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

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 $f : \text{dom } f \to \mathbb{R}$, consists of solving at each iteration the following subproblem:

$$
\begin{aligned}
    x_{k+1} &= \arg\min_x \left\{ a_{k+1} f(x) + \frac{1}{2} \| x - x_k \|^2 \right\}, \\
    k &\geq 0,
\end{aligned}
$$

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 strong convex. Therefore, we can hope that computing an (inexact) proximal step is usually simpler than solving the initial problem.

For a function $f \in \mathcal{F}_L^{1,1}$ (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 $\mathcal{F}_L^{1,1}$ [17]. An accelerated
It can be improved up to the level by a direct acceleration scheme, utilizing the notion of estimating sequences \([2, 22]\). Note that this rate can also be achieved by the high-order tensor methods \([22]\). Thus, we use a sequence of auxiliary points \(\{v_k\}_{k \geq 0}\) and the scaling coefficients \(A_k \equiv \sum_{i=1}^k a_i\). 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:

\[
\frac{1}{2}||x_0 - x||^2 + A_k f(x) \geq \frac{1}{2}||v_k - x||^2 + A_k f(x_k), \quad x \in \text{dom} \ f.
\]

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

\[
(1.4) \quad f(x_k) - f^* \leq \frac{||x_0 - x^*||^2}{2A_k}.
\]

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

\[
\frac{1}{2}||x_0 - x||^2 + A_{k+1} f(x) = \frac{1}{2}||x_0 - x||^2 + A_k f(x) + a_{k+1} f(x)
\geq \frac{1}{2}||v_k - x||^2 + A_k f(x_k) + a_{k+1} f(x)
\geq \frac{1}{2}||v_k - x||^2 + A_{k+1} f\left(\frac{a_k x_k + A_k x_k}{A_{k+1}}\right) \equiv h_{k+1}(x),
\]

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:

\[
(1.5) \quad v_{k+1} = \arg\min_{x \in \mathbb{E}} h_{k+1}(x).
\]

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 \( f \in \mathcal{F}_F^{1,1} \), the recommended choice of \( a_{k+1}^2 \) corresponds to the quadratic equation \([18]\):

\[
a_{k+1}^2 = \frac{1}{f}(a_{k+1} + A_k).
\]

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,\(^2\)

\[
\nabla^2 h_{k+1}(x) = I + \frac{a_{k+1}^2}{A_{k+1}} \nabla^2 f \left( \frac{a_{k+1}x_k + A_k x_k}{A_{k+1}} \right) \overset{(1.6)}{\leq} 2I.
\]

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 the 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 Algorithm 1). 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 \( p \)-th derivative (\( p \geq 1 \)) of the smooth part of the objective is Lipschitz continuous. We present a fully defined optimization scheme (Algorithm 2), 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.

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 \( \langle s, x \rangle \).

\(^{1}\)Hence, these bounds should take into account the efficiency of the auxiliary minimization scheme used for solving the problem (1.5).
Let us fix some arbitrary (possibly non-Euclidean) norm \( \| \cdot \| \) on space \( E \) and define the dual norm \( \| \cdot \|_* \) on \( E^* \) in the standard way:

\[
\|s\|_* \overset{\text{def}}{=} \sup_{h \in E} \{ \langle s, h \rangle : \|h\| \leq 1 \}.
\]

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

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

\[
\|D^p f(x)\| \overset{\text{def}}{=} \sup_{h_1, \ldots, h_p \in E} \{ D^p f(x)[h_1, \ldots, h_p] : \|h_i\| \leq 1, i = 1, \ldots, p \}.
\]

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

Our goal is to solve the following composite minimization problem:

\[
(2.1) \min_{x \in \text{dom } F} \left\{ F(x) \equiv f(x) + \psi(x) \right\},
\]

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 \( p \)-th derivative is Lipschitz continuous for some \( p \geq 1 \), and \( \psi : E \to \mathbb{R} \cup \{+\infty\} \) is a proper closed convex function, which we assume to be simple, but possibly nondifferentiable, with \( \text{dom } \psi \subset \text{dom } f \). We also assume that solution \( x^* \in \text{dom } F \) of problem (2.1) does exist, denoting \( F^* = F(x^*) \).

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

\[
\beta_d(x;y) \overset{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle.
\]

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 \( x, y \in \text{dom } d \):

\[
(2.2) \beta_d(x;y) \geq \frac{\sigma_{p+1}(d)}{p+1} \|x - y\|^{p+1}.
\]

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

\[
(2.3) \beta_d(x;y) \geq \frac{2^{1-p}}{p+1} \|x - y\|^{p+1}, \quad x, y \in E.
\]
For more examples of available prox functions see [3, 15].

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

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

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

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

$$
\beta_\psi(x, \psi'(x); y) \geq \sigma_d(\psi) \| y - x \|^2.
$$

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 $f : \text{dom } f \to \mathbb{R}$ be convex, with Lipschitz continuous third derivative:

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

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

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

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

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

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

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

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 $x, y \in \text{dom } (f_1 + f_2)$ we have

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

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

$$
\beta_\ell(x; y) = 0.
$$

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

$$
\beta_f(x; y) = \beta_d(x; y),
$$

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:

$$
h(y) = g(y) + a \psi(y) + \gamma \beta_d(z; y), \quad a, \gamma \geq 0,
$$
where $g$ and $\psi$ are arbitrary closed convex functions, and $\psi$ is strongly convex with respect to $d$ with some constant $\sigma_d(\psi) \geq 0$. Then we have, for every $x, y \in \text{dom } h$ and every $h'(x) \in \partial h(x)$

$$h(y) - h(x) - \langle h'(x), y - x \rangle = \beta_h(x; y)$$

(2.8)  

$$(\text{2.5), (2.7)} \quad \beta_g(x; y) + a\beta_{\psi}(x; y) + \gamma\beta_d(x; y)$$

$$\geq (a\sigma_d(\psi) + \gamma)\beta_d(x; y).$$

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

(2.9)  

