

OPTIMIZATION METHODS FOR FULLY COMPOSITE PROBLEMS

Nikita Doikov, Yurii Nesterov

LIDAM Discussion Paper CORE
2021 / 01

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/core/core-discussion-papers.html>

CORE DISCUSSION PAPER

2021/01

Optimization Methods for Fully Composite Problems ^{*}

Nikita Doikov [†] Yurii Nesterov [‡]

March 20, 2021

Abstract

In this paper, we propose a new Fully Composite Formulation of convex optimization problems. It includes, as a particular case, the problems with functional constraints, max-type minimization problems, and problems of Composite Minimization, where the objective can have simple nondifferential components. We treat all these formulations in a unified way, highlighting the existence of very natural optimization schemes of different order. We prove the global convergence rates for our methods under the most general conditions. Assuming that the upper-level component of our objective function is subhomogeneous, we develop efficient modification of the basic Fully Composite first-order and second-order Methods, and propose their accelerated variants.

Keywords: Convex Optimization, Constrained Optimization, Nonsmooth Optimization, Gradient Methods, High-order Methods, Accelerated Algorithms.

^{*}Research results presented in this paper were obtained in the framework of ERC Advanced Grant 788368.

[†]Institute of Information and Communication Technologies, Electronics and Applied Math. (ICTEAM), Catholic University of Louvain (UCL). E-mail: Nikita.Doikov@uclouvain.be. ORCID: 0000-0003-1141-1625.

[‡]Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL). E-mail: Yurii.Nesterov@uclouvain.be. ORCID: 0000-0002-0542-8757.

1 Introduction

Motivation. Development of the numerical methods for solving different optimization problems heavily depends on the *model of the problem* used by the method’s designer. In modern Optimization Theory, the diversity of problem formulations is sufficiently big. We can speak about problems with functional constraints, or with simple feasible set. The problems can be posed with differentiable or non-differentiable components. Sometimes we speak about problems in (additive) composite form (e.g. [22]). Or, we can speak about optimization of max-type functions (e.g. Section 2.3 in [24]).

All these formulations have quite specific properties and usually they need development of the specific methods. In this paper, we are going to take a step back in this picture and consider a very general problem formulation which covers practically all variants of the existing problem settings. The main advantage of our formulation (we call it *Fully Composite Optimization Problem*) is that for justification of the corresponding numerical schemes we can use only very basic properties of our objects (convexity, monotonicity). Thus, we highlight the generic reasons for existence of the efficient methods for many different problem classes.

As an immediate consequence of our results, we get, in particular, new high-order methods with global linear rate of convergence for convex minimization with functional constraints. Our new first-, second-, and third-order methods can be implemented in practice using the existing polynomial-time technique [25].

Contents. In Section 2, we study uniformly convex smooth functions. We prove two new inequalities based on high-order Taylor polynomials, which provide these functions with the improved global lower bounds. This gives us the main tool for justifying the global convergence rates of our methods.

In Section 3, we present our Fully Composite Optimization Framework, and give several examples, which cover all popular composite settings. Then, in Section 4, we develop basic high-order optimization methods (starting from the first-order methods) for solving Fully Composite Problems. Assuming that the smooth component of our problem is uniformly convex of a certain degree, we establish a global linear rate of convergence of the new methods.

In Section 5, we demonstrate that it is possible to use a simple regularization technique within our framework. It converts any convex problem into uniformly convex one, and thus our basic methods can be applied to solve them.

Section 6 is devoted to *subhomogeneous* functions. We provide the definition and list several properties of such functions. Then we show that for subhomogeneous fully composite formulations, the global convergence of the basic methods holds in a more general convex setting.

We study efficient modifications of the first-order and second-order methods for the subhomogeneous fully composite problems in Sections 7 and 8 respectively. In particular, we establish the accelerated $O(k^{-2})$ global rate of convergence for the Fast Gradient Method [19], and the same global rate for the modifications of the Newton’s Method [27, 7].

In Section 9, we accelerate our fully composite second-order methods up to the level $O(k^{-3})$ using inexact contracting proximal iterations [6].

Notation. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and

by $\mathbb{E}^*$ its dual space composed by linear functions on $\mathbb{E}$. For such a function $s \in \mathbb{E}^*$, we denote by $\langle s, x \rangle$ its value at $x \in \mathbb{E}$. Using a self-adjoint positive-definite operator $B : \mathbb{E} \rightarrow \mathbb{E}^*$ (notation $B = B^* \succ 0$), we define the *conjugate Euclidean norms*:

$$\|x\| = \langle Bx, x \rangle^{1/2}, \quad x \in \mathbb{E}, \quad \|g\|_* = \langle g, B^{-1}g \rangle^{1/2}, \quad g \in \mathbb{E}^*.$$

For a smooth function $f : \text{dom } f \rightarrow \mathbb{R}$ with convex and open domain $\text{dom } f \subseteq \mathbb{E}$, denote by $\nabla f(x)$ its gradient, and by $\nabla^2 f(x)$ its Hessian evaluated at point $x \in \text{dom } f \subseteq \mathbb{E}$. Then

$$\nabla f(x) \in \mathbb{E}^*, \quad \nabla^2 f(x)h \in \mathbb{E}^*, \quad x \in \text{dom } f, \quad h \in \mathbb{E}.$$

In what follows, we often work with directional derivatives. For $p \geq 1$, denote by

$$D^p f(x)[h_1, \dots, h_p]$$

the directional derivative of function f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$. Note that $D^p f(x)[\cdot]$ is a *symmetric p -linear form*. Its *norm* is defined in the standard way:

$$\|D^p f(x)\| = \max_{h_1, \dots, h_p} \{D^p f(x)[h_1, \dots, h_p] : \|h_i\| \leq 1, i = 1, \dots, p\}. \quad (1.1)$$

For example, for any $x \in \text{dom } f$ and $h_1, h_2 \in \mathbb{E}$, we have

$$Df(x)[h_1] = \langle \nabla f(x), h_1 \rangle, \quad D^2 f(x)[h_1, h_2] = \langle \nabla^2 f(x)h_1, h_2 \rangle.$$

Thus, for the Hessian, our definition corresponds to the *spectral norm* of self-adjoint linear operator (maximal module of all eigenvalues computed with respect to operator B).

If all directions $h_1, \dots, h_p$ are the same, we apply the notation $D^p f(x)[h]^p$, $h \in \mathbb{E}$. Then, Taylor approximation of function $f(\cdot)$ at $x \in \text{dom } f$ can be written as follows:

$$\begin{aligned} f(y) &= \Omega_p(f, x; y) + o(\|y - x\|^p), \quad y \in \text{dom } f, \\ \Omega_p(f, x; y) &\stackrel{\text{def}}{=} f(x) + \sum_{k=1}^p \frac{1}{k!} D^k f(x)[y - x]^k, \quad y \in \mathbb{E}. \end{aligned} \quad (1.2)$$

Note that in general, we have (see, for example, Appendix 1 in [26])

$$\|D^p f(x)\| = \max_h \left\{ \left| D^p f(x)[h]^p \right| : \|h\| \leq 1 \right\}. \quad (1.3)$$

Similarly, since for $x, y \in \text{dom } f$ being fixed, the form $D^p f(x)[\cdot, \dots, \cdot] - D^p f(y)[\cdot, \dots, \cdot]$ is p -linear and symmetric, we also have

$$\|D^p f(x) - D^p f(y)\| = \max_h \left\{ \left| D^p f(x)[h]^p - D^p f(y)[h]^p \right| : \|h\| \leq 1 \right\}. \quad (1.4)$$

In this paper, we consider functions from the problem classes $\mathcal{F}_p$, which are convex and p times continuously differentiable on $\mathbb{E}$. Denote by L_p the uniform bound for the Lipschitz constant of p th derivative:

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|, \quad x, y \in \text{dom } f, \quad p \geq 1. \quad (1.5)$$

Sometimes, if an ambiguity could arise, we use notation $L_p(f)$.

Assuming that $f \in \mathcal{F}_p$ and $L_p < +\infty$, by the standard integration arguments we can bound the residual between function value and its Taylor approximation:

$$|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad x, y \in \text{dom } f. \quad (1.6)$$

2 Uniform Convexity of Smooth Functions

Let us couple our smoothness assumption with *uniform convexity* of certain degree. Namely, let us assume that for $p \geq 1$ there exists a constant $\sigma_{p+1}(f) > 0$ such that

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \sigma_{p+1}(f) \|y - x\|^{p+1}, \quad x, y \in \text{dom } f. \quad (2.1)$$

By simple integration, this inequality ensures the following functional growth:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_{p+1}(f)}{p+1} \|y - x\|^{p+1}, \quad x, y \in \text{dom } f. \quad (2.2)$$

Let us consider the uniformly convex functions of degree $p + 1$, whose p th derivative is Lipschitz continuous. We introduce the following constant:

$$\gamma_p(f) \stackrel{\text{def}}{=} \frac{\sigma_{p+1}(f)}{L_p(f)},$$

called the *condition number of degree p* of function f . Combining (1.6) and (2.2), we get

$$\frac{\sigma_{p+1}(f)}{p+1} \|y - x\|^{p+1} \leq \sum_{k=2}^p \frac{1}{k!} D^k f(x) [y - x]^k + \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad x, y \in \text{dom } f.$$

Thus, in the case of unbounded domain, we have

$$\gamma_p(f) \leq \frac{1}{p!}. \quad (2.3)$$

Let us prove now the main inequalities of our problem class. For $\alpha \geq 0$, denote

$$\beta_p(f, \alpha) = \frac{(p! \gamma_p(f))^{\frac{1}{p}}}{(1+\alpha)^{\frac{1}{p}} + (p! \gamma_p(f))^{\frac{1}{p}}}. \quad (2.4)$$

Theorem 1 *For any $p \geq 1$, $\alpha \geq 0$, and all $x, y \in \text{dom } f$, we have*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \sum_{k=2}^p \frac{\beta^{k-1}}{(k-1)!} D^k f(x) [y - x]^k + \frac{\alpha L_p \beta^p}{p!} \|y - x\|^{p+1}, \quad (2.5)$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \sum_{k=2}^p \frac{\beta^{k-1}}{k!} D^k f(x) [y - x]^k + \frac{\alpha L_p \beta^p}{(p+1)!} \|y - x\|^{p+1}, \quad (2.6)$$

where $0 \leq \beta \leq \beta_p(f, \alpha)$.

