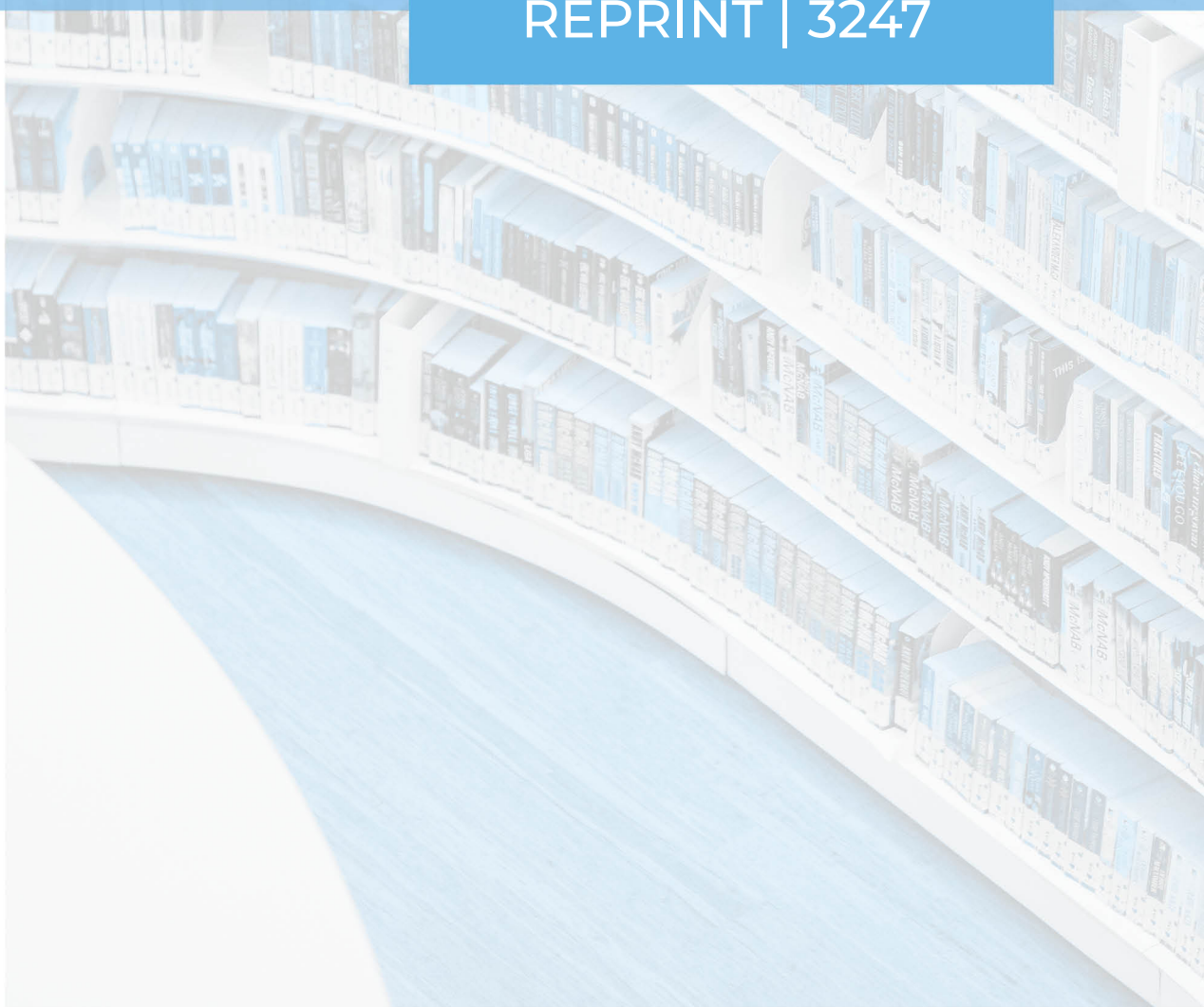
CONTRACTING PROXIMAL METHODS FOR SMOOTH CONVEX OPTIMIZATION

Nikita Doikov, Yurii Nesterov

REPRINT | 3247



UCLouvain

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

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email:lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html

SIAM J. OPTIM.
Vol. 30, No. 4, pp. 3146–3169
© 2020 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

CONTRACTING PROXIMAL METHODS FOR SMOOTH CONVEX OPTIMIZATION*

NIKITA DOIKOV† AND YURII NESTEROV‡

Abstract. In this paper, we propose new accelerated methods for smooth convex optimization, called contracting proximal methods. At every step of these methods, we need to minimize a contracted version of the objective function augmented by a regularization term in the form of Bregman divergence. We provide global convergence analysis for a general scheme admitting inexactness in solving the auxiliary subproblem. In the case of using for this purpose high-order tensor methods, we demonstrate an acceleration effect for both convex and uniformly convex composite objective functions. Thus, our construction explains acceleration for methods of any order starting from one. The augmentation of the number of calls of oracle due to computing the contracted proximal steps is limited by the logarithmic factor in the worst-case complexity bound.

Key words. convex optimization, proximal method, accelerated methods, global complexity bounds, high-order algorithms

AMS subject classifications. 49M15, 49M37, 65K05, 90C25, 90C30

DOI. 10.1137/19M130769X

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

1. Introduction.

One of the classical iterative methods in theoretical optimization is the proximal point algorithm [24]. This method, as applied to minimizing a convex function $f : \operatorname{dom} f \to \mathbb{R}$, consists of solving at each iteration the following subproblem:

$$x_{k+1} = \operatorname*{argmin}_x \left\{ a_{k+1} f(x) + \frac{1}{2}\|x_k - x\|^2 \right\}, \qquad k \geq 0, \tag{1.1}$$

where $\|\cdot\|$ is the standard Euclidean norm, and $\{a_k\}_{k\geq 1}$ is a sequence of positive coefficients. In general, we can hope only to use an inexact solution of the subproblem (1.1) (see [10, 27, 26] for the convergence analysis). An important observation is that the regularized objective in (1.1) is *strongly convex*. Therefore, we can hope that computing an (inexact) proximal step is usually simpler than solving the initial problem.

For a function $f \in \mathscr{F}_L^{1,1}$ (convex differentiable functions with Lipschitz continuous gradients), we can set all values of the coefficients a_k equal to a positive constant. This gives a global sublinear rate of convergence of the iterations (1.1) in functional residual of the order $O(1/k)$. This rate is the same rate as that of the gradient method [21].

For the same class of functions, we can get a faster rate of convergence of the order $O(1/k^2)$ using the accelerated gradient method [18]. This is the best possible rate achievable for the first-order black-box optimization on $\mathscr{F}_L^{1,1}$ [17]. An accelerated

*Received by the editors December 18, 2019; accepted for publication (in revised form) August 5, 2020; published electronically November 10, 2020.
https://doi.org/10.1137/19M130769X
Funding: The research results of this paper were obtained in the framework of ERC Advanced Grant 788368.

†Institute of Information and Communication Technologies, Electronics and Applied Mathematics (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).

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

variant of the proximal point algorithm with the optimal rate of convergence was proposed in [11] (see also [25, 13, 14, 12] for extensions and some applications).

In this paper, we present a new family of proximal-type algorithms for smooth convex optimization called *contracting proximal methods*, which includes an accelerated algorithm from [11] as a particular case and provides a systematic way for constructing faster proximal accelerated methods for high-order optimization. Thus, for the class of convex functions, whose pth derivative is Lipschitz continuous ($p \geq 1$), our new methods achieve the $O(1/k^{p+1})$-rate of convergence for the outer proximal iterations, while the inner subproblems can be efficiently solved up to desired accuracy by the high-order tensor methods [22]. Note that this rate can also be achieved by a direct acceleration scheme, utilizing the notion of estimating sequences [2, 22]. It can be improved up to the level $O(1/k^{\frac{3p+1}{2}})$ by using a special line-search on each iteration [16, 7]. The latter rate was shown to be the optimal one [1].

The main difference between contracting proximal methods and the classical approach (1.1) consists in employing the *contracted* objective function (which provides the methods with their name) and the *Bregman divergence* (notation $\beta_d(x;y)$) instead of the usual Euclidean norm. The exact form of our method is very simple:

$$\left.\begin{array}{l} v_{k+1} = \operatorname{argmin}_x \left\{ A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + \beta_d(v_k; x) \right\} \\ x_{k+1} = \frac{a_{k+1}v_{k+1} + A_k x_k}{A_{k+1}} \end{array}\right\}, \quad k \geq 0. \tag{1.2}$$

Thus, we use a sequence of auxiliary points $\{v_k\}_{k\geq 0}$ and the scaling coefficients $A_k \stackrel{\text{def}}{=} \sum_{i=1}^k a_i$.

Let us illustrate the basic idea behind this construction by the simplest *Euclidean setting*, when $\beta_d(x;y) \equiv \frac{1}{2}\|x-y\|^2$. We are going to ensure at each iteration $k \geq 0$ the following condition:

$$\frac{1}{2}\|x_0 - x\|^2 + A_k f(x) \geq \frac{1}{2}\|v_k - x\|^2 + A_k f(x_k), \qquad x \in \operatorname{dom} f. \tag{1.3}$$

A direct consequence of (1.3) is the global convergence bound

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2A_k}. \tag{1.4}$$

We can propagate inequality (1.3) to the next iteration by a trivial observation:

$$\begin{aligned} \frac{1}{2}\|x_0 - x\|^2 + A_{k+1} f(x) &= \frac{1}{2}\|x_0 - x\|^2 + A_k f(x) + a_{k+1} f(x) \\ &\overset{(1.3)}{\geq} \frac{1}{2}\|v_k - x\|^2 + A_k f(x_k) + a_{k+1} f(x) \\ &\geq \frac{1}{2}\|v_k - x\|^2 + A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) \equiv h_{k+1}(x), \end{aligned}$$

where the last inequality is due to convexity of the objective. Note that the first step of contracting proximal method (1.2) is defined exactly as follows:

$$v_{k+1} = \operatorname*{argmin}_{x \in \mathbb{E}} h_{k+1}(x). \tag{1.5}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Hence, by strong convexity of $h_{k+1}(\cdot)$, we finally justify that

$$h_{k+1}(x) \geq h_{k+1}(v_{k+1}) + \frac{1}{2}\|v_{k+1} - x\|^2 \geq A_{k+1} f(x_{k+1}) + \frac{1}{2}\|v_{k+1} - x\|^2.$$

Thus, for the Euclidean setting, iteration (1.2) immediately results in the convergence guarantee (1.4). However, we are still free in the choice of coefficients $\{a_k\}_{k\geq 1}$. The only reason for bounding their growth consists in keeping the complexity of the optimization problem (1.5) at an acceptable level.[1] For $f \in \mathscr{F}_L^{1,1}$, the recommended choice of a_{k+1} corresponds to the quadratic equation [18]:

$$a_{k+1}^2 = \frac{1}{L}(a_{k+1} + A_k). \tag{1.6}$$

It is easy to see that this choice results in the optimal $O(1/k^2)$-rate of convergence for the method. On the other hand, it makes the *condition number* of the problem (1.5) equal to an absolute constant. Let us assume for simplicity that f is two times continuously differentiable. Then, in view of the presence of the regularization term, $\nabla^2 h_{k+1}(x) \succeq I$. On the other hand,

$$\nabla^2 h_{k+1}(x) = I + \frac{a_{k+1}^2}{A_{k+1}} \nabla^2 f\big(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\big) \overset{(1.6)}{\preceq} 2I.$$

Hence, we are able to solve the problem (1.5) very efficiently by a usual gradient method (see the details in section 4).