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

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

(3.1)  

$\gamma_0\beta_d(x_0; x) + A_kF(x) \geq \gamma_k\beta_d(v_k; x) + A_kF(x_k) + C_k(x),$

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

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

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

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

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,

$$F(x_k) - F^* \leq \frac{\gamma_0\beta_d(x_0, x^*)}{A_0}, \quad k \geq 1,$$

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 discuss a simple possibility for propagating relation (3.1) to the next iteration.

$$\gamma_0\beta_d(x_0; x) + A_{k+1}F(x)$$

$$= \gamma_0\beta_d(x_0; x) + A_kF(x) + a_{k+1}F(x)$$

(3.2)  

$$\geq \gamma_k\beta_d(v_k; x) + A_kF(x_k) + a_{k+1}F(x) + C_k(x)$$

$$\geq \gamma_k\beta_d(v_k; x) + A_{k+1}f\left(\frac{a_{k+1}x_k + A_kx_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + A_k\psi(x_k) + C_k(x),$$

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

(3.3)  

$h_{k+1}(x) \stackrel{\text{def}}{=} A_{k+1}f\left(\frac{a_{k+1}x_k + A_kx_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + \gamma_k\beta_d(v_k; x).$
This function is strongly convex with respect to $d(\cdot)$ with parameter
\begin{equation}
\sigma_{d}(h_{k+1}) \geq \gamma_{k+1} \overset{\text{def}}{=} a_{k+1}\sigma_{d}(\psi) + \gamma_{k}.
\end{equation}
If we are able to compute the exact minimum
\begin{equation}
T = \arg\min_{x \in \mathbb{R}} h_{k+1}(x),
\end{equation}
then by (2.9) we see that
\begin{align*}
h_{k+1}(x) &\geq h_{k+1}(T) + \gamma_{k+1}\beta_{d}(T; x) + A_{k}\psi(x_{k}) \\
&= A_{k+1}f\left(\frac{a_{k+1}T + A_{k}x_{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}) \\
&\geq A_{k+1}F\left(\frac{a_{k+1}T + A_{k}x_{k}}{A_{k+1}}\right) + \gamma_{k}\beta_{d}(v_{k}; T) + \gamma_{k+1}\beta_{d}(T; x).
\end{align*}
And it is natural to set $v_{k+1} = T$ and
\begin{equation}
x_{k+1} \overset{\text{def}}{=} \frac{a_{k+1}v_{k+1} + A_{k}x_{k}}{A_{k+1}}.
\end{equation}
Thus we would obtain guarantee (3.1) for the next step, with
\begin{equation}
C_{k+1}(x) = C_{k}(x) + \gamma_{k}\beta_{d}(v_{k}; v_{k+1}) = \sum_{i=1}^{k} \gamma_{i}\beta_{d}(v_{i}; v_{i+1}) \geq 0.
\end{equation}

Now, instead of computing the exact minimum (3.5), let us relax $v_{k+1} \in \text{dom } \psi$ to be a point with a small norm of subgradient:
\begin{equation}
\|s\|_{*} \leq \delta_{k+1} \quad \text{for some } s \in \partial h_{k+1}(v_{k+1}).
\end{equation}
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.

\begin{algorithm}
\caption{Contracting Proximal Method}
\begin{algorithmic}
\Require Choose $x_{0} \in \text{dom } \psi$, $\gamma_{0} > 0$, 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:
\begin{equation*}
h_{k+1}(x) := 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).
\end{equation*}
3: Choose accuracy $\delta_{k+1} \geq 0$.
4: Find $v_{k+1} \in \text{dom } \psi$ such that $\exists s \in \partial h_{k+1}(v_{k+1}) : \|s\|_{*} \leq \delta_{k+1}$.
5: Set $x_{k+1} := \frac{a_{k+1}v_{k+1} + A_{k}x_{k}}{A_{k+1}}$.
6: Set $\gamma_{k+1} := \gamma_{k} + a_{k+1}\sigma_{d}(\psi)$.
\end{algorithmic}
\end{algorithm}

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

**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 Algorithm 1 at all iterations \( k \geq 0 \) we have

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

where

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

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

\[ A_k(F(x_k) - F(x)) + \gamma_k \beta_d(v_k; x) + \sum_{i=1}^{k} \gamma_i \beta_d(v_{i-1}; v_i) \]

\[ \leq \gamma_0 \beta_d(x_0; x) + \sum_{i=1}^{k} \langle s_i, v_i - x \rangle, \quad x \in \text{dom } \psi, \]

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

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

Therefore, we have

\[ \gamma_0 \beta_d(x_0; x) + A_{k+1}F(x) \]

\[ \overset{(3.2)}{\geq} h_{k+1}(x) + A_k \psi(x_k) + C_k(x) \]

\[ \overset{(2.8)}{\geq} h_{k+1}(v_{k+1}) + \langle s_{k+1}, x - v_{k+1} \rangle + \gamma_{k+1} \beta_d(v_{k+1}; x) + A_k \psi(x_k) + C_k(x) \]

\[ = A_{k+1}f(x_{k+1}) + a_{k+1} \psi(v_{k+1}) + \gamma_{k+1} \beta_d(v_{k+1}; x) + A_k \psi(x_k) + C_{k+1}(x) \]

\[ \geq A_{k+1} F(x_{k+1}) + \gamma_{k+1} \beta_d(v_{k+1}; x) + C_{k+1}(x). \]

