline{\gamma}}{c\varepsilon} \max\left\{1, \frac{A}{\varepsilon^\alpha}\right\}.$$

Proof Since $\{\beta_t\} \subset \mathbb{R}_+$, the relation (27) implies that $\{\gamma_t\}$ is non-increasing, which together with $\{\gamma_t\} \subset \mathbb{R}_+$ further implies that $\underline{\gamma} = \lim_{t\to\infty} \gamma_t$ exists. In addition, by $\beta_t^* = \min_{0\leq i\leq t} \beta_i$ and (27), one can obtain that

$$\gamma_{t+1} \leq \gamma_t - c\beta_t^* \min\{1, (\beta_t^*)^\alpha / A\}, \quad \forall t \geq 0. \tag{28}$$

Summing up these inequalities yields

$$\gamma_{t+1} \leq \gamma_0 - (t+1)c\beta_t^* \min\{1, (\beta_t^*)^\alpha / A\}, \quad \forall t \geq 0.$$

By this and $\gamma_{t+1} \geq \underline{\gamma}$, we have $\beta_t^* \min\{1, (\beta_t^*)^\alpha / A\} \leq (\gamma_0 - \underline{\gamma})/(c(t+1))$, which implies the desired assertions. ■

Lemma 16 *Suppose that $\{\beta_t\}$ and $\{\gamma_t\}$ are sequences of nonnegative real numbers such that the recurrence (27) holds for some constants $c \in (0,1)$, $\alpha \geq 0$ and $A > 0$. Assume additionally that $\beta_t \geq \gamma_t$ for all $t \geq 0$. Let $\beta_t^* = \min_{0\leq i\leq t} \beta_i$. Then the following statements hold.*

(i) If $\alpha = 0$, then we have $\gamma_t \leq \overline{\gamma}_t$ for all $t \geq 0$ and $\beta_t^* \leq \overline{\gamma}_{\lfloor (t+2)/2 \rfloor}$ for all $t \geq 2\max\{1, A\}/c$, where

$$\overline{\gamma}_t = \gamma_0 \exp(-c \min\{1, A^{-1}\}\, t).$$

Consequently, we have $\beta_t^* \leq \varepsilon$ whenever

$$t \geq \frac{2}{c} \max\{1, A\} \max\left\{1, \log \frac{\gamma_0}{\varepsilon}\right\}.$$

m	k	Average CPU time (sec) Algorithm 2	CGM-LS	Average number of iterations Algorithm 2	CGM-LS
100	5	0.71	6.17	172.4	1139.1
	10	0.34	2.62	55.7	355.1
200	5	2.26	17.09	329.8	2086.4
	10	1.35	9.97	124.9	803.7
300	5	6.41	49.24	624.1	4035.1
	10	3.36	30.67	194.3	1398.2
400	5	9.89	62.94	796.0	4208.8
	10	4.87	41.01	230.3	1474.6
500	5	14.74	179.24	904.3	7131.6
	10	6.35	79.91	263.3	2588.0

Table 3: Numerical results for problem (26)

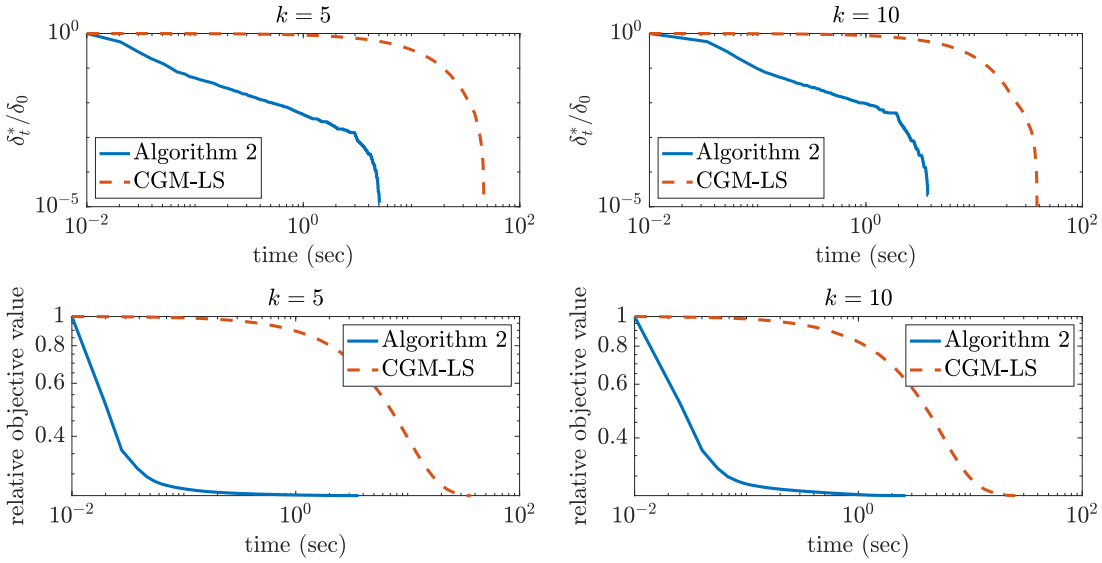

Figure 3: Numerical results on a single random instance of problem (26) with $m = n = 300$, $\lambda = 0.01$, $\alpha = 2$, and $k = 5, 10$, respectively. These sub-figures illustrate the behavior of the best relative Frank-Wolfe gap $\delta_t^*/\delta_0 := \min_{0 \le i \le t} \delta_i/\delta_0$ and the relative function value $\varphi(U^t, V^t)/\varphi(U^0, V^0)$ with respect to CPU time in seconds.

