Complexity bounds for primal-dual methods minimizing the model of objective function

Yu. Nesterov *

July 14, 2016

Abstract

We provide Frank-Wolfe (≡ Conditional Gradients) method with a convergence analysis allowing to approach a primal-dual solution of convex optimization problem with composite objective function. Additional properties of complementary part of the objective (strong convexity) significantly accelerate the scheme. We also justify a new variant of this method, which can be seen as a trust-region scheme applying to the linear model of objective function. For this variant, we prove also the rate of convergence for the total variation of linear model of composite objective over the feasible set.

Our analysis works also for quadratic model, allowing to justify the global rate of convergence for a new second-order method as applied to a composite objective function. To the best of our knowledge, this is the first trust-region scheme supported by the worst-case complexity analysis both for the functional gap and for the variation of local quadratic model over the feasible set.

Keywords: convex optimization, complexity bounds, linear optimization oracle, conditional gradient method, trust-region method.

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: Yurii.Nnesterov@uclouvain.be.
The research results presented in this paper have been supported by the grant "Action de recherche concertè ARC 14/19-060" from the "Direction de la recherche scientifique - Communautè française de Belgique". Scientific responsibility rests with the author.

1 Introduction

Motivation. In the last years, we can see an increasing interest to Frank-Wolfe algorithm [3, 2, 11], which sometimes is called *Conditional Gradient Method* (CGM) [5, 7, 9, 8]. At each iteration of this scheme, we need to solve an auxiliary problem of minimizing a linear function over a convex feasible set. In some situations, mainly when the feasible set is a simple polytope, the complexity of this subproblem is much lower than that of the standard projection technique (e.g. [12]). The standard complexity results for this method are related to convex objective function with Lipschitz-continuous gradient. In this situation, CGM converges as $O(\frac{1}{k})$, where k is the number of iterations. Moreover, it appears that this rate of convergence is optimal for methods with linear optimization oracle [10].

For nonsmooth functions, CGM cannot converge (we give a simple example in Section 2). Therefore, it is interesting to study the dependence of the rate of convergence of CGM on the level of smoothness of the objective function. On the other hand, sometimes nonsmoothness of the objective function results from a complementary regularization term. This situation can be treated in the framework of *composite minimization* [14]. However, the performance of CGM for this structure of the objective function was not studied yet. Finally, by its spirit, CGM is a *primal-dual method*. Indeed, it generates the lower bounds on the optimal value of the objective function, which converge to the optimal value [4]. Therefore, it would be natural to extract from this method an approximate solution of the dual problem. These questions served as the main motivations for this paper.

Contents. In Section 2, we introduce our main problem of interest, where the objective function has a composite form. Our main assumption is that the problem of minimizing a linear function augmented by a simple convex complementary term is solvable.[1)] We assume that the smooth part of the objective has Hölder-continuous gradients. For proving efficiency estimate for CGM, we apply the technique of *estimating sequences* (e.g. [12]) in its extended form [18]. As a result, we get a significant freedom in the choice of step-size coefficients. In the next Section 3 we consider a new variant of CGM, which can be seen as a trust-region method with linear model of the objective. For this scheme, the trust region is formed by contracting the feasible set towards the current test point. This method ensures an appropriate rate of convergence for total variation of the linear model, computed at the sequence of test points.

Analytical form of our bounds for the primal-dual gap is similar to the bounds obtained in [4]. However, we estimate the difference of the current value of the objective function and the minimal value of the accumulated linear model. In Section 4, we explain how to extract from our convergence results an upper bound on the duality gap for some feasible primal-dual solution. In our technique we use an explicit max-representation of the smooth part of the objective function.

In Section 5, we show that the additional properties of complementary part of the objective (strong convexity) significantly accelerate the scheme. Finally, in Section 6, we apply our technique for a new second-order trust-region method, where the quadratic approximation of our objective function is minimized on a trust region formed by a con-

[1)] Performance of CGM as applied to objective function regularized by a norm was studied in [6].

tracted feasible set. To the best of our knowledge, this is the first trust-region scheme [1] supported by the worst-case complexity analysis.

Notation. In what follows, we consider optimization problems over finite-dimensional linear space $\mathbb{E}$ with the dual space $\mathbb{E}^*$. The value of linear function $s \in \mathbb{E}^*$ at $x \in \mathbb{E}$ is denoted by $\langle s, x \rangle$. In E, we fix a norm $\|\cdot\|$, which defines the conjugate norm

$$\|s\|_* \;=\; \max_x \{\langle s, x\rangle : \; \|x\| \le 1\}, \quad s \in \mathbb{E}^*.$$

For a linear operator $A : \mathbb{E} \to \mathbb{E}_1^*$, its *conjugate* operator $A^* : \mathbb{E}_1 \to \mathbb{E}^*$ is defined by identity

$$\langle Ax, y\rangle \;=\; \langle A^* y, x\rangle, \quad x \in \mathbb{E}, \; y \in \mathbb{E}_1.$$

We call operator $B : \mathbb{E} \to \mathbb{E}^*$ self-conjugate if $B = B^*$. It is *positive-semidefinite* if $\langle Bx, x\rangle \ge 0$ for all $x \in \mathbb{E}$. For a linear operator $B : \mathbb{E} \to \mathbb{E}^*$, we define its *operator norm* in the usual way:

$$\|B\| \;=\; \max_x \{\|Bx\|_* : \; \|x\| \le 1\}.$$

For a differentiable function $f(x)$ with $\operatorname{dom} f \subseteq \mathbb{E}$, we denote by $\nabla f(x) \in \mathbb{E}^*$ its *gradient*, and by $\nabla^2 f(x) : \mathbb{E} \to \mathbb{E}^*$ its *Hessian*. Note that $\left(\nabla^2 f(x)\right)^* = \nabla^2 f(x)$.

In the sequel, we often need to estimate the partial sums of different series. For that, it is convenient to use the following trivial lemma.

Lemma 1 *Let function $\xi(\tau)$, $\tau \in \mathbb{R}$, be decreasing and convex. Then, for any two integers a and b, such that $[a - \frac{1}{2}, b + 1] \subset \operatorname{dom} \xi$, we have*

$$\int\limits_a^{b+1} \xi(\tau) d\tau \;\le\; \sum_{k=a}^{b} \xi(k) \;\le\; \int\limits_{a-1/2}^{b+1/2} \xi(\tau)\, d\tau. \tag{1.1}$$

For example, for any $t \ge 0$ and $p \ge -t$, we have

$$\sum_{k=t}^{2t+p} \tfrac{1}{k+p+1} \overset{(1.1)}{\ge} \int\limits_t^{2t+p+1} \tfrac{1}{\tau+p+1} d\tau \;=\; \ln(\tau + p + 1)\Big|_t^{2t+p+1} \;=\; \ln \tfrac{2t+2p+2}{t+p+1} \;=\; \ln 2. \tag{1.2}$$

On the other hand, if $t \ge 1$, then

$$\begin{aligned} \sum_{k=t}^{2t+1} \tfrac{1}{(k+2)^2} &\overset{(1.1)}{\le} \int\limits_{t-1/2}^{2t+3/2} \tfrac{1}{(\tau+2)^2} d\tau \;=\; -\tfrac{1}{\tau+2}\Big|_{t-1/2}^{2t+3/2} \;=\; \tfrac{1}{t+3/2} - \tfrac{1}{2t+7/2} \\ &=\; \tfrac{4t+8}{(2t+3)(4t+7)} \;\le\; \tfrac{12}{11(2t+3)}. \end{aligned} \tag{1.3}$$

In this paper, we will use some special dual functions. Let $Q \subset E$ be a bounded closed convex set. For a closed convex function $F(\cdot)$ with $\operatorname{dom} F \supseteq \operatorname{int} Q$, we define its *restricted dual function*, (with respect to a central point $\bar{x} \in Q$), as follows:

$$F^*_{\bar{x},Q}(s) \;=\; \max_{x \in Q} \{\langle s, \bar{x} - x\rangle + F(\bar{x}) - F(x)\}, \quad s \in \mathbb{E}^*. \tag{1.4}$$

Clearly, this function is well defined for all $s \in \mathbb{E}^*$. Moreover, it is convex and nonnegative on $\mathbb{E}^*$.