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

\[ \gamma_k \beta_d(v_k; x^*) \leq \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \langle s_i, v_i - x^* \rangle. \]
Now, we need to estimate the right-hand side from above. Using the uniform convexity (2.2), we conclude that for every $k \geq 0$

$$\frac{\gamma k \sigma_{p+1}(d)}{p+1} \|v_k - x^*\|^{p+1} \leq \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \|s_i\| \cdot \|v_i - x^*\|$$

(3.11)

$$\leq \gamma_0 \beta_d(x_0; x^*) + \sum_{i=1}^{k} \delta_i \|v_i - x^*\| \equiv \alpha_k.$$\hspace{1cm} (3.7)

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

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

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

$$\alpha_k^{\frac{p}{p+1}} \leq \frac{\alpha_{k-1}}{\alpha_k^{\frac{1}{p+1}}} + \delta_k \left(\frac{p+1}{\gamma k \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} \leq \alpha_{k-1}^{\frac{p}{p+1}} + \delta_k \left(\frac{p+1}{\gamma k \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}.$$\hspace{1cm} (3.11)

Finally, from the last inequality we obtain

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

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

We see that accuracies $\delta_k$ 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 $\delta_k$ are the same.

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

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

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

$$k \geq \left(\frac{\gamma_0 \delta_d(x_0; x^*)}{c e}\right)^{\frac{1}{p+1}} 2^{\frac{1}{p}}$$

and

$$\delta \leq \frac{(c e)^{\frac{1}{p+1}}}{2} \left(\frac{\gamma_0 \sigma_{p+1}(d)}{p+1}\right)^{\frac{1}{p+1}},$$

we have

$$R_k(p, \delta) \leq \varepsilon A_k.$$\hspace{1cm} (3.14)

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

**Proof.** Indeed,

$$\left(\frac{\gamma_0 \delta_d(x_0; x^*)}{A_k}\right)^{\frac{1}{p+1}} \leq \left(\frac{\gamma_0 \delta_d(x_0; x^*)}{c e}\right)^{\frac{1}{p+1}} 2^{\frac{1}{p}} k^p \leq \varepsilon^{\frac{1}{p+1}} \frac{c}{2}.$$
and
\[
\left( \frac{p+1}{\sigma p+1} \right)^{\frac{p+1}{p+1}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{(p+1)}} \leq \left( \frac{p+1}{\gamma_0 p+1} \right)^{\frac{p+1}{p+1}} k \delta \leq \left( \frac{p+1}{\gamma_0 p+1} \right)^{\frac{p+1}{p+1}} \frac{p+1}{p+1} \kappa \delta.
\]

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

\[ \Box \]

**Corollary 3.4.** Let \( \delta_k = \delta > 0 \) for all \( k \geq 1 \). Let the coefficients \( A_k \) grow linearly:
\[
A_k \geq A_1 \exp(\omega(k-1)), \quad k \geq 1,
\]
with some constant \( 0 < \omega \leq 1 \) and initial \( A_1 > 0 \). Then for every
\[
k \geq 1 + \frac{1}{\omega} \log \left( \frac{\gamma_0 \beta d (x_0 x^*)^2 (p+1)/p}{A_1 \varepsilon} \right)
\]
and
\[
\delta \leq \left( A_1 \varepsilon \right)^{\frac{p}{p+1}} \omega \cdot \left( \frac{\gamma_0 \sigma d (x_0 x^*)}{p+1} \right)^{\frac{p}{p+1}}
\]
we have
\[
R_k(p, \delta) \leq \varepsilon A_k.
\]

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

**Proof.** Indeed,
\[
\left( \frac{\gamma_0 \beta d (x_0 x^*)}{A_k} \right)^{\frac{p}{p+1}} \leq \left( \frac{\gamma_0 \beta d (x_0 x^*)}{A_1 \exp(\omega(k-1))} \right)^{\frac{p}{p+1}} \leq \varepsilon \frac{p}{p+1}.
\]

Now, note that the following inequality holds for all \( x \geq 0 \):
\[
\exp(x) \geq 1 + x.
\]

Therefore,
\[
A_k^{\frac{p}{p+1}} \geq A_1^{\frac{p}{p+1}} \exp\left( \frac{p}{k} \omega(k-1) \right) \geq A_1^{\frac{p}{p+1}} \left( 1 + \frac{p}{k} \omega(k-1) \right) > \frac{p}{p+1} A_1^{\frac{p}{p+1}} \omega.
\]

And we obtain
\[
\left( \frac{\gamma_0 \sigma p+1 (x^*)}{A_k} \right)^{\frac{p+1}{p+1}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{(p+1)}} \leq \left( \frac{\gamma_0 \sigma p+1 (x^*)}{c A_k} \right)^{\frac{p+1}{p+1}} k \delta \leq \left( \frac{\gamma_0 \sigma p+1 (x^*)}{c A_k} \right)^{\frac{p+1}{p+1}} \frac{p+1}{p+1} \kappa \delta
\]
\[
\leq \varepsilon \frac{p}{p+1}.
\]

\[ \Box \]
Estimates (3.13) and (3.17) show that the bound for the inner accuracy \(\delta\) has a reasonable dependency on the absolute accuracy \(\varepsilon\) 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 \(s \in \partial h_{k+1}(v_{k+1})\):

\[
\|s\|_* \leq O\left(\varepsilon^{\frac{1}{p+1}}\right) \iff \|s\|_*^{\frac{p+1}{p}} \leq O(\varepsilon).
\]

