Lecture 14: Newton and Hessian Geometry

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From First Order to Second Order

First-order recap (Lectures 7–13). Local model = linear, with quadratic upper bound:

$$f(y) \;\le\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L}{2}\|y - x\|^2.$$

One global constant L; same step size everywhere.

Lecture 14: Newton's method. Local model = the actual quadratic Taylor expansion:

$$m_x(d) = f(x) + \langle \nabla f(x), d \rangle + \tfrac{1}{2}\langle \nabla^2 f(x) d, d \rangle.$$

The Newton step is the *exact minimizer* of m_x.

Trade-offs.

- **Cost.** Each step needs the Hessian (or Hessian-vector products) and a linear solve.
- **Gain.** Affine-invariant; *quadratic* local convergence; one step exact on quadratics.

Hessian Geometry: A Metric That Moves

Definition 14.1 (Hessian local norms). Let f be twice differentiable at x with $\nabla^2 f(x) \succ 0$.

- **Primal local norm:** $\|u\|_x := \sqrt{\langle \nabla^2 f(x) u, u \rangle}$, $u \in E$.
- **Dual local norm:** $\|g\|_{x,*} := \sqrt{\langle g, (\nabla^2 f(x))^{-1} g \rangle}$, $g \in E^*$.

Same norm/dual-norm structure as Lectures 7–10, except the norm *moves with* x.

Convexity test (Lemma 14.1). f convex on open convex $U \Leftrightarrow \nabla^2 f(x) \succeq 0$ for all $x \in U$.

The *strict* version $\nabla^2 f(x) \succ 0$ makes the local quadratic model strictly convex — this is what makes the Newton direction uniquely defined.

Newton Direction and Newton Decrement

Definition 14.2 (Newton direction). At x with $\nabla^2 f(x) \succ 0$:

$$d_f(x) := -\big(\nabla^2 f(x)\big)^{-1} \nabla f(x).$$

Newton decrement: $\lambda_f(x) := \|\nabla f(x)\|_{x,*} = \sqrt{\langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \rangle}$.

Two key identities.

- $\|d_f(x)\|_x^2 = \lambda_f(x)^2$ ("Newton-step size in the local metric")
- $\langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2$ ("descent rate per unit decrement")

Damped Newton step. For $t \in (0, 1]$: $x^+ = x + t\, d_f(x)$. Full step: $t = 1$.

Stopping rule. $\lambda_f(x)^2/2 =$ predicted gap to the local quadratic minimizer.

Quadratic Model and Model Improvement

Lemma 14.2 (Newton minimizes the local quadratic model). With $m_x(d) := f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2}\langle \nabla^2 f(x) d, d \rangle$:

$$d_f(x) = \operatorname*{argmin}_d m_x(d) = \operatorname*{argmin}_d \left\{ \langle \nabla f(x), d \rangle + \tfrac{1}{2}\|d\|_x^2 \right\},$$

and

$$\boxed{m_x(d_f(x)) = f(x) - \tfrac{1}{2}\lambda_f(x)^2.}$$

Three takeaways.

- Newton is *steepest descent in the Hessian metric* — with the metric chosen by the objective.
- $\frac{1}{2}\lambda_f(x)^2$ is the *model gap*: how much the local quadratic predicts f will drop.
- *Caveat*: this is a model gap, not a true gap to $f^\star$ — without extra structure (e.g., self-concordance, Lecture 16), it does not certify global suboptimality.

What the Decrement Measures

Lemma 14.3 (Decrement vs gradient norm). If $\mu I \preceq \nabla^2 f(x) \preceq MI$, then

$$\frac{1}{\sqrt{M}}\|\nabla f(x)\|_2 \;\le\; \lambda_f(x) \;\le\; \frac{1}{\sqrt{\mu}}\|\nabla f(x)\|_2.$$

The decrement is the gradient norm *measured in the inverse-Hessian geometry*.

- Well-conditioned region $\mu \approx M$: $\lambda_f(x) \approx \|\nabla f(x)\|_2/\sqrt{M}$.
- $\lambda_f(x) = 0 \Leftrightarrow \nabla f(x) = 0$. So λ_f is a valid stopping signal.