In the theory of Nonsmooth Convex Optimization, the dual functions are often used for analyzing the behavior of first-order methods. In these applications, function F in (1.4) is assumed to be a multiple of a strongly convex prox-function (see, for example, [13]). Such an assumption provides us with a possibility to control the step sizes in the corresponding optimization schemes. In this paper, we are going to drop the assumption on strong convexity. Hence, we need to develop another mechanism for controlling the sizes of our moves. This is the reason why we introduce in construction (1.4) an additional scaling parameter $\tau \in [0, 1]$. We call function

$$F^*_{\tau,\bar{x},Q}(s) \;=\; \max_{x \in Q} \{\langle s, \bar{x} - y\rangle + F(\bar{x}) - F(y) : \; y = (1-\tau)\bar{x} + \tau x\}, \quad s \in \mathbb{E}^*, \tag{1.5}$$

the *scaled restricted dual* of function F.

Lemma 2 *For any $s \in \mathbb{E}^*$ and $\tau \in [0, 1]$, we have*

$$F^*_{\bar{x},Q}(s) \;\geq\; F^*_{\tau,\bar{x},Q}(s) \;\geq\; \tau\, F^*_{\bar{x},Q}(s). \tag{1.6}$$

Proof:
Since for any $x \in Q$, the point $y = (1-\tau)\bar{x} + \tau x$ belongs to Q, the first inequality is trivial. On the other hand,

$$\begin{aligned} F^*_{\tau,\bar{x},Q}(s) &= \max_{x \in Q} \{\langle s, \tau(\bar{x} - x)\rangle + F(\bar{x}) - F(y) : \; y = (1-\tau)\bar{x} + \tau x \,\} \\ &\geq \max_{x \in Q} \{\langle s, \tau(\bar{x} - x)\rangle + F(\bar{x}) - (1-\tau)F(\bar{x}) - \tau F(x) \,\} \\ &= \tau F^*_{\bar{x},Q}(s). \end{aligned}$$

□

2 Conditional gradient method

In this paper we consider numerical methods for solving the following *composite* minimization problem:

$$\min_x \left\{ \bar{f}(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) \right\}, \tag{2.1}$$

where Ψ is a *simple* closed convex function with bounded domain $Q \subset \mathbb{E}$, and f is a convex function, which is subdifferentiable on Q. Denote by x_* one of the optimal solutions of (2.1), and $D \stackrel{\text{def}}{=} \mathrm{diam}(Q)$. Our assumption on simplicity of function Ψ means that some auxiliary optimization problems related to Ψ are easily solvable. Complexity of these problems will be always discussed for corresponding optimization schemes.

The most important examples of function Ψ are as follows.

- Ψ is an indicator function of a closed convex set Q:

$$\Psi(x) \;=\; \mathrm{Ind}_Q(x) \stackrel{\text{def}}{=} \begin{cases} 0, & x \in Q, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.2}$$

- Ψ is a self-concordant barrier for a closed convex set Q (see [16, 12]).
- Ψ is a nonsmooth convex function with simple structure (e.g. $\Psi(x) = \|x\|_1$).

We assume that function f is represented by a black-box oracle. If it is a *first-order oracle*, we assume its gradients satisfy the following *Hölder condition*:

$$\|\nabla f(x) - \nabla f(y)\|_* \quad \leq \quad G_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{2.3}$$

Constant G_ν is formally defined for any $\nu \in (0, 1]$. For some values of ν it can be $+\infty$. Note that for any x and y in Q we have

$$f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{G_\nu}{1+\nu} \|y - x\|^{1+\nu}. \tag{2.4}$$

If this is a *second-order oracle*, we assume that its Hessians satisfy Hölder condition

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad H_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{2.5}$$

In this case, for any x and y in Q we have

$$f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \tfrac{H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}. \tag{2.6}$$

For the first variant of *Conditional Gradient Methods* (CGM1) our assumption on simplicity of function Ψ means exactly the following.

Assumption 1 *For any $s \in \mathbb{E}^*$, the auxiliary problem*

$$\min_{x \in Q} \{\langle s, x \rangle + \Psi(x)\} \tag{2.7}$$

is easily solvable. Denote by $v_\Psi(s) \in Q$ one of its optimal solutions.

In the case (2.2), this assumption implies that we are able to solve the problem

$$\min_x \{\langle s, x \rangle : x \in Q\}.$$

Note that point $v_\Psi(s)$ is characterized by the following variational principle:

$$\langle s, x - v_\Psi(s) \rangle + \Psi(x) \quad \geq \quad \Psi(v_\Psi(s)), \quad x \in Q. \tag{2.8}$$

In order to solve problem (2.1), we apply the following method.

Conditional Gradient Method, Type I

1. Choose an arbitrary point $x_0 \in Q$.

2. **For** $t \geq 0$ **iterate:** a) Compute $v_t = v_\Psi(\nabla f(x_t))$. (2.9)

b) Choose $\tau_t \in (0, 1]$ and set $x_{t+1} = (1 - \tau_t) x_t + \tau_t v_t$.

It is clear that this method can minimize only functions with continuous gradient.

Example 1 *Let* $\Psi(x) = \mathrm{Ind}_Q(x)$ *with* $Q = \{x \in \mathbb{R}^2 : (x^{(1)})^2 + (x^{(2)})^2 \leq 1\}$. *Define*

$$f(x) \;=\; \max\{x^{(1)}, x^{(2)}\}.$$

Then clearly $x_* = \left(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)^T$. *Let us choose in (2.9)* $x_0 \neq x_*$.

