Newton's method

Cong Ma

University of Chicago, Winter 2026

Second-Order Methods

- Up to now, we focused on *first-order methods*, i.e., algorithms that rely on gradient information.
- Starting today, we turn to *second-order methods*, which use both the gradient and the *Hessian* (the matrix of second derivatives).
- For this discussion, we will restrict attention to unconstrained problems of the form

$$\min_x \; f(x) \qquad \text{with} \qquad x \in \mathbb{R}^n,$$

- where $f(x)$ is assumed to be *twice differentiable*.

Second-Order Taylor Approximation

At an iterate x^t, a quadratic approximation of f is

$$f(x) \approx f(x^t) + \langle \nabla f(x^t),\, x - x^t \rangle + \frac{1}{2} \langle x - x^t,\, \nabla^2 f(x^t)(x - x^t) \rangle.$$

- Compared to a linear model, the quadratic model can be a much more faithful local approximation.
- Minimizing the quadratic model yields a closed-form step (when $\nabla^2 f(x^t)$ is invertible).

Newton Direction from the Quadratic Model

Differentiate the quadratic model w.r.t. x and set to zero:

$$\nabla f(x^t) + \nabla^2 f(x^t)(x - x^t) = 0 \quad \Longrightarrow \quad x = x^t - [\nabla^2 f(x^t)]^{-1} \nabla f(x^t).$$

Descent directions

Gradient direction: $d_t = -\nabla f(x^t)$, Newton direction: $d_t = -[\nabla^2 f(x^t)$

Newton as Preconditioning

- The Newton direction multiplies the negative gradient by an inverse Hessian:

$$d_t = -\underbrace{[\nabla^2 f(x^t)]^{-1}}_{\text{preconditioner}} \nabla f(x^t).$$

- This is an example of *preconditioning*: reshaping the geometry of the problem via curvature.
- One important consequence is *affine invariance*: under a reparametrization $x = Ay + b$, the Newton direction transforms consistently with the change of coordinates.

Affine Invariance of Newton's Method (Statement)

Consider an invertible affine change of variables

$$x = Ay + b, \qquad A \in \mathbb{R}^{n\times n} \text{ invertible, } b \in \mathbb{R}^n,$$

and define the reparametrized objective

$$g(y) := f(Ay + b).$$

Claim (affine invariance)

Let $x^t = Ay^t + b$. If Newton's method is applied to f at x^t and to g at y^t, then the iterates match under the change of variables:

$$x^{t+1} = Ay^{t+1} + b.$$

Equivalently, the Newton step in x-coordinates equals A times the Newton step in y-coordinates.

Affine Invariance (Proof): How Gradients and Hessians Transform

Let $x = Ay + b$ and $g(y) = f(x)$ with x depending on y.

Gradient transformation (chain rule)

$$\nabla g(y) = A^\top \nabla f(Ay + b).$$

In particular, at y^t with $x^t = Ay^t + b$,

$$\nabla g(y^t) = A^\top \nabla f(x^t).$$

Hessian transformation

Differentiate again:

$$\nabla^2 g(y) = A^\top \nabla^2 f(Ay + b)\, A,$$

hence

$$\nabla^2 g(y^t) = A^\top \nabla^2 f(x^t)\, A.$$

Affine Invariance (Proof): Newton Steps Match

Newton directions:

$$d_x^t := -[\nabla^2 f(x^t)]^{-1}\nabla f(x^t), \qquad d_y^t := -[\nabla^2 g(y^t)]^{-1}\nabla g(y^t).$$

Compute the Newton direction in y-space

Using $\nabla g(y^t) = A^\top \nabla f(x^t)$ and $\nabla^2 g(y^t) = A^\top \nabla^2 f(x^t) A$,

$$\begin{aligned} d_y^t &= -\left(A^\top \nabla^2 f(x^t) A\right)^{-1} A^\top \nabla f(x^t) \\ &= -A^{-1}\left[\nabla^2 f(x^t)\right]^{-1}\left(A^\top\right)^{-1} A^\top \nabla f(x^t) \\ &= -A^{-1}\left[\nabla^2 f(x^t)\right]^{-1}\nabla f(x^t) = A^{-1} d_x^t. \end{aligned}$$

