# AFFINE-INVARIANT CONTRACTING-POINT METHODS FOR CONVEX OPTIMIZATION

Nikita Doikov, Yurii Nesterov

REPRINT | 3240

UCLouvain

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

Mathematical Programming
https://doi.org/10.1007/s10107-021-01761-9

FULL LENGTH PAPER

Series A

# Affine-invariant contracting-point methods for Convex Optimization

**Nikita Doikov**[1] · **Yurii Nesterov**[2]

Received: 16 November 2020 / Accepted: 19 December 2021


**Abstract**
In this paper, we develop new affine-invariant algorithms for solving composite convex minimization problems with bounded domain. We present a general framework of Contracting-Point methods, which solve at each iteration an auxiliary subproblem restricting the smooth part of the objective function onto contraction of the initial domain. This framework provides us with a systematic way for developing optimization methods of different order, endowed with the global complexity bounds. We show that using an appropriate affine-invariant smoothness condition, it is possible to implement one iteration of the Contracting-Point method by one step of the pure tensor method of degree $p \geq 1$. The resulting global rate of convergence in functional residual is then $\mathcal{O}(1/k^p)$, where $k$ is the iteration counter. It is important that all constants in our bounds are *affine-invariant*. For $p = 1$, our scheme recovers well-known Frank–Wolfe algorithm, providing it with a new interpretation by a general perspective of tensor methods. Finally, within our framework, we present efficient implementation and total complexity analysis of the inexact second-order scheme ($p = 2$), called Contracting Newton method. It can be seen as a proper implementation of the *trust-region idea*. Preliminary numerical results confirm its good practical performance both in the number of iterations, and in computational time.

**Keywords** Convex optimization · Frank–Wolfe algorithm · Newton method · Trust region methods · Tensor methods · Global complexity bounds

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 788368).

✉ Nikita Doikov
Nikita.Doikov@uclouvain.be

Yurii Nesterov
Yurii.Nesterov@uclouvain.be

[1] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

[2] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

Published online: 10 January 2022

**Mathematics Subject Classification** 90C25 · 90C06 · 65K05

## 1 Introduction

**Motivation** In the last years, we can see an increasing interest in new frameworks for derivation and justification of different methods for Convex Optimization, provided with a worst-case complexity analysis (see, for example, [3,4,6,11,14,15,18,20–22]). It appears that the accelerated proximal tensor methods [2,20] can be naturally explained through the framework of high-order proximal-point schemes [21] requiring solution of nontrivial auxiliary problem at every iteration.

This possibility serves as a departure point for the results presented in this paper. Indeed, the main drawback of proximal tensor methods consists in necessity of using a fixed Euclidean structure for measuring distances between points. However, the multi-dimensional Taylor polynomials are defined by directional derivatives, which are affine-invariant objects. Can we construct a family of tensor methods, which do not depend on the choice of the coordinate system in the space of variables? The results of this paper give a positive answer on this question.

Our framework extends the initial results presented in [8,18]. In [18], it was shown that the classical Frank–Wolfe algorithm can be generalized onto the case of the composite objective function [17] using a contraction of the feasible set towards the current test point. This operation was used there also for justifying a second-order method with contraction, which looks similar to the classical trust-region methods [5], but with asymmetric trust region. The convergence rates for the second-order methods with contractions were significantly improved in [8]. In this paper, we extend the contraction technique onto the whole family of tensor methods. However, in the vein of [21], we start first from analysing a conceptual scheme solving at each iteration an auxiliary optimization problem formulated in terms of the initial objective function.

The results of this work can be also seen as an affine-invariant counterpart of Contracting Proximal Methods from [6]. In the latter algorithms, one needs to fix the prox function which is suitable for the geometry of the problem, in advance. The parameters of the problem class are also usually required. Last but not least, all methods from this work do not fix a particular prox function and they are parameter-free.

**Contents** The paper is organized as follows.

In Sect. 2, we present a general framework of Contracting-Point methods. We provide two conceptual variants of our scheme for different conditions of inexactness for the solution of the subproblem: using a point with small residual in the function value, and using a stronger condition which involves the gradients. For both schemes we establish global bounds for the functional residual of the initial problem. These bounds lead to global convergence guarantees under a suitable choice of the parameters. For the scheme with the second condition of inexactness, we also provide a computable accuracy certificate. It can be used to estimate the functional residual directly within the method.

Section 3 contains smoothness conditions, which are useful to analyse affine-invariant high-order schemes. We present some basic inequalities and examples, related to the new definitions.

In Sect. 4, we show how to implement one iteration of our methods by computing an (inexact) affine-invariant tensor step. For the methods of degree $p \geq 1$, we establish global convergence in the functional residual of the order $\mathcal{O}(1/k^p)$, where $k$ is the iteration counter. For $p = 1$, this recovers a well-known result about global convergence of the classical Frank–Wolfe algorithm [10,18]. For $p = 2$, we obtain Contracting-Domain Newton Method from [8]. Thus, our analysis also extends the results from these works to the case, when the corresponding subproblem is solved inexactly.

In Sect. 5, we present a two-level optimization scheme, called Inexact Contracting Newton Method. This is an implementation of the inexact second-order method, in which the steps are computed by the first-order Conditional Gradient Method. For the resulting algorithm, we establish global complexity $\mathcal{O}(1/\varepsilon^{1/2})$ calls of the *smooth part oracle* (computing gradient and Hessian of the smooth part of the objective), and $\mathcal{O}(1/\varepsilon)$ calls of the *linear minimization oracle* of the composite part, where $\varepsilon > 0$ is the required accuracy in the functional residual. Additionally, we address effective implementation of our method for optimization over the standard simplex.

Section 6 contains numerical experiments.

In Sect. 7, we discuss our results and highlight some open questions for the future research.

**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 $\langle s, x \rangle$.

For a smooth function $f : \operatorname{dom} f \to \mathbb{R}$, where $\operatorname{dom} f \subseteq \mathbb{E}$, we denote by $\nabla f(x)$ its gradient and by $\nabla^2 f(x)$ its Hessian, evaluated at point $x \in \operatorname{dom} f \subseteq \mathbb{E}$. Note that

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

for all $x \in \operatorname{dom} f$ and $h \in \mathbb{E}$. For $p \geq 1$, we denote by $D^p f(x)[h_1, \ldots, h_p]$ the $p$th directional derivative of $f$ along directions $h_1, \ldots, h_p \in \mathbb{E}$. Note that $D^p f(x)$ is a $p$-linear symmetric form on $\mathbb{E}$. If $h_i = h$ for all $1 \leq i \leq p$, a shorter notation $D^p f(x)[h]^p$ is used. For its gradient in $h$, we use the following notation:

$$D^p f(x)[h]^{p-1} \stackrel{\text{def}}{=} \frac{1}{p} \nabla D^p f(x)[h]^p \in \mathbb{E}^*, \quad h \in \mathbb{E}.$$

In particular, $D^1 f(x)[h]^0 \equiv \nabla f(x)$, and $D^2 f(x)[h]^1 \equiv \nabla^2 f(x)h$.

## 2 Contracting-point methods

Consider the following composite minimization problem

$$F^* \stackrel{\text{def}}{=} \min_{x \in \operatorname{dom} \psi} \Big[ F(x) = f(x) + \psi(x) \Big], \tag{1}$$

where $\psi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is a *simple* proper closed convex function with *bounded domain*, and function $f(x)$ is convex and $p$ ($\geq 1$) times continuously differentiable at every point $x \in \operatorname{dom} \psi$.

The main requirement is that $\psi$ should have a *simple* structure, which means that corresponding auxiliary subproblems are efficiently solvable. We will see examples of subproblems when discussing implementation of the methods. Typically, we substitute some polynomial model for $f$, while the composite component remains unchanged.

