CrossMark

Conditional gradient type methods for composite nonlinear and stochastic optimization

Saeed Ghadimi[1]

Abstract In this paper, we present a conditional gradient type (CGT) method for solving a class of composite optimization problems where the objective function consists of a (weakly) smooth term and a (strongly) convex regularization term. While including a strongly convex term in the subproblems of the classical conditional gradient method improves its rate of convergence, it does not cost per iteration as much as general proximal type algorithms. More specifically, we present a unified analysis for the CGT method in the sense that it achieves the best known rate of convergence when the weakly smooth term is nonconvex and possesses (nearly) optimal complexity if it turns out to be convex. While implementation of the CGT method requires explicitly estimating problem parameters like the level of smoothness of the first term in the objective function, we also present a few variants of this method which relax such estimation. Unlike general proximal type parameter free methods, these variants of the CGT method do not require any additional effort for computing (sub)gradients of the objective function and/or solving extra subproblems at each iteration. We then generalize these methods under stochastic setting and present a few new complexity results. To the best of our knowledge, this is the first time that such complexity results are presented for solving stochastic weakly smooth nonconvex and (strongly) convex optimization problems.

✉ Saeed Ghadimi
    sghadimi@princeton.edu

[1] Princeton University, Princeton, NJ, USA

1 Introduction

In this paper, we study a class of composite nonlinear programming problems given by

$$\Psi^* := \min_{x \in X}\{\Psi(x) := f(x) + h(x)\}, \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set, $f : X \to \mathbb{R}$ is continuously differentiable but possibly nonconvex such that Ψ^* is finite, and h is a possibly non-differentiable (strongly) convex function with parameter $\mu \geq 0$ with known structure i.e.,

$$h(\alpha x + (1 - \alpha)y) \leq \alpha h(x) + (1 - \alpha)h(y) - \frac{\alpha(1 - \alpha)\mu}{2}\|x - y\|^2 \tag{1.2}$$

for all $x, y \in X$ and $\alpha \in [0, 1]$. Note that we allow $\mu = 0$ to simply include the case when h is only convex. Moreover, we assume that f has Hölder continuous gradient on X i.e., there exists $L_\nu > 0$ for any $\nu \in (0, 1]$ such that

$$\|f'(y) - f'(x)\|_* \leq L_\nu \|y - x\|^\nu \quad \forall x, y \in X, \tag{1.3}$$

where $f'(x)$ is the gradient of f at x and $\| \cdot \|_*$ denotes the dual norm. It can be easily verified that the above inequality implies that

$$|f(y) - f(x) - \langle f'(x), y - x \rangle| \leq \frac{L_\nu}{1 + \nu}\|y - x\|^{1+\nu} \quad \forall x, y \in X. \tag{1.4}$$

The assumption in (1.3) covers a wide range class of functions including smooth ($\nu = 1$) and weakly smooth ones ($\nu \in (0, 1)$). Different applications arise in the form of problem (1.1). In several machine learning problems, the objective function is given as summation of a loss function and a (strongly) convex regularization term where the loss function can be convex (see e.g., [14,35]) or nonconvex (see e.g., [2,26]). Signal processing problems using a (strongly) convex regularization term to avoid data overfitting also satisfy in the above setting (see e.g., [19]).

Problem (1.1) has been well-studied under specific assumptions. Nesterov [32] showed that a variant of his well-known accelerated gradient (AG) method, originally proposed for smooth unconstrained problems, achieves the optimal iteration complexities $\mathcal{O}(1/\sqrt{\epsilon})$ and $\mathcal{O}(\ln 1/\epsilon)$ for finding an ϵ-optimal solution of problem (1.1) i.e., a solution $\bar{x} \in X$ s.t. $\Psi(\bar{x}) - \Psi(x^*) \leq \epsilon$, when f is a smooth convex function, $\mu = 0$, and $\mu > 0$, respectively. Lan [23] presented bundle-level type methods achieving the optimal complexity $\mathcal{O}(1/\epsilon^{\frac{2}{1+3\nu}})$ for solving weakly smooth convex problems. Nemirovski and Nesterov [28] also proposed an optimal proximal gradient

type method which possesses the optimal complexity bound $\mathcal{O}(1/\epsilon^{\frac{1-\nu}{1+3\nu}})$ for solving weakly smooth strongly convex problems. This bound is also recently obtained in [4,17]. When f is smooth and possibly nonconvex and h is only convex, the (projected) gradient type methods [1,10,31] have been proposed achieving the best known complexity $\mathcal{O}(1/\epsilon)$ to find an ϵ-stationary point of problem (1.1) i.e., having $\|g_{X,k}\|^2 \le \epsilon$ for at least one iteration k [see (2.3) for definition of the so-called gradient mapping $g_{X,k}$ at the k-th iteration]. Recently, Nesterov's AG method has been also generalized for nonconvex problems achieving the aforementioned complexity bound [9,11]. It seems that for general first-order methods, strong convexity of h does not improve the complexity results when f is possibly nonconvex. Ghadimi et al. [9] also generalized the AG method and bundle-level type methods which possess the complexity bound $\mathcal{O}(1/\epsilon^{\frac{1+\nu}{2\nu}})$ for finding ϵ-stationary point of weakly smooth nonconvex problems. These methods are also uniformly optimal in the sense that they automatically achieve the optimal complexity bound if the problem turns out to be convex. Such uniform methods have been developed before only for convex programming [23,30].

All of the above-mentioned methods except the bundle-level type methods, generally known as proximal type methods, require solving subproblems in the form of

$$\min_{u \in X}\left\{\langle f'(x), u\rangle + h(u) + \frac{1}{\gamma}V(u,x)\right\} \tag{1.5}$$

at each iteration, where γ is a stepsize and $V(u,x)$ is a strongly convex function usually referred to Bregman distance function given by $V(u,x) = \omega(u) - \omega(x) - \langle w'(x), u - x\rangle$ for a strongly convex so-called prox-function ω.

Different class of methods include subproblems in which only linear approximation of the objective function is minimized over a bounded feasible set. The classical conditional gradient (CG) method seems to be the eldest one first proposed by Frank and Wolfe [7] and later widely used in the literature due to its low iteration cost (see e.g., [5,6,8,16,18,22,25]). While this class of algorithms generally achieves the iteration complexity of $\mathcal{O}(1/\epsilon)$ for solving smooth strongly convex problems, some of them exhibit linear rate of convergence under specific conditions on the feasible set (see e.g., [8,15,22,33]). Recently, Nesterov [29] showed that the CG method achieves the iteration complexity $\mathcal{O}(1/\epsilon^{\frac{1}{\nu}})$ for minimizing weakly smooth convex problems. In particular, he considered the class of problems given by (1.1) and showed that by modifying the linear optimization in the subproblems of the CG method to

$$\min_{u \in X}\{\langle f'(x), u\rangle + h(u)\}, \tag{1.6}$$

we can obtain the complexity bound $\mathcal{O}(1/\epsilon^{\frac{1}{2\nu}})$ for solving problem (1.1) when f is weakly smooth and convex (Ψ is strongly convex). Jiang and Zhang [20] also presented an algorithm involving subproblems in the form of (1.6) for solving problem (1.1) when f is possibly nonconvex and h is convex. This method exhibits iteration complexity of $\mathcal{O}(1/\epsilon^{\frac{1+\nu}{\nu}})$ for finding at least one k such that $g_k \le \epsilon$, where g_k, defined in (2.6), is a different termination criterion for finding approximate stationary points (see our

discussion after Algorithm 1 about the relation between g_k and the aforementioned gradient mapping $g_{X,k}$).

Our goal in this paper is to make a bridge between the classical CG method and general first-order methods including proximal type ones. Note that solving subproblem (1.5) can be more difficult than the above one due to the choice of ω. For example, while subproblem (1.6) has a closed form solution for some choices of h, we may need an iterative algorithm to solve subproblem (1.5) due to the existence of Bregman distance function. On the other hand, since variants of the classical CG method require only solving linear programming in each iteration, their iteration costs can be cheaper than solving subproblem (1.6) in general. It is also worth noting that while subproblems of the bundle-level type methods may consist of objective functions similar to that of (1.6), their feasible sets at each iteration are changed and hence are more complicated than X (see e.g., [23]). Indeed, iteration costs of these methods are computationally more expensive than the proximal type methods.

There are several choices of h where (1.6) is solved more efficiently than (1.5). A few examples are as follows:

- $h = I_X$, where I_X denotes the indicator function of the bounded set X.
- $h(x) = \lambda \|x\|_1$ or $h(x) = \lambda \|x\|_\infty$ for some $\lambda > 0$.
- $h(x) = \sum_{i=1}^n x_i \log x_i$ which is strongly convex w.r.t $\| \cdot \|_1$ with parameter 1 over the standard simplex.
- $h(x) = \frac{1}{2}\|x\|_p^2$ which is strongly convex w.r.t $\| \cdot \|_p$ with parameter $p - 1$, for $p \in (1, 2]$.
- $h(x) = \frac{1}{2}\|A\|_{S(p)}^2$, where $\|A\|_{S(p)} = \left(\sum_{i=1}^{\min(n,m)} x(A)_i^p \right)^{\frac{1}{p}}$ is the Schatten l_p norm for a matrix A with singular values $\{x(A)_i\}_{i=1}^{\min(n,m)}$. For any $p \in (1, 2]$, h is strongly convex w.r.t the Schatten l_p norm with parameter $p - 1$ ([21]).

It is worth noting that for the first two examples, subproblem (1.6) is reduced to a linear programming by possibly adding more variables and/or linear constraints and hence it is equivalent to the subproblems of the classical CG method. This type of subproblems can be solved very efficiently for special cases of the feasible set like polytopes. Also, for the rest of the examples where h is strongly convex, subproblem (1.6) has a closed form solution over their corresponding feasible sets i.e., simplex for the entropy function and the whole space for both squared l_p norm and the squared Schatten norm. For the latter, it is trivial that if X is any subset of the space containing the solution of the unconstrained case, it also admits that closed form solution. Note also that unlike the classical setting of the CG method, we can relax the boundedness assumption of the feasible set when h is strongly convex.

We should point out that when h is strongly convex and the prox-function is set to h, difficulty of solving (1.6) and (1.5) are comparable. However, the prox-function ω is usually considered to be continuously differentiable which may not be true for h in our setting. For examples, squared l_p norms ($p \in (1, 2)$) are not differentiable at points with at least one zero component. Moreover, except for very common choices of strongly convex function h like $\| \cdot \|_2^2$, there seems no convergence analysis for this kind of algorithms when applied to (stochastic) composite weakly smooth nonconvex problem. It is also worth noting the stepsize γ in (1.5) usually depends on the problem

parameters. On the other hand, since no stepsize is involved in (1.6), we can design a parameter free algorithm which is computationally much cheaper per iteration, when h is strongly convex, than those include subproblems in the form of (1.5) regardless of choice of the prox-function (see Algorithm 2 and the discussion after that for more details).

While the iteration cost of the methods including subproblems in the form of (1.6) is comparable or even better than that of the classical CG method for several examples of h and X (which some of them are mentioned above), it generally lies between that of having linear programming in the subproblems and general proximal type methods including subproblems in the form of (1.5). As we will show, the same fact holds for the iteration complexity of these methods. However, methods including (1.6) can also achieve the iteration complexity of the general proximal type methods under specific assumptions on problem (1.1) which consequently implies a better total computational complexity. For example, assume that f is a smooth nonconvex function, h is the squared l_p norm, X is set to $\mathbb{R}^n$, or the unit l_p norm ball. Then, our proposed methods, as summarized in Table 1, achieve iteration complexity of $\mathcal{O}(1/\epsilon)$ to find an ϵ-approximate stationary point of problem (1.1) while their subproblems have closed form solutions. Hence, the total number of (sub)gradient computations is still bounded by $\mathcal{O}(1/\epsilon)$. On the other hand, a few proximal type algorithms like the (projected) gradient method achieve the same $\mathcal{O}(1/\epsilon)$ iteration complexity ([9]). However, when the prox-function in subproblem (1.5) is set to a continuously differentiable strongly convex function like l_2 norm squared, then one needs an iterative algorithm to solve the subproblems. It can be easily verified that if the subproblems are solved within ϵ accuracy, then the iteration complexity of the (projected) gradient method does not change. Since the objective function of the subproblem in this example is nonsmooth and strongly convex, the (projected) subgradient method can solve it within $\mathcal{O}(1/\epsilon)$ number of iterations (subgradient computations of h). Therefore, the total number of (sub)gradient computations for this method is bounded by $\mathcal{O}(1/\epsilon^2)$ which is worse than the above-mentioned $\mathcal{O}(1/\epsilon)$ complexity for our proposed methods.