For function f, *we can apply an oracle, which returns at any* $x \in Q$ *a subgradient* $\nabla f(x) \in \{(1,0)^T, (0,1)^T\}$. *Then, for any feasible* x, *the point* $v_\Psi(\nabla f(x))$ *is equal either to* $y_1 = (-1,0)^T$, *or to* $y_2 = (0,-1)^T$. *Therefore, all points of the sequence* $\{x_t\}_{t\geq 0}$, *generated by method (2.9), belong to the triangle* $\mathrm{Conv}\{x_0, y_1, y_2\}$, *which does not contain the optimal point* x_*. □

In order to justify the rate of convergence of method (2.9) for functions with Hölder continuous gradients, we apply the estimating sequences technique [12] in its relaxed form [18]. For that, it is convenient to introduce in (2.9) new control variables. Consider a sequence of nonnegative weights $\{a_t\}_{t\geq 0}$. Define

$$A_t \;=\; \sum_{k=0}^{t} a_k, \quad \tau_t \;=\; \tfrac{a_{t+1}}{A_{t+1}}, \quad t \geq 0. \tag{2.10}$$

From now on, we assume that parameter τ_t in method (2.9) is chosen in accordance to the rule (2.10). Denote

$$\begin{aligned} V_0 \;&=\; \max_x \left\{\langle \nabla f(x_0), x_0 - x\rangle + \Psi(x_0) - \Psi(x)\right\}, \\ B_{\nu,t} \;&=\; a_0 V_0 + \left(\sum_{k=1}^{t} \tfrac{a_k^{1+\nu}}{A_k^{\nu}}\right) G_\nu D^{1+\nu}, \quad t \geq 0. \end{aligned} \tag{2.11}$$

It is clear that

$$\begin{aligned} V_0 \;&\overset{(2.6)}{\leq}\; \max_x \left\{f(x_0) - f(x) + \tfrac{G_\nu}{1+\nu}\|x - x_0\|^{1+\nu} + \Psi(x_0) - \Psi(x)\right\} \\ &\leq\; \bar{f}(x_0) - \bar{f}(x_*) + \tfrac{G_\nu D^{1+\nu}}{1+\nu} \;\overset{\mathrm{def}}{=}\; \Delta(x_0) + \tfrac{G_\nu D^{1+\nu}}{1+\nu}. \end{aligned} \tag{2.12}$$

Theorem 1 *Let sequence* $\{x_t\}_{t\geq 0}$ *be generated by method* (2.9). *Then, for any* $\nu \in (0,1]$, *any step* $t \geq 0$, *and any* $x \in Q$ *we have*

$$A_t(f(x_t) + \Psi(x_t)) \;\leq\; \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] + B_{\nu,t}. \tag{2.13}$$

Proof:

Indeed, in view of definition (2.11), for $t = 0$ inequality (2.13) is satisfied. Assume that it is valid for some $t \geq 0$. Then

$$\sum_{k=0}^{t+1} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] + B_{\nu,t}$$

$$\stackrel{(2.13)}{\geq} A_t(f(x_t) + \Psi(x_t)) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \Psi(x)]$$

$$\geq A_{t+1} f(x_{t+1}) + A_t \Psi(x_t) + \langle \nabla f(x_{t+1}), a_{t+1}(x - x_{t+1}) + A_t(x_t - x_{t+1}) \rangle + a_{t+1}\Psi(x)$$

$$\stackrel{(2.9)_b}{=} A_{t+1} f(x_{t+1}) + A_t \Psi(x_t) + a_{t+1}\left[\Psi(x) + \langle \nabla f(x_{t+1}), x - v_t \rangle\right]$$

$$\stackrel{(2.9)_b}{\geq} A_{t+1}\left(f(x_{t+1}) + \Psi(x_{t+1})\right) + a_{t+1}\left[\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle\right]$$

It remains to note that

$$\begin{aligned} \Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle &\stackrel{(2.8)}{\geq} \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t \rangle \\ &\stackrel{(2.3)}{\geq} -\tau_t^\nu L_\nu D^{1+\nu}. \end{aligned}$$

Thus, for keeping (2.13) valid for the next iteration, it is enough to choose

$$B_{\nu,t+1} \;=\; B_{\nu,t} + \tfrac{a_{t+1}^{1+\nu}}{A_{t+1}^{\nu}} G_\nu D^{1+\nu}.$$

□

Corollary 1 *For any $t \geq 0$ with $A_t > 0$, and any $\nu \in (0, 1]$ we have*

$$\bar{f}(x_t) - \bar{f}(x_*) \;\leq\; \tfrac{1}{A_t} B_{\nu,t}. \tag{2.14}$$

Let us discuss now the possible variants for choosing the weights $\{a_t\}_{t \geq 0}$.

1. *Constant weights.* Let us choose $a_t \equiv 1$, $t \geq 0$. Then $A_t = t + 1$, and for $\nu \in (0, 1)$ we have

$$\begin{aligned} B_{\nu,t} \;&=\; V_0 + \left(\sum_{k=1}^{t} \tfrac{1}{(1+k)^\nu}\right) G_\nu D^{1+\nu} \stackrel{(1.1)}{\leq} V_0 + G_\nu D^{1+\nu} \tfrac{1}{1-\nu}(1+\tau)^{1-\nu}\Big|_{1/2}^{t+1/2} \\ &\stackrel{(2.12)}{\leq} \Delta(x_0) + G_\nu D^{1+\nu}\left[\tfrac{1}{1+\nu} + \left(\tfrac{3}{2}\right)^{1-\nu} \tfrac{1}{1-\nu}\left(\left(1 + \tfrac{2}{3}t\right)^{1-\nu} - 1\right)\right] \end{aligned}$$

   Thus, for $\nu \in (0, 1)$, we have $\tfrac{1}{A_t} B_{\nu,t} \leq O(t^{-\nu})$. For the most important case $\nu = 1$, we have $\lim\limits_{\nu \to 1} \tfrac{1}{1-\nu}\left(\left(1 + \tfrac{2}{3}t\right)^{1-\nu} - 1\right) = \ln(1 + \tfrac{2}{3}t)$. Therefore,

$$\bar{f}(x_t) - \bar{f}(x_*) \;\leq\; \tfrac{1}{t+1}\left(\Delta(x_0) + G_1 D^2\left[\tfrac{1}{2} + \ln(1 + \tfrac{2}{3}t)\right]\right). \tag{2.15}$$

   In this situation, in method (2.9) we take $\tau_t \stackrel{(2.10)}{=} \tfrac{1}{t+1}$.

2. *Linear weights.* Let us choose $a_t \equiv t$, $t \geq 0$. Then $A_t = \frac{t(t+1)}{2}$, and for $\nu \in (0,1)$ with $t \geq 1$ we have

$$B_{\nu,t} = \left(\sum_{k=1}^{t} \frac{2^\nu k^{1+\nu}}{k^\nu (1+k)^\nu}\right) G_\nu D^{1+\nu} \leq \left(\sum_{k=1}^{t} 2^\nu k^{1-\nu}\right) G_\nu D^{1+\nu}$$

$$\overset{(1.1)}{\leq} G_\nu D^{1+\nu} \frac{2^\nu}{2-\nu} \tau^{2-\nu} \Big|_{1/2}^{t+1/2} = \frac{2^\nu}{2-\nu} \left[\left(t+\tfrac{1}{2}\right)^{2-\nu} - \left(\tfrac{1}{2}\right)^{2-\nu}\right] G_\nu D^{1+\nu}.$$

Thus, for $\nu \in (0,1)$, we again have $\frac{1}{A_t} B_{\nu,t} \leq O(t^{-\nu})$. For the case $\nu = 1$, we get the following bound:

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{4}{t+1} G_1 D^2, \quad t \geq 1. \tag{2.16}$$

As we can see, this rate of convergence is better than (2.15). In this case, in method (2.9) we take $\tau_t \overset{(2.10)}{=} \frac{2}{t+2}$, which is a standard recommendation for CGM1 (2.9).