In this section, we propose a conceptual optimization scheme for solving (1) and provide the motivation for the idea. At each step of our method, we choose a contracting coefficient $\gamma_k \in (0, 1]$ restricting the nontrivial part of our objective $f(\cdot)$ onto a *contracted domain*. At the same time, the domain for the composite part remains unchanged.

Namely, at point $x_k \in \operatorname{dom} \psi$, define

$$S_k(y) \stackrel{\text{def}}{=} \gamma_k \psi\big(x_k + \frac{1}{\gamma_k}(y - x_k)\big), \quad y = x_k + \gamma_k(v - x_k), \quad v \in \operatorname{dom} \psi.$$

Note that $S_k(y) = \gamma_k \psi(v)$. Consider the following *exact* iteration:

$$\boxed{\begin{array}{l} v_{k+1}^* \in \operatorname{Argmin}_v \Big\{ f(y) + S_k(y) : \; y = (1 - \gamma_k)x_k + \gamma_k v, \; v \in \operatorname{dom} \psi \Big\}, \\ x_{k+1}^* = (1 - \gamma_k)x_k + \gamma_k v_{k+1}^*. \end{array}} \tag{2}$$

Of course, when $\gamma_k = 1$, exact step from (2) solves the initial problem. However, we are going to look at the *inexact* minimizer. In this case, the choice of $\{\gamma_k\}_{k \geq 0}$ should take into account the efficiency of solving the auxiliary subproblem.

Let us consider the function $v \mapsto g_k(v) := f((1 - \gamma_k)x_k + \gamma_k v)$. Note that its derivatives are as follows:

$$D^q g_k(v) = \gamma_k^q D^q f((1 - \gamma_k)x_k + \gamma_k v), \quad q \geq 1. \tag{3}$$

The smoothness characteristics of the objective (i.e. the Lipschitz constants) are defined by using the derivatives. Hence, we can hope that the smoothness properties of $g_k(\cdot)$ can be better than those of $f(\cdot)$, when $\gamma_k < 1$. We see from (3) that for smaller $\gamma_k$ we have smaller derivatives. The idea is to choose $\gamma_k$ to make a trade-off between the smoothness and the quality of approximation of the initial objective. The result of employing the contracted objective should be combined with the progress made by an optimization algorithm up to the current iterate $x_k$.

Denote by $F_k(\cdot)$ the objective in the auxiliary problem (2), that is

$$F_k(y) \stackrel{\text{def}}{=} f(y) + S_k(y), \quad y = (1 - \gamma_k)x_k + \gamma_k v, \quad v \in \operatorname{dom} \psi.$$

Let us fix a point $\bar{v}_{k+1} \in \operatorname{dom} \psi$ that is an approximate minimizer of $F_k$ in $v$. Thus, we assume that the point $\bar{x}_{k+1} = (1 - \gamma_k)x_k + \gamma_k \bar{v}_{k+1}$ have a *small residual* in the function value:

$$F_k(\bar{x}_{k+1}) - F_k(x_{k+1}^*) \leq \delta_{k+1}, \tag{4}$$

with some fixed $\delta_{k+1} \geq 0$.

**Lemma 1** *For all $k \geq 0$ and $v \in \operatorname{dom} \psi$, we have*

$$F(\bar{x}_{k+1}) \leq (1-\gamma_k)F(x_k) + \gamma_k F(v) + \delta_{k+1}. \tag{5}$$

***Proof*** Indeed, for any $v \in \operatorname{dom} \psi$, we have

$$\begin{aligned} F_k(\bar{x}_{k+1}) &\overset{(4)}{\leq} F_k(x^*_{k+1}) + \delta_{k+1} \\ &\overset{(2)}{\leq} f((1-\gamma_k)x_k + \gamma_k v) + S_k((1-\gamma_k)x_k + \gamma_k v) + \delta_{k+1} \\ &\leq (1-\gamma_k)f(x_k) + \gamma_k f(v) + \gamma_k \psi(v) + \delta_{k+1}. \end{aligned}$$

Therefore,

$$\begin{aligned} F(\bar{x}_{k+1}) &= F_k(\bar{x}_{k+1}) + \psi(\bar{x}_{k+1}) - \gamma_k \psi(\bar{v}_{k+1}) \\ &\leq (1-\gamma_k)f(x_k) + \gamma_k F(v) + \delta_{k+1} + \psi(\bar{x}_{k+1}) - \gamma_k \psi(\bar{v}_{k+1}) \\ &\leq (1-\gamma_k)F(x_k) + \gamma_k F(v) + \delta_{k+1}. \end{aligned}$$

□

Let us write down our method in an algorithmic form.

**Conceptual Contracting − Point Method, I** (6)

**Initialization**. Choose $x_0 \in \operatorname{dom} \psi$.

**Iteration** $k \geq 0$.

1: Choose $\gamma_k \in (0, 1]$.

2: For some $\delta_{k+1} \geq 0$, find $\bar{x}_{k+1}$ satisfying(4).

3: If $F(\bar{x}_{k+1}) \leq F(x_k)$, then set $x_{k+1} = \bar{x}_{k+1}$. Else choose $x_{k+1} = x_k$.

In Step 3 of this method, we add a simple test for ensuring monotonicity in the function value. This step is optional. Moreover, looking at algorithm (6), one may think that we are forgetting the points $\bar{x}_{k+1}$ when the function value is increasing: $F(\bar{x}_{k+1}) > F(x_k)$, and thus we are losing some computations. However, even if point $\bar{x}_{k+1}$ has not been taken as $x_{k+1}$, we shall use it internally as a starting point for computing the next $\bar{x}_{k+2}$ (see also [9] for the concept of monotone inexact step).

It is more convenient to describe the rate of convergence of this scheme with respect to another sequence of parameters. Let us introduce an arbitrary sequence of positive

numbers $\{a_k\}_{k\geq 1}$ and denote $A_k \overset{\text{def}}{=} \sum_{i=1}^k a_i$. Then, we can define the contracting coefficients as follows

$$\gamma_k \overset{\text{def}}{=} \frac{a_{k+1}}{A_{k+1}}. \tag{7}$$

**Theorem 1** *For all points of the sequence $\{x_k\}_{k\geq 0}$, generated by process (6), we have the following relation:*

$$A_k F(x_k) \leq A_k F^* + B_k, \quad \textit{with} \quad B_k \overset{\text{def}}{=} \sum_{i=1}^k A_i \delta_i. \tag{8}$$

***Proof*** Indeed, for $k = 0$, we have $A_k = 0$, $B_k = 0$. Hence, (8) is valid. Assume it is valid for some $k \geq 0$. Then

$$\begin{aligned} A_{k+1} F(x_{k+1}) &\overset{\text{Step 3}}{\leq} A_{k+1} F(\bar{x}_{k+1}) \\ &\leq A_{k+1}\Big((1-\gamma_k) F(x_k) + \gamma_k F^* + \delta_{k+1}\Big) \\ &\overset{(7)}{=} A_k F(x_k) + a_{k+1} F^* + A_{k+1}\delta_{k+1} \\ &\overset{(8)}{\leq} A_{k+1} F^* + B_{k+1}. \end{aligned}$$

□

From bound (8), we can see, that

$$F(x_k) - F^* \leq \frac{1}{A_k} \sum_{i=1}^k A_i \delta_i, \quad k \geq 1. \tag{9}$$

Hence, the actual rate of convergence of method (6) depends on the growth of coefficients $\{A_k\}_{k\geq 1}$ *relatively* to the level of inaccuracies $\{\delta_k\}_{k\geq 1}$. Potentially, this rate can be arbitrarily high. Since we did not assume anything yet about our objective function, this means that we just retransmitted the complexity of solving the problem (1) onto a lower level, the level of computing the point $\bar{x}_{k+1}$, satisfying the condition (4). We are going to discuss different possibilities for that in Sects. 4 and 5.