Newton updates are

$$y^{t+1} = y^t + d_y^t, \qquad x^{t+1} = x^t + d_x^t.$$

Multiply the y-update by A and add b:

$$Ay^{t+1} + b = Ay^t + b + Ad_y^t = x^t + d_x^t = x^{t+1}.$$

Affine Invariance: What Changes (and What Doesn't)

- The Newton step is *coordinate-consistent*: reparameterizing by $x = Ay + b$ produces the same sequence of points in $\mathbb{R}^n$.
- In contrast, (plain) gradient descent is *not* affine invariant: under $x = Ay + b$, the gradient direction transforms as $-\nabla g(y) = -A^\top \nabla f(x)$, which is not generally mapped to $-\nabla f(x)$ by a fixed linear map unless A is orthogonal (up to scaling).

Damped Newton and Newton's Method

Damped Newton update

$$x^{t+1} = x^t - \eta \left[\nabla^2 f(x^t)\right]^{-1} \nabla f(x^t).$$

- $\eta > 0$ is a stepsize (damping).
- The choice $\eta = 1$ gives the classical *Newton's method*.

Failure Mode: Lack of Curvature

- Newton steps can be unstable when the Hessian is singular or nearly singular.
- Near-zero curvature can lead to extremely large steps (via $[\nabla^2 f(x^t)]^{-1}$), and the local quadratic model may be unreliable.

Illustrative example (1D)

$$f(x) = \log(e^{2x} + e^{-2x}).$$

Newton's method can converge very fast from some initial points, but diverge from others that are relatively close.

Example: When Newton Can Diverge

Consider the 1D function

$$f(x) = \log(e^{2x} + e^{-2x}).$$

- Smooth, convex, symmetric, minimized at $x^\star = 0$.
- Yet Newton's method can behave very differently for nearby initial points.

Compute Gradient and Hessian

Write $f(x) = \log(e^{2x} + e^{-2x})$.

$$f'(x) = \frac{2e^{2x} - 2e^{-2x}}{e^{2x} + e^{-2x}} = 2 \cdot \frac{e^{4x} - 1}{e^{4x} + 1}.$$

$$f''(x) = \frac{16e^{4x}}{(e^{4x} + 1)^2}.$$

- $f'(0) = 0$ and $f''(0) = 4$.
- For large x, $f'(x) \to 2$ while $f''(x) \to 0$ (curvature vanishes).

Newton vs Gradient Descent Updates (1D)

Newton's method:

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}.$$

Gradient descent with stepsize η:

$$x_{t+1} = x_t - \eta f'(x_t).$$

- Newton rescales the gradient by $1/f''(x_t)$.
- When $f''(x_t)$ is tiny, the Newton step can become enormous.

Run 1: Start at $x_0 = 0.5$ (Newton is Extremely Fast)

Newton iterates (first few):

$$0.5000, \quad -0.4067, \quad 0.2047, \quad -0.0237, \quad 3.53 \times 10^{-5}, \ \ldots \ \to 0.$$

Gradient descent with $\eta = 0.1$ (first few):

$$0.5000, \quad 0.3477, \quad 0.2274, \quad 0.1422, \quad 0.0868, \ \ldots \ \to 0.$$

- Newton quickly enters a neighborhood where the quadratic model is accurate, leading to very rapid (locally quadratic) convergence.
- GD decreases steadily but much more slowly.

