Gradient Descent Efficiently Finds the Cubic-Regularized Non-Convex Newton Step

Yair Carmon
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA
yairc@stanford.edu

John Duchi
Department of Statistics
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA
jduchi@stanford.edu

Abstract

We consider the minimization of non-convex quadratic forms regularized by a cubic term, which exhibit multiple saddle points and poor local minima. Nonetheless, we prove that, under mild assumptions, gradient descent approximates the global minimum to within $\varepsilon$ accuracy in $O(\varepsilon^{-1} \log(1/\varepsilon))$ steps for large $\varepsilon$ and $O(\log(1/\varepsilon))$ steps for small $\varepsilon$ (compared to a condition number we define), with at most logarithmic dependence on the problem dimension.

1 Introduction

We study the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{2} x^\top A x + b^\top x + \frac{\rho}{3} \|x\|^3,$$  

(1)

where the matrix $A$ is symmetric and possibly indefinite. The problem (1) arises in Newton’s method with cubic regularization, first proposed by Nesterov and Polyak [14]. The method consists of the iterative procedure

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ \nabla g(x_t)^\top (x - x_t) + \frac{1}{2} (x - x_t)^\top \nabla^2 g(x_t) (x - x_t) + \frac{\rho}{3} \|x - x_t\|^3 \right\}$$  

(2)

for (approximately) minimizing a general smooth function $g$, requiring sequential solutions of problems of the form (1). The Nesterov-Polyak scheme (2) falls into the broader framework of trust-region methods [5, 3]. Such methods are among the most practically successful and theoretically sound approaches to non-convex optimization [5, 14, 3]. In particular, as Nesterov and Polyak first showed, under certain smoothness assumptions on $g$, it is possible to establish a rate at which $\nabla^2 g$ converges to a positive semidefinite matrix.

Standard methods for solving the problem (1) exactly require either factorization or inversion of the matrix $A$. However, the cost of these operations scales poorly with the problem dimensionality; in very large scale problems (such as, for example, training deep neural networks) even computing all the entries of the Hessian may be infeasible. Moreover, in methods that rely on matrix inversion or factorization it is often difficult to exploit matrix structure such as sparsity. In contrast, matrix-free methods, which access $A$ only through matrix-vector products, often scale well to high dimensions and leverage structure in $A$ (c.f. [18]), particularly when $A$ is a Hessian. Even without special structure, the product $\nabla^2 g(x)v$ often admits the finite difference approximation $\delta^{-1}(\nabla g(x + \delta v) - \nabla g(x))$, which requires only two gradient evaluations. In neural networks and other arithmetic circuits, back-propogation-like methods allow exact computation of Hessian-vector products at a

* This is an extended abstract. A full version containing all the proofs is available online on arXiv.
similar cost [16, 17]. It is thus of practical and theoretical interest to explore matrix-free methods guaranteed to solve (1) efficiently.

In this paper, we prove that gradient descent, perhaps the simplest matrix-free method, efficiently converges to the global minimum of the (“substantially non-convex” [14]) cubic problem (1), and we provide empirical evidence supporting our theoretical guarantees. More precisely, we show that gradient descent solves problem (1) to accuracy $\varepsilon$ in at most $O(1/\varepsilon)$ iterations, and that it exhibits linear convergence when $\varepsilon$ is small with respect to problem conditioning. Our result implies that when gradient descent is used to approximate (2), the rate of convergence of $\nabla^2 g$ is maintained, thus demonstrating a first-order method with second-order convergence—this is discussed in detail in the full version of the paper.