Proof:

Let us fix some $x, y \in \text{dom } f$ and consider $z_t \stackrel{\text{def}}{=} x + t(y - x)$, $0 \leq t \leq 1$. Then,

$$\begin{aligned} \langle \nabla f(y), y - x \rangle &= \frac{1}{1-t} \langle \nabla f(y), y - z_t \rangle \\ &\stackrel{(2.1)}{\geq} \frac{1}{1-t} \langle \nabla f(z_t), y - z_t \rangle + (1-t)^p \sigma_{p+1}(f) \|y - x\|^{p+1} \\ &= \langle \nabla f(z_t), y - x \rangle + (1-t)^p \sigma_{p+1}(f) \|y - x\|^{p+1}. \end{aligned}$$

Consider now the function $\phi(t) = \langle \nabla f(z_t), y - x \rangle$. Then, by Taylor's formula, we have

$$\begin{aligned}
\phi(t) &= \phi(0) + \sum_{k=1}^{p-1} \frac{t^k}{k!} \phi^{(k)}(0) + \frac{1}{(p-1)!} \int_0^t (t-\lambda)^{p-1} \phi^{(p)}(\lambda) d\lambda \\
&= \sum_{k=1}^p \frac{t^{k-1}}{(k-1)!} D^k f(x) [y-x]^k + \frac{1}{(p-1)!} \int_0^t (t-\lambda)^{p-1} D^{p+1} f(z_\lambda) [y-x]^{p+1} d\lambda \\
&\geq \sum_{k=1}^p \frac{t^{k-1}}{(k-1)!} D^k f(x) [y-x]^k - \frac{t^p}{p!} L_p \|y-x\|^{p+1}.
\end{aligned}$$

Adding these two inequalities, we get

$$\begin{aligned}
\langle \nabla f(y) - \nabla f(x), y - x \rangle &\geq \sum_{k=2}^p \frac{t^{k-1}}{(k-1)!} D^k f(x) [y-x]^k \\
&\quad + \left((1-t)^p \sigma_{p+1}(f) - \frac{t^p}{p!} L_p \right) \|y-x\|^{p+1}.
\end{aligned}$$

Let us choose now t from the inequality

$$(1-t)^p \sigma_{p+1}(f) - \frac{1}{p!} t^p L_p \geq \frac{\alpha}{p!} t^p L_p \Leftrightarrow \frac{p!}{1+\alpha} \gamma_p(f) \geq \left(\frac{t}{1-t} \right)^p.$$

Then it is enough to take $t \stackrel{(2.4)}{\leq} \beta_p(f, \alpha)$. Hence, inequality (2.5) is proved.

The remaining inequality (2.6) can be proved by integration. Indeed

$$\begin{aligned}
&f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\
&= \int_0^1 \frac{1}{\tau} \langle \nabla f(x + \tau(y-x)) - f(x), \tau(y-x) \rangle d\tau \\
&\stackrel{(2.5)}{\geq} \int_0^1 \left(\sum_{k=2}^p \frac{\beta^{k-1} \tau^{k-1}}{(k-1)!} D^k f(x) [y-x]^k + \frac{\alpha L_p(f) \beta^p \tau^p}{p!} \|y-x\|^{p+1} \right) d\tau \\
&= \sum_{k=2}^p \frac{\beta^{k-1}}{k!} D^k f(x) [y-x]^k + \frac{\alpha L_p(f) \beta^p}{(p+1)!} \|y-x\|^{p+1}. \quad \square
\end{aligned}$$

Remark 1 For $\alpha \geq p$, the right-hand side of inequality (2.6) is convex in y . Indeed, let us introduce new variables $z = x + \beta(y-x)$. Then this right-hand side is transformed to the following function:

$$f(x) + \frac{1}{\beta} \left[\langle \nabla f(x), z - x \rangle + \sum_{k=2}^p \frac{1}{k!} D^k f(x) [z-x]^k + \frac{\alpha L_p}{(p+1)!} \|z-x\|^{p+1} \right].$$

Since $\alpha \geq p$, it is convex in z in view of Theorem 1 in [25].

For the optimization schemes developed in this paper, inequality (2.6) serves as the main justification tool. Let us present now the general model of our optimization problems.

3 Fully Composite Optimization Problem

Let $F(\cdot, \cdot)$ be a function from $\mathbb{E} \times \mathbb{R}^m$ to $\mathbb{R} \cup \{+\infty\}$. Hence, $\text{dom } F = \{(x, u) \in \mathbb{E} \times \mathbb{R}^m : F(x, u) < +\infty\}$. For each $x \in \mathbb{E}$, denote

$$\mathcal{D}(x) = \{u \in \mathbb{R}^m : (x, u) \in \text{dom } F\}.$$

Our assumptions on function F are as follows.

Assumption 1 *Function $F(\cdot, \cdot)$ is closed and convex on its domain. Moreover, for any $x \in \mathbb{E}$ with $\mathcal{D}(x) \neq \emptyset$, function $F(x, u)$ is closed, convex and monotone in $u \in \mathcal{D}(x)$.*

Consider now a vector function $f(x) = (f_1(x), \dots, f_m(x))^T : \text{dom } f \rightarrow \mathbb{R}^m$.

Assumption 2 *All components of function f are closed and convex.*

In our framework, all information about function f can be collected by the calls of oracle of certain degree.

Let us call *fully composite* the following optimization problem:

$$\boxed{\varphi^* = \min_{x \in \text{dom } \varphi} \left\{ \varphi(x) \stackrel{\text{def}}{=} F(x, f(x)) \right\}}, \quad (3.1)$$

where $\text{dom } \varphi = \{x \in \text{dom } f : (x, f(x)) \in \text{dom } F\}$. We denote by x^* a solution to problem (3.1): $\varphi(x^*) = \varphi^*$, assuming that it exists.

Of course, problem (3.1) is tractable only if the function F is simple. This is our third assumption.

Assumption 3 *Structure of function F is simple enough for allowing an efficient solution of some auxiliary optimization problems based on approximations of function f .*

We will see soon what kind of auxiliary problems with function F we need to solve. At this moment, let us give several examples of fully composite optimization problems.

1. *Optimization with functional constraints.* Consider the following problem:

$$\min_{x \in Q} \{f_1(x) : f_i(x) \leq 0, i = 2, \dots, m\}, \quad (3.2)$$

where Q is a closed convex set and f satisfies Assumption 2. Then this problem can be written in form (3.1) with

$$F(x, u) = u^{(1)} + \sum_{i=2}^m \text{Ind}_{\mathbb{R}_-}(u^{(i)}) + \text{Ind}_Q(x),$$

where $\text{Ind}_X(\cdot) : \mathbb{E} \rightarrow \{0, +\infty\}$ is the indicator function of the set $X \subseteq \mathbb{E}$.

2. *Additive composite minimization* [22]. Consider the following minimization problem:

$$\min_x \{f_1(x) + \psi(x)\}, \quad (3.3)$$

where $f(x) \equiv \{f_1(x)\}$ satisfies Assumption 2 and $\psi(\cdot)$ is a simple closed convex function. Then we can take

$$F(x, u) = u^{(1)} + \psi(x).$$

3. *Functional composite minimization* (e.g. [19, 20]). Minimization problem

$$\min_x F(f(x)), \quad (3.4)$$

where F is a closed convex monotone function with $\text{dom } F \subseteq \mathbb{R}_+^m$, and f satisfies Assumption 2, is clearly in the form (3.1).

Note that the Taylor polynomial (1.2) is defined in terms of directional derivatives. Therefore we can extend its meaning onto the vector functions without changing notation. Similarly, we will use the following constant vectors

$$\begin{aligned} L_p(f) &= (L_p(f_1), \dots, L_p(f_m))^T, \\ \sigma_{p+1}(f) &= (\sigma_{p+1}(f_1), \dots, \sigma_{p+1}(f_m))^T, \\ \gamma_p(f) &= (\gamma_p(f_1), \dots, \gamma_p(f_m))^T. \end{aligned}$$

Denote $\hat{\beta}_p(f) = \min_{1 \leq i \leq m} \beta_p(f_i, p)$. Then inequality (2.6) can be rewritten in a vector form:

$$f(y) \geq f(x) + \sum_{k=1}^p \frac{\beta^{k-1}}{k!} D^k f(x) [y - x]^k + \frac{p L_p(f) \beta^p}{(p+1)!} \|y - x\|^{p+1}, \quad (3.5)$$

for all $\beta \in [0, \hat{\beta}_p(f))$. The right-hand side of this inequality provides us with the auxiliary problem we need to solve at each iteration of our schemes:

$$F\left(y, f(x) + \sum_{k=1}^p \frac{\beta^{k-1}}{k!} D^k f(x) [y - x]^k + \frac{p L_p(f) \beta^p}{(p+1)!} \|y - x\|^{p+1}\right) \rightarrow \min_y, \quad (3.6)$$

where $\beta = \hat{\beta}_p(f)$. We explain the sense of this operation in the next section.

4 Basic High-order Optimization Methods

Let $\bar{x} \in \text{dom } \varphi$. For inequality (2.6), let us choose $\alpha = p$ and $\beta \in (0, \hat{\beta}_p(f)]$. Then

$$\begin{aligned} (1 - \beta)\varphi(\bar{x}) + \beta\varphi^* &= \min_{v \in \text{dom } \varphi} \left[(1 - \beta)F(\bar{x}, f(\bar{x})) + \beta F(v, f(v)) \right] \\ &\geq \min_{v \in \text{dom } \varphi} F\left((1 - \beta)\bar{x} + \beta v, (1 - \beta)f(\bar{x}) + \beta f(v)\right) \\ &\stackrel{(3.5)}{\geq} \min_{v \in \text{dom } \varphi} F\left((1 - \beta)\bar{x} + \beta v, f(\bar{x}) + \sum_{k=1}^p \frac{\beta^k}{k!} D^k f(\bar{x}) [v - \bar{x}]^k + \frac{p L_p(f) \beta^{p+1}}{(p+1)!} \|v - \bar{x}\|^{p+1}\right) \\ &= \min_{\substack{y = \bar{x} + \beta(v - \bar{x}) \\ v \in \text{dom } \varphi}} F\left(y, \Omega_p(f, \bar{x}; y) + \frac{p L_p(f)}{(p+1)!} \|y - \bar{x}\|^{p+1}\right) \stackrel{\text{def}}{=} \tilde{M}_{p, \beta}^*(\bar{x}). \end{aligned}$$

By Remark 1, the second argument in the objective function of the latter problem is a component-wise convex function. Hence, by Assumption 1, this objective is convex in y .