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,

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

Therefore, we have

\[
R_k(p, \delta) \leq \left( (\gamma_0 \beta_d(x_0; x^*) )^{\frac{1}{p+1}} + \left( \frac{p+1}{\gamma_0 \sigma_{p+1}(d)} \right)^{\frac{1}{p+1}} \right)^{\frac{p+1}{p}}.
\]

**4. Application of tensor methods.** In this section, let us incorporate the high-order tensor methods [22] into Algorithm 1 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 \(B = B^* > 0\)) 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 \(p\)-th derivative of the smooth component of the objective function is Lipschitz continuous:

\[
\|D^p f(x) - D^p f(y)\| \leq L_p(f)\|x - y\|, \quad x, y \in \text{dom} f,
\]

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

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

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

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 \(x \in \text{dom} f\):

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

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

\[
|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p(f)}{(p+1)}\|y - x\|^{p+1},
\]
(4.4) \[ \| \nabla f(y) - \nabla y \Omega_p(f, x; y) \|_p \leq \frac{L_p(f)}{p^2} \| y - x \|_p. \]

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

\[
h_{k+1}(x) = A_{k+1} f \left( \frac{a_{k+1} x + A_k x_k}{A_{k+1}} \right) + a_{k+1} \psi(x) + \gamma_k \beta(\psi(x); x).
\]

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

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

\[
\min_{x \in \text{dom } h} \left\{ h(x) \equiv g(x) + \phi(x) \right\}
\]

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

\[
h(y) - h(x) - \langle h'(x), y - x \rangle \geq \sigma_d(h) \beta_d(x; y) \geq \frac{\sigma_d(h) 2^{1-p}}{p+1} \| y - x \|^{p+1}.
\]

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

\[
T_M(h; x) \equiv \argmin_{y \in \mathbb{R}} \left\{ \Omega_p(g, x; y) + \frac{M}{(p+1)^2} \| y - x \|^{p+1} + \phi(y) \right\},
\]

and consider the following iteration process,

\[
\begin{align*}
z_{t+1} &= T_M(h; z_t), & t &\geq 0.
\end{align*}
\]

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 \( M \geq p L_p(g) \) the auxiliary optimization problem in (4.8) is convex for all \( p \geq 1 \) (see Theorem 1 in [22]). 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:

\[
\left\langle \nabla_y \Omega_p(g, x; T) + \frac{M}{p^2} \| T - x \|^{p-1} B(T - x), y - T \right\rangle + \phi(y) \geq \phi(T), \quad y \in \text{dom } \phi.
\]

Therefore,

\[
\phi'(T) \equiv -\nabla_y \Omega_p(g, x; T) - \frac{M}{p^2} \| T - x \|^{p-1} B(T - x) \in \partial \phi(T).
\]

This inclusion justifies notation \( h'(T) \equiv \nabla g(T) + \phi'(T) \in \partial h(T) \).
In order to work with these objects, we use the following result (see Lemma 2 in [4]).

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

\[
\langle h'(T), x - T \rangle \geq \left( \frac{p!}{(p+1)!} \right)^{\frac{1}{p}} \cdot \|h'(T)\|_p \cdot \frac{\beta^{p+1}}{p} \cdot \frac{p}{(p^2-1)^{\frac{p}{2}}}.
\]

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

\[
\langle h'(T), x - T \rangle \geq \left( \frac{p!}{(p+1)!} \right)^{\frac{1}{p}} \cdot \|h'(T)\|_p^{p+1}.
\]

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 $x, y \in \text{dom } h$ we have

\[
h(T_M(x)) \leq h(y) + \frac{\beta + 1}{p(p+1)!} \|y - x\|^{p+1}.
\]

**Proof.** Indeed,

\[
h(T_M(x)) = g(T_M(x)) + \phi(T_M(x))
\]

\[
\leq \Omega_p(g, x; T_M(x)) + \frac{M}{p(p+1)!} \|T_M(x) - x\|^{p+1} + \phi(T_M(x))
\]

\[
\leq \Omega_p(g, x; y) + \frac{M}{p(p+1)!} \|y - x\|^{p+1} + \phi(y)
\]

\[
\leq g(y) + \frac{M + L_p(g)}{p(p+1)!} \|y - x\|^{p+1} + \phi(y)
\]

\[
= h(y) + \frac{\beta + 1}{p(p+1)!} \|y - x\|^{p+1}.
\]

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 $y \in \text{dom } h$ we have

\[
\|h'(z_{t+2})\|_{\star}^{\frac{p+1}{p}} \leq \exp \left( -t \cdot \min \left\{ 1, \left[ \frac{p! \sigma_d(h) 2^{1-p}}{p(p+1)! L_p(g)} \right]^{\frac{1}{p}} \right\} \cdot \frac{p}{p+1} \right)
\]

\[
\cdot \left( \frac{p+1}{p!} L_p(g) \right)^{\frac{1}{p}} \cdot \left( h(y) - h^* + \frac{L_p(g)}{p!} \|y - z_0\|^{p+1} \right).
\]

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

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

for any $y \in \text{dom } h$. 

