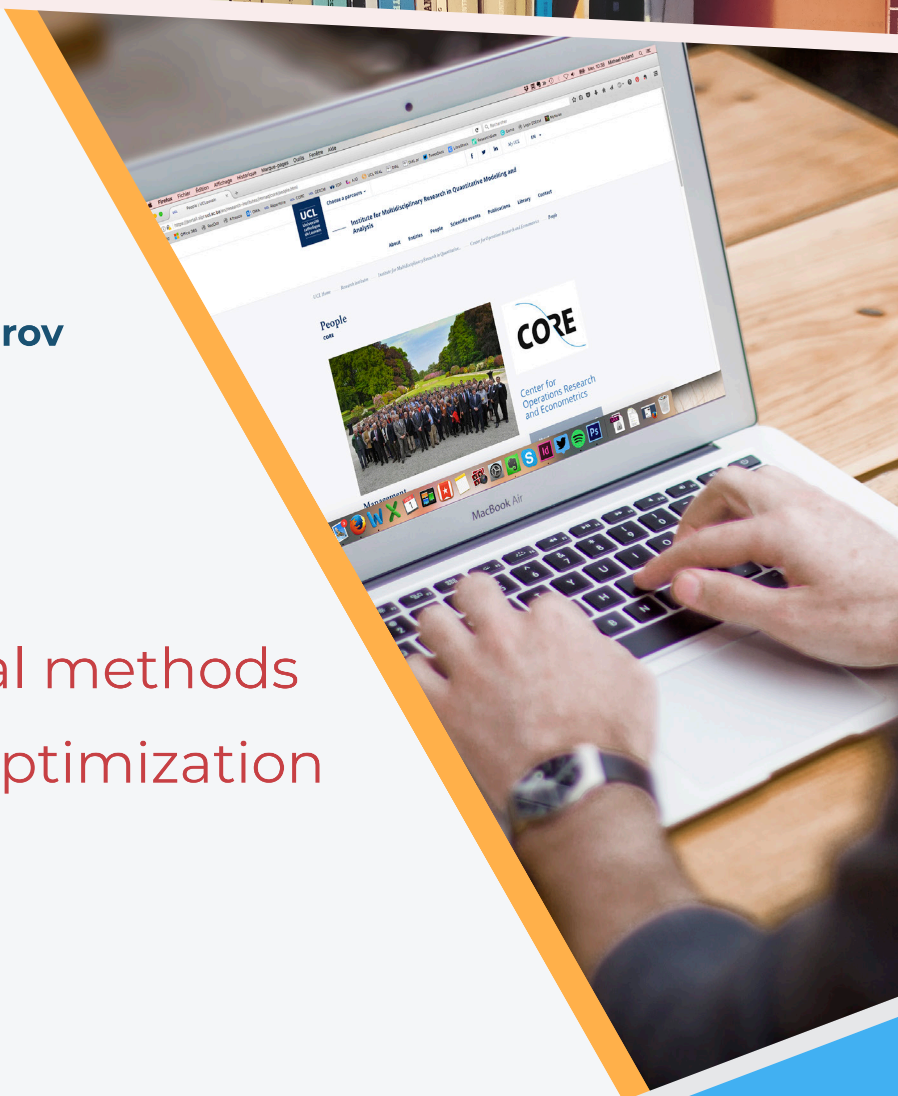
2019/27

DP

Nikita Doikov and Yurii Nesterov



Contracting proximal methods for smooth convex optimization

CORE CENTER FOR OPERATIONS RESEARCH AND ECONOMETRICS

CORE
Voie du Roman Pays 34, L1.03.01
B-1348 Louvain-la-Neuve
Tel (32 10) 47 43 04
Email: immaq-library@uclouvain.be
https://uclouvain.be/en/research-institutes/lidam/core/discussion-papers.html

CORE DISCUSSION PAPER

2019/27

Contracting Proximal Methods for Smooth Convex Optimization *

Nikita Doikov †   Yurii Nesterov ‡

December 16, 2019

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 function. 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.

Keywords: convex optimization, proximal method, accelerated methods, global complexity bounds, high-order algorithms

*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); e-mail: Nikita.Doikov@uclouvain.be.

‡Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL); e-mail: Yurii.Nesterov@uclouvain.be.