Our contribution in this paper consists of the following three aspects. First, we present a conditional gradient type (CGT) method for solving problem (1.1) which only differs from the classical CG method in incorporating function h into the subproblems as shown in (1.6). We show that the number of iterations performed by the CGT method to find an ϵ-approximate stationary point of problem (1.1), when g_k is used as the termination criterion, is bounded by

$$\mathcal{O}\left\{\frac{L^{\frac{1}{\nu}}}{\epsilon^{\frac{1+\nu}{\nu}}}\right\}, \quad \mathcal{O}\left\{\left(\frac{L_\nu}{\epsilon}\right)^{\frac{1}{\nu}}\log\frac{1}{\epsilon}\right\}, \tag{1.7}$$

when f is nonconvex or convex, respectively. In the former case, the first bound is similar to the one obtained in [20] for a variant of the CGT method with different stepsizes. In the latter, the above second bound also guarantees finding an ϵ-optimal solution of (1.1) and is in the same order, up to a logarithmic factor, of the one obtained in [29] with a different stepsize policy. However, it is still worse than the optimal

complexity bound of $\mathcal{O}(\epsilon^{-\frac{2}{1+3\nu}})$ for the general first-order methods. Moreover, if h is strongly convex, then by using a different stepsize policy, the above-mentioned bounds are, respectively, improved to

$$\mathcal{O}\left\{\left(\frac{L_\nu^2}{(\mu^\nu\epsilon)^{1+\nu}}\right)^{\frac{1}{2\nu}}\right\}, \quad \mathcal{O}\left\{\left(\frac{L_\nu^2}{\mu^{\nu(1+\nu)}\epsilon^{1-\nu}}\right)^{\frac{1}{2\nu}}\log\frac{1}{\epsilon}\right\}. \tag{1.8}$$

The first bound, disregarding μ, is the best-known iteration complexity achieved by the first-order methods when applied to (composite) weakly smooth nonconvex problems. Recently, Ghadimi et al. [9] obtained a similar bound for a class of uniformly optimal first-order methods. Note however, that while function h in (1.1) is assumed to be only convex in [9], their algorithms include subproblems in the form of (1.5) for the choice of $\omega(x) = \frac{1}{2}\|x\|_2^2$ and are more complicated than the CGT method. On the other hand, when f is convex, the second bound is worse than the optimal complexity bound of $\mathcal{O}(\epsilon^{\frac{\nu-1}{1+3\nu}})$ for first-order methods for minimizing the class of weakly smooth strongly convex functions ([4,17,28]). Note however, that if f is smooth ($\nu = 1$), it is reduced to the optimal complexity bound $\mathcal{O}(\log\frac{1}{\epsilon})$ for solving smooth strongly convex problems. It should be also mentioned that we have a unified analysis of the CGT method to obtain these bound in the sense that the same stepsize policy is used regardless of the convexity of f. However, knowing the strong convexity of h helps us to get better results by modifying the stepsize policy.

Second, while obtaining the above-mentioned complexity bounds requires explicitly estimating problem parameters L_ν, ν, and μ, we also present two variants of the CGT method which relax such estimations. In particular, when $\mu > 0$, we equip the CGT method with a line search procedure which only uses the target accuracy as a parameter in its implementation. On the other hand, since there is no stepsize involved in the subproblem (1.6), this line search does not require any additional effort for computing extra gradients of f and/or solving more subproblems. In this sense, this variant of the CGT method is computationally much cheaper per iteration than the other parameter free algorithms like the ones in [4,9,30] using subproblems in the form of (1.5) for solving weakly smooth problems. It is worth noting that such a parameter free algorithm is much more desirable in a black-box optimization where only first-order information of f in (1.1) is given through an oracle and its convexity and level of smoothness may not be exactly known. This variant of the CGT method achieves the same complexity bounds in (1.8). Furthermore, we present another variant of the CGT method which does not require h to be strongly convex, any line search procedure, and the target accuracy in advance. However, its complexity bounds are slightly worse than the aforementioned ones for the CGT method.

Finally, we consider problem (1.1) under stochastic setting where only noisy first-order information of f is available via subsequent calls to a stochastic oracle ($\mathcal{SO}$). More specifically, when $\mathcal{SO}$ receives $x_k \in X$ as the input at the k-th call, it outputs stochastic gradient $G(x_k, \xi_k)$, where $\{\xi_k\}_{k\geq 1}$ are random vectors whose distributions P_k are supported on $\Xi_k \subseteq \mathbb{R}^d$. The following assumption is made for the stochastic gradients returned by the $\mathcal{SO}$.

Assumption 1 For any $x \in X$ and $k \geq 1$, we have

$$\text{(a)} \quad \mathbb{E}[G(x, \xi_k)] = f'(x), \tag{1.9}$$

$$\text{(b)} \quad \mathbb{E}\left[\|G(x, \xi_k) - f'(x)\|_*^2 \right] \leq \sigma^2. \tag{1.10}$$

We present a randomized stochastic CGT (RSCGT) method for solving problem (1.1) and show that the total number of calls to the $\mathcal{SO}$ performed by this method to have $\mathbb{E}[g_R] \leq \epsilon$ is bounded by

$$\mathcal{O}\left\{ \sigma^2 \left(\frac{L_\nu^{1+\nu}}{\epsilon^{1+3\nu}} \right)^{\frac{1}{\nu}} \right\}, \quad \mathcal{O}\left\{ \sigma^2 \left(\frac{L_\nu}{\epsilon^{1+2\nu}} \right)^{\frac{1}{\nu}} \right\}, \tag{1.11}$$

when f is nonconvex or convex, respectively. Here, the expectation is taken w.r.t ξ and R, where R is a discrete random variable whose probability mass function is supported on $\{1, 2, \ldots, N\}$ for some given iteration limit $N \geq 1$. This type of randomized framework for stochastic optimization first was proposed by Ghadimi and Lan [12] and later used in [3,10,11], for general stochastic approximation algorithms applying to smooth (un)constrained problems. To the best of our knowledge, the complexity bounds in (1.11) seems to be the first ones for the conditional gradient type methods when applied to stochastic weakly smooth nonconvex and convex problems. Note that the first bound is reduced to $\mathcal{O}(1/\epsilon^4)$, when $\nu = 1$, which is also recently obtained by [34] for the classical CG method ($h \equiv 0$) when applied to stochastic smooth nonconvex problems.[1] Moreover, when h is strongly convex, the above bounds are improved to

$$\mathcal{O}\left\{ \sigma^2 \left(\frac{L_\nu^{1+2\nu}}{(\mu\epsilon)^{\frac{1+4\nu}{2}}} \right)^{\frac{1}{\nu}} \right\}, \quad \mathcal{O}\left\{ \sigma^2 \left(\frac{L_\nu}{(\mu\epsilon)^{\frac{1+2\nu}{2}}} \right)^{\frac{1}{\nu}} \right\}. \tag{1.12}$$

The first bound is better than $\mathcal{O}(\epsilon^{-\frac{(1+\nu)^2}{\nu}})$ obtained in [20] (when h is convex) for a stochastic gradient type method including subproblems in the form of (1.5) in which the Bregman distance function is replaced by $\| \cdot \|^{1+\nu}$ and under stronger assumption than Assumption 1(b). In addition, when $\nu = 1$ and $\mu > 0$, by using different stepsize policy we can further improve the complexity bounds in (1.12) to

$$\mathcal{O}\left\{ \frac{\sigma^2 L_1}{\mu^2 \epsilon^2} \right\}, \quad \mathcal{O}\left\{ \frac{\sigma^2 L_1}{\mu^2 \epsilon} \log \frac{1}{\epsilon} \right\}.$$

Note that the first bound is in the same order of the one obtained in [10,11] for stochastic gradient type algorithms, and the second one is optimal up to a logarithmic constant for solving stochastic smooth strongly convex optimization problems in terms of dependence on the accuracy ϵ (see e.g., [13,27]). We should mention that the bounds

---

[1] Reddi et al. [34] was released several months after releasing the first version of this work.

Table 1 Complexity results for algorithms including subproblems in the form of (1.6) for finding an ϵ-approximate stationary point (or optimal point when f is convex) of problem (1.1) when $\nu \in (0, 1]$

Algorithms	f	h	σ	Complexity
[20]	Nonconvex	Convex	0	$\mathcal{O}\left(\epsilon^{-\frac{1+\nu}{\nu}}\right)$
[29]	Convex	Convex	0	$\mathcal{O}\left(\epsilon^{-\frac{1}{\nu}}\right)$
		Strongly convex		$\mathcal{O}\left(\epsilon^{-\frac{1}{2\nu}}\right)$
This paper	Nonconvex	Convex	0	$\mathcal{O}\left(\epsilon^{-\frac{1+\nu}{\nu}}\right)$
		Strongly convex		$\mathcal{O}\left(\epsilon^{-\frac{1+\nu}{2\nu}}\right)$
		Convex	> 0	$\mathcal{O}\left(\epsilon^{-\frac{1+3\nu}{\nu}}\right)$
		Strongly convex		$\mathcal{O}\left(\epsilon^{-\frac{1+4\nu}{2\nu}}\right)$
	Convex	Convex	0	$\mathcal{O}\left(\epsilon^{-\frac{1}{\nu}}\log\frac{1}{\epsilon}\right)$
		Strongly convex		$\mathcal{O}\left(\epsilon^{\frac{\nu-1}{2\nu}}\log\frac{1}{\epsilon}\right)$
		Convex	> 0	$\mathcal{O}\left(\epsilon^{-\frac{1+2\nu}{\nu}}\right)$
		Strongly convex		$\mathcal{O}\left(\epsilon^{-\frac{1+2\nu}{2\nu}}\right)$

in are also obtained in [24] for a method using only linear optimization in subproblems when applied to smooth stochastic optimization. A brief summary of comparing our results with the ones for conditional gradient type methods when applied to weakly smooth optimization problems is given in Table 1.

Rest of the paper is organized as follows. In Sect. 2, we present the CGT method and its variants with establishing their convergence properties for solving problem (1.1). In Sect. 3, we generalize some of our results to the stochastic setting of problem (1.1). Finally, we conclude the paper by some remarks in Sect. 4.

1.1 Notation

For a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, $h'(x)$ is the gradient of h at x. More generally, when h is a proper convex function, $p_h(x) \in \partial h(x)$ denotes a subgradient of h in the subdifferential set of h at x. For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, $\langle x, y \rangle$ is the standard inner product in $\mathbb{R}^n$, and $\|\cdot\|$ denotes a general norm unless otherwise is mentioned. For any real number r, $\lceil r \rceil$ and $\lfloor r \rfloor$ denote the nearest integer to r from above and below, respectively. We also denote the diameter of set X (whenever it is assumed to

be bounded) by $D_X = \max_{x,y \in X} \|x - y\|$. When f is convex, an optimal solution of (1.1) is denoted by $x^* \in X$ and $\Psi^* = \Psi(x^*)$. $\mathcal{O}(1)$ also stands for a positive constant.

2 Conditional gradient type methods for composite optimization

Our goal in this section is to present a conditional gradient type (CGT) method for solving problem (1.1). Note that the classical CG method cannot be directly applied to this problem since h is not necessarily differentiable. However, it is shown that if h is Lipschitz continuous, then it can be approximated by using some smoothing techniques and hence the CG method can be used for solving problem (1.1) when f is smooth and convex [22]. Different from the classical CG method, we minimize summation of the (strongly) convex function h and a linear approximation of function f in the subproblems of the CGT method.

Below we precisely describe the conditional gradient type (CGT) method.

