HIGH-ORDER OPTIMIZATION METHODS FOR FULLY COMPOSITE PROBLEMS

Nikita Doikov, Yurii Nesterov

REPRINT | 3241

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

SIAM J. OPTIM.
Vol. 32, No. 3, pp. 2402–2427

HIGH-ORDER OPTIMIZATION METHODS FOR FULLY COMPOSITE PROBLEMS*

NIKITA DOIKOV† AND YURII NESTEROV‡

Abstract. In this paper, we study a fully composite formulation of convex optimization problems, which includes, as a particular case, the problems with functional constraints, max-type minimization problems, and problems with simple nondifferentiable components. We treat all these formulations in a unified way, highlighting the existence of very natural optimization schemes of different order $p \geq 1$. As the result, we obtain new high-order ($p \geq 2$) optimization methods for composite formulation. We prove the global convergence rates for them 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.

Key words. convex optimization, constrained optimization, nonsmooth optimization, gradient methods, high-order methods, accelerated algorithms

MSC codes. 65K05, 90C25, 90C30, 49M37, 49M15

DOI. 10.1137/21M1410063

1. Introduction.

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 [50, 51] or with a simple feasible set. The problems can be posed with differentiable or nondifferentiable components. Sometimes we speak about problems in additive composite form (e.g., [43]). Or, we can speak about optimization of max-type functions (e.g., section 2.3 in [45]).

All these formulations have quite specific properties and usually they need development of the specific methods. In this work, 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 this formulation (we call it the *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, it is possible to highlight the generic reasons for existence of the efficient methods for many different types of problems.

The study of a general class of composite optimization problems has a long and rich history (see, e.g., [9, 10, 14, 12, 13, 18, 25, 26, 11]). A more specific study of *convex* composite formulation can also be found in [49, 7, 4, 5, 6, 3, 15].

*Received by the editors April 5, 2021; accepted for publication (in revised form) July 12, 2022; published electronically September 26, 2022.

https://doi.org/10.1137/21M1410063

Funding: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 788368). The research of the second author was also supported by the Multidisciplinary Institute in Artificial Intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

†Institute of Information and Communication Technologies, Electronics and Applied Math (ICTEAM), Catholic University of Louvain (UCL), 1348 Louvain-la-Neuve, Belgium (Nikita.Doikov@uclouvain.be).

‡Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 1348 Louvain-la-Neuve, Belgium (Yurii.Nesterov@uclouvain.be).

In this paper, we develop new high-order ($p \geq 2$) methods for solving fully composite convex problems and study their global complexity. 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. For $p = 1$ (first-order methods), our framework brings back a well-known class of composite gradient methods [9, 35] and recovers their global rates for convex minimization [38].

Our first-, second-, and third-order methods can be implemented in practice using the existing polynomial-time technique [46].

Contents. The rest of the paper is organized as follows. Section 2 contains our notation.

In section 3, 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 4, we present our fully composite optimization framework and give several examples, which cover all popular composite settings. Then, in section 5, 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 6, we demonstrate that it is possible to use a simple regularization technique within our framework. It converts any convex problem into a uniformly convex one, and thus our basic methods can be applied to solve them.

Section 7 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 8 and 9, respectively. In particular, we establish the accelerated $O(k^{-2})$ global rate of convergence for the fast gradient method [39] and the same global rate for the modifications of the Newton's method [48, 22].

In section 10, we accelerate our fully composite second-order methods up to the level $O(k^{-3})$ using inexact contracting proximal iterations [21].

2. 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 $B : \mathbb{E} \to \mathbb{E}^*$ (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 several times differentiable function $f : \operatorname{dom} f \to \mathbb{R}$ with convex and open domain $\operatorname{dom} f \subseteq \mathbb{E}$, denote by $\nabla f(x)$ its gradient and by $\nabla^2 f(x)$ its Hessian evaluated at point $x \in \operatorname{dom} f \subseteq \mathbb{E}$. Then

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

In what follows, we often work with directional derivatives. For $p \geq 1$, assuming that f is p-times differentiable, we denote by

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

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:

$$\|D^p f(x)\| \;=\; \max_{h_1,\dots,h_p} \{D^p f(x)[h_1,\dots,h_p] :\; \|h_i\| \le 1,\, i = 1,\dots,p\}\,. \tag{2.1}$$

For example, for any $x \in \operatorname{dom} f$ 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

$$D^p f(x)[h]^p \;\stackrel{\text{def}}{=}\; D^p f(x)[h,\dots,h], \qquad h \in \mathbb{E}.$$

Then, the Taylor approximation of function $f(\cdot)$ at $x \in \operatorname{dom} f$ can be written as follows:

$$\begin{aligned} f(y) \;&=\; \Omega_p(f,x;y) + o(\|y-x\|^p), \quad y \in \operatorname{dom} f, \\ \Omega_p(f,x;y) \;&\stackrel{\text{def}}{=}\; f(x) + \textstyle\sum\limits_{k=1}^{p} \frac{1}{k!} D^k f(x)[y-x]^k, \quad y \in \mathbb{E}. \end{aligned} \tag{2.2}$$

Note that in general, we have (see, for example, Appendix 1 in [47])

$$\|D^p f(x)\| \;=\; \max_h \left\{ \left| D^p f(x)[h]^p \right| :\; \|h\| \le 1 \right\}. \tag{2.3}$$

Similarly, since for $x, y \in \operatorname{dom} f$ 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

$$\|D^p f(x) - D^p f(y)\| \;=\; \max_h \left\{ \left| D^p f(x)[h]^p - D^p f(y)[h]^p \right| :\; \|h\| \le 1 \right\}. \tag{2.4}$$

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 pth derivative:

$$\|D^p f(x) - D^p f(y)\| \;\le\; L_p \|x-y\|, \quad x, y \in \operatorname{dom} f, \quad p \ge 1. \tag{2.5}$$

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 the function value and its Taylor approximation:

$$|f(y) - \Omega_p(f,x;y)| \;\le\; \tfrac{L_p}{(p+1)!} \|y-x\|^{p+1}, \quad x, y \in \operatorname{dom} f. \tag{2.6}$$

3. Uniform convexity of smooth functions. Let us couple our smoothness assumption (2.5) with *uniform convexity* of certain degree. Namely, let us assume that for $p \ge 1$ there exists a constant $\sigma_{p+1}(f) > 0$ such that

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \;\ge\; \sigma_{p+1}(f) \|y-x\|^{p+1}, \quad x, y \in \operatorname{dom} f. \tag{3.1}$$

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

$$f(y) \;\ge\; f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{\sigma_{p+1}(f)}{p+1} \|y-x\|^{p+1}, \quad x, y \in \operatorname{dom} f. \tag{3.2}$$

Let us consider the uniformly convex functions of degree $p+1$, whose pth derivative is Lipschitz continuous. We introduce the 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 (2.6) and (3.2), we get

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

Thus, in the case of unbounded domain, by dividing both sides of the previous inequality by $\|y-x\|^{p+1}$ and taking the limit $\|y\| \to +\infty$, we have $\frac{D^k f(x)[y-x]^k}{\|y-x\|^{p+1}} \to 0$ for all $1 \leq k \leq p$, and therefore

$$\gamma_p(f) \leq \frac{1}{p!}. \tag{3.3}$$

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

$$\beta_p(f,\alpha) \stackrel{\text{def}}{=} \frac{(p!\,\gamma_p(f))^{\frac{1}{p}}}{(1+\alpha)^{\frac{1}{p}} + (p!\,\gamma_p(f))^{\frac{1}{p}}}. \tag{3.4}$$

THEOREM 3.1. *Let $p \geq 1$. Assume that $f : \operatorname{dom} f \to \mathbb{R}$ is uniformly convex of degree $p+1$, and its pth derivative is Lipschitz continuous. Then, for any $\alpha \geq 0$, and all $x, y \in \operatorname{dom} f$, we have*

$$\langle \nabla f(y) - \nabla f(x), y-x \rangle \geq \sum_{k=2}^{p} \frac{\beta^{k-1}}{(k-1)!} D^k f(x)[y-x]^k + \frac{\alpha L_p \beta^p}{p!}\|y-x\|^{p+1}, \tag{3.5}$$

$$\begin{aligned} f(y) \geq\ & f(x) + \langle \nabla f(x), y-x \rangle + \sum_{k=2}^{p} \frac{\beta^{k-1}}{k!} D^k f(x)[y-x]^k \\ & + \frac{\alpha L_p \beta^p}{(p+1)!}\|y-x\|^{p+1}, \end{aligned} \tag{3.6}$$

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

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

$$\begin{aligned} \langle \nabla f(y), y-x \rangle &= \frac{1}{1-t}\langle \nabla f(y), y - z_t \rangle \\ &\stackrel{(3.1)}{\geq} \frac{1}{1-t}\langle \nabla f(z_t), y - z_t \rangle + (1-t)^p \sigma_{p+1}(f)\|y-x\|^{p+1} \\ &= \langle \nabla f(z_t), y-x \rangle + (1-t)^p \sigma_{p+1}(f)\|y-x\|^{p+1}. \end{aligned}$$

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

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

Let us choose now t from the inequality

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

Then it is enough to take $t \overset{(3.4)}{\leq} \beta_p(f, \alpha)$. Hence, inequality (3.5) is proved.

