

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

1. Motivate Newton's method through linear vs quadratic convergence, and through first- vs second-order critical points;
2. Interpret Newton's method as inverse-Hessian preconditioned gradient descent and explain why it solves strictly convex quadratics in one step;
3. Prove local quadratic convergence with full-step Newton and understand its numerous limitations.

1 Motivation

1.1 Linear vs quadratic convergence

Last block, we proved **linear convergence** by gradient descent. Recall that a differentiable function f is said to have L_1 -Lipschitz gradient (i.e. L_1 -smooth) if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_1 \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

If f is twice differentiable, then f is μ -strongly convex if

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

Theorem 1. Linear convergence. Let f be μ -strongly convex and have L_1 -Lipschitz gradient. Then gradient descent

$$x_+ = x - \frac{1}{L_1} \nabla f(x)$$

satisfies

$$\|x_+ - x_\star\|_2 \leq \left(1 - \frac{\mu}{L_1}\right) \|x - x_\star\|_2.$$

Unfortunately, convex problems are often *ill-conditioned*, with $\mu \ll L_1$. So linear convergence can still be too slow.

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

Solution: First, taking logarithm, note that

$$\log \left[\left(1 - \frac{\mu}{L_1} \right)^k \right] = k \log \left(1 - \frac{\mu}{L_1} \right) \leq -k \left(\frac{\mu}{L_1} \right).$$

After k iterations, we have

$$\begin{aligned} \|x_k - x_\star\|_2 &\leq \left(1 - \frac{\mu}{L_1} \right)^k \|x_0 - x_\star\|_2 \\ &\leq \exp \left(-k \frac{\mu}{L_1} \right) \|x_0 - x_\star\|_2 \end{aligned}$$

Therefore,

$$k \geq \frac{L_1}{\mu} \log \left(\frac{\|x_0 - x_\star\|_2}{\epsilon} \right) \implies \|x_k - x_\star\|_2 \leq \epsilon.$$

Today, we prove **quadratic convergence** by Newton's method. By analogy, a twice differentiable function is said to have L_2 -*Lipschitz Hessian* if

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

The size of L_2 controls the accuracy of the second-order Taylor expansion. For example, if $L_2 = 0$, then f is exactly quadratic.

Theorem 2. Quadratic convergence. Let f be μ -strongly convex and have L_2 -Lipschitz Hessian. Then full-step Newton's method

$$x_+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$$

satisfies

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

Question 2. In general, if initial point satisfies

$$\|x_0 - x_\star\| \leq \frac{\mu}{L_2},$$

how many iterations for $\|x_k - x_\star\|_2 \leq \epsilon$?

Solution: After one iteration, we have

$$\frac{L_2}{2\mu} \|x_+ - x_\star\|_2 \leq \left(\frac{L_2}{2\mu} \|x - x_\star\|_2 \right)^2.$$

After k iterations, it is then

$$\frac{L_2}{2\mu} \|x_k - x_\star\| \leq \left(\frac{L_2}{2\mu} \|x_0 - x_\star\| \right)^{2^k} \leq \left(\frac{1}{2} \right)^{2^k}.$$

Therefore

$$k \geq \log_2 \log_2 \left(\frac{2\mu}{L_2 \epsilon} \right) \implies \|x_k - x_\star\| \leq \epsilon.$$

Question 3. If f is exactly quadratic, how many iterations for $\|x_k - x_\star\|_2 \leq \epsilon$?

Solution: Since $L_2 = 0$, need just 1 iteration starting from anywhere.

1.2 First vs second-order critical points

Last block, we proved that gradient descent always converges to **first-order** critical points. The proof used the fact that L_1 -Lipschitz gradient implies the following quadratic upper-bound

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

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

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

Theorem 3. 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

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

But nonconvex problems often have *first-order saddle points*, i.e. non-global first-order points.

Question 4. How many iterations for $\|\nabla f(x_{\text{best}})\| \leq \epsilon$?

Solution: Solve directly

$$k \geq \frac{2L [f(x_0) - f_{\text{inf}}]}{\epsilon^2} = \frac{\text{const}}{\epsilon^2}.$$

This week, we will prove that *variants* of Newton's method always converge to **second-order** critical points. By analogy, L_2 -Lipschitz Hessian implies the following *cubic* upper-bound

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

The *cubic regularized* Newton's method generates its iterates by optimizing this upper-bound model.

Theorem 4. 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

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

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

Able to escape saddle points even with $\nabla f(x) = 0$, so long as $\nabla^2 f(x) \not\preceq 0$.

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.