Algorithm 1 The conditional gradient type (CGT) algorithm

Input: $x_0 \in X$, $\{\alpha_k\}_{k \geq 1} \in (0, 1]$.
0. Set $y_0 = x_0$ and $k = 1$.
1. Compute $f'(y_{k-1})$ and set

$$x_k = w(f'(y_{k-1})) := \arg\min_{u \in X} \left\{ \langle f'(y_{k-1}), u \rangle + h(u) \right\}, \qquad (2.1)$$

$$y_k = (1 - \alpha_k)y_{k-1} + \alpha_k x_k, \qquad (2.2)$$

2. Set $k \leftarrow k + 1$ and go to step 1.

We now add a few remarks about the above algorithm. First, note that as mentioned above, the CGT method only differs from the classical CG method in the existence of h in subproblem (2.1). As we will show, while this may make the subproblems harder to solve, it improves rate of convergence of the CG method when h is strongly convex. Second, as mentioned in Sect. 1, there are several examples of h where solving (2.1) is not harder than subproblems of the classical CG methods. Moreover, when h is strongly convex, by properly choosing stepsizes α_k, we can relax the boundedness assumption of feasible set X required in the setting of iterative methods minimizing only linear approximation of the objective function at each iteration. Third, note that Algorithm 1 is essentially the same as the ones proposed in [20,29]. However, we consider different assumptions on the objective function and discuss different choices of stepsize policies for this algorithm which improve its existing complexity bounds for solving both (strongly) convex and nonconvex problems. Finally, since f is possibly nonconvex in (1.1), we need to define a different termination criterion to establish the convergence of Algorithm 1. To do so, we can define gradient mapping at point $x_k \in X$ as

$$g_{X,k} = g_X(w(f'(y_{k-1}))) := y_{k-1} - x_k. \qquad (2.3)$$

Note that the extra notation of $w(\cdot)$ will be helpful in our convergence analysis when we consider the stochastic setting of problem (1.1). The equivalent definitions of the gradient mapping have been widely used based on the updating step of different algorithms (see e.g., [11,32]). Note that by the optimality condition of (2.1) and for all $x \in X$, we have

$$\langle -[f'(y_{k-1}) + p_h(x_k)], x - x_k \rangle \le 0 \to \langle -p_\Psi(x_k), x - x_k \rangle$$
$$\le \langle f'(y_{k-1}) - f'(x_k), x - x_k \rangle. \quad (2.4)$$

Moreover, if X is bounded, then by Cauchy–Schwarz inequality and (1.3), we have

$$\langle f'(y_{k-1}) - f'(x_k), x - x_k \rangle \le \|f'(y_{k-1}) - f'(x_k)\|_* \|x - x_k\|$$
$$\le L_v D_X \|y_{k-1} - x_k\|^v = L_v D_X \|g_{X,k}\|^v. \quad (2.5)$$

Hence, if $\|g_{X,k}\| \le [\epsilon/(L_v D_X)]^{\frac{1}{v}}$, then x_k is an ϵ-approximate stationary point of problem (1.1) i.e.,

$$\langle -p_\Psi(x_k), x - x_k \rangle \le \epsilon \ \forall \, x \in X.$$

When f is convex, the above relation also implies $\Psi(x_k) - \Psi(x^*) \le \epsilon$.

Another termination criterion can be defined as

$$g_k = g(w(f'(y_{k-1}))) := \langle f'(y_{k-1}), y_{k-1} \rangle$$
$$+ h(y_{k-1}) + \max_{x \in X}\{-\langle f'(y_{k-1}), x \rangle - h(x)\}$$
$$= \langle f'(y_{k-1}), y_{k-1} - x_k \rangle + h(y_{k-1}) - h(x_k) \ge 0, \quad (2.6)$$

where the last equality is due to (2.1) and the inequality follows from convexity of h and (2.4). Similarly, for any $x \in X$, we have

$$\langle -f'(y_{k-1}), x - y_{k-1} \rangle + h(y_{k-1}) - h(x) \le g_k,$$
$$\langle -f'(y_{k-1}), x - y_{k-1} \rangle + h(y_{k-1}) - h(x) \le \langle -p_\Psi(y_{k-1}), x - y_{k-1} \rangle. \quad (2.7)$$

Observe that if y_{k-1} is a stationary point of problem (1.1), then RHS of the above inequality is nonpositive for all $x \in X$. Hence, setting $x = x_k$, the LHS equals g_k which becomes nonpositive and together with (2.6) imply that g_k is zero. Therefore, $g_k = 0$ is a necessary condition for y_{k-1} to be a stationary point. Note, however, that (2.4) implies that $\|g_{X,k}\| = 0$ is a sufficient condition for $x_k(= y_{k-1})$ to be a stationary point. It can be also seen that when $h \equiv 0$, both criteria are equivalent (at different points). In this case, g_k will be the Frank–Wolf gap defined in the literature of conditional gradient methods. Hence, it makes sense to consider it as a generalized Frank–Wolf gap when h exists.

On the other hand, when f is convex, by (2.6) we have

$$g_k = h(y_{k-1}) + \max_{x \in X}\{\langle f'(y_{k-1}), y_{k-1} - x \rangle - h(x)\}$$
$$\geq h(y_{k-1}) + \max_{x \in X}\{f(y_{k-1}) - [f(x) + h(x)]\} = \Psi(y_{k-1}) - \Psi(x^*), \quad (2.8)$$

which implies that g_k provides an upper bound on the optimality gap $\Psi(y_{k-1}) - \Psi(x^*)$. Next result summarizes the relation between g_k and $g_{X,k}$. It should be pointed out that a similar result stating the relation between g_k and the gradient mapping for proximal methods have been proposed in [20] when h is only convex.

Lemma 1 *Let $\{x_k, y_k\}_{k \geq 0}$ be generated by Algorithm 1. Also, suppose that $g_{X,k}$ and g_k are defined in (2.3) and (2.6), respectively.*

(a) *If there exist positive constants M_f and M_h such that $\|f'(x)\|_* \leq M_f$ and $\|p_h(x)\|_* \leq M_h$ for any $x \in X$, then we have $g_k \leq (M_f + M_h)\|g_{X,k}\| \ \forall \, k \geq 1$.*
(b) *If h is strongly convex with parameter $\mu > 0$, then we have $\|g_{X,k}\|^2 \leq 2g_k/\mu \ \forall \, k \geq 1$.*

Proof First, note that by (2.6), convexity of h, Cauchy-Schwarz, and triangle inequalities, we have

$$g_k \leq \langle f'(y_{k-1}) + p_h(y_{k-1}), y_{k-1} - x_k \rangle \leq \|f'(y_{k-1}) + p_h(y_{k-1})\|_* \|y_{k-1} - x_k\|$$
$$\leq (\|f'(y_{k-1})\|_* + \|p_h(y_{k-1})\|_*)\|y_{k-1} - x_k\|,$$

which together with (2.3) and boundedness assumption on (sub)gradients of f and h imply part a). Second, by (2.6), strong convexity of h, and optimality condition of (2.1), we have

$$g_k \geq \langle f'(y_{k-1}) + p_h(x_k), y_{k-1} - x_k \rangle + \frac{\mu}{2}\|y_{k-1} - x_k\|^2 \geq \frac{\mu}{2}\|y_{k-1} - x_k\|^2,$$

which together with (2.3) imply part (b). □

While, g_k denotes a weaker notation of approximate stationary points than $\|g_{X,k}\|$ when X is bounded and h is only convex, the latter cannot be theoretically guaranteed to be convergent. Furthermore, when h is strongly convex with $\mu > 0$, Lemma 1(b) implies that g_k is an upper bound on $\|g_{X,k}\|^2$. In this case, $g_k = 0$ is a necessary and sufficient condition for y_{k-1} to be a stationary point of problem (1.1). Therefore, to include more general cases of h and X, we provide convergence analysis of Algorithm 1 and its variants based on g_k throughout the paper.

Below we establish main convergence properties of Algorithm 1.

Theorem 1 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 1.*

(a) *If stepsizes $\alpha_k \in (0, 1]$ are chosen such that*

$$\alpha_k^\nu \leq \frac{(1 + \nu)g_k}{4L_\nu\|x_k - y_{k-1}\|^{1+\nu}} \quad \forall \, k \geq 1, \quad (2.9)$$

then for any $N \geq 1$, we have

$$\sum_{k=1}^{N} \alpha_k g_k \leq \frac{4}{3}[\Psi(x_0) - \Psi^*]. \tag{2.10}$$

If, in addition, f is convex, then we have

$$\frac{\Psi(y_N) - \Psi^*}{A_N} + \sum_{k=1}^{N} \frac{\alpha_k g_k}{4A_k} \leq \Psi(x_0) - \Psi^*, \tag{2.11}$$

where

$$A_0 = 1, \quad A_N = \prod_{k=1}^{N} \left(1 - \frac{\alpha_k}{2}\right) \quad \forall N \geq 1. \tag{2.12}$$

(b) *If h is strongly convex with parameter $\mu > 0$ and stepsizes $\alpha_k \in (0, 1]$ are chosen such that*

$$\alpha_k^{\nu} \leq \frac{(1 + \nu)g_k^{\frac{1-\nu}{2}}}{4L_\nu} \left(\frac{\mu}{2}\right)^{\frac{1+\nu}{2}} \quad \forall k \geq 1, \tag{2.13}$$

then for any $N \geq 1$, we still obtain (2.10) and (2.11) when f is nonconvex or convex, respectively.

Proof We first show part (a). Noting (1.4) and (2.2), we have

$$
\begin{aligned}
f(y_k) &\leq f(y_{k-1}) + \langle f'(y_{k-1}), y_k - y_{k-1}\rangle + \frac{L_\nu}{1+\nu}\|y_k - y_{k-1}\|^{1+\nu} \\
&= f(y_{k-1}) - \alpha_k\left[\langle f'(y_{k-1}), y_{k-1} - x_k\rangle - \frac{L_\nu \alpha_k^{\nu}}{1+\nu}\|x_k - y_{k-1}\|^{1+\nu}\right].
\end{aligned}
\tag{2.14}
$$

Also, noting (1.2) and (2.2), we have

$$h(y_k) \leq (1 - \alpha_k)h(y_{k-1}) + \alpha_k h(x_k) - \frac{\alpha_k(1 - \alpha_k)\mu}{2}\|x_k - y_{k-1}\|^2. \tag{2.15}$$

Adding the above two inequalities and noting (2.6), we obtain

$$\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k\left[g_k - \frac{L_\nu \alpha_k^{\nu}}{1+\nu}\|x_k - y_{k-1}\|^{1+\nu} + \frac{(1 - \alpha_k)\mu}{2}\|x_k - y_{k-1}\|^2\right], \tag{2.16}$$

which after re-arranging the terms and noting (2.9) imply that

$$\frac{3\alpha_k g_k}{4} \le \Psi(y_{k-1}) - \Psi(y_k) \ \forall \, k \ge 1.$$

Summing up the above inequalities and noting the fact that $\Psi^* \le \Psi(y_k) \ \forall \, k \ge 0$, we obtain (2.10). Moreover, if f is convex, then by subtracting Ψ^* from both sides of (2.16), we have

$$
\begin{aligned}
\Psi(y_k) - \Psi^* &\le \left(1 - \frac{\alpha_k}{2}\right) \left[\Psi(y_{k-1}) - \Psi^*\right] + \frac{\alpha_k}{2} \left[\Psi(y_{k-1}) - \Psi^*\right] \\
&\quad - \alpha_k \left[g_k - \frac{L_\nu \alpha_k^\nu}{1+\nu} \|x_k - y_{k-1}\|^{1+\nu}\right] \\
&\le \left(1 - \frac{\alpha_k}{2}\right) \Psi(y_{k-1}) - \alpha_k \left[\frac{g_k}{2} - \frac{L_\nu \alpha_k^\nu}{1+\nu} \|x_k - y_{k-1}\|^{1+\nu}\right] \\
&\le \left(1 - \frac{\alpha_k}{2}\right) \left[\Psi(y_{k-1}) - \Psi^*\right] - \frac{\alpha_k g_k}{4},
\end{aligned}
\tag{2.17}
$$

where the second and third inequalities are followed by (2.8) and (2.9), respectively. Dividing both sides of the above inequality by A_k, noting (2.12), summing them up, and re-arranging the terms, we obtain (2.11).

We now show part (b). Suppose that h is strongly convex with parameter $\mu > 0$. Then, by (2.3), Lemma 1(b), and (2.16), we have

$$\Psi(y_k) \le \Psi(y_{k-1}) - \alpha_k \left[g_k - \frac{L_\nu \alpha_k^\nu}{1+\nu} \left(\frac{2g_k}{\mu}\right)^{\frac{1+\nu}{2}}\right], \tag{2.18}$$

which together with (2.13) and similarly to part (a), imply (2.10) and (2.11). $\qquad\square$

Observe that the convergence results in Theorem 1 rely on the choice of stepsizes $\{\alpha_k\}_{k\ge1}$ which should satisfy (2.9) or (2.13). Below we specialize these results for specific choices of stepsizes.