Let us look at the solution of the above minimization problem, that is

$$\tilde{y}_{p,\beta}^*(\bar{x}) \stackrel{\text{def}}{=} \underset{y}{\operatorname{argmin}} \left\{ F\left(y, \Omega_p(f, \bar{x}; y) + \frac{pL_p(f)}{(p+1)!} \|y - \bar{x}\|^{p+1}\right) : \bar{x} + \frac{1}{\beta}(y - \bar{x}) \in \operatorname{dom} \varphi \right\}.$$

Note that in view of Assumption 1, we have

$$\begin{aligned} \tilde{M}_{p,\beta}^*(\bar{x}) &= F\left(\tilde{y}_{p,\beta}^*(\bar{x}), \Omega_p(f, \bar{x}; \tilde{y}_{p,\beta}^*(\bar{x})) + \frac{pL_p(f)}{(p+1)!} \|\tilde{y}_{p,\beta}^*(\bar{x}) - \bar{x}\|^{p+1}\right) \\ &\stackrel{(1.6)}{\geq} F\left(\tilde{y}_{p,\beta}^*(\bar{x}), f(\tilde{y}_{p,\beta}^*(\bar{x}))\right) = \varphi\left(\tilde{y}_{p,\beta}^*(\bar{x})\right). \end{aligned}$$

Thus, we can estimate now the rate of convergence of the following method.

Restricted pth order Basic Method	
Choose $x_0 \in \operatorname{dom} \varphi$ and $\beta \in (0, \hat{\beta}_p(f)]$. For $k \geq 0$ iterate: $x_{k+1} = \tilde{y}_{p,\beta}^*(x_k)$.	(4.1)

We have proved for this method the following theorem.

Theorem 2 *Let sequence $\{x_k\}_{k \geq 0}$ be generated by the method (4.1). Then for all $k \geq 0$ we have*

$$\varphi(x_k) - \varphi^* \leq (1 - \beta)^k (\varphi(x_0) - \varphi^*). \quad (4.2)$$

Therefore, the rate of convergence is linear, and the contraction parameter β can reach the condition number.

Note that method (4.1) could move with bigger steps. Indeed,

$$\tilde{M}_{p,\beta}^*(\bar{x}) \geq \min_{y \in \operatorname{dom} \varphi} F\left(y, \Omega_p(f, \bar{x}; y) + \frac{pL_p(f)}{(p+1)!} \|y - \bar{x}\|^{p+1}\right) \stackrel{\text{def}}{=} M_p^*(\bar{x}).$$

Denote

$$y_p^*(\bar{x}) \stackrel{\text{def}}{=} \underset{y \in \operatorname{dom} \varphi}{\operatorname{argmin}} F\left(y, \Omega_p(f, \bar{x}; y) + \frac{pL_p(f)}{(p+1)!} \|y - \bar{x}\|^{p+1}\right).$$

By the same reasons as before, $M_p^*(\bar{x}) \geq \varphi(y_p^*(\bar{x}))$. Hence, we can estimate the rate of convergence of the following method.

Full-step pth order Basic Method	
Choose $x_0 \in \operatorname{dom} \varphi$. For $k \geq 0$ iterate: $x_{k+1} = y_p^*(x_k)$.	(4.3)

Theorem 3 Let sequence $\{x_k\}_{k \geq 0}$ be generated by (4.3). Then for all $k \geq 0$ we have

$$\varphi(x_k) - \varphi^* \leq (1 - \hat{\beta}_p(f))^k (\varphi(x_0) - \varphi^*) \quad (4.4)$$

In view of potentially bigger steps, method (4.3) is often faster in practice.

Example 1 Let us look at implementation of method (4.3) for a particular class of optimization problems with functional constraints (3.2) with $Q = \mathbb{E}$. This is

$$\min_{x \in \mathbb{E}} \{f_1(x) : f_i(x) \leq 0, i = 2, \dots, m\}.$$

Then, for $p = 1$, each iteration of the method (4.3) can be represented as follows:

$$y_1^*(\bar{x}) = \bar{x} - \left(\sum_{i=1}^m \lambda_*^{(i)} L_1(f_i) B \right)^{-1} g(\lambda_*),$$

where $g(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^m \lambda^{(i)} \nabla f_i(\bar{x})$, and $\lambda_* \in \mathbb{R}_+^m$ is a solution to the corresponding dual problem

$$\max_{\substack{\lambda \in \mathbb{R}_+^m \\ \lambda^{(1)}=1}} \left\{ \sum_{i=1}^m \lambda^{(i)} f_i(\bar{x}) - \frac{1}{2 \sum_{i=1}^m \lambda^{(i)} L_1(f_i)} \|g(\lambda)\|_*^2 \right\}. \quad (4.5)$$

On the other hand, for $p = 2$, one iteration of the method (4.3) is as follows:

$$y_2^*(\bar{x}) = \bar{x} - H(\lambda_*, \tau_*)^{-1} g(\lambda_*),$$

with operator $H(\lambda, \tau) \stackrel{\text{def}}{=} \sum_{i=1}^m \lambda^{(i)} \nabla^2 f_i(\bar{x}) + \tau B$. The optimal $\lambda_* \in \mathbb{R}_+^m$ and $\tau_* \in \mathbb{R}_+$ can be computed from the following concave optimization problem

$$\max_{\substack{\lambda \in \mathbb{R}_+^m, \tau \in \mathbb{R}_+ \\ \lambda^{(1)}=1}} \left\{ \sum_{i=1}^m \lambda^{(i)} f_i(\bar{x}) - \frac{3\tau^3}{2[\sum_{i=1}^m \lambda^{(i)} L_2(f_i)]^2} - \frac{1}{2} \langle H(\lambda, \tau)^{-1} g(\lambda), g(\lambda) \rangle \right\}. \quad (4.6)$$

Note that typically the dimension of the problems (4.5) and (4.6) is not big. Hence, they can be solved, for example, by the Interior-Point Methods [26] very efficiently.

5 General Regularization Scheme

In the previous sections, we discussed two methods for solving problem (3.1) under assumption of uniform convexity of functional components: $\hat{\beta}_p(f) > 0$. If this assumption is not valid, we still can apply methods of Section 4 to a special *regularized* problem.

Let us present a general regularization framework for fully composite problem (3.1). For that, we use a component-wise convex regularizing vector function

$$d(x) = (d_1(x), \dots, d_m(x))^T : \text{dom } f \rightarrow \mathbb{R}^m.$$

It is related to the starting point of our process $x_0 \in \text{dom } \varphi$ in the following way:

$$d(x_0) = f(x_0), \quad (5.1)$$

$$d(x) \geq f(x), \quad \forall x \in \text{dom } f. \quad (5.2)$$

We discuss a simple possibility for choosing such a regularizer in the end of this section.

Let $\mu \geq 0$ be a regularizing parameter. Define the following regularized function:

$$\varphi_\mu(x) = F(x, (1 - \mu)f(x) + \mu d(x)), \quad x \in \text{dom } \varphi. \quad (5.3)$$

It is convenient to assume that the function $d(\cdot)$ satisfies the following assumption.

Assumption 4 *Vector function $d(x) - f(x)$ is component-wise convex on $\text{dom } f$.*

In this case, the regularized function $\varphi_\mu(\cdot)$ is convex for any $\mu \geq 0$.

Note that $\varphi_\mu(x_0) \stackrel{(5.1)}{=} \varphi(x_0)$. Clearly, for all $x \in \text{dom } \varphi$ we have

$$\varphi(x) \stackrel{(5.2)}{\leq} \varphi_\mu(x). \quad (5.4)$$

Our regularized problem looks now as follows:

$$\boxed{\varphi_\mu^* = \min_{x \in \text{dom } \varphi} \varphi_\mu(x).} \quad (5.5)$$

At this moment, let us assume that we are able to generate an approximate solution to this perturbed problem by one of the methods of Section 4. Namely, assume that, for certain $\delta > 0$, we have a point $\bar{x} \in \text{dom } \varphi$, satisfying the following inequality:

$$\begin{aligned} \varphi_\mu(\bar{x}) - \varphi_\mu^* &\leq \delta (\varphi_\mu(x_0) - \varphi_\mu^*) \stackrel{(5.1)}{=} \delta (\varphi(x_0) - \varphi_\mu^*) \\ &\stackrel{(5.4)}{\leq} \delta (\varphi(x_0) - \varphi^*). \end{aligned} \quad (5.6)$$

We need to understand now how good is this point for our initial problem (3.1).

In order to answer this question, we need to introduce a *local measure* for non-negative vectors $g \in \mathbb{R}_+^m$ with respect to some point $x \in \text{dom } \varphi$ and a functional level A :

$$\xi_A(x; g) = \min_{\lambda > 0} \left\{ \lambda : f(x) + \frac{1}{\lambda} g \in \mathcal{D}(x), F\left(x, f(x) + \frac{1}{\lambda} g\right) \leq A \right\}. \quad (5.7)$$

Clearly, this measure is well defined at least at all points $x \in \text{dom } \varphi$ with $\varphi(x) < A$.