1 Introduction

One of the classical iterative methods in theoretical optimization is the Proximal Point Algorithm [18]. This method, as applied to minimizing a convex function $f : \mathrm{dom}\, f \to \mathbb{R}$, consists of solving at each iteration the following subproblem:

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

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 [6, 21, 20] 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 [15].

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 [12]. This is the best possible rate achievable for the first-order black-box optimization on $\mathscr{F}_L^{1,1}$ [11]. An accelerated variant of the Proximal Point Algorithm with the optimal rate of convergence was proposed in [7] (see also [19, 9, 8] 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 [7] 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, which p-th 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 [16].

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:

$$\left.\begin{array}{rcl} v_{k+1} & = & \operatorname*{argmin}_x \Big\{ A_{k+1} f\big(\tfrac{a_{k+1}x + A_k x_k}{A_{k+1}}\big) + \beta_d(v_k; x) \Big\} \\[2ex] x_{k+1} & = & \tfrac{a_{k+1}v_{k+1} + A_k x_k}{A_{k+1}} \end{array}\right\}, \quad k \geq 0. \tag{1.2}$$

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

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:

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

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

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

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

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

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:

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

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

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

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) on an acceptable level.[1] For $f \in \mathscr{F}_L^{1,1}$, the recommended choice of a_{k+1} corresponds to the quadratic equation [12]:

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

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. Indeed, in view of the presence of the regularization term, $\nabla^2 h_{k+1}(x) \succeq I$. On the other hand,

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

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 Euclidean setting (described above) produces the same sequence of points as the accelerated Proximal Point Algorithm

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

from [7]. However, now we can employ also 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 1 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 assumption that p-th derivative ($p \geq 1$) of the smooth part of the objective is Lipschitz continuous. We present fully-defined optimization scheme (Algorithm 2), with incorporated steps of Tensor Method of a certain degree. Resulting algorithm achieves the accelerated rate of convergence, with an additional logarithmic factor for the number of total oracle calls. Final complexity estimate for this scheme is given by Theorems 3 and 4.

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:

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

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

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

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

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

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

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

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

The main example, which naturally appears in Tensor Methods (see [16]) and which we use in Section 4, is the following prox function.

Example 1

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

for some $p \geq 1$. For 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 Lemma 5 in [3]), so it holds:

$$\beta_d(v; x) \;\; \geq \;\; \tfrac{2^{1-p}}{p+1}\|v - x\|^{p+1}, \qquad v, x \in \mathbb{E}. \tag{2.2}$$

For more examples of available prox functions see [1, 10].

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

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

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

We say that function ψ is *strongly convex with respect to* d (see [22, 1, 10]) with constant $\sigma_d(\psi) > 0$, if it holds for all $x, y \in \mathrm{dom}\,\psi$ and for all $\psi'(x) \in \partial\psi(x)$

$$\beta_\psi(x, \psi'(x); y) \;\; \geq \;\; \sigma_d(\psi)\beta_d(x; y). \tag{2.3}$$

Inequality (2.3) 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 [16]).

Example 2 *Let $f : \mathrm{dom}\, f \to \mathbb{R}$ be convex, with Lipschitz continuous third derivative:*

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

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

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

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

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

Then, for Euclidean norm, the function $g(\cdot)$ is strongly convex with respect to the following prox function (see Lemma 4 in [16]):

$$d(h) \;\; \equiv \;\; \tfrac{1}{2}\left(1 - \tfrac{1}{\tau}\right)\langle \nabla^2 f(x)h, h\rangle + \tfrac{\tau(\tau-1)L_3}{8}\|h\|^4.$$

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

$$\beta_{a_1 f_1 + a_2 f_2}(x; y) \quad = \quad a_1 \beta_{f_1}(x; y) + a_2 \beta_{f_2}(x; y), \qquad a_1, a_2 \geq 0. \tag{2.4}$$

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

$$\beta_\ell(x; y) \quad = \quad 0. \tag{2.5}$$

Therefore, from (2.4) and (2.5) we conclude, that

$$\beta_f(x; y) \quad = \quad \beta_d(x; y), \tag{2.6}$$

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:

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

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 $x, y \in \mathrm{dom}\, h$ and every $h'(x) \in \partial h(x)$

$$\begin{array}{rcl} h(y) - h(x) - \langle h'(x), y - x \rangle & = & \beta_h(x; y) \\ & \overset{(2.4),(2.6)}{=} & \beta_g(x; y) + a\beta_\psi(x; y) + \gamma \beta_d(x; y) \\ & \geq & (a\sigma_d(\psi) + \gamma)\beta_d(x; y). \end{array} \tag{2.7}$$

