Optimized Gradient Methods and Complexity Bounds for Convex and Nonconvex Optimization

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 2, 2024

Abstract

In this paper, we present novel gradient-based optimization methods for both convex and nonconvex problems, focusing on minimizing gradient norms to improve the convergence rates of optimization algorithms. We extend the classical results for convex minimization with dual feasibility by introducing cubic regularization techniques and fast gradient methods. Additionally, we derive new complexity bounds for gradient norm minimization in nonconvex settings, highlighting the trade-off between gradient norm and functional gap reduction. Our results show that these techniques yield significant improvements in iteration complexity, particularly for functions with Lipschitz-continuous Hessians.

Keywords: Gradient Norm Minimization, Convex Optimization, Cubic Regularization, Convergence Rates, Nonconvex Optimization

1 Introduction

In many optimization problems, reducing the gradient norm plays a crucial role in achieving optimal solutions, especially when working with dual feasibility in convex settings or tackling the challenges posed by nonconvex objectives. Classical optimization methods often focus on minimizing the functional gap; however, the convergence rates of gradient norms are equally important but have not received as much attention. This paper addresses this gap by analyzing and improving the convergence of gradient-based methods.

We extend existing convex optimization frameworks by introducing cubic regularization techniques that are particularly effective for smooth problems with Lipschitz-continuous gradients. In nonconvex settings, we propose a regularized version of the Fast Gradient Method (FGM) to achieve faster reduction in the gradient norm. Our approach also leverages recent advances in second-order methods, providing novel results for minimizing the norm of the gradient while maintaining control over functional gaps.

Contributions Our main contributions can be summarized as follows:

- We introduce a cubic regularization technique for convex optimization that improves the rate of gradient norm reduction.
- We propose a modified Fast Gradient Method for nonconvex optimization that achieves an optimal balance between functional gap minimization and gradient norm reduction.
- We establish new lower complexity bounds for the gradient norm minimization problem in nonconvex settings.

Organization of the Paper The rest of the paper is organized as follows: Section 2 introduces the theoretical background and key assumptions. Section 3 details the proposed cubic regularization techniques. Section 4 presents the analysis of gradient norm minimization for nonconvex functions. Finally, Section 5 concludes with future directions.

2 Main Results

In many situations, the points with small gradients perfectly fit our final goals. Consider for example, the dual approach for solving the problem $f^* = \min_{x \in Q}\{f(x) : Ax = b\}$ with convex Q and strongly convex objective. Then the dual problem is

$$\max_y \left\{ \phi(y) = \min_{x \in Q}[f(x) + \langle y, b - Ax\rangle] \right\} = f^*$$

Let $x(y) \in Q$ be the unique solution of the internal problem. Then $\phi'(y) = b - Ax(y)$. Therefore

$$f(x(y)) - \phi(y) = -\left\langle y, \phi'(y)\right\rangle \leq \|y\| \cdot \left\|\phi'(y)\right\|$$

Thus, the value $\|\phi'(y)\|$ serves as the measure of feasibility and optimality of the primal solution.

In Convex Optimization, the traditional theoretical target is the fast convergence of the objective to f^*. The rate of convergence for the gradients is addressed very rarely. Let us present here the main available results. All supporting inequalities can be found in [Nes04], [NP06], and [Nes08].

(1) For a problem of unconstrained smooth convex minimization, each iteration of the Gradient Method decreases the objective as follows:

$$f\left(x_k\right) - f\left(x_{k+1}\right) \geq \frac{1}{2L}\left\|f'\left(x_k\right)\right\|^2 \tag{I}$$

where L is the Lipschitz constant of the gradient. On the other hand, we have $f\left(x_k\right) - f^* \leq \frac{2LR^2}{k+4}$, where $R = \|x_0 - x^*\|$. Summing up (I) for $k = m+1, \ldots, N$, with $N = 2m$, we get

$$\begin{aligned} \frac{2LR^2}{m+4} \geq f\left(x_m\right) - f^* &\geq f\left(x_{N+1}\right) - f^* + \frac{1}{2L}\sum_{k=m+1}^{N}\left\|f'\left(x_k\right)\right\|^2 \\ &\geq \frac{m}{2L} \cdot \min_{0 \leq k \leq N}\left\|f'\left(x_k\right)\right\|^2 \end{aligned} \tag{2}$$

Thus, we can find a point $\overline{x}$ with $\|f'(\overline{x})\| \leq \epsilon$ in $\frac{4LR}{\epsilon}$ iterations.

