Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Near-optimal tensor methods for minimizing gradient norm

Pavel Dvurechensky[1], Alexander Gasnikov[2,3], Petr Ostroukhov[2], Cesar A. Uribe[4], Anastasiya Ivanova[2,5]

submitted: February 19, 2020

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: pavel.dvurechensky@wias-berlin.de

[2] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny, Moscow Region
141701 Russian Federation
E-Mail: gasnikov@yandex.ru
ostroukhov@phystech.edu
anastasiya.s.ivanova@phystech.edu

[3] Institute for Information Transmission Problems of RAS
Bolshoy Karetny per. 19, build.1
127051 Moscow
Russian Federation
E-Mail: gasnikov@yandex.ru

[4] Laboratory for Information and Decision Systems (LIDS)
Institute for Data, Systems, and Society (IDSS)
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA
E-Mail: cauribe@mit.edu

[5] Higher School of Economics
11, Pokrovsky boulevard
109028 Moscow
Russian Federation
E-Mail: anastasiya.s.ivanova@phystech.edu

No. 2694
Berlin 2020

2010 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25.

Key words and phrases. Convex optimization, tensor methods, gradient norm, nearly optimal methods.

The work of A. Gasnikov and C.A.Uribe was partially supported by the Yahoo! Research Faculty Engagement Program. The work by P. Dvurechensky was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Near-optimal tensor methods for minimizing gradient norm

Pavel Dvurechensky, Alexander Gasnikov, Petr Ostroukhov, Cesar A. Uribe, Anastasiya Ivanova

Abstract

Motivated by convex problems with linear constraints and, in particular, by entropy-regularized optimal transport, we consider the problem of finding approximate stationary points, i.e. points with the norm of the objective gradient less than small error, of convex functions with Lipschitz p-th order derivatives. Lower complexity bounds for this problem were recently proposed in [Grapiglia and Nesterov, arXiv:1907.07053]. However, the methods presented in the same paper do not have optimal complexity bounds. We propose two optimal up to logarithmic factors methods with complexity bounds with respect to the initial objective residual and the distance between the starting point and solution respectively.

1 Introduction

Although, the idea of using higher order derivatives in optimization methods is known at least since 1970's, see [16], recently these methods started to gain an increased research interest [3, 19, 4, 7, 1, 2] in optimization. Before [17] the main bottleneck was the auxiliary problem of minimizing the regularized Taylor expansion of the objective, which potentially can be a non-convex problem. Nesterov showed that an appropriate regularization makes this a convex problem and proposes an efficient method for solving this subproblem for the third-order method.

This motivated recent research in order to propose optimal high-order methods for convex optimization [12, 13, 20, 15, 6, 5].

In this paper, we consider the following unconstrained convex optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where f has p-th Lipschitz-continuous derivative with constant M_p. Contrary to existing approaches, where the objective is to find an ε-approximate solution $\bar{x}$ such that $f(\bar{x}) - f^* \leq \varepsilon$, we will focus on the problem of finding approximate stationary point $\|\nabla f(\bar{x})\|_* \leq \varepsilon$.

In [12, 13], the authors proposed a class of near-optimal methods up to a logarithmic factor for the solution of problems of the class (1) in the general convex setting and under additional assumption of uniform convexity. In the latter case, however, computational complexity was expressed in terms of the initial objective residual or optimality gap $f(x_0) - f^*$, where x_0 is the starting point. At the same time it is interesting to understand, how the complexity depends on the initial distance to the solution $\|x_0 - x^*\|$. Recently in [14], the authors proposed a set of methods for problems of the form (1) to find approximate stationary points. In [14, Theorem 4.2], the authors showed that in order to find a point $\bar{x}$ such that $\|\nabla f(\bar{x})\|_* \leq \varepsilon$, their proposed method require $O\big(\varepsilon^{-(p+1)/(p(p+2))}\big)$ iterations. Such complexity bound does not match the corresponding lower bounds proposed in [14, Theorem 6.6] and [14, Theorem 6.8], where the number of iterations required to find an ε-approximation is of the order $\Omega\big(\varepsilon^{-2(p+1)/(3p+1)}\big)$ with respect to the initial functional residual and $\Omega\big(\varepsilon^{-2/(3p+1)}\big)$ with respect to the

DOI 10.20347/WIAS.PREPRINT.2694 Berlin 2020

Property	Lower Bound	Upper Bound
Initial function residual	$\Omega\left(\varepsilon^{-2(p+1)/(3p+1)}\right)$	$O\left(\varepsilon^{-1}\right)$
	$\Omega\left(\varepsilon^{-2(p+1)/(3p+1)}\right)$	$\tilde{O}\left(\varepsilon^{-2(p+1)/(3p+1)}\right)$
Initial argument residual	$\Omega\left(\varepsilon^{-2/(3p+1)}\right)$	$O\left(\varepsilon^{-(p+1)/(p(p+2))}\right)$
	$\Omega\left(\varepsilon^{-2/(3p+1)}\right)$	$\tilde{O}\left(\varepsilon^{-2/(3p+1)}\right)$

Table 1: Complexity of minimizing the gradient norm from [14] and ours.

initial argument residual. As a related work, we also mention [4, 7], who study complexity bounds for tensor methods for finding approximate stationary points in the non-convex setting, thus being not directly related to our convex setting.

In this paper, we use the framework developed in [12, 13] to propose a near-optimal method to find an approximate stationary point of a convex function with high-order smoothness. The bound for our method matches up to a logarithmic multiplier the lower bound from [14]. Our contributions in terms of the complexity can be summarized in the Table 1. Besides that we present a variant of near optimal tensor method for minimization of uniformly high-order smooth functions with complexity bound depending on the initial distance to the solution $\|x_0 - x^*\|$ as opposed to objective residual in [12]. We also explain, how our methods can be extended to obtain near-optimal methods for functions with Hölder-continuous high-order derivatives.