Corollary 1 *Suppose that the sequence $\{x_k, y_k\}_{k\ge0}$ is generated by Algorithm 1 and a target accuracy $\epsilon > 0$ is given.*

(a) *Let X be bounded and stepsizes $\{\alpha_k\}$ set to*

$$\alpha_k = \left(\frac{g_k}{g_k + \frac{4L_\nu}{1+\nu}\|x_k - y_{k-1}\|^{1+\nu}}\right)^{\frac{1}{\nu}} \ \forall \, k \ge 1, \tag{2.19}$$

then for any $N \ge 1$, we have

$$g_{\hat{N}} \le \max\left\{\frac{4\left[\Psi(x_0) - \Psi^*\right]}{3N\bar{\alpha}_{\epsilon,\nu}}, \epsilon\right\}, \tag{2.20}$$

where

$$\hat{N} = \arg \min_{k \in \{1, \ldots, N\}} g_k, \tag{2.21}$$

and

$$\bar{\alpha}_{\epsilon, \nu} = \left(\frac{\epsilon}{\epsilon + \frac{4L_\nu}{1+\nu} D_X^{1+\nu}} \right)^{\frac{1}{\nu}}. \tag{2.22}$$

If, in addition, f is convex, then we have

$$\Psi(y_{\hat{N}-1}) - \Psi^* \leq g_{\hat{N}} \leq \max \left\{ \frac{4[\Psi(x_0) - \Psi^*]\left(1 - \frac{\bar{\alpha}_{\epsilon, \nu}}{2}\right)^N}{\bar{\alpha}_{\epsilon, \nu}}, \epsilon \right\}. \tag{2.23}$$

(b) *If h is strongly convex with parameter $\mu > 0$ and stepsizes $\{\alpha_k\}$ are set to*

$$\alpha_k = \left(\frac{g_k^{\frac{1-\nu}{2}}}{g_k^{\frac{1-\nu}{2}} + \frac{4L_\nu}{1+\nu}\left(\frac{2}{\mu}\right)^{\frac{1+\nu}{2}}} \right)^{\frac{1}{\nu}} \quad \forall \, k \geq 1, \tag{2.24}$$

then we have (2.20) in which $\bar{\alpha}_{\epsilon, \nu}$ is replaced by

$$\alpha_{\epsilon, \nu} = \left(\frac{\epsilon^{\frac{1-\nu}{2}}}{\epsilon^{\frac{1-\nu}{2}} + \frac{4L_\nu}{(1+\nu)}\left(\frac{2}{\mu}\right)^{\frac{1+\nu}{2}}} \right)^{\frac{1}{\nu}}. \tag{2.25}$$

With this replacement, we also obtain (2.23) when f is convex.

Proof First, observe that the choice of stepsizes in (2.19) satisfies (2.9). Now, if there exists a $k \in \{1, \ldots, N\}$ such that $g_k \leq \epsilon$, then both (2.20) and (2.23) clearly hold due to (2.21). If this is not the case, then we must have $g_k > \epsilon$ $k = 1, \ldots, N$ which together with (2.22), and boundedness of X imply that

$$\alpha_k \geq \bar{\alpha}_{\epsilon, \nu} \; k = 1, \ldots, N, \quad \sum_{k=1}^{N} \alpha_k g_k \geq \sum_{k=1}^{N} \bar{\alpha}_{\epsilon, \nu} g_k \geq N \bar{\alpha}_{\epsilon, \nu} g_{\hat{N}}. \tag{2.26}$$

Hence, (2.20) still holds by noting (2.10). Second, noting (2.12) and the first inequality in (2.26), we have

$$A_N = \prod_{k=1}^{N}\left(1 - \frac{\alpha_k}{2}\right) \leq \prod_{k=1}^{N}\left(1 - \frac{\bar{\alpha}_{\epsilon,\nu}}{2}\right) = \left(1 - \frac{\bar{\alpha}_{\epsilon,\nu}}{2}\right)^N,$$

$$\sum_{k=1}^{N} \frac{\alpha_k}{2A_k} = \sum_{k=1}^{N}\left[\frac{1}{A_k} - \frac{1 - \frac{\alpha_k}{2}}{A_k}\right] = \sum_{k=1}^{N}\left[\frac{1}{A_k} - \frac{1}{A_{k-1}}\right] = \frac{1}{A_N} - 1$$

$$\geq \frac{\bar{\alpha}_{\epsilon,\nu}}{2}\left(1 - \frac{\bar{\alpha}_{\epsilon,\nu}}{2}\right)^{-N}. \tag{2.27}$$

Combining the above relations with (2.8) under convexity of f, (2.11), and noting that $\Psi(y_N) \geq \Psi^*$, we obtain (2.23). Third, note that when $\mu > 0$, the choice of stepsizes in (2.24) satisfies (2.13) and then (2.26), (2.27) hold by replacing $\bar{\alpha}_{\epsilon,\nu}$ with $\alpha_{\epsilon,\nu}$ due to (2.25). Hence, (2.20) and (2.23) still hold with this replacement. □

We now add a few remarks about the above results. First, observe that by (2.22), (2.25), and convexity of the univariate function $x \to x^r (r > 1)$ on $(0, \infty)$, we have

$$\frac{1}{\bar{\alpha}_{\epsilon,\nu}} = \left[1 + \frac{2L_\nu D_X^{1+\nu}}{(1+\nu)\epsilon}\right]^{\frac{1}{\nu}} \leq 2^{\frac{1-\nu}{\nu}}\left[1 + \left(\frac{2L_\nu D_X^{1+\nu}}{(1+\nu)\epsilon}\right)^{\frac{1}{\nu}}\right], \quad \frac{1}{\alpha_{\epsilon,\nu}}$$

$$\leq 2^{\frac{1-\nu}{\nu}}\left[1 + \left(\frac{4L_\nu}{(1+\nu)\epsilon^{\frac{1-\nu}{2}}}\right)^{\frac{1}{\nu}}\left(\frac{2}{\mu}\right)^{\frac{1+\nu}{2}}\right], \tag{2.28}$$

which together with (2.20), imply that if stepsizes are set to (2.19), then the total number of iterations performed by Algorithm 1 to have $g_k \leq \epsilon$ for at least one k, is bounded by

$$\mathcal{O}(1) \max\left\{1, \left(\frac{L_\nu D_X^{1+\nu}}{\epsilon}\right)^{\frac{1}{\nu}}\right\}\left\{\frac{\Psi(x_0) - \Psi^*}{\epsilon}\right\}. \tag{2.29}$$

This bound (when the second term is the dominating one) is also obtained in [20] for Algorithm 1 with different choice of stepsizes. Moreover, when f is convex, the above complexity bound is improved to

$$\mathcal{O}(1) \max\left\{1, \left(\frac{L_\nu D_X^{1+\nu}}{\epsilon}\right)^{\frac{1}{\nu}}\right\}\left\{\log\frac{\Psi(x_0) - \Psi^*}{\epsilon} + \frac{1}{\nu}\log\frac{L_\nu D_X^{1+\nu}}{\epsilon}\right\}, \tag{2.30}$$

which also guarantees fining a solution $\bar{x} \in X$ such that $\Psi(\bar{x}) - \Psi(x^*) \leq \epsilon$. This bound, up to a logarithmic factor, is similar to the bound $\mathcal{O}(\epsilon^{-\frac{1}{\nu}})$ bound obtained

by Nesterov [29] for Algorithm 1 with different stepsize policy not depending on the problem parameters. However, it is worse than the optimal complexity bound of $\mathcal{O}(\epsilon^{-\frac{2}{1+3\nu}})$ for general first-order methods applied to weakly smooth convex optimization problems [27].

Second, if $\mu > 0$ and stepsizes are set to (2.19), then the complexity bound (2.29) is improved to

$$\mathcal{O}(1) \max \left\{ 1, \left(\frac{L_\nu^2}{\mu^{\nu(1+\nu)} \epsilon^{1-\nu}} \right)^{\frac{1}{2\nu}} \right\} \left\{ \frac{\Psi(x_0) - \Psi^*}{\epsilon} \right\}. \tag{2.31}$$

This bound (when the second term is the dominating one) is obtained by Ghadimi et al. [9] for a class of gradient type methods when applied to weakly smooth nonconvex problems. However, their methods do not require the knowledge of problem parameters in advance. It should be also mentioned that when $\nu = 1$, (2.31) is reduced to the best known complexity $\mathcal{O}(1/\epsilon)$ for first-order methods when applied to smooth nonconvex problems. Furthermore, observe that if f is convex (Ψ is strongly convex), then the bounds in (2.30) and (2.31) are improved to

$$\mathcal{O}(1) \max \left\{ 1, \left(\frac{L_\nu^2}{\mu^{\nu(1+\nu)} \epsilon^{1-\nu}} \right)^{\frac{1}{2\nu}} \right\} \left\{ \log \frac{\Psi(x_0) - \Psi^*}{\epsilon} + \frac{1}{\nu} \log \frac{L_\nu^2}{\mu^{\nu(1+\nu)} \epsilon^{1-\nu}} \right\}, \tag{2.32}$$

which also guarantees fining a solution $\bar{x} \in X$ such that $\Psi(\bar{x}) - \Psi(x^*) \leq \epsilon$. This bound also much better than the complexity of $\mathcal{O}(\epsilon^{-\frac{1}{2\nu}})$ obtained by Nesterov [29] for Algorithm 1 for solving strongly convex problem. However, it is worse than the optimal complexity bound $\mathcal{O}(\epsilon^{\frac{\nu-1}{1+3\nu}})$ for minimizing weakly smooth strongly convex problems which is achieved by a few more complicated first-order methods [4,17]. It should be also mentioned that these methods do not consider the composite setting. On the other hand (2.32) guarantees linear rate of convergence for Algorithm 1 when f is smooth, convex and h is strongly convex.

Finally, observe that our choices of stepsizes in (2.19) and (2.24) do not depend on the convexity of f which provide unified analysis for Algorithm 1. Moreover, strong convexity of h significantly improves the existing complexity bounds obtained by conditional gradient methods for both convex and nonconvex problems and relaxes the boundedness assumption of the feasible set. On the other hand, these stepsize policies depend on the problem parameters like ν and L_ν which may not be known exactly. Below we present a variant of Algorithm 1 which does not require such knowledge of parameters in advance. It should be mentioned that convergence of this algorithm can be only guaranteed when h is strongly convex.

Note that in the framework of the CGT method, y_k is indeed a weighted average of x_i $i = 0, 1, \ldots, k$ due to (2.2). Hence, purpose of the line search procedure defined in (2.33) and (2.34) is to only find the appropriate choice of α_k to specify these weights

---

Algorithm 2 The conditional gradient type algorithm with line search

---

Input: $x_0 \in X$, $\{\alpha_k\}_{k \geq 1} \in (0, 1]$, $\gamma \in (0, 1)$, and $\delta > 0$.

0. Set $y_0 = x_0$ and $k = 1$.

1. Compute $f'(y_{k-1})$ and set x_k to (2.1).

2. Set y_k to (2.2) with the choice of

$$\alpha_k = \gamma^{t_k} \quad \forall \, k \geq 1, \qquad (2.33)$$

where $t_k \geq 1$ is the smallest integer number such that

$$\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k g_k + \delta \alpha_k \qquad (2.34)$$

holds, where g_k is defined in (2.6).

3. Set $k \leftarrow k + 1$ and go to step 1.

---

and subproblem (2.1) is solve once at each iteration. Similar procedures have been also used for weakly smooth convex and nonconvex problems in [9,30]. For each run of these line search procedures, a new subproblem is to be solved and gradient computation of the objective function at a new point may be required. Note however, that our procedure does not require this additional effort and hence it is computationally much cheaper per iteration.

In the next result, we show that the line search procedure defined in Step 2 of Algorithm 2 is run only for a finite number of times and hence the complexity of this algorithm is similar to that of Algorithm 1 when h is strongly convex.

Theorem 2 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 2 and h is strongly convex with parameter $\mu > 0$. Then, the following statements hold.*

(a) *The procedure (2.34) is well-defined in the sense that t_k in (2.33) is finite for any $k \geq 1$.*

(b) *For any $N \geq 1$, we have*

$$\sum_{k=1}^{N} \alpha_k g_k \leq \Psi(x_0) - \Psi^* + \delta \sum_{k=1}^{N} \alpha_k. \qquad (2.35)$$

If, in addition, f is convex, then for any $N \geq 1$, we have

$$\frac{\Psi(y_N) - \Psi^*}{A_N} + \sum_{k=1}^{N} \frac{\alpha_k g_k}{2A_k} \leq \Psi(x_0) - \Psi^* + \delta \sum_{k=1}^{N} \frac{\alpha_k}{A_k}, \qquad (2.36)$$

where A_k is defined in (2.12).