Now, let us endow the method (6) with a computable *accuracy certificate*. For this purpose, for a sequence of given test points $\{\bar{x}_k\}_{k\geq 1} \subset \operatorname{dom} \psi$, we introduce the following *Estimating Function* (see [19]):

$$\varphi_k(v) \overset{\text{def}}{=} \sum_{i=1}^k a_i \big[ f(\bar{x}_i) + \langle \nabla f(\bar{x}_i), v - \bar{x}_i \rangle + \psi(v) \big].$$

By convexity of $f(\cdot)$, we have $A_k F(v) \geq \varphi_k(v)$ for all $v \in \operatorname{dom}\psi$. Hence, for all $k \geq 1$, we can get the following bound for the functional residual:

$$F(x_k) - F^* \leq \ell_k \overset{\text{def}}{=} F(x_k) - \frac{1}{A_k}\varphi_k^*, \quad \varphi_k^* \overset{\text{def}}{=} \min_{v \in \operatorname{dom}\psi} \varphi_k(v). \tag{10}$$

The complexity of computing the value of $\ell_k$ usually does not exceed the complexity of computing the next iterate of our method since it requires just one call of the *linear minimization oracle*. Let us show that an appropriate rate of decrease of the estimates $\ell_k$ can be guaranteed by sufficiently accurate steps of the method (2). For that, we need a stronger condition on point $\bar{x}_{k+1}$, that is

$$\begin{aligned} &\langle \nabla f(\bar{x}_{k+1}), v - \bar{v}_{k+1}\rangle + \psi(v) \geq \psi(\bar{v}_{k+1}) - \frac{1}{\gamma_k}\delta_{k+1}, \quad v \in \operatorname{dom}\psi, \\ &\bar{x}_{k+1} = (1-\gamma_k)x_k + \gamma_k \bar{v}_{k+1}, \end{aligned} \tag{11}$$

with some $\delta_{k+1} \geq 0$. Note that, for $\delta_{k+1} = 0$, condition (11) ensures the exactness of the corresponding step of method (2).

Let us consider now the following algorithm.

**Conceptual Contracting – Point Method, II** (12)

**Initialization**. Choose $x_0 \in \operatorname{dom}\psi$.

**Iteration** $k \geq 0$.

1: Choose $\gamma_k \in (0, 1]$.

2: For some $\delta_{k+1} \geq 0$, find $\bar{x}_{k+1}$ satisfying (11).

3: If $F(\bar{x}_{k+1}) \leq F(x_k)$, then set $x_{k+1} = \bar{x}_{k+1}$. Else choose $x_{k+1} = x_k$.

This scheme differs from the previous method (6) only in the characteristic condition (11) for the next test point.

**Theorem 2** *For all points of the sequence $\{x_k\}_{k\geq 0}$, generated by the process (12), we have*

$$\varphi_k^* \geq A_k F(x_k) - B_k, \quad k \geq 0. \tag{13}$$

***Proof*** For $k = 0$, relation (13) is valid since both sides are zeros. Assume that (13) holds for some $k \geq 0$. Then, for any $v \in \operatorname{dom}\psi$, we have

$$\varphi_{k+1}(v) \equiv \varphi_k(v) + a_{k+1}\big[f(\bar{x}_{k+1}) + \langle \nabla f(\bar{x}_{k+1}), v - \bar{x}_{k+1}\rangle + \psi(v)\big]$$

$$
\begin{aligned}
&\overset{(13)}{\geq} A_k F(x_k) - B_k + a_{k+1}\big[f(\bar{x}_{k+1}) + \langle \nabla f(\bar{x}_{k+1}), v - \bar{x}_{k+1}\rangle + \psi(v)\big] \\
&\overset{(*)}{\geq} A_{k+1}\big[f(\bar{x}_{k+1}) + \langle \nabla f(\bar{x}_{k+1}), \frac{a_{k+1}v + A_k x_k}{A_{k+1}} - \bar{x}_{k+1}\rangle\big] \\
&\quad + A_k \psi(x_k) + a_{k+1}\psi(v) - B_k \\
&= A_{k+1} f(\bar{x}_{k+1}) + a_{k+1}\big[\langle \nabla f(\bar{x}_{k+1}), v - \bar{v}_{k+1}\rangle + \psi(v)\big] \\
&\quad + A_k \psi(x_k) - B_k \\
&\overset{(11)}{\geq} A_{k+1} f(\bar{x}_{k+1}) + a_{k+1}\psi(\bar{v}_{k+1}) + A_k \psi(x_k) - B_{k+1} \\
&\overset{(**)}{\geq} A_{k+1} F(\bar{x}_{k+1}) - B_{k+1} \overset{\text{Step 3}}{\geq} A_{k+1} F(x_{k+1}) - B_{k+1}.
\end{aligned}
$$

Here, the inequalities $(*)$ and $(**)$ are justified by convexity of $f(\cdot)$ and $\psi(\cdot)$, correspondingly. Thus, (13) is proved for all $k \geq 0$. □

Combining now (10) with (13), we obtain

$$
F(x_k) - F^* \leq \ell_k \quad \leq \quad \frac{1}{A_k} \sum_{i=1}^{k} A_i \delta_i, \quad k \geq 1. \tag{14}
$$

We see that the right hand side in (14) is the same, as that one in (9). However, this convergence is stronger, since it provides a bound for the accuracy certificate $\ell_k$.

## 3 Affine-invariant high-order smoothness conditions

We are going to study complexity of solving the auxiliary problem in (2), and how it depends on the contracting parameter. For that we use *affine-invariant* characteristics of variation of function $f(\cdot)$ over the compact convex sets. For a convex set $Q$, define

$$
\Delta_Q^p(f) \overset{\text{def}}{=} \sup_{\substack{x,v \in Q, \\ t \in (0,1]}} \frac{1}{t^{p+1}} \Big| f(x + t(v - x)) - f(x) - \sum_{i=1}^{p} \frac{t^i}{i!} D^i f(x)[v - x]^i \Big|. \tag{15}
$$

Note, that for $p = 1$ this characteristic was considered in [13] for the analysis of the classical Frank–Wolfe algorithm.

In many situations, it is more convenient to use an upper bound for $\Delta_Q^p(f)$, which is a full variation of its $(p+1)$th derivative over the set $Q$:

$$
\mathcal{V}_Q^{p+1}(f) \overset{\text{def}}{=} \sup_{x,y,v \in Q} \Big| D^{p+1} f(y)[v - x]^{p+1} \Big|. \tag{16}
$$

Indeed, by Taylor formula, we have

$$
\frac{1}{t^{p+1}} \Big[ f(x + t(v - x)) - f(x) - \sum_{i=1}^{p} \frac{t^i}{i!} D^i f(x)[v - x]^i \Big]
$$

$$= \frac{1}{p!} \int_0^1 (1-\tau)^p D^{p+1} f(x + \tau t(v-x))[v-x]^{p+1} d\tau.$$

Hence,

$$\Delta_Q^p(f) \le \frac{1}{(p+1)!} \mathcal{V}_Q^{p+1}(f). \tag{17}$$

Sometimes, in order to exploit a *primal-dual* structure of the problem, we need to work with the dual objects (gradients), as in method (12). In this case, we need a characteristic of variation of the gradient $\nabla f(\cdot)$ over the set $Q$:

$$\Gamma_Q^p(f) \stackrel{\text{def}}{=} \sup_{\substack{x,y,v \in Q, \\ t \in (0,1]}} \frac{1}{t^p} \Big| \langle \nabla f(x + t(v-x)) - \nabla f(x) - \sum_{i=2}^{p} \frac{t^{i-1}}{(i-1)!} D^i f(x)[v-x]^{i-1}, v - y \rangle \Big|. \tag{18}$$