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

$$h(y) \quad \geq \quad h(T) + (a\sigma_d(\psi) + \gamma)\beta_d(T; y). \tag{2.8}$$

3 Contracting Proximal Method

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

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

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

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

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

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

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:

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

and the rate of convergence is determined by the growth of coefficients A_k towards 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.

$$\begin{array}{rl} & \gamma_0 \beta_d(x_0; x) + A_{k+1} F(x) \\ = & \gamma_0 \beta_d(x_0; x) + A_k F(x) + a_{k+1} F(x) \\ \stackrel{(3.1)}{\geq} & \gamma_k \beta_d(v_k; x) + A_k F(x_k) + a_{k+1} F(x) + C_k(x) \\ \geq & \gamma_k \beta_d(v_k; x) + A_{k+1} f\left(\tfrac{a_{k+1} x + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + A_k \psi(x_k) + C_k(x), \end{array} \tag{3.2}$$

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

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

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

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

If we are able to compute the *exact minimum*

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

then by (2.8) we see, that

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

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

$$x_{k+1} \quad \stackrel{\text{def}}{=} \quad \tfrac{a_{k+1} v_{k+1} + A_k x_k}{A_{k+1}}. \tag{3.6}$$

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

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

Now, instead of computing the exact minimum (3.5), let us relax $v_{k+1} \in \mathrm{dom}\,\psi$ 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 small subgradient of h_{k+1} (see [16]).

Thus, we come to the following general scheme.

Algorithm 1: Contracting Proximal Method

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

Iteration $k \geq 0$.

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

2: Denote contracted objective with regularizer:

$$h_{k+1}(x) := A_{k+1} f\left(\tfrac{a_{k+1}x + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x) + \gamma_k \beta_d(v_k; x).$$

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

4: Find $v_{k+1} \in \mathrm{dom}\,\psi$ 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 $\gamma_{k+1} := \gamma_k + a_{k+1}\sigma_d(\psi)$.

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 1 *For some $p \geq 1$, prox-function $d(\cdot)$ is uniformly convex of degree $p+1$ with parameter $\sigma_{p+1}(d) > 0$:*

$$\beta_d(v; x) \;\;\geq\;\; \tfrac{\sigma_{p+1}(d)}{p+1}\|v - x\|^{p+1}, \qquad v, x \in \mathbb{E}. \tag{3.8}$$

Let us write down the convergence guarantees of the method.

Theorem 1 (Convergence of Contracting Proximal Method) *Let Assumption 1 hold. Then for Algorithm 1 at all iterations $k \geq 0$ we have:*

$$A_k\left(F(x_k) - F^*\right) + \gamma_k \beta_d(v_k; x^*) + \sum_{i=1}^{k} \gamma_i \beta_d(v_{i-1}; v_i) \;\;\leq\;\; R_k(p, \delta), \tag{3.9}$$

where

$$R_k(p, \delta) \;\;\stackrel{\mathrm{def}}{=}\;\; \left(\left(\gamma_0 \beta_d(x_0; x^*)\right)^{\frac{p}{p+1}} + \left(\tfrac{p+1}{\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{1/(p+1)}}\right)^{\frac{p+1}{p}}. \tag{3.10}$$

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

$$\begin{array}{rl} A_k\left(F(x_k) - F(x)\right) + \gamma_k \beta_d(v_k; x) + \sum\limits_{i=1}^{k} \gamma_i \beta_d(v_{i-1}; v_i) & \\ \leq \quad \gamma_0 \beta_d(x_0; x) + \sum\limits_{i=1}^{k} \langle s_i, v_i - x \rangle, \quad x \in \mathrm{dom}\, \psi, & \end{array} \tag{3.11}$$

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 (3.11) is exactly (3.1) with

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

Therefore, we have

$$\begin{array}{rl} \gamma_0 \beta_d(x_0; x) + A_{k+1} F(x) & \\ \stackrel{(3.2)}{\geq} & h_{k+1}(x) + A_k \psi(x_k) + C_k(x) \\ \stackrel{(2.7)}{\geq} & h_{k+1}(v_{k+1}) + \langle s_{k+1}, x - v_{k+1} \rangle + \gamma_{k+1} \beta_d(v_{k+1}; x) + A_k \psi(x_k) + C_k(x) \\ = & A_{k+1} f(x_{k+1}) + a_{k+1} \psi(v_{k+1}) + \gamma_{k+1} \beta_d(v_{k+1}; x) + A_k \psi(x_k) + C_{k+1}(x) \\ \geq & A_{k+1} F(x_{k+1}) + \gamma_{k+1} \beta_d(v_{k+1}; x) + C_{k+1}(x). \end{array}$$

This is (3.11) for the next step.