Why use λ_f instead of $\|\nabla f\|_2$?

- $\|\nabla f\|_2$ depends on the choice of coordinates.
- λ_f is *affine-invariant* (next slide).

Affine Invariance

Proposition 14.4 (Affine invariance of Newton). Let $T(z) = Az + b$ with A invertible, and $g(z) := f(Tz)$. Setting $x = Tz$:

$$d_g(z) = A^{-1} d_f(x), \qquad \lambda_g(z) = \lambda_f(x).$$

Hence full and damped Newton trajectories are linked by T:

$$x_t = T(z_t) \ \ \forall t.$$

Why this matters. Coordinate changes / preconditioning do not change Newton's behavior.

- GD: scaling axes changes step size, conditioning, convergence rate.
- Newton: identical iterate trajectories under any linear isomorphism.

Slogan. Newton chooses the "right" metric automatically, in every coordinate system.

Affine Invariance: Proof Sketch

Setup. $x = Tz = Az + b$, $g(z) := f(Tz)$. Pullback $A^* : E^* \to F^*$, $A^*\xi := \xi \circ A$.

Step 1 (chain rule). $\nabla g(z) = A^* \nabla f(x), \quad \nabla^2 g(z) = A^* \nabla^2 f(x) A$. The Hessian of g stays positive definite because A is an isomorphism.

Step 2 (Newton direction). Verify $A^{-1} d_f(x)$ solves the Newton equation for g:

$$\nabla^2 g(z)\big(A^{-1} d_f(x)\big) = A^* \nabla^2 f(x) d_f(x) = -A^* \nabla f(x) = -\nabla g(z).$$

By uniqueness of the Newton direction, $d_g(z) = A^{-1} d_f(x)$.

Step 3 (decrement). $(\nabla^2 g(z))^{-1} \nabla g(z) = A^{-1} (\nabla^2 f(x))^{-1} \nabla f(x)$, hence

$$\lambda_g(z)^2 = \langle \nabla g(z), (\nabla^2 g(z))^{-1} \nabla g(z) \rangle = \langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \rangle = \lambda_f(x)^2.$$

Step 4 (trajectories). $T(z_t + \eta d_g(z_t)) = x_t + \eta d_f(x_t)$. The Armijo test is invariant because $\langle \nabla g, d_g \rangle = \langle \nabla f, d_f \rangle$ and $g(z + \eta d_g) = f(x + \eta d_f)$. Hence backtracking accepts identical step sizes. $\square$

Quadratic Exactness: One Step

Theorem 14.5 (Quadratic exactness). For $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with $Q \succ 0$, one full Newton step from any x lands at the unique minimizer:

$$x + d_f(x) = -Q^{-1}b = x^\star.$$

Why. $\nabla f(x) = Qx + b$, $\nabla^2 f(x) = Q$, so $d_f(x) = -Q^{-1}(Qx + b) = -x - Q^{-1}b$, and $x + d_f(x) = -Q^{-1}b$.

What this tells us. On a quadratic, the local quadratic model *is* the global function. The Hessian-Lipschitz constant is 0. *Newton has nothing to approximate.*

For non-quadratic f: each step is exact on the local quadratic; Hessian-Lipschitz controls the discrepancy from one step to the next. This is the source of *quadratic* local convergence (next slides).

Example: Ridge Least Squares Solved in One Step

Problem. $f(w) = \frac{1}{2}\|Aw - b\|_2^2 + \frac{\sigma}{2}\|w\|_2^2, \quad \sigma > 0.$

Newton ingredients.

$$\nabla f(w) = A^\top(Aw - b) + \sigma w, \qquad \nabla^2 f(w) = A^\top A + \sigma I.$$

One Newton step from any w.

$$w + d_f(w) = (A^\top A + \sigma I)^{-1} A^\top b.$$

This is exactly the ridge-regression closed form.

Contrast with gradient descent.

$$w^+ = w - \eta\big(A^\top(Aw - b) + \sigma w\big).$$

Same residual; *scalar* η vs. *matrix* $(A^\top A + \sigma I)^{-1}$. Newton replaces the step size with the curvature inverse.