The remaining inequality (3.6) can be proved by integration. Indeed

$$\begin{aligned} & f(y) - f(x) - \langle \nabla f(x), y - x\rangle \\ = \;& \int_0^1 \tfrac{1}{\tau} \langle \nabla f(x + \tau(y-x)) - f(x), \tau(y-x)\rangle d\tau \\ \overset{(3.5)}{\geq} \;& \int_0^1 \left(\sum_{k=2}^{p} \tfrac{\beta^{k-1}\tau^{k-1}}{(k-1)!} D^k f(x)[y-x]^k + \tfrac{\alpha L_p(f) \beta^p \tau^p}{p!} \|y-x\|^{p+1} \right) d\tau \\ = \;& \sum_{k=2}^{p} \tfrac{\beta^{k-1}}{k!} D^k f(x)[y-x]^k + \tfrac{\alpha L_p(f) \beta^p}{(p+1)!} \|y-x\|^{p+1}. \end{aligned}$$

Thus we obtain (3.6). ∎

Remark 3.2. For $\alpha \geq p$, the right-hand side of inequality (3.6) 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:

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

Since $\alpha \geq p$, it is convex in z in view of Theorem 1 in [46].

For the optimization schemes developed in this paper, inequality (3.6) serves as the main justification tool. Let us present now the general model of our optimization problems.

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

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

Our assumptions on function F are as follows.

Assumption 4.1. 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 $u \in \mathcal{D}(x)$ in a componentwise ordering, i.e., for any $u, v \in \mathcal{D}(x)$ such that for all $i (u_i \leq v_i)$, it holds that

$$F(x, u) \;\leq\; F(x, v).$$

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

Assumption 4.2. All components of function f are closed and convex.

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

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

$$\boxed{\varphi^* \;=\; \min_{x \in \operatorname{dom}\varphi} \left\{ \varphi(x) \stackrel{\text{def}}{=} F(x, f(x)) \right\},} \tag{4.1}$$

where $\operatorname{dom}\varphi = \{x \in \operatorname{dom} f \;:\; (x, f(x)) \in \operatorname{dom} F\}$. We denote by x^* a solution to problem (4.1): $\varphi(x^*) = \varphi^*$, assuming that it exists.

Of course, problem (4.1) is tractable only if the function F is simple. Thus, we require that the structure of function F is *simple enough* for allowing an efficient solution to 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:

$$\min_{x \in Q}\{f_1(x) :\; f_i(x) \le 0,\; i = 2, \dots, m\}, \tag{4.2}$$

where Q is a closed convex set and f satisfies Assumption 4.2. Then this problem can be written in form (4.1) with

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

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

2. *Additive composite minimization* [43]. Consider the following minimization problem:

$$\min_x\{f_1(x) + \psi(x)\}, \tag{4.3}$$

where $f(x) \equiv \{f_1(x)\}$ satisfies Assumption 4.2 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., [39, 40]). Minimization problem

$$\min_x F(f(x)), \tag{4.4}$$

where F is a closed convex monotone function with $\operatorname{dom} F \subseteq \mathbb{R}^m$, and f satisfies Assumption 4.2, is clearly in the form (4.1).

4. *Functional and additive composition.* Note that monotonicity in the first argument of the outer part is not assumed. Combining two previous examples, we can consider the problems of the form (e.g., [26]):

$$\min_x\big\{F(f(x)) + \psi(x)\big\}.$$

5. *Composition with linear mapping.* Note that for any fully composite problem (4.1) and for arbitrary linear operator $A : \mathbb{E} \to \mathbb{E}$, the following problem is also fully composite:

$$\min_x F(Ax, f(x)).$$

For a vector-valued function $f(x) = (f_1(x), \dots, f_m(x))^T : \operatorname{dom} f \to \mathbb{R}^m$ whose components are p-times differentiable, we denote its directional derivative as the vector of directional derivatives of its components:

$$D^p f(x)[h]^p \overset{\text{def}}{=} \left(D^p f_1(x)[h]^p, \dots, D^p f_m(x)[h]^p\right)^T \in \mathbb{R}^m, \quad x \in \operatorname{dom} f, \quad h \in \mathbb{E}.$$

Since the Taylor polynomial (2.2) is defined in terms of directional derivatives, we can extend its meaning onto vector-valued 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 $\hat{\beta}_p(f) = \min_{1 \le i \le m} \beta_p(f_i, p)$. Then inequality (3.6) can be rewritten in a vector form:

$$f(y) \ge f(x) + \sum_{k=1}^{p} \tfrac{\beta^{k-1}}{k!} D^k f(x)[y-x]^k + \tfrac{pL_p(f)\beta^p}{(p+1)!}\|y-x\|^{p+1} \tag{4.5}$$

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:

$$F\left(y, f(x) + \sum_{k=1}^{p} \tfrac{\beta^{k-1}}{k!} D^k f(x)[y-x]^k + \tfrac{pL_p(f)\beta^p}{(p+1)!}\|y-x\|^{p+1}\right) \to \min_y, \tag{4.6}$$

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

5. Basic high-order optimization methods. Let $\bar{x} \in \operatorname{dom} \varphi$. For inequality (3.6), let us choose $\alpha = p$ and $\beta \in (0, \hat{\beta}_p(f)]$. Then

$$\begin{aligned} (1-\beta)\varphi(\bar{x}) + \beta\varphi^* &= \min_{v \in \operatorname{dom}\varphi} \left[(1-\beta)F(\bar{x}, f(\bar{x})) + \beta F(v, f(v))\right] \\ &\ge \min_{v \in \operatorname{dom}\varphi} F\Big((1-\beta)\bar{x} + \beta v, (1-\beta)f(\bar{x}) + \beta f(v)\Big) \\ &\overset{(4.5)}{\ge} \min_{v \in \operatorname{dom}\varphi} F\Big((1-\beta)\bar{x} + \beta v, f(\bar{x}) + \sum_{k=1}^{p} \tfrac{\beta^k}{k!} D^k f(\bar{x})[v-\bar{x}]^k + \tfrac{pL_p(f)\beta^{p+1}}{(p+1)!}\|v-\bar{x}\|^{p+1}\Big) \\ &= \min_{\substack{y = \bar{x} + \beta(v-\bar{x}) \\ v \in \operatorname{dom}\varphi}} F\Big(y,\ \Omega_p(f, \bar{x}; y) + \tfrac{pL_p(f)}{(p+1)!}\|y-\bar{x}\|^{p+1}\Big) \overset{\text{def}}{=} \tilde{M}^*_{p,\beta}(\bar{x}). \end{aligned}$$

By Remark 3.2, the second argument in the objective function of the latter problem is a componentwise convex function. Hence, by Assumption 4.1, this objective is convex in y.

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

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

Note that in view of Assumption 4.1, we have

$$\begin{aligned}\tilde{M}^*_{p,\beta}(\bar{x}) &= F\left(\tilde{y}^*_{p,\beta}(\bar{x}),\ \Omega_p(f,\bar{x};\tilde{y}^*_{p,\beta}(\bar{x})) + \tfrac{pL_p(f)}{(p+1)!}\|\tilde{y}^*_{p,\beta}(\bar{x})-\bar{x}\|^{p+1}\right)\\ &\overset{(2.6)}{\geq} F\left(\tilde{y}^*_{p,\beta}(\bar{x}), f(\tilde{y}^*_{p,\beta}(\bar{x}))\right) = \varphi\left(\tilde{y}^*_{p,\beta}(\bar{x})\right).\end{aligned}$$

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

(5.1)

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:** $x_{k+1} = \tilde{y}^*_{p,\beta}(x_k)$.