Proof We first show part (a). Noting Lemma 1(b) and Hölder inequality $ab \leq a^p/p + b^q/q$ with $p = \frac{2}{1+\nu}$ and $q = \frac{2}{1-\nu}$ for $\nu \in (0, 1)$, we have

$$\frac{L_\nu(\alpha_k \|x_k - y_{k-1}\|)^{1+\nu}}{1+\nu} = \alpha_k \left[\frac{L_\nu \alpha_k^\nu \|x_k - y_{k-1}\|^{1+\nu}}{1+\nu} \left(\frac{(1-\nu)}{2\delta} \right)^{\frac{1-\nu}{2}} \right] \left[\frac{2\delta}{(1-\nu)} \right]^{\frac{1-\nu}{2}}$$

$$\leq \frac{\hat{L}_{\nu,\delta} \alpha_k^{\frac{1+3\nu}{1+\nu}}}{2} \|x_k - y_{k-1}\|^2 + \delta\alpha_k, \tag{2.37}$$

where

$$\hat{L}_{\nu,\delta} = \left\{ \frac{L_\nu}{\left[\frac{2(1+\nu)\delta}{1-\nu} \right]^{\frac{1-\nu}{2}}} \right\}^{\frac{2}{1+\nu}} \quad \text{and} \quad \hat{L}_{1,\delta} = \lim_{\nu \to 1} \hat{L}_{\nu,\delta} = L_1. \tag{2.38}$$

Combining the above relation with (2.16), we obtain

$$\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k g_k - \frac{\alpha_k \mu}{2} \left[1 - \alpha_k - \frac{\hat{L}_{\nu,\delta} \alpha_k^{\frac{2\nu}{1+\nu}}}{\mu} \right] \|x_k - y_{k-1}\|^2 + \delta\alpha_k. \tag{2.39}$$

Noting $\alpha_k \leq \alpha_k^{\frac{2\nu}{1+\nu}}$ due to the facts that $\nu, \alpha_k \in (0, 1]$, if

$$\alpha_k \leq \left(\frac{1}{1 + \frac{\hat{L}_{\nu,\delta}}{\mu}} \right)^{\frac{1+\nu}{2\nu}} := \alpha_{\delta,\nu}, \tag{2.40}$$

then (2.39) clearly implies (2.34). Observe that the above upper bound on α_k is fixed and less than 1. Hence, this observation and (2.33) imply that the line search procedure defined in (2.34) is terminated in finite steps. Moreover, assuming that (2.34) holds, part b) follows similarly to the proof of Theorem 1 and hence we skip the details. □

Observe that the line search procedure defined in Step 2 of Algorithm 2 is performed up to a constant number $\lceil \log \alpha_{\delta,\nu} / \log \gamma \rceil$ per iteration due to (2.40). Hence, if γ is small, then the line search terminates earlier. However, this may lead to a conservative choice of α_k. Therefore, a moderate choice like $\gamma = 0.5$ would perform well in practice.

We specialize convergence of Algorithm 2 in the next result.

Corollary 2 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 2 and h is strongly convex with parameter $\mu > 0$. Then, for any $N \geq 1$, we have*

$$g_{\hat{N}} \leq \frac{\Psi(x_0) - \Psi^*}{\gamma N \alpha_{\delta,\nu}} + \delta, \tag{2.41}$$

where $\hat{N}$ and $\alpha_{\delta,\nu}$ are defined in (2.21) and (2.40), respectively.

If, in addition, f is convex, then we have

$$\Psi(y_{\hat{N}-1}) - \Psi^* \leq g_{\hat{N}} \leq 2\left[\frac{(1 - \frac{\gamma\alpha_{\delta,\nu}}{2})^N[\Psi(x_0) - \Psi^*]}{\gamma\alpha_{\delta,\nu}} + \delta\right]. \qquad (2.42)$$

Proof Noting that by (2.33) and (2.40), we have $\alpha_k \geq \gamma\alpha_{\delta,\nu}$, the results follow similarly to the proof of Corollary 1 in which $\bar{\alpha}_{\epsilon,\nu}$ is replaced by $\gamma\alpha_{\delta,\nu}$. $\qquad \square$

Observe that by (2.40) and similar to (2.28), we have

$$\alpha_{\delta,\nu}^{-1} \leq 2^{\frac{1-\nu}{\nu}}\left[1 + \left(\frac{\hat{L}_{\nu,\delta}}{\mu}\right)^{\frac{1+\nu}{2\nu}}\right],$$

which together with the choice of $\delta = \epsilon/4$, (2.38), and (2.41) imply that the number of iterations performed by Algorithm 2, to have $g_k \leq \epsilon$ for at least one k is bounded by (2.31) and (2.32) when f is nonconvex or convex, respectively. However, these results are obtain without knowing problem parameters ν and L_ν. Also note that only positivity of μ is to be known and its value is not required in the implementation of Algorithm 2. It should be also mentioned that, while such a prior knowledge of problem parameters is not required in [9] for general first-order methods to obtain the similar bound for weakly smooth nonconvex problems, their line search procedure is computationally more expensive than the one in (2.34) as mentioned before.

Below we present another variant of Algorithm 1 whose implementation is totally independent of all problem parameters and we do not need to know whether h is strongly convex or fix the target accuracy in advance. We should point out that while we do not need any knowledge about the problem parameters, the complexity bounds of this algorithm are worse than the ones we obtained for Algorithm 2.

Algorithm 3 The folded conditional gradient type (FCGT) algorithm

Input: $x_0 \in X$, $\{\alpha_k\}_{k\geq 1} \in (0, 1]$, and $\{\beta_k\}_{k\geq 1} \in (0, 1]$.
0. Set the initial points $y_0 = x_0$ and $k = 1$.
1. Compute $f'(y_{k-1})$ and set x_k to (2.1).
2. Set $y_k = \arg\min_{x \in \{\hat{y}_k, \tilde{y}_k\}} \Psi(x)$, where

$$\begin{aligned}\hat{y}_k &= (1 - \alpha_k)y_{k-1} + \alpha_k x_k, \\ \tilde{y}_k &= (1 - \beta_k)y_{k-1} + \beta_k x_k.\end{aligned} \qquad (2.43)$$

2. Set $k \leftarrow k + 1$ and go to step 1.

Note that the above algorithm only differs from Algorithm 1 in updating the sequence $\{y_k\}_{k\geq 0}$. In particular, two folded sequences $\{\hat{y}_k\}_{k\geq 0}$ and $\{\tilde{y}_k\}_{k\geq 0}$ are generated by taking different convex combination of y_{k-1} and x_k and y_k is set to the one with

smaller objective value. This way of updating y_k, as shown in the next result, allows us to have stepsize policies which are totally independent of the problem parameters.

Theorem 3 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm* 3 *and X is bounded.*

(a) *For any $N \geq 1$, we have*

$$\sum_{k=1}^{N} \beta_k g_k \leq \Psi(x_0) - \Psi^* + \bar{C}_\nu \sum_{k=1}^{N} \beta_k^{1+\nu}, \tag{2.44}$$

where

$$\bar{C}_\nu = \frac{L_\nu D_X^{1+\nu}}{(1+\nu)}. \tag{2.45}$$

If, in addition, f is convex, then for any $N \geq 1$, we have

$$\frac{\Psi(y_N) - \Psi^*}{A_N} + \sum_{k=1}^{N} \frac{\alpha_k g_k}{2 A_k} \leq \Psi(x_0) - \Psi^* + \bar{C}_\nu \sum_{k=1}^{N} \frac{\alpha_k^{1+\nu}}{A_k}, \tag{2.46}$$

where A_k is defined in (2.12).

(b) *If h is strongly convex with parameter $\mu > 0$, then for any $N \geq 1$, we have*

$$\sum_{k=1}^{N} \beta_k (1 - \beta_k) g_k \leq \Psi(x_0) - \Psi^* + \frac{\mu C_\nu^2}{2} \sum_{k=1}^{N} \beta_k^{1+2\nu}, \tag{2.47}$$

where

$$C_\nu = \frac{L_\nu D_X^\nu}{\mu(1+\nu)}. \tag{2.48}$$

If, in addition, f is convex and stepsizes are chosen such that

$$\alpha_k \leq \frac{6}{7} \ \forall \, k \geq 1, \tag{2.49}$$

then for any $N \geq 1$, we have

$$\frac{\Psi(y_N) - \Psi^*}{A_N} + \sum_{k=1}^{N} \frac{\alpha_k g_k}{2 A_k} \leq \Psi(x_0) - \Psi^* + \frac{7 \mu C_\nu^2}{2} \sum_{k=1}^{N} \frac{\alpha_k^{1+2\nu}}{A_k}. \tag{2.50}$$

Proof We first show part a). Noting (2.43), the facts that $\Psi(y_k) \leq \Psi(\hat{y}_k)$, $\Psi(y_k) \leq \Psi(\tilde{y}_k)$, boundedness of X, similar to (2.16), and (2.17), we obtain

$$\Psi(y_k) \leq \Psi(\tilde{y}_k) \leq \Psi(y_{k-1}) - \beta_k g_k + \frac{L_\nu \beta_k^{1+\nu}}{1+\nu} D_X^{1+\nu},$$

$$\Psi(y_k) - \Psi^* \leq \Psi(\hat{y}_k) - \Psi^* \leq \left(1 - \frac{\alpha_k}{2}\right)[\Psi(y_{k-1}) - \Psi(x^*)] - \frac{\alpha_k}{2} g_k$$

$$+ \frac{L_\nu \alpha_k^{1+\nu}}{1+\nu} D_X^{1+\nu}.$$

when f is nonconvex and convex, respectively. Then, (2.44) and (2.46) are followed similarly to the proof of Theorem 1(a) by noting (2.45).

We now show part (b). By the fact that $ab \leq a^2/2 + b^2/2$, we have

$$\frac{L_\nu m_1^{1+\nu}}{1+\nu} \|x_k - y_{k-1}\|^{1+\nu} \leq \frac{(m_2 L_\nu)^2 m_1^{1+2\nu}}{2\mu(1+\nu)^2} \|x_k - y_{k-1}\|^{2\nu}$$

$$+ \frac{\mu m_1}{2m_2^2} \|x_k - y_{k-1}\|^2 \quad (2.51)$$

for some $m_1 > 0$ and $m_2 \in \mathbb{R}$. Combining the above relation with the choice of $m_1 = \beta_k$, $m_2 = 1$, (2.43), and (2.16) in which α_k is replaced with β_k, we have

$$\Psi(y_k) \leq \Psi(\tilde{y}_k) \leq \Psi(y_{k-1}) - \beta_k g_k + \frac{\mu \beta_k^2}{2} \|x_k - y_{k-1}\|^2$$

$$+ \frac{L_\nu^2 \beta_k^{1+2\nu}}{2\mu(1+\nu)^2} \|x_k - y_{k-1}\|^{2\nu}$$

$$\leq \Psi(y_{k-1}) - \beta_k(1 - \beta_k) g_k + \frac{L_\nu^2 \beta_k^{1+2\nu} D_X^{2\nu}}{2\mu(1+\nu)^2}, \quad (2.52)$$

where the second inequality follows from Lemma 1(b) and boundedness of X. Summing up the above inequalities and re-arranging the terms, and in the view of (2.48), we obtain (2.47). Moreover, if f is convex, then by (2.43), (2.51) with the choice of $m_1 = \alpha_k$, $m_2 = \sqrt{7}$, and similar to (2.17), we obtain

$$\Psi(y_k) - \Psi^* \leq \left(1 - \frac{\alpha_k}{2}\right)[\Psi(y_{k-1}) - \Psi(x^*)] - \frac{\alpha_k}{2} g_k$$

$$- \frac{\alpha_k(6 - 7\alpha_k)\mu}{14} \|x_k - y_{k-1}\|^2 + \frac{7L_\nu^2 \alpha_k^{1+2\nu}}{2\mu(1+\nu)^2} \|x_k - y_{k-1}\|^{2\nu}$$

$$(2.53)$$

Dividing both sides of the above inequality by A_k, summing them up, noting boundedness of X, (2.49), and in the view of (2.48), we have (2.50). □

In the next result, we specify the choice of stepsizes β_k and α_k to specialize the rate of convergence of Algorithm 3 when applied to nonconvex or (strongly) convex problems defined in (1.1).