(ii) If $\alpha > 0$, then we have $\gamma_t \le \overline{\gamma}_t$ for all $t \ge t_0$ and $\beta_t^* \le (1+\alpha)^{\frac{1}{1+\alpha}} \overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor} \le e^{\frac{1}{e}} \overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor}$ for all $t \ge t_0 + 2A/(c\gamma_{t_0}^{\alpha})$, where

$$\overline{\gamma}_t = \left(\frac{1}{\gamma_{t_0}^{-\alpha} + A^{-1} c\alpha (t - t_0)} \right)^{\frac{1}{\alpha}}, \quad t_0 = \left\lceil \frac{1}{c} \left(\log \frac{\gamma_0}{cA^{1/\alpha}} \right)_+ \right\rceil. \tag{29}$$

Consequently, we have $\beta_t^* \le \varepsilon$ whenever

$$t \ge t_0 + \frac{2A}{c\gamma_{t_0}^{\alpha}} \max \left\{ 1, \ \frac{1}{\alpha} \left[\left(\frac{e^{\frac{1}{e}} \gamma_{t_0}}{\varepsilon} \right)^{\alpha} - 1 \right] \right\}. \tag{30}$$

Proof (i) Consider the case $\alpha = 0$. By (27) and $\beta_t \ge \gamma_t \ge 0$ for all $t \ge 0$, one can obtain for all $t \ge 0$ that

$$\gamma_{t+1} \le \gamma_t - c\gamma_t \min\{1, \gamma_t^{\alpha}/A\} = \gamma_t (1 - c \min\{1, A^{-1}\}) \le \gamma_t \exp(-c \min\{1, A^{-1}\}),$$

and hence $\gamma_t \le \gamma_0 \exp(-c \min\{1, A^{-1}\} t) = \overline{\gamma}_t$ for all $t \ge 0$. By this, $\{\gamma_t\} \subset \mathbb{R}_+$, and (28) with $\alpha = 0$, one has that for any $k \le t$,

$$c \min\{1, A^{-1}\}(t - k + 1)\beta_t^* \le \sum_{i=k}^{t} (c \min\{1, A^{-1}\} \beta_i^*) \le \sum_{i=k}^{t} (\gamma_i - \gamma_{i+1}) = \gamma_k - \gamma_{t+1} \le \overline{\gamma}_k. \tag{31}$$

For convenience, let $T_0 = 2\max\{1, A\}/c$ and $T_1 = \max\{1, A\} \log(\gamma_0/\varepsilon)/c$. For any $t \ge T_0$, letting $k = \lfloor (t+2)/2 \rfloor$ in (31), we obtain $\beta_t^* \le \overline{\gamma}_{\lfloor (t+2)/2 \rfloor}$ due to

$$c \min\{1, A^{-1}\}(t - k + 1) \ge c \min\{1, A^{-1}\} t/2 = t/T_0 \ge 1.$$

Moreover, since $\overline{\gamma}_{\lfloor (t+2)/2 \rfloor} \le \varepsilon$ holds if $\lfloor (t+2)/2 \rfloor \ge T_1$, we have $\beta_t^* \le \varepsilon$ whenever

$$t \ge \max\{T_0, 2T_1\} = \frac{2}{c} \max\{1, A\} \max \left\{ 1, \log \frac{\gamma_0}{\varepsilon} \right\}.$$

Hence, statement (i) holds.

(ii) We now consider the case $\alpha > 0$. It follows from the relation $\gamma_t \le \beta_t$ and the monotonicity of the sequence $\{\gamma_t\}$ that $\gamma_t \le \beta_t^*$. As long as $\beta_t^* > A^{1/\alpha}$, the relation (28) implies

$$\gamma_{t+1} \le \gamma_t - c\beta_t^* \le (1 - c)\gamma_t \le \gamma_t \exp(-c). \tag{32}$$

Claim that $\beta_{t_0}^* \le A^{1/\alpha}$ for t_0 defined in (29). Suppose for contradiction that $\beta_{t_0}^* > A^{1/\alpha}$. Then, as (32) holds for $t = 0, \ldots, t_0$, we have $\gamma_{t_0} \le \gamma_0 \exp(-ct_0)$ and thus $\gamma_{t_0} \le cA^{1/\alpha}$ follows by the expression of t_0. However, since $\{\gamma_t\}$ is nonnegative, the first inequality of (32) implies $\beta_{t_0}^* \le c^{-1}\gamma_{t_0} \le A^{1/\alpha}$, which leads to a contradiction. Hence, $\beta_{t_0}^* \le A^{1/\alpha}$ holds as claimed.

It follows from the monotonicity of $\{\beta_t^*\}$ that $\gamma_t \le \beta_t^* \le A^{1/\alpha}$ for all $t \ge t_0$. By this and (28), one can obtain the recurrence

$$\gamma_{t+1} \le \gamma_t - c(\beta_t^*)^{1+\alpha}/A \le \gamma_t - c\gamma_t^{1+\alpha}/A, \quad \forall t \ge t_0. \tag{33}$$

Then, by (Borwein et al., 2014, Lemma 4.1), this recurrence implies the assertion

$$\gamma_t \le (\gamma_{t_0}^{-\alpha} + A^{-1}c\alpha(t-t_0))^{-1/\alpha} = \overline{\gamma}_t, \quad \forall t \ge t_0. \tag{34}$$

We next show that $\beta_t^* \le (1+\alpha)^{\frac{1}{1+\alpha}}\overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor}$ for all $t \ge t_0 + \frac{2A}{c\gamma_{t_0}^{\alpha}}$. It follows from the first inequality in (33) and the monotonicity of $\{\beta_t^*\}$ that

$$c(\beta_t^*)^{1+\alpha}/A \le \gamma_i - \gamma_{i+1}, \quad \forall t \ge t_0, \ \ i \le t. \tag{35}$$

Let $k = \lfloor (t+t_0+1)/2 \rfloor$. Observe that $t-k+1 \ge k-t_0$. When $t \ge t_0+1$, we have $t \ge k \ge t_0+1$ and thus $\gamma_k \le \overline{\gamma}_k$ holds from (34). By these relations, $\gamma_{t+1} \ge 0$, and summing up the inequality (35) for $i = k, \ldots, t$, one has

$$(k-t_0)c(\beta_t^*)^{1+\alpha}/A \le (t-k+1)c(\beta_t^*)^{1+\alpha}/A \le \gamma_k - \gamma_{t+1} \le \gamma_k \le \overline{\gamma}_k, \quad \forall t \ge t_0+1.$$

In view of this and the expression of $\overline{\gamma}_t$, we obtain that for all $t \ge t_0+1$,

$$\beta_t^* \le \left(\frac{\overline{\gamma}_k}{A^{-1}c(k-t_0)}\right)^{\frac{1}{1+\alpha}} = \left(\frac{\gamma_{t_0}^{-\alpha} + A^{-1}c\alpha(k-t_0)}{A^{-1}c(k-t_0)}\right)^{\frac{1}{1+\alpha}} \overline{\gamma}_k = \theta_k \overline{\gamma}_k, \tag{36}$$

where $\theta_k = \left(\frac{\gamma_{t_0}^{-\alpha} + A^{-1}c\alpha(k-t_0)}{A^{-1}c(k-t_0)}\right)^{\frac{1}{1+\alpha}}$. Observe that θ_k is non-increasing and $\theta_k \downarrow 1$ as $k \to \infty$. For convenience, let $T_2 = t_0 + 2A/(c\gamma_{t_0}^{\alpha})$. Claim that $\theta_k \le (1+\alpha)^{\frac{1}{1+\alpha}}$ whenever $t \ge T_2$. Indeed, fix any $t \ge T_2$. By this and the expression of k, one can observe that $k \ge t_0 + A/(c\gamma_{t_0}^{\alpha})$, which along with the expression of θ_k implies that $\theta_k \le (1+\alpha)^{\frac{1}{1+\alpha}}$ holds as claimed. Using this and (36), we conclude that

$$\beta_t^* \le (1+\alpha)^{\frac{1}{1+\alpha}}\overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor} \le e^{\frac{1}{e}}\overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor}, \quad \forall t \ge T_2,$$

where the second inequality follows from $(1+\alpha)^{\frac{1}{1+\alpha}} \le \max_{\theta>0} \theta^{-\theta} = e^{\frac{1}{e}}$.