We have proved for this method the following theorem.

THEOREM 5.1. *Assume that all components of f are uniformly convex of degree $p+1$, and pth derivatives of its components are Lipschitz continuous. Let sequence $\{x_k\}_{k\geq 0}$ be generated by method* (5.1). *Then for all $k \geq 0$ we have*

$$\varphi(x_k) - \varphi^* \leq (1-\beta)^k(\varphi(x_0) - \varphi^*). \tag{5.2}$$

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

Note that method (5.1) could move with bigger steps. Indeed,

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

Denote

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

By the same reasons as before, $M^*_p(\bar{x}) \geq \varphi\left(y^*_p(\bar{x})\right)$. Hence, we can estimate the rate of convergence of the following method:

(5.3)

Full-step pth order basic method
Choose $x_0 \in \operatorname{dom}\varphi$.
For $k \geq 0$ **iterate:** $x_{k+1} = y^*_p(x_k)$.

THEOREM 5.2. *Assume that all components of f are uniformly convex of degree $p+1$, and pth derivatives of its components are Lipschitz continuous. Let sequence $\{x_k\}_{k\geq 0}$ be generated by* (5.3). *Then for all $k \geq 0$ we have*

$$\varphi(x_k) - \varphi^* \leq (1-\hat{\beta}_p(f))^k(\varphi(x_0) - \varphi^*). \tag{5.4}$$

In view of potentially bigger steps, method (5.3) may be faster in practice. For $p = 1$, this method was studied recently in [3], where the global rate of convergence of order $O(1/k)$ was established for a general convex case.

Example 5.3. Let us look at implementation of method (5.3) for a particular class of optimization problems with functional constraints (4.2) with $Q = \mathbb{E}$. This is

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

Then, for $p = 1$, each iteration of method (5.3) 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

$$\max_{\substack{\lambda \in \mathbb{R}_+^m \\ \lambda^{(1)} = 1}} \left\{ \sum_{i=1}^m \lambda^{(i)} f_i(\bar{x}) - \tfrac{1}{2 \sum_{i=1}^m \lambda^{(i)} L_1(f_i)} \|g(\lambda)\|_*^2 \right\}. \tag{5.5}$$

Note that this is the moving ball approximation algorithm introduced in [2].

On the other hand, for $p = 2$, one iteration of method (5.3) is as follows:

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

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

$$\max_{\substack{\lambda \in \mathbb{R}_+^m, \tau \in \mathbb{R}_+ \\ \lambda^{(1)} = 1}} \left\{ \sum_{i=1}^m \lambda^{(i)} f_i(\bar{x}) - \tfrac{\tau^3}{6 [\sum_{i=1}^m \lambda^{(i)} L_2(f_i)]^2} - \tfrac{1}{2} \langle H(\lambda, \tau)^{-1} g(\lambda), g(\lambda) \rangle \right\}. \tag{5.6}$$

Note that typically the dimension of problems (5.5) and (5.6) is not big. Hence, they can be solved, for example, by the interior-point methods [47] very efficiently.

6. General regularization scheme. In the previous sections, we discussed two methods for solving problem (4.1) 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 5 to a special *regularized* problem.

Let us present a general regularization framework for fully composite problem (4.1). For that, we use a componentwise convex regularizing vector function

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

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

$$d(x_0) \;\; = \;\; f(x_0), \tag{6.1}$$

$$d(x) \; \ge \; f(x) \quad \forall x \in \operatorname{dom} f. \tag{6.2}$$

A simple possibility for such a regularizer is the following choice.

Example 6.1. For $1 \le i \le m$,

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

with a certain $c_i > 0$.

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

$$\varphi_\mu(x) \;=\; F(x, (1-\mu)f(x) + \mu d(x)), \quad x \in \operatorname{dom}\varphi. \tag{6.3}$$

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

Assumption 6.2. Vector function $d(x) - f(x)$ is componentwise convex on the domain of f.

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

Note that $\varphi_\mu(x_0) \overset{(6.1)}{=} \varphi(x_0)$. Clearly, for all $x \in \operatorname{dom}\varphi$ we have

$$\varphi(x) \overset{(6.2)}{\le} \varphi_\mu(x). \tag{6.4}$$

Our regularized problem looks now as follows:

$$\boxed{\varphi_\mu^* \;=\; \min_{x \in \operatorname{dom}\varphi} \varphi_\mu(x).} \tag{6.5}$$

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 5. Namely, assume that, for certain $\delta > 0$, we have a point $\bar{x} \in \operatorname{dom}\varphi$, satisfying the following inequality:

$$\begin{aligned} \varphi_\mu(\bar{x}) - \varphi_\mu^* \;&\le\; \delta\left(\varphi_\mu(x_0) - \varphi_\mu^*\right) \overset{(6.1)}{=} \delta\left(\varphi(x_0) - \varphi_\mu^*\right) \\ &\overset{(6.4)}{\le} \delta\left(\varphi(x_0) - \varphi^*\right). \end{aligned} \tag{6.6}$$

We need to understand now how good this point is for our initial problem (4.1).

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

$$\xi_A(x; g) \;=\; \inf_{\lambda > 0} \left\{ \lambda :\; f(x) + \tfrac{1}{\lambda} g \in \mathcal{D}(x),\; F\Big(x, f(x) + \tfrac{1}{\lambda} g\Big) \le A \right\}. \tag{6.7}$$

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

LEMMA 6.3. *Suppose that, for some points x_0 and $\bar{x}$ from $\operatorname{dom}\varphi$, we have $\varphi(x_0) \le A$ and $\varphi(\bar{x}) \le A$. Suppose that the local measure is finite at x_0: $\xi_A(x_0; g) < +\infty$. Then, for any $g \in \mathbb{R}^m_+$ and any coefficient $\tau \in [0, 1]$, it holds that*

$$\xi_A((1-\tau)x_0 + \tau\bar{x}; g) \;\le\; \tfrac{1}{1-\tau}\, \xi_A(x_0; g). \tag{6.8}$$

Proof. Let $\lambda > 0$ be such that $f(x_0) + \frac{1}{\lambda} g \in \mathcal{D}(x_0)$ and $F(x_0, f(x_0) + \frac{1}{\lambda} g) \le A$. Since $f(\bar{x}) \in \mathcal{D}(\bar{x})$, we have

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

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

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

At the same time,

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

Thus, $\xi_A((1-\tau)x_0+\tau\bar{x}; g) \le \frac{\lambda}{1-\tau}$. To complete the proof, it remains to take the minimum in λ. ∎

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

LEMMA 6.4. *Let $A \ge \varphi(x_0)$ and the regularizing function $d(\cdot)$ satisfy Assumption* 6.2 *and conditions* (6.1), (6.2). *Assume that parameters α and τ from $[0,1]$ are chosen as follows:*

$$\frac{\tau}{1-\tau} \ \le\ \frac{\alpha}{\mu\,\xi_0^*}. \tag{6.9}$$

Then

$$\varphi_\mu^* - \varphi^* \ \le\ (1-\tau)(1-\alpha)(\varphi(x_0)-\varphi^*)+\alpha(A-\varphi^*). \tag{6.10}$$

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

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

Let us choose now $x = (1-\tau)x_0+\tau x^*$ with arbitrary $\tau \in [0,1]$. Then in view of Assumption 6.2, we have

$$\begin{aligned} g_x\ &\le\ (1-\tau)(d(x_0)-f(x_0))+\tau(d(x^*)-f(x^*)) \\ &\overset{(6.1)}{=}\ \tau(d(x^*)-f(x^*))\ =\ \tau g_*. \end{aligned} \tag{6.11}$$

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

$$\frac{\mu\tau}{\alpha}\ \le\ \frac{1-\tau}{\xi_0^*}\ \overset{(6.8)}{\le}\ \frac{1}{\xi_A(x;g^*)}.$$

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

$$\varphi_\mu^* - \varphi^*\ \le\ (1-\tau)(1-\alpha)(\varphi(x_0)-\varphi^*)+\alpha(A-\varphi^*),$$

which completes the proof. ∎

Let us put in (6.10) the best values of parameters. If τ satisfies (6.9) 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 $A \ge \varphi(x_0)$, we get the following estimate:

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

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

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

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

$$\tfrac{\beta(1-\alpha_*)}{\beta+\alpha_*} + \alpha_* \;=\; \tfrac{2\beta}{\beta+\sqrt{\beta+\beta^2}} \;\le\; 2\sqrt{\beta}.$$

In other words, we get the following bound:

$$\varphi_\mu^* - \varphi^* \;\le\; 2\sqrt{\mu\xi_0^*}\,(A-\varphi^*). \tag{6.12}$$

Therefore, if we have an approximate solution $\bar{x}$ to the regularized problem, which satisfies (6.6), we can ensure the bound for the original problem

$$\begin{aligned}
\varphi(\bar{x}) - \varphi^* &\overset{(6.12)}{\le} \varphi(\bar{x}) - \varphi_\mu^* + 2\sqrt{\mu\xi_0^*}(A-\varphi^*) \\
&\overset{(6.4)}{\le} \varphi_\mu(\bar{x}) - \varphi_\mu^* + 2\sqrt{\mu\xi_0^*}(A-\varphi^*) \\
&\overset{(6.6)}{\le} \delta(\varphi(x_0)-\varphi^*) + 2\sqrt{\mu\xi_0^*}(A-\varphi^*),
\end{aligned}$$

and the regularization parameter should be of the following order:

$$\boxed{\mu \;\approx\; \tfrac{\delta^2}{\xi_0^*}.}$$

Now, let us consider the regularizer from Example 6.1:

$$d_i(x) \;:=\; f_i(x) + \tfrac{c_i}{p+1}\|x-x_0\|^{p+1}, \qquad c_i > 0.$$

This function is uniformly convex of degree $p+1$ with parameter $\sigma_{p+1}(d_i) = \tfrac{c_i}{2^{p-1}}$ (see, e.g., Lemma 2.5 in [24]). Moreover, its pth derivative is Lipschitz continuous with constant $L_p(d_i) = L_p(f_i) + c_i \cdot p!$ (see Theorem 7.1 in [52]).

Hence, applying method (5.3) to the regularized objective, we obtain the linear rate

$$\varphi_\mu(x_k) - \varphi_\mu^* \overset{(5.4)}{\le} (1-\beta)^k(\varphi_\mu(x_0) - \varphi_\mu^*),$$

where the condition number is equal to

$$\beta \;=\; \hat{\beta}_p((1-\mu)f + \mu d) \;=\; \min_{1\le i\le m} \tfrac{(p!\gamma_p((1-\mu)f_i+\mu d_i))^{\frac{1}{p}}}{(1+p)^{\frac{1}{p}} + (p!\gamma_p((1-\mu)f_i+\mu d_i))^{\frac{1}{p}}} \tag{6.13}$$

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

$$\beta \;=\; \Big[1 + \big((1+p)2^{p-1}\big(\tfrac{1}{\mu p!}+1\big)\big)^{\frac{1}{p}}\Big]^{-1}.$$

Thus, parameter μ plays a crucial role in the complexity of regularized problem (6.5).

7. 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 necessarily uniformly convex. At the same time, we demonstrate that our methods can be accelerated.

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

$$f(\gamma x) \quad \leq \quad \gamma f(x). \tag{7.1}$$

THEOREM 7.1. *A closed and convex function $f(\cdot)$ is subhomogeneous if and only if it satisfies one of the following three conditions:*

$$\langle g_x, x\rangle \quad \leq \quad f(x), \qquad x \in \operatorname{int}(\operatorname{dom} f), \quad g_x \in \partial f(x), \tag{7.2}$$

$$\langle g_y, x\rangle \quad \leq \quad f(x), \qquad x, y \in \operatorname{int}(\operatorname{dom} f), \quad g_y \in \partial f(y), \tag{7.3}$$

$$f(x+ty) \quad \leq \quad f(x) + tf(y), \qquad x, y \in \operatorname{int}(\operatorname{dom} f),\ x + ty \in \operatorname{dom} f, \quad t \geq 0. \tag{7.4}$$

Proof. Assume that (7.1) is true. Then

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

and this is (7.2).

Assume (7.2) is true. Since f is convex, for any $x, y \in \operatorname{int}(\operatorname{dom} f)$ and $g_y \in \partial f(y)$, we have

$$f(x) - \langle g_y, x\rangle \quad \geq \quad f(y) - \langle g_y, y\rangle \quad \overset{(7.2)}{\geq} \quad 0.$$

This is relation (7.3).

Finally, assume that (7.3) is true. For any $y \in \operatorname{dom} f$ denote by $g(y)$ a particular subgradient in $\partial f(y)$. Then we have, by applying the mean-value theorem (see, e.g., Theorem 2.3.4 in [33]), for any $x \in \operatorname{int}(\operatorname{dom} f)$ and $\gamma > 1$ such that $\gamma x \in \operatorname{dom} f$,

$$f(\gamma x) \quad = \quad f(x) + \textstyle\int_0^{\gamma-1} \langle g(x+\tau x), x\rangle d\tau \quad \overset{(7.3)}{\leq} \quad f(x) + \int_0^{\gamma-1} f(x) d\tau \quad = \quad f(x).$$

And this is (7.1). Thus, conditions (7.1), (7.2), and (7.3) are equivalent.

In order to justify equivalence with (7.4), note that it can be rewritten as

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

Therefore,

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

and this is (7.3). On the other hand, if (7.3) is true, then

$$f(x+ty) - f(x) \quad = \quad \textstyle\int_0^t \langle g(x+\tau y), y\rangle d\tau \quad \overset{(7.3)}{\leq} \quad \int_0^t f(y) d\tau \quad = \quad tf(y).$$

And this is (7.4). ∎

Example 7.2. Clearly, function $f(x) = \max_{i=1}^n x^{(i)}$ is subhomogeneous.

Example 7.3. Consider the following function:

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

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

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

since $\eta(u) \leq 0$ for all $u \in \Delta_n$. Thus, function $f(\cdot)$ is subhomogeneous since condition (7.2) is satisfied.

Example 7.4. Let function f be subhomogeneous. Then, for a linear operator A, function

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

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

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

with additional normalizing constraint $\tau = 1$.

Remark 7.5. Let $F : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function. Assume that F is monotone on its domain. Consequently, with any point $x \in \operatorname{dom} F$, we have

$$x - \mathbb{R}^m_+ \quad \subseteq \quad \operatorname{dom} F.$$

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

$$0 \quad \in \quad \operatorname{int}(\operatorname{dom} F). \tag{7.5}$$

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

$$F(x + ty) \quad \leq F(x) + tF(y). \tag{7.6}$$

Combining (7.5) and (7.6), we conclude $\operatorname{dom} F = \mathbb{R}^m$.

Now, let us introduce our additional assumption.

Assumption 7.6. For any $x \in \operatorname{dom} \varphi$, the function $F(x, u)$ is subhomogeneous in $u \in \mathbb{R}^m$. Thus, for any $x \in \operatorname{dom} \varphi$, it holds that

$$F(x, u + tv) \quad \leq \quad F(x, u) + tF(x, v), \quad u, v \in \mathbb{R}^m, \quad t \geq 0. \tag{7.7}$$

We are ready to analyze convergence of the full-step pth-order basic method (5.3) in a general convex case, when uniform convexity is absent. Let us use the following notation:

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

THEOREM 7.7. *Let the initial level set be bounded:*

$$D_0 \quad \stackrel{\text{def}}{=} \quad \sup_x \Big\{ \|x - x^*\| \ : \ \varphi(x) \leq \varphi(x_0) \Big\} \quad < \quad +\infty. \tag{7.8}$$

Assume that pth derivatives of the components of f are Lipschitz continuous. Then, for the iterations $\{x_k\}_{k \geq 1}$ of method (5.3)*, we have*

$$\varphi(x_k) - \varphi^* \quad \leq \quad \frac{(p+1)^{p+1} F(L_p(f)) D_0^{p+1}}{p!} \cdot k^{-p}. \tag{7.9}$$