It is remarkable that exactly the same reasoning justifies the accelerated versions of *all* high-order tensor methods ($p \geq 2$). The only difference consists in the degree of the proximal term, which must be compatible with the order of optimization scheme used for solving the problem (1.5).

Our first-order contracting proximal method for a Euclidean setting (described above) produces the same sequence of points as the accelerated proximal point algorithm from [11]. However, now we can also employ the Bregman divergence, which sometimes is more suitable to the topology of our function and ensures faster convergence.

The rest of the paper is organized as follows. Section 2 introduces notation used throughout the paper and describes our problem of interest in the composite form. We also give a definition of Bregman divergence and mention some of its properties.

In section 3, we introduce a general contracting proximal method (formulated as section 3). We present its convergence analysis for a problem in composite form and arbitrary Bregman divergence. We study both convex and strongly convex cases under inexactness in proximal steps. Theorem 3.2 specifies how the parameters of the algorithm and inner accuracy affect the convergence rate.

In section 4, we discuss implementation of one iteration of our method, under the assumption that the pth derivative ($p \geq 1$) of the smooth part of the objective is Lipschitz continuous. We present a fully defined optimization scheme (section 4), with incorporated steps of the tensor method of a certain degree. The resulting algorithm achieves the accelerated rate of convergence, with an additional logarithmic factor for the number of total oracle calls. The final complexity estimate for this scheme is given by Theorems 4.6 and 4.7.

Section 5 contains numerical experiments. Section 6 has some final remarks.

[1] Hence, these bounds should take into account the efficiency of the auxiliary minimization scheme used for solving the problem (1.5).

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

2. Notation. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space and by $\mathbb{E}^*$ its dual space, which is a space of linear functions on $\mathbb{E}$. The value of function $s \in \mathbb{E}^*$ at point $x \in \mathbb{E}$ is denoted by $\langle s, x \rangle$.

Let us fix some arbitrary (possibly non-Euclidean) norm $\|\cdot\|$ on space $\mathbb{E}$ and define the dual norm $\|\cdot\|_*$ on $\mathbb{E}^*$ in the standard way:

$$\|s\|_* \stackrel{\text{def}}{=} \sup_{h \in \mathbb{E}} \{\langle s, h \rangle : \|h\| \le 1\}.$$

For a smooth function f, its gradient at point x is denoted by $\nabla f(x)$, and its Hessian is $\nabla^2 f(x)$. Note that

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

Higher derivatives are denoted as $D^p f(x)[\cdot]$, which are p-linear symmetric forms on $\mathbb{E}$, and the norm is induced:

$$\|D^p f(x)\| \stackrel{\text{def}}{=} \sup_{h_1, \dots, h_p \in \mathbb{E}} \left\{ D^p f(x)[h_1, \dots, h_p] : \; \|h_i\| \le 1, i = 1, \dots, p \right\}.$$

For convex but not necessary differentiable function ψ, we denote by $\partial \psi(x) \subset \mathbb{E}^*$ its subdifferential at point $x \in \operatorname{dom} \psi$.

Our goal is to solve the following composite minimization problem:

$$\min_{x \in \operatorname{dom} F} \Big\{ F(x) \equiv f(x) + \psi(x) \Big\}, \tag{2.1}$$

where f is several times differentiable on its open domain convex function, with some reasonable assumptions on the growth of its derivatives (for example, that its pth derivative is Lipschitz continuous for some $p \ge 1$), and $\psi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function, which we assume to be *simple*, but possibly nondifferentiable, with $\operatorname{dom} \psi \subset \operatorname{dom} f$. We also assume that solution $x^* \in \operatorname{dom} F$ of problem (2.1) does exist, denoting $F^* = F(x^*)$.

Let us fix arbitrary differentiable strictly convex function $d : \operatorname{dom} \psi \to \mathbb{R}$, which we call the *prox function*. Then, we denote by $\beta_d(x; y)$ the corresponding *Bregman divergence*, centered at x:

$$\beta_d(x; y) \stackrel{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

We say that function d is *uniformly convex* of degree $p + 1$ (with respect to the norm $\|\cdot\|$) with constant $\sigma_{p+1}(d) > 0$ if it holds for all $x, y \in \operatorname{dom} d$:

$$\beta_d(x; y) \ge \frac{\sigma_{p+1}(d)}{p+1} \|x - y\|^{p+1}. \tag{2.2}$$

The main example, which naturally appears in tensor methods (see [22]) and which we use in section 4, is the following prox function.

Example 2.1.

$$d(x) \equiv \frac{1}{p+1} \|x - x_0\|^{p+1}$$

for some $p \ge 1$. For a Euclidean norm (when $\|x\| \equiv \langle Bx, x \rangle^{1/2}$ for a fixed postive-definite linear operator $B = B^* \succ 0$) this prox function is *uniformly convex* of degree $p + 1$ with constant 2^{1-p} (see [5, Lemma 5]), so it holds that

$$\beta_d(x; y) \ge \frac{2^{1-p}}{p+1} \|x - y\|^{p+1}, \qquad x, y \in \mathbb{E}. \tag{2.3}$$

For more examples of available prox functions see [3, 15].

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

The definition of Bregman divergence can be extended onto nondifferentiable function ψ by specifying a particular subgradient $\psi'(x) \in \partial\psi(x)$:

$$\beta_\psi(x, \psi'(x); y) \stackrel{\text{def}}{=} \psi(y) - \psi(x) - \langle \psi'(x), y - x \rangle.$$

However, we will use simpler notation $\beta_\psi(x; y)$ if no ambiguity arises.

We say that function ψ is *strongly convex with respect to* d (see [28, 3, 15]) with constant $\sigma_d(\psi) > 0$ if it holds for all $x, y \in \operatorname{dom} \psi$ and for all $\psi'(x) \in \partial\psi(x)$ that

$$\beta_\psi(x, \psi'(x); y) \geq \sigma_d(\psi)\beta_d(x; y). \tag{2.4}$$

Inequality (2.4) always holds with $\sigma_d(\psi) = 0$ just by convexity. An interesting illustration of this concept is given by a regularized Taylor polynomial of degree 3 for convex function (see [22]).

Example 2.2. Let $f : \operatorname{dom} f \to \mathbb{R}$ be convex, with Lipschitz continuous third derivative:

$$\|D^3 f(y) - D^3 f(x)\| \leq L_3 \|y - x\|, \qquad x, y \in \operatorname{dom} f.$$

Denote by $\Omega_3(f, x; y)$ its Taylor approximation of degree 3 around some fixed point x,

$$\Omega_3(f, x; y) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} D^2 f(x)[y - x]^2 + \frac{1}{6} D^3 f(x)[y - x]^3,$$

and consider its regularization of degree 4, with some $\tau > 1$:

$$g(y) \equiv \Omega_3(f, x; y) + \frac{\tau^2 L_3}{8} \|y - x\|^4.$$

Then, for a Euclidean norm, the function $g(\cdot)$ is strongly convex with respect to the following prox function (see [22, Lemma 4]):

$$d(h) \equiv \frac{1}{2} \left(1 - \frac{1}{\tau}\right) D^2 f(x)[h]^2 + \frac{\tau(\tau - 1) L_3}{8} \|h\|^4.$$

Let us summarize some basic properties of Bregman divergence, which follow directly from its definition. For any pair f_1, f_2 of convex functions and all $x, y \in \operatorname{dom}(f_1 + f_2)$ we have

$$\beta_{a_1 f_1 + a_2 f_2}(x; y) = a_1 \beta_{f_1}(x; y) + a_2 \beta_{f_2}(x; y), \qquad a_1, a_2 \geq 0. \tag{2.5}$$

For any linear function $\ell(x) = a + \langle g, x \rangle$ we have

$$\beta_\ell(x; y) = 0. \tag{2.6}$$

Therefore, from (2.5) and (2.6) we conclude that

$$\beta_f(x; y) = \beta_d(x; y), \tag{2.7}$$

when $f(y) = \beta_d(z; y)$ for some fixed z. Now, consider the following simple but general construction, which we use in a core of our analysis. Let h be a regularized composite objective:

$$h(y) = g(y) + a\psi(y) + \gamma \beta_d(z; y), \qquad a, \gamma \geq 0,$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

where g and ψ are arbitrary closed convex functions, and ψ is strongly convex with respect to d with some constant $\sigma_d(\psi) \geq 0$. Then we have, for every $x, y \in \operatorname{dom} h$ and every $h'(x) \in \partial h(x)$

$$\begin{aligned} h(y) - h(x) - \langle h'(x), y - x \rangle &= \beta_h(x; y) \\ &\overset{(2.5),(2.7)}{=} \beta_g(x; y) + a\beta_\psi(x; y) + \gamma\beta_d(x; y) \\ &\geq (a\sigma_d(\psi) + \gamma)\beta_d(x; y). \end{aligned} \tag{2.8}$$