Now, plugging $x \equiv x^*$ into (3.11) and taking into account nonnegativity of all terms in the left-hand side, we get

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

Now, we need to estimate the right-hand side from above. Using uniform convexity (3.8), we conclude that, for every $k \geq 0$

$$\begin{array}{rcl} \frac{\gamma_k \sigma_{p+1}(d)}{p+1} \| v_k - x^* \|^{p+1} & \leq & \gamma_0 \beta_d(x_0; x^*) + \sum\limits_{i=1}^{k} \| s_i \|_* \cdot \| v_i - x^* \| \\ & \stackrel{(3.7)}{\leq} & \gamma_0 \beta_d(x_0; x^*) + \sum\limits_{i=1}^{k} \delta_i \| v_i - x^* \| \quad \equiv \quad \alpha_k. \end{array} \tag{3.12}$$

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

$$\alpha_k \quad = \quad \alpha_{k-1} + \delta_k \| v_k - x^* \| \quad \stackrel{(3.12)}{\leq} \quad \alpha_{k-1} + \delta_k \left(\frac{p+1}{\gamma_k \sigma_{p+1}(d)} \right)^{\frac{1}{p+1}} \alpha_k^{\frac{1}{p+1}}.$$

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

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

Finally, from the last inequality we obtain

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

which is the right-hand side of (3.9). $\square$

We see that accuracies δ_k for subgradients of the subproblems appears in (3.10) 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 (3.9) small enough. Let us consider the simplest case, when all δ_k are the same.

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

$$A_k \quad \geq \quad c k^{p+1}, \quad k \geq 1, \tag{3.13}$$

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

$$k \quad \geq \quad \left(\frac{\gamma_0 \beta_d(x_0; x_*)}{c\varepsilon}\right)^{\frac{1}{p+1}} 2^{\frac{1}{p}} \qquad \textit{and} \qquad \delta \quad \leq \quad \frac{(c\varepsilon)^{\frac{p}{p+1}}}{2} \left(\frac{\gamma_0 \sigma_{p+1}(d)}{p+1}\right)^{\frac{1}{p+1}} \tag{3.14}$$

we have

$$R_k(p, \delta) \quad \leq \quad \varepsilon A_k. \tag{3.15}$$

Consequently, by (3.9) *we have $F(x_k) - F^* \leq \varepsilon$.*

Proof:
Indeed,

$$\left(\frac{\gamma_0 \beta_d(x_0; x^*)}{A_k}\right)^{\frac{p}{p+1}} \quad \overset{(3.13)}{\leq} \quad \left(\frac{\gamma_0 \beta_d(x_0; x^*)}{c}\right)^{\frac{p}{p+1}} k^p \quad \overset{(3.14)}{\leq} \quad \frac{\varepsilon^{\frac{p}{p+1}}}{2},$$

and

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

$$\leq \quad \frac{\left(\frac{p+1}{\gamma_0 \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} \delta}{c^{\frac{p}{p+1}}} \quad \overset{(3.14)}{\leq} \quad \frac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

Summing up these two inequalities we obtain (3.15). $\square$

Corollary 2 *Let $\delta_k = \delta > 0$ for all $k \geq 1$. Let the coefficients A_k grow linearly:*

$$A_k \;\; \geq \;\; A_1 \exp\bigl(\omega(k-1)\bigr), \quad k \geq 1, \tag{3.16}$$

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

$$k \;\; \geq \;\; 1 + \tfrac{1}{\omega} \log\left(\tfrac{\gamma_0 \beta_d(x_0; x^*)}{A_1 \varepsilon} 2^{(p+1)/p} \right) \tag{3.17}$$

and

$$\delta \;\; \leq \;\; \tfrac{(A_1 \varepsilon)^{\frac{p}{p+1}} \omega}{2} \cdot \tfrac{p}{p+1} \cdot \left(\tfrac{\gamma_0 \sigma_{p+1}(d)}{p+1} \right)^{\frac{1}{p+1}} \tag{3.18}$$

we have

$$R_k(p, \delta) \;\; \leq \;\; \varepsilon A_k. \tag{3.19}$$

Consequently, by (3.9) we have $F(x_k) - F^* \leq \varepsilon$.