Proof. By the definition of the method step, we have

$$\begin{aligned}
\varphi(x_{k+1}) &= F(x_{k+1}, f(x_{k+1})) \\
&\overset{(2.6)}{\leq} F\Big(x_{k+1}, \Omega_p(f, x_k; x_{k+1}) + \tfrac{pL_p(f)}{(p+1)!}\|x_{k+1} - x_k\|^{p+1}\Big) \\
&\leq F\Big(y, \Omega_p(f, x_k; y) + \tfrac{pL_p(f)}{(p+1)!}\|y - x_k\|^{p+1}\Big) \\
&\overset{(2.6)}{\leq} F\Big(y, f(y) + \tfrac{L_p(f)}{p!}\|y - x_k\|^{p+1}\Big) \\
&\overset{(7.7)}{\leq} \varphi(y) + \tfrac{F(L_p(f))}{p!}\|y - x_k\|^{p+1}
\end{aligned} \tag{7.10}$$

for all $y \in \operatorname{dom}\varphi$. 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 $A_k := k \cdot (k+1) \cdot \cdots \cdot (k+p)$, and $a_{k+1} := A_{k+1} - A_k = \frac{p+1}{k+p+1}A_{k+1}$. Therefore, we obtain by convexity

$$\begin{aligned}
\varphi(x_{k+1}) &\leq \tfrac{a_{k+1}}{A_{k+1}}\varphi^* + \tfrac{A_k}{A_{k+1}}\varphi(x_k) + \big(\tfrac{a_{k+1}}{A_{k+1}}\big)^{p+1} \cdot \tfrac{F(L_p(f))\|x^* - x_k\|^{p+1}}{p!} \\
&\leq \tfrac{a_{k+1}}{A_{k+1}}\varphi^* + \tfrac{A_k}{A_{k+1}}\varphi(x_k) + \big(\tfrac{p+1}{k+p+1}\big)^{p+1} \cdot \tfrac{F(L_p(f))D_0^{p+1}}{p!}.
\end{aligned} \tag{7.11}$$

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

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

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

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

This is (7.9). ∎

8. 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). Recall that for a vector-valued function $f(x) = (f_1(x), \ldots, f_m(x))^T \in \mathbb{R}^m$, we use the following notation for the directional derivative:

$$\langle \nabla f(x), h\rangle \equiv (\langle \nabla f_1(x), h\rangle, \ldots, \langle \nabla f_m(x), h\rangle)^T \in \mathbb{R}^m, \quad x \in \operatorname{dom} f, \quad h \in \mathbb{E}.$$

We start with the basic scheme. The methods of this type have been extensively studied in recent decades (see [9, 41, 16, 35]).

Basic gradient method (8.1)

Choose $x_0 \in \operatorname{dom}\varphi$ and $M = \alpha F(L_1(f))$, $\alpha \geq 1$.

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

$$x_{k+1} = \operatorname*{argmin}_{y \in \operatorname{dom}\varphi}\Big\{F\big(y, f(x_k) + \langle \nabla f(x_k), y - x_k\rangle\big) + \tfrac{M}{2}\|y - x_k\|^2\Big\}.$$

Contrary to the scheme (5.3), the regularization term in method (8.1) is *outside* of the composite part. Therefore, an implementation of each step can be much simpler.

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

$$\begin{aligned}
\varphi(x_{k+1}) &\overset{(2.6)}{\leq} F\Big(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \tfrac{L_1(f)}{2}\|x_{k+1} - x_k\|^2\Big) \\
&\overset{(7.7)}{\leq} F\Big(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle\Big) + \tfrac{\alpha F(L_1(f))}{2}\|x_{k+1} - x_k\|^2 \\
&\overset{(8.1)}{\leq} F\Big(y, f(x_k) + \langle \nabla f(x_k), y - x_k\rangle\Big) + \tfrac{\alpha F(L_p(f))\|y - x_k\|^2}{2} \\
&\leq \varphi(y) + \tfrac{\alpha F(L_p(f))\|y - x_k\|^2}{2},
\end{aligned}$$

where we used componentwise convexity of f and monotonicity of F in the last inequality. By the same arguments as in the proof of Theorem 7.7, we get the following result.

THEOREM 8.1. *Let the initial level set be bounded* (7.8)*. Assume that the gradients of the components of f are Lipschitz continuous. Then, for the sequence $\{x_k\}_{k\geq 1}$ generated by method* (8.1)*, it holds that*

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

For the problems with bounded domain, we can propose the following alternative scheme, which is a generalization of the classical Frank–Wolfe algorithm [27, 44]. 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 (8.1).

(8.2)

Contracting conditional gradient method

Choose $x_0 \in \operatorname{dom}\varphi$ and $\{\gamma_k\}_{k\geq 0}$.

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

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ F\big(y, f(x_k) + \langle \nabla f(x_k), y - x_k\rangle\big) : x_k + \tfrac{1}{\gamma_k}(y - x_k) \in \operatorname{dom}\varphi \Big\}.$$

THEOREM 8.2. *Let $\operatorname{dom}\varphi$ be a bounded convex set. Denote its diameter by*

$$\mathscr{D} \overset{\text{def}}{=} \sup_{x,y\in\operatorname{dom}\varphi} \|x - y\| < +\infty. \tag{8.3}$$

Assume that the gradients of the components of f are Lipschitz continuous. Set $\gamma_k := \frac{2}{k+2}$. Then, for the iterations $\{x_k\}_{k\geq 1}$ of method (8.2)*, it holds that*

$$\varphi(x_k) - \varphi^* \leq \tfrac{4F(L_1(f))\mathscr{D}^2}{k}. \tag{8.4}$$

Proof. Let us denote the point $v_{k+1} \overset{\text{def}}{=} x_k + \frac{1}{\gamma_k}(x_{k+1} - x_k) \overset{(8.2)}{\in} \operatorname{dom}\varphi$. Hence,

$$\|x_{k+1} - x_k\| \;=\; \gamma_k \|v_{k+1} - x_k\| \;\le\; \gamma_k \mathscr{D}. \tag{8.5}$$

Now, considering one iteration of the method, we obtain

$$\begin{aligned}
\varphi(x_{k+1}) &\overset{(2.6)}{\le} F\Big(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \tfrac{L_1(f)}{2}\|x_{k+1} - x_k\|^2\Big) \\
&\overset{(7.7)}{\le} F\Big(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle\Big) + \tfrac{F(L_1(f))}{2}\|x_{k+1} - x_k\|^2 \\
&\overset{(8.5)}{\le} F\Big(x_{k+1}, f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle\Big) + \tfrac{\gamma_k^2 F(L_1(f))\mathscr{D}^2}{2} \\
&\overset{(8.2)}{\le} F\Big(y, f(x_k) + \langle \nabla f(x_k), y - x_k\rangle\Big) + \tfrac{\gamma_k^2 F(L_1(f))\mathscr{D}^2}{2} \\
&\le \varphi(y) + \tfrac{\gamma_k^2 F(L_1(f))\mathscr{D}^2}{2}
\end{aligned}$$

for every $y \in \operatorname{dom}\varphi$. Substituting $y := \gamma_k x^* + (1 - \gamma_k)x_k$, and using convexity of φ, we obtain an inequality very similar to (7.11) from the proof of Theorem 7.7 for $p = 1$. Hence, by the same arguments, we establish the rate (8.4). ∎

Finally, we can present an accelerated method (see [39, 45, 19]).

