

Learning objectives. By the end of class, you should be able to

1. Explain why computing second-order critical points is an important goal for modern nonconvex optimization.
2. Derive cubic-regularized Newton from the second-order Taylor bound, and prove that one cubic-regularized step gives sufficient decrease, a small gradient, and almost-positive-semidefinite curvature.
3. Use these one-step bounds to prove convergence to approximate second-order critical points.

1 Recap: Globalizing Newton

Let f have L_2 -Lipschitz Hessian, meaning

$$\|[\nabla^2 f(x) - \nabla^2 f(y)]v\|_2 \leq L_2 \|x - y\|_2 \|v\|_2 \quad \text{for all } x, y, v \in \mathbb{R}^n.$$

Additionally, let f be twice differentiable and μ -strongly convex, such that

$$z^T \nabla^2 f(x) z \geq \mu \|z\|^2 \quad \text{for all } x, z \in \mathbb{R}^n.$$

Last time, we learned that Newton's method $x_+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$ “essentially” solves the problem in just 5–7 full steps, due to quadratic convergence

$$\|x_k - x_\star\|_2 \leq \frac{\mu}{2L_2} \|x_k - x_\star\|_2^2.$$

Question 1. How many iterations to $\|x_k - x_\star\|_2 \leq \epsilon$?

Solution: If good initial point $\|x_0 - x_\star\|_2 \leq \frac{\mu}{L_2}$, then at most

$$k \leq \left\lceil \log_2 \log_2 \left(\frac{2\mu}{L_2 \epsilon} \right) \right\rceil \text{ iterations.}$$

If bad initial point $\|x_0 - x_\star\|_2 > \frac{2\mu}{L_2}$, no guarantee.

If the initial point is not sufficiently good, however, then full-step Newton may diverge or fail. In order to allow Newton to start at any initial point, simple remedy is to backtrack.

Backtracking Newton: Compute Newton search direction

$$v = -\nabla^2 f(x)^{-1} \nabla f(x),$$

Given parameters $c_1 \in (0, \frac{1}{2})$ and $c_2 \in (0, 1)$, pick stepsize α by testing $\alpha \in \{1, c_2, c_2^2, c_2^3, \dots\}$ until the Armijo condition is satisfied

$$f(x + \alpha v) \leq f(x) + c_1 \alpha \nabla f(x)^T v,$$

and then accept $x_+ = x + \alpha v$.

Question 2. Why not just fix $\alpha = 0.01$ for example?

Solution: Quadratic convergence requires $\alpha = 1$.

Theorem 1. Let f be L_1 -Lipschitz gradient, L_2 -Lipschitz Hessian, and μ -strongly convex. Define

$$\eta = \frac{\mu^{3/2}}{L_2} \min\{1, 3(1 - 2c_1)\}, \quad \gamma = c_1 \eta^2 \min\left\{1, \frac{2c_2(1 - c_1)\mu}{L_1}\right\}.$$

Backtracking Newton converges to an iterate satisfying

$$\delta(x_k) := \sqrt{\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)} \leq \epsilon$$

in at most

$$k \leq \left\lceil \frac{f(x_0) - f(x_*)}{\gamma} + \log_2 \log_2 \left(\frac{2\eta}{\epsilon} \right) \right\rceil \quad \text{iterations.}$$

Pro: Relatively simple. Completely resolves μ -strongly convex case.

Con: Fundamentally unable to handle nonconvexity.

Question 3. Why is backtracking Newton fundamentally unable to handle nonconvexity?

Solution: Line search only shortens v . It never changes its direction.

If $\nabla^2 f(x)$ is singular, then v is undefined.

If $\nabla^2 f(x) \not\geq 0$, then v may point uphill:

$$\nabla f(x)^T v = -v^T \nabla^2 f(x) v > 0 \quad \text{is possible.}$$

Then $f(x + \alpha v) = f(x) + \alpha \nabla f(x)^T v + o(\alpha) > f(x) + c_1 \alpha \nabla f(x)^T v$ for all small $\alpha > 0$, so the Armijo test never passes.

If $\nabla f(x) = 0$ at a saddle point, then $v = 0$ and we are stuck.

Simplest remedy is to use GD globally and switch to Newton for fast convergence within strongly convex basin. Problem is that it can get stuck at spurious first-order points.