Since

$$\begin{aligned} &\frac{1}{t} \Big[ f(x + t(v-x)) - f(x) - \sum_{i=1}^{p} \frac{t^i}{i!} D^i f(x)[v-x]^i \Big] \\ &= \frac{1}{t} \Big[ \int_0^1 \langle \nabla f(x + \tau t(v-x)), t(v-x) \rangle d\tau - \sum_{i=1}^{p} \frac{t^i}{i!} D^i f(x)[v-x]^i \Big] \\ &= \int_0^1 \langle \nabla f(x + \tau t(v-x)) - \sum_{i=1}^{p} \frac{(\tau t)^{i-1}}{(i-1)!} D^i f(x)[v-x]^{i-1}, v - x \rangle d\tau, \end{aligned}$$

we conclude that

$$\Delta_Q^p(f) \le \frac{1}{p+1} \Gamma_Q^p(f). \tag{19}$$

At the same time, by Taylor formula, we get

$$\begin{aligned} &\frac{1}{t^p} \Big[ \nabla f(x + t(v-x)) - \nabla f(x) - \sum_{i=2}^{p} \frac{t^{i-1}}{(i-1)!} D^i f(x)[v-x]^{i-1} \Big] \\ &\quad = \frac{1}{(p-1)!} \int_0^1 (1-\tau)^{p-1} D^{p+1} f(x + \tau t(v-x))[v-x]^p d\tau. \end{aligned} \tag{20}$$

Therefore, again we have an upper bound in terms of the variation of $(p+1)$th derivative, that is

$$\Gamma_Q^p(f) \overset{(20)}{\leq} \frac{1}{p!} \sup_{x,y,z,v \in Q} \left| \langle D^{p+1} f(z)[v-x]^p, v-y \rangle \right| \;\leq\; \frac{2(p+1)^p}{(p!)^2} \mathcal{V}_Q^{p+1}(f).$$

The last inequality can be justified by simple arguments from Linear Algebra (see also Section 2.3 in [16]). Hence, the value of $\mathcal{V}_Q^{p+1}(f)$ is the biggest one. However, in many cases it is more convenient.

***Example 1*** For a fixed self-adjoint positive-definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$, define the corresponding Euclidean norm as $\|x\| := \langle Bx, x \rangle^{1/2}$, $x \in \mathbb{E}$. Let $Q \subset \mathbb{E}$ be a compact set with diameter

$$\mathscr{D} = \mathscr{D}_{\|\cdot\|}(Q) \overset{\text{def}}{=} \max_{x,y \in Q} \|x - y\| \;<\; +\infty.$$

Let $W$ be an open set containing it: $Q \subset W \subseteq \mathbb{E}$. Assume that function $f$ is $(p+1)$-times continuously differentiable on $W$, and its $p$-th derivative is Lipschitz continuous on $W$ (w.r.t. $\|\cdot\|$):

$$\|D^p f(x) - D^p f(y)\| \overset{\text{def}}{=} \max_{h \in \mathbb{E}: \|h\| \leq 1} |(D^p f(x) - D^p f(y))[h]^p| \;\leq\; L_p \|y - x\|,$$

for all $x, y \in W$.

Then, we have

$$\mathcal{V}_Q^{p+1}(f) \leq L_p \mathscr{D}^{p+1}.$$

□

In some situations we can obtain much better estimates.

***Example 2*** Let $A \succeq 0$, and $f(x) = \frac{1}{2}\langle Ax, x \rangle$ with

$$x \in \mathbb{S}_n \overset{\text{def}}{=} \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^{(i)} = 1\}.$$

For measuring distances in the standard simplex, we choose $\ell_1$-norm:

$$\|h\| = \sum_{i=1}^n |h^{(i)}|, \quad h \in \mathbb{R}^n.$$

In this case, $\mathscr{D} = \mathscr{D}_{\|\cdot\|}(\mathbb{S}_n) = 2$, and $L_1 = \max\limits_{1 \leq i \leq n} A^{(i,i)}$. On the other hand,

$$\mathcal{V}^2_{\mathbb{S}_n}(f) = \max_{1 \leq i,j \leq n} \langle A(e_i - e_j), e_i - e_j \rangle$$

$$\leq \max_{1\leq i,j\leq n} [2\langle Ae_i, e_i\rangle + 2\langle Ae_j, e_j\rangle] \;=\; 4L_1,$$

where $e_k$ denotes the $k$th coordinate vector in $\mathbb{R}^n$. Thus, $\mathcal{V}^2_{\mathbb{S}_n} \leq L_1\mathscr{D}^2$.

However, for some matrices, the value $\mathcal{V}^2_{\mathbb{S}_n}(f)$ can be much smaller than $L_1\mathscr{D}^2$. Indeed, let $A = aa^T$ for some $a \in \mathbb{R}^n$. Then $L_1 = \max\limits_{1\leq i\leq n} (a^{(i)})^2$, and

$$\mathcal{V}^2_{\mathbb{S}_n}(f) = \left[\max_{1\leq i\leq n} a^{(i)} - \min_{1\leq i\leq n} a^{(i)}\right]^2,$$

which can be much smaller than $4L_1$. □

***Example 3*** For given vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, consider the objective

$$f(x) = \ln\left(\sum_{k=1}^m e^{\langle a_k, x\rangle}\right), \quad x \in \mathbb{S}_n.$$

Then, it holds

$$\begin{aligned}\langle \nabla^2 f(x)h, h\rangle &\leq \max_{1\leq k,l\leq m} \langle a_k - a_l, h\rangle^2 \\ &\leq \max_{1\leq k,l\leq m} \|a_k - a_l\|^2_\infty \|h\|^2_1, \quad h \in \mathbb{R}^n\end{aligned}$$

(see Example 1 in [7] for the first inequality). Therefore, in $\ell_1$-norm we have

$$L_1 = \max_{1\leq k,l\leq m} \max_{1\leq i\leq n} \left[a^{(i)}_k - a^{(i)}_l\right]^2.$$

At the same time,

$$\begin{aligned}\mathcal{V}^2_{\mathbb{S}_n}(f) &= \sup_{x\in\mathbb{S}_n} \max_{1\leq i,j\leq n} \langle \nabla^2 f(x)(e_i - e_j), e_i - e_j\rangle \\ &\leq \max_{1\leq k,l\leq m} \max_{1\leq i,j\leq n} \left[\left(a^{(i)}_k - a^{(j)}_k\right) - \left(a^{(i)}_l - a^{(j)}_l\right)\right]^2.\end{aligned}$$

The last expression is the maximal difference between variations of the coordinates. It can be much smaller than $L_1\mathscr{D}^2 = 4L_1$.

Moreover, we have (see Example 1 in [7]):

$$|D^3 f(x)[h]^3| \leq \max_{1\leq k,l\leq m} |\langle a_k - a_l, h\rangle|^3, \quad h \in \mathbb{R}^n.$$

Hence, we obtain

$$\mathcal{V}^3_{\mathbb{S}_n}(f) \leq \max_{1\leq k,l\leq m} \max_{1\leq i,j\leq n} \left|\left(a^{(i)}_k - a^{(j)}_k\right) - \left(a^{(i)}_l - a^{(j)}_l\right)\right|^3.$$

□

## 4 Contracting-point tensor methods

In this section, we show how to implement Contracting-point methods, by using affine-invariant tensor steps. At each iteration of (2), we approximate $f(\cdot)$ by Taylor's polynomial of degree $p \geq 1$ around the current point $x_k$:

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

Thus, we need to solve the following auxiliary problem:

$$\min_{v \in \text{dom}\, \psi} \left\{ M_k(y) \overset{\text{def}}{=} \Omega_p(f, x_k; y) + S_k(y) : \; y = (1 - \gamma_k)x_k + \gamma_k v \right\}. \tag{21}$$

Note that this global minimum $M_k^*$ is well defined since $\text{dom}\, \psi$ is bounded. Let us take

$$\bar{x}_{k+1} = (1 - \gamma_k)x_k + \gamma_k \bar{v}_{k+1},$$

where $\bar{v}_{k+1}$ is an inexact solution to (21) in the following sense:

$$M_k(\bar{x}_{k+1}) - M_k^* \leq \xi_{k+1}. \tag{22}$$

Then, this point serves as a good candidate for the inexact step of our method.