This paper is organized as follows. We first start in Section 2 with a motivating example for the problem of finding approximate stationary points of convex functions. We describe the entropy regularized optimal transport problem and show that its structure provides a natural justification of tensor methods that exploit the smoothness properties of the corresponding dual problems. Section3 presents some results from other works, which we use in our paper. Section 4.1 presents the near-optimal algorithm for finding approximate stationary points, with respect to the initial objective residual; near-optimal complexity bounds are shown explicitly. Section 4.2 shows the corresponding near-optimal algorithm with respect to the initial argument residual; near-optimal complexity bounds are shown as well. In Section 5 we discuss possible extensions of the proposed methods, in particular for problems with Hölder-continuous higher-order derivatives. Section 6 shows some numerical results on the proposed algorithms for the logistic regression problem and minimization of "bad"functions which give the lower bounds for the considered problem class. Finally, conclusions and future work is presented in Section 7.

1.1 Notation

For $p \geq 1$, we denote by $\nabla^p f(x)[h_1, ..., h_p]$ the directional derivative of function f at x along directions $h_i \in \mathbb{R}^n$, $i = 1, ..., p$. $\nabla^p f(x)[h_1, ..., h_p]$ is symmetric p-linear form and its norm is defined as

$$\|\nabla^p f(x)\|_2 = \max_{h_1,...,h_p \in \mathbb{R}^n} \{\nabla^p f(x)[h_1, ..., h_p] : \|h_i\|_2 \leq 1, i = 1, ..., p\}$$

or equivalently

$$\|\nabla^p f(x)\|_2 = \max_{h \in \mathbb{R}^n} \{|\nabla^p f(x)[h, ..., h]| : \|h\|_2 \leq 1, i = 1, ..., p\}.$$

Here, for simplicity, $\|\cdot\|_2$ is standard Euclidean norm, but our algorithm and derivations can be generalized for the Euclidean norm given by general a positive semi-definite matrix B. We consider

convex, p times differentiable on $\mathbb{R}$ functions satisfying Lipschitz condition for p-th derivative

$$\|\nabla^p f(x) - \nabla^p f(y)\|_2 \leq M_p \|x - y\|_2, x, y \in \mathbb{R}^n. \tag{2}$$

Given a function f, numbers $p \geq 1$ and $M \geq 0$, define

$$T^f_{p,M}(x) \in \text{Arg}\min_{y \in \mathrm{R}^n} \left\{ \sum_{r=0}^{p} \frac{1}{r!} \nabla^r f(x) \underbrace{[y-x, ..., y-x]}_{r} + \frac{M}{(p+1)!} \|y - x\|_2^{p+1} \right\}, \tag{3}$$

and given a number $L \geq 0$ and point $z \in \mathbb{R}^n$, we define

$$F_{L,z}(x) \triangleq f(x) + \frac{L}{2} \|x - z\|_2^2. \tag{4}$$

2 A motivating example: problems with linear constraints

Let us consider a convex optimization problem with linear constraints

$$\min_{x \in Q \subseteq E} \{f(x) : Ax = b\}, \tag{5}$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, A is a given linear operator from E to some finite-dimensional real vector space H, $b \in H$ is given, $f(x)$ is a convex function on Q with respect to some chosen norm $\|\cdot\|_E$ on E.

The Lagrange dual problem for (5), written as a minimization problem, is

$$\min_{\lambda \in H^*} \left\{ \varphi(\lambda) := \langle \lambda, b \rangle + \max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right) \right\}. \tag{6}$$

We assume that the dual objective is smooth. In this case, by the Demyanov-Danskin theorem, $\nabla \varphi(\lambda) = b - Ax(\lambda)$, where

$$x(\lambda) := \arg\min_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right).$$

Proposition 2.1 (Lemma 1 in [11]). *Assume that for some λ*

$$-\langle \lambda, \nabla \varphi(\lambda) \rangle \leq \varepsilon_f, \quad \|\varphi(\lambda)\|_H \leq \varepsilon_{eq}.$$

Then

$$f(x(\lambda)) - f^* \leq \varepsilon_f, \quad \|Ax(\lambda) - b\|_H \leq \varepsilon_{eq}.$$

This means that if there is a method for the dual problem, which generates a bounded sequence of iterates λ_k and a point λ_k s.t. the gradient of the dual objective is small, then, using the relation $x(\lambda_k)$ we can reconstruct a nearly feasible, nearly optimal solution to the primal problem. This is a general motivation for convex optimization methods for minimizing the objective gradient norm. Moreover, the complexity bound for the dual method directly translates to the complexity for solving the primal problem without any overhead.

To further motivate the high-order methods for minimization of the objective gradient norm, we present a particular example of smooth dual objective with with high-order Lipschitz derivatives. This example

is the Entropy-regularized optimal transport problem [8, 9]. Next we provide a brief description of the problem and the properties of the dual objective.

Consider two histograms $p, q \in \Sigma_n$ on a support of size n, where Σ_n is the standard simplex. Also, consider a matrix $M \in \mathbb{R}_+^{n\times n}$ which is symmetric and accounts to the "cost" of transportation such that M_{ij} is the cost of moving a unit of mass from bin i to bin j. For example, given support points $(x_i)_{1\le i\le n}$ on the Euclidean space, one can consider $M_{ij} = \|x_i - x_j\|_2^2$, which corresponds to 2-Wasserstein distance. The entropy-regularized optimal transport problem is defined as:

$$W_\gamma(p,q) \triangleq \min_{X\in U(p,q)} \langle M, X\rangle - \gamma E(X), \tag{7}$$

where $\gamma \ge 0$ is a regularization parameter, $E(X) \triangleq -\sum_{i,j} X_{ij}\ln(X_{ij})$, and U is the transport polytope such that,

$$U(p,q) \triangleq \{X \in \mathbb{R}_+^{n\times n} \mid X\mathbf{1}_n = p, X^T\mathbf{1}_n = q\}.$$

It is known that the problem (7) is strongly convex and admits a unique optimal solution X^*[9]. If $\gamma = 0$ and $M_{ij} = \|x_i - x_j\|^r$, (7) is know as the r-th power of r-Wasserstein distance between p and q.