2 First- vs Second-order Optimality

Last block, we proved that gradient descent always converges to first-order critical points. The proof used the following basic fact.

Lemma 2. If f has L_1 -Lipschitz gradient, then for all $x, v \in \mathbb{R}^n$,

$$|f(x+v) - [f(x) + \nabla f(x)^T v]| \leq \frac{L_1}{2} \|v\|^2.$$

This implies the following quadratic upper-bound

$$f(y) \leq \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{L_1}{2} \|y-x\|^2}_{u_x(y)} \quad \text{for all } x, y \in \mathbb{R}^n.$$

Our proof interpreted GD as optimizing this upper-bound model

$$x_+ = \arg \min_{y \in \mathbb{R}^n} u_x(y) = x - \frac{1}{L_1} \nabla f(x).$$

and used model progress to upper-bound actual progress.

Lemma 3. Descent lemma. Let f have L_1 -Lipschitz gradient. Then $x_+ = \arg \min_y u_x(y)$ satisfies

$$f(x_+) \leq f(x) - \frac{1}{2L_1} \|\nabla f(x)\|^2.$$

Theorem 4. First-order optimality. Let f have L_1 -Lipschitz gradient. Then gradient descent

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} u_{x_k}(y)$$

attains an iterate $x_{\text{best}} \in \{x_1, x_2, \dots, x_k\}$ satisfying

$$\|\nabla f(x_{\text{best}})\| \leq \sqrt{\frac{2L_1[f(x_0) - f_{\text{inf}}]}{k}}.$$

Many real-world nonconvex functions turn out to be *benign*, meaning that

$$\nabla f(x) = 0, \quad \nabla^2 f(x) \succeq 0 \quad \Longleftrightarrow \quad f(x) = \inf_{y \in \mathbb{R}^n} f(y).$$

Examples include numerous machine learning tasks; recommendation systems, phase-based imaging, robot navigation systems, power system estimation, etc.

Unfortunately, these do tend to have spurious first-order critical points. The following is $f(x) = \|\mathcal{A}(xx^T - zz^T)\|^2$ for $n = 1$.

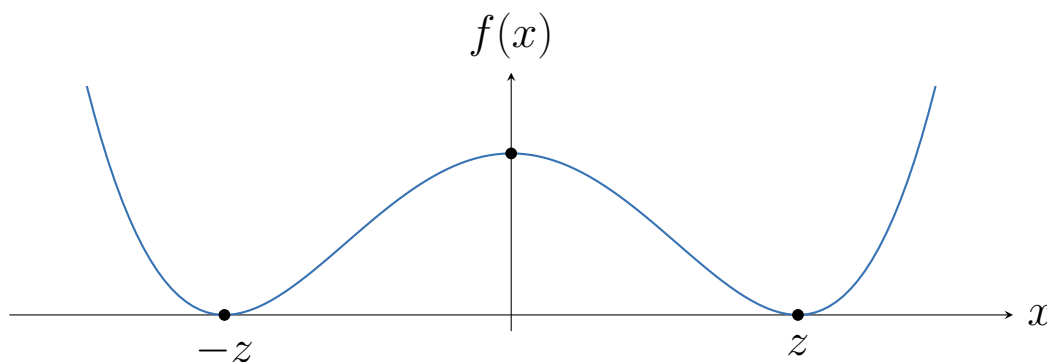
Example 5. Low-rank matrix recovery with $n = 1$. Fix $z \neq 0$ and let $f(x) = (x^2 - z^2)^2$ for $x \in \mathbb{R}$.

- (a) Find all first-order critical points.
- (b) Which are second-order critical, i.e. also satisfy $f''(x) \geq 0$?
- (c) Where does gradient descent go from $x_0 = 0$?

Solution: (a) $f'(x) = 4x(x^2 - z^2) = 0 \iff x \in \{0, z, -z\}$.

(b) $f''(x) = 12x^2 - 4z^2$. At $x = 0$: $f''(0) = -4z^2 < 0$, a local maximum. First-order but not second-order critical. At $x = \pm z$: $f''(\pm z) = 8z^2 > 0$ and $f(\pm z) = 0 = f_{\inf}$. Global minimizers.

(c) $f'(0) = 0$, so gradient descent stays at $x_0 = 0$ forever.



3 Cubic-regularized Newton

We viewed GD as minimizing a *quadratic* upper-bound of a linear model, and proved convergence to *first-order optimality*.