Question 5. If L_2 -Lipschitz Hessian, how many iterations for

$$\|\nabla f(x_{\text{best}})\| \leq \epsilon_1 \text{ and } \nabla^2 f(x_{\text{best}}) \succeq -\epsilon_2 I?$$

Solution: Solve directly

$$\begin{aligned} k &\geq \max \left\{ \frac{12L_2^{1/2} [f(x_0) - f_{\text{inf}}]}{\epsilon_1^{3/2}}, \frac{12L_2^2 [f(x_0) - f_{\text{inf}}]}{(\frac{1}{2}\epsilon_2)^3} \right\} \\ &\geq \frac{\text{const}}{\epsilon_1^{3/2}} + \frac{\text{const}}{\epsilon_2^3}. \end{aligned}$$

1.3 Why so many Newton-type methods?

Newton-type methods include Cubic-regularized, trust-region, truncated CG, BFGS, L-BFGS, AdaGrad, Shampoo, Muon. Muon has recently been demonstrated at frontier language-model scale.

Question 6. What's wrong with naive Newton's method

$$x_+ = x - \alpha \nabla^2 f(x)^{-1} \nabla f(x)?$$

Solution: It is really expensive and really badly behaved.

- For quadratic convergence, need $\alpha = 1$; this is extremely aggressive.
- If $\nabla^2 f(x)$ is singular, then method fails.
- If $\nabla^2 f(x) \not\preceq 0$ is nonsingular, method doesn't point downhill; climbs uphill to converge to saddle point. Consider

$$f(x, y) = \frac{1}{2}x^2 - \frac{1}{2}y^2 = \frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Gradient descent correctly yields $(x, y) \rightarrow (0, \pm\infty)$

$$\begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \alpha \left(\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} (1 - \alpha)x \\ (1 + \alpha)y \end{bmatrix},$$

While Newton's method converges to saddle $(x, y) \rightarrow (0, 0)$

$$\begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \alpha \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) = (1 - \alpha) \begin{bmatrix} x \\ y \end{bmatrix}$$

Not only does not escape saddle, but *converges to it*.

- Hessian expensive to evaluate / form; need to solve large dense linear equation.

The main remedies:

1. **Backtracking line search.** Set $v = -\nabla^2 f(x)^{-1} \nabla f(x)$ and pick $\alpha \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots\}$ until the following holds

$$f(x + \alpha v) \leq f(x) + 10^{-3} \cdot \alpha \nabla f(x)^T v.$$

Pro: Easy to implement. Con: Fail if $\nabla^2 f(x) \not\succeq 0$; cannot escape saddle. Maybe probe is still too expensive.

2. **Trust-region Newton's method.** Solve subproblem at each iteration

$$\begin{aligned} \text{minimize} \quad & f(x) + \nabla f(x)^T (x_+ - x) \\ & + \frac{1}{2} (x_+ - x)^T \nabla^2 f(x) (x_+ - x) \\ \text{subject to} \quad & \|x_+ - x\| \leq \Delta \end{aligned}$$

Pro: Robust for any $\nabla^2 f(x)$ and escapes saddle. Con: Must solve subproblem; must adjust radius.

3. **Cubic-regularized Newton's method.** Solve subproblem at each iteration

$$\begin{aligned} \text{minimize} \quad & f(x) + \nabla f(x)^T (x_+ - x) \\ & + \frac{1}{2} (x_+ - x)^T \nabla^2 f(x) (x_+ - x) + \frac{M}{6} \|x_+ - x\|^3. \end{aligned}$$

Pro: Robust for any $\nabla^2 f(x)$ and escapes saddle; better guarantees; easy to analyze. Con: Must solve subproblem. In practice, slightly more iterations than trust-region.

4. **Quasi-Newton methods.** Approximate the true Hessian by $B \approx \nabla^2 f(x)$ and then proceed with one of the methods above. Examples: BFGS, L-BFGS, Shampoo, Muon, etc. *Heavy overlap with the idea of preconditioning.*

2 Newton as Preconditioned GD

The homework considered *preconditioned* gradient descent

$$x_+ = x - \alpha B \nabla f(x) \quad \text{where } B = B^T \succ 0.$$

The analysis used B -norms for gradients and displacements:

$$\begin{aligned} \|\nabla f(x)\|_B^2 &= \nabla f(x)^T B \nabla f(x), \\ \|x - x_\star\|_{B^{-1}}^2 &= (x - x_\star)^T B^{-1} (x - x_\star). \end{aligned}$$

Using the B -norm versions of L -smoothness and μ -PL inequality

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_{B^{-1}}^2, \\ \frac{1}{2} \|\nabla f(x)\|_B^2 &\geq \mu [f(x) - f_{\inf}], \end{aligned}$$

we obtained analogs of classic linear convergence ($\alpha = \frac{1}{L}$)

$$f(x_k) - f_{\inf} \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_{\inf}].$$