In particular, for the exact minimum $T = \operatorname*{argmin}_{y \in \mathbb{E}} h(y)$, we have

$$h(y) \geq h(T) + (a\sigma_d(\psi) + \gamma)\beta_d(T; y). \tag{2.9}$$

3. Contracting proximal method. In our general scheme, we are going to maintain the following inequality, for every $x \in \operatorname{dom} \psi$ and $k \geq 0$:

$$\gamma_0\beta_d(x_0; x) + A_k F(x) \geq \gamma_k\beta_d(v_k; x) + A_k F(x_k) + C_k(x), \tag{3.1}$$

where $\{x_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$ are sequences of points from $\operatorname{dom} \psi$, $\{A_k\}_{k\geq 0}$ is a sequence of increasing numbers,

$$a_{k+1} \overset{\text{def}}{=} A_{k+1} - A_k \ > \ 0, \qquad A_0 = 0,$$

and $\{\gamma_k\}_{k\geq 0}$ is a sequences of nondecreasing proximal coefficients,

$$\gamma_{k+1} \geq \gamma_k, \qquad \gamma_0 > 0.$$

We would prefer to have functions $C_k(x)$ as big as possible. Thus, if it happens to be $C_k(x^*) \geq 0$ for all $k \geq 1$, then from (3.1) we have a convergence guarantee,

$$F(x_k) - F^* \leq \frac{\gamma_0\beta_d(x_0, x^*)}{A_k}, \qquad k \geq 1,$$

and the rate of convergence is determined by the growth of coefficients A_k toward infinity. However, generally $C_k(x)$ may have arbitrary sign.

Let us discus a simple possibility for propagating relation (3.1) to the next iteration.

(3.2)

$$\begin{aligned} &\gamma_0\beta_d(x_0; x) + A_{k+1}F(x) \\ &\quad = \ \gamma_0\beta_d(x_0; x) + A_k F(x) + a_{k+1}F(x) \\ &\quad \overset{(3.1)}{\geq} \ \gamma_k\beta_d(v_k; x) + A_k F(x_k) + a_{k+1}F(x) + C_k(x) \\ &\quad \geq \ \gamma_k\beta_d(v_k; x) + A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + A_k\psi(x_k) + C_k(x), \end{aligned}$$

where the last inequality is due to the convexity of f. Let us consider a contracted objective with a regularizer from the last step:

$$h_{k+1}(x) \overset{\text{def}}{=} A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + \gamma_k\beta_d(v_k; x). \tag{3.3}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

This function is strongly convex with respect to $d(\cdot)$ with parameter

$$\sigma_d(h_{k+1}) \geq \gamma_{k+1} \stackrel{\text{def}}{=} a_{k+1}\sigma_d(\psi) + \gamma_k. \tag{3.4}$$

If we are able to compute the *exact minimum*

$$T = \operatorname*{argmin}_{x \in \mathbb{E}} h_{k+1}(x), \tag{3.5}$$

then by (2.9) we see that

$$\begin{aligned} &h_{k+1}(x) + A_k\psi(x_k) \\ &\quad \geq h_{k+1}(T) + \gamma_{k+1}\beta_d(T;x) + A_k\psi(x_k) \\ &\quad = A_{k+1}f\left(\frac{a_{k+1}T + A_kx_k}{A_{k+1}}\right) + a_{k+1}\psi(T) + \gamma_k\beta_d(v_k;T) + \gamma_{k+1}\beta_d(T;x) + A_k\psi(x_k) \\ &\quad \geq A_{k+1}F\left(\frac{a_{k+1}T + A_kx_k}{A_{k+1}}\right) + \gamma_k\beta_d(v_k;T) + \gamma_{k+1}\beta_d(T;x). \end{aligned}$$

And it is natural to set $v_{k+1} = T$ and

$$x_{k+1} \stackrel{\text{def}}{=} \frac{a_{k+1}v_{k+1} + A_kx_k}{A_{k+1}}. \tag{3.6}$$

Thus we would obtain guarantee (3.1) for the next step, with

$$C_{k+1}(x) \equiv C_k(x) + \gamma_k\beta_d(v_k;v_{k+1}) \equiv \sum_{i=1}^{k} \gamma_i\beta_d(v_i;v_{i+1}) \geq 0.$$

Now, instead of computing the exact minimum (3.5), let us relax $v_{k+1} \in \operatorname{dom}\psi$ to be a point with a *small norm of subgradient*:

$$\|s\|_* \leq \delta_{k+1} \quad \text{for some} \quad s \in \partial h_{k+1}(v_{k+1}). \tag{3.7}$$

Note that condition (3.7) can be easily verified algorithmically since in composite setting we are able to compute points with a small subgradient of h_{k+1} (see [22]).

Thus, we come to the following general scheme.

Algorithm 1: Contracting Proximal Method

Require: Choose $x_0 \in \operatorname{dom}\psi$, $\gamma_0 > 0$, set $v_0 := x_0$, $A_0 := 0$.

Ensure: $k \geq 0$.

1: Choose $a_{k+1} > 0$. Set $A_{k+1} := A_k + a_{k+1}$.

2: Denote contracted objective with regularizer:

$$h_{k+1}(x) := A_{k+1}f\big(\tfrac{a_{k+1}x + A_kx_k}{A_{k+1}}\big) + a_{k+1}\psi(x) + \gamma_k\beta_d(v_k;x).$$

3: Choose accuracy $\delta_{k+1} \geq 0$.

4: Find $v_{k+1} \in \operatorname{dom}\psi$ such that $\exists\, s \in \partial h_{k+1}(v_{k+1}) :\ \|s\|_* \leq \delta_{k+1}$.

5: Set $x_{k+1} := \frac{a_{k+1}v_{k+1} + A_kx_k}{A_{k+1}}$.

6: Set $\gamma_{k+1} := \gamma_k + a_{k+1}\sigma_d(\psi)$.

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

At this moment, we need one additional assumption. It relates the dual norm $\|\cdot\|_*$ (used at step 4) with the Bregman divergence $\beta_d(v;x)$.

Assumption 3.1. For some $p \geq 1$, prox-function $d(\cdot)$ is uniformly convex of degree $p+1$ with respect to the primal norm $\|\cdot\|$ with parameter $\sigma_{p+1}(d) > 0$ (see inequality (2.2)).

Let us write down the convergence guarantees of the method.

THEOREM 3.2 (convergence of contracting proximal method). *Let Assumption* 3.1 *hold. Then for section* 3 *at all iterations* $k \geq 0$ *we have*

$$A_k\left(F(x_k) - F^*\right) + \gamma_k \beta_d(v_k; x^*) + \sum_{i=1}^{k} \gamma_i \beta_d(v_{i-1}; v_i) \leq R_k(p, \delta), \tag{3.8}$$

where

$$R_k(p,\delta) \overset{\text{def}}{=} \left(\left(\gamma_0 \beta_d(x_0; x^*)\right)^{\frac{p}{p+1}} + \left(\frac{p+1}{\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{1/(p+1)}} \right)^{\frac{p+1}{p}}. \tag{3.9}$$

Proof. First, let us ensure by induction in $k \geq 0$ the following inequality:

$$\begin{aligned} &A_k\left(F(x_k) - F(x)\right) + \gamma_k \beta_d(v_k; x) + \sum_{i=1}^{k} \gamma_i \beta_d(v_{i-1}; v_i) \\ &\leq \gamma_0 \beta_d(x_0; x) + \sum_{i=1}^{k} \langle s_i, v_i - x\rangle, \quad x \in \operatorname{dom}\psi, \end{aligned} \tag{3.10}$$

where $s_i \in \partial h_i(v_i)$. It is obviously true for $k = 0$. Let it hold for some $k \geq 0$ and consider the next. Note that (3.10) is exactly (3.1) with

$$C_k(x) \equiv \sum_{i=1}^{k} \Big[\gamma_i \beta_d(v_{i-1}; v_i) + \langle s_i, x - v_i\rangle \Big].$$

Therefore, we have

$$\begin{aligned} &\gamma_0 \beta_d(x_0; x) + A_{k+1} F(x) \\ &\overset{(3.2)}{\geq} h_{k+1}(x) + A_k \psi(x_k) + C_k(x) \\ &\overset{(2.8)}{\geq} h_{k+1}(v_{k+1}) + \langle s_{k+1}, x - v_{k+1}\rangle + \gamma_{k+1}\beta_d(v_{k+1}; x) + A_k\psi(x_k) + C_k(x) \\ &= A_{k+1} f(x_{k+1}) + a_{k+1}\psi(v_{k+1}) + \gamma_{k+1}\beta_d(v_{k+1}; x) + A_k\psi(x_k) + C_{k+1}(x) \\ &\geq A_{k+1} F(x_{k+1}) + \gamma_{k+1}\beta_d(v_{k+1}; x) + C_{k+1}(x). \end{aligned}$$

This is (3.10) for the next step.