Lemma 6. If f has L_2 -Lipschitz Hessian, then for all $x, v \in \mathbb{R}^n$,

$$\left| f(x+v) - \left[f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \right] \right| \leq \frac{L_2}{6} \|v\|^3,$$

$$\left\| \nabla f(x+v) - [\nabla f(x) + \nabla^2 f(x) v] \right\| \leq \frac{L_2}{2} \|v\|^2,$$

Implies the following upper-bound

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \underbrace{\frac{1}{2} (y-x)^T \nabla^2 f(x) (y-x) + \frac{L_2}{6} \|y-x\|^3}_{c_x(y)} \text{ for all } x, y \in \mathbb{R}^n.$$

By analogy, minimizing a *cubic* upper-bound of a quadratic model, then the resulting method converges to *second-order optimality*

$$x_+ = \arg \min_{y \in \mathbb{R}^n} c_x(y).$$

Pro: Easy to analyze. Handles nonconvexity; converges to second-order optimality. Quadratic local convergence for μ -PL.

Con: Cubic-regularized subproblem is difficult to solve.

Lemma 7. Descent lemma. Let f have L_2 -Lipschitz Hessian.

Let $x_+ = \arg \min_y c_x(y)$ and $v = x_+ - x$. Then

$$\begin{aligned} f(x_+) &\leq f(x) - \frac{L_2}{12} \|v\|^3, & (\text{Decrease}) \\ \|\nabla f(x_+)\| &\leq L_2 \|v\|^2, & (\text{Gradient}) \\ \nabla^2 f(x_+) &\succeq -2L_2 \|v\| I. & (\text{Curvature}) \end{aligned}$$

Theorem 8. Second-order optimality. Let f have L_2 -Lipschitz Hessian. Then cubic-regularized Newton

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} c_{x_k}(y)$$

attains an iterate $x_{\text{best}} \in \{x_1, x_2, \dots, x_k\}$ satisfying

$$\|\nabla f(x_{\text{best}})\| \leq \left(\frac{12L_2^{1/2} [f(x_0) - f_{\text{inf}}]}{k} \right)^{2/3},$$

$$\lambda_{\min}(\nabla^2 f(x_{\text{best}})) \geq -2 \left(\frac{12L_2^2 [f(x_0) - f_{\text{inf}}]}{k} \right)^{1/3}.$$

Proof: Write $D = f(x_0) - f_{\text{inf}}$ and $v_i = x_{i+1} - x_i$. Summing (Decrease) over $i = 0, \dots, k-1$ gives

$$\frac{L_2}{12} \sum_{i=0}^{k-1} \|v_i\|^3 \leq \sum_{i=0}^{k-1} [f(x_i) - f(x_{i+1})] = f(x_0) - f(x_k) \leq D.$$

So the shortest step v_j satisfies

$$\|v_j\|^3 \leq \frac{1}{k} \sum_{i=0}^{k-1} \|v_i\|^3 \leq \frac{12D}{L_2 k}.$$

Take $x_{\text{best}} = x_{j+1}$. By (Gradient),

$$\|\nabla f(x_{\text{best}})\| \leq L_2 \|v_j\|^2 \leq L_2 \left(\frac{12D}{L_2 k} \right)^{2/3} = \left(\frac{12L_2^{1/2} D}{k} \right)^{2/3}.$$

By (Curvature),

$$\begin{aligned} \lambda_{\min}(\nabla^2 f(x_{\text{best}})) &\geq -2L_2 \|v_j\| \\ &\geq -2L_2 \left(\frac{12D}{L_2 k} \right)^{1/3} = -2 \left(\frac{12L_2^2 D}{k} \right)^{1/3}. \end{aligned}$$

□

3.1 Proof of Decrease

We begin by noting that the global minimum x_+ of c_x is also a second-order optimal point.

Lemma 9. Second-order Optimality.

$$\begin{aligned}\nabla c_x(x_+) &= \nabla f(x) + \nabla^2 f(x)v + \frac{L_2}{2}\|v\|v = 0, \\ \nabla^2 c_x(x_+) &= \nabla^2 f(x) + \frac{L_2}{2}\|v\| \left(I + \frac{vv^T}{v^T v} \right) \succeq 0.\end{aligned}$$

We omit the proof because it is straightforward but annoying multivariate calculus.

