■

# Complexity Bounds for Primal-dual Methods Minimizing the Model of Objective Function

Yurii Nesterov

CORE

CORE DISCUSSION PAPER
2015/3

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

Yurii NESTEROV[*]

February 2, 2015

## Abstract

We provide Frank-Wolfe ($\equiv$ 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 the linear model of objective function. Our analysis works also for a quadratic model, allowing to justify the global rate of convergence for a new second-order method. To the best of our knowledge, this is the first trust-region scheme supported by the worst-case complexity bound.

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

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

Yu. Nesterov [*]

February 2, 2015

## Abstract

We provide Frank-Wolfe ($\equiv$ 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 the linear model of objective function. Our analysis works also for a quadratic model, allowing to justify the global rate of convergence for a new second-order method. To the best of our knowledge, this is the first trust-region scheme supported by the worst-case complexity bound.

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

# 1    Introduction

**Motivation.** In the last years, we can see an increasing interest to Frank-Wolfe algorithm [3, 2, 10], which sometimes is called *Conditional Gradient Method* (CGM) [5, 7, 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. [11]). 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 [9].

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* [12]. 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 motivation 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 *estimate sequences* (e.g. [11]) in its extended form [16]. As a result, we get a significant freedom in the choice of step-size coefficients. In this section we also consider a variant of augmented 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.

Our bounds for the primal-dual gap are very similar to bounds in [4]. However, they are obtained for the difference of the current value of the objective function and the minimal value of the accumulated linear model. In Section 3 we explain how to extract from this information 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 4, we show that the additional properties of complementary part of the objective (strong convexity) significantly accelerate the scheme. Finally, in Section 5, 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 contracted feasible set. To the best of our knowledge, this is the first trust-region scheme [1] supported by the worst-case complexity analysis.

---

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

1

**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

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

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 $\langle Bx, x \rangle \geq 0$ 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:

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

For a differentiable function $f(x)$ with 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 from above 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 convex. Then, for any two integers $a$ and $b$, such that $[a - \frac{1}{2}, b + \frac{1}{2}] \subset \mathrm{dom}\, \xi$, we have*

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

# 2 Conditional gradient methods

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

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

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

The most important examples of function $\Psi$ are as follows.

- $\Psi$ is an indicator function of a closed convex set $Q$:

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

- $\Psi$ is a self-concordant barrier for a closed convex set $Q$ (see [14, 11]).

- $\Psi$ 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*:

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

Constant $L_\nu$ is formally defined for any $\nu \in (0, 1]$. For some values of $\nu$ it can be $+\infty$. Note that for any $x$ and $y$ in $\operatorname{dom} \Psi$ we have

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

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

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

In this case, for any $x$ and $y$ in $\operatorname{dom} \Psi$ we have

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

For *Conditional Gradient Methods* (CGM) our assumption on simplicity of function $\Psi$ means exactly the following.

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

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

*is easily solvable. Denote by $v_\Psi(s) \in \operatorname{dom} \Psi$ one of its optimal solutions.*

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

$$\min_x \{\langle s, x \rangle : x \in \operatorname{dom} \Psi\}.$$

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

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

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

<div align="center">

**Conditional Gradient Method, Type I**

</div>

    **1.**  Choose an arbitrary point $x_0 \in \operatorname{dom} \Psi$.

$$\tag{2.9}$$

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

          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 $x \in \mathrm{dom}\,\Psi$ 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 triangle $\mathrm{Conv}\{x_0, y_1, y_2\}$, which does not contain $x_*$.* $\square$

In order to justify the rate of convergence of method (2.9) for functions with Hölder continuous gradients, we apply the estimate sequences technique [11] in its relaxed form [16]. 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

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

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

$$V_0 \;\; = \;\; \max_x \left\{ \langle \nabla f(x_0), x_0 - x \rangle + \Psi(x_0) - \Psi(x) \right\},$$

$$B_{\nu,t} \;\; = \;\; a_0 V_0 + \left( \sum_{k=1}^t \frac{a_k^{1+\nu}}{A_k^\nu} \right) G_\nu D^{1+\nu}, \quad t \geq 0. \tag{2.11}$$