Lemma 1 *Let for some points x_0 and $\bar{x}$ from $\text{dom } \varphi$ we have $\varphi(x_0) \leq A$ and $\varphi(\bar{x}) \leq A$. Then, for any $g \in \mathbb{R}_+^m$ and any coefficient $\tau \in [0, 1]$ we have*

$$\xi_A((1 - \tau)x_0 + \tau\bar{x}; g) \leq \frac{1}{1 - \tau} \xi_A(x_0; g). \quad (5.8)$$

Proof:

Let $f(x_0) + \frac{1}{\lambda} g \in \mathcal{D}(x_0)$ for some $\lambda > 0$, and $F(x_0, f(x_0) + \frac{1}{\lambda} g) \leq A$. Since $f(\bar{x}) \in \mathcal{D}(\bar{x})$, we have

$$(1 - \tau)(f(x_0) + \frac{1}{\lambda} g) + \tau f(\bar{x}) \in \mathcal{D}((1 - \tau)x_0 + \tau\bar{x}).$$

By convexity, $(1 - \tau)f(x_0) + \tau f(\bar{x}) \geq f((1 - \tau)x_0 + \tau\bar{x})$, and we conclude that

$$f((1 - \tau)x_0 + \tau\bar{x}) + \frac{1 - \tau}{\lambda} g \in \mathcal{D}((1 - \tau)x_0 + \tau\bar{x}).$$

At the same time,

$$\begin{aligned}
& F((1-\tau)x_0 + \tau\bar{x}, f((1-\tau)x_0 + \tau\bar{x})) + \frac{1-\tau}{\lambda}g) \\
& \leq F((1-\tau)x_0 + \tau\bar{x}, (1-\tau)(f(x_0) + \frac{1}{\lambda}g) + \tau f(\bar{x})) \\
& \leq (1-\tau)F(x_0, f(x_0) + \frac{1}{\lambda}g) + \tau F(\bar{x}, f(\bar{x})) \leq A.
\end{aligned}$$

Thus, $\xi_A((1-\tau)x_0 + \tau\bar{x}; g) \leq \frac{\lambda}{1-\tau}$. $\square$

Denote $g_* = d(x^*) - f(x^*)$ and $\xi_0^* = \xi_A(x_0; g_*)$.

Lemma 2 *Let $A \geq \varphi(x_0)$ and the regularizing function $d(\cdot)$ satisfy Assumption 4 and conditions (5.1), (5.2). Assume that parameters α and τ from $[0, 1]$ are chosen as follows:*

$$\frac{\tau}{1-\tau} \leq \frac{\alpha}{\mu \xi_0^*}. \quad (5.9)$$

Then

$$\varphi_\mu^* - \varphi^* \leq (1-\tau)(1-\alpha)(\varphi(x_0) - \varphi^*) + \alpha(A - \varphi^*). \quad (5.10)$$

Proof:

Let us fix a point $x \in \text{dom } \varphi$ with $\varphi(x) \leq A$. Denote by $g_x = d(x) - f(x) \geq 0$. Then, for any $\alpha \in [0, 1]$ we have

$$\begin{aligned}
\varphi_\mu^* \leq \varphi_\mu(x) &= F(x, f(x) + \mu g_x) = F(x, (1-\alpha)f(x) + \alpha(f(x) + \frac{\mu}{\alpha}g_x)) \\
&\leq (1-\alpha)\varphi(x) + \alpha F(x, f(x) + \frac{\mu}{\alpha}g_x).
\end{aligned}$$

Let us choose now $x = (1-\tau)x_0 + \tau x^*$ with arbitrary $\tau \in [0, 1]$. Then in view of Assumption 4, we have

$$\begin{aligned}
g_x &\leq (1-\tau)(d(x_0) - f(x_0)) + \tau(d(x^*) - f(x^*)) \\
&\stackrel{(5.1)}{=} \tau(d(x^*) - f(x^*)) = \tau g_*.
\end{aligned} \quad (5.11)$$

Hence, $\varphi_\mu^* \leq (1-\tau)(1-\alpha)\varphi(x_0) + \tau(1-\alpha)\varphi^* + \alpha F(x, f(x) + \frac{\mu\tau}{\alpha}g_*)$. Let our parameters satisfy inequality $\frac{\mu\tau}{\alpha(1-\tau)} \leq \frac{1}{\xi_0^*}$. Then

$$\frac{\mu\tau}{\alpha} \leq \frac{1-\tau}{\xi_0^*} \stackrel{(5.8)}{\leq} \frac{1}{\xi_A(x; g_*)}.$$

This means that $F(x, f(x) + \frac{\mu\tau}{\alpha}g_*) \leq A$, and we get

$$\varphi_\mu^* - \varphi^* \leq (1-\tau)(1-\alpha)(\varphi(x_0) - \varphi^*) + \alpha(A - \varphi^*). \quad \square$$

Let us put in (5.10) the best values of parameters. If τ satisfies (5.9) as equality, then

$$\tau = \frac{\alpha}{\alpha + \mu \xi_0^*}, \quad 1 - \tau = \frac{\mu \xi_0^*}{\alpha + \mu \xi_0^*}.$$

Using the upper bound $A \geq \varphi(x_0)$, we get the following estimate:

$$\varphi_\mu^* - \varphi^* \leq \left[\frac{\mu\xi_0^*(1-\alpha)}{\alpha+\mu\xi_0^*} + \alpha \right] (A - \varphi^*) = \frac{\mu\xi_0^* + \alpha^2}{\mu\xi_0^* + \alpha} (A - \varphi^*).$$

Denoting $\beta = \mu\xi_0^*$, we can find the optimal α_* from the equation

$$\frac{2\alpha_*}{\beta + \alpha_*^2} = \frac{1}{\beta + \alpha_*}.$$

Thus, $\alpha_* + \beta = \sqrt{\beta + \beta^2}$. This means that $\alpha_* = \frac{\beta}{\beta + \sqrt{\beta + \beta^2}}$. Hence,

$$\frac{\beta(1-\alpha_*)}{\beta + \alpha_*} + \alpha_* = \frac{2\beta}{\beta + \sqrt{\beta + \beta^2}} \leq 2\sqrt{\beta}.$$

In other words, we get the following bound:

$$\varphi_\mu^* - \varphi^* \leq 2\sqrt{\mu\xi_0^*} (A - \varphi^*). \quad (5.12)$$

Therefore, if we have an approximate solution $\bar{x}$ to the regularized problem, which satisfies (5.6), we can ensure the following bound for the original problem

$$\begin{aligned} \varphi(\bar{x}) - \varphi^* &\stackrel{(5.12)}{\leq} \varphi(\bar{x}) - \varphi_\mu^* + 2\sqrt{\mu\xi_0^*} (A - \varphi^*) \stackrel{(5.4)}{\leq} \varphi_\mu(\bar{x}) - \varphi_\mu^* + 2\sqrt{\mu\xi_0^*} (A - \varphi^*) \\ &\stackrel{(5.6)}{\leq} \delta(\varphi(x_0) - \varphi^*) + 2\sqrt{\mu\xi_0^*} (A - \varphi^*), \end{aligned}$$

and the regularization parameter should be of the following order:

$$\mu \approx \frac{\delta^2}{\xi_0}$$

Now, let us discuss a possible choice for the regularization functions. Our goal is to have a uniformly convex smooth part of the objective. Thus, the following candidate for the regularizer is the most natural:

$$d_i(x) := f_i(x) + \frac{c_i}{p+1} \|x - x_0\|^{p+1},$$

for a certain $c_i > 0$. This function is uniformly convex of degree $p+1$ with parameter $\sigma_{p+1}(d_i) = \frac{c_i}{2^{p-1}}$ (see e.g. Lemma 2.5 in [9]). Moreover, its p th derivative is Lipschitz continuous with constant $L_p(d_i) = L_p(f_i) + c_i \cdot p!$ (see Theorem 7.1 in [28]).

Hence, applying method (4.3) to the regularized objective, we obtain the linear rate

$$\varphi_\mu(x_k) - \varphi_\mu^* \stackrel{(4.4)}{\leq} (1 - \beta)^k (\varphi_\mu(x_0) - \varphi_\mu^*),$$

where the condition number is equal to

$$\beta = \hat{\beta}_p((1 - \mu)f + \mu d) = \min_{1 \leq i \leq m} \frac{(p! \gamma_p((1 - \mu)f_i + \mu d_i))^{\frac{1}{p}}}{(1 + p)^{\frac{1}{p}} + (p! \gamma_p((1 - \mu)f_i + \mu d_i))^{\frac{1}{p}}}, \quad (5.13)$$

with $\gamma_p((1 - \mu)f_i + \mu d_i) = \frac{\sigma_{p+1}((1 - \mu)f_i + \mu d_i)}{L_p((1 - \mu)f_i + \mu d_i)} = \frac{\mu c_i}{2^{p-1}(L_p(f_i) + \mu c_i p!)}$. We see that it is natural to set $c_i := L_p(f_i)$. In this case, we have

$$\beta = \left[1 + ((1 + p)2^{p-1}(\frac{1}{\mu p!} + 1))^{\frac{1}{p}} \right]^{-1}.$$

Thus, parameter μ plays a crucial role in the complexity of regularized problem (5.5).

6 Subhomogeneous Functions

In this section, we consider a finer problem class by adding some additional assumption on the outer component of the fully composite objective. We show that for such problems, it is possible to prove the global convergence rates for the methods in a general convex case, when the smooth part is not necessary uniformly convex. At the same time, we demonstrate that our methods can be accelerated.

A closed convex function $f : \text{dom } f \rightarrow \mathbb{R}$ is called *subhomogeneous* if for any $x \in \text{dom } f$ and $\gamma \geq 1$ such that $\gamma x \in \text{dom } f$, we have

$$f(\gamma x) \leq \gamma f(x). \quad (6.1)$$

Theorem 4 *Closed and convex function $f(\cdot)$ is subhomogeneous if and only if it satisfies one of the following three conditions:*

$$\langle g_x, x \rangle \leq f(x), \quad x \in \text{dom } f, \quad g_x \in \partial f(x), \quad (6.2)$$

$$\langle g_y, x \rangle \leq f(x), \quad x, y \in \text{dom } f, \quad g_y \in \partial f(y), \quad (6.3)$$

$$f(x + ty) \leq f(x) + tf(y), \quad x, y, x + ty \in \text{dom } f, \quad t \geq 0. \quad (6.4)$$

Proof:

Assume that (6.1) is true. Then

$$\gamma f(x) \geq f(\gamma x) \geq f(x) + \langle g_x, (\gamma - 1)x \rangle,$$

and this is (6.2).

Assume (6.2) is true. Since f is convex, for any $x, y \in \text{dom } f$ and $g_y \in \partial f(y)$, we have

$$f(x) - \langle g_y, x \rangle \geq f(y) - \langle g_y, y \rangle \stackrel{(6.2)}{\geq} 0.$$

This is relation (6.3).