Fully composite fast gradient method (8.6)

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

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

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

THEOREM 8.3. *Assume that the gradients of the components of f are Lipschitz continuous. Then, for the sequence $\{x_k\}_{k\ge 1}$, generated by method* (8.6)*, it holds that*

$$A_k\varphi(x_k) + \tfrac{M}{2}\|x - v_k\|^2 \;\le\; A_k\varphi(x) + \tfrac{M}{2}\|x - x_0\|^2, \qquad x \in \operatorname{dom}\varphi. \tag{8.7}$$

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

$$\varphi(x_k) - \varphi^* \;\le\; \tfrac{M\|x^* - x_0\|^2}{2A_k} \;\le\; \tfrac{2M\|x^* - x_0\|^2}{k^2}. \tag{8.8}$$

Proof. Let us establish (8.7) by induction. Assume that it holds for the current iterate, and consider the next step. We fix an arbitrary $x \in \operatorname{dom}\varphi$ and denote the

convex combination $y := \frac{a_{k+1}x + A_k x_k}{A_{k+1}}$. Then

$$
\begin{aligned}
\tfrac{M}{2}\|x - x_0\|^2 + A_{k+1}\varphi(x) &= \tfrac{M}{2}\|x - x_0\|^2 + A_k\varphi(x) + a_{k+1}\varphi(x) \\
&\overset{(8.7)}{\geq} \tfrac{M}{2}\|x - v_k\|^2 + A_k\varphi(x_k) + a_{k+1}\varphi(x) \\
&\geq \tfrac{M}{2}\|x - v_k\|^2 + A_{k+1}\varphi(y) = A_{k+1}\Big(\tfrac{M}{2}\|y - y_k\|^2 + \varphi(y)\Big) \\
&\geq A_{k+1}\Big(\tfrac{M}{2}\|y - y_k\|^2 + F\big(y, f(y_k) + \langle \nabla f(y_k), y - y_k\rangle\big)\Big),
\end{aligned} \tag{8.9}
$$

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

The function in the right-hand side of (8.9) is *strongly convex* in y. Hence, we obtain

$$
\begin{aligned}
&\tfrac{M}{2}\|y - y_k\|^2 + F\big(y, f(y_k) + \langle \nabla f(y_k), y - y_k\rangle\big) \\
&\overset{(8.6)}{\geq} \tfrac{M}{2}\|y - x_{k+1}\|^2 + \tfrac{M}{2}\|x_{k+1} - y_k\|^2 + F\big(x_{k+1}, f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle\big) \\
&\overset{(7.7)}{\geq} \tfrac{M}{2}\|y - x_{k+1}\|^2 + F\big(x_{k+1}, f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \tfrac{L_1(f)}{2}\|x_{k+1} - y_k\|^2\big) \\
&\overset{(2.6)}{\geq} \tfrac{M}{2}\|y - x_{k+1}\|^2 + \varphi(x_{k+1}) = \tfrac{M}{2A_{k+1}}\|x - v_{k+1}\|^2 + \varphi(x_{k+1}).
\end{aligned}
$$

Thus we establish (8.7) 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 (8.8). ∎

9. 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 [48].

Fully composite cubic Newton method (9.1)

Choose $x_0 \in \operatorname{dom}\varphi$ and $M = \alpha F(L_2(f))$, $\alpha \geq 1$.

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

$$
x_{k+1} = \operatorname*{argmin}_{y \in \operatorname{dom}\varphi}\Big\{F\big(y, \Omega_2(f, x_k; y)\big) + \tfrac{M}{6}\|y - x_k\|^3\Big\}.
$$

Iterations of this type are simpler than those in the general method (5.3) with $p = 2$ since the regularization term is *outside* the composite part. At the same time, for any

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

$$
\begin{aligned}
\varphi(x_{k+1}) &\overset{(2.6)}{\leq} F\big(x_{k+1}, \Omega_2(f, x_k; x_{k+1}) + \tfrac{L_2(f)}{6}\|x_{k+1} - x_k\|^3\big) \\
&\overset{(7.7)}{\leq} F\big(x_{k+1}, \Omega_2(f, x_k; x_{k+1})\big) + \tfrac{M}{6}\|x_{k+1} - x_k\|^3 \\
&\overset{(9.1)}{\leq} F\big(y, \Omega_2(f, x_k; y)\big) + \tfrac{M}{6}\|y - x_k\|^3 \\
&\overset{(2.6)}{\leq} F\big(y, f(y) + \tfrac{L_2(f)}{6}\|y - x_k\|^3\big) + \tfrac{M}{6}\|y - x_k\|^3 \\
&\overset{(7.7)}{\leq} \varphi(y) + \tfrac{(1+\alpha)F(L_2(f))}{6}\|y - x_k\|^3.
\end{aligned} \tag{9.2}
$$

Hence, using the same reasoning as in Theorem 7.7, we prove the global convergence result.

THEOREM 9.1. *Let the initial level set be bounded* (7.8)*. Assume that the Hessians of the components of f are Lipschitz continuous. Then, for the sequence $\{x_k\}_{k\geq 1}$ generated by method* (9.1)*, we have*

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

In papers [22, 20], 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 (8.2).

Fully composite contracting Newton method (9.3)

Choose $x_0 \in \operatorname{dom}\varphi$ and $\{\gamma_k\}_{k\geq 0}$.

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

$$
x_{k+1} \;=\; \operatorname*{argmin}_y \Big\{ F\big(y, \Omega_2(f, x_k; y)\big) \;:\; x_k + \tfrac{1}{\gamma_k}(y - x_k) \in \operatorname{dom}\varphi \Big\}.
$$

Repeating the previous reasoning, we obtain the following result.

THEOREM 9.2. *Let the size of* $\operatorname{dom}\varphi$ *be bounded by diameter* $\mathscr{D}$ (8.3)*. Assume that the Hessians of the components of f are Lipschitz continuous. Define $\gamma_k := \frac{3}{k+3}$. Then, for the sequence $\{x_k\}_{k\geq 1}$, generated by method* (9.3)*, we have*

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

10. 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 [42], 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 [21, 31, 36]) 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 [23]).

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

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

It is well known that this function is *uniformly convex* of degree 3 (see, e.g., [24]), so it holds that

$$\rho_d(x; y) \;\stackrel{\text{def}}{=}\; d(y) - d(x) - \langle \nabla d(x), y - x\rangle \;\geq\; \tfrac{\alpha}{6}\|y - x\|^3, \qquad x, y \in \mathbb{E}. \tag{10.1}$$

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

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

$$h(x) \;\geq\; h(T) + \rho_d(T; x), \qquad x \in \mathbb{E}. \tag{10.2}$$

   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 that

$$\rho_d(\bar{v}; x) \;=\; \rho_d(v; x) + \rho_d(\bar{v}; v) + \langle \nabla d(v) - \nabla d(\bar{v}), x - v\rangle. \tag{10.3}$$

We use this prox function in the following general method.

Fully composite contracting proximal-point scheme (10.4)

Choose $x_0 \in \operatorname{dom}\varphi$ 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 objective $h_{k+1}(x) = A_{k+1}\varphi(\gamma_k x + (1 - \gamma_k)x_k) + \rho_d(v_k; x)$.
3. Find a point v_{k+1} s.t. $h_{k+1}(v_{k+1}) - h^*_{k+1} \leq \delta$ by basic method (9.1).
4. $x_{k+1} = \gamma_k v_{k+1} + (1 - \gamma_k)x_k$.

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

THEOREM 10.1. *Assume that the Hessians of the components of f are Lipschitz continuous. Then, for the iterations of method* (10.4)*, we have*

$$A_k\varphi(x_k) + \rho_d(v_k; x) \;\leq\; A_k\varphi(x) + \rho_d(x_0; x) + C_k(x), \qquad x \in \mathbb{E}, \tag{10.5}$$

where

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

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

$$\begin{aligned} \delta \quad &\geq \quad h_{k+1}(v_{k+1}) - h_{k+1}(v_{k+1}^*) \quad \stackrel{(10.2)}{\geq} \quad \rho_d(v_{k+1}^*; v_{k+1}) \\ &\stackrel{(10.1)}{\geq} \quad \tfrac{\alpha}{6}\|v_{k+1} - v_{k+1}^*\|^3. \end{aligned}$$

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

$$\|v_{k+1} - v_{k+1}^*\| \quad \leq \quad \left(\tfrac{6\delta}{\alpha}\right)^{\frac{1}{3}}. \tag{10.6}$$

Now, assume that (10.5) holds 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

$$\begin{aligned} \rho_d(x_0; x) + A_{k+1}\varphi(x) \quad &= \quad \rho_d(x_0; x) + A_k\varphi(x) + a_{k+1}\varphi(x) \\ &\stackrel{(10.5)}{\geq} \quad \rho_d(v_k; x) + A_k\varphi(x_k) + a_{k+1}\varphi(x) - C_k(x) \\ &\geq \quad \rho_d(v_k; x) + A_{k+1}\varphi(\gamma_k x + (1-\gamma_k)x_k) - C_k(x) \\ &= \quad h_{k+1}(x) - C_k(x) \\ &\stackrel{(10.2)}{\geq} \quad \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}^*) - C_k(x), \end{aligned}$$