Finally, by the expression of $\overline{\gamma}_t$, one can see that the relation $e^{\frac{1}{e}}\overline{\gamma}_{\lfloor (t+t_0+1)/2 \rfloor} \le \varepsilon$ holds if $t \ge T_3$, where

$$T_3 = t_0 + \frac{2A}{c\gamma_{t_0}^{\alpha}\alpha}\left[\left(\frac{e^{\frac{1}{e}}\gamma_{t_0}}{\varepsilon}\right)^{\alpha} - 1\right].$$

It then follows that $\beta_t^* \le \varepsilon$ holds whenever $t \ge \max\{T_2, T_3\}$ and hence (30) holds. This completes the proof of statement (ii). ■

The following lemma establishes the weak smoothness of the function $\frac{1}{p}\|Ax-b\|_p^p$ for $p \in (1,2]$ which has been used in Section 5.

Lemma 17 *For $p \in (1,2]$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, the function $\phi(x) = \frac{1}{p}\|Ax-b\|_p^p$ satisfies*

$$\|\nabla\phi(x) - \nabla\phi(y)\|_2 \le M_{p-1}\|x-y\|_2^{p-1}, \quad \forall x, y \in \mathbb{R}^n,$$

where $M_{p-1} = 2^{2-p}m^{\frac{(p-1)(2-p)}{2p}}\|A\|_2^p$ and $\|A\|_2 = \max_{\|x\|_2\le 1}\|Ax\|_2$.

Proof We first consider the univariate function $g(\tau) = \frac{1}{p}|\tau|^p$ for $\tau \in \mathbb{R}$. Its derivative is given by $g'(\tau) = |\tau|^{p-1}\operatorname{sign}(\tau)$. Claim that

$$|g'(\tau) - g'(\tau')| \leq 2^{2-p}|\tau - \tau'|^{p-1}, \quad \forall \tau, \tau' \in \mathbb{R}. \tag{37}$$

Indeed, it suffices to show that

$$|\alpha^{p-1} - \beta^{p-1}| \leq |\alpha - \beta|^{p-1}, \quad \alpha^{p-1} + \beta^{p-1} \leq 2^{2-p}(\alpha + \beta)^{p-1}$$

for every $\alpha, \beta \geq 0$. The first inequality follows from the fact $(x+y)^{p-1} \leq x^{p-1} + y^{p-1}$ for $x, y \geq 0$ and the second one holds due to the concavity property $[(\alpha+\beta)/2]^{p-1} \geq (\alpha^{p-1} + \beta^{p-1})/2$. Hence, (37) holds as claimed.

Let $h(z) = \frac{1}{p}\|z\|_p^p$ for any $z \in \mathbb{R}^m$. Notice that $h(z) = \sum_{i=1}^m g(z_i)$, which together with (37) implies that $\|\nabla h(x) - \nabla h(y)\|_p \leq 2^{2-p}\|x-y\|_p^{p-1}$. Also, observe that $\phi(x) = h(Ax-b)$. Using these and $\|z\|_2 \leq \|z\|_p \leq m^{1/p-1/2}\|z\|_2$ for any $z \in \mathbb{R}^m$, we obtain that

$$\begin{aligned}
\|\nabla\phi(x) - \nabla\phi(y)\|_2 &\leq \|A^T\|_2 \|\nabla h(Ax-b) - \nabla h(Ay-b)\|_2 \leq \|A\|_2 \|\nabla h(Ax-b) - \nabla h(Ay-b)\|_p \\
&\leq 2^{2-p}\|A\|_2 \|A(x-y)\|_p^{p-1} \leq 2^{2-p} m^{(p-1)(1/p-1/2)} \|A\|_2 \|A(x-y)\|_2^{p-1} \\
&\leq 2^{2-p} m^{(p-1)(2-p)/(2p)} \|A\|_2^p \|x-y\|_2^{p-1}.
\end{aligned}$$

Hence, the conclusion holds as desired. ■

6.2 Proof of the Main Results in Section 3

In this subsection, we prove Theorems 5 and 7. Before proceeding, we establish a descent property for the sequence $\{\varphi(x_t)\}$.

Lemma 18 *Let the sequences $\{x_t\}$, $\{\delta_t\}$ and $\{v_t\}$ be generated in Algorithm 1. Suppose that Assumption 1 holds and that $\delta_t > 0$ for all $t \geq 0$.*[7] *Then we have*

$$\varphi(x_{t+1}) \leq \varphi(x_t) - \frac{\nu}{1+\nu}\delta_t \min\left\{1, \left(\frac{\delta_t}{M_\nu \|x_t - v_t\|^{1+\nu}}\right)^{\frac{1}{\nu}}\right\}, \quad \forall t \geq 0. \tag{38}$$

Proof By the Hölder continuity of ∇f (see (6)), we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{M_\nu}{1+\nu}\|x-y\|^{1+\nu}, \quad \forall x, y \in \operatorname{dom} g. \tag{39}$$

7. If $\delta_t = 0$ for some $t \geq 0$, x_t is already a stationary point of problem (1) and Algorithm 1 shall be terminated.