Finally, assume that (6.3) is true. For any $y \in \text{dom } f$ denote by $g(y)$ a particular subgradient in $\partial f(y)$. Then, for any $x \in \text{dom } f$ and $\gamma > 1$ such that $\gamma x \in \text{dom } f$, we have

$$f(\gamma x) = f(x) + \int_0^{\gamma-1} \langle g(x + \tau x), x \rangle d\tau \stackrel{(6.3)}{\leq} f(x) + \int_0^{\gamma-1} f(x) d\tau = f(x).$$

And this is (6.1). Thus, conditions (6.1), (6.2) and (6.3) are equivalent.

In order to justify equivalence with (6.4), note that it can be rewritten as

$$\frac{1}{t}[f(x + ty) - f(x)] \leq f(y).$$

Therefore,

$$\max_{g \in \partial f(x)} \langle g, y \rangle = \lim_{t \rightarrow +0} \frac{1}{t}[f(x + ty) - f(x)] \leq f(y),$$

and this is (6.3). On the other hand, if (6.3) is true, then

$$f(x + ty) - f(x) = \int_0^t \langle g(x + \tau y), y \rangle d\tau \stackrel{(6.3)}{\leq} \int_0^t f(y) d\tau = tf(y).$$

And this is (6.4). □

Example 2 Clearly, function $f(x) = \max_{i=1}^n x^{(i)}$ is subhomogeneous.

Example 3 Consider the following function:

$$f(x) = \ln \left(\sum_{i=1}^n e^{x^{(i)}} \right) = \max_{u \in \Delta_n} \{ \langle u, x \rangle - \eta(u) \},$$

where $\Delta_n \in \mathbb{R}_+^n$ is the standard simplex, and $\eta(u) = \sum_{i=1}^n u^{(i)} \ln u^{(i)}$ is the negative entropy. Denote by $u(x)$ the unique optimal solution to this problem. Then $\nabla f(x) = u(x)$, and we conclude that

$$\langle \nabla f(x), x \rangle = \langle u(x), x \rangle \leq \langle u(x), x \rangle - \eta(u(x)) = f(x),$$

since $\eta(u) \leq 0$ for all $u \in \Delta_n$. Thus, function $f(\cdot)$ is subhomogeneous since condition (6.2) is satisfied. $\square$

Example 4 Let function f be subhomogeneous. Then, for a linear operator A , function

$$\bar{f}(x) = f(Ax)$$

is subhomogeneous too. Further, we can handle any affine transformation $Ax + b$, by incorporating into our problem an auxiliary variable $\tau \in \mathbb{R}$:

$$\bar{f}(x, \tau) = f(Ax + \tau b),$$

with additional normalizing constraint $\tau = 1$. $\square$

Remark 2 Let $F : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper closed convex. Assume that F is monotone on its domain. Consequently, with any point $x \in \text{dom } F$, we have

$$x - \mathbb{R}_+^m \subseteq \text{dom } F.$$

Hence, if the domain contains a vector with strictly positive entries, we have

$$0 \in \text{int}(\text{dom } F). \quad (6.5)$$

Assume that for all $x, y \in \text{dom } F$ and any $t \geq 0$, it holds

$$F(x + ty) \leq F(x) + tF(y). \quad (6.6)$$

Combining (6.5) and (6.6), we conclude $\text{dom } F = \mathbb{R}^m$. $\square$

Now, let us introduce our additional assumption.

Assumption 5 For any $x \in \text{dom } \varphi$, the function $F(x, u)$ is subhomogeneous in $u \in \mathbb{R}^m$. Thus, for any $x \in \text{dom } \varphi$, it holds

$$F(x, u + tv) \leq F(x, u) + tF(x, v), \quad u, v \in \mathbb{R}^m, \quad t \geq 0. \quad (6.7)$$

We are ready to analyze convergence of the Full-step p th order Basic Method (4.3) in a general convex case, when uniform convexity is absent. Let us use the following notation:

$$F(L_p(f)) \stackrel{\text{def}}{=} \sup_{x \in \text{dom } \varphi} F(x, L_p(f)).$$

Theorem 5 *Let the initial level set be bounded*

$$D_0 \stackrel{\text{def}}{=} \sup_x \left\{ \|x - x^*\| : \varphi(x) \leq \varphi(x_0) \right\} < +\infty. \quad (6.8)$$

Then, for the iterations $\{x_k\}_{k \geq 1}$ of the method (4.3), we have

$$\varphi(x_k) - \varphi^* \leq \frac{(p+1)^{p+1} F(L_p(f)) D_0^{p+1}}{p!} \cdot k^{-p}. \quad (6.9)$$

Proof:

By the definition of the method step, we have

$$\begin{aligned} \varphi(x_{k+1}) &= F(x_{k+1}, f(x_{k+1})) \\ &\stackrel{(1.6)}{\leq} F\left(x_{k+1}, \Omega_p(f, x_k; x_{k+1}) + \frac{pL_p(f)}{(p+1)!} \|x_{k+1} - x_k\|^{p+1}\right) \\ &\leq F\left(y, \Omega_p(f, x_k; y) + \frac{pL_p(f)}{(p+1)!} \|y - x_k\|^{p+1}\right) \\ &\stackrel{(1.6)}{\leq} F\left(y, f(y) + \frac{L_p(f)}{p!} \|y - x_k\|^{p+1}\right) \\ &\stackrel{(6.7)}{\leq} \varphi(y) + \frac{F(L_p(f))}{p!} \|y - x_k\|^{p+1}, \end{aligned} \quad (6.10)$$

for all $y \in \text{dom } \varphi$. Substituting $y := x_k$, we conclude

$$\varphi(x_{k+1}) \leq \varphi(x_k).$$

Hence, the method is monotone. Let us take a convex combination $y := \frac{a_{k+1}}{A_{k+1}} x^* + \frac{A_k}{A_{k+1}} x_k$, where $A_k := k \cdot (k+1) \cdot \dots \cdot (k+p)$, and $a_{k+1} := A_{k+1} - A_k = \frac{p+1}{k+p+1} A_{k+1}$. Therefore, we obtain by convexity

$$\begin{aligned} \varphi(x_{k+1}) &\leq \frac{a_{k+1}}{A_{k+1}} \varphi^* + \frac{A_k}{A_{k+1}} \varphi(x_k) + \left(\frac{a_{k+1}}{A_{k+1}}\right)^{p+1} \cdot \frac{F(L_p(f)) \|x^* - x_k\|^{p+1}}{p!} \\ &\leq \frac{a_{k+1}}{A_{k+1}} \varphi^* + \frac{A_k}{A_{k+1}} \varphi(x_k) + \left(\frac{p+1}{k+p+1}\right)^{p+1} \cdot \frac{F(L_p(f)) D_0^{p+1}}{p!}. \end{aligned} \quad (6.11)$$

Multiplying both sides by A_{k+1} , we get

$$\begin{aligned} A_{k+1}(\varphi(x_{k+1}) - \varphi^*) &\leq A_k(\varphi(x_k) - \varphi^*) + A_{k+1} \left(\frac{p+1}{k+p+1}\right)^{p+1} \cdot \frac{F(L_p(f)) D_0^{p+1}}{p!} \\ &\leq A_k(\varphi(x_k) - \varphi^*) + \frac{(p+1)^{p+1} F(L_p(f)) D_0^{p+1}}{p!}. \end{aligned}$$

Summing up the last inequality for different iterations, we finally obtain, for $k \geq 1$:

$$\varphi(x_k) - \varphi^* \leq \frac{1}{A_k} \cdot \frac{k(p+1)^{p+1} F(L_p(f)) D_0^{p+1}}{p!} \leq \frac{(p+1)^{p+1} F(L_p(f)) D_0^{p+1}}{p!} \cdot k^{-p}. \quad \square$$

7 Fully Composite Gradient Methods

Let us consider a more efficient version of the Fully Composite Methods for the particular case $p = 1$ (first-order algorithms). We start with the basic scheme.

Basic Gradient Method	
Choose $x_0 \in \text{dom } \varphi$ and $M = \alpha F(L_1(f))$, $\alpha \geq 1$. For $k \geq 0$ iterate: $x_{k+1} = \underset{y \in \text{dom } \varphi}{\operatorname{argmin}} \left\{ F\left(y, f(x_k) + \langle \nabla f(x_k), y - x_k \rangle\right) + \frac{M}{2} \ y - x_k\ ^2 \right\}.$	(7.1)

Contrary to the scheme (4.3), the regularization term in the method (7.1) is *outside* of the composite part. Therefore, an implementation of each step can be much simpler.

For one iteration of the method, for all $y \in \text{dom } \varphi$, we have

$$\begin{aligned}
 \varphi(x_{k+1}) &\stackrel{(1.6)}{\leq} F\left(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_1(f)}{2} \|x_{k+1} - x_k\|^2\right) \\
 &\stackrel{(6.7)}{\leq} F\left(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle\right) + \frac{\alpha F(L_1(f))}{2} \|x_{k+1} - x_k\|^2 \\
 &\stackrel{(7.1)}{\leq} F\left(y, f(x_k) + \langle \nabla f(x_k), y - x_k \rangle\right) + \frac{\alpha F(L_p(f)) \|y - x_k\|^2}{2} \\
 &\leq \varphi(y) + \frac{\alpha F(L_p(f)) \|y - x_k\|^2}{2},
 \end{aligned}$$

where we used component-wise convexity of f and monotonicity of F in the last inequality. By the same arguments as in the proof of Theorem 5, we get the following result.

Theorem 6 *Let the initial level set be bounded (6.8). Then, for the sequence $\{x_k\}_{k \geq 1}$ generated by the method (7.1), it holds*

$$\varphi(x_k) - \varphi^* \leq \frac{4\alpha F(L_1(f)) D_0^2}{k}.$$

For the problems with bounded domain, we can propose the following alternative scheme, which is a generalization of the classical Frank-Wolfe algorithm [10, 23]. In this method, we do not use an explicit regularizer. Thus, the cost of each step is usually even cheaper than in the Gradient Method (7.1).