A standard way to deal with the optimization problem (7) is to write its dual.

$$\begin{aligned}
&\min_{X\in U(p,q)} \langle M, X\rangle + \gamma\langle X, \ln X\rangle \\
&\quad= \min_{X\in\Sigma_{n^2}} \langle M, X\rangle + \max_{\xi,\eta} \left\{\langle \xi, p - X\mathbf{1}_n\rangle + \langle \eta, q - X^T\mathbf{1}_n\rangle\right\} \\
&\quad= \max_{\xi,\eta} \left\{\langle \xi, p\rangle + \langle \eta, q\rangle + \min_{X\in\Sigma_{n^2}} \left\{\langle M + \xi{\mathbf{1}_n}^T + \mathbf{1}_n\mu^T + \gamma\ln X, X\rangle\right\}\right\} \\
&\quad= \max_{\xi,\eta} -\gamma\ln \sum_{i,j=1}^n \exp\left(-\frac{1}{\gamma}(M_{ij} - \xi_i - \eta_j)\right) + \langle \xi, p\rangle + \langle \eta, q\rangle
\end{aligned} \tag{8}$$

In this case the explicit dependence of the primal solution from the dual variables is given by

$$X(\xi,\eta) = \frac{\operatorname{diag}(\mathrm{e}^{\frac{\xi}{\gamma}})\mathrm{e}^{-\frac{M}{\gamma}}\operatorname{diag}(\mathrm{e}^{\frac{\eta}{\gamma}})}{\mathrm{e}^{\frac{\xi}{\gamma}}\mathrm{e}^{-\frac{M}{\gamma}}\mathrm{e}^{\frac{\eta}{\gamma}}}, \tag{9}$$

where the exponent is applied componentwise to vectors and matrices. We underline that as opposed to the standard dual problem [8], we consider X to lie not in $\mathbb{R}_+^{n\times n}$, but rather in the standard simplex of the size n^2, the latter bein the corollary of the marginal constraints $X\mathbf{1}_n = p$, $X^T\mathbf{1}_n = q$ since $p, q \in \Sigma_n$. This allows us to obtain a high-order smooth dual objective which has a softmax form. On the contrary, the dual problem in [8] has sum of exponents in the dual objective, meaning that the derivatives are not Lipschitz-continuous.

To show the correspondence to a general primal dual pair of problems (5)–(6), let us define $E = \mathbb{R}^{n^2}$, $\|\cdot\|_E = \|\cdot\|_1$, and variable $x = \operatorname{vec}(X) \in \mathbb{R}^{n^2}$ to be the vector obtained from a matrix X by writing each column of X below the previous column. Also we set $f(x) = \langle M, X\rangle + \gamma\langle X, \ln X\rangle$, $Q = \Sigma_{n^2}$, $b^T = (p^T, q^T)$, $A : \mathbb{R}^{n^2} \to \mathbb{R}^{2n}$ defined by the identity $(A\operatorname{vec}(X))^T = ((X\mathbf{1}_n)^T, (X^T\mathbf{1}_n)^T)$, and $\lambda^T = (\xi^T, \eta^T)$.
matrix A has the form

$$A = \begin{pmatrix} I_nI_n & I_n\dots \\ \mathbf{1}_n^T\mathbf{0}_n^T & \mathbf{0}_n^T\dots \\ \mathbf{0}_n^T\mathbf{1}_n^T & \mathbf{0}_n^T\dots \\ \dots\dots & \dots\dots \end{pmatrix},$$

where I_n is the identity matrix, $\mathbf{0}_n^T$ is the vector of all zeros. Using these notations, we can write the dual problem (8) as

$$\begin{aligned}&\max_{\lambda} -\gamma \ln \sum_{i,j=1}^{n} \exp\left(-\frac{[M - A^T\lambda]_{ij}}{\gamma}\right) + \langle \lambda, b\rangle \\ &= \max_{\lambda} -\mathbf{smax}_\gamma(A^T\lambda - M) + \langle \lambda, b\rangle, \end{aligned} \tag{10}$$

where

$$\mathbf{smax}_\gamma(y) \triangleq \gamma \log\left(\sum_{i=1}^{m} \exp(y_i/\gamma)\right).$$

More importantly, the following property holds.

Proposition 2.2 ([6, Theorem 3.4])**.** *Let $z \in \mathbb{R}^n$, $c \in \mathbb{R}^m$ and $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$. Then the function $\boldsymbol{smax}_\gamma(\mathcal{A}z - c)$ is (order 3) $\frac{15}{\gamma^3}$-smooth with respect to $\|\cdot\|_{\mathcal{A}^T\mathcal{A}}$.*

As a corollary, the dual objective in (10) is order 3 $\frac{15}{\gamma^3}$-Lipschitz-continuous w.r.t. $\|\cdot\|_{AA^T}$.

We can conclude that minimizing the norm of the gradient of the dual function provides an estimate for the optimality gap of the corresponding primal problem, the estimate of the optimal transport cost in this case. Thus, having a fast method that exploits the high-order smoothness of the dual problem can provide efficient algorithms for the computation of entropy regularized optimal transport plans.

3 Preliminaries

To make the paper more self-contained, in this section we recall the near-optimal tensor methods for minimization of convex objective functions with Lipschitz-continupus p-th derivative [12].