By the convexity of g, and (39) with $y = (1-\tau)x_t + \tau v_t$ and $x = x_t$, one can obtain that for any $\tau \in [0,1]$,

$$\begin{aligned}
&\varphi((1-\tau)x_t + \tau v_t) \\
&\leq f(x_t) + \langle \nabla f(x_t), (1-\tau)x_t + \tau v_t - x_t \rangle + \frac{M_\nu}{1+\nu} \|(1-\tau)x_t + \tau v_t - x_t\|^{1+\nu} \\
&\quad + g((1-\tau)x_t + \tau v_t) \\
&\leq f(x_t) - \tau \langle \nabla f(x_t), x_t - v_t \rangle + \tau^{1+\nu} \frac{M_\nu}{1+\nu} \|x_t - v_t\|^{1+\nu} + (1-\tau)g(x_t) + \tau g(v_t) \\
&= \varphi(x_t) - \tau\delta_t + \tau^{1+\nu} \frac{M_\nu}{1+\nu} \|x_t - v_t\|^{1+\nu}. \qquad (40)
\end{aligned}$$

Letting $\tau = \tau_t$ in (40), and using the expression of τ_t and x_{t+1}, we obtain that for any $t \geq 0$,

$$\begin{aligned}
\varphi(x_{t+1}) &\leq \varphi(x_t) - \tau_t\delta_t + \tau_t^{1+\nu} \frac{M_\nu}{1+\nu} \|x_t - v_t\|^{1+\nu} \\
&\leq \varphi(x_t) - \frac{\nu}{1+\nu}\delta_t \min\left\{1, \left(\frac{\delta_t}{M_\nu \|x_t - v_t\|^{1+\nu}}\right)^{\frac{1}{\nu}}\right\}.
\end{aligned}$$

■

We are now ready to prove Theorems 5 and 7.

Proof of Theorem 5 Let the sequences $\{x_t\}$ and $\{v_t\}$ be generated in Algorithm 1. One can observe that $x_t, v_t \in \operatorname{dom} g$ for all $t \geq 0$. It then follows that $\|x_t - v_t\| \leq D_g$. By this and Lemma 18, one can obtain

$$\varphi(x_{t+1}) \leq \varphi(x_t) - \frac{\nu}{1+\nu}\delta_t \min\left\{1, \left(\frac{\delta_t}{M_\nu D_g^{1+\nu}}\right)^{\frac{1}{\nu}}\right\}, \quad \forall t \geq 0. \qquad (41)$$

(i) It follows from (41) that $\{\varphi(x_t)\}$ is non-increasing, which, together with the fact that $\varphi(x_t) \geq \varphi^*$ for all $t \geq 0$, implies that $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ exists. In addition, one can observe from (41) that the recurrence (27) holds for $\beta_t = \delta_t$, $\gamma_t = \varphi(x_t) - \varphi_*$, $\alpha = 1/\nu$, $c = \nu/(1+\nu)$, and $A = M_\nu^{\frac{1}{\nu}} D_g^{\frac{1+\nu}{\nu}}$. The inequality (12) then directly follows from Lemma 15.

(ii) One can observe from (41) that the recurrence (27) holds for $\beta_t = \delta_t$, $\gamma_t = \varphi(x_t) - \varphi^*$, $\alpha = 1/\nu$, $c = \nu/(1+\nu)$, and $A = M_\nu^{\frac{1}{\nu}} D_g^{\frac{1+\nu}{\nu}}$. In addition, $\beta_t \geq \gamma_t$ due to Lemma 3. The conclusion of this statement then immediately follows from Lemma 16 (ii). ■

Proof of Theorem 7 Let the sequences $\{x_t\}$ and $\{v_t\}$ be generated in Algorithm 1. One can observe that $x_t, v_t \in \operatorname{dom} g$ for all $t \geq 0$. By this, the expression of δ_t, and Assumption 2, one has

$$\delta_t = \langle \nabla f(x_t), x_t \rangle + g(x_t) - \langle \nabla f(x_t), v_t \rangle - g(v_t) \geq \frac{\kappa}{\rho} \|x_t - v_t\|^\rho, \quad \forall t \geq 0. \qquad (42)$$

Using this inequality, we can obtain that

$$\frac{\delta_t}{M_\nu \|x_t - v_t\|^{1+\nu}} = \frac{1}{M_\nu} \left(\frac{\delta_t}{\|x_t - v_t\|^\rho}\right)^{\frac{1+\nu}{\rho}} \delta_t^{1-\frac{1+\nu}{\rho}} \geq \frac{1}{M_\nu} \left(\frac{\kappa}{\rho}\right)^{\frac{1+\nu}{\rho}} \delta_t^{1-\frac{1+\nu}{\rho}},$$

which together with (38) yields

$$\varphi(x_{t+1}) \leq \varphi(x_t) - \frac{\nu}{1+\nu}\delta_t \min\left\{1, \left(\frac{\kappa^{\frac{1+\nu}{\rho}}\delta_t^{1-\frac{1+\nu}{\rho}}}{\rho^{\frac{1+\nu}{\rho}} M_\nu}\right)^{\frac{1}{\nu}}\right\}, \quad \forall t \geq 0. \tag{43}$$

(i) By (43) and a similar argument as in the proof of Theorem 5 (i), one can see that $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t\to\infty}\varphi(x_t)$ exists. In addition, one can observe from (43) that the recurrence (27) holds for $\beta_t = \delta_t$, $\gamma_t = \varphi(x_t) - \varphi_*$, $\alpha = (\rho - 1 - \nu)/(\rho\nu)$, $c = \nu/(1+\nu)$, and $A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} M_\nu^{\frac{1}{\nu}}$. The inequality (13) then directly follows from Lemma 15.

(ii) Let $\alpha = (\rho - 1 - \nu)/(\rho\nu)$ and $c = \nu/(1+\nu)$. One can observe from (43) that the recurrence (27) holds for such α, c, $\beta_t = \delta_t$, $\gamma_t = \varphi(x_t) - \varphi^*$, and $A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} M_\nu^{\frac{1}{\nu}}$. Also, $\beta_t \geq \gamma_t$ due to Lemma 3. In addition, by $\nu \in (0,1]$, $\rho \geq 2$, and the expression of α, it is not hard to see that $\alpha = 0$ if and only if $\nu = 1$ and $\rho = 2$. Also, one can see from the expression of c and A that $c = 1/2$ and $A = 2M_1/\kappa$ when $\nu = 1$ and $\rho = 2$. The conclusion of this statement then follows from these observations and Lemma 16. ■

6.3 Proof of the Main Results in Section 4

In this subsection, we prove Theorems 9, 10, and 12.