Lemma 10. Decrease. $f(x_+) \leq f(x) - \frac{L_2}{12}\|v\|^3.$

Proof: First, we confirm $-\nabla f(x)^T v \geq 0$. Indeed,

$$0 \leq c_x(x - v) - c_x(x + v) = -2\nabla f(x)^T v.$$

Model progress upper-bounds actual progress:

$$\begin{aligned}f(x) - f(x_+) &\geq c_x(x) - c_x(x + v) \\ &= -\nabla f(x)^T v - \frac{1}{2}v^T \nabla^2 f(x)v - \frac{L_2}{6}\|v\|^3.\end{aligned}$$

Taking inner product of first-order optimality (Lemma 9) with v gives

$$\frac{1}{2}v^T \nabla^2 f(x)v = -\frac{1}{2}\nabla f(x)^T v - \frac{L_2}{4}\|v\|^3.$$

Substituting,

$$f(x) - f(x_+) \geq -\frac{1}{2}\nabla f(x)^T v + \left(\frac{L_2}{4} - \frac{L_2}{6} \right) \|v\|^3 \geq \frac{L_2}{12}\|v\|^3.$$

□

3.2 Proof of Gradient and Curvature

Lemma 11. Gradient. $\|\nabla f(x_+)\| \leq L_2\|v\|^2$.

Proof: By first-order optimality (Lemma 9)

$$\nabla f(x) + \nabla^2 f(x)v = -\frac{L_2}{2}\|v\|v.$$

Then

$$\begin{aligned} \|\nabla f(x+v)\| &\leq \|\nabla f(x+v) - [\nabla f(x) + \nabla^2 f(x)v]\| \\ &\quad + \|\nabla f(x) + \nabla^2 f(x)v\| \\ &\leq \frac{L_2}{2}\|v\|^2 + \left\| -\frac{L_2}{2}\|v\|v \right\| = L_2\|v\|^2 \end{aligned}$$

□

Lemma 12. Curvature. $\nabla^2 f(x_+) \succeq -2L_2\|v\|I$.

Proof: By second-order optimality (Lemma 9)

$$0 \preceq \nabla^2 f(x) + \frac{L_2}{2}\|v\| \left(I + \frac{vv^T}{v^T v} \right) \preceq \nabla^2 f(x) + L_2\|v\|I.$$

Since f has L_2 -Lipschitz Hessian, for every u ,

$$\begin{aligned} u^T [\nabla^2 f(x+v) - \nabla^2 f(x)]u &\geq -\|u\| \|\nabla^2 f(x+v) - \nabla^2 f(x)\|u\| \\ &\geq -L_2\|v\|\|u\|^2. \end{aligned}$$

Combining the two bounds

$$\nabla^2 f(x+v) \succeq \nabla^2 f(x) - L_2\|v\|I \succeq -2L_2\|v\|I.$$

□

4 Summary

Second-order critical points satisfy

$$\nabla f(x) = 0, \quad \nabla^2 f(x) \succeq 0.$$

They are a natural target for nonconvex optimization because first-order critical points may still be saddles or local maxima. On a benign landscape, every second-order critical point is globally optimal.

When f has L_2 -Lipschitz Hessian, its quadratic Taylor model has cubic error. Cubic-regularized Newton therefore minimizes the global upper-bound

$$f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \frac{L_2}{6} \|v\|^3.$$

Unlike ordinary Newton, this model remains well behaved even when $\nabla^2 f(x)$ is indefinite.

If $v = x_+ - x$ is a cubic-regularized Newton step, then

$$\begin{aligned} f(x_+) &\leq f(x) - \frac{L_2}{12} \|v\|^3, \\ \|\nabla f(x_+)\| &\leq L_2 \|v\|^2, \\ \nabla^2 f(x_+) &\succeq -2L_2 \|v\| I. \end{aligned}$$

These three inequalities contain the entire convergence argument.

The decrease bound implies that the steps cannot all remain large. After k iterations, at least one step satisfies

$$\|v\|^3 \leq \frac{12[f(x_0) - f_{\inf}]}{L_2 k}.$$

For that iterate, the gradient and curvature bounds then certify approximate second-order optimality:

$$\|\nabla f(x_{\text{best}})\| = O(k^{-2/3}), \quad \lambda_{\min}(\nabla^2 f(x_{\text{best}})) \geq -O(k^{-1/3}).$$