Denote \(x_h^* \overset{\text{def}}{=} \arg\min_{y \in \mathcal{E}} h(y)\) and consider \(y = z_t + \alpha(x_h^* - z_t)\) for \(\alpha \in [0, 1]\). Then we have

\[
h(z_{t+1}) - h^* \leq h(z_t) - h^* - \alpha (h(z_t) - h^*) + \alpha \frac{p+1}{p} L_p(g) \|x_h^* - z_t\|^{p+1}
\]

\[(4.15)\]

Using (4.7),

\[
(1 - \alpha + \frac{\alpha}{p+1} \frac{(p+1)L_p(g)}{p\sigma_d(h_2\eta)}} \cdot (h(z_t) - h^*)
\]

The minimum of the right-hand side is attained at

\[
\alpha^* = \min \left\{ 1, \left[ \frac{p\sigma_d(h_2\eta)}{(p+1)L_p(g)} \right]^{\frac{1}{p}} \right\}
\]

Plugging it into (4.15) gives

\[
h(z_{t+1}) - h^* \leq \left( 1 - \alpha^* \frac{p}{p+1} \right) \cdot (h(z_t) - h^*)
\]

\[(4.16)\]

Therefore, for every \(t \geq 0\) we have

\[
h(z_{t+1}) - h^* \overset{(4.16)}{\leq} \exp \left(-t \alpha^* \frac{p}{p+1} \right) \cdot (h(z_t) - h^*)
\]

\[(4.14)\]

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

\[
h(z_{t+1}) - h^* \geq h(z_{t+1}) - h(z_{t+2})
\]

\[
\geq \langle h'(z_{t+2}), z_{t+1} - z_{t+2} \rangle
\]

\[(4.11)\]

\[
\geq \left( \frac{pL_p(g)}{(p+1)L_p(g)} \right)^{\frac{1}{p}} \cdot \|h'(z_{t+2})\|^{\frac{p+1}{p}}.
\]

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

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

\[
z_{t+1} = T_M(h_{k+1}; z_t), \quad t \geq 0,
\]

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

\[
\|h'_{k+1}(z_{t_k})\|_* \leq \delta_{k+1}
\]
Then for
\[ (4.17) \quad t_k \geq 2 + \max \left\{ 1, \frac{\ell_{k+1}}{\mu_{k+1}} \right\} \cdot \frac{p+1}{p} \cdot \log \left( \frac{\ell_{k+1} D_{k+1}}{\delta_{k+1}} \right), \]
where
\[ (4.18) \quad \ell_{k+1} \overset{\text{def}}{=} \left( \frac{(p+1) L_p(g_{k+1})}{p!} \right)^{\frac{1}{p}}, \quad \mu_{k+1} \overset{\text{def}}{=} \left( g_{k+1} 2^{1-p} \right)^{\frac{1}{p}}, \]
and
\[ (4.19) \quad D_{k+1} \overset{\text{def}}{=} A_k (F(x_k) - F^*) + \gamma_k \beta_d(v_k; x^*) + \left( \frac{\ell_{k+1}}{\mu_{k+1}} \right)^{\beta_d(v_k; x^*)} \]
Proof. By definition, for all \( x \in \text{dom } \psi \), we have
\[ h_{k+1}(x) + A_k \psi(x_k) = A_{k+1} f \left( \frac{a_{k+1} x + A_k x_k}{A_{k+1}} \right) + a_{k+1} \psi(x) + \gamma_k \beta_d(v_k; x) + A_k \psi(x_k) \]
\[ \geq A_{k+1} F \left( \frac{a_{k+1} x + A_k x_k}{A_{k+1}} \right) + \gamma_k \beta_d(v_k; x) \geq A_{k+1} F^*. \]
Therefore,
\[ (4.20) \quad -h^*_{k+1} - A_k \psi(x_k) \leq -A_{k+1} F^*. \]
Then for \( y \equiv x^* \overset{\text{def}}{=} \arg \min_{y \in \mathbb{R}} F(y) \) we obtain
\[ h_{k+1}(y) - h^*_{k+1} + \frac{L_p(g_{k+1})}{p!} \| y - z_0 \|^{p+1} = h_{k+1}(x^*) - h^*_{k+1} + \frac{L_p(g_{k+1})}{p!} \| x^* - v_k \|^{p+1} \]
\[ = A_{k+1} f \left( \frac{a_{k+1} x + A_k x_k}{A_{k+1}} \right) + a_{k+1} \psi(x^*) - h^*_{k+1} + \gamma_k \beta_d(v_k; x^*) \]
\[ + \frac{L_p(g_{k+1})}{p!} \| x^* - v_k \|^{p+1} \]
\[ \leq a_{k+1} F^* + A_k F(x_k) - h^*_{k+1} - A_k \psi(x_k) + \gamma_k \beta_d(v_k; x^*) \]
\[ + \frac{L_p(g_{k+1})}{p!} \| x^* - v_k \|^{p+1} \]
\[ \leq A_k (F(x_k) - F^*) + \gamma_k \beta_d(v_k; x^*) + \frac{L_p(g_{k+1})}{p!} \| x^* - v_k \|^{p+1} \overset{\text{(2.3)}}{\leq} D_{k+1}. \]
It remains to use this bound together with \( (4.13) \) and the following estimation of the strong convexity parameter: \( \sigma_d(h_{k+1}) \overset{\text{def}}{=} \gamma_{k+1} \).