Proof of Theorem 9 (i) For any $\tau \in [0,1]$, by the convexity of g, the expression of δ_t, and (15) with $x = x_t$, $y = (1-\tau)x_t + \tau v_t$ and $\varepsilon = \tau\delta_t/2$, one has

$$\begin{aligned}
&\varphi((1-\tau)x_t + \tau v_t) = f((1-\tau)x_t + \tau v_t) + g((1-\tau)x_t + \tau v_t)\\
&\leq f(x_t) + \langle \nabla f(x_t), (1-\tau)x_t + \tau v_t - x_t\rangle + \frac{L(\tau\delta_t/2)}{2}\left\|(1-\tau)x_t + \tau v_t - x_t\right\|^2 + \frac{\tau\delta_t}{2}\\
&\quad + g((1-\tau)x_t + \tau v_t)\\
&\leq f(x_t) - \tau\left\langle \nabla f(x_t), x_t - v_t\right\rangle + \tau^2\frac{L(\tau\delta_t/2)}{2}\left\|x_t - v_t\right\|^2 + (1-\tau)g(x_t) + \tau g(v_t) + \frac{\tau\delta_t}{2}\\
&= \varphi(x_t) - \frac{1}{2}\tau\delta_t + \tau^2\frac{L(\tau\delta_t/2)}{2}\left\|x_t - v_t\right\|^2.
\end{aligned}$$

Letting $\tau = \tau_t^{(i)}$ in this inequality yields

$$\varphi(x_{t+1}^{(i)}) \leq \varphi(x_t) - \frac{1}{2}\tau_t^{(i)}\delta_t + (\tau_t^{(i)})^2\frac{L(\tau_t^{(i)}\delta_t/2)}{2}\left\|x_t - v_t\right\|^2.$$

Hence, (14) holds if

$$L_t^{(i)} \geq L(\tau_t^{(i)}\delta_t/2) = \max\left\{L(\delta_t/2), L\left(\frac{\delta_t^2}{4L_t^{(i)}\left\|x_t - v_t\right\|^2}\right)\right\}, \tag{44}$$

where the equality follows from the expression of $\tau_t^{(i)}$ and the fact that $L(\cdot)$ is non-increasing. By (16), one can verify that

$$L_t^{(i)} \geq L\left(\frac{\delta_t^2}{4L_t^{(i)}\left\|x_t - v_t\right\|^2}\right) \iff L_t^{(i)} \geq L\left(\frac{\delta_t^2}{4\left\|x_t - v_t\right\|^2}\right)^{\frac{1+\nu}{2\nu}},$$

which together with (44) implies that

$$L_t^{(i)} \geq L(\tau_t^{(i)}\delta_t/2) \iff L_t^{(i)} \geq \max\left\{L(\delta_t/2), L\left(\frac{\delta_t^2}{4\|x_t - v_t\|^2}\right)^{\frac{1+\nu}{2\nu}}\right\} = \widetilde{L}_t.$$

Hence, (14) holds if $L_t^{(i)} \geq \widetilde{L}_t$.

(ii) In the t-th outer iteration, let $i_t \geq 0$ denote the final iteration counter for the adaptive line search loop, and let $\widetilde{L}_t^* = \max_{0\leq i\leq t} \widetilde{L}_i$. For $t \geq 0$ and $s \in \{0, \ldots, t\}$, we first show that

$$L_{s-1} \leq 2\widetilde{L}_t^* \implies L_s \leq 2\widetilde{L}_t^*. \tag{45}$$

Indeed, if $i_s = 0$, one can observe that $L_s = L_{s-1}/2$, which immediately implies that (45) holds. Now we suppose $i_s > 0$. It then follows that the adaptive line search loop fails to terminate at the inner iteration $i_s - 1$, which along with statement (i) implies that $L_s/2 = L_s^{(i_s-1)} < \widetilde{L}_s$. It then follows that $L_s < 2\widetilde{L}_s \leq 2\widetilde{L}_t^*$. Hence, (45) holds as desired. By these arguments, one can also observe that $L_t = L_{t-1}/2$ whenever $t \notin \mathcal{T}$, where

$$\mathcal{T} = \{t \in \mathbb{Z}_+ : L_{t-1} \leq 2\widetilde{L}_t^*\}.$$

Due to this observation and the fact that $\widetilde{L}_t^* \geq \widetilde{L}_0 > 0$ for all $t \geq 0$, the set $\mathcal{T}$ must be nonempty. Then, $\tilde{t}_0 = \min\{t : t \in \mathcal{T}\}$ is well-defined. Since $t \in \mathcal{T}$ implies $t + 1 \in \mathcal{T}$ due to (45), we see that $\mathcal{T} = \{\tilde{t}_0, \tilde{t}_0 + 1, \ldots\}$. In addition, (45) implies $\mathcal{T} \subset \{t \in \mathbb{Z}_+ : L_t \leq 2\widetilde{L}_t^*\}$. Hence, we obtain that

$$L_t \leq 2\widetilde{L}_t^*, \quad \forall t \geq \tilde{t}_0.$$

To complete the proof, it suffices to show $\tilde{t}_0 \leq (\log_2(L_{-1}/\widetilde{L}_0))_+$. Indeed, it holds trivially if $\tilde{t}_0 = 0$. Now we suppose $\tilde{t}_0 > 0$. Recall that $L_t = L_{t-1}/2$ whenever $t \notin \mathcal{T}$. It then follows that $L_t = L_{t-1}/2$ for $t = 0, \ldots, \tilde{t}_0 - 1$. Hence, we have

$$L_{-1}/2^{\tilde{t}_0-1} = L_{\tilde{t}_0-2} > 2\widetilde{L}_{\tilde{t}_0-1}^* \geq 2\widetilde{L}_0,$$

which yields $\tilde{t}_0 \leq \log_2(L_{-1}/\widetilde{L}_0)$. Thus, $\tilde{t}_0 \leq (\log_2(L_{-1}/\widetilde{L}_0))_+$ holds as desired.

(iii) We first show $\widetilde{L}_t \leq \overline{L}(\delta_t)$. Indeed, by (16) and the expression of $\widetilde{L}_t$, one has

$$\widetilde{L}_t = \max\left\{\left(\frac{1-\nu}{1+\nu}\frac{1}{\delta_t}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \left(\frac{2(1-\nu)}{1+\nu}\right)^{\frac{1-\nu}{2\nu}} \left(\frac{\|x_t - v_t\|}{\delta_t}\right)^{\frac{1-\nu}{\nu}} M_\nu^{\frac{1}{\nu}}\right\}. \tag{46}$$

Since $x_t, v_t \in \operatorname{dom} g$, one can observe that $\|x_t - v_t\|/\delta_t \leq D_g/\delta_t$ if $\operatorname{dom} g$ is bounded. Also, if Assumption 2 holds, it follows from (42) that

$$\frac{\|x_t - v_t\|}{\delta_t} = \left(\frac{\|x_t - v_t\|^\rho}{\delta_t}\right)^{\frac{1}{\rho}} \delta_t^{\frac{1}{\rho}-1} \leq \left(\frac{\rho}{\kappa}\right)^{\frac{1}{\rho}} \delta_t^{\frac{1}{\rho}-1}.$$

