Cubic Regularization Literature Review

Aekus Bhathal
University of California, Berkeley
abhathal@berkeley.edu

Abstract

We do an in-depth analysis of Nesterov and Polyak's seminal paper "Cubic regularization of Newton method and its global performance". We motivate cubic regularization by drawing the connection between Newton's method and Gradient Descent, and show that cubic regularization stabilizes Newton's method. Then, we examine the algorithm for cubic-regularized Newton's method proposed by Nesterov and Polyak, providing analysis of the convergence bounds proven by the paper. We take care to consider the the non-degenerate case as a special condition, providing intuition for why relaxing the cubic coefficient is appropriate. Finally, we prove that the cubic regularization objective is only locally non-convex with strict bounds. We use this to show that an update step of the cubic regularization problem must be in a convex region of the function.

1 Motivation

1.1 Local Models and Upper Bounds

We begin with a discussion of second-order methods, specifically Newton's method and it's descendants. Given an unconstrained optimization problem,

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

Newton's method completes an update step

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

From a local model view, the motivation behind Newton's method is clear: we minimize a second-order Taylor series around $f(x_k)$.

Definition 1. *Suppose f is n-times differentiable. Let $\tilde{f}_{x_k}(x; n)$ be the n-th order Taylor expansion around x_k.*

Then, we can define the update step of Newton's method as

$$\tilde{f}_{x_k}(x; 2) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \tag{1}$$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left[\tilde{f}_{x_k}(x; 2) \right] \tag{2}$$

Indeed, this update step looks very similar to the local model view of gradient descent. Suppose the gradient of f is Lipschitz continuous. Then,

$$\|\nabla f(x) - \nabla f(y)\| \leq \frac{L}{2} \|x - y\| \tag{3}$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We find the gradient descent with step $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ is a solution of the problem

$$\tilde{f}_{x_k}(x;1) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \tag{4}$$

$$\bar{f}_{x_k}(x;1) \triangleq \tilde{f}_{x_k}(x;1) + \frac{L}{2}\|x - x_k\|^2 \tag{5}$$

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left[\bar{f}_{x_k}(x;1)\right] \tag{6}$$

Lemma 1. *Suppose $f(x)$ has Lipschitz continuous gradient as defined in (3). Then, for any iterate x_k in the gradient descent scheme defined in (6), $f(x_{k+1}) \leq f(x_k)$.*

Proof. It is trivial that for all $x \in \mathbb{R}^n$, $\bar{f}_{x_k}(x;1) \geq \tilde{f}_{x_k}(x;1)$. With the added Lipschitz continuous gradient assumption, Nesterov and Polyak show that $\bar{f}_{x_k}(x;1)$ is also a global first-order upper bound of $f(x)$. It follows, then that

$$\begin{aligned} f(x_{k+1}) &= f\Big(\operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left[\bar{f}_{x_k}(x;1)\right]\Big) \\ &\leq \bar{f}_{x_k}\Big(\operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left[\bar{f}_{x_k}(x;1)\right];1\Big) \\ &= \min_{x \in \mathbb{R}^n} \bar{f}_{x_k}(x;1) \\ &\leq \bar{f}_{x_k}(x_k;1) \\ &= f(x_k) \end{aligned}$$

□

This shows that gradient descent always descends. With a bit more analysis, we see it also converges.

1.2 Newton's Method as Anisotropic Gradient Descent

Taking a slight detour, the similarity between (6) and (2) suggests a different view of this analysis that conveys the relationship between gradient descent and Newton's method and why we might care about second order methods.

Proposition 1. *Suppose that f is twice differentiable. Then,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \iff -LI \preceq \nabla^2 f(x) \preceq LI$$

Applying Proposition 1 to the second order taylor expansion about x_k, we have

$$\tilde{f}_{x_k}(x;2) \leq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 = \bar{f}_{x_k}(x;1) \tag{7}$$

If $\nabla^2 f(x) = LI$ (i.e. the Hessian is isotropic and strictly convex), we in fact have an equivalence $\tilde{f}_{x_k}(x;2) = \bar{f}_{x_k}(x;1)$, meaning that Newton's Method is equivalent to gradient descent with step size $\frac{1}{L}$. The largest difference between Newton's Method and gradient descent, and perhaps the advantage of Newton's method, seems to be when the hessian is largely anisotropic. In such cases, gradient descent (if chosen with step size $\frac{1}{L}$) approximates an isotropic hessian with the largest eigenvalue, losing much information.