A number of researchers have considered low-complexity methods for solving the problem (1). Cartis et al. [3] propose solution methods working in small Krylov subspaces, and Bianconcini et al. [2] apply a matrix-free gradient method (NMGRAD) in conjunction with an early stopping criterion for the problem. Both approaches exhibit strong practical performance, and they enjoy first-order convergence guarantees for the overall optimization method (in which problem (1) is an iteratively solved sub-problem). In both works, however, it appears challenging to give convergence guarantees for the iterative subproblem solvers, and they do not provide the second-order convergence rates of Nesterov and Polyak’s Newton method (2). In their paper [4], Cartis et al. give sufficient conditions for a low-complexity approximate subproblem solution to guarantee such second-order rates, but it is not clear how to construct a first-order method fulfilling these conditions. Closely related and intensely studied is the quadratic trust-region problem [5, 8, 9, 6], where one replaces the regularizer $(\rho/3)\|x\|^3$ with the constraint $\|x\| \leq R$ for some $R > 0$. Classical low complexity methods for this problem include subspace methods and the heuristic Steihaug-Toint truncated conjugate gradient method; we know of no convergence rates in either case.

Recently, fast matrix-free approaches were proposed for the trust-region problem [10, 11] and, concurrently to this work, for the problem (1) [1]. These approaches reduce the problems to short sequences of approximate eigenvector computation and convex optimization problems; thus obtaining accelerated rates of $O(1/\sqrt{\varepsilon})$, which are better than those we achieve when $\varepsilon$ is large relative to problem conditioning. However, while these works indicate that solving (1) is never harder than approximating the bottom eigenvector of $A$, the regime of linear convergence we identify shows that it is sometimes much easier. In addition, we believe that our results provide interesting insights on the potential of gradient descent and other direct methods for non-convex problems.

Another related line of work is the study of the behavior of gradient descent around saddle-points and its ability to escape them [7, 12, 15, 13]. A common theme in these works is an “exponential growth” mechanism that pushes the gradient descent iterates away from critical points with negative curvature, similar to the amplification of large eigen-directions in the classical power method. This mechanism plays a prominent role in our analysis as well, highlighting the implications of negative curvature for the dynamics of gradient descent.

## 2 Preliminaries and basic convergence guarantees

Before continuing we provide some notation to our approach to problem (1), where $\rho > 0$, $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$ is symmetric and possibly indefinite, and $\|\cdot\|$ denotes the Euclidean norm. The eigenvalues of $A$ are $\lambda^{(1)}(A) \leq \lambda^{(2)}(A) \leq \cdots \leq \lambda^{(d)}(A)$, where any of the eigenvalues $\lambda^{(i)}(A)$ may be negative. We define the eigengap of $A$ by $\gamma = \lambda^{(1)}(A) - \lambda^{(1)}(A)$ where $k$ is the first eigenvalue of $A$ strictly larger than $\lambda^{(1)}(A)$. For any vector $v, v^{(i)}$ denotes the $i$th coordinate of $v$ in the eigen-basis of $A$.

We let $\|\cdot\|_{\text{op}}$ be the standard $\ell_2$-operator norm, so $\|A\|_{\text{op}} = \max_{\|u\| = 1} \|Au\|$, and define

$$\gamma = -\lambda^{(1)}(A), \quad \gamma_+ = \gamma \vee 0, \quad \text{and} \quad \beta = \|A\|_{\text{op}} = \max\{|\lambda^{(1)}(A)|, |\lambda^{(d)}(A)|\},$$

so that the function $f$ is non-convex if and only if $\gamma > 0$ (and is convex when $\gamma_+ = 0$). We remark that our results continue to hold when $\beta$ is an upper bound on $\|A\|_{\text{op}}$ rather than its exact value.

We let $s$ denote a solution to problem (1), i.e. a global minimizer of $f$ that has the characterization [14, Section 5] as the solution to the equality and inequality

$$\nabla f(s) = (A + \rho \|s\| I)s + b = 0 \quad \text{and} \quad \rho \|s\| \geq \gamma,$$  

(3)
and is unique whenever $\rho \|s\| > \gamma$.

The global minimizer admits the following equivalent characterization whenever the vector $b$ is not orthogonal to the eigenspace associated with $\lambda^{(1)}(A)$.