Local Quadratic Convergence

Theorem 14.6 (Local quadratic convergence). Suppose $\nabla f(x^\star) = 0$, $\nabla^2 f(x^\star) \succ 0$, and $\nabla^2 f$ is ρ-Lipschitz on a ball $B(x^\star, r)$. Then for x_0 close enough,

$$\|x_{t+1} - x^\star\|_2 \;\le\; C\|x_t - x^\star\|_2^2.$$

Doubling-precision-per-step. Once in the basin of attraction, the error *squares* every step.

- One step: $10^{-2} \to 10^{-4} \to 10^{-8} \to 10^{-16}$. **Doubling-precision-per-step.**
- Compare GD: $\|x_{t+1} - x^\star\|_2 \le \rho\,\|x_t - x^\star\|_2$ for some $\rho < 1$ — linear, not quadratic.

Caveat. This is a *local* statement. Outside the basin, Newton can diverge or take unboundedly large steps. Globalization (next part) handles this.

Local Convergence: Proof Sketch

Setup: $H_\star := \nabla^2 f(x^\star) \succ 0$. Near $x^\star$, $\nabla^2 f(x) \succeq (\mu/2) I$, so $\|\nabla^2 f(x)^{-1}\|_2 \le 2/\mu$.

Newton residual. Write $e := x - x^\star$, $x^+ = x + d_f(x)$:

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

Hessian-Lipschitz: replace $\nabla^2 f(x^\star + se)$ with $\nabla^2 f(x)$.

$$\nabla^2 f(x) e - (\nabla f(x) - \nabla f(x^\star)) = \int_0^1 \big(\nabla^2 f(x) - \nabla^2 f(x^\star + se)\big) e \, ds.$$

ρ-Lipschitz $\Rightarrow$ integrand has norm $\le \rho(1-s)\|e\|_2^2$. Integrating: $\le \frac{\rho}{2}\|e\|_2^2$.

Combine. $\|x^+ - x^\star\|_2 \le \frac{\rho}{\mu}\|e\|_2^2$. Set $C = \rho/\mu$. □

Heart. The integral remainder of Taylor's theorem; Hessian-Lipschitz controls it; $\|H^{-1}\| \le 2/\mu$ converts norm-of-residual into norm-of-error.

Logistic Regression: Newton as IRLS

Problem. $f(w) = \sum_{i=1}^{m} \log(1 + \exp(-b_i a_i^\top w)) + \frac{\sigma}{2}\|w\|_2^2$.
Hessian. Let $q_i(w) := 1/(1 + \exp(b_i a_i^\top w))$. Then

$$\nabla^2 f(w) = A^\top D(w) A + \sigma I, \quad D(w) = \operatorname{diag}(q_i(w)(1 - q_i(w))).$$

Newton step. Solve

$$\left(A^\top D(w) A + \sigma I\right) d = -\nabla f(w).$$

This is **iteratively reweighted least squares (IRLS)**. Weights $q_i(1 - q_i)$:

- Large near uncertain examples ($q_i \approx 1/2$).
- Small for confident examples ($q_i \approx 0$ or ≈ 1).

Curvature analysis. $\nabla^2 f \succeq \sigma I$ (regularization gives strong convexity). $\nabla^2 f$ is Hessian-Lipschitz with constant $\propto \sum_i \|a_i\|_2^3$.

Cost Model: Iterations vs Wall Clock

Compare per iteration.

- Gradient step: $O(n)$ + first-order oracle.
- Newton step: Hessian (or H-vector products) + linear solve $Hd = -g$.

Cost of the linear solve.

- Dense direct (Cholesky): $O(n^3)$ flops, $O(n^2)$ memory.
- Iterative (CG): $O(n^2)\times$ (effective condition number).
- Structured (sparse, low-rank, Kronecker): can be near-linear.

When Newton wins.

- Small/medium n (linear solve cheap).
- Hessian has structure (sparse, low-rank, $A^\top DA + \sigma I$).
- Need high accuracy ($\log\log$ phase pays off).
- Ill-conditioned Hessian where GD struggles.

When GD/AGD wins. Large-scale learning where n^2 memory is impossible and modest accuracy suffices.