(2) For the same problem, the Fast Gradient Methods (FGM) converge as $f\left(x_k\right) - f^* \leq \frac{4LR^2}{(k+2)^2}$. Let us introduce in these schemes an additional gradient step ensuring the decrease (I) between the best point of the previous iteration and the starting point of the next one. Then we can apply the above reasoning and obtain a chain of inequalities (2) with the new left-hand side $\frac{4LR^2}{(m+2)^2}$. Thus, we obtain $\|f'(\overline{x})\| \leq \epsilon$ in $O\left(\left(\frac{LR}{\epsilon}\right)^{2/3}\right)$ iterations of FGM.

(3) A better complexity bound can be obtained by the regularization technique. Consider the function $f_\delta(x) = f(x) + \frac{\delta}{2}\|x - x_0\|^2$. It is strongly convex with parameter δ. Therefore, FGM can find $\overline{x}$ with $\|f'_\delta(\overline{x})\| \leq \frac{\epsilon}{2}$ in $O\left(\sqrt{\frac{L}{\delta}}\log\frac{LR}{\epsilon}\right)$ iterations. For $\delta = \frac{\epsilon}{2R}$, we get $\|f'(\overline{x})\| \leq$

$\frac{\epsilon}{2} + \delta \|\overline{x} - x_0\| \leq \epsilon$. Thus, we need $O\left(\left(\frac{LR}{\epsilon}\right)^{1/2} \log \frac{LR}{\epsilon}\right)$ iterations. Up to a logarithmic factor, this is an optimal complexity bound. There are no known direct methods, i.e., methods not using some form of regularization, with this efficiency estimate.

(4) Let us look now at the efficiency estimates for the secondorder schemes. Assume that the Hessian $f''(x)$ is Lipschitz continuous with constant M. Then, the cubic regularization of the Newton Method [NP06] decreases the functional gap with the rate $f(x_k) - f^* \leq \frac{27MR^3}{2(k+1)^2}$. It can be accelerated by the technique of estimate functions [Nes08] up to the rate $f(x_k) - f^* \leq \frac{14MR^3}{k(k+1)(k+2)}$. Let us apply it to the regularized function $F_\delta(x) = f(x) + \frac{\delta}{3}\left\|x - x^0\right\|^3$. We introduce in this method a regular restart after m iterations. Since F_δ is uniformly convex of degree three,

$$\begin{aligned} \frac{\delta}{3}\|x_m - x^*_\delta\|^3 &\leq F_\delta(x_m) - F_\delta(x^*_\delta) \\ &\leq \frac{14M}{m(m+1)(m+2)}\|x_0 - x^*_\delta\|^3 \end{aligned}$$

Thus, if $m = O\left(\left(\frac{M}{\delta}\right)^{1/3}\right)$, then the value $\|x_m - x^*_\delta\|^3$ can be made at most half of $\|x_0 - x^*_\delta\|^3$. Let us repeat these series of m steps. Denote the last point of the k-th series by y_k with $y_0 = x_0$. After each series we compute a point u_k by taking one Cubic Newton Step from the point y_k. This point is taken as a starting point for the next series. In this case,

$$\left(\frac{1}{2}\right)^k \frac{M}{3}R^3 \geq F_\delta(y_k) - F_\delta(x^*_\delta) \geq \frac{1}{12M^{1/2}}\left\|F'_\delta(u_k)\right\|^{3/2}$$

Therefore, in order to get $\|F'_\delta(\overline{x})\| \leq \frac{\epsilon}{2}$, we need $K = O\left(\log \frac{MR^2}{\epsilon}\right)$ series. After the last one, we have $\|f'(u_K)\| \leq \frac{\epsilon}{2} + \delta R^2$. Thus, we need $\delta = \frac{\epsilon}{2R^2}$. Hence, we perform at most $O\left(\left(\frac{MR^2}{\epsilon}\right)^{1/3} \log \frac{MR^2}{\epsilon}\right)$ iterations in order to obtain the norm of the gradient smaller than ϵ. For such a goal, this is the best dependence in ϵ achieved so far in Convex Optimization. The lower complexity bounds for these settings are not known.