**Theorem 3** *Let $\xi_{k+1} \leq c\gamma_k^{p+1}$, for some constant $c \geq 0$. Then*

$$F_k(\bar{x}_{k+1}) - F_k^* \leq \delta_{k+1},$$

*for $\delta_{k+1} = (c + 2\Delta^p_{\text{dom}\, \psi}(f))\gamma_k^{p+1}$.*

***Proof*** Indeed, for $y = x_k + \gamma_k(v - x_k)$ with arbitrary $v \in \text{dom}\, \psi$, we have

$$\begin{aligned} F_k(y) &= f(y) + S_k(y) \\ &\overset{(15)}{\geq} \Omega_p(f, x_k; y) + S_k(y) - \Delta^p_{\text{dom}\, \psi}(f)\gamma_k^{p+1} \\ &\overset{(22)}{\geq} \Omega_p(f, x_k; \bar{x}_{k+1}) + S_k(\bar{x}_{k+1}) - (c + \Delta^p_{\text{dom}\, \psi}(f))\gamma_k^{p+1} \\ &\overset{(15)}{\geq} f(\bar{x}_{k+1}) + S_k(\bar{x}_{k+1}) - (c + 2\Delta^p_{\text{dom}\, \psi}(f))\gamma_k^{p+1} \\ &= F_k(\bar{x}_{k+1}) - \delta_{k+1}. \end{aligned}$$

□

Thus, we come to the following minimization scheme.

**Contracting-Point Tensor Method, I** (23)

**Initialization**. Choose $x_0 \in \operatorname{dom} \psi$, $c \geq 0$.

**Iteration** $k \geq 0$.

1: Choose $\gamma_k \in (0, 1]$.

2: For some $\xi_{k+1} \leq c\gamma_k^{p+1}$, find $\bar{x}_{k+1}$ satisfying (22).

3: If $F(\bar{x}_{k+1}) \leq F(x_k)$, then set $x_{k+1} = \bar{x}_{k+1}$. Else choose $x_{k+1} = x_k$.

For $p = 1$ and $\psi(\cdot)$ being an indicator function of a compact convex set, this is well-known Frank–Wolfe algorithm [10]. For $p = 2$, this is Contracting-Domain Newton Method from [8].

Straightforward consequence of our observations is the following

**Theorem 4** *Let* $A_k \stackrel{\text{def}}{=} k \cdot (k+1) \cdot \ldots \cdot (k+p)$, *and consequently* $\gamma_k = \frac{p+1}{k+p+1}$. *Then, for all iterations* $\{x_k\}_{k \geq 1}$ *generated by method* (23), *we have*

$$F(x_k) - F^* \leq (p+1)^{p+1} \cdot (c + 2\Delta^p_{\operatorname{dom}\psi}) \cdot k^{-p}.$$

***Proof*** Combining (9) with Theorem 3, we have

$$F(x_k) - F^* \leq \frac{(c + 2\Delta^p_{\operatorname{dom}\psi}(f))}{A_k} \sum_{i=1}^{k} \frac{a_i^{p+1}}{A_i^p}, \quad k \geq 1.$$

Since

$$\frac{1}{A_k} \sum_{i=1}^{k} \frac{a_i^{p+1}}{A_i^p} = \frac{1}{A_k} \sum_{i=1}^{k} \frac{(p+1)^{p+1} A_i}{(p+i)^{p+1}} \leq \frac{(p+1)^{p+1} k}{A_k} \leq \frac{(p+1)^{p+1}}{k^p},$$

we get the required inequality. □

It is important that the required level of accuracy $\xi_{k+1}$ for solving the subproblem is not static: it is changing with iterations. Indeed, from the practical perspective, there is no need to use high accuracy during the first iterations, but it is natural to improve our precision while approaching the optimum. Inexact proximal-type tensor methods with dynamic inner accuracies were studied in [9].

Let us note that the objective $M_k(y)$ from (21) is generally nonconvex for $p \geq 3$, and it may be nontrivial to look for its global minimum. Because of that, we propose an alternative condition for the next point. It requires just to find a point satisfying an (inexact) *first-order necessary condition for local optimality* of $\Omega_p(f, x_k; y)$. That is a point $\bar{x}_{k+1}$, satisfying for all $v \in \operatorname{dom} \psi$

$$\begin{aligned} &\langle \nabla \Omega_p(f, x_k; \bar{x}_{k+1}), v - \bar{v}_{k+1} \rangle + \psi(v) \geq \psi(\bar{v}_{k+1}) - \frac{1}{\gamma_k} \xi_{k+1}, \\ &\bar{x}_{k+1} = (1 - \gamma_k) x_k + \gamma_k \bar{v}_{k+1}, \end{aligned} \tag{24}$$

for some tolerance value $\xi_{k+1} \geq 0$.

**Theorem 5** *Let point $\bar{x}_{k+1}$ satisfy condition (24) with*

$$\xi_{k+1} \leq c \gamma_k^{p+1},$$

*for some constant $c \geq 0$. Then it satisfies inexact condition (11) of the Conceptual Contracting-Point Method with*

$$\delta_{k+1} = (c + \Gamma^p_{\operatorname{dom} \psi}(f)) \gamma_k^{p+1}.$$

***Proof*** Indeed, for any $v \in \operatorname{dom} \psi$, we have

$$\begin{aligned} &\langle \nabla f(\bar{x}_{k+1}), v - \bar{v}_{k+1} \rangle + \psi(v) \\ &\quad = \langle \nabla \Omega_p(f, x_k; \bar{x}_{k+1}), v - \bar{v}_{k+1} \rangle + \psi(v) \\ &\qquad + \langle \nabla f(\bar{x}_{k+1}) - \Omega_p(f, x_k; \bar{x}_{k+1}), v - \bar{v}_{k+1} \rangle \\ &\quad \overset{(24)}{\geq} \psi(\bar{v}_{k+1}) - c \gamma_k^p + \langle \nabla f(\bar{x}_{k+1}) - \Omega_p(f, x_k; \bar{x}_{k+1}), v - \bar{v}_{k+1} \rangle \\ &\quad \overset{(18)}{\geq} \psi(\bar{v}_{k+1}) - (c + \Gamma^p_{\operatorname{dom} \psi}(f)) \gamma_k^p \quad = \quad \psi(\bar{v}_{k+1}) - \frac{1}{\gamma_k} \delta_{k+1}. \end{aligned}$$

□

Note the appearance of $\gamma_k^{p+1}$ in both Theorems 3 and 5. It comes from the form of the derivatives for contracted objective (3), where we substitute $q = p + 1$ to bound the error of $p$-th order Taylor approximation.

Now, changing inexactness condition (22) in method (23) by condition (24), we come to the following algorithm.

**Contracting − Point Tensor Method, II**

**Initialization**. Choose $x_0 \in \operatorname{dom} \psi$, $c \geq 0$.

**Iteration** $k \geq 0$. (25)

1: Choose $\gamma_k \in (0, 1]$.

2: For some $\xi_{k+1} \leq c\gamma_k^{p+1}$, find $\bar{x}_{k+1}$ satisfying (24).

3: If $F(\bar{x}_{k+1}) \leq F(x_k)$, then set $x_{k+1} = \bar{x}_{k+1}$. Else choose $x_{k+1} = x_k$.

Its convergence analysis is straightforward.