By representation \( (4.5) \), we have a simple relations between Lipschitz constants of the derivatives for function \( g_{k+1}(\cdot) \) and \( f(\cdot) \):
\[ (4.21) \quad L_p(g_{k+1}) = \frac{a_{k+1}^{p+1}}{A_{k+1}^p} L_p(f), \quad p \geq 1. \]
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

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

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 \( c \equiv \frac{p!\gamma_0}{2^{p-1}(p+1)p^p\gamma_0 L_p(f)} \) and \( a_k \equiv c(p+1)k^p \). Then we have

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

and we get

\[
(4.22) \quad \frac{A_{k+1}}{A_k} \leq \frac{c(p+1)^{p+1}}{\frac{p!\gamma_0}{2^{p-1}(p+1)p^p\gamma_0 L_p(f)}}.
\]

Thus we obtain

\[
(4.23) \quad \frac{\ell_{k+1}}{\mu_{k+1}} = \left( \frac{A_{k+1}^{p+1}}{A_k} \cdot \frac{2^{p-1}(p+1)L_p(f)}{p!\gamma_0} \right)^{\frac{1}{p}} \overset{(4.22)}{\leq} \frac{1}{c}.
\]

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

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

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

\[
(4.25) \quad \frac{\ell_{k+1}}{\mu_{k+1}} \leq \left( \frac{a_{k+1}^{p+1}}{A_{k+1}^{p+1}} \cdot \frac{L_p(f)(p+1)2^{p-1}}{p!\sigma_d(\psi)} \right)^{\frac{1}{p}} \leq \omega \cdot \left( \frac{L_p(f)(p+1)2^{p-1}}{p!\sigma_d(\psi)} \right)^{\frac{1}{p+1}} \leq 1.
\]

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.
We are ready to specify the whole optimization procedure.

**Algorithm 2: Contracting Proximal Tensor Method**

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

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

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

$c := \frac{p! \gamma_0}{2^p - 1} \frac{\|x - x_0\|_{p+1}}{L_p(f)}$, $\omega := \min\{\left(\frac{\sigma_d(x_0; x)}{L_p(f)(p+1)^{2p-1}}\right)^{\frac{1}{p+1}}, \frac{1}{2}\}$.

**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:

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

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

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

4. Solve inner subproblem by tensor method up to accuracy $\delta$:

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

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

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

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

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

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 $\delta$ be fixed as follows:

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

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

$$(4.26) \quad K = \left[1 + 2^p \left(\frac{p! (p+1)^{p+2} L_p(f) \beta_d(x_0; x_*)}{\varepsilon p! (p+1)^p (p+1)^{2p+2}}\right)^{\frac{1}{p+1}} \right]$$

iterations of Algorithm 2. The total number of oracle calls $N_K := \sum_{k=1}^K t_k$ is bounded.
as

\[ N_K \leq K \cdot \left( 3 + \frac{p+1}{p} \log \left( 4 \left( 1 + \frac{1}{70} \right) (p+1) \frac{3}{2} K^p \right) \right). \]

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

\[ c = \frac{p^3 \gamma_0}{2(p-1)(p+1)^p \log L_p(f)}. \]

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

\[
t_k \leq 3 + \max \left\{ 1, \frac{\ell_{k+1}}{p_{k+1}} \right\} \cdot \frac{p+1}{p} \cdot \log \left( \frac{\ell_{k+1} D_{k+1}}{\delta + p} \right).
\]

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 \( A_k \):

\[ A_k = c(p+1) \sum_{i=1}^{k} i^p \leq c(p+1) \int_0^{k+1} x^p \, dx = c(k+1)^{p+1}. \]

Therefore, for every \( 0 \leq k \leq K \):

\[
\frac{R_k(p, \delta)}{\delta + p} = \left( \frac{(\gamma_0 \beta_d(x_0; x^*))^{\frac{p}{p+1}}}{\delta} + \frac{(p+1)^{2^p-1}}{\gamma_0} \right) \frac{p+1}{p}.
\]

\[
\leq \left( \frac{(\gamma_0 \beta_d(x_0; x^*))^{\frac{p}{p+1}}}{\delta} + \frac{(p+1)^{2^p-1}}{\gamma_0} \right) \frac{p+1}{p} K.
\]

\[
= \left( \frac{(L_p(f) \beta_d(x_0; x^*) \gamma_0}{p!} \frac{2^p(p+1)^{p+1}}{\gamma_0} + \frac{(p+1)^{2^p-1}}{\gamma_0} \right) \frac{p+1}{p} K.
\]

\[
\leq \left( \frac{(p+1)^{2^p-1}}{\gamma_0} \right) \frac{p+1}{p} K^p \leq 4 \left( \frac{p+1}{\gamma_0} \right) \frac{3}{2} K^p.
\]

This completes the proof. \( \square \)

Now, let us discuss the overall dependence of \( \delta \) and \( K \) on \( p \), given by the claim of Theorem 4.6. For simplicity, we fix \( L_p(f) \gamma^2 = 1 \), \( \beta_d(x_0; x^*) = 1 \), and \( \gamma_0 = 1 \). Thus, we observe the functions

\[ \delta(p) := \frac{(p+1)^{2^p-1}}{2(p+1)^{p+1}} \gamma_0, \quad K(p) := 1 + 2 \frac{1}{p!} \left( \frac{2^p-1}{p!} (p+1)^{p+2} \right). \]

One can see that \( \log_2 \delta(p) \leq -p \). 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)

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

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 \( p \).
CONTRACTING PROXIMAL METHODS

Fig. 1: The dependence of $\delta$ and $K$ on $p$, while $\frac{L_{p}(f)}{\varepsilon}$ and $\beta_{d}(x_{0}; x^{*})$ are fixed.