Proof:
Indeed,

$$\left(\tfrac{\gamma_0 \beta_d(x_0; x^*)}{A_k} \right)^{\frac{p}{p+1}} \stackrel{(3.16)}{\leq} \left(\tfrac{\gamma_0 \beta_d(x_0; x^*)}{A_1 \exp\left(\omega(k-1)\right)} \right)^{\frac{p}{p+1}} \stackrel{(3.18)}{\leq} \tfrac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

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

$$\exp(x) \;\; \geq \;\; 1 + x. \tag{3.20}$$

Therefore,

$$\tfrac{A_k^{\frac{p}{p+1}}}{k} \stackrel{(3.16)}{\geq} \tfrac{A_1^{\frac{p}{p+1}} \exp\left(\frac{p}{p+1}\omega(k-1)\right)}{k} \stackrel{(3.20)}{\geq} \tfrac{A_1^{\frac{p}{p+1}} \left(1 + \frac{p}{p+1}\omega(k-1)\right)}{k} \;\; > \;\; \tfrac{p}{p+1} A_1^{\frac{p}{p+1}} \omega. \tag{3.21}$$

And we obtain

$$\frac{\left(\frac{p+1}{\sigma_{p+1}(d)}\right)^{\frac{1}{p+1}}}{A_k^{\frac{p}{p+1}}} \sum_{i=1}^{k} \frac{\delta_i}{\gamma_i^{1/(p+1)}} \;\; \leq \;\; \frac{\left(\frac{p+1}{\gamma_0 \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} k\delta}{A_k^{\frac{p}{p+1}}} \stackrel{(3.21)}{<} \frac{\left(\frac{p+1}{\gamma_0 \sigma_{p+1}(d)}\right)^{\frac{1}{p+1}} p+1\delta}{A_1^{\frac{p}{p+1}} p\omega} \stackrel{(3.18)}{\leq} \frac{\varepsilon^{\frac{p}{p+1}}}{2}.$$

$\square$

Estimates (3.14) and (3.18) 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, on step 4 of the algorithm we need to find a point v_{k+1} with subgradient $s \in \partial h_{k+1}(v_{k+1})$:

$$\|s\|_* \;\leq\; O\left(\varepsilon^{\frac{p}{p+1}}\right) \;\; \Leftrightarrow \;\; \|s\|_*^{\frac{p+1}{p}} \;\leq\; O(\varepsilon).$$

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

$$\sum_{i=1}^{k} \delta_i \;=\; c\Big(1 + \sum_{i=2}^{k} \tfrac{1}{i^s}\Big) \;\le\; c\Big(1 + \int\limits_{1}^{+\infty} \tfrac{dx}{x^s}\Big) \;=\; \tfrac{cs}{s-1}.$$

Therefore, we have

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

4 Application of Tensor Methods

In this section, let us incorporate the high-order Tensor methods [16] 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 $B = B^* \succ 0$) 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 2 *For fixed* $p \ge 1$, *the* p*-th derivative of the smooth component of the objective function is Lipschitz continuous:*

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

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

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

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

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

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

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

By Assumption 2, we are able to bound its accuracy in the following way: for all $x, y \in \mathrm{dom}\, f$ it holds

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

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

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

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

This is a sum of two convex functions: smooth component g_{k+1}, and possibly nonsmooth 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:

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

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 $x, y \in \mathrm{dom}\, h$ and all $h'(x) \in \partial h(x)$ we have:

$$h(y) - h(x) - \langle h'(x), y - x \rangle \quad \geq \quad \sigma_d(h) \beta_d(x; y) \overset{(2.2)}{\geq} \tfrac{\sigma_d(h) 2^{1-p}}{p+1} \|y - x\|^{p+1}. \tag{4.7}$$

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

$$T_M(h; x) \quad \overset{\mathrm{def}}{=} \quad \operatorname*{argmin}_{y \in \mathbb{E}} \Big\{ \Omega_p(g, x; y) + \tfrac{M}{(p+1)!} \|y - x\|^{p+1} + \phi(y) \Big\}, \tag{4.8}$$

and consider the following iteration process:

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

For $p = 1$, the point (4.8) is used in the Composite Gradient Method [14]. For $p = 2$, this is a step of Composite Cubic Newton [4, 5]. It can be shown that for $M \geq pL_p(g)$ the auxiliary optimization problem in (4.8) is convex for *all* $p \geq 1$ (see Theorem 1 in [16]). Therefore it can be efficiently solved by different techniques of Convex Optimization and Linear Algebra (see also [17, 16]).