Run 2: Start at $x_0 = 0.7$ (Newton Blows Up)

Newton iterates (first few):

$$0.7000, \quad -1.3480, \quad 26.1045, \quad -2.79 \times 10^{44}, \quad \text{diverged.}$$

Gradient descent with $\eta = 0.1$ (first few):

$$0.7000, \quad 0.5229, \quad 0.3669, \quad 0.2418, \quad 0.1520, \ \ldots \ \to 0.$$

- A small change in initialization ($0.5 \to 0.7$) radically changes Newton's behavior.
- GD remains stable because its step length is directly controlled by η.

Why Does Newton Diverge Here? (Mechanism)

Key observation: for large $|x|$,

$$f'(x) \approx 2\,\mathrm{sign}(x), \qquad f''(x) = \frac{16e^{4x}}{(e^{4x}+1)^2} \approx 16e^{-4|x|} \ \ \text{(tiny)}.$$

So the Newton step size behaves like

$$\left|\frac{f'(x)}{f''(x)}\right| \approx \frac{2}{16e^{-4|x|}} = \frac{1}{8}e^{4|x|},$$

which grows explosively once an iterate lands in a low-curvature region.

- Newton is powerful near the minimizer (good curvature).
- Newton can be dangerous far away (near-flat curvature $\Rightarrow$ huge steps).

Takeaway: Stabilizing Newton

This example motivates damping / safeguards:

- **Damped Newton:** $x_{t+1} = x_t - \eta \, \frac{f'(x_t)}{f''(x_t)}$ with $\eta \in (0, 1]$.
- **Line search / trust region:** accept a Newton-like step only if it decreases f sufficiently.
- **Regularization:** replace $f''(x_t)$ by $f''(x_t) + \lambda$ (or use cubic regularization).

Key Lemma: Error Recursion for Damped Newton

Damped Newton:

$$x^{t+1} = x^t - \eta\,[\nabla^2 f(x^t)]^{-1}\nabla f(x^t), \qquad \eta > 0,$$

and let $x^\star$ be a local minimizer with $\nabla f(x^\star) = 0$, and assume $\nabla^2 f(x^t)$ invertible.

Lemma 1

Define

$$H_t \;:=\; [\nabla^2 f(x^t)]^{-1} \int_0^1 \nabla^2 f(x^\star + \lambda(x^t - x^\star))\, d\lambda.$$

Then

$$x^{t+1} - x^\star \;=\; (I - \eta H_t)(x^t - x^\star).$$

Proof of the Newton Lemma (Simplified)

Using $\nabla f(x^\star) = 0$, the fundamental theorem of calculus gives

$$\nabla f(x^t) = \int_0^1 \nabla^2 f\big(x^\star + \lambda(x^t - x^\star)\big)\,(x^t - x^\star)\,d\lambda.$$

Multiply both sides by $[\nabla^2 f(x^t)]^{-1}$ and define

$$H_t := [\nabla^2 f(x^t)]^{-1} \int_0^1 \nabla^2 f\big(x^\star + \lambda(x^t - x^\star)\big)\,d\lambda,$$

so that

$$[\nabla^2 f(x^t)]^{-1}\nabla f(x^t) = H_t(x^t - x^\star).$$

Plug into damped Newton:

$$x^{t+1} - x^\star = (x^t - x^\star) - \eta[\nabla^2 f(x^t)]^{-1}\nabla f(x^t) = (I - \eta H_t)(x^t - x^\star).$$

Setup: Strong Curvature + Lipschitz Hessian

Let $x^\star$ be a local minimum of f with *strong curvature*:

$$\nabla f(x^\star) = 0, \qquad \nabla^2 f(x^\star) \succeq \mu I \quad (\mu > 0). \tag{2}$$

Assume the Hessian is M-*Lipschitz* (in spectral norm):

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_s \leq M\|x - y\|_2. \tag{3}$$

High-level idea: near $x^\star$, the Hessian stays well-conditioned and cannot change too fast, so $I - H_t$ becomes small. This will imply a strong local contraction for Newton.