**Theorem 6** *Let* $A_k \stackrel{\text{def}}{=} k \cdot (k+1) \cdot \ldots \cdot (k+p)$, *and consequently* $\gamma_k = \frac{p+1}{k+p+1}$. *Then, for all iterations* $\{x_k\}_{k \geq 1}$ *of method* (25), *we have*

$$F(x_k) - F^* \leq \ell_k \;\; \leq \;\; (p+1)^{p+1} \cdot (c + \Gamma^p_{\operatorname{dom}\psi}(f)) \cdot k^{-p}.$$

***Proof*** Combining inequality (14) with the statement of Theorem 5, we have

$$F(x_k) - F^* \leq \ell_k \;\; \leq \;\; \frac{c + \Gamma^p_{\operatorname{dom}\psi}(f)}{A_k} \sum_{i=1}^{k} \frac{a_i^{p+1}}{A_i^p}, \quad k \geq 1.$$

It remains to use the same reasoning, as in the proof of Theorem 4. □

Finally, let us discuss a trust-region interpretation of our methods. In the exact form ($\xi_k \equiv 0$), iterations of the Contracting-Point Tensor Methods can be rewritten as follows, for $k \geq 0$:

$$x_{k+1} \in \operatorname*{Argmin}_x \left\{ \Omega_p(f, x_k; x) + \gamma_k \psi\left(x_k + \frac{1}{\gamma_k}(x - x_k)\right) \right\}.$$

For $\psi(x) \equiv \operatorname{Ind}_Q(x)$, where $Q$ is a bounded convex set, this method can be seen as a *trust-region* scheme [5] with $p$-th order Taylor model of the objective function, regularized by the contraction of the feasible set $Q$.

## 5 Inexact contracting Newton method

In this section, let us present an implementation of our method (23) for $p = 2$, when at each step we solve the subproblem inexactly by a variant of first-order Conditional

Gradient Method. The entire algorithm looks as follows.

**Inexact Contracting Newton Method**

**Initialization**. Choose $x_0 \in \operatorname{dom} \psi$, $c > 0$.

**Iteration** $k \geq 0$.

1: Choose $\gamma_k \in (0, 1]$.

2: Denote $g_k(v) = \langle \nabla f(x_k), v - x_k \rangle + \frac{\gamma_k}{2} \langle \nabla^2 f(x_k)(v - x_k), v - x_k \rangle$.

3: Initialize inner method $t = 0$, $z_0 = x_k$, $\phi_0(w) \equiv 0$.

4-a: Set $\alpha_t = \frac{2}{t+2}$.

4-b: Set $\phi_{t+1}(w) = \alpha_t \big[ g_k(z_t) + \langle \nabla g_k(z_t), w - z_t \rangle + \psi(w) \big]$

$$+ (1 - \alpha_t)\phi_t(w).$$

4-c: Compute $w_{t+1} \in \operatorname*{Argmin}_w \phi_{t+1}(w)$.

4-d: Set $z_{t+1} = \alpha_t w_{t+1} + (1 - \alpha_t) z_t$.

4-e: If $g_k(z_{t+1}) + \psi(z_{t+1}) - \phi_{t+1}(w_{t+1}) > c\gamma_k^2$, then

Set $t = t + 1$ and go to 4-a, else go to 5.

5: Set $\bar{x}_{k+1} = \gamma_k z_{t+1} + (1 - \gamma_k) x_k$.

6: If $F(\bar{x}_{k+1}) \leq F(x_k)$, then set $x_{k+1} = \bar{x}_{k+1}$.Else choose $x_{k+1} = x_k$.

(26)

We provide an analysis of the total number of *oracle calls* for $f$ (step 2) and the total number of *linear minimization oracle calls* for the composite component $\psi$ (step 4-c), required to solve problem (1) up to the given accuracy level.

**Theorem 7** *Let $\gamma_k = \frac{3}{k+3}$. Then, for iterations $\{x_k\}_{k \geq 1}$ generated by method (26), we have*

$$F(x_k) - F^* \leq 27 \cdot (c + 2\Delta^{(2)}_{\operatorname{dom} \psi}) \cdot k^{-2}. \tag{27}$$

*Therefore, for any $\varepsilon > 0$, it is enough to perform*

$$K = \left\lceil \sqrt{\frac{27(c + 2\Delta^{(2)}_{\operatorname{dom} \psi}(f))}{\varepsilon}} \right\rceil \tag{28}$$

*iteration of the method, in order to get $F(x_K) - F^* \le \varepsilon$. And the total number $N_K$ of linear minimization oracle calls during these iterations is bounded as*

$$N_K \le 2 \cdot \Big(1 + \frac{2\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{c}\Big) \cdot \Big(1 + \frac{27(c + 2\Delta^{(2)}_{\mathrm{dom}\,\psi}(f))}{\varepsilon}\Big). \tag{29}$$

***Proof*** Let us fix arbitrary iteration $k \ge 0$ of our method and consider the following objective:

$$\begin{aligned} m_k(v) &= g_k(v) + \psi(v) \\ &= \langle \nabla f(x_k), v - x_k\rangle + \frac{\gamma_k}{2}\langle \nabla^2 f(x_k)(v - x_k), v - x_k\rangle + \psi(v). \end{aligned}$$

We need to find the point $\bar{v}_{k+1}$ such that

$$m_k(\bar{v}_{k+1}) - m_k^* \le c\gamma_k^2. \tag{30}$$

Note that if we set $\bar{x}_{k+1} := \gamma_k \bar{v}_{k+1} + (1-\gamma_k)x_k$, then from (30) we obtain bound (22) satisfied with $\xi_{k+1} = c\gamma_k^3$. Thus we would obtain iteration of Algorithm (23) for $p = 2$, and Theorem 4 gives the required rate of convergence (27). We are about to show that steps 4-a – 4-e of our algorithm are aiming to find such point $\bar{v}_{k+1}$.

Let us introduce auxiliary sequences $A_t \overset{\text{def}}{=} t \cdot (t+1)$ and $a_{t+1} \overset{\text{def}}{=} A_{t+1} - A_t$ for $t \ge 0$. We use these sequences for analysing the *inner* method.

Then, $\alpha_t \equiv \frac{a_{t+1}}{A_{t+1}}$, and we have the following representation of the *Estimating Functions*, for every $t \ge 0$

$$\phi_{t+1}(w) = \frac{1}{A_{t+1}} \sum_{i=0}^{t} a_{i+1}\Big[g_k(z_i) + \langle \nabla g_k(z_i), w - z_i\rangle + \psi(w)\Big].$$

By convexity of $g_k(\cdot)$, we have

$$m_k(w) \ge \phi_{t+1}(w), \quad w \in \mathrm{dom}\,\psi.$$

Therefore, we obtain the following upper bound for the residual (30), for any $v \in \mathrm{dom}\,\psi$

$$m_k(v) - m_k^* \le m_k(v) - \phi_{t+1}^*, \tag{31}$$

where $\phi_{t+1}^* = \min_w \phi_{t+1}(w) = \phi_{t+1}(w_{t+1})$.

Now, let us show by induction, that

$$A_t\phi_t^* \ge A_t m_k(z_t) \; - \; B_t, \quad t \ge 0, \tag{32}$$

for $B_t := \frac{\gamma_k \mathcal{V}^2_{\mathrm{dom}\,\psi}(f)}{2} \sum_{i=0}^{t} \frac{a_{i+1}^2}{A_{i+1}}$. It obviously holds for $t = 0$. Assume that it holds for some $t \ge 0$. Then,

$$A_{t+1}\phi_{t+1}^* = A_{t+1}\phi_{t+1}(w_{t+1})$$

$$
\begin{aligned}
&= A_t \phi_t(w_{t+1}) + a_{t+1}\big[g_k(z_t) + \langle \nabla g_k(z_t), w_{t+1} - z_t\rangle + \psi(w_{t+1})\big] \\
&\overset{(32)}{\geq} A_t m_k(z_t) + a_{t+1}\big[g_k(z_t) + \langle \nabla g_k(z_t), w_{t+1} - z_t\rangle + \psi(w_{t+1})\big] - B_t \\
&= A_{t+1}\Big[g_k(z_t) + \alpha_t \langle \nabla g_k(z_t), w_{t+1} - z_t\rangle + \alpha_t \psi(w_{t+1}) + (1-\alpha_t)\psi(z_t)\Big] - B_t \\
&\geq A_{t+1}\Big[g_k(z_t) + \alpha_t \langle \nabla g_k(z_t), w_{t+1} - z_t\rangle + \psi(z_{t+1})\Big] - B_t.
\end{aligned}
$$

Note that

$$
\begin{aligned}
g_k(z_{t+1}) &= g_k(z_t + \alpha_t(w_{t+1} - z_t)) \\
&= g_k(z_t) + \alpha_t \langle \nabla g_k(z_t), w_{t+1} - z_t\rangle \\
&\quad + \frac{\alpha_t^2 \gamma_k}{2} \langle \nabla^2 f(x_k)(w_{t+1} - z_t), w_{t+1} - z_t\rangle.
\end{aligned}
$$

Therefore, we obtain

$$
A_{t+1}\phi_{t+1}^* \geq A_{t+1} m_k(z_{t+1}) - B_t - \frac{a_{t+1}^2}{A_{t+1}} \cdot \frac{\gamma_k \mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{2},
$$

and this is (32) for the next step. Therefore, we have (32) established for all $t \geq 0$.