Proof: Plug $x_+ = x - \frac{1}{L} B \nabla f(x)$ into L -smoothness yields

$$\begin{aligned} f(x_+) &\leq f(x) + \left\langle \nabla f(x), -\frac{1}{L} B \nabla f(x) \right\rangle + \frac{L}{2} \left\| -\frac{1}{L} B \nabla f(x) \right\|_{B^{-1}}^2 \\ &= f(x) - \frac{1}{2L} \|\nabla f(x)\|_B^2. \end{aligned}$$

Now subtract $f_{\inf}$ and plug in μ -PL

$$f(x_+) - f_{\inf} \leq \left(1 - \frac{\mu}{L}\right) [f(x) - f_{\inf}].$$

Repeat over k iterations. □

The quantity $\kappa = L/\mu$ is commonly known as the *condition number*. Roughly, GD converges to ϵ -accuracy in $O(\kappa \log(1/\epsilon))$ iters.

Question 7. What is the point of the preconditioner B ?

Solution: It can significantly reduce the number κ , which in turn, can significantly reduce the number of iterations.

Example 5. Consider the quadratic objective and starting point

$$f(x) = \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x = \begin{bmatrix} 9 \\ 1 \end{bmatrix}.$$

(a) For regular GD, $x_+ = x - \alpha \nabla f(x)$, find the stepsize α that minimizes $f(x_+)$.

(b) With preconditioned GD, $x_+ = x - \alpha B \nabla f(x)$, find a stepsize α and preconditioner B that minimize $f(x_+)$.

Solution: We have $\nabla f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} 9 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 9 \end{bmatrix}$.

(a) Ordinary gradient descent gives

$$x_+ = \begin{bmatrix} 9 \\ 1 \end{bmatrix} - \alpha \begin{bmatrix} 9 \\ 9 \end{bmatrix} = \begin{bmatrix} 9 - 9\alpha \\ 1 - 9\alpha \end{bmatrix}.$$

Therefore $f(x_+) = \frac{1}{2}(9 - 9\alpha)^2 + \frac{9}{2}(1 - 9\alpha)^2$. Differentiating with respect to α gives $f'(x_+) = -162 + 810\alpha$, so the minimizing stepsize is $\alpha = \frac{1}{5}$.

(b) Choose $B = \begin{bmatrix} 1 & 0 \\ 0 & 1/9 \end{bmatrix}$ and $\alpha = 1$. Then

$$B \nabla f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1/9 \end{bmatrix} \begin{bmatrix} 9 \\ 9 \end{bmatrix} = \begin{bmatrix} 9 \\ 1 \end{bmatrix},$$

so $x_+ = x - B \nabla f(x) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = x_*$. Hence $f(x_+) = f_{\min}$.

Question 8. What is the *best* preconditioner B ?

Solution: For strictly convex f , i.e. $\nabla^2 f(x) \succ 0$ for all x , the best preconditioner is $B = \nabla^2 f(x)^{-1}$. (For nonconvex f , not super obvious.)

Example 6. Consider the general quadratic objective

$$f(x) = c + g^T x + \frac{1}{2} x^T H x$$

where we assume $H \succ 0$.

- (a) For regular GD, $x_+ = x - \alpha \nabla f(x)$, find the stepsize α that minimizes $f(x_+)$.
- (b) With preconditioned GD, $x_+ = x - \alpha B \nabla f(x)$, find a stepsize α and preconditioner B that minimize $f(x_+)$.

Solution: (a) Since $\nabla f(x) = g + Hx$ and $\nabla^2 f(x) = H$, ordinary gradient descent gives $x_+ = x - \alpha \nabla f(x)$. Expanding the quadratic along this direction,

$$f(x_+) = f(x) - \alpha \|\nabla f(x)\|_2^2 + \frac{\alpha^2}{2} \nabla f(x)^T \nabla^2 f(x) \nabla f(x).$$

Differentiating with respect to α and setting the derivative equal to zero gives the exact line search

$$\alpha = \frac{\|\nabla f(x)\|_2^2}{\nabla f(x)^T \nabla^2 f(x) \nabla f(x)}.$$

(b) The minimizer satisfies $\nabla f(x_*) = g + Hx_* = 0$, so $x_* = -H^{-1}g$. Choose $B = H^{-1}$ and $\alpha = 1$. Then

$$x_+ = x - H^{-1}(g + Hx) = -H^{-1}g = x_*.$$

Hence the inverse Hessian preconditioner reaches the minimizer of any strictly convex quadratic in one step.

3 Full-step Newton's method

So Newton converges in 1 full step for exact quadratics. What happens if f is not exactly quadratic? Under suitable assumption, our convergence rate is so fast that we “essentially” solve in just 5-7 full steps.

Recall that a twice differentiable function is said to have L_2 -Lipschitz Hessian if

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

If f is twice differentiable, then f is μ -strongly convex if

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

We now prove the advertised quadratic convergence theorem.