It is clear that

$$V_0 \;\; \overset{(2.6)}{\leq} \;\; \max_x \left\{ f(x_0) - f(x) + \frac{G_\nu}{1+\nu} \|x - x_0\|^{1+\nu} + \Psi(x_0) - \Psi(x) \right\}$$

$$\leq \;\; \bar{f}(x_0) - \bar{f}(x_*) + \frac{G_\nu D^{1+\nu}}{1+\nu} \;\; \overset{\text{def}}{=} \;\; \Delta(x_0) + \frac{G_\nu D^{1+\nu}}{1+\nu}. \tag{2.12}$$

**Lemma 2** *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 $x \in \mathrm{dom}\,\Psi$ 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

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

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

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

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

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

It remains to note that

$$\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle \overset{(2.8)}{\geq} \quad \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t \rangle$$

$$\overset{(2.3)}{\geq} \quad -\tau_t^\nu L_\nu D^{1+\nu}.$$

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

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

$\square$

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

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

Let us discuss now 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

$$B_{\nu,t} \quad = \quad V_0 + \left( \sum_{k=1}^{t} \tfrac{1}{(1+k)^\nu} \right) G_\nu D^{1+\nu} \overset{(1.1)}{\leq} V_0 + G_\nu D^{1+\nu} \tfrac{1}{1-\nu} (1+\tau)^{1-\nu} \Big|_{1/2}^{t+1/2}$$

$$\overset{(2.12)}{\leq} \quad \Delta(x_0) + G_\nu D^{1+\nu} \left[ \tfrac{1}{1+\nu} + \left( \tfrac{3}{2} \right)^{1-\nu} \tfrac{1}{1-\nu} \left( \left( 1 + \tfrac{2}{3}t \right)^{1-\nu} - 1 \right) \right]$$

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 $\lim_{\nu \to 1} \tfrac{1}{1-\nu} \left( \left( 1 + \tfrac{2}{3}t \right)^{1-\nu} - 1 \right) = \ln(1 + \tfrac{2}{3}t)$. Therefore,

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

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

5

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 CGM (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 $\tau_t$ in this situation is $\tau_t \overset{(2.10)}{=} \frac{6(t+1)}{(t+2)(2t+3)}$. It can be easily checked that the 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 [13]). 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.

To conclude this section, let us consider a variant of method (2.9). For $\Psi(x) \equiv \text{Ind}_Q(x)$ these two methods coincide. Otherwise, they generate different minimization sequences.

### Conditional Gradient Method, Type II

1. Choose an arbitrary point $x_0 \in \text{dom}\,\Psi$.

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

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

(2.18)

6

This method can be seen as a *Trust-Region Scheme* [1] with linear model of objective function. The trust region in (2.18) is formed by a contraction of the initial feasible set. In Section 5, we consider a more traditional trust-region method with quadratic model of the objective.

Note that point $x_{t+1}$ in method (2.18) is characterized by the following variational principle:

$$x_{t+1} = (1 - \tau_t)x_t + \tau_t v_t, \ v_t \in \text{dom} \, \Psi,$$

$$\Psi((1 - \tau_t)x_t + \tau_t x) + \tau_t \langle \nabla f(x_t), x - x_t \rangle \qquad (2.19)$$

$$\geq \ \Psi(x_{t+1}) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle, \quad x \in \text{dom} \, \Psi.$$

Let us choose somehow the sequence of nonnegative weights $\{a_t\}_{t \geq 0}$, and define in (2.18) the coefficients $\tau_t$ in accordance to (2.10). Define now the estimate functional sequence $\{\phi_t(x)\}_{t \geq 0}$ as follows:

$$\phi_0(x) = a_0 \bar{f}(x),$$

$$\phi_{t+1}(x) = \phi_t(x) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)], \quad t \geq 0. \qquad (2.20)$$

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

$$\phi_t(x) \leq A_t \bar{f}(x), \quad x \in \text{dom} \, \Psi. \qquad (2.21)$$

Denote

$$C_{\nu,t} = a_0 \Delta(x_0) + \tfrac{1}{1+\nu} \left( \sum_{k=1}^{t} \frac{a_k^{1+\nu}}{A_k^{\nu}} \right) G_\nu D^{1+\nu}, \quad t \geq 0. \qquad (2.22)$$