Now, plugging $x \equiv x^*$ into (3.10) and taking into account the nonnegativity of all terms in the left-hand side, we get

$$\gamma_k \beta_d(v_k; x^*) \leq \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \langle s_i, v_i - x^*\rangle.$$

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Now, we need to estimate the right-hand side from above. Using the uniform convexity (2.2), we conclude that for every $k \geq 0$

$$\begin{aligned} \frac{\gamma_k \sigma_{p+1}(d)}{p+1} \|v_k - x^*\|^{p+1} &\leq \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \|s_i\|_* \cdot \|v_i - x^*\| \\ &\overset{(3.7)}{\leq} \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \delta_i \|v_i - x^*\| \quad \equiv \quad \alpha_k. \end{aligned} \tag{3.11}$$

In order to finish the proof, it is enough to bound from above the value α_k, for which we have the following recurrence:

$$\alpha_k = \alpha_{k-1} + \delta_k \|v_k - x^*\| \overset{(3.11)}{\leq} \alpha_{k-1} + \delta_k \left(\frac{p+1}{\gamma_k \sigma_{p+1}(d)} \right)^{\frac{1}{p+1}} \alpha_k^{\frac{1}{p+1}}.$$

Dividing both sides by $\alpha_k^{\frac{1}{p+1}}$ and using the monotonicity of this sequence, we get

$$\alpha_k^{\frac{p}{p+1}} \leq \frac{\alpha_{k-1}}{\alpha_k^{1/(p+1)}} + \delta_k \left(\frac{p+1}{\gamma_k \sigma_{p+1}(d)} \right)^{\frac{1}{p+1}} \leq \alpha_{k-1}^{\frac{p}{p+1}} + \delta_k \left(\frac{p+1}{\gamma_k \sigma_{p+1}(d)} \right)^{\frac{1}{p+1}}.$$

Finally, from the last inequality we obtain

$$\alpha_k \leq \left(\alpha_0^{\frac{p}{p+1}} + \left(\frac{p+1}{\sigma_{p+1}(d)} \right)^{\frac{1}{p+1}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{1/(p+1)}} \right)^{\frac{p+1}{p}},$$

which is the right-hand side of (3.8). ∎

We see that accuracies δ_k for subgradients of the subproblems appear in (3.9) in an additive form, weighted by the coefficients $\gamma_k^{-\frac{1}{p+1}}$. They should be chosen in a way making the right-hand side of (3.8) small enough. Let us consider the simplest case, when all δ_k are the same.

COROLLARY 3.3. *Let $\delta_k = \delta > 0$ for all $k \geq 1$. Assume that the coefficients A_k grow* sublinearly*:*

$$A_k \geq c k^{p+1}, \quad k \geq 1, \tag{3.12}$$

with some constant $c > 0$. Then for every

$$k \geq \left(\frac{\gamma_0 \beta_d(x_0; x_*)}{c\varepsilon} \right)^{\frac{1}{p+1}} 2^{\frac{1}{p}} \qquad \text{and} \qquad \delta \leq \frac{(c\varepsilon)^{\frac{p}{p+1}}}{2} \left(\frac{\gamma_0 \sigma_{p+1}(d)}{p+1} \right)^{\frac{1}{p+1}} \tag{3.13}$$

we have

$$R_k(p, \delta) \leq \varepsilon A_k. \tag{3.14}$$

Consequently, by (3.8) *we have $F(x_k) - F^* \leq \varepsilon$.*

Proof. Indeed,

$$\left(\frac{\gamma_0 \beta_d(x_0; x^*)}{A_k} \right)^{\frac{p}{p+1}} \overset{(3.12)}{\leq} \left(\frac{\gamma_0 \beta_d(x_0; x^*)}{c} \right)^{\frac{p}{p+1}} k^p \overset{(3.13)}{\leq} \frac{\varepsilon^{\frac{p}{p+1}}}{2}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

and

$$\frac{\left(\frac{p+1}{\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}}{A_k^{\frac{p}{p+1}}}\sum_{i=1}^{k}\frac{\delta_i}{\gamma_i^{1/(p+1)}} \le \frac{\left(\frac{p+1}{\gamma_0\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}k\delta}{A_k^{\frac{p}{p+1}}} \overset{(3.12)}{\le} \frac{\left(\frac{p+1}{\gamma_0\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}\delta}{c^{\frac{p}{p+1}}k^{p+1}}$$
$$\le \frac{\left(\frac{p+1}{\gamma_0\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}\delta}{c^{\frac{p}{p+1}}} \overset{(3.13)}{\le} \frac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

Summing up these two inequalities we obtain (3.14). ∎

COROLLARY 3.4. *Let $\delta_k = \delta > 0$ for all $k \ge 1$. Let the coefficients A_k grow* linearly*:*

$$A_k \ge A_1\exp\big(\omega(k-1)\big), \quad k \ge 1, \tag{3.15}$$

with some constant $0 < \omega \le 1$ and initial $A_1 > 0$. Then for every

$$k \ge 1 + \frac{1}{\omega}\log\left(\frac{\gamma_0\beta_d(x_0;x^*)}{A_1\varepsilon}2^{(p+1)/p}\right) \tag{3.16}$$

and

$$\delta \le \frac{(A_1\varepsilon)^{\frac{p}{p+1}}\omega}{2}\cdot\frac{p}{p+1}\cdot\left(\frac{\gamma_0\sigma_{p+1}(d)}{p+1}\right)^{\frac{1}{p+1}} \tag{3.17}$$

we have

$$R_k(p,\delta) \le \varepsilon A_k. \tag{3.18}$$

Consequently, by (3.8) *we have $F(x_k) - F^* \le \varepsilon$.*

Proof. Indeed,

$$\left(\frac{\gamma_0\beta_d(x_0;x^*)}{A_k}\right)^{\frac{p}{p+1}} \overset{(3.15)}{\le} \left(\frac{\gamma_0\beta_d(x_0;x^*)}{A_1\exp\big(\omega(k-1)\big)}\right)^{\frac{p}{p+1}} \overset{(3.17)}{\le} \frac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

Now, note that the following inequality holds for all $x \ge 0$:

$$\exp(x) \ge 1 + x. \tag{3.19}$$

Therefore,

$$\frac{A_k^{\frac{p}{p+1}}}{k} \overset{(3.15)}{\ge} \frac{A_1^{\frac{p}{p+1}}\exp\left(\frac{p}{p+1}\omega(k-1)\right)}{k} \overset{(3.19)}{\ge} \frac{A_1^{\frac{p}{p+1}}\left(1+\frac{p}{p+1}\omega(k-1)\right)}{k} > \frac{p}{p+1}A_1^{\frac{p}{p+1}}\omega. \tag{3.20}$$

And we obtain

$$\frac{\left(\frac{p+1}{\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}}{A_k^{\frac{p}{p+1}}}\sum_{i=1}^{k}\frac{\delta_i}{\gamma_i^{1/(p+1)}} \le \frac{\left(\frac{p+1}{\gamma_0\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}k\delta}{A_k^{\frac{p}{p+1}}} \overset{(3.20)}{<} \frac{\left(\frac{p+1}{\gamma_0\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}p+1\delta}{A_1^{\frac{p}{p+1}}p\omega}$$
$$\overset{(3.17)}{\le} \frac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

∎

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Estimates (3.13) and (3.17) show that the bound for the inner accuracy δ has a reasonable dependency on the absolute accuracy ε required for the initial problem (2.1). Thus, in both cases, in step 4 of the algorithm we need to find a point v_{k+1} with subgradient $s \in \partial h_{k+1}(v_{k+1})$:

$$\|s\|_* \leq O\Big(\varepsilon^{\frac{p}{p+1}}\Big) \quad \Leftrightarrow \quad \|s\|_*^{\frac{p+1}{p}} \leq O\big(\varepsilon\big).$$

This is a reachable goal, especially for methods minimizing $h_{k+1}(\cdot)$ with a linear rate of convergence.

In practice, it may be reasonable not to use very small inner accuracy on a first stage but to decrease it over the iterations. Then, the following simple choice of $\{\delta_k\}_{k\geq 0}$ can work.

COROLLARY 3.5. *Let us define* $\delta_k \equiv \frac{c}{k^s}$ *with fixed absolute constants* $c > 0$ *and* $s > 1$. *Then,*

$$\sum_{i=1}^{k} \delta_i = c\left(1 + \sum_{i=2}^{k} \frac{1}{i^s}\right) \leq c\left(1 + \int_1^{+\infty} \frac{dx}{x^s}\right) = \frac{cs}{s-1}.$$

Therefore, we have

$$R_k(p,\delta) \leq \left(\big(\gamma_0 \beta_d(x_0;x^*)\big)^{\frac{p}{p+1}} + \left(\frac{p+1}{\gamma_0 \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} \frac{cs}{s-1}\right)^{\frac{p+1}{p}}.$$

4. Application of tensor methods. In this section, let us incorporate the high-order tensor methods [22] into section 3 for solving the corresponding inner subproblem (3.5). From now on, we restrict our attention to Euclidean norms. Let us fix symmetric positive-definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$ (notation $B = B^* \succ 0$) and use the following norm for the primal space: $\|x\| \equiv \langle Bx, x\rangle^{1/2}$, $x \in \mathbb{E}$. The norms for multilinear forms on $\mathbb{E}$ are induced in the standard way (see section 2).

Assumption 4.1. For fixed $p \geq 1$, the pth derivative of the smooth component of the objective function is Lipschitz continuous:

$$\|D^p f(x) - D^p f(y)\| \leq L_p(f)\|x - y\|, \quad x, y \in \operatorname{dom} f, \tag{4.1}$$

with some constant $0 < L_p(f) < +\infty$.

For this setup, we use the following simple prox function:

$$d(x) \equiv \frac{1}{p+1}\|x - x_0\|^{p+1}. \tag{4.2}$$

Thus, the choice of prox function (4.2) is strictly related to the preferable degree $p \geq 1$ of smoothness of function f.

Let us define the Taylor approximation $\Omega_p(f, x; y)$ of function f around the point $x \in \operatorname{dom} f$:

$$\Omega_p(f, x; y) \stackrel{\text{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y - x]^i.$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

By Assumption 4.1, we are able to bound its accuracy in the following way: for all $x, y \in \operatorname{dom} f$ it holds that

$$|f(y) - \Omega_p(f, x; y)| \le \frac{L_p(f)}{(p+1)!}\|y - x\|^{p+1}, \tag{4.3}$$

$$\|\nabla f(y) - \nabla_y \Omega_p(f, x; y)\|_* \le \frac{L_p(f)}{p!}\|y - x\|^p. \tag{4.4}$$

Let us look at our regularized objective $h_{k+1}(\cdot)$, which needs to be minimized at every step $k \ge 0$:

$$h_{k+1}(x) = \underbrace{A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right)}_{\stackrel{\text{def}}{=} g_{k+1}(x)} + \underbrace{a_{k+1}\psi(x) + \gamma_k \beta_d(v_k; x)}_{\stackrel{\text{def}}{=} \phi_{k+1}(x)}. \tag{4.5}$$

This is a sum of two convex functions: smooth component g_{k+1} and possibly non-smooth but simple component ϕ_{k+1}, which is strongly convex with respect to d.