Contracting Conditional Gradient Method
Choose $x_0 \in \text{dom } \varphi$ and $\{\gamma_k\}_{k \geq 0}$. For $k \geq 0$ iterate: $x_{k+1} = \underset{y}{\operatorname{argmin}} \left\{ F(y, f(x_k) + \langle \nabla f(x_k), y - x_k \rangle) \right. \\ \left. : x_k + \frac{1}{\gamma_k}(y - x_k) \in \text{dom } \varphi \right\}.$

(7.2)

Theorem 7 Let $\text{dom } \varphi$ be a bounded convex set. Denote its diameter by

$$\mathcal{D} \stackrel{\text{def}}{=} \sup_{x, y \in \text{dom } \varphi} \|x - y\| < +\infty. \quad (7.3)$$

Set $\gamma_k := \frac{2}{k+2}$. Then, for the iterations $\{x_k\}_{k \geq 1}$ of the method (7.2), it holds

$$\varphi(x_k) - \varphi^* \leq \frac{4F(L_1(f))\mathcal{D}^2}{k}. \quad (7.4)$$

Proof:

Let us denote the point $v_{k+1} \stackrel{\text{def}}{=} x_k + \frac{1}{\gamma_k}(x_{k+1} - x_k) \stackrel{(7.2)}{\in} \text{dom } \varphi$. Hence,

$$\|x_{k+1} - x_k\| = \gamma_k \|v_{k+1} - x_k\| \leq \gamma_k \mathcal{D}. \quad (7.5)$$

Now, considering one iteration of the method, we obtain

$$\begin{aligned} \varphi(x_{k+1}) &\stackrel{(1.6)}{\leq} F\left(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_1(f)}{2} \|x_{k+1} - x_k\|^2\right) \\ &\stackrel{(6.7)}{\leq} F\left(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle\right) + \frac{F(L_1(f))}{2} \|x_{k+1} - x_k\|^2 \\ &\stackrel{(7.5)}{\leq} F\left(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle\right) + \frac{\gamma_k^2 F(L_1(f)) \mathcal{D}^2}{2} \\ &\stackrel{(7.2)}{\leq} F\left(y, f(x_k) + \langle \nabla f(x_k), y - x_k \rangle\right) + \frac{\gamma_k^2 F(L_1(f)) \mathcal{D}^2}{2} \\ &\leq \varphi(y) + \frac{\gamma_k^2 F(L_1(f)) \mathcal{D}^2}{2}, \end{aligned}$$

for every $y \in \text{dom } \varphi$. Substituting $y := \gamma_k x^* + (1 - \gamma_k)x_k$, and using convexity of φ , we obtain inequality very similar to (6.11), from the proof of Theorem 5 for $p = 1$. Hence, by the same arguments, we establish the rate (7.4). $\square$

Finally, we can present an accelerated method (see [19, 24, 4]).

Fully Composite Fast Gradient Method

Choose $x_0 \in \text{dom } \varphi$ and $M = \alpha F(L_1(f))$, $\alpha \geq 1$. Set $v_0 = x_0$, $A_0 = 0$.

For $k \geq 0$ **iterate**:

1. Find a_{k+1} from the equation $A_k + a_{k+1} = a_{k+1}^2$. Set $A_{k+1} = A_k + a_{k+1}$.
2. $y_k = \frac{a_{k+1}v_k + A_k x_k}{A_{k+1}}$.
3. $x_{k+1} = \underset{y}{\operatorname{argmin}} \left\{ F(y, f(y_k) + \langle \nabla f(y_k), y - y_k \rangle) + \frac{M}{2} \|y - y_k\|^2 \right\}$.
4. $v_{k+1} = x_{k+1} + \frac{A_k}{a_{k+1}}(x_{k+1} - x_k)$.

(7.6)

Theorem 8 For the sequence $\{x_k\}_{k \geq 1}$, generated by the method (7.6), it holds

$$A_k \varphi(x_k) + \frac{M}{2} \|x - v_k\|^2 \leq A_k \varphi(x) + \frac{M}{2} \|x - x_0\|^2, \quad x \in \text{dom } \varphi. \quad (7.7)$$

Consequently, in the case $x = x^*$, we get the following convergence guarantees:

$$\varphi(x_k) - \varphi^* \leq \frac{M \|x^* - x_0\|^2}{2A_k} \leq \frac{2M \|x^* - x_0\|^2}{k^2}. \quad (7.8)$$

Proof:

Let us establish (7.7) by induction. Assume that it holds for the current iterate, and consider the next step. We fix an arbitrary $x \in \text{dom } \varphi$ and denote the convex combination $y := \frac{a_{k+1}x + A_k x_k}{A_{k+1}}$. Then

$$\begin{aligned} \frac{M}{2} \|x - x_0\|^2 + A_{k+1} \varphi(x) &= \frac{M}{2} \|x - x_0\|^2 + A_k \varphi(x) + a_{k+1} \varphi(x) \\ &\stackrel{(7.7)}{\geq} \frac{M}{2} \|x - v_k\|^2 + A_k \varphi(x_k) + a_{k+1} \varphi(x) \\ &\geq \frac{M}{2} \|x - v_k\|^2 + A_{k+1} \varphi(y) = A_{k+1} \left(\frac{M}{2} \|y - y_k\|^2 + \varphi(y) \right) \\ &\geq A_{k+1} \left(\frac{M}{2} \|y - y_k\|^2 + F(y, f(y_k) + \langle \nabla f(y_k), y - y_k \rangle) \right), \end{aligned} \quad (7.9)$$

where in the last inequality we used convexity of components of f and monotonicity of F .

The function in the right hand side of (7.9) is *strongly convex* in y . Hence, we obtain

$$\begin{aligned}
& \frac{M}{2} \|y - y_k\|^2 + F(y, f(y_k) + \langle \nabla f(y_k), y - y_k \rangle) \\
& \stackrel{(7.6)}{\geq} \frac{M}{2} \|y - x_{k+1}\|^2 + \frac{M}{2} \|x_{k+1} - y_k\|^2 + F(x_{k+1}, f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle) \\
& \stackrel{(6.7)}{\geq} \frac{M}{2} \|y - x_{k+1}\|^2 + F\left(x_{k+1}, f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L_1(f)}{2} \|x_{k+1} - y_k\|^2\right) \\
& \stackrel{(1.6)}{\geq} \frac{M}{2} \|y - x_{k+1}\|^2 + \varphi(x_{k+1}) = \frac{M}{2A_{k+1}} \|x - v_{k+1}\|^2 + \varphi(x_{k+1}).
\end{aligned}$$

Thus we establish (7.7) for all $k \geq 0$.

Note that $\sqrt{A_{k+1}} - \sqrt{A_k} = \frac{a_{k+1}}{\sqrt{A_{k+1}} + \sqrt{A_k}} \geq \frac{a_{k+1}}{2\sqrt{A_{k+1}}} = \frac{1}{2}$. Hence, $A_k \geq \frac{k^2}{4}$, and this proves the rate of convergence (7.8). $\square$

8 Fully Composite Newton Methods

In this section, we analyze different variants of the second-order methods for the fully composite formulation. Let us start with cubic regularization of the Newton's method [27].

Fully Composite Cubic Newton Method
Choose $x_0 \in \text{dom } \varphi$ and $M = \alpha F(L_2(f))$, $\alpha \geq 1$. For $k \geq 0$ iterate: $x_{k+1} = \underset{y \in \text{dom } \varphi}{\operatorname{argmin}} \left\{ F(y, \Omega_2(f, x_k; y)) + \frac{M}{6} \ y - x_k\ ^3 \right\}.$

(8.1)

Iterations of this type are simpler than those in the general method (4.3) with $p = 2$ since the regularization term is *outside* the composite part. At the same time, for any

$y \in \text{dom } \varphi$, we have the following guarantee:

$$\begin{aligned}
\varphi(x_{k+1}) &\stackrel{(1.6)}{\leq} F(x_{k+1}, \Omega_2(f, x_k; x_{k+1}) + \frac{L_2(f)}{6} \|x_{k+1} - x_k\|^3) \\
&\stackrel{(6.7)}{\leq} F(x_{k+1}, \Omega_2(f, x_k; x_{k+1})) + \frac{M}{6} \|x_{k+1} - x_k\|^3 \\
&\stackrel{(8.1)}{\leq} F(y, \Omega_2(f, x_k; y)) + \frac{M}{6} \|y - x_k\|^3 \\
&\stackrel{(1.6)}{\leq} F(y, f(y) + \frac{L_2(f)}{6} \|y - x_k\|^3) + \frac{M}{6} \|y - x_k\|^3 \\
&\stackrel{(6.7)}{\leq} \varphi(y) + \frac{(1+\alpha)F(L_2(f))}{6} \|y - x_k\|^3.
\end{aligned} \tag{8.2}$$

Hence, using the same reasoning as in Theorem 5, we prove the global convergence result.

Theorem 9 *Let the initial level set be bounded (6.8). Then, for the sequence $\{x_k\}_{k \geq 1}$ generated by the method (8.1), we have*

$$\varphi(x_k) - \varphi^* \leq \frac{9(1+\alpha)F(L_2(f))D_0^3}{2k^2}.$$

In papers[7, 5], we looked at another modification of the Newton's method, based on the *contracting idea*. Let us present the corresponding version of the algorithm for the Fully Composite Formulation. This method can be seen as a second-order counterpart of the Conditional Gradient Method (7.2).

Fully Composite Contracting Newton Method
Choose $x_0 \in \text{dom } \varphi$ and $\{\gamma_k\}_{k \geq 0}$. For $k \geq 0$ iterate: $x_{k+1} = \underset{y}{\operatorname{argmin}} \left\{ F(y, \Omega_2(f, x_k; y)) \right. \\ \left. : x_k + \frac{1}{\gamma_k}(y - x_k) \in \text{dom } \varphi \right\}.$

(8.3)

Repeating the previous reasoning, we obtain the following result.

Theorem 10 *Let the size of $\text{dom } \varphi$ be bounded by diameter $\mathcal{D}$ (7.3). Define $\gamma_k := \frac{3}{k+3}$. Then, for the sequence $\{x_k\}_{k \geq 1}$, generated by the method (8.3), we have*

$$\varphi(x_k) - \varphi^* \leq \frac{9F(L_2(f))\mathcal{D}^3}{k^2}.$$

9 Fully Composite Contracting Proximal Scheme

In this section, we develop an accelerated second-order method.