Combining (31) with (32), we get the following guarantee for the inner steps 4-a – 4-e:

$$
\begin{aligned}
m_k(z_{t+1}) - m_k^* \leq m_k(z_{t+1}) - \phi_{t+1}^* &\leq \frac{\gamma_k \mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{2A_{t+1}} \sum_{i=0}^{t} \frac{a_{i+1}^2}{A_{i+1}} \\
&\leq \frac{2\gamma_k \mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{t+1}.
\end{aligned}
$$

Therefore, all iterations of our method is well-defined. We exit from the inner loop on step 4-e after

$$
t \geq \frac{2\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{c\gamma_k} - 1 = \frac{2(k+3)\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{3c} - 1, \tag{33}
$$

and the point $\bar{v}_{k+1} \equiv z_{t+1}$ satisfies (30).

Hence, we obtain (27) and (28). The total number of linear minimization oracle calls can be estimated as follows

$$
\begin{aligned}
N_K &\overset{(33)}{\leq} \sum_{k=0}^{K-1} \Big(1 + \frac{2(k+3)\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{3c}\Big) = K\Big(1 + \frac{\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{3c}(K+5)\Big) \\
&\leq K^2\Big(1 + \frac{2\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{c}\Big) \leq 2 \cdot \Big(1 + \frac{2\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)}{c}\Big) \cdot \Big(1 + \frac{27(c + 2\Delta^{(2)}_{\mathrm{dom}\,\psi}(f))}{\varepsilon}\Big).
\end{aligned}
$$

□

According to the result of Theorem 7, in order to solve problem (1) up to $\varepsilon > 0$ accuracy, we need to perform $\mathcal{O}(\frac{1}{\varepsilon})$ total computations of step 4-c of the method (estimate (29)). This is the same amount of linear minimization oracle calls, as required in the classical Frank–Wolfe algorithm [18]. However, this estimate can be over-pessimistic for our method. Indeed, it comes as the product of the worst-case complexity bounds for the outer and the inner optimization schemes. It seems to be very rare to meet with the worst-case instance at the both levels simultaneously. Thus, the practical performance of our method can be much better.

At the same time, the total number of gradient and Hessian computations is only $\mathcal{O}(\frac{1}{\varepsilon^{1/2}})$ (estimate (28)). This can lead to a significant acceleration over first-order Frank–Wolfe algorithm, when the gradient computation is a bottleneck (see our experimental comparison in the next section).

The only parameter which remains to choose in method (26), is the tolerance constant $c > 0$. Note that the right hand side of (29) is convex in $c$. Hence, its approximate minimization provides us with the following choice

$$c = 2\sqrt{\mathcal{V}^{(2)}_{\mathrm{dom}\,\psi}(f)\,\Delta^{(2)}_{\mathrm{dom}\,\psi}(f)}.$$

In practical applications, we may not know some of these constants. However, in many cases they are small. Therefore, an appropriate choice of $c$ is a small constant.

Finally, let us discuss effective implementation of our method, when the composite part is $\{0, +\infty\}$-indicator of the standard simplex:

$$\mathrm{dom}\,\psi \;=\; \mathbb{S}_n \stackrel{\mathrm{def}}{=} \Big\{x \in \mathbb{R}^n_+ \;:\; \sum_{i=1}^n x^{(i)} = 1\Big\}. \tag{34}$$

This is an example of a set with a finite number of *atoms*, which are the standard coordinate vectors in this case:

$$\mathbb{S}_n = \mathrm{Conv}\,\{e_1, \ldots, e_n\}.$$

See [13] for more examples of atomic sets in the context of Frank–Wolfe algorithm. The maximization of a convex function over such sets can be implemented very efficiently, since the maximum is always at the corner (one of the atoms).

At iteration $k \geq 0$ of method (26), we need to minimize over $\mathbb{S}_n$ the quadratic function

$$g_k(v) = \langle \nabla f(x_k), v - x_k\rangle + \frac{\gamma_k}{2}\langle \nabla^2 f(x_k)(v - x_k), v - x_k\rangle,$$

whose gradient is

$$\nabla g_k(v) = \nabla f(x_k) + \gamma_k \nabla^2 f(x_k)(v - x_k).$$

Assume that we keep the vector $\nabla g_k(z_t) \in \mathbb{R}^n$ for the current point $z_t$, $t \geq 0$ of the inner process, as well as its aggregation

$$h_t \overset{\text{def}}{=} \alpha_t \nabla g_k(z_t) + (1 - \alpha_t) h_{t-1}, \quad h_{-1} \overset{\text{def}}{=} 0 \in \mathbb{R}^n.$$

Then, at step 4-c we need to compute a vector

$$w_{t+1} \in \operatorname*{Argmin}_{w \in \mathbb{S}_n} \langle h_t, w \rangle \;=\; \operatorname{Conv}\left\{ e_j \;:\; j \in \operatorname*{Argmin}_{1 \leq j \leq n} h_t^{(j)} \right\}.$$

It is enough to find an index $j$ of a minimal element of $h_t$ and to set $w_{t+1} := e_j$. The new gradient is equal to

$$\begin{aligned} \nabla g_k(z_{t+1}) &\overset{\text{Step 4-d}}{=} \nabla g_k(\alpha_t w_{t+1} + (1 - \alpha_t) z_t) \\ &= \alpha_t \Big( \nabla f(x_k) + \gamma_k \nabla^2 f(x_k)(e_j - x_k) \Big) + (1 - \alpha_t) \nabla g_k(z_t), \end{aligned}$$

and the function value can be expressed using the gradient as follows

$$g_k(z_{t+1}) = \frac{1}{2} \langle \nabla f(x_k) + \nabla g_k(z_{t+1}), z_{t+1} - x_k \rangle.$$

The product $\nabla^2 f(x_k) e_j$ is just $j$-th column of the matrix. Hence, preparing in advance the following objects: $\nabla f(x_k) \in \mathbb{R}^n$, $\nabla^2 f(x_k) \in \mathbb{R}^{n \times n}$ and the Hessian-vector product $\nabla^2 f(x_k) x_k \in \mathbb{R}^n$, we are able to perform iteration of the inner loop (steps 4-a – 4-e) very efficiently in $\mathcal{O}(n)$ arithmetical operations.

## 6 Numerical experiments

Let us consider the problem of minimizing the log-sum-exp function (SoftMax)

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

over the standard simplex $\mathbb{S}_n$ (34). Coefficients $\{a_i\}_{i=1}^m$ and $b$ are generated randomly from the uniform distribution on $[-1, 1]$. We compare the performance of Inexact Contracting Newton Method (26) with that one of the classical Frank–Wolfe algorithm, for different values of the parameters. The results are shown on Figs. 1, 2 and 3.