Then, $\widetilde{L}_t \leq \overline{L}(\delta_t)$ holds due to (46) and the last two inequalities. By this, $\min_{0\leq i\leq t} \delta_i \geq \varepsilon$, and the fact that $\overline{L}(\cdot)$ is non-increasing, we obtain that

$$\widetilde{L}_t^* = \max_{0\leq i\leq t} \widetilde{L}_i \leq \max_{0\leq i\leq t} \overline{L}(\delta_i) \leq \overline{L}(\varepsilon).$$

In addition, from the proof of statement (ii), we can observe that $L_t \leq \max\{L_{-1}/2, 2\widetilde{L}_t^*\}$ for all $t \geq 0$. Also, by the definition of i_s, one can see that $L_s = 2^{i_s-1}L_{s-1}$ for $s \geq 0$, and the total number of inner loops performed by the adaptive line search procedure until the t-th iteration of Algorithm 2 is given by $\sum_{s=0}^{t}(1+i_s)$. By these observations, one can have

$$\sum_{s=0}^{t}(1+i_s) = \sum_{s=0}^{t}\left(2+\log_2\frac{L_s}{L_{s-1}}\right) = 2(t+1)+\log_2\frac{L_t}{L_{-1}}$$
$$\leq 2(t+1)+\log_2\frac{\max\{L_{-1}/2, 2\widetilde{L}_t^*\}}{L_{-1}} \leq 2t+2+[\log_2(2\overline{L}(\varepsilon)/L_{-1})]_+,$$

and hence the conclusion holds. ■

Before proving Theorems 10 and 12, we establish a lemma that will be used shortly.

Lemma 19 *Let the sequences $\{x_t\}$, $\{\delta_t\}$ and $\{v_t\}$ be generated in Algorithm 2. Suppose that Assumption 1 holds and that $\delta_t > 0$ for all $t \geq 0$. Let $\tilde{t}_0 = \lceil(\log_2(L_{-1}/\widetilde{L}_0))_+\rceil$, $\delta_t^* = \min_{0\leq i\leq t}\delta_t$, and $\widetilde{L}_t$ and $\overline{L}(\cdot)$ be defined in* (17) *and* (18)*, respectively. Then it holds that*

$$\varphi(x_{t+1}) \leq \varphi(x_t) - \frac{\delta_t^*}{4}\min\{1, C_t\}, \quad \forall t \geq \tilde{t}_0, \tag{47}$$

where

$$C_t := \frac{1}{2\overline{L}(\delta_t^*)}\min_{0\leq i\leq t}\frac{\delta_i}{\|x_i - v_i\|^2}, \quad \forall t \geq 0. \tag{48}$$

Proof One can observe from Algorithm 2 that

$$\tau_t = \min\left\{1, \frac{\delta_t}{2L_t\|x_t - v_t\|^2}\right\}, \quad \varphi(x_{t+1}) \leq \varphi(x_t) - \frac{1}{2}\tau_t\delta_t + \frac{1}{2}L_t\tau_t^2\|x_t - v_t\|^2, \quad \forall t \geq 0.$$

It then follows that

$$\varphi(x_{t+1}) \leq \varphi(x_t) - \frac{\delta_t}{4}\min\left\{1, \frac{\delta_t}{L_t\|x_t - v_t\|^2}\right\}, \quad \forall t \geq 0. \tag{49}$$

Recall from the proof of Theorem 9 (iii) that $\widetilde{L}_t \leq \overline{L}(\delta_t)$ for all $t \geq 0$. Using this, Theorem 9 (ii), $\delta_t^* = \min_{0\leq i\leq t}\delta_t$, and the monotonicity of $\overline{L}(\cdot)$, we have

$$L_t \leq 2\max_{0\leq i\leq t}\widetilde{L}_i \leq 2\max_{0\leq i\leq t}\overline{L}(\delta_i) = 2\overline{L}(\delta_t^*), \quad \forall t \geq \tilde{t}_0.$$

The conclusion then follows from this, (49), and $\delta_t^* = \min_{0\leq i\leq t}\delta_t$. ■

We are now ready to prove Theorems 10 and 12.

Proof of Theorem 10 One can observe that $x_t, v_t \in \operatorname{dom} g$. It then follows that $\|x_t - v_t\| \leq D_g$ for all $t \geq 0$. By this and $\delta_t^* = \min_{0\leq i\leq t}\delta_i$, one has

$$\frac{\delta_i}{\|x_i - v_i\|^2} \geq \frac{\delta_t^*}{D_g^2}, \quad \forall 0 \leq i \leq t. \tag{50}$$

Let C_t be defined in (48). We next bound C_t from below by considering two cases in view of the definition of $\overline{L}(\cdot)$ in (18).

Case 1) $\overline{L}(\delta_t^*) = \left(\frac{2(1-\nu)}{1+\nu}\right)^{\frac{1-\nu}{2\nu}} M_\nu^{\frac{1}{\nu}} \left(\frac{D_g}{\delta_t^*}\right)^{\frac{1-\nu}{\nu}}$. By this, (48) and (50), we obtain that

$$C_t \geq \frac{1}{2}\left(\frac{2(1-\nu)}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}} M_\nu^{-\frac{1}{\nu}} \left(\frac{D_g}{\delta_t^*}\right)^{-\frac{1-\nu}{\nu}} \cdot \frac{\delta_t^*}{D_g^2} = 2^{-\frac{1+\nu}{2\nu}} \left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}} \left(\frac{\delta_t^*}{M_\nu D_g^{1+\nu}}\right)^{\frac{1}{\nu}} =: D_t.$$

Case 2) $\overline{L}(\delta_t^*) = \left(\frac{1-\nu}{1+\nu}\frac{1}{\delta_t^*}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$. By this, (48) and (50), one has

$$C_t \geq \frac{1}{2}\left(\frac{1-\nu}{1+\nu}\frac{1}{\delta_t^*}\right)^{-\frac{1-\nu}{1+\nu}} M_\nu^{-\frac{2}{1+\nu}} \cdot \frac{\delta_t^*}{D_g^2} = \frac{1}{2}\left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{1+\nu}} \left(\frac{\delta_t^*}{M_\nu D_g^{1+\nu}}\right)^{\frac{2}{1+\nu}} = D_t^{\frac{2\nu}{1+\nu}}.$$

