Cutting plane methods can be extended into nonconvex optimization[1]

Oliver Hinder OHINDER@STANFORD.EDU
Management Science and Engineering Department. Stanford.

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We show that it is possible to obtain an $O(\epsilon^{-4/3})$ runtime — including computational cost — for finding ϵ-stationary points of nonconvex functions using cutting plane methods. This improves on the best known epsilon dependence achieved by cubic regularized Newton of $O(\epsilon^{-3/2})$ as proved by Nesterov and Polyak (2006). Our techniques utilize the convex until proven guilty principle proposed by Carmon, Duchi, Hinder, and Sidford (2017).

Keywords: optimization, cutting plane methods, nonconvex, stationary point, local minima

We consider the problem of finding an ϵ-stationary point x of the function $f : \mathbf{R}^d \to \mathbf{R}$ starting from some point $x^{(0)}$, i.e.,

$$\|\nabla f(x)\| \leq \epsilon$$

under the assumptions that $f(x^{(0)}) - \inf_z f(z)$ is bounded below and the function has Lipschitz first and third derivatives. It is well-known that gradient descent achieves an ϵ^{-2} runtime when the first derivatives are Lipschitz. This was improved to $\epsilon^{-3/2}$ by Nesterov and Polyak (2006) using cubic regularized Newton when the second derivatives are Lipschitz. However, each iteration of cubic regularized Newton is more expensive — it requires Hessian evaluations and solving a linear system. This observation has inspired a line of work developing gradient based methods that improve on the worst-case runtime of gradient descent (Agarwal, Allen-Zhu, Bullins, Hazan, and Ma, 2017; Carmon, Duchi, Hinder, and Sidford, 2016, 2017a; Jin, Netrapalli, and Jordan, 2017; Royer and Wright, 2017). These methods have cheap 'gradient' iteration costs and runtime bounds worse than cubic regularization but better than gradient descent. If low accuracy is desired in high dimensions these gradient based methods are preferable.

What about the regime where the dimension is low but we want to obtain high accuracy? In this case it might be acceptable to have iteration costs that scale polynomially with the dimension if that enables an algorithm with significantly less iterations. Under the assumption that the first and third derivatives to be Lipschitz, our main result is an algorithm that takes

$$\tilde{O}((T_1 + d^\omega) d \epsilon^{-4/3})$$

time to find an ϵ-stationary point, where T_p which refers to cost of one evaluating the function and its first p derivatives and $O(d^\omega)$ denotes the runtime for a linear system solve. To prove our result we utilize the 'convex until proven guilty' principle proposed by Carmon et al. (2017a) to adapt cutting plane methods to find stationary points of nonconvex functions. Cutting plane methods are traditionally used to optimize general convex functions in low dimensions to high accuracy.

1. Extended abstract. Full paper version appears arXiv:1805.08370.

Our result can be contrasted with the results of Birgin, Gardenghi, Martínez, Santos, and Toint (2017) who gives a runtime of $O((T_3+?)\epsilon^{-(p+1)/p})$ where the ? denotes the cost of finding a stationary point of a pth order regularized problem. Letting $p = 1$ gives gradient descent and $p = 2$ cubic regularized Newton. However, for $p > 2$ all known methods for solving pth order have ϵ-dependencies that cause the computational runtime to scale at best with $O(\epsilon^{-3/2})$ corresponding to cubic regularized Newton. Therefore our major contribution is to show it is possible to obtain an $O(\epsilon^{-4/3})$ runtime — including computational cost — for finding ϵ-stationary points of nonconvex functions. See Table 1 for a comparison of our results with existing results.

Lipschitz	method	runtime	dimension-free lower bound (Carmon et al., 2017b,c)
∇f	gradient descent	$T_1\epsilon^{-2}$	$T_1\epsilon^{-2}$
$\nabla f, \nabla^2 f$	Carmon et al. (2017a)	$T_1\epsilon^{-7/4}$	$T_1\epsilon^{-12/7}$
$\nabla f, \nabla^3 f$	Carmon et al. (2017a)	$T_1\epsilon^{-5/3}$	$T_1\epsilon^{-8/5}$
$\nabla^2 f$	cubic reg. Nesterov and Polyak (2006)	$(T_2 + d^\omega)\epsilon^{-3/2}$	$T_2\epsilon^{-3/2}$
$\nabla^p f$	pth reg. Birgin et al. (2017).	$(T_p+?)\epsilon^{-\frac{p+1}{p}}$	$T_p\epsilon^{-(p+1)/p}$
$\nabla f, \nabla^3 f$	This paper. Thm 1.	$((T_1 + d^\omega)d + T_2)\epsilon^{-4/3}$	
$\nabla f, \nabla^3 f$	This paper. Thm 2.	$(T_3 + d^4)\epsilon^{-4/3}$	$T_3\epsilon^{-4/3}$

Table 1: Comparison of the runtime of different algorithms for finding stationary points of nonconvex functions. The question mark is a placeholder for the time to solve a pth order regularization problem.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *Symposium on Theory of Computing*, 2017.

Ernesto G Birgin, JL Gardenghi, José Mario Martínez, Sandra Augusta Santos, and Ph L Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2):359–368, 2017.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. 'Convex until proven guilty': dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of 34th International Conference on Machine Learning*, pages 654–663, 2017a.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *arXiv preprint arXiv:1710.11606*, 2017b.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: First-order methods. *arXiv preprint arXiv:1711.00841*, 2017c.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *arXiv preprint arXiv:1706.03131*, 2017.