Corollary 3 *Suppose that the sequence* $\{x_k, y_k\}_{k \geq 0}$ *is generated by Algorithm 3 and* X *is bounded.*

(a) *If stepsizes are set to*

$$\beta_k = \frac{1}{2N^q} \quad k = 1, 2, \ldots, N \tag{2.54}$$

for some $q \in (0, 1)$, *then for any* $N \geq 1$, *we have*

$$g_{\hat{N}} \leq \frac{2[\Psi(x_0) - \Psi^*]}{N^{1-q}} + \frac{\bar{\mathcal{C}}_\nu}{(2N^q)^\nu}, \tag{2.55}$$

where $\hat{N}$ *is defined in* (2.21). *If, in addition,* f *is convex and stepsizes are set*

$$\alpha_1 = \frac{6}{7}, \quad \alpha_k = \frac{6}{k+5} \quad \forall \, k \geq 2, \tag{2.56}$$

then for any $N \geq 1$, *we have*

$$\Psi(y_{\hat{N}-1}) - \Psi^* \leq g_{\hat{N}} \leq \frac{160[\Psi(x_0) - \Psi^*]}{(N+3)^3} + \frac{27 \cdot 6^\nu \bar{\mathcal{C}}_\nu}{(N+5)^\nu}. \tag{2.57}$$

(b) *If* h *is strongly convex with parameter* $\mu > 0$ *and* β_k *is set to* (2.54), *then we have*

$$g_{\hat{N}} \leq \frac{4[\Psi(x_0) - \Psi^*]}{N^{1-q}} + \frac{\mu \mathcal{C}_\nu^2}{(2N^q)^{2\nu}}. \tag{2.58}$$

If, in addition, f *is convex and* (2.56) *holds, then we have*

$$\Psi(y_{\hat{N}-1}) - \Psi^* \leq g_{\hat{N}} \leq \frac{160[\Psi(x_0) - \Psi^*]}{(N+3)^3} + \frac{14 \cdot 6^{2\nu} \mu \mathcal{C}_\nu^2}{(N+5)^{2\nu}}. \tag{2.59}$$

Proof First, note that by (2.54), we have

$$\sum_{k=1}^{N} \beta_k = \frac{N^{1-q}}{2}, \quad \sum_{k=1}^{N} \beta_k(1 - \beta_k) = \sum_{k=1}^{N} \frac{1}{2N^q}\left(1 - \frac{1}{2N^q}\right) \geq \sum_{k=1}^{N} \frac{1}{4N^q} = \frac{N^{1-q}}{4},$$

$$\sum_{k=1}^{N} \beta_k^{1+\nu} = \sum_{k=1}^{N} \frac{N^{1-q-q\nu}}{2^{1+\nu}}, \quad \sum_{k=1}^{N} \beta_k^{1+2\nu} = \frac{N^{1-q-2q\nu}}{2^{1+2\nu}},$$

which together with (2.44) and (2.47) imply (2.55) and (2.58), respectively.

Second, noting (2.56) and (2.12), we obtain

$$A_k = \frac{4}{7} \prod_{i=2}^{k} \frac{i+2}{i+5} = \frac{480}{7(k+3)(k+4)(k+5)},$$

$$\sum_{k=1}^{N} \frac{\alpha_k}{2A_k} = \frac{1}{A_N} - 1 \geq \frac{(N+3)(N+4)(N+5)}{160},$$

$$\sum_{k=1}^{N} \frac{\alpha_k^{1+2\nu}}{A_k} \leq \sum_{k=1}^{N} \frac{7 \cdot 6^{2\nu}(k+3)(k+4)}{80(k+5)^{2\nu}} \leq \frac{6^{2\nu}}{11} \sum_{k=1}^{N} (k+4)^{2-2\nu}$$

$$\leq \frac{6^{2\nu}}{11} \int_{1}^{N+1} (x+4)^{2-2\nu} dx$$

$$= \frac{6^{2\nu}[(N+5)^{3-2\nu} - 5^{3-2\nu}]}{11(3-2\nu)} \leq 6^{2\nu-1}(N+3)(N+4)(N+5)^{1-2\nu},$$

$$\sum_{k=1}^{N} \frac{\alpha_k^{1+\nu}}{A_k} \leq \sum_{k=1}^{N} \frac{7 \cdot 6^{\nu}(k+3)(k+4)}{80(k+5)^{\nu}} \leq 6^{\nu-1}(N+3)(N+4)(N+5)^{1-\nu},$$

$$(2.60)$$

which together with (2.46) and (2.50) imply (2.57) and (2.59), respectively. □

We now add a few remarks about the above results. First, note that rate of convergence of Algorithm 3 when f is nonconvex depends on the choice of q. In particular, (2.55) implies that the total number of iterations required by this method to have $g_k \leq \epsilon$ for at least one k, is bounded by

$$\mathcal{O}(1) \left\{ \left(\frac{\Psi(x_0) - \Psi^*}{\epsilon} \right)^{\frac{1}{1-q}} + \left(\frac{L_\nu D_X^{1+\nu}}{\epsilon} \right)^{\frac{1}{q\nu}} \right\}. \tag{2.61}$$

Observe that the best choice is $q = \frac{1}{1+\nu}$ which reduces the above bound to the order of the one in (2.29). However, Algorithm 3 is assumed to be implemented without knowing any problem parameters and hence (2.61) would be worse than (2.29). It should be also mentioned that the above observation suggests to confine the range of q and have $q \in [1/2, 1)$ to get better a complexity bound. Moreover, when f is convex, then (2.57) implies that the iteration complexity of Algorithm 3 to find and ϵ-optimal solution of problem (1.1) is bounded by

$$\mathcal{O}(1) \left\{ \left(\frac{\Psi(x_0) - \Psi^*}{\epsilon} \right)^{\frac{1}{3}} + \left(\frac{L_\nu D_X^{1+\nu}}{\epsilon} \right)^{\frac{1}{\nu}} \right\}, \tag{2.62}$$

which is in the same order of (2.30). However, the above bound is obtained without knowing any problem parameters. It should be mentioned that this bound (disregarding

the first term) is also obtained by Nesterov in [29] for Algorithm 1 when f is convex. However, our choice of stepsizes and analysis in the MCGT method as a variant of Algorithm 1 is slightly different.

Second, when f is possibly nonconvex and h is strongly convex, (2.58) implies that the bound in (2.61) is improved to

$$\mathcal{O}(1) \left\{ \left(\frac{\Psi(x_0) - \Psi^*}{\epsilon} \right)^{\frac{1}{1-q}} + \left(\frac{L_\nu D_X^\nu}{\mu\sqrt{\epsilon}} \right)^{\frac{1}{q\nu}} \right\}, \qquad (2.63)$$

which is worse than the one in (2.31). Similar to the aforementioned observation, (2.63) also suggests to confine the range of q to $q \in [1/3, 1)$. Furthermore, when f is convex, then (2.50) implies that that the bound in (2.61) is improved to

$$\mathcal{O}(1) \left\{ \left(\frac{\Psi(x_0) - \Psi^*}{\epsilon} \right)^{\frac{1}{3}} + \left(\frac{L_\nu D_X^\nu}{\mu\sqrt{\epsilon}} \right)^{\frac{1}{\nu}} \right\}. \qquad (2.64)$$

This bound, when the second term is the dominating one, is also obtained in [29]. Although, the above bound is still worse than the one in (2.32), it is obtained without knowing any problem parameter or doing any line search procedure. It is worth noting that in this case, we need boundedness of the feasible set even when h is strongly convex.

On the other hand, if we know the problem parameters, it is possible to modify Algorithm 3 and obtain better results when $\mu > 0$. In particular, we can properly shrink the feasible set and restart Algorithm 3 to achieve the complexity bound in (2.32) with the logarithmic factor. However, it requires knowing problem parameters and solving subproblems of (2.1) over the intersection of the feasible set and a ball around the current search point. Hence, this algorithm has limitation in practice and we skip its presentation for sake of simplicity.

3 Stochastic conditional gradient type method for composite optimization

In this section, we would like to generalize the algorithms presented in Sect. 2 for solving problem (1.1) under stochastic setting. In particular, we assume that only noisy information of f is available through the stochastic oracle ($\mathcal{SO}$) satisfying Assumption 1. Moreover, we consider $\| \cdot \|$ as the norm associated with the inner product in $\mathbb{R}^n$ throughout this section.

Below we present a stochastic variant of Algorithm 1 for solving (1.1).

Note that the above algorithm differs from Algorithm 1 in two aspects. First, replacing the true gradient of f by its stochastic approximation given in (3.1). While this average of stochastic gradients is an unbiased estimator for the true gradient of f, it has reduced variance with respect to the individual ones. More specially, denoting $\Delta_{i,k} = G(y_{k-1}, \xi_{i,k}) - f'(y_{k-1})$ and under Assumption 1, we have

Algorithm 4 The randomized stochastic conditional gradient type (RSCGT) algorithm

Input: $x_0 \in X$, $\{\alpha_k\}_{k\geq 1} \in (0, 1]$, an iteration limit $N \geq 1$, probability mass function (PMF) $P_R(\cdot)$ supported on $\{1, 2, \ldots, N\}$, and sequence of positive integer numbers $\{b_k\}_{k\geq 1}$.

0. Set $y_0 = x_0$, $k = 1$, and generate random variable R according to $P_R(\cdot)$.

For $k = 1, 2, \ldots, R$:
    1. Call the stochastic oracle b_k times to compute

$$\bar{G}(y_{k-1}, \xi_k) = \frac{\sum_{i=1}^{b_k} G(y_{k-1}, \xi_{i,k})}{b_k}. \tag{3.1}$$

    2. Set

$$x_k = w(\bar{G}(y_{k-1}, \xi_k)) = \arg\min_{u \in X} \left\{ \langle \bar{G}(y_{k-1}, \xi_k), u \rangle + h(u) \right\}, \tag{3.2}$$

    and y_k to (2.2).
End

$$\mathbb{E}[\bar{G}(y_{k-1}, \xi_k)] = \frac{\sum_{i=1}^{b_k} \mathbb{E}[G(y_{k-1}, \xi_{i,k})]}{b_k} = f'(y_{k-1}),$$

$$\mathbb{E}\left[\|\bar{G}(y_{k-1}, \xi_k) - f'(y_{k-1})\|_*^2 \right] = \frac{1}{b_k^2} \mathbb{E}\left[\left\| \sum_{i=1}^{b_k} \Delta_{i,k} \right\|_*^2 \right] \leq \frac{\sigma^2}{b_k}, \tag{3.3}$$

where the last inequality follows from the fact that $\mathbb{E}\left[\langle \sum_{i=1}^{b_k-1} \Delta_{i,k}, \Delta_{b_k,k} \rangle | \sum_{i=1}^{b_k-1} \Delta_{i,k} \right] = 0$.

Second, Algorithm 4 is terminated according to the random variable R since $\hat{N}$ in (2.21) is not computable under stochastic setting. Hence, this algorithm consists of two levels of randomness: ξ coming from the stochastic oracle and R imposed for terminating the algorithm. This randomization scheme is the current practice for establishing rate of convergence of stochastic algorithms for nonconvex problems. As mentioned in Sect. 1, it was first proposed by Ghadimi and Lan in [12] for smooth unconstrained stochastic optimization and later used for constrained and composite optimization (see e.g., [3,10,11]).

Noting subproblem (3.2), we can define stochastic generalized Frank-Wolf gap ($\tilde{g}_k$) which will not be useful in our analysis. Hence, we still use (2.6) with $w(f'(y_{k-1}))$ to establish convergence of Algorithm 4. Note however, that this quantity is not available under stochastic setting and it is only used to provide convergence analysis. Moreover, we cannot generalize the analysis used in Theorem 1 for the weakly smooth stochastic problems due to the dependence of the stepsizes on $\tilde{g}_k$. Therefore, we use different stepsize policies to establish convergence analysis of Algorithm 4. It is also worth noting that, when $\nu = 1$ and h is strongly convex, then stepsizes similar to those used in Theorem 1 only depend on problem parameters and give us a better rate of convergence. Hence, in addition to the general results, we also present specific rates of convergence when $\nu = 1$ and $\mu > 0$.

In the next result, we provide main convergence analysis of Algorithm 4.

Theorem 4 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 4 and X is bounded.*

(a) *Assume that $D_X \geq 1$ (for simplicity) and let R be a discrete random variable whose PMF is given by*

$$P_R(R = k) = \frac{\alpha_k}{\sum_{k=1}^{N} \alpha_k} \quad k = 1, \ldots, N \tag{3.4}$$

for some given iteration limit $N \geq 1$. Then, we have

$$\mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{\Psi(x_0) - \Psi^* + \sum_{k=1}^{N} \left[\frac{2L_\nu D_X^2 \alpha_k^{1+\nu}}{(1+\nu)} + \frac{(1+\nu)\sigma^2 \alpha_k^{1-\nu}}{4L_\nu b_k} \right]}{\sum_{k=1}^{N} \alpha_k}. \tag{3.5}$$

If f is convex and $P_R(\cdot)$ is given by

$$P_R(R = k) = \frac{\alpha_k A_k^{-1}}{\sum_{k=1}^{N} \alpha_k A_k^{-1}} \quad k = 1, \ldots, N, \tag{3.6}$$

then we have

$$\mathbb{E}_{\xi_{[N]}, R}\left[\Psi(y_{R-1}) - \Psi^*\right]$$
$$\leq \mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{2\left\{ \Psi(x_0) - \Psi(x^*) + \sum_{k=1}^{N} \left[\frac{2L_\nu D_X^2 \alpha_k^{1+\nu}}{(1+\nu)A_k} + \frac{(1+\nu)\sigma^2 \alpha_k^{1-\nu}}{4L_\nu b_k A_k} \right] \right\}}{\sum_{k=1}^{N} \alpha_k A_k^{-1}}, \tag{3.7}$$

where A_k is defined in (2.12).