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

$$\Big\langle \nabla_y \Omega_p(g, x; T) + \tfrac{M}{p!} \|T - x\|^{p-1} B(T - x), y - T \Big\rangle + \phi(y) \quad \geq \quad \phi(T), \qquad y \in \mathrm{dom}\, \phi.$$

Therefore,

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

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

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

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

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

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

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

Next lemma describes the global behavior of method (4.9).

Lemma 2 *Let $\beta \geq 1$ and $M = \beta L_p(g)$. Then for any $x, y \in \mathrm{dom}\, h$ we have:*

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

Proof:
Indeed,

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

□

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

Theorem 2 (Convergence of Tensor Method) *Let $M = pL_p(g)$. Then, for every $t \geq 0$ and $y \in \mathrm{dom}\, h$ we have*

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

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

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

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

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

The minimum of the right-hand side is attained at

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

Plugging it into (4.15) gives

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

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

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

for every $y \in \operatorname{dom} h$. It remains to use Lemma 1 and finish the proof:

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

□

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 *Let we 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

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

for

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

where

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

and

$$
\begin{aligned}
D_{k+1} &\overset{\text{def}}{=} A_k(F(x_k)-F^*) + \gamma_k \beta_d(v_k;x^*) + \left(\tfrac{\ell_{k+1}}{\mu_{k+1}}\right)^p \beta_d(v_k;x^*) \\
&\overset{(3.9)}{\leq} R_k(p,\delta)\cdot \left(1+\tfrac{1}{\gamma_0}\left(\tfrac{\ell_{k+1}}{\mu_{k+1}}\right)^p\right).
\end{aligned}
\tag{4.19}
$$

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

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

Therefore,

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

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

$$
\begin{aligned}
& h_{k+1}(y) - h_{k+1}^* + \tfrac{L_p(g_{k+1})}{p!}\|y-z_0\|^{p+1} \\
&= h_{k+1}(x^*) - h_{k+1}^* + \tfrac{L_p(g_{k+1})}{p!}\|x^*-v_k\|^{p+1} \\
&= A_{k+1} f\left(\tfrac{a_{k+1}x^* + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(x^*) - h_{k+1}^* + \gamma_k\beta_d(v_k;x^*) \\
&\quad + \tfrac{L_p(g_{k+1})}{p!}\|x^*-v_k\|^{p+1} \\
&\leq a_{k+1}F^* + A_kF(x_k) - h_{k+1}^* - A_k\psi(x_k) + \gamma_k\beta_d(v_k;x^*) + \tfrac{L_p(g_{k+1})}{p!}\|x^*-v_k\|^{p+1} \\
&\overset{(4.20)}{\leq} A_k(F(x_k)-F^*) + \gamma_k\beta_d(v_k;x^*) + \tfrac{L_p(g_{k+1})}{p!}\|x^*-v_k\|^{p+1} \quad \overset{(2.2)}{\leq} \quad D_{k+1}.
\end{aligned}
$$

It remains to use this bound together with (4.13) and the following estimation of strong convexity parameter:

$$
\sigma_d(h_{k+1}) \quad \overset{(3.4)}{\geq} \quad \gamma_{k+1}.
$$

□

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

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

Therefore, we can control the condition number of our objective. Indeed, by (4.17), the main complexity factor in 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

$$A_k \;=\; c(p+1)\sum_{i=1}^{k} i^p \;\geq\; c(p+1)\int\limits_0^k x^p dx \;=\; ck^{p+1},$$

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

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

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

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

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.

We are ready to specify the whole optimization procedure.

Algorithm 2: Contracting Proximal Tensor Method

Initialization. Choose $x_0 \in \mathrm{dom}\, F$, inner accuracy $\delta > 0$, $\gamma_0 > 0$.

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

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

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

Iteration $k \geq 0$.