**Lemma 3** *Let sequence $\{x_t\}_{t \geq 0}$ be generated by method (2.18). Then, for any $\nu \in (0,1]$ and any step $t \geq 0$, we have*

$$A_t \bar{f}(x_t) \leq \phi_t(x) + C_{\nu,t}, \quad x \in \text{dom} \, \Psi. \qquad (2.23)$$

**Proof:**
For $t = 0$, we have $C_{\nu,0} = a_0[\bar{f}(x_0) - \bar{f}(x_*)]$. Thus, (2.23) follows from (2.21).

Assume now that (2.23) is valid for some $t \geq 0$. In view of definition (2.10), optimality condition (2.19) 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 $x \in \text{dom} \, \Psi$. Therefore,

$$\phi_{t+1}(x) + C_{\nu,t} = \phi_t(x) + C_{\nu,t} + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)]$$

$$\overset{(2.23)}{\geq} A_t[f(x_t) + \Psi(x_t)] + a_{t+1}[f(x_t) + \Psi(x)]$$

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

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

$$\overset{(2.4)}{\geq} A_{t+1}\left[\bar{f}(x_{t+1}) - \tfrac{1}{1+\nu}G_\nu \|x_{t+1} - x_t\|^{1+\nu}\right].$$

7

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

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

$\square$

In view of (2.21), inequality (2.23) results in the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \le \quad \tfrac{1}{A_t} C_{\nu,t}, \quad t \ge 0. \tag{2.24}$$

Note that for linearly growing weights $a_t = t$, $A_t = \frac{t(t+1)}{2}$, $t \ge 0$, we have already estimated

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

Therefore, for $\nu = 1$, we get the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \le \quad \tfrac{2}{t+1} G_1 D^2, \quad t \ge 1. \tag{2.25}$$

# 3   Computing the primal-dual solution

Note that both methods (2.9) and (2.18) admit computable accuracy certificates. For the first method, denote

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

Clearly,

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \le \quad \bar{f}(x_t) - \ell_t \overset{(2.13)}{\le} \tfrac{1}{A_t} B_{\nu,t}. \tag{3.1}$$

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 $\hat{\ell}_t = \tfrac{1}{A_t} \min_x \{ \phi_t(x) : \ x \in \mathrm{dom}\, \Psi \}$, we also have

$$\bar{f}(x_t) - \bar{f}(x_*) \quad \le \quad \bar{f}(x_t) - \hat{\ell}_t \overset{(2.13)}{\le} \tfrac{1}{A_t} C_{\nu,t}, \quad t \ge 1. \tag{3.2}$$

Accuracy certificates (3.1) and (3.2) justify that both methods (2.9) and (2.18) 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:

$$f(x) \quad = \quad \max_u \{ \langle Ax, u \rangle - g(u) : \ u \in Q_d \}, \tag{3.3}$$

8

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 $Q_d$:

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

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

It is well known (e.g. [13]) 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 (3.3).

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

$$\min_x \{\bar{f}(x) : \ x \in \text{dom } \Psi\} \ \stackrel{(3.3)}{=} \ \min_x \left\{ \Psi(x) + \max_u \{\langle Ax, u \rangle - g(u) : \ u \in Q_d\} \right\}$$

$$\geq \ \max_{u \in Q_d} \left\{ -g(u) + \min_x \{\langle A^*u, x \rangle + \Psi(x)\} \right\}.$$

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

$$\max_{u \in Q_d} \left\{ \bar{g}(u) \stackrel{\text{def}}{=} -g(u) + \Phi(u) \right\}. \tag{3.4}$$

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

$$\bar{f}(x) - \bar{g}(u) \ \geq \ 0, \quad x \in \text{dom } \Psi, \ u \in Q_d. \tag{3.5}$$

Let us show that both methods (2.9) and (2.18) are able to approximate the optimal solution to the dual problem (3.4).