Theorem (L17.2): Bounding $\|I - H_t\|_s$ Locally

Recall

$$H_t := [\nabla^2 f(x_t)]^{-1} \int_0^1 \nabla^2 f(x^\star + \lambda(x_t - x^\star))\, d\lambda.$$

Lemma 2

If (2) and (3) hold, then whenever

$$\|x_t - x^\star\|_2 \leq \frac{\mu}{2M},$$

we have the bound

$$\|I - H_t\|_s \leq \frac{M}{\mu} \|x_t - x^\star\|_2.$$

Interpretation: once x_t is close enough to $x^\star$, H_t is close to I, and the Newton map is strongly contractive.

Local Analysis I: Bounding $\|I - H_t\|_s$ via Lipschitz Hessian

Assume near $x^\star$:

$$\nabla^2 f(x^\star) \succeq \mu I, \qquad \|\nabla^2 f(x) - \nabla^2 f(y)\|_s \leq M\|x - y\|_2.$$

Goal (for x^t close enough to $x^\star$):

$$\|I - H_t\|_s \;\leq\; \frac{M}{\mu}\,\|x^t - x^\star\|_2.$$

This estimate plus the recursion implies very fast local contraction.

Local Analysis II: Proof Sketch of $\|I - H_t\|_s$ Bound

Rewrite $I - H_t$ to expose a Hessian difference:

$$I - H_t = [\nabla^2 f(x^t)]^{-1} \int_0^1 \Big(\nabla^2 f(x^t) - \nabla^2 f(x^\star + \lambda(x^t - x^\star)) \Big)\, d\lambda.$$

Take spectral norms:

$$\|I - H_t\|_s \le \|[\nabla^2 f(x^t)]^{-1}\|_s \int_0^1 \|\nabla^2 f(x^t) - \nabla^2 f(x^\star + \lambda(x^t - x^\star))\|_s \, d\lambda.$$

Use Lipschitzness and $\|x^t - (x^\star + \lambda(x^t - x^\star))\|_2 = (1-\lambda)\|x^t - x^\star\|_2$:

$$\int_0^1 \cdots d\lambda \le \int_0^1 M(1-\lambda)\, d\lambda \, \|x^t - x^\star\|_2 = \frac{M}{2}\|x^t - x^\star\|_2.$$

If $\|x^t - x^\star\|_2 \le \mu/(2M)$, then

$$\nabla^2 f(x^t) \succeq \nabla^2 f(x^\star) - M\|x^t - x^\star\|_2 I \succeq \frac{\mu}{2} I \quad \Rightarrow \quad \|[\nabla^2 f(x^t)]^{-1}\|_s \le \frac{2}{\mu}.$$

Combine the last two displays to get

$$\|I - H_t\|_s \le \frac{M}{\mu}\|x^t - x^\star\|_2.$$

Quadratic Convergence of Newton's Method

Theorem 3 (Local quadratic convergence)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with M-Lipschitz continuous Hessian, and let $x^\star$ be a local minimum of f with strong curvature, i.e.,

$$\nabla f(x^\star) = 0, \qquad \nabla^2 f(x^\star) \succeq \mu I$$

for some $\mu > 0$. Then, as long as we start Newton's method from a point x_0 with

$$\|x_0 - x^\star\|_2 \leq \frac{\mu}{2M},$$

the distance to optimality of the iterates x_t generated by Newton's method decays as

$$\frac{\|x_{t+1} - x^\star\|_2}{\mu/M} \leq \left(\frac{\|x_t - x^\star\|_2}{\mu/M} \right)^2.$$

Proof

From the Newton lemma with $\eta = 1$:

$$\|x^{t+1} - x^\star\|_2 \leq \|I - H_t\|_s \, \|x^t - x^\star\|_2.$$

Using $\|I - H_t\|_s \leq \frac{M}{\mu}\|x^t - x^\star\|_2$ (valid when $\|x^t - x^\star\|_2 \leq \mu/(2M)$),

$$\|x^{t+1} - x^\star\|_2 \;\leq\; \frac{M}{\mu}\,\|x^t - x^\star\|_2^2.$$

Equivalently,

$$\frac{M}{\mu}\|x^{t+1} - x^\star\|_2 \;\leq\; \Big(\frac{M}{\mu}\|x^t - x^\star\|_2\Big)^2.$$