Let us drop unnecessary indices and consider the subproblem in a general form:

$$\min_{x \in \operatorname{dom} h} \Big\{ h(x) \equiv g(x) + \phi(x) \Big\} \tag{4.6}$$

with g having bounded Lipschitz constant for some $p \ge 1$: $0 < L_p(g) < +\infty$. Since we assume the objective to be strongly convex with respect to d from (4.2) with parameter $\sigma_d(h) > 0$, for every $x, y \in \operatorname{dom} h$ and all $h'(x) \in \partial h(x)$ we have

$$h(y) - h(x) - \langle h'(x), y - x \rangle \ge \sigma_d(h)\beta_d(x; y) \stackrel{(2.3)}{\ge} \frac{\sigma_d(h) 2^{1-p}}{p+1}\|y - x\|^{p+1}. \tag{4.7}$$

Bound (4.3) motivates us to define the following point,

$$T_M(h; x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ \Omega_p(g, x; y) + \frac{M}{(p+1)!}\|y - x\|^{p+1} + \phi(y) \right\}, \tag{4.8}$$

and consider the following iteration process,

$$\boxed{z_{t+1} = T_M(h; z_t), \qquad t \ge 0.} \tag{4.9}$$

For $p = 1$, the point (4.8) is used in the composite gradient method [20]. For $p = 2$, this is a step of composite cubic Newton [6, 9]. It can be shown that for $M \ge pL_p(g)$ the auxiliary optimization problem in (4.8) is convex for *all* $p \ge 1$ (see [22, Theorem 1]). Therefore it can be efficiently solved by different techniques of convex optimization and linear algebra (see also [23, 22]).

Let us mention some properties of point $T \equiv T_M(h; x)$. Its characteristic condition is as follows:

$$\left\langle \nabla_y \Omega_p(g, x; T) + \frac{M}{p!}\|T - x\|^{p-1} B(T - x), y - T \right\rangle + \phi(y) \ge \phi(T), \quad y \in \operatorname{dom} \phi.$$

Therefore,

$$\phi'(T) \stackrel{\text{def}}{=} -\nabla_y \Omega_p(g, x; T) - \frac{M}{p!}\|T - x\|^{p-1} B(T - x) \in \partial\phi(T).$$

This inclusion justifies notation $h'(T) \stackrel{\text{def}}{=} \nabla g(T) + \phi'(T) \in \partial h(T)$.

In order to work with these objects, we use the following result (see [4, Lemma 2]).

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

LEMMA 4.2. *Let $\beta \geq 1$ and $M = \beta L_p(g)$. Then*

$$\langle h'(T), x - T\rangle \geq \left(\frac{p!}{(p+1)L_p(g)}\right)^{\frac{1}{p}} \cdot \|h'(T)\|_*^{\frac{p+1}{p}} \cdot \frac{(\beta^2-1)^{\frac{p-1}{2p}}}{\beta} \cdot \frac{p}{(p^2-1)^{\frac{p-1}{2p}}}. \tag{4.10}$$

In particular, if $\beta = p$, then

$$\langle h'(T), x - T\rangle \geq \left(\frac{p!}{(p+1)L_p(g)}\right)^{\frac{1}{p}} \cdot \|h'(T)\|_*^{\frac{p+1}{p}}. \tag{4.11}$$

The next lemma describes the global behavior of method (4.9).

LEMMA 4.3. *Let $\beta \geq 1$ and $M = \beta L_p(g)$. Then for any $x, y \in \operatorname{dom} h$ we have*

$$h(T_M(x)) \leq h(y) + \frac{(\beta+1)L_p(g)}{(p+1)!}\|y - x\|^{p+1}. \tag{4.12}$$

Proof. Indeed,

$$\begin{aligned}
h(T_M(x)) &= g(T_M(x)) + \phi(T_M(x)) \\
&\overset{(4.3)}{\leq} \Omega_p(g, x; T_M(x)) + \frac{M}{(p+1)!}\|T_M(x) - x\|^{p+1} + \phi(T_M(x)) \\
&\overset{(4.8)}{\leq} \Omega_p(g, x; y) + \frac{M}{(p+1)!}\|y - x\|^{p+1} + \phi(y) \\
&\overset{(4.3)}{\leq} g(y) + \frac{M + L_p(g)}{(p+1)!}\|y - x\|^{p+1} + \phi(y) \\
&= h(y) + \frac{(\beta+1)L_p(g)}{(p+1)!}\|y - x\|^{p+1}.
\end{aligned}$$

∎

Now, we are ready to prove a convergence result on the iteration process (4.9).

THEOREM 4.4 (convergence of tensor method). *Let $M = pL_p(g)$. Then, for every $t \geq 0$ and $y \in \operatorname{dom} h$ we have*

$$\begin{aligned}
\|h'(z_{t+2})\|_*^{\frac{p+1}{p}} &\leq \exp\left(-t \cdot \min\left\{1, \left[\frac{p!\,\sigma_d(h)2^{1-p}}{(p+1)L_p(g)}\right]^{\frac{1}{p}}\right\} \cdot \frac{p}{p+1}\right) \\
&\quad \cdot \left(\frac{(p+1)L_p(g)}{p!}\right)^{\frac{1}{p}} \cdot \left(h(y) - h^* + \frac{L_p(g)}{p!}\|y - z_0\|^{p+1}\right).
\end{aligned} \tag{4.13}$$

Proof. Let us consider the point $z_{t+1} = T_M(z_t)$. By (4.12), we have

$$h(z_{t+1}) \leq h(y) + \frac{L_p(g)}{p!}\|y - z_t\|^{p+1} \tag{4.14}$$

for any $y \in \operatorname{dom} h$. Denote $x_h^* \overset{\text{def}}{=} \operatorname{argmin}_{y \in \mathbb{E}} h(y)$ and consider $y = z_t + \alpha(x_h^* - z_t)$ for $\alpha \in [0, 1]$. Then we have

$$\begin{aligned}
h(z_{t+1}) - h^* &\leq h(z_t) - h^* - \alpha\,(h(z_t) - h^*) + \alpha^{p+1}\frac{L_p(g)}{p!}\|x_h^* - z_t\|^{p+1} \\
&\overset{(4.7)}{\leq} \left(1 - \alpha + \alpha^{p+1}\frac{(p+1)L_p(g)}{p!\,\sigma_d(h)2^{1-p}}\right) \cdot (h(z_t) - h^*)\,.
\end{aligned} \tag{4.15}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

The minimum of the right-hand side is attained at

$$\alpha^* = \min\left\{1, \left[\frac{p!\,\sigma_d(h)2^{1-p}}{(p+1)L_p(g)}\right]^{\frac{1}{p}}\right\}.$$

Plugging it into (4.15) gives

$$\begin{aligned} h(z_{t+1}) - h^* &\le \left(1 - \alpha^*\frac{p}{p+1}\right)\cdot(h(z_t) - h^*) \\ &\le \exp\left(-\alpha^*\frac{p}{p+1}\right)\cdot(h(z_t) - h^*). \end{aligned} \tag{4.16}$$

Therefore, for every $t \ge 0$ we have

$$\begin{aligned} h(z_{t+1}) - h^* &\overset{(4.16)}{\le} \exp\left(-t\alpha^*\frac{p}{p+1}\right)\cdot(h(z_1) - h^*) \\ &\overset{(4.14)}{\le} \exp\left(-t\alpha^*\frac{p}{p+1}\right)\cdot\left(h(y) - h^* + \frac{L_p(g)}{p!}\|y - z_0\|^{p+1}\right) \end{aligned}$$

for every $y \in \operatorname{dom} h$. It remains to use Lemma 4.2 and finish the proof:

$$\begin{aligned} h(z_{t+1}) - h^* &\ge h(z_{t+1}) - h(z_{t+2}) \\ &\ge \langle h'(z_{t+2}), z_{t+1} - z_{t+2}\rangle \\ &\overset{(4.11)}{\ge} \left(\frac{p!}{(p+1)L_p(g)}\right)^{\frac{1}{p}}\cdot\|h'(z_{t+2})\|_*^{\frac{p+1}{p}}. \end{aligned}$$

∎

Thus, we can see that, applying tensor method (4.9) of degree $p \ge 1$ on step 4 of the general contracting proximal method (section 3), we obtain fast linear convergence for the norms of subgradients. Hence, we can estimate the total number of inner steps t_k at iteration $k \ge 0$ as follows.

COROLLARY 4.5. *Let us minimize function $h_{k+1}(\cdot)$ by iterations,*

$$z_{t+1} = T_M(h_{k+1}; z_t), \quad t \ge 0,$$

using $M := pL_p(g_{k+1})$ and $z_0 := v_k$. Then we have

$$\|h'_{k+1}(z_{t_k})\|_* \le \delta_{k+1}$$

for

$$t_k \ge 2 + \max\left\{1, \frac{\ell_{k+1}}{\mu_{k+1}}\right\}\cdot\frac{p+1}{p}\cdot\log\left(\frac{\ell_{k+1}\,D_{k+1}}{\delta_{k+1}^{\frac{p+1}{p}}}\right), \tag{4.17}$$

where

$$\ell_{k+1} \overset{\text{def}}{=} \left(\frac{(p+1)L_p(g_{k+1})}{p!}\right)^{\frac{1}{p}}, \qquad \mu_{k+1} \overset{\text{def}}{=} \left(\gamma_{k+1}2^{1-p}\right)^{\frac{1}{p}} \tag{4.18}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

and

$$D_{k+1} \overset{\text{def}}{=} A_k(F(x_k) - F^*) + \gamma_k \beta_d(v_k; x^*) + \left(\frac{\ell_{k+1}}{\mu_{k+1}}\right)^p \beta_d(v_k; x^*)$$
$$\overset{(3.8)}{\leq} R_k(p,\delta) \cdot \left(1 + \frac{1}{\gamma_0}\left(\frac{\ell_{k+1}}{\mu_{k+1}}\right)^p\right). \tag{4.19}$$

Proof. By definition, for all $x \in \operatorname{dom}\psi$, we have

$$\begin{aligned} &h_{k+1}(x) + A_k\psi(x_k) \\ &\quad = A_{k+1} f\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + \gamma_k \beta_d(v_k; x) + A_k\psi(x_k) \\ &\quad \geq A_{k+1} F\left(\frac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + \gamma_k \beta_d(v_k; x) \geq A_{k+1}F^*. \end{aligned}$$