The cubic regularization of the Newton's method was accelerated in [21], using the *Estimating Functions* technique. It is based on accumulating the gradients at the new points of the optimization process into a global linear model. However, for the fully composite problems, we can guarantee only the progress in terms of the *objective function*, and the good properties of the gradients are not easily available.

Therefore, we use inexact Contracting Proximal-Point iterations (see [6, 14, 17]) as the basis of our accelerated scheme. For simplicity, we consider the case $p = 2$ (second-order methods). Generalization to arbitrary $p \geq 1$ is more or less straightforward (see also [8]).

Let us choose a prox-function, suitable for our problem class:

$$d(x) := \frac{\alpha}{3} \|x - x_0\|^3, \quad x, x_0 \in \mathbb{E}, \quad \alpha := F(L_2(f)).$$

It is well known that this function is *uniformly convex* of degree 3 (see e.g. [9]), so it holds:

$$\rho_d(x; y) \stackrel{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq \frac{\alpha}{6} \|y - x\|^3, \quad x, y \in \mathbb{E}. \quad (9.1)$$

We need the following facts on Bregman divergence. They can be checked in a direct way.

1. For a closed convex function $\psi : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ and a fixed prox center $v \in \mathbb{E}$, denote $h(x) := \psi(x) + \rho_d(v; x)$. Then, for the optimum $T := \arg\min_x h(x)$, it holds

$$h(x) \geq h(T) + \rho_d(T; x), \quad x \in \mathbb{E}. \quad (9.2)$$

In other words, the function $h(\cdot)$ is *strongly convex with respect to* $d(\cdot)$.

2. For any $\bar{v}, v \in \mathbb{E}$ and $x \in \mathbb{E}$, it holds

$$\rho_d(\bar{v}; x) = \rho_d(v; x) + \rho_d(\bar{v}; v) + \langle \nabla d(v) - \nabla d(\bar{v}), x - v \rangle. \quad (9.3)$$

We use this prox function in the following general method.

Fully Composite Contracting Proximal-Point Scheme	
Choose $x_0 \in \text{dom } \varphi$ and $\delta > 0$. Set $v_0 = x_0$, $A_0 = 0$. For $k \geq 0$ iterate: 1. Choose $A_{k+1} = \left(\frac{k+1}{3}\right)^3$. Set $\gamma_k = \frac{A_{k+1} - A_k}{A_{k+1}}$. 2. Form the subproblem $h_{k+1}(x) = A_{k+1}\varphi(\gamma_k x + (1 - \gamma_k)x_k) + \rho_d(v_k; x)$. 3. Find a point v_{k+1} s.t. $h_{k+1}(v_{k+1}) - h_{k+1}^* \leq \delta$ by the basic method (8.1). 4. $x_{k+1} = \gamma_k v_{k+1} + (1 - \gamma_k)x_k$.	(9.4)

Let us justify first its rate of convergence. Then we discuss the efficiency of implementation of the Step 3.

Theorem 11 For the iterations of method (9.4), we have

$$A_k \varphi(x_k) + \rho_d(v_k; x) \leq A_k \varphi(x) + \rho_d(x_0; x) + C_k(x), \quad x \in \mathbb{E}, \quad (9.5)$$

where

$$C_k(x) \stackrel{\text{def}}{=} k\delta + c_1(x) \cdot \sum_{i=1}^k \|x - v_i\| + c_2 \cdot \sum_{i=1}^k \|x - v_i\|^2,$$

with $c_1(x) \stackrel{\text{def}}{=} \alpha^{1/3}(6\delta)^{2/3} + 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - x_0\|$ and $c_2 \stackrel{\text{def}}{=} 2\alpha^{2/3}(6\delta)^{1/3}$.

Proof:

Let us denote by v_{k+1}^* the *exact* minimizer of $h_{k+1}(\cdot)$: $v_{k+1}^* \stackrel{\text{def}}{=} \underset{x}{\operatorname{argmin}} h_{k+1}(x)$. Then, by the presence in the objective of the uniformly convex term, we get

$$\begin{aligned} \delta &\geq h_{k+1}(v_{k+1}) - h_{k+1}(v_{k+1}^*) \stackrel{(9.2)}{\geq} \rho_d(v_{k+1}^*; v_{k+1}) \\ &\stackrel{(9.1)}{\geq} \frac{\alpha}{6} \|v_{k+1} - v_{k+1}^*\|^3. \end{aligned}$$

Hence, we can bound the distance between the exact minimizer and the approximate point v_{k+1} obtained by the inner method, as follows:

$$\|v_{k+1} - v_{k+1}^*\| \leq \left(\frac{6\delta}{\alpha}\right)^{\frac{1}{3}}. \quad (9.6)$$

Now, assume that (9.5) holds by for the current $k \geq 0$, and consider the next step of the method. Let us denote $a_{k+1} := A_{k+1} - A_k$. Then, for arbitrary $x \in \mathbb{E}$, we have

$$\begin{aligned} \rho_d(x_0; x) + A_{k+1} \varphi(x) &= \rho_d(x_0; x) + A_k \varphi(x) + a_{k+1} \varphi(x) \\ &\stackrel{(9.5)}{\geq} \rho_d(v_k; x) + A_k \varphi(x_k) + a_{k+1} \varphi(x) - C_k(x) \\ &\geq \rho_d(v_k; x) + A_{k+1} \varphi(\gamma_k x + (1 - \gamma_k)x_k) - C_k(x) \quad (9.7) \\ &= h_{k+1}(x) - C_k(x) \\ &\stackrel{(9.2)}{\geq} \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}^*) - C_k(x), \end{aligned}$$

where the last inequality follows from *strong convexity* of $h_{k+1}(\cdot)$ with respect to $d(\cdot)$.

We can continue using our inexact solution v_{k+1} to the auxiliary problem:

$$\begin{aligned} \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}^*) &\stackrel{\text{Step 3}}{\geq} \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}) - \delta \\ &\stackrel{(9.3)}{=} \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - \delta + \rho_d(v_{k+1}^*; v_{k+1}) \\ &\quad + \langle \nabla d(v_{k+1}) - \nabla d(v_{k+1}^*), x - v_{k+1} \rangle \\ &\geq \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - \delta \\ &\quad - \|\nabla d(v_{k+1}) - \nabla d(v_{k+1}^*)\|_* \cdot \|x - v_{k+1}\|. \end{aligned}$$

Now, since $\nabla^2 d(x) = \|x - x_0\|B + \frac{1}{\|x - x_0\|}B(x - x_0)(x - x_0)^*B \preceq 2\alpha\|x - x_0\|B$ for all $x \in \mathbb{E}$, we can bound the difference between the gradients, as follows:

$$\begin{aligned}
\|\nabla d(v_{k+1}) - \nabla d(v_{k+1}^*)\|_* &= \left\| \int_0^1 \nabla^2 d(v_{k+1}^* + \tau(v_{k+1} - v_{k+1}^*)) d\tau (v_{k+1} - v_{k+1}^*) \right\|_* \\
&\leq 2\alpha \cdot \|v_{k+1} - v_{k+1}^*\| \cdot \int_0^1 \|v_{k+1}^* + \tau(v_{k+1} - v_{k+1}^*) - x_0\| d\tau \\
&\leq 2\alpha \cdot \|v_{k+1} - v_{k+1}^*\| \cdot \left(\|v_{k+1} - x_0\| + \frac{1}{2}\|v_{k+1} - v_{k+1}^*\| \right) \\
&\leq 2\alpha \cdot \|v_{k+1} - v_{k+1}^*\| \cdot \|x - v_{k+1}\| \\
&\quad + 2\alpha \cdot \|v_{k+1} - v_{k+1}^*\| \cdot \|x - x_0\| \\
&\quad + \alpha \cdot \|v_{k+1} - v_{k+1}^*\|^2 \\
&\stackrel{(9.6)}{\leq} 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - v_{k+1}\| + 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - x_0\| \\
&\quad + \alpha^{1/3}(6\delta)^{2/3}.
\end{aligned}$$

Therefore, combining all component together, we conclude that

$$\begin{aligned}
\rho_d(x_0; x) + A_{k+1}\varphi(x) &\geq \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - C_k(x) - \delta \\
&\quad - \|x - v_{k+1}\| \cdot \left(\alpha^{1/3}(6\delta)^{2/3} + 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - x_0\| \right) \\
&\quad - \|x - v_{k+1}\|^2 \cdot \left(2\alpha^{2/3}(6\delta)^{1/3} \right) \\
&\equiv \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - C_{k+1}(x) \\
&\geq \rho_d(v_{k+1}; x) + A_{k+1}\varphi(x_{k+1}) - C_{k+1}(x).
\end{aligned}$$

Thus, (9.5) is justified for all $k \geq 0$. $\square$

Let us substitute into (9.5) $x = x^*$, and fix some $K \geq 0$. For all $0 \leq k \leq K$, we get

$$\begin{aligned}
\frac{\alpha}{6}\|x^* - v_k\|^3 &\stackrel{(9.1)}{\leq} \rho_d(v_k; x^*) + A_k(\varphi(x_k) - \varphi^*) \\
&\stackrel{(9.5)}{\leq} \rho_d(x_0; x^*) + K\delta + c_1(x^*) \sum_{i=1}^k \|x^* - v_i\| + c_2 \sum_{i=1}^k \|x^* - v_i\|^2 \stackrel{\text{def}}{=} R_k,
\end{aligned} \tag{9.8}$$

and we need to estimate the quantities R_k from above. Note that

$$\begin{aligned}
R_{k+1} &= R_k + c_1(x^*)\|x^* - v_{k+1}\| + c_2\|x^* - v_{k+1}\|^2 \\
&\stackrel{(9.8)}{\leq} R_k + aR_{k+1}^{1/3} + bR_{k+1}^{2/3},
\end{aligned} \tag{9.9}$$

where

$$\begin{aligned}
a &\stackrel{\text{def}}{=} c_1(x^*) \cdot \left(\frac{6}{\alpha}\right)^{1/3} = 6\delta^{2/3} + 12\delta^{1/3}\rho_d(x_0; x^*)^{1/3}, \\
b &\stackrel{\text{def}}{=} c_2 \cdot \left(\frac{6}{\alpha}\right)^{2/3} = 12\delta^{1/3}.
\end{aligned} \tag{9.10}$$

Dividing (9.9) by $R_{k+1}^{1/3}$ and using monotonicity of the sequence, we obtain quadratic inequality with respect to $R_{k+1}^{1/3}$, that is

$$(R_{k+1}^{1/3})^2 - bR_{k+1}^{1/3} - (R_k^{2/3} + a) \leq 0.$$

It can be resolved as follows:

$$R_{k+1}^{1/3} \leq \frac{b + \sqrt{b^2 + 4(R_k^{2/3} + a)}}{2} \leq b + \sqrt{R_k^{2/3} + a} \leq R_k^{1/3} + b + \sqrt{a}.$$

Hence, telescoping the last inequality, we get

$$R_k \leq (R_0^{1/3} + (b + \sqrt{a})k)^3 \leq 9(R_0 + k^3b^3 + k^3a^{3/2}).$$

Substituting the actual values of the parameters, we come to the following conclusion.