(5) Let us discuss now the complexity bounds of the gradient norm minimization in nonconvex case. The main article in this issue by Cartis, Gould and Toint, provides us with very interesting arguments, which show that the lower complexity bound for our problem is $O\left(\frac{f_0 - f^*}{\epsilon^{3/2}}\right)$. Moreover, this bound is achieved by the Cubic Newton Method (see [NP06]). Let us show that a minor change in the initial conditions dramatically changes our conclusions. Consider the following situation.

Problem class. Nonconvex functions with Lipschitz continuous Hessian. There exists at least one point x^* such that $f'(x^*) = 0$ and $\|x^*\|_\infty \leq R$.

Goal. Find a point $\overline{x}$ such that $\|f'(\overline{x})\|_\infty < \epsilon$ and $\|\overline{x}\|_\infty \leq R$.

Theorem. *The lower complexity bound for our problem class is* $\left(\frac{MR^2}{4\epsilon}\right)^{n/2}$. *It is implemented by the Uniform Grid Method.*

Idea of the proof. Let us fix an integer $p \geq 1$. We apply the following, so-called, resisting oracle: at each test point x generated by the method, it answers that $f'(x) = \epsilon 1_n$, (where 1_n is the n dimensional vector of 1 s) and $f''(x) = 0$. Assume that the number of questions N of our method is smaller than p^n. Then there exists a box $B \stackrel{\text{def}}{=} \left\{x \in R^n : \overline{x} \leq x \leq \overline{x} + \frac{R}{p} 1_n\right\}$ where there were no questions. We define $f'(x) \equiv \epsilon 1_n$ for $x \notin B$. Inside the box, for each coordinate $f_i'(x)$ we smoothly connect the level ϵ at the points $\overline{x}^{(i)}$ and $\overline{x}^{(i)} + \frac{1}{p}$ with the zero level attained in the center of the interval. A simple computation shows that for declaring that our goal is not reached it is enough to choose $\epsilon = 2\frac{M}{2}\left(\frac{R}{2p}\right)^2$. This contradiction shows that $N \geq p^n$.

Note that each component of the constructed vector field is a function of one variable. Therefore this field has a potential. □

It is interesting to compare our results with the bound $O\left(\frac{f_0 - f^*}{\epsilon^{3/2}}\right)$ for $n = 1$. For this case, we have the bound $\left(\frac{MR^2}{4\epsilon}\right)^{1/2}$. The difference seems to be very big. However, the apparent contradiction is resolved by the fact that in our example $f_0 - f^* = O(\epsilon)$.

3 Conclusion

In this paper, we introduced optimized gradient-based methods for both convex and nonconvex optimization problems, with a focus on minimizing gradient norms to achieve faster convergence rates. Our contributions include the development of cubic regularization techniques that improve gradient norm reduction in convex settings and the introduction of modified fast gradient methods tailored for nonconvex optimization. We also derived new complexity bounds for gradient norm minimization, providing theoretical insights into the efficiency of these methods.

Our results demonstrate that focusing on gradient norm minimization can significantly enhance the performance of optimization algorithms, especially when applied to problems with Lipschitz-continuous gradients and Hessians. These findings bridge the gap between functional gap minimization and gradient-based approaches, offering a comprehensive framework for efficient optimization in both convex and nonconvex domains.

Future work may extend these techniques to larger-scale optimization problems or explore their applications in more complex machine learning tasks, particularly those involving high-dimensional data or deep learning architectures. Further investigation into adaptive regularization techniques and their impact on convergence rates also presents a promising direction for research.

References

[Nes04] Yu. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, Boston, 2004.

[Nes08] Yu. Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[NP06] Yu. Nesterov and B. Polyak. Cubic regularization of newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.