1.3 Derivation of Cubic Regularization

We now formulate cubic regularization. Suppose f has Lipschitz continuous Hessian under the operator norm. Then,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_{\text{op}} \leq \frac{L}{2}\|x - y\|$$

Nesterov and Polyak [1] show that

$$\bar{f}_{x_k}(x;2) = \tilde{f}_{x_k}(x_k;2) + \frac{L}{6}\|x - x_k\|^3 \tag{8}$$

is a global second-order upper bound of f. Adapting Lemma 1 to the case of a Lipschitz continuous Hessian shows that

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left[\bar{f}_{x_k}(x; 2) \right] \tag{9}$$

defines a scheme that also guarantees convergence. Thus, it becomes plausible to consider cubic regularization as a more stable second-order optimization method. We explore what this means in the following section.

1.4 Instability of Newton's Method

Nesterov and Polyak [1] highlight a few issues pertaining to Newton's method. The most apparent issue is degeneracy in the Hessian, making $[\nabla^2 f(x)]^{-1}$ not exist. This is solved by Levenberg-Marquardt regularization by augmenting the hessian with some small μI. By making the Hessian more isotropic, we can leverage the stability of gradient descent. Later, we will show that cubic regularization solved with Lagrangian duality uses a similar regularization to avoid issues pertaining to Hessian degeneracy.

An additional issue with Newton's method is the tendency to converge to saddle points and local maxima. We explore such a case in Example 1.

Example 1. *Suppose f has Lipschitz continuous Hessian, is locally a concave quadratic, and obtains a local maximum in that region.*

$$\begin{gathered} \forall x \in \mathcal{R},\ f(x) = \frac{1}{2} x^T A x + b^T x \\ A \prec 0 \\ -A^{-1} b \in \mathcal{R} \end{gathered}$$

If $x_k \in \mathcal{R}$, $\tilde{f}_{x_k}(x; 2) = f(x)$, so the next step of Newton's method would take $x_{k+1} = -A^{-1}b$, the local maximum. $\nabla f(x_{k+1}) = 0$, so Newton's method would get stuck here.

The problem in Example 1 occurs because Newton's method solves first-order optimality conditions, so it indiscriminately converges to stationary points. Cubic regularization as defined in (9) would not have this instability since cubic regularization descends over the function family of lipschitz continuous hessians. Given this, we might also expect cubic regularization to be more likely to descend other function families.

2 Convergence Rates

2.1 Cubic Regularization of Newton's Method

The equation (6) is not precisely the objective being minimized by Nesterov and Polyak [1]. Instead, we relax the cubic coefficient. There are a few reasons this makes sense. For one, it may not be tractable to find a tight Lipschitz bound on the Hessian. When working with nonconvex function families, some other empirically found coefficient might obtain better results. It may also be true that the Lipschitz bound locally is tighter than the global Lipschitz bound globally, so an adaptive scheme where coefficient is free to change every iteration might generalize to more function families. Define

$$T_M(x) \triangleq \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left[\tilde{f}_x(y; 2) + \frac{M}{6} \|y - x\|^3 \right] \tag{10}$$

$$\hat{f}_M(x) \triangleq \min_{y \in \mathbb{R}^n} \left[\tilde{f}_x(y; 2) + \frac{M}{6} \|y - x\|^3 \right] \tag{11}$$

If $M = 0$, this is precisely the Newton step. If $M = L$, this relates to minimizing $\bar{f}_x(y; 2)$. If $M \geq L$, under the Lipschitz Hessian assumption, we have a second order upper bound on $f(x)$, meaning setting $x_{k+1} = T_M(x_k)$ is a descent algorithm:

$$\begin{aligned} f(x_{k+1}) &\leq \bar{f}_{x_k}(x_{k+1}; 2) \\ &\leq \bar{f}_{x_k}(x_{k+1}; 2) + \frac{M-L}{6}\|x_{k+1} - x_k\|^3 \\ &= \hat{f}_M(x_k) \\ &\leq f(x_k) \end{aligned}$$