**Claim 1.** If $b^{(1)} \neq 0$, $s$ is the unique point that satisfies $\nabla f(s) = 0$ and $b^{(1)}s^{(1)} \leq 0$.

Additionally, the norm of $s$ is upper bounded by

$$
\|s\| \leq \frac{\gamma}{2\rho} + \sqrt{\left(\frac{\gamma}{2\rho}\right)^2 + \frac{\|b\|^2}{\rho}} \leq \frac{\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{\|b\|^2}{\rho}} \leq R,
$$

(4)

and admits the lower bound

$$
\|s\| \geq R_c \triangleq \frac{-b^T Ab}{2\rho \|b\|^2} + \sqrt{\left(\frac{b^T Ab}{2\rho \|b\|^2}\right)^2 + \frac{\|b\|^2}{\rho}} \geq -\frac{\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{\|b\|^2}{\rho}}.
$$

(5)

The gradient descent method begins at some initializations $x_0 \in \mathbb{R}^d$ and generates iterates $x_1, x_2, \ldots$ according to

$$
x_{t+1} = x_t - \eta \nabla f(x_t) = (I - \eta A - \rho \eta \|x_t\| I)x_t - \eta b.
$$

(6)

Throughout our analysis we make the following assumptions about the step size $\eta$ and $x_0$.

**Assumption A.** The step size $\eta$ in (6) satisfies $0 < \eta \leq \frac{1}{4(\beta + \rho R)}$.

**Assumption B.** The initialization of (6) satisfies $x_0 = -rb/\|b\|$, with $0 \leq r \leq R_c$.

A key to our analysis is the fact that $\|x_t\|$ is monontic.

**Lemma 1.** Let Assumptions A and B hold. Then for all $t \geq 0$, the iterates (6) of gradient descent satisfy $x_t^T \nabla f(x_t) \leq 0$, and the norms $\|x_t\|$ are non-decreasing and satisfy $\|x_t\| \leq R$.

Assumption B guarantees that $b^{(1)}x_0^{(1)} \leq 0$ while Assumption A and the norm bound in Lemma 1 together guarantee that $b^{(1)}x_t^{(1)} \leq c_t b^{(1)}x_{t-1}^{(1)}$ for some $c_t \geq 0$, and thus $b^{(1)}x_t^{(1)} \leq 0$ for every $t$ by induction. By standard arguments, limit points of gradient descent are critical point. Hence, when $b^{(1)} \neq 0$ every partial limit of gradient descent satisfies Claim 1 and is therefore the unique global minimum $s$, which implies $x_t \to s$ as $t \to \infty$. This also allows us to strengthen the norm bound of Lemma 1 to $\|x_t\| \leq \|s\|$ for every $t$.

### 3 Main result: non-asymptotic convergence rates

**Theorem 1.** Let Assumptions A and B hold, $b^{(1)} \neq 0$, and $\varepsilon > 0$. Then $f(x_t) \leq f(s) + \varepsilon$ for all

$$
t \geq T_c \triangleq \frac{1}{\eta} \left( \tau_{\text{grow}}(b^{(1)}) + \tau_{\text{converge}}(\varepsilon) \right) \min \left\{ \frac{1}{\rho \|s\| - \gamma}, \frac{10 \|s\|^2}{\varepsilon}, \sqrt{\frac{10 \|s\|^2}{\varepsilon}} \right\},
$$

(7)

where $\text{gap}' = (\text{gap} \wedge \rho \|s\|) \mathbb{I}_{\{\varepsilon \geq 10 \|s\|^2 (\rho \|s\| - \gamma)\}}$ and

$$
\tau_{\text{grow}}(b^{(1)}) = 6 \log \left( 1 + \frac{\gamma^2}{4\rho \rho R_c} \right) \quad \text{and} \quad \tau_{\text{converge}}(\varepsilon) = 6 \log \left( \frac{(\beta + 2\rho \|s\|) \|s\|^2}{\varepsilon} \right).
$$

Theorem 1 shows that the rate of convergence changes from roughly $O(1/\varepsilon)$ to $O(\log(1/\varepsilon))$ as $\varepsilon$ decreases, with an intermediate gap-dependent rate of $O(1/\sqrt{\varepsilon})$. The constants $\tau_{\text{grow}}$ and $\tau_{\text{converge}}$ correspond to a period ($\tau_{\text{grow}}$) in which $\|x_t\|$ grows exponentially until reaching the basin of attraction to the global minimum and a period ($\tau_{\text{converge}}$) of linear convergence to $s$. The term $\tau_{\text{grow}} = 0$ when the problem is convex (as $\gamma \leq 0$), so we see that the period of exponential growth is directly related to the negative curvature in $f$.