Globalization: Damped Newton with Backtracking

Algorithm (Backtracking Damped Newton). Parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$.

1. Compute $d_t = d_f(x_t)$, $\lambda_t = \lambda_f(x_t)$.
2. If $\lambda_t = 0$: stop.
3. Set $\eta = 1$.
4. **While** $x_t + \eta d_t \notin \operatorname{dom} f$ **or** $f(x_t + \eta d_t) > f(x_t) + \alpha\eta\langle \nabla f(x_t), d_t \rangle$: $\eta \leftarrow \beta\eta$.
5. $x_{t+1} = x_t + \eta d_t$.

Two checks.

- *Domain test*: keep x_t inside $\operatorname{dom} f$. Critical for log barriers (Lecture 16).
- *Armijo test*: actual descent matches α-fraction of the predicted descent.

Slogan. Backtracking does *not* change the Newton direction. It only checks whether the local quadratic model is reliable at the current scale.

Why Backtracking Terminates

Proposition 14.7 (Descent + finite Armijo). If f is differentiable and $\nabla^2 f(x) \succ 0$, $\nabla f(x) \neq 0$:

- Newton direction is strict descent: $\langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2 < 0$.
- For every Armijo $\alpha \in (0,1)$, there is $\bar{\eta} > 0$ such that every $\eta \in (0, \bar{\eta}]$ satisfies the domain and Armijo conditions.

Backtracking eventually accepts: tested values $\eta = 1, \beta, \beta^2, \ldots$ reach $\bar{\eta}$ in finite steps.

Caveat: positive curvature is essential. If $\nabla^2 f(x)$ has a negative direction, $d_f(x)$ may point to a *saddle* or *maximum* of the quadratic model. The descent identity uses positive definiteness. **Today's lecture is the convex regime.**

Nonconvex variants modify H, use trust regions, or regularize: see cubic regularization below.

Two-Phase Damped Newton Theorem

Theorem 14.8 (Globalized convergence). Assume $\mu I \preceq \nabla^2 f$ and $\nabla^2 f$ is ρ-Lipschitz globally; $f^\star > -\infty$. Define $\lambda_{\mathrm{N}} := \min\{1, 3(1-2\alpha)\}\,\mu^{3/2}/\rho$ and $\gamma_{\mathrm{N}} := \alpha\beta\lambda_{\mathrm{N}}^2$. Then:

1. **Damped phase** ($\lambda_t \geq \lambda_{\mathrm{N}}$): $f(x_{t+1}) \leq f(x_t) - \gamma_{\mathrm{N}}$.
2. **Quadratic phase** ($\lambda_t < \lambda_{\mathrm{N}}$): backtracking accepts $\eta = 1$, and $\lambda_{t+1} \leq \frac{\rho}{2\mu^{3/2}}\lambda_t^2 \leq \frac{1}{2}\lambda_t$.

Total complexity.

- Damped phase iterations: $\leq \lceil (f(x_0) - f^\star)/\gamma_{\mathrm{N}} \rceil$ — each guarantees a constant drop.
- Quadratic phase: $\log\log(1/\varepsilon)$ iterations to reach decrement $\leq \varepsilon$.

Slogan. Constant-drop damped phase + doubling-precision quadratic phase.

Helper Lemma: Hessian-Lipschitz Taylor Estimate

Lemma 14.4 (Hessian-Lipschitz Taylor). If $\nabla^2 f$ is ρ-Lipschitz on $\mathbb{R}^n$, then for all x, s:

$$\boxed{f(x+s) \;\le\; f(x) + \langle \nabla f(x), s\rangle + \tfrac{1}{2} s^\top \nabla^2 f(x) s + \tfrac{\rho}{6}\|s\|_2^3.}$$

Proof. Let $\phi(\theta) := f(x+\theta s)$, $\phi''(\theta) = s^\top \nabla^2 f(x+\theta s) s$. Integral Taylor:

$$f(x+s) = f(x) + \phi'(0) + \int_0^1 (1-\theta)\, \phi''(\theta)\, d\theta.$$

Add/subtract $s^\top \nabla^2 f(x) s$. Hessian-Lipschitz: $s^\top(\nabla^2 f(x+\theta s) - \nabla^2 f(x))s \le \rho\theta\|s\|_2^3$. Then $\int_0^1 \theta(1-\theta)d\theta = \frac{1}{6}$. $\square$

What this says. The L-smoothness upper bound ($\frac{L}{2}\|s\|^2$) is replaced by the *true* quadratic at x + a cubic correction — sharper whenever the Hessian varies slowly.