3. *Aggressive weights.* Let us choose, for example, $a_t \equiv t^2$, $t \geq 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$. Note that for $k \geq 0$ we have $\frac{k^{2+\nu}}{(k+1)^\nu (2k+1)^\nu} \leq \frac{k^{2-\nu}}{2^\nu}$. Therefore, for $\nu \in (0,1)$ with $t \geq 1$ we obtain

$$B_{\nu,t} = \left(\sum_{k=1}^{t} \frac{6^\nu k^{2(1+\nu)}}{k^\nu (1+k)^\nu (2k+1)^\nu}\right) G_\nu D^{1+\nu} \leq \left(\sum_{k=1}^{t} 3^\nu k^{2-\nu}\right) G_\nu D^{1+\nu}$$

$$\overset{(1.1)}{\leq} G_\nu D^{1+\nu} \frac{3^\nu}{3-\nu} \tau^{3-\nu} \Big|_{1/2}^{t+1/2} = \frac{3^\nu}{3-\nu} \left[\left(t+\tfrac{1}{2}\right)^{3-\nu} - \left(\tfrac{1}{2}\right)^{3-\nu}\right] G_\nu D^{1+\nu}.$$

For $\nu \in (0,1)$, we get again $\frac{1}{A_t} B_{\nu,t} \leq O(t^{-\nu})$. For $\nu = 1$, we obtain

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{9}{2t+1} G_1 D^2, \quad t \geq 1., \tag{2.17}$$

which is slightly worse than (2.16). The rule for choosing the coefficients τ_t in this situation is $\tau_t \overset{(2.10)}{=} \frac{6(t+1)}{(t+2)(2t+3)}$. It can be easily checked that further increase of the rate of growth of coefficients a_t makes the rate of convergence of method (2.9) even worse.

Note that the above rules for choosing the coefficients $\{\tau_t\}_{t \geq 0}$ in method (2.9) do not depend on the smoothness parameter $\nu \in (0,1]$. In this sense, method (2.9) is a *universal method* for solving the problem (2.1) (see [15]). Moreover, this method does not depend on the choice of the norm in $\mathbb{E}$. Hence, its rate of convergence can be established with respect to the best norm describing the geometry of the feasible set.

3 Trust-region variant of Frank-Wolfe Method

Let us consider a variant of method (2.9), which takes into account the composite form of the objective function in problem (2.1). For $\Psi(x) \equiv \mathrm{Ind}_Q(x)$, these two methods coincide.

Otherwise, they generate different minimization sequences.

Conditional Gradient Method, Type II

1. Choose an arbitrary point $x_0 \in Q$.

2. **For** $t \geq 0$ **iterate:** Choose coefficient $\tau_t \in (0, 1]$ and compute (3.1)

$$x_{t+1} = \arg\min_y \{\langle \nabla f(x_t), y\rangle + \Psi(y) : \; y = (1-\tau_t)x_t + \tau_t x, \; x \in Q\}.$$

This method can be seen as a *Trust-Region Scheme* [1] with linear model of the objective function. Trust region in method (3.1) is formed by a contraction of the initial feasible set. In Section 6, we will consider a more traditional trust-region method with quadratic model of the objective.

Note that point x_{t+1} in method (3.1) is characterized by the following variational principle:

$$\begin{aligned} x_{t+1} &= (1-\tau_t)x_t + \tau_t v_t, \; v_t \in Q, \\ &\\ &\Psi((1-\tau_t)x_t + \tau_t x) + \tau_t \langle \nabla f(x_t), x - x_t\rangle \\ &\\ &\geq \Psi(x_{t+1}) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle, \quad x \in Q. \end{aligned} \tag{3.2}$$

Let us choose somehow the sequence of nonnegative weights $\{a_t\}_{t\geq 0}$, and define in (3.1) the coefficients τ_t in accordance to (2.10). Define now the estimating functional sequence $\{\phi_t(x)\}_{t\geq 0}$ as follows:

$$\begin{aligned} \phi_0(x) &= a_0 \bar{f}(x), \\ \phi_{t+1}(x) &= \phi_t(x) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \Psi(x)], \quad t \geq 0. \end{aligned} \tag{3.3}$$

Clearly, for all $t \geq 0$ we have

$$\phi_t(x) \leq A_t \bar{f}(x), \quad x \in Q. \tag{3.4}$$

Denote

$$C_{\nu,t} = a_0 \Delta(x_0) + \frac{1}{1+\nu} \left(\sum_{k=1}^{t} \frac{a_k^{1+\nu}}{A_k^\nu} \right) G_\nu D^{1+\nu}, \quad t \geq 0. \tag{3.5}$$

Let us introduce

$$\delta(x) \stackrel{\text{def}}{=} \max_{y \in Q} \{\langle \nabla f(x), x - y\rangle + \Psi(x) - \Psi(y)\} \stackrel{(1.4)}{\equiv} \Psi^*_{x,Q}(\nabla f(x)). \tag{3.6}$$

For problem (2.1), this value measures the level of satisfaction of the first-order optimality conditions at point $x \in Q$. For any $x \in Q$, we have

$$\delta(x) \geq \bar{f}(x) - \bar{f}(x_*) \geq 0. \tag{3.7}$$

We call $\delta(x)$ the *total variation* of linear model of the composite objective function in problem (2.1) over the feasible set.

Theorem 2 *Let sequence $\{x_t\}_{t\geq 0}$ be generated by method (3.1). Then, for any $\nu \in (0,1]$ and any step $t \geq 0$, we have*

$$A_t \bar{f}(x_t) \;\leq\; \phi_t(x) + C_{\nu,t}, \quad x \in Q. \tag{3.8}$$

Moreover, for any $t \geq 0$ we have

$$\bar{f}(x_t) - \bar{f}(x_{t+1}) \;\geq\; \tau_t \delta(x_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu} \tau_t^{1+\nu}. \tag{3.9}$$

Proof:
Let us prove inequality (3.8). For $t = 0$, we have $C_{\nu,0} = a_0[\bar{f}(x_0) - \bar{f}(x_*)]$. Thus, in this case (3.8) follows from (3.4).

Assume now that (3.8) is valid for some $t \geq 0$. In view of definition (2.10), optimality condition (3.2) can written in the following form:

$$a_{t+1}\langle \nabla f(x_t), x - x_t \rangle \;\geq\; A_{t+1}\left[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle\right]$$

for all $x \in Q$. Therefore,

$$\begin{aligned}
\phi_{t+1}(x) + C_{\nu,t} &= \phi_t(x) + C_{\nu,t} + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)] \\
&\overset{(3.2),(3.8)}{\geq} A_t[f(x_t) + \Psi(x_t)] + a_{t+1}[f(x_t) + \Psi(x)] \\
&\quad + A_{t+1}\left[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle\right] \\
&\geq A_{t+1}\left[f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \Psi(x_{t+1})\right] \\
&\overset{(2.4)}{\geq} A_{t+1}\left[\bar{f}(x_{t+1}) - \tfrac{1}{1+\nu} G_\nu \|x_{t+1} - x_t\|^{1+\nu}\right].
\end{aligned}$$