Corollary 1 *For the iterations of the method (9.4), for all $k \geq 1$, it holds*

$$A_k(\varphi(x_k) - \varphi^*) + \rho_d(v_k; x^*) \leq O\left(\rho_d(x_0; x^*) + k^3\delta + k^3\delta^{1/2}\rho_d(x_0; x^*)^{1/2}\right). \quad (9.11)$$

Hence,

$$\varphi(x_k) - \varphi^* \leq O\left(\frac{\rho_d(x_0; x^*)}{k^3} + \delta + \delta^{1/2}\rho_d(x_0; x^*)^{1/2}\right).$$

And to solve the initial problem with ε -accuracy, we need to pick up

$$\delta \approx \min\left\{\varepsilon, \frac{\varepsilon^2}{\rho_d(x_0; x^*)}\right\}. \quad \square \quad (9.12)$$

Let us apply now the Basic Cubic Newton Method (8.1) for solving the subproblems at Step 3. Denote the contracted smooth part by

$$\bar{f}(x) := f(\gamma_k x + (1 - \gamma_k)x_k),$$

and the new outer part by

$$\bar{F}(x, u) := A_{k+1}F(\gamma_k x + (1 - \gamma_k)x_k, u) + \rho_d(v_k; x).$$

Hence, the objective in the subproblem can be represented as follows:

$$h_{k+1}(x) \equiv \bar{F}(x, \bar{f}(x)) \rightarrow \min_{x \in \text{dom } \varphi} \quad (9.13)$$

Then iterations of the method (8.1) applied to (9.13) are:

$$z_{t+1} := \operatorname{argmin}_{x \in \text{dom } \varphi} \left\{ \bar{F}(x, \Omega_2(\bar{f}, z_t; x)) + \frac{\bar{F}(L_2(\bar{f}))}{6} \|x - z_t\|^3 \right\}, \quad t \geq 0,$$

and let us start with $z_0 := v_k$.

Note that $L_2(\bar{f}) = \gamma_k^3 L_2(f)$. Consequently,

$$\begin{aligned} \bar{F}(L_2(\bar{f})) &\leq A_{k+1}F(L_2(\bar{f})) \stackrel{(6.1)}{\leq} \gamma_k^3 A_{k+1}F(L_2(f)) \\ &= \frac{a_{k+1}^3}{A_{k+1}^2} F(L_2(f)) = \frac{((k+1)^3 - k^3)^3}{3^3(k+1)^6} F(L_2(f)) \leq F(L_2(f)). \end{aligned} \quad (9.14)$$

The guarantee (8.2) of one step ensures that, for any $x \in \text{dom } \varphi$, it holds

$$\begin{aligned} h_{k+1}(z_{t+1}) &\leq h_{k+1}(x) + \frac{\bar{F}(L_2(\bar{f}))}{3} \|x - z_t\|^3 \\ &\stackrel{(9.14)}{\leq} h_{k+1}(x) + \frac{F(L_2(f))}{3} \|x - z_t\|^3. \end{aligned} \tag{9.15}$$

Let $x = \tau v_{k+1}^* + (1 - \tau)z_t$, with $\tau := \frac{1}{\sqrt{6}}$ and v_{k+1}^* being the minimizer of (9.13). Then

$$\begin{aligned} h_{k+1}(z_{t+1}) &\leq \tau h_{k+1}^* + (1 - \tau)h_{k+1}(z_t) + \frac{\tau^3 F(L_2(f))}{3} \|v_{k+1}^* - z_t\|^3 \\ &\stackrel{(9.1)}{\leq} \tau h_{k+1}^* + (1 - \tau)h_{k+1}(z_t) + 2\tau^3 (h_{k+1}(z_t) - h_{k+1}^*). \end{aligned}$$

Therefore,

$$\begin{aligned} h_{k+1}(z_{t+1}) - h_{k+1}^* &\leq (1 - \tau + 2\tau^3) \cdot (h_{k+1}(z_t) - h_{k+1}^*) \\ &= \left(1 - \frac{2}{3\sqrt{6}}\right) \cdot (h_{k+1}(z_t) - h_{k+1}^*) \leq \frac{3}{4} (h_{k+1}(z_t) - h_{k+1}^*). \end{aligned}$$

We see that our subsolver has a fast *linear* rate of convergence, which does not depend on any condition number. Let us estimate the residual after one step of the method. Substituting $x = x^*$ (the solution to the original problem) into (9.15), we get

$$\begin{aligned} h_{k+1}(z_1) - h_{k+1}^* &\stackrel{(9.15)}{\leq} h_{k+1}(x^*) - h_{k+1}^* + \frac{F(L_2(f))}{3} \|x^* - v_k\|^3 \\ &\leq a_{k+1}\varphi^* + A_k\varphi(x_k) + \rho_d(v_k; x^*) - h_{k+1}^* + \frac{F(L_2(f))}{3} \|x^* - v_k\|^3 \\ &\stackrel{(*)}{\leq} A_k(\varphi(x_k) - \varphi^*) + \rho_d(v_k; x^*) + 2\rho_d(v_k; x^*) \\ &\stackrel{(9.11), (9.12)}{\leq} O\left(\rho_d(x_0; x^*) + k^3\varepsilon\right), \end{aligned}$$

where we used in $(*)$ the uniform convexity of the prox-function and the following bound:

$$\begin{aligned} h_{k+1}(x) &\geq A_{k+1}F(\gamma_k x + (1 - \gamma_k)x_k, f(\gamma_k x + (1 - \gamma_k)x_k)) \\ &\geq \min_{y \in \text{dom } \varphi} A_{k+1}F(y, f(y)) = A_{k+1}\varphi^*. \end{aligned}$$

Combining these bounds together, we come to the following final conclusion.

Corollary 2 *For solving the initial problem with ε -accuracy:*

$$\varphi(x_K) - \varphi^* \leq \varepsilon,$$

we need to perform $K = O\left(\left[\frac{F(L_2(f))\|x_0 - x^\|^3}{\varepsilon}\right]^{1/3}\right)$ iterations of the proximal-point scheme (9.4). At each iteration, it requires no more than*

$$N = O\left(1 + \log\left[\frac{F(L_2(f))\|x_0 - x^*\|^3}{\varepsilon}\right]\right)$$

steps of the basic method (8.1).

We see that the price to pay for the level of generality is an additional logarithmic term in the final complexity estimate. It remains to be an open theoretical question: whether we can develop a *direct* accelerated high-order method for Fully Composite Formulation, which do not need inexact proximal iterations. It would also help in constructing *optimal* high-order methods [18, 11, 16, 2], matching the existing lower complexity bounds [1, 24].

Another interesting research direction is the development of *universal* [12, 13] and *randomized* [3, 15] variants of the Fully Composite Methods.

References

- [1] Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019.
- [2] Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. In *Conference on Learning Theory*, pages 492–507. PMLR, 2019.
- [3] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018.
- [4] Alexandre d’Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *arXiv preprint arXiv: 2101.09545*, 2021.
- [5] Nikita Doikov and Yurii Nesterov. Affine-invariant contracting-point methods for convex optimization. *CORE Discussion Papers 2020/29*, 2020.
- [6] Nikita Doikov and Yurii Nesterov. Contracting proximal methods for smooth convex optimization. *SIAM Journal on Optimization*, 30(4):3146–3169, 2020.
- [7] Nikita Doikov and Yurii Nesterov. Convex optimization based on global lower second-order models. *Advances in Neural Information Processing Systems*, 33:16546–16556, 2020.
- [8] Nikita Doikov and Yurii Nesterov. Inexact tensor methods with dynamic accuracies. In *International Conference on Machine Learning*, pages 2577–2586. PMLR, 2020.
- [9] Nikita Doikov and Yurii Nesterov. Minimizing uniformly convex functions by cubic regularization of newton method. *Journal of Optimization Theory and Applications*, pages 1–23, 2021.
- [10] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.
- [11] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César A Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In *Conference on Learning Theory*, pages 1374–1391. PMLR, 2019.
- [12] Geovani Grapiglia and Yurii Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017.

- [13] Geovani Grapiglia and Yurii Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. *SIAM Journal on Optimization*, 30(4):2750–2779, 2020.
- [14] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.
- [15] Filip Hanzely, Nikita Doikov, Peter Richtárik, and Yurii Nesterov. Stochastic subspace cubic Newton method. In *International Conference on Machine Learning*, pages 4027–4038. PMLR, 2020.
- [16] Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. In *Conference on Learning Theory*, pages 1799–1801. PMLR, 2019.
- [17] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.
- [18] Renato DC Monteiro and Benar F Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.
- [19] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.
- [20] Yurii Nesterov. *Effective methods in nonlinear programming*. Moscow, Radio i Svyaz, 1989.
- [21] Yurii Nesterov. Accelerating the cubic regularization of Newton’s method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.
- [22] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
- [23] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1-2):311–330, 2018.
- [24] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- [25] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, pages 1–27, 2019.
- [26] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.
- [27] Yurii Nesterov and Boris Polyak. Cubic regularization of Newton’s method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
- [28] Anton Rodomanov and Yurii Nesterov. Smoothness parameter of power of Euclidean norm. *Journal of Optimization Theory and Applications*, 185(2):303–326, 2020.