Now, we see the cubic regularization algorithm presented in the paper [1]:

Algorithm 1 Cubic Regularization of Newton's Method

```
choose $x_0 \in \mathbb{R}^n$
choose $L_0 \in [0, L]$
$k \leftarrow 1$
while end condition not met do
    Choose $M_k \in [L_0, 2L]$ s.t. $f(T_{M_k}(x_k)) \leq \hat{f}_M(x_k)$
    $x_{k+1} \leftarrow T_{M_k}(x_k)$
    $k \leftarrow k + 1$
end while
return $x_k$
```

We still haven't discussed how we can efficiently solve $T_M(x)$, which we'll leave to a later part.

2.2 Convergence

At any local optimum x^*, for some twice differentiable $f(x)$, we must have that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$. This inspires Nesterov and Polyak [1] to give a measure of local optimality:

$$\mu_M(x) \triangleq \max\left\{\sqrt{\frac{2}{L+M}\|\nabla f(x)\|}, -\frac{2}{2L+M}\lambda_n(\nabla^2 f(x))\right\} \tag{12}$$

Proposition 2. *For some twice differentiable $f(x)$ and $M \geq 0$,*

$$\mu_M(x) = 0 \iff \|\nabla f(x)\| = 0,\ \nabla^2 f(x) \succeq 0$$

If $f(x)$ is convex in some region $\mathcal{C} \subseteq \mathbb{R}^n$, we also have that

$$\forall x \in \mathcal{C}, \mu(x) = \sqrt{\frac{2}{L+M}\|\nabla f(x)\|} \tag{13}$$

For $M = L$, $\mu_M(x_k)^2 = \|x_{k+1} - x_k\|$ for a gradient descent scheme with stepsize $\frac{1}{L}$. One interpretation of $\mu(x)$ in a locally convex set is the square root step distance for gradient descent. If $M \in [L, 2L]$, then $\mu_M(x_k) \leq \sqrt{\|x_{k+1} - x_k\|}$.

Nesterov and Polyak [1] show that $\mu_M(T_M(x)) \leq \|x - T_M(x)\|$, therefore for x_k that Algorithm 1 chooses, $\mu_M(x_{k+1}) \leq \|x_{k+1} - x_k\|$. Fixing the convergence of $\mu_M(x_k)$, the distance between iterates converges quadratically with Algorithm 1 with respect to the distance between iterates of gradient descent.

Using the bound on $\mu_M(T_M(x))$, the paper [1] provide a global convergence bound.

$$\min_{1\leq i\leq k} \mu_L(x_i) \leq \frac{8}{3}\Big(\frac{3(f(x_0) - f^*)}{2k \cdot L_0}\Big)^{1/3} \tag{14}$$

Importantly, since $u_L(x_i) \propto \sqrt{\|\nabla f(x_i)\|}$, we have that

$$\min_{1\leq i\leq k} \|\nabla f(x_i)\| \leq O(k^{-2/3}) \tag{15}$$

Let $\bar{x}_k$ be the best iterate so far,

$$\bar{x}_k = \arg\min_{1\leq i\leq k} \mu_L(x) \tag{16}$$

Then, we have $\lim_{k\to\infty} \mu_L(\bar{x}_k) = 0$. Hence per Proposition 2,

$$\lim_{k\to\infty} \|\nabla f(\bar{x}_k)\| = 0, \ \lim_{k\to\infty} \lambda_n(\nabla^2 f(\bar{x}_k)) \geq 0 \tag{17}$$

With assumptions about the second order differentiability of $f(x)$, and since Nesterov and Polyak [1] show the sequence $\{\bar{x}_k\}$ converges to a limit point x^* (which is a direct consequence of $\{x_k\}$ converging to x^*), Algorithm 1 converges to a local minimum:

$$\lim_{k\to\infty} \bar{x}_i = x^*, \ \|\nabla f(x^*)\| = 0, \ \nabla^2 f(x^*) \succeq 0, \ f(x^*) = f^* \tag{18}$$

Importantly, this means that Algorithm 1 will not converge to a non-degenerate saddle point or local maximum, a notion Nesterov and Polyak [1] formalizes in Lemma 6 of their paper.