Note that for any $\bar{x} \in \text{dom } \Psi$ we have

$$f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \ \stackrel{(3.3)}{=} \ \langle A\bar{x}, u(\bar{x}) \rangle - g(u(\bar{x})) + \langle A^*u(\bar{x}), x - \bar{x} \rangle$$

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

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

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

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

Thus, we get

$$0 \ \stackrel{(3.5)}{\leq} \ \bar{f}(x_t) - \bar{g}(u_t) \ \leq \ \bar{f}(x_t) - \ell_t \ \stackrel{(3.1)}{\leq} \ \frac{1}{A_t} B_{\nu,t}, \quad t \geq 0. \tag{3.6}$$

For the second method (2.18), we need to choose $a_0 = 0$ and take $u_t = \frac{1}{A_t} \sum_{k=1}^{t} a_k u(x_{k-1})$. In this case, by the similar reasoning, we get

$$0 \ \stackrel{(3.5)}{\leq} \ \bar{f}(x_t) - \bar{g}(u_t) \ \leq \ \bar{f}(x_t) - \hat{\ell}_t \ \stackrel{(3.2)}{\leq} \ \frac{1}{A_t} C_{\nu,t}, \quad t \geq 1. \tag{3.7}$$

9

# 4 Strong convexity of function $\Psi$

In this section, we assume that function $\Psi$ in problem (2.1) is *strongly convex* (see, for example, Section 2.1.3 in [11]). This means that there exists a positive constant $\sigma_\Psi$ such that

$$\Psi(\tau x + (1-\tau)y) \quad \leq \quad \tau\Psi(x) + (1-\tau)\Psi(y) - \tfrac{1}{2}\sigma_\Psi\tau(1-\tau)\|x-y\|^2 \qquad (4.1)$$

for all $x, y \in \operatorname{dom}\Psi$ 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 $\Psi$, the variational principle (2.8), characterizing the point $v_t$ in method (2.9) can be strengthen:

$$\Psi(x) + \langle \nabla f(x_t), x - v_t \rangle \quad \geq \quad \Psi(v_t) + \tfrac{1}{2}\sigma_\psi\|x - v_t\|^2, \quad x \in \operatorname{dom}\Psi. \qquad (4.2)$$

Let $V_0$ be defined as in (2.11). Denote

$$\hat{B}_{\nu,t} \quad = \quad a_0 V_0 + \left(\sum_{k=1}^{t} \frac{a_k^{1+2\nu}}{A_k^{2\nu}}\right)\frac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}, \quad t \geq 0. \qquad (4.3)$$

**Lemma 4** *Let sequence $\{x_t\}_{t\geq 0}$ be generated by method (2.9), and function $\Psi$ is strongly convex. Then, for any $\nu \in (0,1]$, any step $t \geq 0$, and any $x \in \operatorname{dom}\Psi$ we have*

$$A_t(f(x_t) + \Psi(x_t)) \quad \leq \quad \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] + \hat{B}_{\nu,t}. \qquad (4.4)$$

**Proof:**
The beginning of the proof of this statement is very similar to that of Lemma 2. Assuming that (4.4) is valid for some $t \geq 0$, we get the following inequality:

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

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

Further.

$$\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t \rangle \quad \overset{(4.2)}{\geq} \quad \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t \rangle + \tfrac{1}{2}\sigma_\Psi\|x - v_t\|^2$$

$$\overset{(2.3)}{\geq} \quad -\tfrac{1}{2\sigma_\Psi}\|\nabla f(x_{t+1}) - \nabla f(x_t)\|_*^2$$

$$\overset{(2.3)}{\geq} \quad -\tfrac{1}{2\sigma_\Psi}\left(\frac{a_{t+1}^\nu}{A_{t+1}^\nu}G_\nu D^\nu\right)^2.$$

Thus, for keeping (4.4) valid for the next iteration, it is enough to choose

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

$\square$

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

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

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

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

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{54}{(t+1)(2t+1)} \cdot \frac{G_1^2 D^2}{2\sigma_\Psi}, \tag{4.5}$$

which is much better than (2.16).