Combining these two cases, and using (18) and (48), we conclude that $C_t \geq \min\{D_t, D_t^{\frac{2\nu}{1+\nu}}\}$. By this and $2\nu/(1+\nu) \leq 1$, one can observe that

$$\min\{1, C_t\} \geq \min\{1, D_t\}, \quad \forall t \geq 0.$$

In view of this, (47), and the expression of D_t, one has

$$\begin{aligned}\varphi(x_{t+1}) &\leq \varphi(x_t) - \frac{\delta_t^*}{4}\min\left\{1, 2^{-\frac{1+\nu}{2\nu}}\left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}}\left(\frac{\delta_t^*}{M_\nu D_g^{1+\nu}}\right)^{\frac{1}{\nu}}\right\},\\ &\leq \varphi(x_t) - \frac{\delta_t^*}{4}\min\left\{1, \left(\frac{\delta_t^*}{2M_\nu D_g^{1+\nu}}\right)^{\frac{1}{\nu}}\right\}, \quad \forall t \geq \tilde{t}_0, \end{aligned} \tag{51}$$

where the last inequality is due to $2^{\frac{1+\nu}{2\nu}} \leq 2^{\frac{1+1}{2\nu}}$ and $\frac{1-\nu}{1+\nu} \leq 1$.

(i) It follows from (49) that $\{\varphi(x_t)\}$ is non-increasing, which, together with the fact that $\varphi(x_t) \geq \varphi^*$ for all $t \geq 0$, implies that $\varphi_* = \lim_{t\to\infty}\varphi(x_t)$ exists. In addition, one can observe from (51) that (27) holds for $\beta_t = \delta^*_{t+\tilde{t}_0}$, $\gamma_t = \varphi(x_{t+\tilde{t}_0}) - \varphi_*$, $\alpha = 1/\nu$, $c = 1/4$, and $A = (2M_\nu D_g^{1+\nu})^{\frac{1}{\nu}}$. The inequality (19) then directly follows from Lemma 15.

(ii) One can observe from (51) that (27) holds for $\beta_t = \delta^*_{t+\tilde{t}_0}$, $\gamma_t = \varphi(x_{t+\tilde{t}_0}) - \varphi^*$, $\alpha = 1/\nu$, $c = 1/4$, and $A = (2M_\nu D_g^{1+\nu})^{\frac{1}{\nu}}$. In addition, by Lemma 3, and the monotonicity of $\{\varphi(x_t)\}$, one has

$$\delta_t^* = \min_{0\leq i\leq t}\delta_i \geq \min_{0\leq i\leq t}\{\varphi(x_i) - \varphi^*\} = \varphi(x_t) - \varphi^*, \quad \forall t \geq 0.$$

Hence, $\beta_t \geq \gamma_t$ for all $t \geq 0$. The conclusion of this statement then follows from Lemma 16 (ii). ■

Proof of Theorem 12 It follows from (42) and $\delta_t^* = \min_{0\leq i\leq t}\delta_i$ that

$$\frac{\delta_i}{\|x_i - v_i\|^2} = \left(\frac{\delta_i}{\|x_i - v_i\|^\rho}\right)^{\frac{2}{\rho}} \delta_i^{1-\frac{2}{\rho}} \geq \left(\frac{\kappa}{\rho}\right)^{\frac{2}{\rho}} \delta_i^{1-\frac{2}{\rho}} \geq \left(\frac{\kappa}{\rho}\right)^{\frac{2}{\rho}} (\delta_t^*)^{1-\frac{2}{\rho}}, \quad \forall 0 \leq i \leq t. \tag{52}$$

Let C_t be defined by (48). We next bound C_t from below by considering two cases in view of the definition of $\overline{L}(\cdot)$ in (18).

Case 1) $\overline{L}(\delta_t^*) = \left(\frac{2(1-\nu)}{1+\nu}\right)^{\frac{1-\nu}{2\nu}} M_\nu^{\frac{1}{\nu}} \left(\frac{\rho}{\kappa(\delta_t^*)^{\rho-1}}\right)^{\frac{1-\nu}{\nu\rho}}$. By this, (48) and (52), we obtain that

$$\begin{aligned} C_t &\geq \frac{1}{2}\left(\frac{2(1-\nu)}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}} M_\nu^{-\frac{1}{\nu}} \left(\frac{\rho}{\kappa(\delta_t^*)^{\rho-1}}\right)^{-\frac{1-\nu}{\nu\rho}} \cdot \left(\frac{\kappa}{\rho}\right)^{\frac{2}{\rho}} (\delta_t^*)^{1-\frac{2}{\rho}} \\ &= 2^{-\frac{1+\nu}{2\nu}} \left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}} \left[\frac{\left(\frac{\kappa}{\rho}\right)^{\frac{1+\nu}{\rho}} (\delta_t^*)^{\frac{\rho-(1+\nu)}{\rho}}}{M_\nu}\right]^{\frac{1}{\nu}} =: E_t. \end{aligned}$$

Case 2) $\overline{L}(\delta_t^*) = \left(\frac{1-\nu}{1+\nu}\frac{1}{\delta_t^*}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$. By this, (48) and (52), one has

$$\begin{aligned} C_t &\geq \frac{1}{2}\left(\frac{1-\nu}{1+\nu}\frac{1}{\delta_t^*}\right)^{-\frac{1-\nu}{1+\nu}} M_\nu^{-\frac{2}{1+\nu}} \cdot \left(\frac{\kappa}{\rho}\right)^{\frac{2}{\rho}} (\delta_t^*)^{1-\frac{2}{\rho}} \\ &= \frac{1}{2}\left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{1+\nu}} \left[\frac{\left(\frac{\kappa}{\rho}\right)^{\frac{1+\nu}{\rho}} (\delta_t^*)^{\frac{\rho-(1+\nu)}{\rho}}}{M_\nu}\right]^{\frac{2}{1+\nu}} = E_t^{\frac{2\nu}{1+\nu}}. \end{aligned}$$

Combining these two cases, and using (18) and (48), we conclude that $C_t \geq \min\{E_t, E_t^{\frac{2\nu}{1+\nu}}\}$. By this and $2\nu/(1+\nu) \leq 1$, one can observe that

$$\min\{1, C_t\} \geq \min\{1, E_t\}, \quad \forall t \geq 0.$$

In view of this, (47), and the expression of E_t, one has

$$\begin{aligned} \varphi(x_{t+1}) &\leq \varphi(x_t) - \frac{\delta_t^*}{4}\min\left\{1, 2^{-\frac{1+\nu}{2\nu}}\left(\frac{1-\nu}{1+\nu}\right)^{-\frac{1-\nu}{2\nu}} \left[\frac{\left(\frac{\kappa}{\rho}\right)^{\frac{1+\nu}{\rho}} (\delta_t^*)^{\frac{\rho-(1+\nu)}{\rho}}}{M_\nu}\right]^{\frac{1}{\nu}}\right\} \\ &\leq \varphi(x_t) - \frac{\delta_t^*}{4}\min\left\{1, \left[\frac{\left(\frac{\kappa}{\rho}\right)^{\frac{1+\nu}{\rho}} (\delta_t^*)^{\frac{\rho-(1+\nu)}{\rho}}}{2M_\nu}\right]^{\frac{1}{\nu}}\right\}, \quad \forall t \geq \tilde{t}_0, \end{aligned} \tag{53}$$

where the last inequality is due to $2^{\frac{1+\nu}{2\nu}} \leq 2^{\frac{1+1}{2\nu}}$ and $\frac{1-\nu}{1+\nu} \leq 1$.