2.3 Non-degenerate Case

Let's turn our discussion to the choice of L_0. If there is some point x_k that is degenerate, then $\lambda_n(\nabla^2 f(x_k)) = 0$. There exists some non-zero $u \in \mathbb{R}^n$ and constant $c \in \mathbb{R}$ such that

$$\langle \nabla^2 f(x_k)cu, cu\rangle = 0 \tag{19}$$

Consider the next step of Algorithm 1:

$$\begin{aligned}
\hat{f}_M(x) &= \min_{y\in\mathbb{R}^n} \left[f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), (y - x_k)\rangle + \frac{M}{6}\|y - x_k\|^3 \right] \\
&\leq f(x_k) + \langle \nabla f(x_k), cu\rangle + \frac{1}{2}\langle \nabla^2 f(x_k)cu, cu\rangle + \frac{Mc^3}{6}\|u\|^3 \\
&= f(x_k) + \langle \nabla f(x_k), u\rangle + \frac{Mc^3}{6}\|u\|^3
\end{aligned}$$

If we take $M = 0$, then if $\langle \nabla f(x_k), u\rangle \neq 0$, we can choose $c \in \{+\infty, -\infty\}$ such that $\langle \nabla f(x_k), cu\rangle = -\infty$ so the problem is ill-defined. With $M > 0$, we effectively penalize selecting large c, so this degeneracy is not an issue.

If we know that our function is some neighbourhood around a local minima is Non-degenerate, then there is no need to enforce $M \in [L_0, 2L]$. Nesterov and Polyak [1] prove that by relaxing this constraint, so that $M \in (0, 2L]$, we obtain quadratic convergence as a function of the smallest eigenvalue. Let x_0 be some point such that $\nabla^2 f(x_0) \succ 0$ and $\frac{L\|\nabla f(x_0)\|}{\lambda_n^2(\nabla^2 f(x_0))} \leq \frac{1}{4}$

$$\|\nabla f(x_k)\| \leq \lambda_n^2(\nabla^2 f(x_0)) \frac{9e^{3/2}}{16L2^{(2^k)}} \tag{20}$$

Most interestingly, the starting assumption is sufficient so that $\nabla f(x_k) \succ 0$, meaning not only does this algorithm never arrive at a degenerate point, but the path it traces is also locally convex.

3 Solving Cubic Regularization

3.1 Standard Method

Nesterov and Polyak [1] have provided convergence rates of cubic-regularized Newton's method in terms of the number of iterations, but we have yet to show that each iteration of cubic-regularized Newton's method is computationally efficient. First, let us explore the form of the objective function in (10). We rewrite it here with new notation for convenience:

$$\begin{aligned}
v_u(h, x) &\triangleq \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + \frac{M}{6}\|h\|^3 \\
T_M(x) &= \operatorname*{arg\,min}_{h\in\mathbb{R}^n} \left[v_u(h, x) \right]
\end{aligned}$$

The benefit of this form over Newton's method is it guarantees a bounded minimum since the cubic term will asymptotically grow faster than the quadratic term.

Lemma 2. *If $f(x)$ is not convex, $v_u(h,x)$ is not convex in some ball $\|h\| \leq r$, but is convex outside some ball $\|h\| > 2r$.*

Proof. We begin by computing the hessian of $v_u(h,x)$

$$\nabla_h^2 \, v_u(h,x) = \nabla_x^2 f(x) + M\frac{hh^T + \|h\|^2 I}{2\|h\|} \tag{21}$$