It remains to note that $\|x_{t+1} - x_t\| = \tau_t \|x_t - v_t\| \overset{(2.10)}{\leq} \frac{a_{t+1}}{A_{t+1}} D$. Thus, we can take

$$C_{\nu,t+1} \;=\; C_{\nu,t} + \tfrac{1}{1+\nu} \tfrac{a_{t+1}^{1+\nu}}{A_{t+1}^{\nu}} G_\nu D^{1+\nu}.$$

In order to prove inequality (3.9), let us introduce the values

$$\begin{aligned}
\delta_t(\tau) &\overset{\text{def}}{=} \max_{x \in Q}\{\langle \nabla f(x_t), x_t - y \rangle + \Psi(x_t) - \Psi(y) : \; y = (1-\tau)x_t + \tau x\} \\
&\overset{(1.5)}{=} \Psi^*_{\tau, x_t, Q}(\nabla f(x_t)), \quad \tau \in [0,1].
\end{aligned}$$

Clearly,

$$\begin{aligned}
-\delta_t(\tau_t) &= \min_{x \in Q}\{\langle \nabla f(x_t), y - x_t \rangle + \Psi(y) - \Psi(x_t) : y = (1-\tau_t)x_t + \tau_t x\} \\
&= \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \Psi(x_{t+1}) - \Psi(x_t) \\
&\overset{(2.4)}{\geq} \bar{f}(x_{t+1}) - \bar{f}(x_t) - \tfrac{G_\nu}{1+\nu} \|x_{t+1} - x_t\|^{1+\nu}.
\end{aligned}$$

Since $\|x_{t+1} - x_t\| \leq \tau_t D$, we conclude that

$$\bar{f}(x_t) - \bar{f}(x_{t+1}) \;\geq\; \delta_t(\tau_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu}\tau_t^{1+\nu} \overset{(1.6)}{\geq} \tau_t \delta(x_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu}\tau_t^{1+\nu}. \quad \square$$

In view of (3.4), inequality (3.8) results in the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \;\leq\; \tfrac{1}{A_t} C_{\nu,t}, \quad t \geq 0. \tag{3.10}$$

For the linearly growing weights $a_t = t$, $A_t = \frac{t(t+1)}{2}$, $t \geq 0$, we have already seen that

$$C_{\nu,t} \;=\; \tfrac{1}{1+\nu} B_{\nu,t} \;\leq\; \tfrac{2^\nu}{(1+\nu)(2-\nu)} \left[\left(t + \tfrac{1}{2}\right)^{2-\nu} - \left(\tfrac{1}{2}\right)^{2-\nu} \right] G_\nu D^{1+\nu}.$$

In the case $\nu = 1$, this results in the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \;\leq\; \tfrac{2}{t+1} G_1 D^2, \quad t \geq 1. \tag{3.11}$$

Let us justify for this case the rate of convergence of the sequence $\{\delta(x_t)\}_{t \geq 1}$. We have $\tau_t \overset{(2.10)}{=} \frac{a_{t+1}}{A_{t+1}} = \frac{2}{t+2}$. On the other hand, for any $T \geq t$,

$$\tfrac{2G_1 D^2}{t+1} \overset{(3.11)}{\geq} \bar{f}(x_t) - \bar{f}(x_*) \overset{(3.9)}{\geq} \sum_{k=t}^{T} \left[\tau_k \delta(x_k) - \tfrac{1}{2} G_1 D^2 \tau_k^2 \right] + \bar{f}(x_{T+1}) - \bar{f}(x_*). \tag{3.12}$$

Denote $\delta_T^* = \min\limits_{0 \leq t \leq T} \delta(x_t)$.[2] Then, choosing $T = 2t + 1$, we get

$$\begin{aligned} 2 \ln 2 \cdot \delta_T^* &\overset{(1.2)}{\leq} \left(\sum_{k=t}^{T} \tfrac{2}{k+2} \right) \delta_T^* \overset{(3.12)}{\leq} 2 G_1 D^2 \left[\tfrac{1}{t+1} + \sum_{k=t}^{T} \tfrac{1}{(k+2)^2} \right] \\ &\overset{(1.3)}{\leq} 2 G_1 D^2 \left[\tfrac{1}{t+1} + \tfrac{12}{11(2t+3)} \right] \;=\; 2 G_1 D^2 \left[\tfrac{2}{T+1} + \tfrac{12}{11(T+2)} \right] \\ &\leq \tfrac{68}{11} \cdot \tfrac{G_1 D^2}{T+1}. \end{aligned}$$

Thus, in the case $\nu = 1$, for odd T, we get the following bound:

$$\delta_T^* \;\leq\; \tfrac{34}{11 \ln 2} \cdot \tfrac{G_1 D^2}{T+1}. \tag{3.13}$$

4 Computing the primal-dual solution

Note that both methods (2.9) and (3.1) admit computable accuracy certificates. For the first method, denote

$$\ell_t \;=\; \tfrac{1}{A_t} \min_x \left\{ \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] : \; x \in Q \right\}.$$

[2] It seems that such line of arguments was used in the first time in Section 7.5 of [8].

Clearly,

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \leq \quad \bar{f}(x_t) - \ell_t \quad \overset{(2.13)}{\leq} \quad \tfrac{1}{A_t} B_{\nu,t}. \tag{4.1}$$

For the second method, let us choose $a_0 = 0$. Then the estimate functions are linear:

$$\phi_t(x) \quad = \quad \sum_{k=1}^{t} a_k [f(x_{k-1}) + \langle \nabla f(x_{k-1}), x - x_{k-1} \rangle + \Psi(x)].$$

Therefore, defining $\hat{\ell}_t = \frac{1}{A_t} \min\limits_x \{\phi_t(x) : \; x \in Q\}$, we also have

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \leq \quad \bar{f}(x_t) - \hat{\ell}_t \quad \overset{(2.13)}{\leq} \quad \tfrac{1}{A_t} C_{\nu,t}, \quad t \geq 1. \tag{4.2}$$

Accuracy certificates (4.1) and (4.2) justify that both methods (2.9) and (3.1) are able to recover some information on the optimal dual solution. However, in order to implement this ability, we need to open the Black Box and introduce an *explicit model* of the function $f(x)$.

Let us assume that function f is representable in the following form:

$$f(x) \quad = \quad \max_u \{\langle Ax, u \rangle - g(u) : \; u \in Q_d\}, \tag{4.3}$$

where $A : \mathbb{E} \to \mathbb{E}_1^*$, Q_d is a closed convex set in a finite-dimensional linear space $\mathbb{E}_2$, and function $g(\cdot)$ is *p-uniformly convex* on Q_d:

$$\langle \nabla g(u_1) - \nabla g(u_2), u_1 - u_2 \rangle \quad \geq \quad \sigma_g \|u_1 - u_2\|^p, \quad u_1, u_2 \in Q_d,$$

where the *convexity degree* $p \geq 2$.

It is well known (e.g. [15]) that in this case, for $\nu = \frac{1}{p-1}$ we have $G_\nu = \left(\frac{1}{\sigma_g}\right)^{\frac{1}{p-1}}$. In view of Danskin Theorem, $\nabla f(x) = A^* u(x)$, where $u(x) \in Q_d$ is the unique optimal solution to optimization problem in the representation (4.3).