Theorem 4.7 (Strongly Convex Case). Let $\sigma_{d}(\psi) > 0$ and condition number $\omega$ be defined as in (4.24). Let for a given $\varepsilon > 0$ the inner accuracy $\delta$ be fixed as follows:

\begin{equation}
\delta = \left(\frac{p! \varepsilon}{L_{p}(f)}\right)^{\frac{p}{p+1}} \frac{\gamma(\omega) \varepsilon}{2^{p}(p+1)!} \frac{(p+1)^{2} + 1}{(p+1)^{2} - 1}.
\end{equation}

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

\begin{equation}
K = \left\lfloor 2 + \frac{1}{\omega} L \right\rfloor
\end{equation}

iterations of Algorithm 2, where

\begin{equation}
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! \varepsilon} \right\} \right).
\end{equation}

The total number of oracle calls $N_{K}$ is bounded as follows:

\begin{equation}
N_{K} \leq K \cdot \left( 3 + \left( \frac{e}{(e-1)p} \right) \cdot (1 + L) \right) + \log\left( \max\left\{ 1, \left( \frac{4\sigma_{d}(\psi)p!}{(p+1)\ell_{p}(f)} \right)^{\frac{1}{p}} \right\} \cdot \left( 1 + \frac{1}{70} \right) \cdot \frac{(p+1)^{p+2}}{p^{p+1}} \cdot 2 \cdot \frac{2e^{2+p+4}}{p} \right).
\end{equation}

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

\begin{equation}
\omega \leq \frac{1}{2} \leq \frac{e-1}{e},
\end{equation}

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

\begin{equation}
1 - e^{-1} \alpha \geq \exp(-\alpha),
\end{equation}

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

\begin{equation}
A_{1} \exp(k \omega) \leq A_{k+1} \leq A_{1} \exp\left(k \omega \frac{e}{e-1}\right).
\end{equation}

Now, estimate (4.31) follows directly from (4.34) and (3.16) by using the value $A_{1} = \frac{p! \gamma_{0}}{2^{p-1}(p+1)!} L_{p}(f)$.
By the choice of \( a_{k+1} \), we have \( \frac{\ell_{k+1}}{\mu_{k+1}} \leq 1 \), and we need only to estimate the value under the logarithm in (4.17). For every \( 0 \leq k \leq K \), we have

\[
\ell_{k+1} D_{k+1} \leq \mu_{k+1} R_k(p, \delta) \left( 1 + \frac{1}{\gamma_0} \right) \delta^{\frac{p+1}{p}}
\]

(4.19)

\[
\leq (\gamma_0 + \sigma_d(\psi) A_{k+1}) \frac{1}{2} 2^\frac{k}{p} - 1 \left( 1 + \frac{1}{\gamma_0} \right)
\]

(4.25)

\[
\leq (\gamma_0 + \sigma_d(\psi) A_{K+1}) \frac{1}{2} 2^\frac{k}{p} - 1 \left( 1 + \frac{1}{\gamma_0} \right)
\]

(4.30)

\[
\leq \left( \frac{\gamma_0 \beta_d(x_0, x^*)}{\delta} \right)^{\frac{p}{p+1}} + \left( \frac{(p+1)2^{p-1}}{\gamma_0} \right)^{\frac{1}{p+1}} K
\]

Let us estimate different terms in this expression separately.

1. By definition of \( \omega \), we have

\[
(4.35) \quad \omega^{p+1} \leq \frac{(p+1)p\sigma_d(\psi)}{\gamma_0} A_1.
\]

Therefore,

\[
(4.36) \quad \gamma_0 + \sigma_d(\psi) A_{K+1} \leq \sigma_d(\psi) A_1 \left( \frac{(p+1)p}{\omega^{p+1}} + \exp \left( K\omega \frac{p}{p+1} \right) \right)
\]

(4.31)

\[
\leq 2\sigma_d(\psi) A_1 \exp \left( K\omega \frac{p}{p+1} \right).
\]

2. Substituting the value for \( \delta \), we obtain

\[
(4.31) \quad \leq \frac{(p+1)^2 \sigma_d(x_{0, x^*})}{p\gamma_0} \exp \left( \frac{p}{p+1} \right)
\]

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

\[
K \leq \frac{p+1}{p\omega} \exp \left( K\omega \frac{p}{p+1} \right).
\]
Therefore,

\[ \frac{\ell_{k+1}D_{k+1}}{\delta_{k+1}} \leq \exp\left( K\omega \frac{c}{(e-1)p} \right) \cdot \left( 2^{2-p}\sigma_d(\psi)A_1 \right)^{\frac{1}{p}} \cdot \left( 1 + \frac{1}{\gamma_0} \right) \]

\[ \cdot \left( \frac{\exp\left( K\omega \frac{c}{(e-1)p} \right)}{p\omega^{\gamma_0}} \right) \cdot \left( p + 1 \right)^{2\frac{p}{p+1}} \cdot \left( p + 1 \right)^{\frac{p+2}{p+1}} \cdot \left( 2^{2-p}\sigma_d(\psi)A_1 \right)^{\frac{1}{p}} \]

\[ \cdot \left( 1 + \frac{1}{\gamma_0} \right) \cdot \left( p + 1 \right)^{\frac{p+2}{p}} \cdot 2^{\frac{p+2}{p+1}} \cdot \frac{2^{p+2}}{p^{p+1}} \cdot \frac{1}{\gamma_0} \cdot \frac{4\sigma_d(\psi)p^2}{(p+1)L_p(f)^2} \]

\[ = \exp\left( K\omega \left( \frac{e}{(e-1)p} + 1 \right) \right) \cdot \max\left\{ 1, \left( \frac{4\sigma_d(\psi)p^2}{(p+1)L_p(f)^2} \right)^{\frac{1}{p}} \right\} \]

\[ \cdot \left( 1 + \frac{1}{\gamma_0} \right) \cdot \left( p + 1 \right)^{\frac{p+2}{p}} \cdot 2^{\frac{p+2}{p+1}} \cdot \frac{2^{p+2}}{p^{p+1}}, \]

and we obtain (4.32). ⊓⊔

According to Theorems 4.6 and 4.7, the rate of convergence for the outer iterations of Algorithm 2 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 numerical performance of the contracting proximal method and the classical proximal point algorithm (1.1) for unconstrained minimization of a convex quadratic function:

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

with \( A = A^* \succeq 0 \). We also run the gradient method and the accelerated gradient method for this problem. A typical behaviour 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 \( \delta_k \) 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.)

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

\[ \lambda_i = \frac{1}{1+\exp\left( \frac{\alpha}{n+1-2i} \right)^2}, \quad 1 \leq i \leq n. \]
Therefore it holds that $\lambda_1 = 1/(1 + \exp(\alpha))$ and $\lambda_n = 1/(1 + \exp(-\alpha))$, so parameter $\alpha$ 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.

![Fig. 2: Convergence of first-order methods on quadratic function.](image)

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

| $n$ | $q$ | iter | mat-vec | iter | mat-vec | iter | mat-vec | iter | mat-vec |
|-----|-----|------|---------|------|---------|------|---------|------|---------|
| 500 | $10^{-2}$ | 339  | 339     | 361  | 1044    | 115  | 229     | 74   | 137     |
|     | $10^{-4}$ | 12815 | 12815   | 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    |

We see that the 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 a performance of second-order methods for unconstrained minimization of the following objective:

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

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

We compare the cubically regularized Newton method [23] and its accelerated variant from [19] with the contracting proximal cubic Newton (Algorithm 2 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, $\|x\| = \ldots$
CONTRACTING PROXIMAL METHODS

\langle Bx, x \rangle^{1/2}, \text{ with matrix } B = \sum_{i=1}^{m} a_i a_i^T, \text{ and fix the regularization parameter equal to 1. The results are shown in Table 2.}

|      | Cubic Newton | Accelerated Cubic Newton | Contracting Proximal Cubic Newton |
|------|--------------|--------------------------|-----------------------------------|
| n=50 | 1            | 389                      | 177                               |
|      | 0.1          | 482                      | 202                               |
|      | 0.05         | 886                      | 343                               |
| n=100| 1            | 834                      | 308                               |
|      | 0.1          | 1210                     | 377                               |
|      | 0.05         | 2598                     | 641                               |

|      | Oracle | Oracle | Oracle | Oracle |
|------|--------|--------|--------|--------|
| 50   | 389    | 353    | 112    | 491    |
| 0.1  | 482    | 403    | 141    | 587    |
| 0.05 | 886    | 685    | 236    | 1129   |
| 100  | 834    | 615    | 189    | 849    |
| 0.1  | 1210   | 753    | 232    | 1021   |
| 0.05 | 2598   | 1281   | 397    | 1740   |

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

We see that the 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. Arievani, O. Shamir, and R. Shiff, Oracle complexity of second-order methods for smooth convex optimization, Mathematical Programming, 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. Boule, and M. Teboulle, A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications, Mathematics of Operations Research, 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, 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.
[7] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, J. Qin, 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, Computational Mathematics and Mathematical Physics, 58 (2018), pp. 48–64.
[9] G. N. Grapiglia and Y. Nesterov, Accelerated regularized Newton methods for minimizing composite convex functions, SIAM Journal on Optimization, 29 (2019), pp. 77–99.
[10] O. Güler, On the convergence of the proximal point algorithm for convex minimization, SIAM Journal on Control and Optimization, 29 (1991), pp. 403–419.
[11] O. Güler, New proximal point algorithms for convex minimization, SIAM Journal on Optimization, 2 (1992), pp. 649–664.
[12] A. Ivanova, D. Grishchenko, A. Gasnikov, and E. Shulgin, Adaptive catalyst for smooth convex optimization, arXiv:1911.11271, (2019).
[13] 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.
[14] H. Lin, J. Mairal, and Z. Harchaoui, Catalyst acceleration for first-order convex optimization: from theory to practice, Journal of Machine Learning Research, 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 Journal on Optimization, 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 Journal on Optimization, 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/k²), in Dokl. akad. nauk SSSR, vol. 269, (1983), pp. 543–547.
[19] Y. Nesterov, Accelerating the cubic regularization of Newton’s method on convex problems, Mathematical Programming, 112 (2008), pp. 159–181.
[20] Y. Nesterov, Gradient methods for minimizing composite functions, Mathematical Programming, 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, Mathematical Programming, doi:10.1007/s10107-019-01449-1 (2019).
[23] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton’s method and its global performance, Mathematical Programming, 108 (2006), pp. 177–205.
[24] R. T. Rockafellar, Monotone operators and the proximal point algorithm, SIAM journal on control and optimization, 14 (1976), pp. 877–898.
[25] S. Salzo and S. Villa, Inexact and accelerated proximal point algorithms, Journal of Convex analysis, 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 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, Numerical functional analysis and optimization, 22 (2001), pp. 1013–1035.
[28] Q. Van Nguyen, Forward-backward splitting with bregman distances, Vietnam Journal of Mathematics, 45 (2017), pp. 519–539.