We see, that the new method works significantly better in terms of the outer iterations (oracle calls). This confirms our theory. At the same time, for many values of the parameters, it shows better performance in terms of total computational time as well[1].

[1] CPU time was evaluated on a machine with Intel Core i5 CPU, 1.6GHz; 8 GB RAM. The methods were implemented in Python.

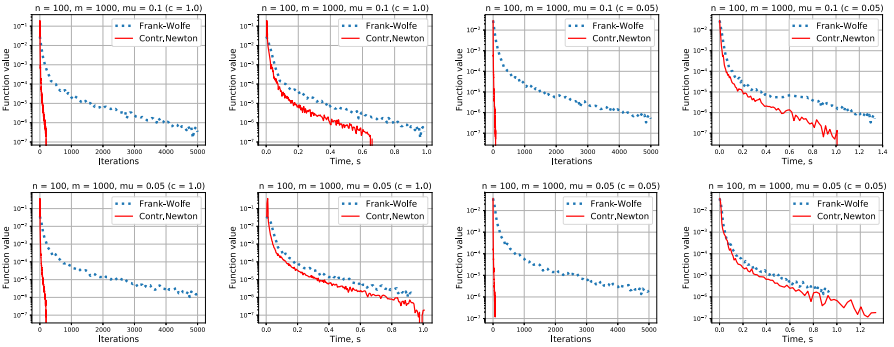


**Fig. 1** $n = 100,\ m = 1000$

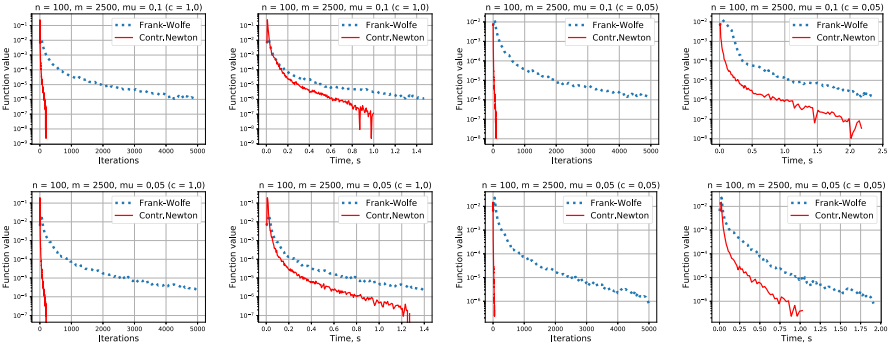


**Fig. 2** $n = 100,\ m = 2500$

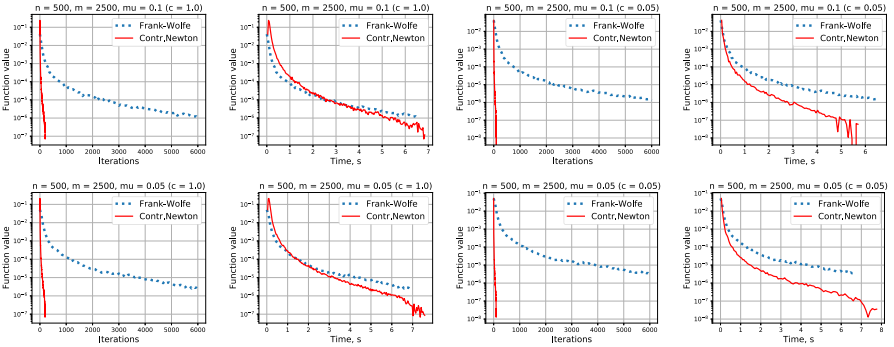


**Fig. 3** $n = 500,\ m = 2500$

## 7 Discussion

In this paper, we present a new general framework of Contracting-Point methods, which can be used for developing affine-invariant optimization algorithms of different order. For the methods of order $p \geq 1$, we prove the following global convergence rate:

$$F(x_k) - F^* \leq \mathcal{O}\big(1/k^p\big), \quad k \geq 1.$$

This is the same rate, as that of the basic high-order Proximal-Point scheme [21]. However, the methods from our paper are free from using the norms or any other characteristic parameters of the problem. This nice property makes Contracting-Point methods favourable for solving optimization problems over the sets with a non-Euclidean geometry (e.g. over the simplex or over a general convex polytope).

At the same time, it is known that in Euclidean case, the prox-type methods can be accelerated, achieving $\mathcal{O}(1/k^{p+1})$ global rate of convergence [2,6,20,21]. Using additional one-dimensional search at each iteration, this rate can be improved up to $\mathcal{O}(1/k^{\frac{3p+1}{2}})$ (see [11,21]). The latter rate is shown to be optimal [1,19]. To the best of our knowledge, the lower bounds for high-order methods in general non-Euclidean case remain unknown. However, the worst-case oracle complexity of the classical Frank–Wolfe algorithm (the case $p = 1$ in our framework) is proven to be near-optimal for smooth minimization over $\|\cdot\|_\infty$-balls [12].

Another open question is a possibility of efficient implementation of our methods for the case $p \geq 3$. In view of absence of explicit regularizer (contrary to the prox-type methods), the subproblem in (21) can be nonconvex. Hence, it seems hard to find its global minimizer. We hope that for some problem classes, it is still feasible to satisfy the inexact necessary condition for local optimality (24) by reasonable amount of computations. We keep this question for further investigation.

**Acknowledgements** We are very thankful to anonymous referees for valuable comments that improved the initial version of this paper.



## References

1. Arjevani, Y., Shamir, O., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization. Math. Program. **178**(1–2), 327–360 (2019)
2. Baes, M.: Estimate Sequence Methods: Extensions and Approximations. Institute for Operations Research, ETH, Zürich (2009)
3. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2016)
4. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Math. Program. **169**(2), 337–375 (2018)
5. Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, Philadelphia (2000)
6. Doikov, N., Nesterov, Y.: Contracting proximal methods for smooth convex optimization. CORE discussion papers 2019/27 (2019)
7. Doikov, N., Nesterov, Y.: Minimizing uniformly convex functions by cubic regularization of Newton method. arXiv preprint arXiv:1905.02671 (2019)

8. Doikov, N., Nesterov, Y.: Convex optimization based on global lower second-order models. arXiv preprint arXiv:2006.08518 (2020)
9. Doikov, N., Nesterov, Y.: Inexact tensor methods with dynamic accuracies. arXiv preprint arXiv:2002.09403 (2020)
10. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Nav. Res. Logist. Q. **3**(1–2), 95–110 (1956)
11. Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., Uribe, C.A., Jiang, B., Wang, H., Zhang, S., Bubeck, S., Qijia, J., Lee, Y.T., Yuanzhi, L., Aaron, S.: Near optimal methods for minimizing convex functions with Lipschitz $p$-th derivatives. In: Conference on Learning Theory, pp. 1392–1393 (2019)
12. Guzmán, C., Nemirovski, A.: On lower complexity bounds for large-scale smooth convex optimization. J. Complex. **31**(1), 1–14 (2015)
13. Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: International Conference on Machine Learning, pp. 427–435 (2013)
14. Kamzolov, D., Gasnikov, A., Dvurechensky, P.: On the optimal combination of tensor optimization methods. arXiv preprint arXiv:2002.01004 (2020)
15. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. **28**(1), 333–354 (2018)
16. Nemirovski, A.: Interior point polynomial time methods in convex programming. Lecture notes (2004)
17. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
18. Nesterov, Y.: Complexity bounds for primal-dual methods minimizing the model of objective function. Math. Program. **171**(1–2), 311–330 (2018)
19. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer, New York (2018)
20. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. Math. Program. **186**, 1–27 (2019)
21. Nesterov, Y.: Inexact accelerated high-order proximal-point methods optimization. CORE discussion papers 2020/8 (2020)
22. Nesterov, Y.: Superfast second-order methods for unconstrained convex optimization. CORE discussion papers 2020/7 (2020)