# 5 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 *Trust-Region Method*.

<div align="center"><u>**Trust-Region Method**</u></div>

1. Choose an arbitrary point $x_0 \in \operatorname{dom} \Psi$.

2. **For $t \geq 0$ iterate:** Define coefficient $\tau_t \in (0,1]$ and choose $\qquad\qquad$ (5.1)

$$x_{t+1} \in \operatorname{Arg\,min}_y \Big\{ \ \langle \nabla f(x_t), y - x_t \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_t)(y-x_t), y - x_t \rangle + \Psi(y) :$$
$$y \in (1-\tau_t)x_t + \tau_t \operatorname{dom} \Psi \ \Big\}.$$

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 quite 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 the complexity of matrix inversion. Thus, method (5.1) can be easily applied to problems of the form

$$\min_x \{ f(x) : \ \|x - x_0\| \leq r \}, \tag{5.2}$$

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

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

$$\langle \nabla^2 f(x)h, h \rangle \leq L\|h\|^2, \quad x \in \operatorname{dom} \Psi, \ h \in \mathbb{E}. \tag{5.3}$$

Let us choose a sequence of nonnegative weights $\{a_t\}_{t \geq 0}$, and define in (5.1) 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 (2.20), where the sequence $\{x_t\}_{t \geq 0}$ is generated by method (5.1). Finally, denote

$$\hat{C}_{\nu,t} = a_0 \Delta(x_0) + \left( \sum_{k=1}^{t} \frac{a_k^{2+\nu}}{A_k^{1+\nu}} \right) \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + \left( \sum_{k=1}^{t} \frac{a_k^2}{2A_k} \right) LD^2. \qquad (5.4)$$

**Lemma 5** *Let sequence $\{x_t\}_{t \geq 0}$ be generated by method (5.1). Then, for any $\nu \in [0, 1]$ and any step $t \geq 0$ we have*

$$A_t \bar{f}(x_t) \leq \phi_t(x) + \hat{C}_{\nu,t}, \quad x \in \operatorname{dom} \Psi. \qquad (5.5)$$

**Proof:**
For $t = 0$, $\hat{C}_{\nu,0} = a_0[\bar{f}(x_0) - \bar{f}(x_*)]$. Therefore, inequality (5.5) is valid.

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

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

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

$$\forall \; y = (1 - \tau_t)x_t + \tau_t x, \quad x \in \operatorname{dom} \Psi.$$

Therefore, in view of definition (2.10), for any $x \in \operatorname{dom} \Psi$ we have

$$
\begin{aligned}
a_{t+1}\langle \nabla f(x_t), x - x_t \rangle \;\geq\;\; & A_{t+1}\langle \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle \\[4pt]
& + a_{t+1}\langle \nabla^2 f(x_t)(x_{t+1} - x_t), x_t - x \rangle \\[4pt]
& + A_{t+1}[\Psi(x_{t+1}) - \Psi((1 - \tau_t)x_t + \tau_t x)] \\[4pt]
\overset{(5.3)}{\geq}\;\; & A_{t+1}\langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle \\[4pt]
& + A_{t+1}[\Psi(x_{t+1}) - \Psi((1 - \tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}}LD^2.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)] \\[6pt]
\geq\;\; & A_t \Psi(x_t) + A_{t+1}[f(x_t) + \langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t \rangle] \\[6pt]
& + a_{t+1}\Psi(x) + A_{t+1}[\Psi(x_{t+1}) - \Psi((1 - \tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}}LD^2 \\[6pt]
\overset{(2.6)}{\geq}\;\; & A_{t+1}[f(x_{t+1}) + \Psi(x_{t+1})] - A_{t+1}\tfrac{H_\nu \|x_{t+1} - x_t\|^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}}LD^2 \\[6pt]
\geq\;\; & A_{t+1}\bar{f}(x_{t+1}) - \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}}LD^2.
\end{aligned}
$$

Thus, if (5.5) is valid for some $t \geq 0$, then

$$\phi_{t+1}(x) + \hat{C}_{\nu,t} \geq A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x)]$$

$$\geq A_{t+1} \bar{f}(x_{t+1}) - \frac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \frac{a_{t+1}^2}{2A_{t+1}} LD^2.$$

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

Thus, inequality (5.5) ensures the following rate of convergence of method (5.1)

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{1}{A_t} \hat{C}_{\nu,t}. \tag{5.6}$$

The particular expression of the right-hand side of this inequality can be obtained exactly in the same way as 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

$$\sum_{k=1}^{t} \frac{a_k^3}{A_k^2} = \sum_{k=1}^{t} \frac{36k^6}{k^2(k+1)^2(2k+1)^2} \leq 18t,$$

$$\sum_{k=1}^{t} \frac{a_k^2}{2A_k} = \sum_{k=1}^{t} \frac{3k^4}{k(k+1)(2k+1)} \leq \frac{3}{2} \sum_{k=1}^{t} k = \frac{3}{4}t(t+1).$$