Algorithm 1 Near-Optimal Tensor Method [12, Algorithm 1]

Input: u_0, y_0 — starting points; N — iteration number; $A_0 = 0$

Output: y^N

1: **for** $k = 0, 1, 2, \ldots, N-1$ **do**

2: Choose L_k such that

$$\frac{1}{2} \le \frac{2(p+1)M_p}{p!L_k}\|y^{k+1} - x^k\|_2^{p-1} \le 1, \tag{11}$$

where

$$a_{k+1} = \frac{1/L_k + \sqrt{1/L_k^2 + 4A_k/L_k}}{2}, \quad A_{k+1} = A_k + a_{k+1}, \quad \{\text{note that } L_k a_k^2 = A_{k+1}\}$$

$$x^k = \frac{A_k}{A_{k+1}} y^k + \frac{a_{k+1}}{A_{k+1}} u^k, \quad y^{k+1} = T_{p,pM_p}^{F_{L_k,x^k}}(x^k).$$

3: $u^{k+1} = u^k - a_{k+1}\nabla f(y^{k+1})$

4: **end for**

5: **return** y^N

Theorem 3.1 (Theorem 1 in [12]). *Let sequence* (x^k, y^k, u^k), $k \geq 0$ *be generated by Algorithm 1. Then*

$$f(y^k) - f^* \leq \frac{cM_p\|y^0 - x_*\|_2^{p+1}}{k^{\frac{3p+1}{2}}}, \quad \text{where} \quad c = \frac{2^{\frac{3(p+1)^2+4}{4}}(p+1)}{p!}.$$

Moreover, each iteration k *requires* $O\left(\ln \frac{1}{\varepsilon}\right)$ *oracle calls.*

The following lemma is a particular case of Lemma 5.2 in [14] with $\nu = 1$, $\theta = 0$, and $\varphi = 0$.

Lemma 3.2 (Lemma 5.2 in [14]). *Let* $M_p < \infty$, $M \geq pM_p$ *and let for some* $x \in \mathbb{R}^n$

$$\tilde{z} = T^f_{p,M}(x).$$

Then

$$f(x) - f(\tilde{z}) \geq \frac{1}{8(p+1)!M^{\frac{1}{p}}}\|\nabla f(\tilde{z})\|_*^{\frac{p+1}{p}}.$$

4 Near-optimal tensor methods for gradient norm minimization

4.1 Near-optimal tensor methods with respect to the initial objective residual

In this section we build up from Algorithm 1 to develop a near optimal algorithm for which we can provide explicit complexity bounds for the approximation of a stationary point. The obtained oracle complexity bound matches that of the lower bound presented in [14] up to a logarithmic factor. The basic assumption is that the starting point x_0 satisfies $f(x_0) - f^* \leq \Delta_0$.

Algorithm 2 Near-optimal algorithm with respect to initial objective residual

1: **Input** $p, M_p, \Delta_0 : f(x_0) - f^* \leq \Delta_0, \varepsilon$.
2: **Define**:

$$M_\mu = pM_p, \quad \mu = \frac{\varepsilon^2}{32\Delta_0}, \quad \tilde{\varepsilon} = \frac{(\varepsilon/2)^{\frac{p}{p+1}}}{8M_\mu^{\frac{1}{p}}(p+1)!}, \quad f_\mu(x) = f(x) + \frac{\mu}{2}\|x - x_0\|_2^2.$$

3: **while** $\Delta_k \geq \tilde{\varepsilon}$, where $\Delta_k = \Delta_0 \cdot 2^{-k}$ **do**
4:

$$\text{Set} \quad \Delta_k = \Delta_0 \cdot 2^{-k} \quad \text{and} \quad N_k = \max\left\{\left\lceil\left(\frac{2cM_p 2^{\frac{p+1}{2}}}{\mu^{\frac{p+1}{2}}}\Delta_k^{\frac{p-1}{2}}\right)^{\frac{2}{3p+1}}\right\rceil, 1\right\}. \tag{12}$$

where $c = 2^{(3(p+1)^2+4)/4}(p+1)/p!$.
5: Set $z_{k+1} = y^{N_k}$ as the output of Algorithm 1 applied to $f_\mu(x)$ starting from z_k and run for N_k steps.
6: $k = k + 1$.
7: **end while**
8: **Find** $\tilde{z} = T^{f_\mu}_{p,\, M_\mu}(z_k)$
9: **Output** $\tilde{z}$.

Theorem 4.1. *Assume the function f is convex, p times differentiable on $\mathbb{R}$ with M_p-Lipschitz p-th derivative. Let $\tilde{z}$ be generated by Algorithm 2. Then*

$$\|\nabla f(\tilde{z})\|_2 \leq \varepsilon,$$

and the total number of iterations of Algorithm 1 required by Algorithm 2 is

$$O\left(\frac{M_p^{\frac{2}{3p+1}}}{\varepsilon^{\frac{2(p+1)}{3p+1}}}\Delta_0^{\frac{2p}{3p+1}} + \log_2 \frac{2^{\frac{4p-3}{p+1}}\Delta_0 (pM_p)^{\frac{1}{p}}(p+1)!}{\varepsilon^{\frac{p}{p+1}}}\right).$$

Moreover, the total oracle complexity is within a $O\left(\ln \frac{1}{\varepsilon}\right)$ factor of the above iteration complexity.

Proof. By definition of $f_\mu(x)$:

$$f_\mu(x_0) - f_\mu(x_\mu^*) = f(x_0) - f(x_\mu^*) - \frac{\mu}{2}\|x_\mu^* - x_0\|_2^2 \leq f(x_0) - f(x_*) \leq \Delta_0,$$

Where x_μ^* is the minimum of $f_\mu(x)$. So, for $k = 0$ we have $f_\mu(z_k) - f_\mu(x_\mu^*) \leq \Delta_k$.