Let us write down the *dual problem* to (2.1).

$$\begin{aligned} \min_x \{\bar{f}(x) : \; x \in Q\} \quad &\overset{(4.3)}{=} \quad \min_x \left\{ \Psi(x) + \max_u \{\langle Ax, u \rangle - g(u) : \; u \in Q_d\} \right\} \\ &\geq \quad \max_{u \in Q_d} \left\{ -g(u) + \min_x \{\langle A^* u, x \rangle + \Psi(x)\} \right\}. \end{aligned}$$

Thus, defining $\Phi(u) = \min\limits_x \{\langle A^* u, x \rangle + \Psi(x)\}$, we get the following dual problem:

$$\max_{u \in Q_d} \left\{ \bar{g}(u) \overset{\text{def}}{=} -g(u) + \Phi(u) \right\}. \tag{4.4}$$

In this problem, the objective function is nonsmooth and uniformly strongly concave of degree p. Clearly, we have

$$\bar{f}(x) - \bar{g}(u) \quad \geq \quad 0, \quad x \in Q, \; u \in Q_d. \tag{4.5}$$

Let us show that both methods (2.9) and (3.1) are able to approximate the optimal solution to the dual problem (4.4).

Note that for any $\bar{x} \in Q$ we have

$$f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \overset{(4.3)}{=} \langle A\bar{x}, u(\bar{x}) \rangle - g(u(\bar{x})) + \langle A^* u(\bar{x}), x - \bar{x} \rangle$$

$$= \langle Ax, u(\bar{x}) \rangle - g(u(\bar{x})).$$

Therefore, denoting for the first method (2.9) $u_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u(x_k)$, we obtain

$$\ell_t = \min_{x \in Q} \left\{ \Psi(x) + \frac{1}{A_t} \sum_{k=0}^{t} a_k [\langle Ax, u(x_k) \rangle - g(u(x_k))] \right\}$$

$$= \Phi(u_t) - \frac{1}{A_t} \sum_{k=0}^{t} a_k g(u(x_k)) \leq \bar{g}(u_t).$$

Thus, we get

$$0 \overset{(4.5)}{\leq} \bar{f}(x_t) - \bar{g}(u_t) \leq \bar{f}(x_t) - \ell_t \overset{(4.1)}{\leq} \frac{1}{A_t} B_{\nu,t}, \quad t \geq 0. \tag{4.6}$$

For the second method (3.1), we need to choose $a_0 = 0$ and take $u_t = \frac{1}{A_t} \sum\limits_{k=1}^{t} a_k u(x_{k-1})$. In this case, by the similar reasoning, we get

$$0 \overset{(4.5)}{\leq} \bar{f}(x_t) - \bar{g}(u_t) \leq \bar{f}(x_t) - \hat{\ell}_t \overset{(4.2)}{\leq} \frac{1}{A_t} C_{\nu,t}, \quad t \geq 1. \tag{4.7}$$

5 Strong convexity of function Ψ

In this section, we assume that function Ψ in problem (2.1) is *strongly convex* (see, for example, Section 2.1.3 in [12]). This means that there exists a positive constant σ_Ψ such that

$$\Psi(\tau x + (1-\tau)y) \leq \tau \Psi(x) + (1-\tau)\Psi(y) - \tfrac{1}{2}\sigma_\Psi \tau(1-\tau)\|x-y\|^2 \tag{5.1}$$

for all $x, y \in Q$ and $\tau \in [0,1]$. Let us show that in this case CG-methods converge much faster. We demonstrate it for method (2.9).

In view of strong convexity of Ψ, the variational principle (2.8), characterizing the point v_t in method (2.9) can be strengthen:

$$\Psi(x) + \langle \nabla f(x_t), x - v_t \rangle \geq \Psi(v_t) + \tfrac{1}{2}\sigma_\psi \|x - v_t\|^2, \quad x \in Q. \tag{5.2}$$

Let V_0 be defined as in (2.11). Denote

$$\hat{B}_{\nu,t} = a_0 V_0 + \left(\sum_{k=1}^{t} \frac{a_k^{1+2\nu}}{A_k^{2\nu}} \right) \frac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}, \quad t \geq 0. \tag{5.3}$$

Theorem 3 *Let sequence $\{x_t\}_{t \geq 0}$ be generated by method (2.9), and function Ψ is strongly convex. Then, for any $\nu \in (0,1]$, any step $t \geq 0$, and any $x \in Q$ we have*

$$A_t(f(x_t) + \Psi(x_t)) \leq \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] + \hat{B}_{\nu,t}. \tag{5.4}$$

Proof:
The beginning of the proof of this statement is very similar to that of Theorem 1. Assuming that (5.4) is valid for some $t \geq 0$, we get the following inequality:

$$\begin{aligned} & \sum_{k=0}^{t+1} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] + B_{\nu,t} \\ \geq \; & A_{t+1} \left(f(x_{t+1}) + \Psi(x_{t+1}) \right) + a_{t+1} \left[\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle \right]. \end{aligned}$$

Further.

$$\begin{aligned} \Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle & \overset{(5.2)}{\geq} \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t \rangle + \tfrac{1}{2}\sigma_\Psi \|x - v_t\|^2 \\ & \overset{(2.3)}{\geq} -\tfrac{1}{2\sigma_\Psi} \|\nabla f(x_{t+1}) - \nabla f(x_t)\|_*^2 \\ & \overset{(2.3)}{\geq} -\tfrac{1}{2\sigma_\Psi} \left(\tfrac{a_{t+1}^\nu}{A_{t+1}^\nu} G_\nu D^\nu \right)^2. \end{aligned}$$

Thus, for keeping (5.4) valid for the next iteration, it is enough to choose

$$\hat{B}_{\nu,t+1} \;=\; \hat{B}_{\nu,t} + \tfrac{1}{2\sigma_\Psi} \tfrac{a_{t+1}^{1+2\nu}}{A_{t+1}^{2\nu}} G_\nu^2 D^{2\nu}.$$

□

It can be easily checked that in our situation, the linear weights strategy $a_t \equiv t$ is not the best one. Let us choose $a_t = t^2$, $t \geq 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$, and we get

$$\begin{aligned} \hat{B}_{\nu,t} \;=\; & \left(\sum_{k=1}^{t} \tfrac{6^{2\nu} k^{2(1+2\nu)}}{k^{2\nu}(k+1)^{2\nu}(2k+1)^{2\nu}} \right) \tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi} \;\leq\; \left(3^{2\nu} \sum_{k=1}^{t} k^{2(1-\nu)} \right) \tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi} \\ \overset{(1.1)}{\leq} \; & \tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi} \cdot \tfrac{3^{2\nu}}{3-2\nu} \tau^{3-2\nu} \Big|_{1/2}^{t+1/2} \;=\; \tfrac{3^{2\nu}}{3-2\nu} \left[(t+\tfrac{1}{2})^{3-2\nu} - \left(\tfrac{1}{2}\right)^{3-2\nu} \right] \tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}. \end{aligned}$$

Thus, for $\nu \in (0,1)$, we get $\frac{1}{A_t}\hat{B}_{\nu,t} \leq O(t^{-2\nu})$. For $\nu = 1$, we obtain

$$\bar{f}(x_t) - \bar{f}(x_*) \;\leq\; \tfrac{54}{(t+1)(2t+1)} \cdot \tfrac{G_1^2 D^2}{2\sigma_\Psi}, \tag{5.5}$$

which is much better than (2.16).