Thus, we get

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{18H_1 D^3}{(t+1)(t+2)} + \frac{9LD^2}{2(t+2)}. \tag{5.7}$$

Note that this rate of convergence is worse than that of the Newton method with cubic regularization [15]. However, to the best of our knowledge, inequality (5.7) gives us the first global rate of convergence obtained so far for an optimization scheme belonging to the family of trust-region methods [1]. In view of inequality (5.5), the optimal solution of the dual problem (3.4) can be approximated by method (5.1) with $a_0 = 0$ in the same way as it was done in Section 3 for Conditional Gradient Methods.

# 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 $30^{th}$ International Conderence on Machine Learning*, Atlanta, Georgia, (2013).

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

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

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

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

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

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

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

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

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

# Recent titles

## CORE Discussion Papers

# Recent titles

## CORE Discussion Papers - continued

2014/59 Manuela BRAIONE and Nicolas K. SCHOLTES. Construction of Value-at-Risk forecasts under different distributional assumptions within a BEKK framework.

2014/60 Jörg BREITUNG and Christian M. HAFNER. A simple model for now-casting volatility series.

2014/61 Timo TERASVIRTA and Yukai YANG. Linearity and misspecification tests for vector smooth transition regression models.

2014/62 Timo TERASVIRTA and Yukai YANG. Specification, estimation and evaluation of vector smooth transition autoregressive models with applications.

2014/63 Axel GAUTIER and Nicolas PETIT. Optimal enforcement of competition policy: the commitments procedure under uncertainty.

2014/64 Sébastien BROOS and Axel GAUTIER. Competing one-way essential complements: the forgotten side of net neutrality.

2014/65 Jean HINDRIKS and Yukihiro NISHIMURA. On the timing of tax and investment in fiscal competition models.

2014/66 Jean HINDRIKS and Guillaume LAMY. Back to school, back to segregation?

2014/67 François MANIQUET et Dirk NEUMANN. Echelles d'équivalence du temps de travail: évaluation de l'impôt sur le revenu en Belgique à la lumière de l'éthique de la responsabilité.

2015/1 Yurii NESTEROV and Vladimir SHIKHMAN. Algorithm of Price Adjustment for Market Equilibrium.

2015/2 Claude d'ASPREMONT and Rodolphe DOS SANTOS FERREIRA. Oligopolistic vs. monopolistic competition: Do intersectoral effects matter?

2015/3 Yuuri NESTEROV. Complexity bounds for primal-dual methods minimizing the model of objective function.

## Books

W. GAERTNER and E. SCHOKKAERT (2012), *Empirical Social Choice*. Cambridge University Press.

L. BAUWENS, Ch. HAFNER and S. LAURENT (2012), *Handbook of Volatility Models and their Applications*. Wiley.

J-C. PRAGER and J. THISSE (2012), *Economic Geography and the Unequal Development of Regions*. Routledge.

M. FLEURBAEY and F. MANIQUET (2012), *Equality of Opportunity: The Economics of Responsibility*. World Scientific.

J. HINDRIKS (2012), *Gestion publique*. De Boeck.

M. FUJITA and J.F. THISSE (2013), *Economics of Agglomeration: Cities, Industrial Location, and Globalization.* (2nd edition). Cambridge University Press.

J. HINDRIKS and G.D. MYLES (2013). *Intermediate Public Economics.* (2nd edition). MIT Press.

J. HINDRIKS, G.D. MYLES and N. HASHIMZADE (2013). *Solutions Manual to Accompany Intermediate Public Economics.* (2nd edition). MIT Press.

J. HINDRIKS (2015). *Quel avenir pour nos pensions ? Les grands défis de la réforme des pensions.* De Boeck.

## CORE Lecture Series

R. AMIR (2002), Supermodularity and Complementarity in Economics.

R. WEISMANTEL (2006), Lectures on Mixed Nonlinear Programming.

A. SHAPIRO (2010), Stochastic Programming: Modeling and Theory.