Let us assume that $f_\mu(z_k) - f_\mu(x_\mu^*) \leq \Delta_k$ and show that $f_\mu(z_{k+1}) - f_\mu(x_\mu^*) \leq \Delta_{k+1}$. From Theorem 3.1 applied to $f_\mu(x)$, since it is μ-strongly convex and has M_p-Lipschitz p-th derivative, it holds that

$$\begin{aligned} f_\mu(z_{k+1}) - f_\mu(x_\mu^*) &\leq \frac{cM_p\|z_k - x_\mu^*\|_2^{p+1}}{N_k^{\frac{3p+1}{2}}} \leq \frac{cM_p}{N_k^{\frac{3p+1}{2}}}\left(\frac{2(f_\mu(z_k) - f_\mu(x_\mu^*))}{\mu}\right)^{\frac{p+1}{2}} \\ &\leq \frac{cM_p}{N_k^{\frac{3p+1}{2}}}\left(\frac{2\Delta_k}{\mu}\right)^{\frac{p+1}{2}} \leq \frac{\Delta_k}{2} = \Delta_{k+1}. \end{aligned} \tag{13}$$

Thus, $f_\mu(z_k) - f_\mu(x_\mu^*) \leq \Delta_k$ for all $k \geq 0$.

According to Lemma 3.2, we have

$$f_\mu(z_k) - f_\mu(\tilde{z}) \geq \frac{1}{8(p+1)!M_\mu^{\frac{1}{p}}}\|\nabla f_\mu(\tilde{z})\|_2^{\frac{p+1}{p}}. \tag{14}$$

At the same time, by the stopping criterion in Algorithm 2,

$$f_\mu(z_k) - f_\mu(\tilde{z}) \leq f_\mu(z_k) - f_\mu(x_\mu^*) \leq \Delta_k \leq \tilde{\varepsilon}. \tag{15}$$

By the definition of $\tilde{\varepsilon}$ and (14), (15), we have that

$$\|\nabla f_\mu(\tilde{z})\|_2 \leq \frac{\varepsilon}{2}. \tag{16}$$

By definition, f_μ is μ-strongly convex and, using (14), we have

$$\frac{\mu}{2}\|x_\mu^* - x_0\|_2^2 \leq f_\mu(x_0) - f_\mu(x_\mu^*) \leq \Delta_0, \tag{17}$$

$$\frac{\mu}{2}\|\tilde{z} - x_\mu^*\|_2^2 \leq f_\mu(\tilde{z}) - f_\mu(x_\mu^*) \leq f_\mu(z_k) - f_\mu(x_\mu^*) \leq \Delta_0. \tag{18}$$

Applying triangle inequality to the sum of (17) and (18), we get

$$\frac{\mu}{2}\|\tilde{z} - x_0\|_2^2 \leq \mu\big(\|x_\mu^* - x_0\|_2^2 + \|\tilde{z} - x_\mu^*\|_2^2\big) \leq 4\Delta_0,$$

and

$$\|\tilde{z} - x_0\|_2 \leq 2\sqrt{\frac{2\Delta_0}{\mu}}.$$

By definition of μ in Algorithm 2, we have

$$\mu\|\tilde{z} - x_0\|_2 \leq \mu \cdot 2\sqrt{\frac{2\Delta_0}{\mu}} = 2\sqrt{2\mu\Delta_0} = \frac{\varepsilon}{2}. \tag{19}$$

Finally, according to the definition of f_μ, (16), (19) and triangle inequality, we get

$$\|\nabla f(\tilde{z})\|_2 \leq \|\nabla f_\mu(\tilde{z})\|_2 + \mu\|\tilde{z} - x_0\|_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

It remains to bound the total number of steps of Algorithm 1. Denote $\tilde{c} = \left(2c2^{\frac{p+1}{2}}\right)^{\frac{2}{3p+1}}$.

$$\begin{aligned}
\sum_{i=0}^{k} N_i &\leq \tilde{c}\frac{M_p^{\frac{2}{3p+1}}}{\mu^{\frac{p+1}{3p+1}}}\sum_{i=0}^{k}(\Delta_0 \cdot 2^{-i})^{\frac{p-1}{3p+1}} + k \leq \tilde{c}\frac{M_p^{\frac{2}{3p+1}}}{\mu^{\frac{p+1}{3p+1}}}\Delta_0^{\frac{p-1}{3p+1}} \cdot \sum_{i=0}^{k} 2^{-i\frac{p-1}{3p+1}} + k \\
&\leq 2\tilde{c}\frac{M_p^{\frac{2}{3p+1}}}{\mu^{\frac{p+1}{3p+1}}}\Delta_0^{\frac{p-1}{3p+1}} + \log_2\frac{\Delta_0}{\tilde{\varepsilon}} \\
&= O\left(\frac{M_p^{\frac{2}{3p+1}}}{\varepsilon^{\frac{2(p+1)}{3p+1}}}\Delta_0^{\frac{2p}{3p+1}} + \log_2\frac{2^{\frac{4p-3}{p+1}}\Delta_0(pM_p)^{\frac{1}{p}}(p+1)!}{\varepsilon^{\frac{p}{p+1}}}\right)
\end{aligned} \tag{20}$$

According to Theorem 3.1, the total number of oracle calls is within the $O\left(\ln\frac{1}{\varepsilon}\right)$ factor from the number of iterations of Algorithm 1. This completes the proof. □

□

4.2 Near-optimal tensor methods with respect to the initial variable residual

In this section we build up from Algorithm 1 to develop a near optimal algorithm for which we provide explicit complexity bounds for the approximation of a stationary point. The obtained oracle complexity bound matches that of the lower bound presented in [14] up to a logarithmic factor. The basic assumption is that the starting point x_0 satisfies $\|x_0 - x_*\|_2 \leq R$.

Theorem 4.2. *Assume the function f is convex, p times differentiable on $\mathbb{R}^n$ with M_p-Lipschitz p-th derivative. Let $\tilde{z}$ be generated by Algorithm 3. Then*

$$\|\nabla f(\tilde{z})\|_2 \leq \varepsilon \tag{22}$$

and the total number of iterations of Algorithm 1 required by Algorithm 3 is

$$O\left(\frac{M_p^{\frac{2}{3p+1}}R^{\frac{2(p-1)}{3p+1}}}{\varepsilon^{\frac{2}{3p+1}}} + \log\frac{2^{\frac{p}{p+1}}(p+1)!(pM_p)^{\frac{1}{p}}}{\varepsilon^{\frac{1}{p+1}}}\right).$$

Moreover, the total oracle complexity is within a $O\left(\ln\frac{1}{\varepsilon}\right)$ factor of the above iteration complexity.

