An Optimal High-Order Tensor Method for Convex Optimization

Bo JIANG ISYEBOJIANG@GMAIL.COM
Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China

Haoyue WANG HAOYUEWANG14@FUDAN.EDU.CN
Department of Mathematics, Fudan University, Shanghai, China

Shuzhong ZHANG ZHANGS@UMN.EDU
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA; joint appointment with Institute of Data and Decision Analytics, The Chinese University of Hong Kong, Shenzhen, China

Editors: Alina Beygelzimer and Daniel Hsu

[1] **Abstract**

This paper is concerned with finding an *optimal* algorithm for minimizing a composite convex objective function. The basic setting is that the objective is the sum of two convex functions: the first function is smooth with up to the d-th order derivative information available, and the second function is possibly non-smooth, but its proximal tensor mappings can be computed approximately in an efficient manner. The problem is to find – in that setting – the best possible (optimal) iteration complexity for convex optimization. Along that line, for the smooth case (without the second non-smooth part in the objective), Nesterov (1983) proposed an optimal algorithm for the first-order methods ($d = 1$) with iteration complexity $O\left(1/k^2\right)$. A high-order tensor algorithm with iteration complexity of $O\left(1/k^{d+1}\right)$ was proposed by Baes (2009) and Nesterov (2018). In this paper, we propose a new high-order tensor algorithm for the general composite case, with the iteration complexity of $O\left(1/k^{(3d+1)/2}\right)$, which matches the lower bound for the d-th order methods as established in Nesterov (2018); Arjevani et al. (2018), and hence is optimal. Our approach is based on the *Accelerated Hybrid Proximal Extragradient* (A-HPE) framework proposed in Monteiro and Svaiter (2013), where a bisection procedure is installed for each A-HPE iteration. At each bisection step a proximal tensor subproblem is approximately solved, and the total number of bisection steps per A-HPE iteration is bounded by a logarithmic factor in the precision required.

Keywords: convex optimization; tensor method; acceleration; iteration complexity.

1. Introduction

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

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + h(x), \tag{1}$$

where f is differentiable and convex, and h is convex but possibly non-smooth. In this context, we assume that convex tensor (polynomial) proximal mappings regarding h can be approximately computed efficiently. Given that structure, a fundamental question is to find an *optimal* algorithm that solves the above problem, using the available derivative information of the smooth part f.

1. Extended abstract. Full version appears as arXiv reference, arXiv:1812.06557v2.

In case $F(x) = f(x)$, and only the gradient information of f is available, Nesterov (1983) proposed a gradient-type algorithm, which achieves the overall iteration complexity of $O(1/k^2)$, matching the lower bound on the iteration complexity of this class of solution methods, hence is known to be an *optimal* algorithm among all the first-order methods. When the Hessian information is available, Nesterov (2008) proposed an acceleration scheme for cubic regularized Newton's method, and he showed that the iteration complexity bound improves from $O\left(1/k^2\right)$ to $O\left(1/k^3\right)$. A few years later, Monteiro and Svaiter (2013) proposed a different acceleration scheme, which they termed as *Accelerated Hybrid Proximal Extragradient Method* (A-HPE) framework, and they proved that if the second-order information is incorporated into the A-HPE framework then the corresponding accelerated Newton proximal extragradient method has a superior iteration complexity bound of $O\left(1/k^{7/2}\right)$ over $O\left(1/k^3\right)$. In 2018, Arjevani et al. (2018) showed that $O\left(1/k^{7/2}\right)$ is actually a lower bound for the oracle complexity of the second-order methods for convex smooth optimization. This shows that the accelerated Newton proximal extragradient method is an optimal second-order method.

As evidenced by the special cases $d = 1$ and $d = 2$, there is a clear tradeoff between the level of derivative information required and the overall iteration complexity improved. Therefore, a natural and important question arises:

What is the *exact* tradeoff relationship between d and the worst-case iteration complexity?

To answer such question, Baes (2009) extended the accelerated cubic regularized Newton method to the general high-order case with the iteration complexity of $O\left(1/k^{d+1}\right)$, where d is the order of derivative information used in the algorithm. Similar extension was recently revisited by Nesterov (2018) with a discussion on the efficient implementation of the method when $d = 3$. Jiang et al. (2018) extended Nesterov's approach to accommodate the composite optimization (1) and relaxed the requirement on the knowledge of problem parameters such as the Lipschitz constants and the requirement on the exact solutions of the subproblems while maintaining the same iteration bound as in Nesterov (2018). Along the line of bounding the worst case iteration complexity using up to the d-th order derivative information, there have also been significant progresses as well. Arjevani et al. (2018) showed that the worst case iteration complexity of any algorithm in that setting cannot be better than $O\left(1/k^{(3d+1)/2}\right)$. A simplified analysis of the bound can be found in Nesterov (2018). So, there was a gap between the achieved iteration bound $O\left(1/k^{d+1}\right)$ and the best possible bound of $O\left(1/k^{(3d+1)/2}\right)$. Clearly at least one of the two bounds is improvable. In this paper, we aim to settle the above theoretical quest by providing a new implementable algorithm whose iteration complexity is precisely $O\left(1/k^{(3d+1)/2}\right)$. As a result, the tradeoff relationship discussed above is pinned down to be exactly $O\left(1/k^{(3d+1)/2}\right)$. We note the independent work Gasnikov et al. (2018) first in Russian and then tranlated to English in late December 2018, and two other independent works Bubeck et al. (2018) and Bullins (2018), which were posted on arxiv in December 2018 as well. They derive similar results to ours with a limitation on the smooth objective functions (or more specifically– quartics in Bullins (2018)), while we allow a composite objective.

Our algorithm is based on the A-HPE framework of Monteiro and Svaiter (2013), and it specifies a way to generate an approximate solution through the use of high order derivative information by Taylor expansion. In each iteration, such approximate solution is computed by means of a bisection process. At each bisection step, a regulated convex tensor (polynomial) optimization subproblem is approximately solved. Moreover, we show that, to implement one A-HPE iteration, the number

of bisection steps – each calling to solve a convex tensor subproblem – is upper bounded by a logarithmic factor in the inverse of the required precision. Our bisection procedure is similar to the one proposed in Monteiro and Svaiter (2013) for the case $d = 2$; however, a key modification is applied which enables the removal of the so-called "bracketing stage" used in Monteiro and Svaiter (2013).

References

Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming, published online*, 2018.

M. Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zrich, Switzerland*, 2009.

S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford. Near-optimal method for highly smooth convex optimization. *ArXiv Preprint: 1812.08026*, 2018.

B. Bullins. Fast minimization of structured convex quartics. *ArXiv Preprint: 1812.10349*, 2018.

A. Gasnikov, P. Dvurechensky, E. Gorbunov, D. Kovalev, A. Mohhamed, E. Chernousova, and C.A. Uribe. The global rate of convergence for optimal tensor methods in smooth convex optimization. *arXiv:1809.00382*, 2018.

B. Jiang, T. Lin, and S. Zhang. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *arxiv:1811.02427*, 2018.

R.D.C. Monteiro and B.F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2): 1092–1125, 2013.

Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Doklady AN SSSR, translated as Soviet Math.Docl.*, 269:543–547, 1983.

Yu. Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Yu. Nesterov. Implementable tensor methods in unconstrained convex optimization. *CORE Discussion Paper 2018/05, Catholic University of Louvain, Center for Operations Research and Econometrics (CORE)*, 2018.