Global Linear Convergence of Damped Newton

Theorem 4 (Corollary L17.1)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, μ-strongly convex, and L-smooth. Then, the distance to optimality of the iterates x_t generated by damped Newton's method with stepsize $\eta \leq \mu/L$ decays exponentially fast at the rates

$$\|x_{t+1} - x^\star\|_2^2 \leq \frac{L}{\mu}\left(1 - \eta\frac{\mu}{L}\right)^t \|x_1 - x^\star\|_2^2$$

and

$$f(x_{t+1}) - f(x^\star) \leq \left(1 - \eta\frac{\mu}{L}\right)^t (f(x_1) - f(x^\star)).$$

Proof of Corollary L17.1: One-Step Decrease

L-smoothness gives, for any step,

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|_2^2.$$

Using $\nabla^2 f(x_t) \succeq \mu I$, we have

$$\|x_{t+1} - x_t\|_2^2 \leq \frac{1}{\mu} \langle x_{t+1} - x_t, \nabla^2 f(x_t)(x_{t+1} - x_t) \rangle.$$

Plug in and use the damped Newton step

$$x_{t+1} - x_t = -\eta \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t),$$

to obtain

$$f(x_{t+1}) \leq f(x_t) - \eta \left\langle \nabla f(x_t), \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t) \right\rangle + \frac{L}{2\mu} \eta^2 \left\langle \nabla f(x_t), \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t) \right\rangle$$

If $\eta \leq \mu/L$, then

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} \left\langle \nabla f(x_t), \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t) \right\rangle. \tag{5}$$

Proof of Corollary L17.1: Lower Bound via Strong Convexity

μ-strong convexity gives, for all $y \in \mathbb{R}^n$,

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{\mu}{2} \|y - x_t\|_2^2.$$

Using $\nabla^2 f(x_t) \preceq LI$ (i.e., L-smoothness), we have

$$\|y - x_t\|_2^2 \geq \frac{1}{L} \langle y - x_t, \nabla^2 f(x_t)(y - x_t) \rangle,$$

hence

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{\mu}{2L} \langle y - x_t, \nabla^2 f(x_t)(y - x_t) \rangle.$$

Minimizing the right-hand side over y yields the minimizer

$$y^\star = x_t - \frac{L}{\mu} [\nabla^2 f(x_t)]^{-1} \nabla f(x_t),$$

and therefore

$$f(x^\star) \geq f(x_t) - \frac{L}{2\mu} \left\langle \nabla f(x_t), [\nabla^2 f(x_t)]^{-1} \nabla f(x_t) \right\rangle,$$

Proof of Corollary L17.1: Linear Rate (Function Values and Distance)

Combine (5) and (6):

$$f(x_{t+1}) \leq f(x_t) - \eta \frac{\mu}{L}(f(x_t) - f(x^\star)),$$

so

$$f(x_{t+1}) - f(x^\star) \leq \left(1 - \eta \frac{\mu}{L}\right)(f(x_t) - f(x^\star)) \leq \left(1 - \eta \frac{\mu}{L}\right)^t (f(x_1) - f(x^\star)).$$

Using strong convexity and smoothness:

$$f(x_{t+1}) - f(x^\star) \geq \frac{\mu}{2}\|x_{t+1} - x^\star\|_2^2, \qquad f(x_1) - f(x^\star) \leq \frac{L}{2}\|x_1 - x^\star\|_2^2,$$

we obtain

$$\|x_{t+1} - x^\star\|_2^2 \leq \frac{L}{\mu}\left(1 - \eta \frac{\mu}{L}\right)^t \|x_1 - x^\star\|_2^2.$$