Algorithm 3 Near-optimal algorithm with respect to initial argument residual

1: **Input** $M_p, x_0,\ R : \|x^* - x_0\|_2^2 \leq R,\ \varepsilon.$

2: **Define**:

$$M_\mu = pM_p,\ \mu = \frac{\varepsilon}{4R},\ \tilde{\varepsilon} = \frac{(\varepsilon/2)^{\frac{p}{p+1}}}{8(p+1)!M_\mu^{\frac{1}{p}}},\ f_\mu(x) = f(x) + \frac{\mu}{2}\|x - x_0\|_2^2,\ z_0 = x_0,\ k = 0.$$

3: **while** $\mu R_k^2/2 \geq \tilde{\varepsilon}$ where $R_k = R \cdot 2^{-k}$ **do**

4:

$$\text{Set } R_k = R \cdot 2^{-k} \text{ and } N_k = \max\left\{\left\lceil\left(\frac{8cM_pR_k^{p-1}}{\mu}\right)^{\frac{2}{3p+1}}\right\rceil, 1\right\}, \tag{21}$$

where $c = 2^{(3(p+1)^2+4)/4}(p+1)/p!$.

5: Set $z_{k+1} = y^{N_k}$ as the output of Algorithm 1 applied to $f_\mu(x)$ starting from z_k and run for N_k steps.

6: $k = k + 1$.

7: **end while**

8: **Find** $\tilde{z} = T^{f_\mu}_{p,\, M_\mu}(z_k)$

9: **Output** $\tilde{z}$.

Proof. By definition of $f_\mu(x)$, we have

$$f(x_\mu^*) + \frac{\mu}{2}\|x_\mu^* - x_0\|_2^2 = f_\mu(x_\mu^*) \leq f_\mu(x^*) = f(x^*) + \frac{\mu}{2}\|x^* - x_0\|_2^2 \leq f(x_\mu^*) + \frac{\mu}{2}\|x^* - x_0\|_2^2.$$

Hence, $\|x_\mu^* - x_0\|_2^2 \leq \|x^* - x_0\|_2^2 \leq R^2$. So, for $k = 0$ we have $\|x_\mu^* - z_k\|_2 \leq R_k$.

Let us assume that $\|x_\mu^* - z_k\|_2 \leq R_k$ and show that $\|x_\mu^* - z_{k+1}\|_2 \leq R_{k+1}$. From Theorem 3.1 applied to $f_\mu(x)$, since it is μ-strongly convex and has M_p-Lipschitz p-th derivative, it holds that

$$\frac{\mu}{2}\|z_{k+1} - x_\mu^*\|_2^2 \leq f_\mu(z_{k+1}) - f_\mu(x_\mu^*) \leq \frac{cM_p\|z_k - x_\mu^*\|_2^{p+1}}{N_k^{\frac{3p+1}{2}}} \leq \frac{\mu(R_k/2)^2}{2} = \frac{\mu R_{k+1}^2}{2}.$$

Thus, $\|z_{k+1} - x_\mu^*\|_2 \leq R_{k+1},\ f_\mu(z_k) - f_\mu(x_\mu^*) \leq \frac{\mu R_k^2}{2}$ for all $k \geq 0$.

From Lemma 3.2, we have

$$f_\mu(z_k) - f_\mu(\tilde{z}) \geq \frac{1}{8(p+1)!M_\mu^{\frac{1}{p}}}\|\nabla f_\mu(\tilde{z})\|_2^{\frac{p}{p+1}}. \tag{23}$$

At the same time,

$$f_\mu(z_k) - f_\mu(\tilde{z}) \leq f_\mu(z_k) - f_\mu(x_\mu^*) \leq \frac{\mu R_k^2}{2} \leq \tilde{\varepsilon}$$

by the stopping criterion of the algorithm. Combining these two inequalities and from the choice of $\tilde{\varepsilon}$ we get that

$$\|\nabla f_\mu(\tilde{z})\|_2 \leq \frac{\varepsilon}{2}.$$

From (23) we also have that

$$\frac{\mu}{2}\|\tilde{z} - x_\mu^*\|_2^2 \leq f_\mu(\tilde{z}) - f_\mu(x_\mu^*) \leq f_\mu(z_k) - f_\mu(x_\mu^*) \leq \frac{\mu R_k^2}{2} = \frac{\mu}{2}(R \cdot 2^{-k})^2 \leq \frac{\mu R^2}{2}.$$

Thus, $\|\tilde{z} - x_\mu^*\|_2 \leq R$. Hence, $\|\tilde{z} - x_0\|_2 \leq \|\tilde{z} - x_\mu^*\|_2 + \|x_\mu^* - x_0\|_2 \leq 2R$.