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 δ:

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

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 3 (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

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

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

$$N_K \leq K \cdot \left(3 + \tfrac{p+1}{p} \log\Bigl(4\bigl(1 + \tfrac{1}{\gamma_0}\bigr)(p+1)^{\frac{1}{p}} K^p \Bigr) \right). \tag{4.27}$$

Proof:
Estimate (4.26) follows directly from (3.14), by substituting the value

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

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

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

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:

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

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

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

This completes the proof. $\square$

Theorem 4 (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:*

$$\delta \;=\; \left(\tfrac{p!\,\varepsilon}{L_p(f)}\right)^{\frac{p}{p+1}} \frac{\gamma_0 p\,\omega}{2^p(p+1)^{((p+1)^2+1)/(p+1)}}. \tag{4.29}$$

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

$$K \;=\; \left\lfloor 2 + \tfrac{1}{\omega}\mathcal{L} \right\rfloor \tag{4.30}$$

iterations of Algorithm 2, where

$$\mathcal{L} \;\stackrel{\mathrm{def}}{=}\; \log\Big(\max\Big\{ \tfrac{(p+1)^p}{\omega^{p+1}}, \tfrac{L_p(f)\beta_d(x_0;x^*)(p+1)^{p+1}2^{p+\frac{1}{p}}}{p!\,\varepsilon} \Big\}\Big).$$

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

$$\begin{aligned} N_K \;\leq\; K \cdot \Bigg(& 3 + \left(1 + \tfrac{e}{(e-1)p}\right) \cdot (1 + \mathcal{L}) \\ & + \log\Bigg(\max\Big\{1, \left(\tfrac{4\sigma_d(\psi)p!}{(p+1)L_p(f)}\right)^{\frac{1}{p}}\Big\} \cdot \left(1 + \tfrac{1}{\gamma_0}\right) \cdot \frac{(p+1)^{\frac{p+2}{p}}}{p^{\frac{p+1}{p}}} \cdot 2^{\frac{2p^2+p+4}{p}}\Bigg)\Bigg). \end{aligned} \tag{4.31}$$

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

$$\omega \;\leq\; \tfrac{1}{2} \;\leq\; \tfrac{e-1}{e}, \tag{4.32}$$

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

$$1 - \tfrac{e-1}{e}\alpha \;\geq\; \exp(-\alpha),$$

taking $\alpha = \omega\frac{e}{e-1} \stackrel{(4.32)}{\leq} 1$ we obtain $A_{k+1} \leq A_k \exp\big(\omega\frac{e}{e-1}\big)$. Therefore we have, for all $k \geq 0$:

$$A_1 \exp\big(k\omega\big) \;\leq\; A_{k+1} \;\leq\; A_1 \exp\Big(k\omega\tfrac{e}{e-1}\Big). \tag{4.33}$$

Now, estimate (4.30) follows directly from (4.33) and (3.17) 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 $\frac{\ell_{k+1}}{\mu_{k+1}} \stackrel{(4.25)}{\leq} 1$, and we need only to estimate the value under the logarithm in (4.17). For every $0 \leq k \leq K$, we have:

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

Let us estimate different terms in this expression separately.

1. By definition of ω, we have

$$\omega^{p+1} \leq \frac{(p+1)^p\sigma_d(\psi)A_1}{\gamma_0}. \tag{4.34}$$

   Therefore,

$$\begin{array}{rcl}
\gamma_0 + \sigma_d(\psi)A_{K+1} & \overset{(4.34),(4.33)}{\leq} & \sigma_d(\psi)A_1\left(\frac{(p+1)^p}{\omega^{p+1}} + \exp\left(K\omega\frac{e}{e-1}\right)\right) \\
& \overset{(4.30)}{\leq} & 2\sigma_d(\psi)A_1\exp\left(K\omega\frac{e}{e-1}\right).
\end{array}$$

2. Substituting the value for δ, we obtain

$$\begin{array}{rcl}
\frac{(\gamma_0\beta_d(x_0;x^*))^{\frac{p}{p+1}}}{\delta} & \overset{(4.29)}{=} & \left(\frac{L_p(f)\beta_d(x_0;x^*)}{p!\,\varepsilon}\right)^{\frac{p}{p+1}} \frac{2^p(p+1)^{((p+1)^2+1)/(p+1)}}{p\,\omega\gamma_0^{\frac{1}{p+1}}} \\
& \overset{(4.30)}{\leq} & \frac{(p+1)^2 2^{(2p^2+p+1)/(p+1)}}{p\,\omega\gamma_0^{\frac{1}{p+1}}}\exp\left(K\omega\frac{p}{p+1}\right).
\end{array}$$