6 Second-order trust-region method

Let us assume now that in problem (2.1) function f is twice continuously differentiable. Then we can apply to this problem the following method.

Contracting Trust-Region Method

1. Choose an arbitrary point $x_0 \in Q$.

2. **For** $t \geq 0$ **iterate:** Define coefficient $\tau_t \in (0, 1]$ and choose (6.1)

$$x_{t+1} \in \mathrm{Arg}\min_y \Big\{ \quad \langle \nabla f(x_t), y - x_t \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_t)(y - x_t), y - x_t \rangle + \Psi(y) : \\ y \in (1 - \tau_t)x_t + \tau_t x,\ x \in Q \quad \Big\}.$$

Note that this scheme is well defined even if the Hessian of function f is positive semidefinite. Of course, in general, the computational cost of each iteration of this scheme can be big. However, in one important case, when $\Psi(x)$ is an indicator function of a Euclidean ball, the complexity of each iteration of this scheme is dominated by complexity of matrix inversion. Thus, method (6.1) can be easily applied to problems of the form

$$\min_x \{ f(x) :\ \|x - x_0\| \leq r \}, \tag{6.2}$$

where the norm $\|\cdot\|$ is Euclidean.

Let $H_\nu < +\infty$ for some $\nu \in (0, 1]$. In this section we assume that

$$\langle \nabla^2 f(x)h, h \rangle \ \leq L\|h\|^2, \quad x \in Q,\ h \in \mathbb{E}. \tag{6.3}$$

Let us choose a sequence of nonnegative weights $\{a_t\}_{t \geq 0}$, and define in (6.1) the coefficients $\{\tau_t\}_{t \geq 0}$ in accordance to (2.10). Define the estimate functional sequence $\{\phi_t(x)\}_{t \geq 0}$ by recurrent relations (3.3), where the sequence $\{x_t\}_{t \geq 0}$ is generated by method (6.1). Finally, denote

$$\hat{C}_{\nu,t} \ = \ a_0 \Delta(x_0) + \left(\sum_{k=1}^{t} \tfrac{a_k^{2+\nu}}{A_k^{1+\nu}} \right) \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + \left(\sum_{k=1}^{t} \tfrac{a_k^2}{2A_k} \right) LD^2. \tag{6.4}$$

In our convergence results, we estimate also the level of the second-order optimality condition for problem (2.1) at the current test points. Let us introduce

$$\theta(x) \stackrel{\mathrm{def}}{=} \max_{y \in Q} \{ \langle \nabla f(x), x - y \rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle + \Psi(x) - \Psi(y) \}. \tag{6.5}$$

For any $x \in Q$ we have $\theta(x) \geq 0$. We call $\theta(x)$ the *total variation* of quadratic model of the composite objective function in problem (2.1) over the feasible set. Denoting

$$F_x(y) \ = \ \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle + \Psi(y),$$

we get $\theta(x) = \Big(F_x \Big)^*_{x,Q} (\nabla f(x))$ (see definition (1.4)).

Theorem 4 *Let sequence $\{x_t\}_{t \geq 0}$ be generated by method* (6.1)*. Then, for any $\nu \in [0, 1]$ and any step $t \geq 0$ we have*

$$A_t \bar{f}(x_t) \ \leq \ \phi_t(x) + \hat{C}_{\nu,t}, \quad x \in Q. \tag{6.6}$$

Moreover, for any $t \geq 0$ we have

$$\bar{f}(x_t) - \bar{f}(x_{t+1}) \ \geq \ \tau_t \theta(x_t) - \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} \tau_t^{2+\nu}. \tag{6.7}$$

Proof:
Let us prove inequality (6.6). For $t = 0$, $\hat{C}_{\nu,0} = a_0[\bar{f}(x_0) - \bar{f}(x_*)]$. Therefore, this inequality is valid.

Note that the point x_{t+1} is characterized by the following variational principle:

$$x_{t+1} = (1-\tau_t)x_t + \tau_t v_t, \quad v_t \in Q,$$

$$\Psi(y) + \langle \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t), y - x_{t+1} \rangle \geq \Psi(x_{t+1}),$$

$$\forall\, y = (1-\tau_t)x_t + \tau_t x, \quad x \in Q.$$

Therefore, in view of definition (2.10), for any $x \in Q$ we have

$$\begin{aligned} a_{t+1}\langle \nabla f(x_t), x - x_t \rangle \quad &\geq \quad A_{t+1}\langle \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle \\ &\quad + a_{t+1}\langle \nabla^2 f(x_t)(x_{t+1} - x_t), x_t - x \rangle \\ &\quad + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] \\ &\overset{(6.3)}{\geq} \quad A_{t+1}\langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle \\ &\quad + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2. \end{aligned}$$

Hence,

$$\begin{aligned} & A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)] \\ \geq \quad & A_t \Psi(x_t) + A_{t+1}[f(x_t) + \langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle] \\ & + a_{t+1}\Psi(x) + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2 \\ \overset{(2.6)}{\geq} \quad & A_{t+1}[f(x_{t+1}) + \Psi(x_{t+1})] - A_{t+1}\tfrac{H_\nu \|x_{t+1} - x_t\|^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2 \\ \geq \quad & A_{t+1}\bar{f}(x_{t+1}) - \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2. \end{aligned}$$

Thus, if (6.6) is valid for some $t \geq 0$, then

$$\begin{aligned} \phi_{t+1}(x) + \hat{C}_{\nu,t} \quad &\geq \quad A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)] \\ &\geq \quad A_{t+1}\bar{f}(x_{t+1}) - \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2. \end{aligned}$$

Therefore, we can take $\hat{C}_{\nu,t+1} = \hat{C}_{\nu,t} + \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + \tfrac{a_{t+1}^2}{2A_{t+1}} LD^2$.

In order to justify inequality (6.7), let us introduce the values

$$\begin{aligned}\theta_t(\tau) &\stackrel{\text{def}}{=} \max_{x\in Q}\{\langle\nabla f(x_t), x_t - y\rangle - \tfrac{1}{2}\langle\nabla^2 f(x_t)(y-x_t), y-x_t\rangle \\ &\qquad +\Psi(x_t) - \Psi(y) :\ y = (1-\tau)x_t + \tau x\} \\ &\stackrel{(1.5)}{=} \left(F_{x_t}\right)^*_{\tau,x_t,Q}(\nabla f(x_t)), \quad \tau\in[0,1].\end{aligned}$$

Clearly,

$$\begin{aligned}-\theta_t(\tau_t) = &\min_{x\in Q}\{\langle\nabla f(x_t), y - x_t\rangle - \tfrac{1}{2}\langle\nabla^2 f(x_t)(y-x_t), y-x_t\rangle \\ &+\Psi(y) - \Psi(x_t) : y = (1-\tau_t)x_t + \tau_t x\} \\ = &\ \langle\nabla f(x_t), x_{t+1} - x_t\rangle - \tfrac{1}{2}\langle\nabla^2 f(x_t)(x_{t+1}-x_t), x_{t+1}-x_t\rangle + \Psi(x_{t+1}) - \Psi(x_t) \\ \stackrel{(2.6)}{\geq} &\ \bar f(x_{t+1}) - \bar f(x_t) - \tfrac{H_\nu}{(1+\nu)(2+\nu)}\|x_{t+1}-x_t\|^{2+\nu}.\end{aligned}$$