(b) *Assume that h is strongly convex with parameter $\mu > 0$, and stepsizes are chosen such that*

$$\alpha_k \leq \frac{6}{7} \quad \forall \, k \geq 1. \tag{3.8}$$

If $P_R(\cdot)$ is given by (3.4), then we have

$$\mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{\Psi(x_0) - \Psi^* + \sum_{k=1}^{N} \left[\frac{7\mu C_\nu^2 \alpha_k^{1+2\nu}}{2} + \frac{\sigma^2 \alpha_k}{2\mu b_k} \right]}{\sum_{k=1}^{N} \alpha_k}, \tag{3.9}$$

where C_ν is defined in (2.48).

If, in addition, f is convex and $P_R(\cdot)$ is given by (3.6), then we have

$$
\mathbb{E}_{\xi_{[N]}, R}\left[\Psi(y_{R-1}) - \Psi(x^*)\right]
$$

$$
\leq \mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{2\left\{\Psi(x_0) - \Psi^* + \sum_{k=1}^{N}\left[\frac{7\mu C_v^2 \alpha_k^{1+2v}}{2A_k} + \frac{\sigma^2 \alpha_k}{2\mu b_k A_k}\right]\right\}}{\sum_{k=1}^{N} \alpha_k A_k^{-1}}.
$$

(3.10)

Proof We first show part a). Noting optimality condition of (3.2), (strong) convexity of h, for any $u \in X$ and some $p_h \in \partial h(x_k)$, we have

$$
\langle f'(y_{k-1}) + \Delta_k + p_h, u - x_k\rangle \geq 0,
$$

$$
h(u) \geq h(x_k) + \langle p_h, u - x_k\rangle + \frac{\mu}{2}\|u - x_k\|^2,
$$

where $\Delta_k = \bar{G}(y_{k-1}, \xi_k) - f'(y_{k-1})$. Summing up the above inequalities, we obtain

$$
\langle f'(y_{k-1}) + \Delta_k, x_k - u\rangle + h(x_k) + \frac{\mu}{2}\|x_k - u\|^2 \leq h(u) \quad \forall u \in X. \quad (3.11)
$$

Letting $u = w(f'(y_{k-1}))$ (the solution of the subproblem when true gradient is used) in the above inequality, multiplying it by α_k, adding it up with (2.14), and (2.15), we obtain

$$
\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k \left[\langle f'(y_{k-1}), y_{k-1} - w(f'(y_{k-1}))\rangle + h(y_{k-1})\right.
$$

$$
\left. - h(w(f'(y_{k-1})))\right] + \frac{L_v \alpha_k^{1+v}}{1+v}\|x_k - y_{k-1}\|^{1+v}
$$

$$
+ \alpha_k \langle \Delta_k, w(f'(y_{k-1})) - x_k\rangle
$$

$$
- \frac{\mu \alpha_k}{2}\left[\|w(f'(y_{k-1})) - x_k\|^2 + (1 - \alpha_k)\|x_k - y_{k-1}\|^2\right]
$$

$$
= \Psi(y_{k-1}) - \alpha_k g_k + \frac{L_v \alpha_k^{1+v}}{1+v}\|x_k - y_{k-1}\|^{1+v}
$$

$$
+ \alpha_k \langle \Delta_k, w(f'(y_{k-1})) - x_k\rangle
$$

$$
- \frac{\mu \alpha_k}{2}\left[\|w(f'(y_{k-1})) - x_k\|^2 + (1 - \alpha_k)\|x_k - y_{k-1}\|^2\right] \quad (3.12)
$$

$$
\leq \Psi(y_{k-1}) - \alpha_k g_k + \frac{L_v \alpha_k^{1+v}}{1+v}\left[\|x_k - y_{k-1}\|^{1+v} + \|w(f'(y_{k-1})) - x_k\|^2\right]
$$

$$
+ \frac{(1+v)\alpha_k^{1-v}\|\Delta_k\|^2}{4L_v}
$$

$$
\leq \Psi(y_{k-1}) - \alpha_k g_k + \frac{2L_v \alpha_k^{1+v} D_X^2}{1+v} + \frac{(1+v)\alpha_k^{1-v}\|\Delta_k\|^2}{4L_v}, \quad (3.13)
$$

where the second inequality follows from the fact that

$$\alpha_k \langle \Delta_k, w(f'(y_{k-1})) - x_k \rangle \leq \alpha_k \|\Delta_k\|_* \|x_k - w(f'(y_{k-1}))\|$$
$$\leq \frac{m_3 \alpha_k^{2m_4}}{2} \|x_k - w(f'(y_{k-1}))\|^2 + \frac{\alpha_k^{2(1-m_4)}}{2m_3} \|\Delta_k\|_*^2,$$
(3.14)

with $m_3 = \frac{2L_\nu}{1+\nu}$, $m_4 = \frac{1+\nu}{2}$, and the last inequality follows from the assumption that $D_X \geq 1$. Taking expectation from both sides of the above inequality, re-arranging the terms, summing them up, diving both sides by $\sum_{k=1}^N \alpha_k$, noting (3.3), and in the view of

$$\mathbb{E}_{\xi_{[N]}, R}[g_R] = \frac{\sum_{k=1}^N \mathbb{E}_{\xi_{[N]}} \alpha_k g_k}{\sum_{k=1}^N \alpha_k},$$

we obtain (3.5). If f is convex, then (3.7) follows similarly to the proof of (2.46) by noting (2.8), (3.13), and (3.6).

We now show part (b). Noting (3.12), (3.14) ($m_3 = \mu$, $m_4 = 1/2$), and (2.51) ($m_1 = \alpha_k$, $m_2 = \sqrt{7}$), we have

$$\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k g_k - \frac{\alpha_k(6 - 7\alpha_k)\mu}{14} \|x_k - y_{k-1}\|^2$$
$$+ \frac{L_\nu^2 \alpha_k^{1+2\nu}}{\mu(1+\nu)^2} \|x_k - y_{k-1}\|^{2\nu} + \frac{\alpha_k \|\Delta_k\|_*^2}{2\mu}$$
$$\leq \Psi(y_{k-1}) - \alpha_k g_k + \frac{7L_\nu^2 \alpha_k^{1+2\nu} D_X^{2\nu}}{2\mu(1+\nu)^2} + \frac{\alpha_k \|\Delta_k\|_*^2}{2\mu},$$
(3.15)

where the second inequality follows from (3.8) and boundedness of X. Rest of the proof is similar to part (a) and hence we skip the details. □

In the next result, we specialize rate of convergence of Algorithm 4 by specifying the choice of stepsizes.

Corollary 4 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 4 and X is bounded.*

(a) *If stepsizes, batch sizes, and the iteration limit are set to*

$$\alpha_k = \frac{1}{N_0(\epsilon)^{\frac{1}{1+\nu}}}, \quad b_k = \bar{b}_\epsilon = \left\lceil \frac{(1+\nu)\sigma^2 D_{f,X}}{L_\nu \epsilon^2} \right\rceil \quad \forall k \geq 1,$$
$$N_0(\epsilon) = \left\lceil \left(\frac{2D_{f,X}}{\epsilon} \right)^{\frac{1+\nu}{\nu}} \right\rceil$$
(3.16)

for some $D_{f,X} \geq \Psi(x_0) - \Psi^* + (2L_\nu D_X^2)/(1+\nu)$ and a given target accuracy $\epsilon > 0$. Then, the total required number of calls to the stochastic oracle to have

$\mathbb{E}[g_R] \leq \epsilon$, where R is uniformly distributed over $\{1, 2, \ldots, N_0(\epsilon)\}$, is bounded by

$$\bar{b}_\epsilon N_0(\epsilon) = \frac{\mathcal{O}(1)\sigma^2 D_{f,X}^{\frac{1+2v}{v}}}{L_v \epsilon^{\frac{1+3v}{v}}}. \tag{3.17}$$

If f is convex, the stepsizes are set to (2.56), $P_R(\cdot)$ is given by (3.6), batch sizes, and the iteration limit are set to

$$b_k = \left\lceil \left(\frac{(1+v)\sigma(k+3)^v}{2^{2v+1.5}L_v D_X}\right)^2 \right\rceil \; \forall\, k \geq 1, \;\; N_1(\epsilon) = \left\lceil 4\left(\frac{216 L_v D_X^2}{(1+v)\epsilon}\right)^{\frac{1}{v}} \right\rceil, \tag{3.18}$$

and assuming $\Psi(x_0) - \Psi^* \leq \mathcal{O}(N_1(\epsilon)^{3-v}L_v D_X^2)$ (for simplicity), then the complexity bound in (3.17) is improved to

$$b_\epsilon N_1(\epsilon) = \mathcal{O}(1)\sigma^2 \left(\frac{L_v D_X^{2+2v}}{\epsilon^{1+2v}}\right)^{\frac{1}{v}}, \tag{3.19}$$

which also guarantees $\mathbb{E}\left[\Psi(y_{R-1}) - \Psi(x^*)\right] \leq \epsilon$.

(b) If $\mu > 0$, stepsizes, batch sizes, and the iteration limit are set to

$$\alpha_k = \frac{6}{7N_2(\epsilon)^{\frac{1}{1+2v}}} \;\; b_k = b_\epsilon = \left\lceil \frac{2\sigma^2}{\mu\epsilon} \right\rceil \; \forall\, k \geq 1,$$

$$N_2(\epsilon) = \left\lceil \left(\frac{14}{3\epsilon}\left[\frac{\Psi_0 - \Psi^*}{3} + \mu\left(\frac{6^v \mathcal{C}_v}{7^v}\right)^2\right]\right)^{\frac{1+2v}{2v}} \right\rceil, \tag{3.20}$$

then the complexity bound in (3.17) is improved to

$$b_\epsilon N_2(\epsilon) = \frac{\mathcal{O}(1)\sigma^2}{\mu\epsilon^{\frac{1+4v}{2v}}}\left[\Psi(x_0) - \Psi^* + \frac{L_v^2 D_X^{2v}}{\mu}\right]^{\frac{1+2v}{2v}}. \tag{3.21}$$

If, f is convex, stepsizes are set to (2.56), the iteration limit is set to

$$N_3(\epsilon) = \left\lceil \left(\frac{56 \cdot 36^v \mu \mathcal{C}_v^2}{\epsilon}\right)^{\frac{1}{2v}} \right\rceil, \tag{3.22}$$

and assuming $\Psi(x_0) - \Psi^* \leq \mathcal{O}(\mu \mathcal{C}_\nu^2 N_3(\epsilon)^{3-2\nu})$ *(for simplicity), then the complexity bound in* (3.19) *is improved to*

$$b_\epsilon N_3(\epsilon) = \frac{\mathcal{O}(1)\sigma^2 L_\nu^{\frac{1}{\nu}} D_X}{(\mu\epsilon)^{\frac{1+2\nu}{2\nu}}}. \tag{3.23}$$

Moreover, this bound guarantees $\mathbb{E}\left[\Psi(y_{R-1}) - \Psi(x^*)\right] \leq \epsilon$.

Proof By (3.4) and choice of α_k in (3.16) we have R is uniformly distributed and together with (3.5), we obtain

$$\mathbb{E}_{\xi_{[N]}, R}\left[g_R\right] \leq \frac{\Psi(x_0) - \Psi^* + 2L_\nu D_X^2/(1+\nu)}{N_0(\epsilon)^{\frac{\nu}{1+\nu}}} + \frac{(1+\nu)\sigma^2 N_0(\epsilon)^{\frac{\nu}{1+\nu}}}{4L_\nu \bar{b}_\epsilon}.$$

Replacing choices of batchsizes and the iteration limit given by (3.16), in the above relation, we clearly have $\mathbb{E}[g_R] \leq \epsilon$. Hence, (3.17) follows from the fact that the total number of calls to the $\mathcal{SO}$ is bounded by $\sum_{k=1}^{N} b_k = \sum_{k=1}^{N_0(\epsilon)} \bar{b}_\epsilon = \bar{b}_\epsilon N_0(\epsilon)$. Moreover, If f is convex, then combining (2.56), (3.6), (3.7), (3.18), and similarly to the proof of Corollary 3.a, we obtain (3.19) which also guarantees $\mathbb{E}\left[\Psi(y_{R-1}) - \Psi(x^*)\right] \leq \epsilon$. Part b) follows similarly and hence we skip the details.   □

We now add a few remarks about the above results. First, note that the framework of Algorithm 3 cannot be directly generalized to the stochastic setting. More specifically, since we do not know the exact value of the objective function Ψ, taking that minimum in Step 2 of this algorithm requires much computational effort under stochastic setting. Hence, we use different stepsize policies for both nonconvex and (strongly) convex cases in the above result which lead to different complexity bounds.