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,

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

and we obtain (4.31). $\square$

5 Numerical Examples

5.1 Quadratic function

Let us compare numerical performance of Contracting Proximal Method and the classical Proximal Point Algorithm (1.1) for unconstrained minimization of a convex quadratic function:

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

with $A = A^* \succeq 0$. We also run the Gradient Method and the Accelerated Gradient Method for this problem. A typical behaviour of the algorithms is shown on Figure 1. 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 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 $\alpha > 0$

$$
\lambda_i \quad = \quad \frac{1}{1+\exp\Big(\frac{\alpha}{n-1}(n+1-2i)\Big)}, \qquad 1 \leq i \leq n.
$$

Therefore it holds: $\lambda_1 = 1/(1+\exp(\alpha))$ 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.

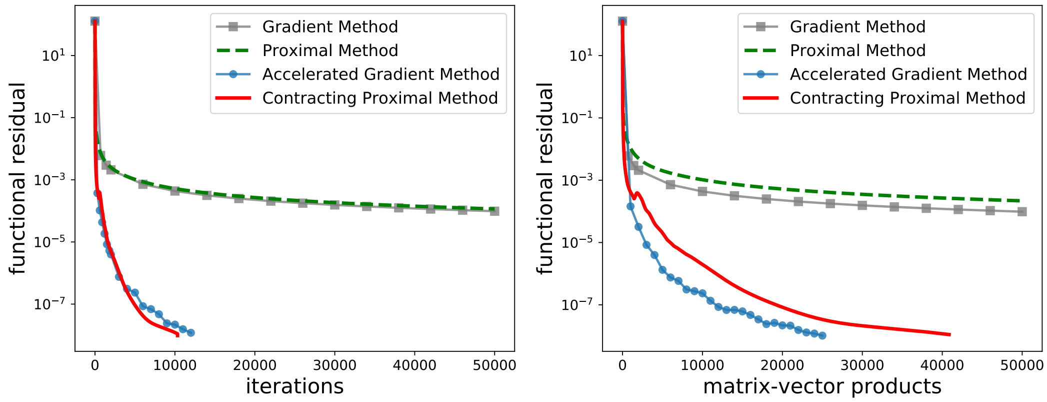


Figure 1: Convergence of first-order methods on quadratic function.

		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

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

We see that Contracting Proximal Method is *always better* than the usual Proximal Algorithm. It requires about the same number of iteration 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 performance of second-order methods for unconstrained minimization of the following objective

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

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

We compare cubically regularized Newton method [17] and its accelerated variant from [13] 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: $\|x\| = \langle Bx, x\rangle^{1/2}$, with matrix $B = \sum_{i=1}^m a_i a_i^T$, and fix regularization parameter being equal 1.

The results are shown in Table 2.

		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

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

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.

References

[1] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[2] Nikita Doikov and Yurii Nesterov. Local convergence of tensor methods. *CORE Discussion Paper 2019/21*, 2019.

[3] Nikita Doikov and Yurii Nesterov. Minimizing uniformly convex functions by cubic regularization of Newton method. *arXiv preprint arXiv:1905.02671*, 2019.

[4] Nikita Doikov and Peter Richtárik. Randomized block cubic Newton method. In *International Conference on Machine Learning*, pages 1289–1297, 2018.

[5] Geovani N Grapiglia and Yurii Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM Journal on Optimization*, 29(1):77–99, 2019.

[6] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[7] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[8] Anastasiya Ivanova, Dmitry Grishchenko, Alexander Gasnikov, and Egor Shulgin. Adaptive catalyst for smooth convex optimization. *arXiv preprint arXiv:1911.11271*, 2019.

[9] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[10] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[11] Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

[12] Yurii Nesterov. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[13] Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[14] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[15] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[16] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, doi:10.1007/s10107-019-01449-1, 2019.

[17] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[18] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[19] Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex analysis*, 19(4):1167–1192, 2012.

[20] Mark Schmidt, Nicolas L Roux, and Francis R Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

[21] Mikhail V Solodov and Benar F Svaiter. A unified framework for some inexact proximal point algorithms. *Numerical functional analysis and optimization*, 22(7-8):1013–1035, 2001.

[22] Quang Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.