(i) In view of (49) and $\varphi(x_t) \geq \varphi^* \in \mathbb{R}$, the sequence $\{\varphi(x_t)\}$ is non-increasing and $\varphi_* = \lim_{t\to\infty} \varphi(x_t)$ exists. In addition, one can observe from (53) that (27) holds for $\beta_t = \delta^*_{t+\tilde{t}_0}$, $\gamma_t = \varphi(x_{t+\tilde{t}_0}) - \varphi_*$, $\alpha = (\rho-1-\nu)/(\rho\nu)$, $c = 1/4$, and $A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} (2M_\nu)^{\frac{1}{\nu}}$. The inequality (20) then directly follows from Lemma 15.

(ii) One can observe from (53) that (27) holds for $\beta_t = \delta^*_{t+\tilde{t}_0}$, $\gamma_t = \varphi(x_{t+\tilde{t}_0}) - \varphi^*$, $\alpha = (\rho-1-\nu)/(\rho\nu)$, $c = 1/4$, and $A = \left(\frac{\rho}{\kappa}\right)^{\frac{1+\nu}{\rho\nu}} (2M_\nu)^{\frac{1}{\nu}}$. The rest of the proof is the same as that of Theorem 10 (ii). ■

7 Concluding Remarks

In this paper we first analyzed iteration complexity of a parameter-dependent conditional gradient method for solving problem (1), whose step sizes depend explicitly on the problem parameters. We then proposed a novel parameter-free conditional gradient method for solving (1) without using any prior knowledge of the problem parameters and showed that it enjoys the same order of iteration complexity as the the parameter-dependent conditional gradient method. Preliminary numerical experiments demonstrate the practical superiority of our parameter-free conditional gradient method over the other variants.

It shall be mentioned that our proposed method requires a pre-specified norm and thus it is norm-dependent. In contrast, some existing conditional gradient methods (e.g., Jaggi, 2013; Lacoste-Julien, 2016; Peña, 2022) are norm-independent. It would be interesting to develop a parameter-free but norm-independent conditional gradient method achieving the same complexity bounds as obtained this paper for solving (1). This is left for future research.

Acknowledgments and Disclosure of Funding

The authors would like to thank the anonymous referees for their valuable comments that improved the quality of the paper. The first author's research is supported by the Grant-in-Aid for Early-Career Scientists (JP21K17711) from Japan Society for the Promotion of Science. The work of the second author was partially supported by NSF Award IIS-2211491.

References

F. Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):115–129, 2015.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31:167–175, 2003.

A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Meth. Oper. Res.*, 59:235–247, 2004.

J. Bolte, L. Glaudin, E. Pauwels, and M. Serrurier. A Hölderian backtracking method for min-max and min-min problems. *arXiv preprint arXiv:2007.08810*, 2020.

J. M. Borwein, G. Li, and L. Yao. Analysis of the convergence rate for the cyclic projection algorithm applied to basic semi-algebraic convex sets. *SIAM J. Optim.*, 24(1):498–527, 2014.

T. Chen, J. B. Lasserre, V. Magron, and E. Pauwels. Semialgebraic optimization for Lipschitz constants of ReLU networks. *Advances in Neural Information Processing Systems*, 33:19189–19200, 2020.

K. L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms*, 6(4):1–30, 2010.

V. F. Demyanov and A. M. Rubinov. *Approximate methods in optimization problems*. American Elsevier Publishing Company, New York, 1970.

J. C. Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM J. Control Optim.*, 17:187–211, 1979.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 3: 95–110, 1956.

R. M. Freund and P. Grigas. New analysis and results for the Frank-Wolfe method. *Math. Program.*, 155:199–230, 2016.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. *Proceedings of the 32nd International Conference on Machine Learning*, 37:541–549, 2015.

S. Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Math. Program.*, 173:431–464, 2019.

C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *J. Complexity*, 31:1–14, 2015.

Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152:75–112, 2015.

E. Hazan. Sparse approximate solutions to semidefinite programs. *Proceedings of Theoretical Informatics, 8th Latin American Symposium*, pages 306–316, 2008.

M. Ito, Z. Lu, and C. He. Adaptive step-size rule for conditional gradient methods minimizing weakly smooth objective functions. SIAM Conference on Optimization, 2021. `http://trout.math.cst.nihon-u.ac.jp/~ito.m/articles/2021/adap-cgm.pdf`.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on Machine Learning*, 28:427–435, 2013.

T. Kerdreux, A. d'Aspremont, and S. Pokutta. Projection-free optimization on uniformly convex sets. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 130:19–27, 2021a.

T. Kerdreux, A. d'Aspremont, and S. Pokutta. Local and global uniform convexity conditions. *arXiv preprint arXiv:2102.05134*, 2021b.

S. Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Comput. Math. and Math. Phys.*, 6(5):1–50, 1966.

Y. Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152:381–404, 2015.

Y. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Math. Program.*, 171:311–330, 2018.

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108:177–205, 2006.

J. Peña. Affine invariant convergence rates of the conditional gradient method. *arXiv preprint arXiv:2112.06727v2*, 2022.

O. V. Thanh, N. Gillis, and F. Lecron. Bounded simplex-structured matrix factorization. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9062–9066. IEEE, 2022.

J.-P. Vial. Strong convexity of sets and functions. *J. Math. Econ.*, 9:187–205, 1982.

R. Zhao and R. M. Freund. Analysis of the Frank-Wolfe method for logarithmically-homogeneous barriers, with an extension. *arXiv preprint arXiv:2010.08999v1*, 2020.

C. Zălinescu. *Convex analysis in general vector spaces*. World Scientific, Singapore, 2002.

P. Zwiernik. Semialgebraic statistics and latent tree models. *Monographs on Statistics and Applied Probability*, 146, 2016.