Second, the complexity bound in (3.17) seems to be the first one for solving stochastic nonconvex weakly smooth optimization problems obtained by the class of algorithms including subproblems in the form (1.6). When $\nu = 1$, this bound is reduced to $\mathcal{O}(\epsilon^{-4})$ which is in the same order of the one achieved by a stochastic CG method in [34] for smooth nonconvex problems ($h \equiv 0$). Moreover, when h is strongly convex, (3.17) is improved to (3.21) which is also better than the complexity of $\mathcal{O}(\epsilon^{-\frac{(1+\nu)^2}{\nu}})$ obtained for a gradient type method including subproblems in the form (1.5) in which the Bregman distance function is replaced by $\|\cdot\|^{1+\nu}$ and under the boundedness assumption of higher momentums of the stochastic error in gradient estimations ([20]).

Third, when f is convex, (3.19) seems to be the first complexity bound of CGT methods for solving stochastic convex weakly smooth optimization problems. When $\nu = 1$, this bound is $\mathcal{O}(\epsilon^{-3})$ which is worse than $\mathcal{O}(\epsilon^{-2})$ obtained in [24] for a conditional gradient method skipping computing stochastic gradients for some iterations. When Ψ is strongly convex, (3.19) is improved to (3.23) and it reduces to $\mathcal{O}(\epsilon^{-1.5})$ for solving stochastic strongly convex smooth optimization problems. Note however, that this bound is still worse than the optimal complexity $\mathcal{O}(\epsilon^{-1})$ for stochastic first-order methods. In the rest of this section, we provide a different choice of stepsizes

for Algorithm 4 to achieve the aforementioned optimal complexity bound when $\nu = 1$ and $\mu > 0$.

Theorem 5 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 4, h is strongly convex with parameter $\mu > 0$, f has Lipschitz continuous gradient i.e., $\nu = 1$ in (1.3) and (1.4), the stepsizes satisfy*

$$\alpha_k \leq \frac{\mu}{L_1 + \mu} \quad \forall\, k \geq 1, \tag{3.24}$$

and $P_R(\cdot)$ is given by (3.4). Then, we have

$$\mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{\Psi(x_0) - \Psi^* + \frac{\sigma^2}{2\mu} \sum_{k=1}^{N} \frac{\alpha_k}{b_k}}{\sum_{k=1}^{N} \alpha_k} \tag{3.25}$$

If, in addition, f is convex and $P_R(\cdot)$ is given by (3.6), then we have

$$\mathbb{E}_{\xi_{[N]}, R}\left[\Psi(y_{R-1}) - \Psi(x^*)\right] \leq \mathbb{E}_{\xi_{[N]}, R}[g_R] \leq \frac{2[\Psi(x_0) - \Psi^*] + \frac{\sigma^2}{\mu} \sum_{k=1}^{N} \frac{\alpha_k}{b_k A_k}}{\sum_{k=1}^{N} \alpha_k A_k^{-1}}, \tag{3.26}$$

where A_k is defined in (2.12)

Proof By (3.12) and (3.14) (with $m_3 = \mu$, $m_4 = 1/2$), we have

$$\Psi(y_k) \leq \Psi(y_{k-1}) - \alpha_k g_k - \frac{\alpha_k[\mu - (\mu + L_1)\alpha_k]}{2}\|x_k - y_{k-1}\|^2 + \frac{\alpha_k\|\Delta_k\|_*^2}{2\mu}$$

$$\leq \Psi(y_{k-1}) - \alpha_k g_k + \frac{\alpha_k\|\Delta_k\|_*^2}{2\mu},$$

where the second inequality follows from (3.24). Rest of the proof is similar to that of Theorem 4.(a) and hence we skip the details. □

In the next result, we specify the appropriate choice of stepsizes for Algorithm 4 when f is smooth.

Corollary 5 *Suppose that the sequence $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 4, h is strongly convex with $\mu > 0$, and f has Lipschitz continuous gradient. If batch sizes are set to (3.20), stepsizes, and the iteration limit are set to*

$$\alpha_k = \bar{\alpha} = \frac{\mu}{L_1 + \mu}, \quad \forall\, k \geq 1, \quad N_4(\epsilon) = \left\lceil \frac{4(\mu + L_1)[\Psi(x_0) - \Psi^*]}{3\mu\epsilon} \right\rceil \tag{3.27}$$

for a given $\epsilon > 0$, then the total required number of calls to the stochastic oracle to have $\mathbb{E}[g_R] \leq \epsilon$, where R is uniformly distributed over $\{1, 2, \ldots, N_4(\epsilon)\}$, is bounded by

$$b_\epsilon N_4(\epsilon) = \mathcal{O}(1) \left(\frac{(\mu + L_1)\sigma^2 [\Psi(x_0) - \Psi^*]}{\mu^2 \epsilon^2} \right). \tag{3.28}$$

If, in addition, f is convex, $P_R(\cdot)$ is given by (3.6), and the iteration limit is set to

$$N_5(\epsilon) = \left\lceil \frac{2(\mu + L_1)}{\mu} \log \left(\frac{4(\mu + L_1)[\Psi(x_0) - \Psi^*]}{\mu\epsilon} \right) \right\rceil, \tag{3.29}$$

then the complexity bound in (3.28) is improved to

$$b_\epsilon N_5(\epsilon) = \mathcal{O}(1) \left(\frac{(\mu + L_1)\sigma^2}{\mu^2 \epsilon} \log \left[\frac{(\mu + L_1)[\Psi(x_0) - \Psi(x^*)]}{\mu\epsilon} \right] \right). \tag{3.30}$$

Moreover, the above bound guarantees $\mathbb{E} \left[\Psi(y_{R-1}) - \Psi(x^*) \right] \le \epsilon$.

Proof The results are followed similarly to that of Corollary 4 by noting (3.27) and (2.27). □

Note that, the bound in (3.28) is better than the one in (3.21) when $\nu = 1$ and after disregarding μ, it is the best known complexity of first-order methods when applied to smooth nonconvex stochastic programming (see e.g., [10–12]). Moreover, when f is convex, the complexity bound in (3.30) is also better than the one in (3.23) and it is indeed optimal up to a logarithmic constant for solving smooth strongly convex stochastic optimization [27].

It should be also pointed out that while all of the complexity bounds in this section are presented for expectation results, we can simply use Markov's inequality to provide similar bounds to have $\text{Prob}\{g_k \ge \epsilon\} \le \Lambda$ in a single run of these algorithms. In particular, these large-deviation bounds are obtained with replacing ϵ by $\epsilon\Lambda$ in their associated expectation complexity bounds. However, we can improve dependence of these bounds on Λ by designing a two-phase variant of the RSCGT method which consists of several runs of the RSCGT method in the first phase and choosing the best solution among these runs as the output in the second phase (see e.g., [9,12] for similar two-phase variants of randomized stochastic (projected) gradient methods for smooth problems in more details).

4 Concluding remarks

In this paper, we present conditional gradient type methods for solving a class of composite nonlinear optimization problems defined in (1.1). We present a unified analysis for both nonconvex and (strongly) convex problems in the sense that our proposed method achieves the best-known complexity for nonconvex problems, and its rate of convergence is (nearly) optimal if f is convex. We also present a few variants of this method which do not require the knowledge of problem parameters. Since there is no stepsize involved in the subproblems of these methods, their computational efforts are much cheaper per iteration than those of the existing parameter free proximal type methods. Moreover, we consider problem (1.1) under stochastic setting where only

noisy information of f is available and provide complexity of our generalized methods to this setting. To the best of our knowledge, this is the first time that such complexity results of CGT methods are presented for stochastic weakly smooth optimization.

References

1. Cartis, C., Gould, N.I.M., Toint, P.L.: On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization. SIAM J. Optim. **20**(6), 2833–2852 (2010)
2. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. J. Mach. Learn. Res. **9**, 203–233 (2008)
3. Dang, C.D., Lan, G.: Stochastic block mirror descent methods for nonsmooth and stochastic optimization. SIAM J. Optim. **25**(2), 856–881 (2015)
4. Devolder, O., Glineur, F., Nesterov, Y.E.: First-order methods with inexact oracle: the strongly convex case. December 2013, CORE Discussion Paper 2013/16
5. Dunn, J.C.: Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. SIAM J. Control Optim. **17**(2), 674–701 (1979)
6. Dunn, J.C.: Convergence rates for conditional gradient sequences generated by implicit step length rules. SIAM J. Control Optim. **18**(5), 473–487 (1980)
7. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Nav. Res. Logist. Q. **3**, 95–110 (1956)
8. Garber, D., Hazan, E.: A Linearly Convergent Conditional Gradient Algorithm with Applications to Online and Stochastic Optimization. arXiv e-prints (2013)
9. Ghadimi, S., Lan, G., Zhang, H.: Generalized uniformly optimal methods for nonlinear programming, manuscript. Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611, USA (August 2015)
10. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. Math. Program. **155**, 267–305 (2016)
11. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic optimization. Math. Program. **156**, 59–99 (2016)
12. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**(4), 2341–2368 (2013)
13. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**, 2061–2089 (2013)
14. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems (NIPS), p. 17 (2005)
15. Guélat, J., Marcotte, P.: Some comments on wolfe's 'away step'. Math. Progr. **35**(1), 110–119 (1986)
16. Harchaoui, Z., Juditsky, A., Nemirovski, A.S.: Conditional gradient algorithms for machine learning. NIPS OPT Workshop (2012)
17. Ito, M.: New results on subgradient methods for strongly convex optimization problems with a unified analysis. Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Japan, Tokyo (April 2015)
18. Jaggi, M.: Revisiting frank-wolfe: projection-free sparse convex optimization. In: The 30th International Conference on Machine Learning (2013)
19. Jensen, T., Jørgensen, J.H., Hansen, P., Jensen, S.: Implementation of an optimal first-order method for strongly convex total variation regularization. BIT Numer. Math. **52**, 329–356 (2012)
20. Jiang, B., Zhang, S.: Iteration Bounds for Finding the ϵ-Stationary Points for Structured Nonconvex Optimization. arXiv e-prints (2014)
21. Kakade, S.M., Shalev-Shwartz, S., Tewari, A.: Regularization techniques for learning with matrices. J. Mach. Learn. Res. **13**, 1865–1890 (2012)

22. Lan, G.: The complexity of large-scale convex programming under a linear optimization oracle. Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611, USA, (June 2013). http://www.optimization-online.org
23. Lan, G.: Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization. Math. Progr. **149**(1), 1–45 (2015)
24. Lan, G., Zhou, Y.: Conditional gradient sliding for convex optimization. SIAM J. Optim. **26**, 1379–1409 (2016)
25. Luss, R., Teboulle, M.: Conditional gradient algorithms for rank one matrix approximations with a sparsity constraint. SIAM Rev. **55**, 65–98 (2013)
26. Mason, L., Baxter, J., Bartlett, P., Frean, M.: Boosting algorithms as gradient descent in function space. Proc. NIPS **12**, 512–518 (1999)
27. Nemirovski, A.S., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
28. Nemirovskii, A.S., Nesterov, Y.E.: Optimal methods for smooth convex minimization. Zh. Vichisl. Mat. Fiz. **25**, 356–369 (1985). (In Russian)
29. Nesterov, Y.E.: Complexity bounds for primal-dual methods minimizing the model of objective function. Technical Report, CORE Discussion Papers, Februray (2015)
30. Nesterov, Y.E.: Universal gradient methods for convex optimization problems. Math. Progr. Ser. A (2014). https://doi.org/10.1007/s10107-014-0790-0
31. Nesterov, Y.E.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston (2004)
32. Nesterov, Y.E.: Gradient methods for minimizing composite objective functions. Math. Progr. Ser. B **140**, 125–161 (2013)
33. Pshenichnyi, B.N., Danilin, I.M.: Numerical Methods in Extremal Problems. Mir Publishers, Moscow (1978)
34. Reddi, S.J., Sra, S., Poczos, B., Smola, A.: Stochastic Frank-Wolfe Methods for Nonconvex Optimization. arXiv e-prints (2016)
35. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)