The dependence of our result on $|b^{(1)}|$ (and the implicit assumption $b^{(1)} \neq 0$) can be eliminated by adding to $b$ a small perturbation uniformly distributed on a sphere of radius $\frac{\rho \varepsilon}{12(\beta + 2\rho \|s\|)}$, and
applying gradient descent on the modified problem instance. As we show in the full version of the paper, this guarantees that with probability at least $1 - \delta$, gradient descent finds an $\varepsilon$-suboptimal solution (to the unperturbed problem), in a number iterations at most a constant times $T_\varepsilon$ defined in (7), with $|b^{(1)}|$ replaced by $\rho \|s\|^2 \delta/(4\sqrt{d})$.

Figure 1 depicts the number of gradient steps required to find a point $x$ satisfying $f(x) - f(s) \leq \varepsilon(f(0) - f(s))$ as a function of $1/\varepsilon$, for random problem instances with a small value of $\rho \|s\| - \gamma$. The slopes in this log-log plot reveal good agreement between theory and experiment, and suggests there exist instances for which our upper bounds are tight up to sub-polynomial factors.

4 Proof outline

We provide a brief overview of the proof of the linear convergence part of Theorem 1. A comprehensive treatment of the proof appears in the full version of this paper. We tacitly let Assumptions A and B hold throughout the section.

Our first result has the structure of a linear convergence guarantee.

**Lemma 2.** For each $t > 0$, we have

$$\|x_t - s\|^2 \leq \left(1 - \eta \left[\rho \|x_t\| - \left(\gamma - \frac{\rho \|s\| - \gamma}{2}\right)\right]\right) \|x_{t-1} - s\|^2 - \eta \rho \left(\|s\| - \|x_{t-1}\|\right)^2 \|s\|.$$  

The above recursion implies geometric decrease in $\|x - s\|$ only when $\rho \|x_t\|$ is larger than $\frac{1}{2} (\rho \|s\| - \gamma)$, which may be non-trivial for non-convex problem instances (with $\gamma > 0$). Using the fact that $\|x_t\|$ is non-decreasing (Lemma 1), Lemma 2 immediately implies the following.

**Lemma 3.** If $\rho \|x_t\| \geq \gamma - \frac{1}{2} (\rho \|s\| - \gamma) + \delta$ for some $t \geq 0$, then for all $\tau \geq 0$,

$$\|x_{t+\tau} - s\|^2 \leq \left(1 - \eta \delta\right)^\tau \|x_t - s\|^2 \leq 2 \|s\|^2 e^{-\eta \delta \tau}.$$  

It remains to show that $\rho \|x_t\|$ will quickly exceed any level below $\gamma$. Fortunately, as long as $\rho \|x_t\|$ is below $\gamma - \delta$, $|x_t^{(1)}|$ grows faster than $\tau_\varepsilon/(1 + \eta \delta)$, which the next lemma leverages.

**Lemma 4.** Let $\delta > 0$. Then $\rho \|x_t\| \geq \gamma - \delta$ for all $t \geq \frac{2}{\eta \delta} \log \left(1 + \frac{\gamma^2}{4 \rho \|s\|}\right)$.  

To give the linear convergence regime of Theorem 1, we first apply Lemma 4 with $\delta = \frac{1}{3} (\rho \|s\| - \gamma)$, which yields $\rho \|x_t\| \geq \gamma - \frac{1}{3} (\rho \|s\| - \gamma)$ for every $t \geq \frac{\tau_{\text{conv}}(\varepsilon)}{\eta (\rho \|s\| - \gamma)}$. We may then apply Lemma 3 and standard smoothness arguments to obtain $f(x_t) \leq f(s) + \varepsilon$ after additional gradient descent iterations.
References