Let u_n be the eigenvector corresponding to the minimum eigenvalue of $\nabla^2 f(x)$. First we show that for $\|h\| \leq r$:

$$\min_y \Big[\langle \nabla_h^2 \, v_u(h,x)y, y\rangle\Big] < 0$$

$$\begin{aligned}
\max_{\|h\|\leq r} \min_y \Big[\langle \nabla_h^2 \, v_u(h,x)y, y\rangle\Big] &= \max_{\|h\|\leq r} \min_y \Big[\langle \nabla_x^2 f(x)y, y\rangle + M\frac{(h^Ty)^2}{2\|h\|} + \frac{M}{2}\|h\|\|y\|^2\Big] \\
&\leq \min_y \max_{\|h\|\leq r} \Big[\langle \nabla_x^2 f(x)y, y\rangle + M\frac{(h^Ty)^2}{2\|h\|} + \frac{M}{2}\|h\|\|y\|^2\Big] \\
&= \min_y \Big[\langle \nabla_x^2 f(x)y, y\rangle + Mr\|y\|^2\Big] \\
&= \|y\|^2(\lambda_n + Mr)
\end{aligned}$$

Hence, $\|h\| \leq r < -\frac{\lambda_n}{M}$ guarantees $v_u(h,x)$ is not convex. Now, we show that for some R and for all $\|h\| \geq R$:

$$\min_y \Big[\langle \nabla_h^2 \, v_u(h,x)y, y\rangle\Big] \geq 0$$

$$\begin{aligned}
\min_{\|h\|\geq R} \min_y \Big[\langle \nabla_h^2 \, v_u(h,x)y, y\rangle\Big] &= \min_y \min_{\|h\|\geq R} \Big[\langle \nabla_x^2 f(x)y, y\rangle + M\frac{(h^Ty)^2}{2\|h\|} + \frac{M}{2}\|h\|\|y\|^2\Big] \\
&= \min_y \Big[\langle \nabla_x^2 f(x)y, y\rangle + \frac{M}{2}R\|y\|^2\Big] \\
&= \|y\|^2(\lambda_n + \frac{MR}{2})
\end{aligned}$$

We see that $R \geq \frac{-2\lambda_n}{M}$ guarantees that $v_u(h,x)$ is convex. Since $R > 2r$ the proof is complete. □

This analysis shows that $v_u(h,x)$ obtains convexity outside of some ball. This means any local optimums found outside of this ball are global optimums outside of this ball. Curiously, we see this ball come up in Nesterov and Polyak's proof of an optimal update step.

Since $v_u(h,x)$ is often non-convex, the paper [1] proposes a dual function

$$v_l(r,x) \triangleq -\frac{1}{2}\langle\Big(\nabla^2 f(x) + \frac{Mr}{2}I\Big)^{-1}\nabla f(x), \nabla f(x)\rangle - \frac{M}{12}r^3 \tag{22}$$

such that strong duality holds

$$\mathcal{D} \triangleq \{r : \nabla^2 f(x) + \frac{Mr}{2}I \succ 0, r \geq 0\} \tag{23}$$

$$\inf_{h\in\mathbb{R}^n} v_u(h,x) = \sup_{r\in D} v_l(r,x) \tag{24}$$

It is clear that any element $r \in \mathcal{D}$ must satisfy $r \geq \max\{\frac{-2\lambda_n}{M}, 0\}$. Nesterov and Polyak [1] prove that an optimal r must also satisfy

$$r = \|\Big(\nabla^2 f(x) + \frac{Mr}{2}I\Big)^{-1}\nabla f(x)\| \tag{25}$$

These are sufficient conditions to guarantee a globally optimal solution to v_u

$$\left(\nabla^2 f(x) + \frac{Mr}{2} I\right)^{-1} \nabla f(x) = \arg\min_h \left[v_u(h, x)\right] \tag{26}$$

What's interesting to note is (26) can be viewed as adaptive Levenberg-Marquardt regularization. Instead of choosing some μ that we hold constant, we select μ based on how non-convex our local quadratic approximation is. Perhaps more importantly, since r is the distance of our update step per (25) and $r \geq \frac{-\lambda_n}{M}$, $r \geq R$ where R is defined as in lemma 2. This means our update step is outside the ball defined by $\|h\| < R$ and thus in a convex region of $v_u(h, x)$.

Solving r is a well-researched problem in the area of trust region optimization, and a solution is efficient, although the details of arithmetic operations required are left out in the paper [1].

References

[1] Nesterov, Y., Polyak, B. Cubic regularization of Newton method and its global performance. Math. Program. 108, 177–205 (2006). https://doi.org/10.1007/s10107-006-0706-8