Finally, from our choice of μ

$$\|\nabla f(\tilde{z})\|_2 \leq \|\nabla f_\mu(\tilde{z})\|_2 + \mu\|\tilde{z} - x_0\|_2 \leq \frac{\varepsilon}{2} + \mu \cdot 2R \leq \varepsilon. \tag{24}$$

It remains to estimate the number of iterations of the Algorithm 1. Summing up the number of operations $N_i,\ i = 0, ..., k$, we obtain

$$\begin{aligned}\sum_{i=0}^{k} N_i &\leq \sum_{i=0}^{k}\left[\left(\frac{8cM_pR_i^{p-1}}{\mu}\right)^{\frac{2}{3p+1}} + 1\right] = \left(\frac{8cM_pR^{p-1}}{\mu}\right)^{\frac{2}{3p+1}}\sum_{i=0}^{k} 2^{\frac{-2i(p-1)}{3p+1}} + k \\ &\leq 2\left(\frac{8cM_pR^{p-1}}{\mu}\right)^{\frac{2}{3p+1}}\sum_{i=0}^{k} 2^{\frac{-2i(p-1)}{3p+1}} + \frac{1}{2}\log_2\frac{\mu R^2}{2\tilde{\varepsilon}} \\ &= O\left(\frac{M_p^{\frac{2}{3p+1}}R^{\frac{2(p-1)}{3p+1}}}{\varepsilon^{\frac{2}{3p+1}}} + \frac{1}{2}\log\frac{2^{\frac{p}{p-1}}(p+1)!M_p^{\frac{1}{p}}}{\varepsilon^{\frac{1}{p+1}}}\right).\end{aligned}$$

According to Theorem 3.1, the total number of oracle calls is within the $O\left(\ln\frac{1}{\varepsilon}\right)$ factor from the number of iterations of Algorithm 1. This completes the proof □ □

5 Extensions

Let us discuss possible extension of the proposed methods. One straightforward generalization is a near-optimal method for minimizing the norm of objective with Hölder-continuous gradient, i.e., for some $\nu \in [0, 1]$ satisfying

$$\|\nabla^p f(x) - \nabla^p f(y)\|_2 \leq M_{p,\nu}\|x - y\|_2^\nu, x, y \in \mathbb{R}^n.$$

The idea is to combine near-optimal tensor method for minimization of functions with Hölder-continuous p-th derivatives [18] with Lemma 5.2 in [14] for general ν. This approach allows to obtain complexity bounds which, up to a logarithmic and constant factors coincide with the lower bounds in [14]. We defer the exact derivations to the next version of the paper.

Another possible extension is inexact solution of the auxiliary subproblems and implementing adaptation to the constant $M_{p,\nu}$ [14]. Importantly, the basic Algorithm 1 is adaptive to M_p. Nevertheless, to apply the regularization technique with parameter μ we need to know the parameter M_p. Thus, it is desirable to overcome this drawback.

6 Numerical analysis

In this section, we present a number of simulations for the proposed near-optimal tensor method. Particularly, we implement Algorithm 2 for the logistic regression problem on both synthetic and real data sets. Also, we show the performance of the Algorithm 2 on a family of functions recently described as are difficult for all tensor methods [17]. We focus on the case where $p = 3$ for which we have efficient methods for the solution of the auxiliary subproblem [17, Section 5]. Finally, we present the performance results for the entropy regularized optimal transport problem.

6.1 Logistic Regression

For the logistic regression problem, we are given a set of d data pairs $\{y_i, w_i\}$ for $1 \leq i \leq d$, where $y_i \in \{1, -1\}$ is the class label of object i, and $w_i \in \mathbb{R}^n$ is the set of features of object i. We are interested in finding a vector x that solves the following optimization problem

$$\frac{1}{d}\sum_{i=1}^{d} \ln\Big(1 + \exp\big(-y_i\langle w_i, x\rangle\big)\Big) \to \min_{x\in\mathbb{R}^n}. \tag{25}$$

Figure 1 shows the gradient norm of the logistic regression function at the points generated by Algorithm 2. Initially, we show the results for synthetic data where $d = 100$ and $n = 10$. We focus on showing the results for different values of ε. We count as iterations each of the iterations of Algorithm 1 [12, Algorithm 1] in Line 3. For implementation simplicity in addition to the N_k upper bound of each of the iteration sin Line 3, if the gradient is not longer decreasing we apply the restarting after 500 iterations.

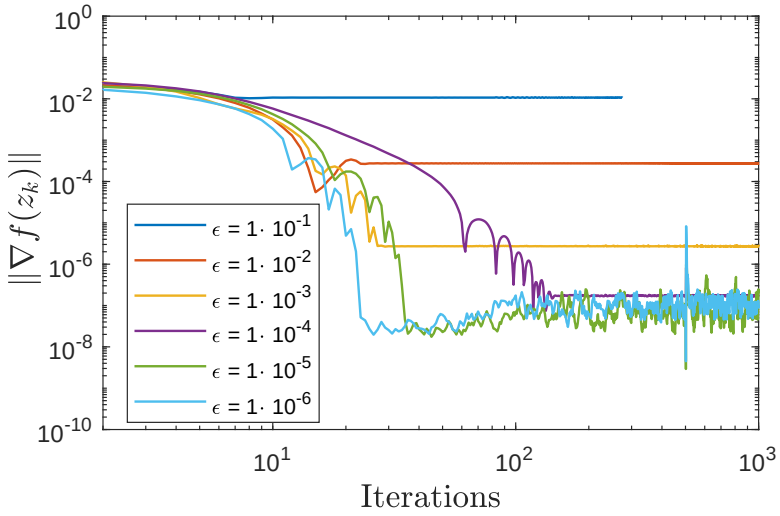

Figure 1: Gradient norm at the interations generated by Algorithm 2 on synthetic data for various values of ε .

Figure 2 shows the gradient norm of the logistic regression function at the points generated by Algorithm 2. In this case, we use the `Mushroom`, `A9A`, `Covertype` and `IJCNN1` datasets from [10] with a fixed value of $\varepsilon = 1 \cdot 10^{-5}$.