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:

$$\begin{aligned} \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}^*) \quad &\stackrel{\text{Step 3}}{\geq} \quad \rho_d(v_{k+1}^*; x) + h_{k+1}(v_{k+1}) - \delta \\ &\stackrel{(10.3)}{=} \quad \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - \delta + \rho_d(v_{k+1}^*; v_{k+1}) \\ &\qquad + \langle \nabla d(v_{k+1}) - \nabla d(v_{k+1}^*), x - v_{k+1} \rangle \\ &\geq \quad \rho_d(v_{k+1}; x) + h_{k+1}(v_{k+1}) - \delta \\ &\qquad - \|\nabla d(v_{k+1}) - \nabla d(v_{k+1}^*)\|_* \cdot \|x - v_{k+1}\|. \end{aligned}$$

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:

$$\begin{aligned}
&\|\nabla d(v_{k+1}) - \nabla d(v^*_{k+1})\|_* \\
&\quad = \quad \| \textstyle\int_0^1 \nabla^2 d(v^*_{k+1} + \tau(v_{k+1} - v^*_{k+1}))d\tau)(v_{k+1} - v^*_{k+1})\|_* \\
&\quad \le \quad 2\alpha \cdot \|v_{k+1} - v^*_{k+1}\| \cdot \textstyle\int_0^1 \|v^*_{k+1} + \tau(v_{k+1} - v^*_{k+1}) - x_0\| d\tau \\
&\quad \le \quad 2\alpha \cdot \|v_{k+1} - v^*_{k+1}\| \cdot \Big(\|v_{k+1} - x_0\| + \tfrac{1}{2}\|v_{k+1} - v^*_{k+1}\|\Big) \\
&\quad \le \quad 2\alpha \cdot \|v_{k+1} - v^*_{k+1}\| \cdot \|x - v_{k+1}\| \\
&\qquad\qquad + 2\alpha \cdot \|v_{k+1} - v^*_{k+1}\| \cdot \|x - x_0\| \\
&\qquad\qquad + \alpha \cdot \|v_{k+1} - v^*_{k+1}\|^2 \\
&\quad \overset{(10.6)}{\le} \quad 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - v_{k+1}\| \;+\; 2\alpha^{2/3}(6\delta)^{1/3} \cdot \|x - x_0\| \\
&\qquad\qquad + \alpha^{1/3}(6\delta)^{2/3}.
\end{aligned}$$

Therefore, combining all components together, we conclude that

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

Thus, (10.5) is justified for all $k \ge 0$. ∎

Let us substitute into (10.5) $x = x^*$ and fix some $K \ge 0$. For all $0 \le k \le K$, we get

$$\begin{aligned}
&\tfrac{\alpha}{6}\|x^* - v_k\|^3 \overset{(10.1)}{\le} \rho_d(v_k; x^*) + A_k(\varphi(x_k) - \varphi^*) \\
&\overset{(10.5)}{\le} \rho_d(x_0; x^*) + K\delta + c_1(x^*) \sum_{i=1}^k \|x^* - v_i\| + c_2 \sum_{i=1}^k \|x^* - v_i\|^2 \;\overset{\text{def}}{=}\; R_k,
\end{aligned} \tag{10.7}$$

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

$$\begin{aligned}
R_{k+1} \quad &= \quad R_k + c_1(x^*)\|x^* - v_{k+1}\| + c_2\|x^* - v_{k+1}\| \\
&\overset{(10.7)}{\le} \quad R_k + aR_{k+1}^{1/3} + bR_{k+1}^{2/3}, \qquad \text{where}
\end{aligned} \tag{10.8}$$

$$\begin{aligned}
a \;&\overset{\text{def}}{=}\; c_1(x^*) \cdot \left(\tfrac{6}{\alpha}\right)^{1/3} \;=\; 6\delta^{2/3} + 12\delta^{1/3}\rho_d(x_0; x^*)^{1/3}, \\
b \;&\overset{\text{def}}{=}\; c_2 \cdot \left(\tfrac{6}{\alpha}\right)^{2/3} \;=\; 12\delta^{1/3}.
\end{aligned} \tag{10.9}$$

Dividing (10.8) 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,

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

It can be resolved as follows:

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

Hence, telescoping the last inequality, we get

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

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

COROLLARY 10.2. *For the iterations of method* (10.4)*, for all* $k \geq 1$*, it holds that*

(10.10)
$$A_k(\varphi(x_k) - \varphi^*) + \rho_d(v_k; x^*) \quad \leq \quad O\Bigl(\rho_d(x_0; x^*) + k^3\delta + k^3\sqrt{\delta\rho_d(x_0; x^*)}\Bigr).$$

Hence,

$$\varphi(x_k) - \varphi^* \quad \leq \quad O\Bigl(\tfrac{\rho_d(x_0;x^*)}{k^3} + \delta + \sqrt{\delta\rho_d(x_0; x^*)}\Bigr).$$

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

(10.11)
$$\delta \quad \approx \quad \min\Bigl\{\varepsilon,\ \tfrac{\varepsilon^2}{\rho_d(x_0;x^*)}\Bigr\}.$$

Let us apply now the basic cubic Newton method (9.1) 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

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

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

(10.12)
$$h_{k+1}(x) \quad \equiv \quad \bar{F}(x, \bar{f}(x)) \quad \to \quad \min_{x\in\operatorname{dom}\varphi}.$$

Then iterations of method (9.1) applied to (10.12) are

$$z_{t+1} \quad := \quad \operatorname*{argmin}_{x\in\operatorname{dom}\varphi}\Bigl\{\bar{F}(x, \Omega_2(\bar{f}, z_t; x)) + \tfrac{\bar{F}(L_2(\bar{f}))}{6}\|x - z_t\|^3\Bigr\}, \qquad t \geq 0,$$

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

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

(10.13)
$$\begin{aligned}\bar{F}(L_2(\bar{f})) \quad &\leq \quad A_{k+1}F(L_2(\bar{f})) \quad \overset{(7.1)}{\leq} \quad \gamma_k^3 A_{k+1}F(L_2(f)) \\ &= \quad \tfrac{a_{k+1}^3}{A_{k+1}^2}F(L_2(f)) \quad = \quad \tfrac{((k+1)^3-k^3)^3}{3^3(k+1)^6}F(L_2(f)) \quad \leq \quad F(L_2(f)).\end{aligned}$$

The guarantee (9.2) of one step ensures that, for any $x \in \operatorname{dom}\varphi$, it holds that

(10.14)
$$\begin{aligned}h_{k+1}(z_{t+1}) \quad &\leq \quad h_{k+1}(x) + \tfrac{\bar{F}(L_2(\bar{f}))}{3}\|x - z_t\|^3 \\ &\overset{(10.13)}{\leq} \quad h_{k+1}(x) + \tfrac{F(L_2(f))}{3}\|x - z_t\|^3.\end{aligned}$$

Let $x = \tau v^*_{k+1} + (1-\tau)z_t$ with $\tau := \frac{1}{\sqrt{6}}$ and v^*_{k+1} being the minimizer of (10.12). Then

$$\begin{aligned} h_{k+1}(z_{t+1}) &\leq \tau h^*_{k+1} + (1-\tau)h_{k+1}(z_t) + \tfrac{\tau^3 F(L_2(f))}{3}\|v^*_{k+1} - z_t\|^3 \\ &\overset{(10.1)}{\leq} \tau h^*_{k+1} + (1-\tau)h_{k+1}(z_t) + 2\tau^3\big(h_{k+1}(z_t) - h^*_{k+1}\big). \end{aligned}$$

Therefore,

$$\begin{aligned} h_{k+1}(z_{t+1}) - h^*_{k+1} &\leq \Big(1 - \tau + 2\tau^3\Big)\cdot\big(h_{k+1}(z_t) - h^*_{k+1}\big) \\ &= \Big(1 - \tfrac{2}{3\sqrt{6}}\Big)\cdot\big(h_{k+1}(z_t) - h^*_{k+1}\big) \leq \tfrac{3}{4}\big(h_{k+1}(z_t) - h^*_{k+1}\big). \end{aligned}$$

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 (10.14), we get

$$\begin{aligned} & h_{k+1}(z_1) - h^*_{k+1} \\ &\overset{(10.14)}{\leq} h_{k+1}(x^*) - h^*_{k+1} + \tfrac{F(L_2(f))}{3}\|x^* - v_k\|^3 \\ &\leq a_{k+1}\varphi^* + A_k\varphi(x_k) + \rho_d(v_k; x^*) - h^*_{k+1} + \tfrac{F(L_2(f))}{3}\|x^* - v_k\|^3 \\ &\overset{(*)}{\leq} A_k(\varphi(x_k) - \varphi^*) + \rho_d(v_k; x^*) + 2\rho_d(v_k; x^*) \\ &\overset{(10.10),(10.11)}{\leq} O\Big(\rho_d(x_0; x^*) + k^3\varepsilon\Big), \end{aligned}$$

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

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

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

COROLLARY 10.3. *For solving the initial problem with ε-accuracy,*

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

we need to perform $K = O(\big[\frac{F(L_2(f))\|x_0 - x^*\|^3}{\varepsilon}\big]^{1/3})$ iterations of the proximal-point scheme (10.4)*. At each iteration, it requires no more than*

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

steps of the basic method (9.1)*.*

We see that the price to pay for the level of generality is an additional logarithmic term in the final complexity estimate. It remains an open theoretical question whether we can develop a *direct* accelerated high-order method for the fully composite formulation, which does not need inexact proximal iterations. It would also help in constructing *optimal* high-order methods [37, 28, 34, 8], matching the existing lower complexity bounds [1, 45].