Therefore,

$$-h^*_{k+1} - A_k\psi(x_k) \leq -A_{k+1}F^*. \tag{4.20}$$

Then for $y \equiv x^* \overset{\text{def}}{=} \operatorname{argmin}_{y\in\mathbb{E}} F(y)$ we obtain

$$\begin{aligned} h_{k+1}(y) - h^*_{k+1} + \frac{L_p(g_{k+1})}{p!}\|y - z_0\|^{p+1} &= h_{k+1}(x^*) - h^*_{k+1} + \frac{L_p(g_{k+1})}{p!}\|x^* - v_k\|^{p+1} \\ &= A_{k+1} f\left(\frac{a_{k+1}x^* + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x^*) - h^*_{k+1} + \gamma_k\beta_d(v_k; x^*) \\ &\quad + \frac{L_p(g_{k+1})}{p!}\|x^* - v_k\|^{p+1} \\ &\leq a_{k+1}F^* + A_k F(x_k) - h^*_{k+1} - A_k\psi(x_k) + \gamma_k\beta_d(v_k; x^*) \\ &\quad + \frac{L_p(g_{k+1})}{p!}\|x^* - v_k\|^{p+1} \\ &\overset{(4.20)}{\leq} A_k(F(x_k) - F^*) + \gamma_k\beta_d(v_k; x^*) + \frac{L_p(g_{k+1})}{p!}\|x^* - v_k\|^{p+1} \overset{(2.3)}{\leq} D_{k+1}. \end{aligned}$$

It remains to use this bound together with (4.13) and the following estimation of a strong convexity parameter: $\sigma_d(h_{k+1}) \overset{(3.4)}{\geq} \gamma_{k+1}$. ∎

By representation (4.5), we have a simple relations between Lipschitz constants of the derivatives for function $g_{k+1}(\cdot)$ and $f(\cdot)$:

$$L_p(g_{k+1}) = \frac{a_{k+1}^{p+1}}{A_{k+1}^p} L_p(f), \quad p \geq 1. \tag{4.21}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Therefore, we can control the condition number of our objective. Indeed, by (4.17), the main complexity factor in the minimization process for $h_{k+1}(\cdot)$ is the ratio

$$\frac{\ell_{k+1}}{\mu_{k+1}} \equiv \left(\frac{(p+1)L_p(g_{k+1})}{p!\,2^{1-p}\gamma_{k+1}}\right)^{\frac{1}{p}} \overset{(4.21),(3.4)}{=} \left(\frac{(p+1)2^{p-1}a_{k+1}^{p+1}L_p(f)}{p!\,A_{k+1}^p(\gamma_0 + A_{k+1}\sigma_d(\psi))}\right)^{\frac{1}{p}}.$$

We are able to keep this ratio small by applying an appropriate growth strategy for coefficients A_k.

Let us consider two cases: $\sigma_d(\psi) = 0$ and $\sigma_d(\psi) > 0$.

1. $\sigma_d(\psi) = 0$. Let us choose $c \equiv \frac{p!\,\gamma_0}{2^{p-1}(p+1)^{p+2}L_p(f)}$ and $a_k \equiv c(p+1)k^p$. Then we have

$$A_k = c(p+1)\sum_{i=1}^{k} i^p \geq c(p+1)\int_0^k x^p dx = ck^{p+1},$$

and we get

$$\frac{a_{k+1}^{p+1}}{A_{k+1}^p} \leq c(p+1)^{p+1} = \frac{p!\,\gamma_0}{2^{p-1}(p+1)L_p(f)}. \tag{4.22}$$

Thus we obtain

$$\frac{\ell_{k+1}}{\mu_{k+1}} = \left(\frac{a_{k+1}^{p+1}}{A_{k+1}^p}\cdot\frac{2^{p-1}(p+1)L_p(f)}{p!\,\gamma_0}\right)^{\frac{1}{p}} \overset{(4.22)}{\leq} 1. \tag{4.23}$$

2. $\sigma_d(\psi) > 0$. For $k = 0$ we pick $a_1 \equiv c(p+1)$ as in the previous case. Now consider $k \geq 1$. Denote

$$\omega \overset{\text{def}}{=} \min\left\{\left(\frac{\sigma_d(\psi)p!}{L_p(f)(p+1)2^{p-1}}\right)^{\frac{1}{p+1}}, \frac{1}{2}\right\} \tag{4.24}$$

and choose a_{k+1} from the equation

$$\frac{a_{k+1}}{A_{k+1}} = \frac{a_{k+1}}{a_{k+1}+A_k} = \omega \quad \Leftrightarrow \quad a_{k+1} = \omega(1-\omega)^{-1}A_k.$$

Therefore

$$\begin{aligned}\frac{\ell_{k+1}}{\mu_{k+1}} &\leq \left(\frac{a_{k+1}^{p+1}}{A_{k+1}^{p+1}}\cdot\frac{L_p(f)(p+1)2^{p-1}}{p!\,\sigma_d(\psi)}\right)^{\frac{1}{p}} \\ &= \omega\cdot\left(\frac{L_p(f)(p+1)2^{p-1}}{p!\,\sigma_d(\psi)}\right)^{\frac{1}{p+1}} \leq 1.\end{aligned} \tag{4.25}$$

Thus, in both cases, at every upper-level step we need to perform a logarithmic number of iterations of the inner method, multiplied by a small constant.

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

We are ready to specify the whole optimization procedure.

Algorithm 2: Contracting Proximal Tensor Method

Require: Choose $x_0 \in \operatorname{dom} F$, inner accuracy $\delta > 0$, $\gamma_0 > 0$.

Set $v_0 := x_0$, $A_0 := 0$.

Fix $d(x) := \frac{1}{p+1}\|x - x_0\|^{p+1}$,

$$c := \frac{p!\,\gamma_0}{2^{p-1}(p+1)^{p+2}L_p(f)}, \quad \omega := \min\left\{\left(\frac{\sigma_d(\psi)p!}{L_p(f)(p+1)2^{p-1}}\right)^{\frac{1}{p+1}}, \frac{1}{2}\right\}.$$

Ensure: $k \geq 0$.

1: **If** $k = 0$ or $\omega = 0$ **Then**

$$a_{k+1} := c(p+1)(k+1)^p.$$

Else

$$a_{k+1} := \omega(1-\omega)^{-1}A_k.$$

2: Set $A_{k+1} := A_k + a_{k+1}$.

3: Denote contracted objective with regularizer:

$$g_{k+1}(x) := A_{k+1}f\left(\tfrac{a_{k+1}x + A_k x_k}{A_{k+1}}\right),$$

$$\phi_{k+1}(x) := a_{k+1}\psi(x) + \gamma_k \beta_d(v_k; x),$$

$$h_{k+1}(x) := g_{k+1}(x) + \phi_{k+1}(x).$$

4: Solve inner subproblem by tensor method up to accuracy δ:

$$z_0 := v_k, \ t_k := 0, \ M := pL_p(f)\tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}.$$

Do $z_{t_k+1} := T_M(h_{k+1}, z_{t_k})$, $t_k := t_k + 1$ **Until** $\|h'_{k+1}(z_{t_k})\|_* \leq \delta$.

$$v_{k+1} := z_{t_k}.$$

5: Set $x_{k+1} := \frac{a_{k+1}v_{k+1} + A_k x_k}{A_{k+1}}$.

6: Set $\gamma_{k+1} := \gamma_k + a_{k+1}\sigma_d(\psi)$.

Let us present global complexity bounds for this method in convex and strongly convex cases.

THEOREM 4.6 (convex case). *Let for a given $\varepsilon > 0$ the inner accuracy δ be fixed as follows:*

$$\delta = \left(\frac{p!\,\varepsilon}{L_p(f)}\right)^{\frac{p}{p+1}} \frac{\gamma_0}{2^p(p+1)^{p+1}}.$$

Then, in order to achieve $F(x_K) - F^* \leq \varepsilon$ it is enough to perform

$$K = \left\lfloor 1 + 2^{\frac{1}{p}}\left(\frac{2^{p-1}(p+1)^{p+2}L_p(f)\,\beta_d(x_0; x^*)}{\varepsilon\, p!}\right)^{\frac{1}{p+1}}\right\rfloor \tag{4.26}$$

iterations of section 4. The total number of oracle calls $N_K \overset{\text{def}}{=} \sum_{k=1}^K t_k$ is bounded

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

as

$$N_K \le K \cdot \left(3 + \frac{p+1}{p} \log\left(4\left(1+\frac{1}{\gamma_0}\right)(p+1)^{\frac{1}{p}} K^p\right)\right). \tag{4.27}$$

Proof. Estimate (4.26) follows directly from (3.13), by substituting the value

$$c = \frac{p!\gamma_0}{2^{p-1}(p+1)^{p+2} L_p(f)}.$$

Now, let us prove (4.27). By (4.17), we have

$$\begin{aligned} t_k &\le 3 + \max\left\{1, \frac{\ell_{k+1}}{\mu_{k+1}}\right\} \cdot \frac{p+1}{p} \cdot \log\left(\frac{\ell_{k+1} D_{k+1}}{\delta^{\frac{p+1}{p}}}\right) \\ &\overset{(4.23),(4.19)}{\le} 3 + \frac{p+1}{p} \cdot \log\left(\frac{\gamma_0^{1/p}(1+\gamma_0^{-1}) R_k(p,\delta)}{\delta^{\frac{p+1}{p}}}\right). \end{aligned}$$

In order to finish the proof, we need to bound the value under the logarithm.