6.2 A family of difficult functions

Next, we analyse the performance of the proposed algorithm on an universal parametric family of objective functions, which are difficult for all tensor methods [17, 14] defined as

$$f_m(x) = \eta_{p+1}\left(A_m x\right) - x_1, \tag{26}$$

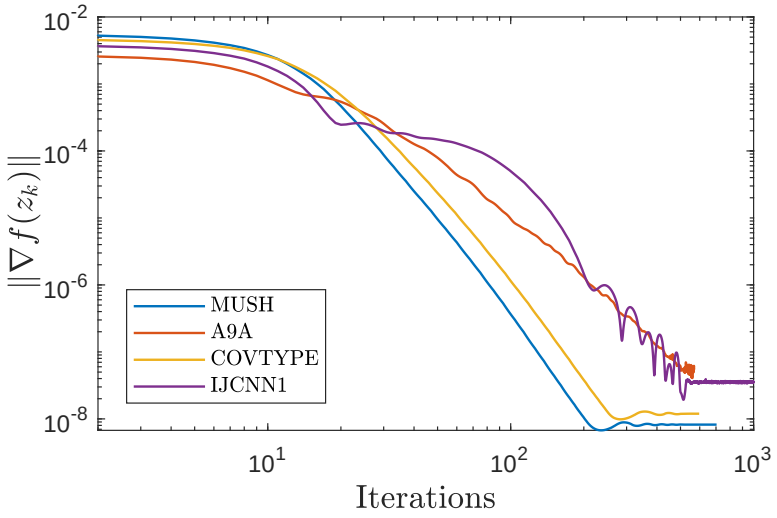

Figure 2: Gradient norm at the interations generated by Algorithm 2 on real data sets from [10] with $\varepsilon = 1 \cdot 10^{-5}$.

where, for integer parameter $p \geq 1$, $\eta_{p+1}(x) = \frac{1}{p+1} \sum\limits_{i=1}^{n} |x_i|^{p+1}$, $2 \leq m \leq n$, $x \in \mathbb{R}^n$, A_m is the $n \times n$ block diagonal matrix:

$$A_m = \begin{pmatrix} U_m & 0 \\ 0 & I_{n-m} \end{pmatrix}, \quad \text{with} \quad U_m = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{27}$$

and I_n is the identity $n \times n$-matrix. For a detailed description of the high-order derivatives of this class of functions, and its optimality properties see [17].

Finally, Figure 3 shows the performance results of Algorithm 2 on the family of functions in (26) with $p = 3$ and various values of parameters $m = n$ with $\varepsilon = 1 \cdot 10^{-5}$.

7 Conclusions

In this paper we consider the problem of minimization of the gradient norm of a convex objective with Lipschitz-continuous p-th derivative. We motivate this problem by minimization problems with linear constraints and, in particular, by Entropy-regularized optimal transport. We propose two algorithms together with their complexity bounds which up to a logarithmic factor coincide with existing lower bounds. Finally, we present preliminary numerical experiments to illustrate the practical performance of the algorithms.

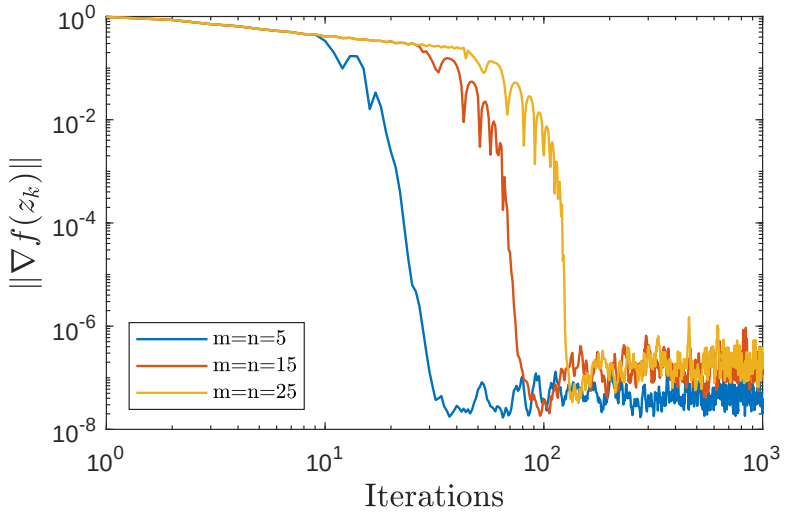

Figure 3: Gradient norm at the interations generated by Algorithm 1 on the family of functions in (26) with $p = 3$ and various values of parameters $m = n$ with $\varepsilon = 1 \cdot 10^{-5}$.

References

[1] Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 774–792. PMLR, 06–09 Jul 2018.

[2] Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, May 2018.

[3] Michel Baes. Estimate sequence methods:extensions and approximations. Technical report, 2009.

[4] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1):359–368, May 2017.

[5] Brian Bullins. Fast minimization of structured convex quartics. *arXiv preprint arXiv:1812.10349*, 2018.

[6] Brian Bullins and Richard Peng. Higher-order accelerated methods for faster non-smooth optimization. *arXiv preprint arXiv:1906.01621*, 2019.

[7] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *arXiv:1708.04044*, 2018.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[9] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[10] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[11] A. V. Gasnikov, E. V. Gasnikova, Yu. E. Nesterov, and A. V. Chernov. Efficient numerical methods for entropy-linear programming problems. *Computational Mathematics and Mathematical Physics*, 56(4):514–524, 2016.

[12] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César A. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1374–1391, Phoenix, USA, 25–28 Jun 2019. PMLR. arXiv:1809.00382.

[13] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A. Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1392–1393, Phoenix, USA, 25–28 Jun 2019. PMLR.

[14] Geovani Nunes Grapiglia and Yurii Nesterov. Tensor methods for finding approximate stationary points of convex functions. *arXiv preprint arXiv:1907.07053*, 2019.

[15] Oliver Hinder, Aaron Sidford, and Nimit Sharad Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *arXiv preprint arXiv:1906.11985*, 2019.

[16] K. H. Hoffmann and H. J. Kornstaedt. Higher-order necessary conditions in abstract mathematical programming. *Journal of Optimization Theory and Applications*, 26(4):533–568, Dec 1978.

[17] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. Technical report, CORE UCL, 2018. CORE Discussion Paper 2018/05.

[18] C. Song and Y. Ma. Towards unified acceleration of high-order algorithms under hölder continuity and uniform convexity. *arXiv:1906.00582*, 2019.

[19] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[20] Ashia Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent. *arXiv preprint arXiv:1902.08825*, 2019.