Another interesting research direction is the development of *universal* [29, 30] and *randomized* [17, 32] variants of the fully composite methods.

Acknowledgments. We are very thankful to the associate editor and two anonymous referees for valuable comments that significantly improved the initial version of this paper.

REFERENCES

[1] Y. Arjevani, O. Shamir, and R. Shiff, *Oracle complexity of second-order methods for smooth convex optimization*, Math. Program., 178 (2019), pp. 327–360.
[2] A. Auslender, R. Shefi, and M. Teboulle, *A moving balls approximation method for a class of smooth constrained minimization problems*, SIAM J. Optim., 20 (2010), pp. 3232–3259.
[3] J. Bolte, Z. Chen, and E. Pauwels, *The multiproximal linearization method for convex composite problems*, Math. Program., 182 (2020), pp. 1–36.
[4] R. I. Boţ, S.-M. Grad, and G. Wanka, *New constraint qualification and conjugate duality for composed convex optimization problems*, J. Optim. Theory Appl., 135 (2007), pp. 241–255.
[5] R. I. Boţ, S.-M. Grad, and G. Wanka, *A new constraint qualification for the formula of the subdifferential of composed convex functions in infinite dimensional spaces*, Math. Nachr., 281 (2008), pp. 1088–1107.
[6] R. I. Boţ, S.-M. Grad, and G. Wanka, *Generalized Moreau–Rockafellar results for composed convex functions*, Optimization, 58 (2009), pp. 917–933.
[7] R. I. Boţ, I. B. Hodrea, and G. Wanka, *Farkas-type results for inequality systems with composed convex functions via conjugate duality*, J. Math. Anal. Appl., 322 (2006), pp. 316–328.
[8] S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford, *Near-optimal method for highly smooth convex optimization*, in Proceedings of the Conference on Learning Theory, PMLR, 2019, pp. 492–507.
[9] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.
[10] J. V. Burke, *Second order necessary and sufficient conditions for convex composite NDO*, Math. Program., 38 (1987), pp. 287–302.
[11] J. V. Burke and A. Engle, *Strong metric (sub) regularity of Karush–Kuhn–Tucker mappings for piecewise linear-quadratic convex-composite optimization and the quadratic convergence of Newton's method*, Math. Oper. Res., 45 (2020), pp. 1164–1192.
[12] J. V. Burke and M. C. Ferris, *A Gauss—Newton method for convex composite optimization*, Math. Program., 71 (1995), pp. 179–194.
[13] J. V. Burke and T. Hoheisel, *Epi-convergent smoothing with applications to convex composite functions*, SIAM J. Optim., 23 (2013), pp. 1457–1479.
[14] J. V. Burke and R. Poliquin, *Optimality conditions for non-finite valued convex composite functions*, Math. Program., 57 (1992), pp. 103–120.
[15] J. V. Burke, H. Tim, and Q. V. Nguyen, *A study of convex convex-composite functions via infimal convolution with applications*, Math. Oper. Res., 46 (2021), pp. 1324–1348.
[16] C. Cartis, N. I. Gould, and P. L. Toint, *On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming*, SIAM J. Optim., 21 (2011), pp. 1721–1739.
[17] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Math. Program., 169 (2018), pp. 337–375.
[18] Y. Cui, J.-S. Pang, and B. Sen, *Composite difference-max programs for modern statistical estimation problems*, SIAM J. Optim., 28 (2018), pp. 3344–3374.
[19] A. d'Aspremont, D. Scieur, and A. Taylor, *Acceleration Methods*, preprint, arXiv:2101.09545, 2021.
[20] N. Doikov and Y. Nesterov, *Affine-Invariant Contracting-Point Methods for Convex Optimization*, CORE Discussion Papers 2020/29, 2020.
[21] N. Doikov and Y. Nesterov, *Contracting proximal methods for smooth convex optimization*, SIAM J. Optim., 30 (2020), pp. 3146–3169.
[22] N. Doikov and Y. Nesterov, *Convex optimization based on global lower second-order models*, Advances in Neural Information Processing Systems, 33 (2020), pp. 16546–16556.
[23] N. Doikov and Y. Nesterov, *Inexact tensor methods with dynamic accuracies*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 2577–2586.
[24] N. Doikov and Y. Nesterov, *Minimizing uniformly convex functions by cubic regularization of Newton method*, J. Optim. Theory Appl., 189 (2021), pp. 317–339.
[25] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948.

[26] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Math. Program., 178 (2019), pp. 503–558.
[27] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.
[28] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. A. Uribe, *Optimal tensor methods in smooth convex and uniformly convexoptimization*, in Proceedings of the Conference on Learning Theory, PMLR, 2019, pp. 1374–1391.
[29] G. Grapiglia and Y. Nesterov, *Regularized Newton methods for minimizing functions with Hölder continuous Hessians*, SIAM J. Optim., 27 (2017), pp. 478–506.
[30] G. Grapiglia and Y. Nesterov, *Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives*, SIAM J. Optim., 30 (2020), pp. 2750–2779.
[31] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim., 2 (1992), pp. 649–664.
[32] F. Hanzely, N. Doikov, P. Richtarik, and Y. Nesterov, *Stochastic subspace cubic Newton method*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 4027–4038.
[33] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, New York, 2004.
[34] B. Jiang, H. Wang, and S. Zhang, *An optimal high-order tensor method for convex optimization*, in Proceedings of the Conference on Learning Theory, PMLR, 2019, pp. 1799–1801.
[35] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, Math. Program., 158 (2016), pp. 501–546.
[36] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Advances in Neural Information Processing Systems, 2015, pp. 3384–3392.
[37] R. D. Monteiro and B. F. Svaiter, *An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods*, SIAM J. Optim., 23 (2013), pp. 1092–1125.
[38] A. Nemirovski, *Information-Based Complexity of Convex Programming*, Lecture notes, Technion – Israel Institute of Technology, 1995.
[39] Y. Nesterov, *A method for solving the convex programming problem with convergence rate O(1/kˆ2)*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.
[40] Y. Nesterov, *Effective Methods in Nonlinear Programming*, Radio i Svyaz, Moscow, 1989.
[41] Y. Nesterov, *Modified Gauss–Newton scheme with worst case guarantees for global performance*, Optim. Methods Softw., 22 (2007), pp. 469–483.
[42] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181.
[43] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[44] Y. Nesterov, *Complexity bounds for primal-dual methods minimizing the model of objective function*, Math. Program., 171 (2018), pp. 311–330.
[45] Y. Nesterov, *Lectures on Convex Optimization*, Springer Optim. Appl. 137, Springer, New York, 2018.
[46] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Math. Program., 186 (2021), pp. 157–183.
[47] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[48] Y. Nesterov and B. Polyak, *Cubic regularization of Newton's method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[49] T. Pennanen, *Graph-convex mappings and k-convex functions*, J. Convex Anal., 6 (1999), pp. 235–266.
[50] M. J. Powell, *Algorithms for nonlinear constraints that use Lagrangian functions*, Math. Program., 14 (1978), pp. 224–248.
[51] M. J. Powell, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Springer, New York, 1978, pp. 144–157.
[52] A. Rodomanov and Y. Nesterov, *Smoothness parameter of power of Euclidean norm*, J. Optim. Theory Appl., 185 (2020), pp. 303–326.