By the choice of a_k, we have an upper bound for A_k:

$$A_k = c(p+1)\sum_{i=1}^{k} i^p \le c(p+1)\int_0^{k+1} x^p dx = c(k+1)^{p+1}. \tag{4.28}$$

Therefore, for every $0 \le k \le K$:

$$\begin{aligned} \frac{R_k(p,\delta)}{\delta^{\frac{p+1}{p}}} &= \left(\frac{(\gamma_0 \beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} + \Big(\frac{(p+1)2^{p-1}}{\gamma_0}\Big)^{\frac{1}{p+1}} k\right)^{\frac{p+1}{p}} \\ &\le \left(\frac{(\gamma_0 \beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} + \Big(\frac{(p+1)2^{p-1}}{\gamma_0}\Big)^{\frac{1}{p+1}} K\right)^{\frac{p+1}{p}} \\ &= \left(\Big(\frac{L_p(f)\beta_d(x_0;x^*)}{p!\,\varepsilon}\Big)^{\frac{p}{p+1}} \frac{2^p (p+1)^{p+1}}{\gamma_0^{\frac{1}{p+1}}} + \Big(\frac{(p+1)2^{p-1}}{\gamma_0}\Big)^{\frac{1}{p+1}} K\right)^{\frac{p+1}{p}} \\ &\overset{(4.26)}{\le} \left(\Big(\frac{(p+1)2^{p-1}}{\gamma_0}\Big)^{\frac{1}{p+1}} (K^p + K)\right)^{\frac{p+1}{p}} \le 4\Big(\frac{p+1}{\gamma_0}\Big)^{\frac{1}{p}} K^p. \end{aligned}$$

This completes the proof. ∎

Now, let us discuss the overall dependence of δ and K on p, given by the claim of Theorem 4.6. For simplicity, we fix $\frac{L_p(f)}{\varepsilon} = 1$, $\beta_d(x_0;x^*) = 1$, and $\gamma_0 = 1$. Thus, we observe the functions

$$\delta(p) := \frac{(p!)^{\frac{p}{p+1}}}{2^p (p+1)^{p+1}}, \qquad K(p) := 1 + 2^{\frac{1}{p}} \left(\frac{2^{p-1}(p+1)^{p+2}}{p!}\right)^{\frac{1}{p+1}}. \tag{4.29}$$

One can see that $\log_2 \delta(p) \le -p$. Therefore, increasing the order of the method, it requires at least to double the precision of solving the subproblem. At the same time, we have (using Stirling's formula)

$$\lim_{p\to+\infty} K(p) = 1 + 2\exp\left(\lim_{p\to+\infty} \frac{(p+2)\log(p+1) - \log p!}{p+1}\right) = 1 + 2\exp(1).$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

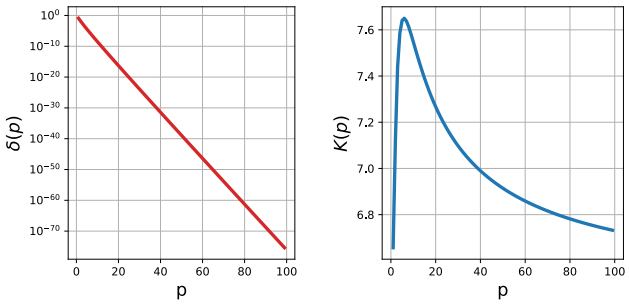


FIG. 1. *The dependence of δ and K on p, while $\frac{L_p(f)}{\varepsilon}$ and $\beta_d(x_0; x^*)$ are fixed.*

Hence, the value of $K(p)$ is bounded from above by an absolute constant. The graphs of the dependence (4.29) are shown in Figure 1. Note that in practice, we are interested rather in small values of p.

THEOREM 4.7 (strongly convex case). *Let $\sigma_d(\psi) > 0$ and condition number ω be defined as in* (4.24). *Let for a given $\varepsilon > 0$ the inner accuracy δ be fixed as follows:*

$$\delta = \left(\frac{p!\,\varepsilon}{L_p(f)}\right)^{\frac{p}{p+1}} \frac{\gamma_0 p\,\omega}{2^p (p+1)^{((p+1)^2+1)/(p+1)}}. \tag{4.30}$$

Then, in order to achieve $F(x_K) - F^* \le \varepsilon$, it is enough to perform

$$K = \left\lfloor 2 + \frac{1}{\omega}\mathcal{L} \right\rfloor \tag{4.31}$$

iterations of section 4, where

$$\mathcal{L} \overset{\text{def}}{=} \log\left(\max\left\{\frac{(p+1)^p}{\omega^{p+1}}, \frac{L_p(f)\beta_d(x_0; x^*)(p+1)^{p+1}2^{p+\frac{1}{p}}}{p!\,\varepsilon}\right\}\right).$$

The total number of oracle calls N_K is bounded as follows:

$$N_K \le K \cdot \left(3 + \left(1 + \frac{e}{(e-1)p}\right) \cdot (1 + \mathcal{L})\right.$$

$$\left. + \log\left(\max\left\{1, \left(\frac{4\sigma_d(\psi)p!}{(p+1)L_p(f)}\right)^{\frac{1}{p}}\right\} \cdot \left(1 + \frac{1}{\gamma_0}\right) \cdot \frac{(p+1)^{\frac{p+2}{p}}}{p^{\frac{p+1}{p}}} \cdot 2^{\frac{2p^2+p+4}{p}}\right)\right). \tag{4.32}$$

Proof. At every iteration $k \ge 1$, we have $A_{k+1} = (1-\omega)^{-1}A_k \ge A_k \exp(\omega)$. At the same time, we know that

$$\omega \le \frac{1}{2} \le \frac{e-1}{e}, \tag{4.33}$$

where $e = \exp(1)$. Since for all $\alpha \in [0, 1]$ it holds that

$$1 - \frac{e-1}{e}\alpha \ge \exp(-\alpha),$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

taking $\alpha = \omega\frac{e}{e-1} \overset{(4.33)}{\leq} 1$ we obtain $A_{k+1} \leq A_k \exp\bigl(\omega\frac{e}{e-1}\bigr)$. Therefore we have, for all $k \geq 0$,

$$A_1 \exp\bigl(k\omega\bigr) \leq A_{k+1} \leq A_1 \exp\Bigl(k\omega\frac{e}{e-1}\Bigr). \tag{4.34}$$

Now, estimate (4.31) follows directly from (4.34) and (3.16) by using the value $A_1 = \frac{p!\,\gamma_0}{2^{p-1}(p+1)^{p+1}L_p(f)}$.

By the choice of a_{k+1}, we have $\frac{\ell_{k+1}}{\mu_{k+1}} \overset{(4.25)}{\leq} 1$, and we need only to estimate the value under the logarithm in (4.17). For every $0 \leq k \leq K$, we have

$$\begin{aligned}
\frac{\ell_{k+1}D_{k+1}}{\delta^{\frac{p+1}{p}}} &\overset{(4.25),(4.19)}{\leq} \frac{\mu_{k+1}R_k(p,\delta)\left(1+\frac{1}{\gamma_0}\right)}{\delta^{\frac{p+1}{p}}} \\
&= (\gamma_0 + \sigma_d(\psi)A_{k+1})^{\frac{1}{p}} 2^{\frac{1}{p}-1}\left(1+\frac{1}{\gamma_0}\right) \\
&\quad \cdot \left(\frac{(\gamma_0\beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} + \Bigl(\frac{(p+1)2^{p-1}}{\gamma_0}\Bigr)^{\frac{1}{p+1}}k\right)^{\frac{p+1}{p}} \\
&\leq (\gamma_0 + \sigma_d(\psi)A_{K+1})^{\frac{1}{p}} 2^{\frac{1}{p}-1}\left(1+\frac{1}{\gamma_0}\right) \\
&\quad \cdot \left(\frac{(\gamma_0\beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} + \Bigl(\frac{(p+1)2^{p-1}}{\gamma_0}\Bigr)^{\frac{1}{p+1}}K\right)^{\frac{p+1}{p}}.
\end{aligned}$$

Let us estimate different terms in this expression separately.

1. By definition of ω, we have

$$\omega^{p+1} \leq \frac{(p+1)^p\sigma_d(\psi)A_1}{\gamma_0}. \tag{4.35}$$

Therefore,

$$\begin{aligned}
\gamma_0 + \sigma_d(\psi)A_{K+1} &\overset{(4.35),(4.34)}{\leq} \sigma_d(\psi)A_1\left(\frac{(p+1)^p}{\omega^{p+1}} + \exp\Bigl(K\omega\frac{e}{e-1}\Bigr)\right) \\
&\overset{(4.31)}{\leq} 2\sigma_d(\psi)A_1\exp\Bigl(K\omega\frac{e}{e-1}\Bigr).
\end{aligned}$$

2. Substituting the value for δ, we obtain

$$\begin{aligned}
\frac{(\gamma_0\beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} &\overset{(4.30)}{=} \left(\frac{L_p(f)\beta_d(x_0;x^*)}{p!\,\varepsilon}\right)^{\frac{p}{p+1}} \frac{2^p(p+1)^{((p+1)^2+1)/(p+1)}}{p\,\omega\gamma_0^{\frac{1}{p+1}}} \\
&\overset{(4.31)}{\leq} \frac{(p+1)^2 2^{(2p^2+p+1)/(p+1)}}{p\,\omega\gamma_0^{\frac{1}{p+1}}}\exp\Bigl(K\omega\frac{p}{p+1}\Bigr).
\end{aligned}$$

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

3. Finally, using that $\exp(x) \geq x$ for all $x \geq 0$, we have

$$K \leq \frac{p+1}{p\,\omega} \exp\Big(K\omega \frac{p}{p+1}\Big).$$

Therefore,

$$\begin{aligned}
\frac{\ell_{k+1} D_{k+1}}{\delta^{\frac{p+1}{p}}} &\leq \exp\Big(K\omega \frac{e}{(e-1)p}\Big) \cdot \Big(2^{2-p}\sigma_d(\psi) A_1\Big)^{\frac{1}{p}} \cdot \Big(1 + \frac{1}{\gamma_0}\Big) \\
&\quad \cdot \left(\frac{\exp\big(K\omega \frac{p}{p+1}\big)}{p\,\omega\,\gamma_0^{1/(p+1)}} \Big((p+1)^2 2^{\frac{2p^2+p+1}{p+1}} + (p+1)^{\frac{p+2}{p+1}} 2^{\frac{p-1}{p+1}}\Big)\right)^{\frac{p+1}{p}} \\
&< \exp\left(K\omega\Big(\frac{e}{(e-1)p} + 1\Big)\right) \cdot \Big(\frac{1}{p\,\omega}\Big)^{\frac{p+1}{p}} \cdot \Big(\frac{\sigma_d(\psi) A_1}{\gamma_0}\Big)^{\frac{1}{p}} \\
&\quad \cdot \Big(1 + \frac{1}{\gamma_0}\Big) \cdot (p+1)^{\frac{2(p+1)}{p}} 2^{\frac{2p^2+p+4}{p}} \\
&= \exp\left(K\omega\Big(\frac{e}{(e-1)p} + 1\Big)\right) \cdot \max\left\{1, \Big(\frac{4\sigma_d(\psi) p!}{(p+1) L_p(f)}\Big)^{\frac{1}{p}}\right\} \\
&\quad \cdot \Big(1 + \frac{1}{\gamma_0}\Big) \cdot \frac{(p+1)^{\frac{p+2}{p}}}{p^{\frac{p+1}{p}}} \cdot 2^{\frac{2p^2+p+4}{p}},
\end{aligned}$$

and we obtain (4.32). ∎

According to Theorems 4.6 and 4.7, the rate of convergence for the outer iterations of section 4 is of the same order as the one of the accelerated tensor method from [22]. However, at each step it uses a logarithmic number of steps of the basic method. It seems to be a reasonable price for the level of generality. Indeed, we are free to choose an arbitrary method as the basic one. The only requirement for it is the possibility of solving the inner subproblem (4.6) efficiently.