Theorem 7. Let f have L_2 -Lipschitz Hessian and be μ -strongly convex. Then the full-step Newton iteration

$$x_+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$$

satisfies

$$\|x_+ - x_\star\| \leq \frac{L_2}{2\mu} \|x - x_\star\|^2$$

Proof: Write $e_+ = x_+ - x_\star$ and $e = x - x_\star$. By fundamental theorem of calculus

$$\nabla f(x) - \underbrace{\nabla f(x_\star)}_{=0} = \int_0^1 \nabla^2 f(x_\star + te) e \, dt.$$

Therefore,

$$\begin{aligned} e_+ &= e - \nabla^2 f(x)^{-1} \nabla f(x) \\ \nabla^2 f(x) e_+ &= \nabla^2 f(x) e - \nabla f(x) \\ &= \int_0^1 [\nabla^2 f(x) - \nabla^2 f(x_\star + te)] e \, dt \end{aligned}$$

So on one hand,

$$\begin{aligned}\|\nabla^2 f(x)e_+\| &= \sqrt{e_+^T [\nabla^2 f(x)]^2 e_+} \\ &\geq \sqrt{\lambda_{\min}([\nabla^2 f(x)]^2) e_+^T e_+} \geq \mu \|e_+\| \end{aligned}$$

on the other hand,

$$\begin{aligned}\|\nabla^2 f(x)e_+\| &\leq \int_0^1 \|[\nabla^2 f(x) - \nabla^2 f(x_\star + te)]e\| \, dt \\ &\leq \int_0^1 L_2 \|x - (x_\star + te)\| \|e\| \, dt \\ &= L_2 \|e\|^2 \int_0^1 (1-t) \, dt = \frac{L_2}{2} \|e\|^2 \end{aligned}$$

Combining the two bounds gives

$$\|e_+\| \leq \frac{L_2}{2\mu} \|e\|^2.$$

□

Unfortunately, full-step Newton can fail outside its convergence radius, even on very simple strictly convex problems.

Example 8. Consider the strictly convex function

$$f(x) = x - \log x, \quad x > 0.$$

- (a) Compute the unique minimizer.
- (b) Perform a full-step Newton at $x_0 = 3$.

Solution: (a) We have $f'(x) = 1 - \frac{1}{x}$ and $f''(x) = \frac{1}{x^2} > 0$, so the unique minimizer is $x_\star = 1$.

(b) Full-step Newton gives

$$x_+ = x - \frac{f'(x)}{f''(x)} = 2x - x^2.$$

Starting from $x_0 = 3$ gives

$$x_1 = 2(3) - 3^2 = -3,$$

which lies outside the domain of f .

4 Preview: Backtracking Newton

To ensure convergence from any initial point, we perform backtracking line search.

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 $x_+ = x + \alpha v$.

Classical analysis is based on the Newton decrement

$$\begin{aligned} \delta(x) &:= \sqrt{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)} \\ &= \|\nabla f(x)\|_B = \|v\|_{B^{-1}}. \end{aligned}$$

Intuitively, the method has two phases:

1. (Fixed decrement) If $\delta(x) > \eta$, each Newton step decreases the objective by some fixed amount $\gamma > 0$.
2. (Quadratic convergence) Once $\delta(x) \leq \eta$, backtracking accepts $\alpha = 1$ forever and achieves quadratic convergence like $\delta(x_+) \leq \frac{1}{2\eta} \delta(x)^2$.

Question 9. How many total iterations?

Solution: Since $f(x) \geq f(x_*)$, the first phase can occur only finitely many times. Eventually, $\delta(x) \leq \eta$, and convergence rate becomes quadratic. Combined, the number of iterations is

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

5 Summary

Newton's method can be viewed as gradient descent with the inverse Hessian as a preconditioner:

$$x_+ = x - \nabla^2 f(x)^{-1} \nabla f(x).$$

For a strictly convex quadratic, the Hessian is constant and one full Newton step reaches the global minimizer exactly.

More generally, suppose that f is μ -strongly convex and has L_2 -Lipschitz Hessian. Then full-step Newton satisfies

$$\|x_+ - x_\star\| \leq \frac{L_2}{2\mu} \|x - x_\star\|^2.$$

Therefore, once $\|x - x_\star\| \leq \frac{\mu}{L_2}$, the error essentially squares at every iteration. This is *quadratic convergence*, requiring only

$$O(\log \log(1/\epsilon))$$

iterations to reach ϵ -accuracy.

The difficulty is that this guarantee is only local. Far from the solution, a full Newton step may diverge or even leave the domain. For nonconvex problems, $\nabla^2 f(x)$ may also be singular or indefinite, so the Newton direction may be undefined or fail to be a descent direction.

These limitations motivate globalized Newton methods: backtracking line search, trust-region methods, and cubic regularization.