Since $\|x_{t+1} - x_t\| \leq \tau_t D$, we conclude that

$$\bar f(x_t) - \bar f(x_{t+1}) \ \geq\ \theta_t(\tau_t) - \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)}\tau_t^{2+\nu} \stackrel{(1.6)}{\geq} \tau_t\theta(x_t) - \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)}\tau_t^{2+\nu}. \quad \square$$

Thus, inequality (6.6) ensures the following rate of convergence of method (6.1)

$$\bar f(x_t) - \bar f(x_*) \ \leq\ \tfrac{1}{A_t}\hat C_{\nu,t}. \tag{6.8}$$

A particular expression of the right-hand side of this inequality for different values of $\nu\in[0,1]$ can be obtained exactly in the same way as it was done in Section 2. Here, we restrict ourselves only by the case $\nu = 1$ and $a_t = t^2$, $t\geq 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$, and

$$\begin{aligned}\sum_{k=1}^t \tfrac{a_k^3}{A_k^2} &= \sum_{k=1}^t \tfrac{36k^6}{k^2(k+1)^2(2k+1)^2} \ \leq\ 18t, \\ \sum_{k=1}^t \tfrac{a_k^2}{2A_k} &= \sum_{k=1}^t \tfrac{3k^4}{k(k+1)(2k+1)} \ \leq\ \tfrac{3}{2}\sum_{k=1}^t k = \tfrac{3}{4}t(t+1).\end{aligned}$$

Thus, we get

$$\bar f(x_t) - \bar f(x_*) \ \leq\ \tfrac{18H_1D^3}{(t+1)(2t+1)} + \tfrac{9LD^2}{2(2t+1)}. \tag{6.9}$$

Note that the rate of convergence (6.9) is worse than the convergence rate of cubic regularization of the Newton method [17]. However, to the best of our knowledge, inequality (6.9) gives us the first global rate of convergence of an optimization scheme belonging to the family of trust-region methods [1]. In view of inequality (6.6), the optimal solution of the dual problem (4.4) can be approximated by method (6.1) with $a_0 = 0$ in the same way as it was suggested in Section 4 for Conditional Gradient Methods.

Let us estimate now the rate of decrease of the values $\theta(x_t)$, $t \geq 0$, in the case $\nu = 1$. Note that $\tau_t \stackrel{(2.10)}{=} \frac{a_{t+1}}{A_{t+1}} = \frac{6(t+1)}{(t+2)(2t+3)}$. It is easy to see that these coefficients satisfy the following inequalities:

$$\frac{3}{t+3} \;\leq\; \tau_t \;\leq\; \frac{6}{2t+5}, \quad t \geq 0. \tag{6.10}$$

Therefore, choosing the total number of steps $T = 2t + 2$, we have

$$\begin{aligned}
\sum_{k=t}^{T} \tau_k \;&\stackrel{(6.10)}{\geq}\; 3 \sum_{k=t}^{2t+2} \frac{1}{k+3} \;\stackrel{(1.2)}{\geq}\; 3 \ln 2, \\
\sum_{k=t}^{T} \tau_k^3 \;&\stackrel{(6.10)}{\leq}\; \sum_{k=t}^{2t+2} \frac{27}{(k+5/2)^3} \;\stackrel{(1.3)}{\leq}\; -\frac{27}{2(k+5/2)^2}\Big|_{t-1/2}^{2t+5/2} \;=\; \frac{27}{2}\left[\frac{1}{(t+2)^2} - \frac{1}{(2t+5)^2}\right] \qquad (6.11) \\
&=\; \frac{27}{2}\left[\frac{4}{(T+2)^2} - \frac{1}{(T+3)^2}\right] \;=\; \frac{27(3T+8)(T+4)}{2(T+2)^2(T+3)^2} \;\leq\; \frac{81}{2(T+1)(T+2)}.
\end{aligned}$$

Now we can use the same trick as in the end of Section 3. Denote $\theta_T^* = \min\limits_{0 \leq t \leq T} \theta(x_t)$. Then

$$\begin{aligned}
\tfrac{36 H_1 D^3}{T(T-1)} + \tfrac{9LD^2}{2(T-1)} \;&\stackrel{(6.9)}{\geq}\; \bar{f}(x_t) - \bar{f}(x_*) \;\geq\; \sum_{k=t}^{T} (\bar{f}(x_k) - \bar{f}(x_{k+1})) \\
&\stackrel{(6.7)}{\geq}\; \theta_T^* \sum_{k=t}^{T} \tau_k - \tfrac{H_1 D^3}{6} \sum_{k=t}^{T} \tau_k^3 \;\stackrel{(6.11)}{\geq}\; 3\theta_T^* \ln 2 - \tfrac{27 H_1 D^3}{4(T+1)(T+2)}.
\end{aligned}$$

Thus, for even T, we get the following bound:

$$\begin{aligned}
\theta_T^* \;&\leq\; \tfrac{3}{\ln 2}\left[\tfrac{4 H_1 D^3}{T(T-1)} + \tfrac{3 H_1 D^3}{4(T+1)(T+2)} + \tfrac{LD^2}{2(T-1)}\right] \\
&\leq\; \tfrac{3}{\ln 2}\left[\tfrac{5 H_1 D^3}{T(T-1)} + \tfrac{LD^2}{2(T-1)}\right].
\end{aligned} \tag{6.12}$$

References

[1] Conn A.B., Gould N.I.M., Toint Ph.L., *Trust Region Methods*, SIAM, Philadelphia, 2000.

[2] J. Dunn. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, **18**(5): 473-487, (1980).

[3] M. Frank, P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, **3**: 149-154 (1956).

[4] R.M. Freund, P. Grigas. New analysis and results for the FrankWolfe method. *Mathematical Programming*, DOI 10.1007/s10107-014-0841-6, (2014).

[5] D. Garber, E. Hazan. A linearly convergent conditional gradient algorithm with application to online and stochastic optimization. arXiv: 1301.4666v5, (2013).

[6] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, DOI 10.1007/s10107-014-0778-9, (2014).

[7] M. Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *Proceedings of the 30th International Conderence on Machine Learning*, Atlanta, Georgia, (2013).

[8] A. Juditsky and A. Nemirovski. Solving variational inequalities with monotone operators on domains given by Linear Minimization Oracles. *Mathemetical Programming*, Ser. A, DOI 10.1007/s10107-015-0876-3.

[9] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization of structural svms. In *Proceedings of the 30th International Conderence on Machine Learning*, Atlanta, Georgia, (2013).

[10] G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. arXiv: 1309.5550v2, (2014).

[11] A. Migdalas. A regularization of the Frank-Wolfe method and unification of certain nonlinear programming methods. *Mathematical Programming*, **63**, 331-345, (1994)

[12] Yu. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, Boston, 2004.

[13] Yu. Nesterov. Primal-dual subgradient methods. (2009)

[14] Yu.Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, **140**(1), 125-161 (2013).

[15] Yu. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, DOI: 10.1007/s10107-014-0790-0, (2014).

[16] Nesterov Yu., Nemirovskii A., *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[17] Yu. Nesterov, B. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, **108**(1), 177-205, (2006).

[18] Yu. Nesterov, V. Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *JOTA*, DOI 10.1007/s10957-014-0677-5, (2014).