Note that an additional feature of our methods is that the sequences of points $\{x_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$ form *triangles* (see the rule (3.6)). A first-order accelerated method with this nice property was discovered in [8].

5. Numerical examples.

5.1. Quadratic function. Let us compare the numerical performance of the contracting proximal method and the classical proximal point algorithm (1.1) for unconstrained minimization of a convex quadratic function:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \qquad x \in \mathbb{R}^n,$$

with $A = A^* \succeq 0$. We also run the gradient method and the accelerated gradient method for this problem. A typical behavior of the algorithms is shown in Figure 2. The contracting proximal method has the same iteration rate as that of the accelerated gradient method, but requires more gradient evaluations (matrix-vector products) per iteration.

To compute every step of the proximal algorithms, we use the gradient method with line search. We try different strategies for choosing inner accuracies δ_k and end up with a simple rule $\delta_k = 1/k^2$, which provides a good balance in the performance of outer proximal iterations and the inner method. (Usually, it requires to do about 4 inner steps per iteration.)

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

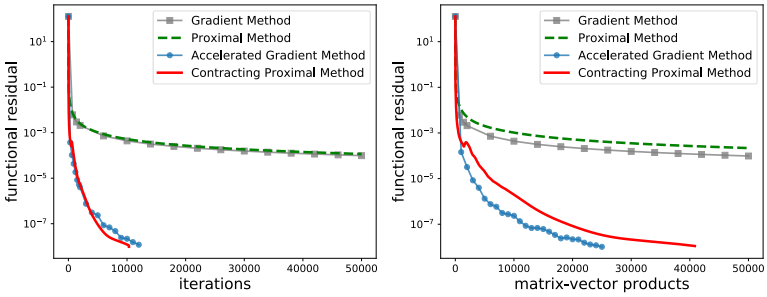


FIG. 2. *Convergence of first-order methods on quadratic function.*

TABLE 1
Minimization of quadratic function, $q = \lambda_{\min}(A)/\lambda_{\max}(A)$.

		Gradient method		Proximal method		Accelerated gradient method		Contracting proximal method	
n	q	iter	mat-vec	iter	mat-vec	iter	mat-vec	iter	mat-vec
500	10^{-2}	339	339	361	1044	115	229	**74**	**137**
	10^{-4}	12158	12158	12842	36731	**350**	**699**	393	1104
	10^{-6}	96072	96072	99269	313795	**854**	**1707**	1081	3780
1000	10^{-2}	338	338	359	1035	110	219	**73**	**135**
	10^{-4}	11884	11884	11912	56996	**360**	**719**	361	1014
	10^{-6}	77675	77675	80758	239508	**755**	**1509**	1117	3957

Data was generated randomly, but the set of eigenvalues of the matrix was fixed according to the sigmoid function, for some given $\alpha > 0$

$$\lambda_i = \frac{1}{1 + \exp\left(\frac{\alpha}{n-1}(n+1-2i)\right)}, \qquad 1 \le i \le n.$$

Therefore it holds that $\lambda_1 = 1/(1+\exp(\alpha))$ and $\lambda_n = 1/(1+\exp(-\alpha))$, so parameter α is related to the *condition number* of the problem.

In Table 1 we demonstrate the number of iterations and the total number of matrix-vector products, which are required for the methods to solve the problem up to $\varepsilon = 10^{-7}$ accuracy in functional residual.

We see that contracting proximal method is *always better* than the usual proximal algorithm. It requires about the same number of iterations as the accelerated gradient methods, but it needs to spend more oracle calls per iteration, which confirms the theory.

5.2. Log-Sum-Exp. In the next example we compare the performance of second-order methods for unconstrained minimization of the following objective:

$$f(x) = \mu \ln\left(\sum_{i=1}^{m} \exp\left(\frac{\langle a_i, x\rangle - b_i}{\mu}\right)\right), \qquad x \in \mathbb{R}^n,$$

where $\mu > 0$ is a parameter, while coefficients of the vectors $\{a_i\}_{i=1}^m$ and b are randomly generated, and we set $m = 6n$.

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

TABLE 2
Comparison of second-order methods on Log-Sum-Exp.

		Cubic Newton		Accelerated cubic Newton		Contracting proximal cubic Newton	
n	μ	iter	oracle	iter	oracle	iter	oracle
50	1	389	389	177	**353**	**112**	491
	0.1	482	482	202	**403**	**141**	587
	0.05	886	886	343	**685**	**236**	1129
100	1	834	834	308	**615**	**189**	849
	0.1	1210	1210	377	**753**	**232**	1021
	0.05	2598	2598	641	**1281**	**397**	1740

We compare a cubically regularized Newton method [23] and its accelerated variant from [19] with the contracting proximal cubic Newton (section 4 with $p = 2$) for minimizing the objective up to $\varepsilon = 10^{-8}$ accuracy in functional residual. In these algorithms we use the following Euclidean norm for the primal space, $\|x\| = \langle Bx, x\rangle^{1/2}$ with matrix $B = \sum_{i=1}^{m} a_i a_i^T$, and fix the regularization parameter equal to 1. The results are shown in Table 2.

We see that contracting proximal method outperforms the direct methods in the number of iterations, but usually requires additional oracle calls for solving the subproblem.

6. Conclusion. In this work, we propose a general acceleration scheme, based on proximal iterations. There are two distinguishing features of our methods: employing the *contraction* of the smooth component of the objective (this provides the acceleration) and flexibility of *prox-function* (its choice should take into account both the geometry of the problem and the order of the smoothness).

One of the recent important applications of the accelerated proximal point methods in machine learning is the universal framework *Catalyst*, applicable to the first-order methods [13, 14]. This is a powerful approach for accelerating many specific optimization methods in a common way. We believe that the results of this paper can help in advancing in this direction, resulting in the faster high-order methods for many practical applications.

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] M. BAES, *Estimate Sequence Methods: Extensions and Approximations*, Institute for Operations Research, ETH, Zürich, Switzerland, 2009.
[3] H. H. BAUSCHKE, J. BOLTE, AND M. TEBOULLE, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Math. Oper. Res., 42 (2016), pp. 330–348.
[4] N. DOIKOV AND Y. NESTEROV, *Local Convergence of Tensor Methods*, CORE Discussion Paper 2019/21, 2019.
[5] N. DOIKOV AND Y. NESTEROV, *Minimizing Uniformly Convex Functions by Cubic Regularization of Newton Method*, preprint, arXiv:1905.02671, 2019.
[6] N. DOIKOV AND P. RICHTÁRIK, *Randomized block cubic Newton method*, in Proceedings of the International Conference on Machine Learning, 2018, pp. 1289–1297.

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

[7] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, J. Qijia, Y. T. Lee, L. Yuanzhi, and S. Aaron, *Near optimal methods for minimizing convex functions with Lipschitz p-th derivatives*, in Proceedings of the Conference on Learning Theory, 2019, pp. 1392–1393.
[8] A. Gasnikov and Y. Nesterov, *Universal method for stochastic composite optimization problems*, Comput. Math. Math. Phys., 58 (2018), pp. 48–64.
[9] G. N. Grapiglia and Y. Nesterov, *Accelerated regularized Newton methods for minimizing composite convex functions*, SIAM J. Optim., 29 (2019), pp. 77–99.
[10] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.
[11] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim., 2 (1992), pp. 649–664.
[12] A. Ivanova, D. Grishchenko, A. Gasnikov, and E. Shulgin, *Adaptive Catalyst for Smooth Convex Optimization*, preprint, arXiv:1911.11271, 2019.
[13] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 3384–3392.
[14] H. Lin, J. Mairal, and Z. Harchaoui, *Catalyst acceleration for first-order convex optimization: From theory to practice*, J. Mach. Learn. Res., 18 (2018), pp. 1–54.
[15] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354.
[16] 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.
[17] A. Nemirovskii and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.
[18] Y. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.
[19] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181.
[20] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[21] Y. Nesterov, *Lectures on Convex Optimization*, Springer Optim. Appl. 137, Springer, New York, 2018.
[22] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Math. Program., 2019, https://doi.org/10.1007/s10107-019-01449-1.
[23] Y. Nesterov and B. T. Polyak, *Cubic regularization of Newton's method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[24] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[25] S. Salzo and S. Villa, *Inexact and accelerated proximal point algorithms*, J. Convex Anal., 19 (2012), pp. 1167–1192.
[26] M. Schmidt, N. L. Roux, and F. R. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Proceedings of Advances in Neural Information Processing Systems, 2011, pp. 1458–1466.
[27] M. V. Solodov and B. F. Svaiter, *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optim., 22 (2001), pp. 1013–1035.
[28] Q. Van Nguyen, *Forward-backward splitting with Bregman distances*, Vietnam J. Math., 45 (2017), pp. 519–539.

Downloaded 07/14/23 to 213.213.220.128 . Redistribution subject to CCBY license

© 2021 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license