Two-Phase Proof (1/3): Apply Lemma 14.4 to the Newton Direction

Setup. Write $g := \nabla f(x)$, $d := d_f(x)$, $\lambda := \lambda_f(x)$. From Lemma 14.2:

$$g^\top d = -\lambda^2, \qquad d^\top \nabla^2 f(x) d = \lambda^2.$$

Bound on $\|d\|_2$. $\nabla^2 f \succeq \mu I \;\Rightarrow\; \mu\|d\|_2^2 \le \lambda^2 \;\Rightarrow\; \|d\|_2 \le \lambda/\sqrt{\mu}$.
Apply Lemma 14.4 with $s = td$, $t \in [0,1]$.

$$f(x+td) \le f(x) + t\underbrace{g^\top d}_{=-\lambda^2} + \frac{t^2}{2}\underbrace{d^\top \nabla^2 f(x) d}_{=\lambda^2} + \frac{\rho t^3}{6}\underbrace{\|d\|_2^3}_{\le \lambda^3/\mu^{3/2}}.$$

Boxed Taylor estimate (along the Newton direction).

$$\boxed{f(x+td) \;\le\; f(x) + t\lambda^2\Big(-1 + \tfrac{t}{2} + \tfrac{\rho t \lambda}{6\mu^{3/2}}\Big), \quad 0 \le t \le 1.}$$

The bracket controls whether Armijo accepts step t.

Two-Phase Proof (2/3): Damped Phase ($\lambda \geq \lambda_{\mathrm{N}}$)

Goal. Show backtracking accepts a step that gives a uniform objective drop $\gamma_{\mathrm{N}} = \alpha\beta\lambda_{\mathrm{N}}^2$.

Pick the trial step $\widehat{t} := \dfrac{\lambda_{\mathrm{N}}}{\lambda} \in (0, 1]$. For any $t \in (0, \widehat{t}]$, $t\lambda \leq \lambda_{\mathrm{N}}$, so the boxed Taylor estimate gives

$$f(x + td) \leq f(x) - t\lambda^2\Big(1 - \underbrace{\tfrac{t}{2}}_{\leq 1/2} - \underbrace{\tfrac{\rho t\lambda}{6\mu^{3/2}}}_{\leq \rho\lambda_{\mathrm{N}}/(6\mu^{3/2})}\Big).$$

Tune λ_{N}**.** Recall $\lambda_{\mathrm{N}} \leq 3(1-2\alpha)\mu^{3/2}/\rho$, so $\frac{\rho\lambda_{\mathrm{N}}}{6\mu^{3/2}} \leq \frac{1-2\alpha}{2}$. Hence

$$\tfrac{t}{2} + \tfrac{\rho t\lambda}{6\mu^{3/2}} \;\leq\; \tfrac{1}{2} + \tfrac{1-2\alpha}{2} \;=\; 1 - \alpha,$$

giving $f(x + td) \leq f(x) - \alpha t\lambda^2 = f(x) + \alpha t\, g^\top d$. **(Armijo at every** $t \leq \widehat{t}$**.)**

Backtracking accepts $t_{\mathrm{acc}} \geq \beta\widehat{t}$**.** Hence $f(x) - f(x^+) \geq \alpha\beta\lambda_{\mathrm{N}}\lambda \geq \alpha\beta\lambda_{\mathrm{N}}^2 = \gamma_{\mathrm{N}}$.

Two-Phase Proof (3/3): Quadratic Phase ($\lambda < \lambda_N$)

Full step accepted. The boxed Taylor estimate at $t = 1$ with $1 \cdot \lambda < \lambda_N$ gives the same Armijo conclusion as in the damped phase, so backtracking accepts $\eta = 1$. Set $x^+ := x + d$.

Newton residual at x^+. Since $g + \nabla^2 f(x)d = 0$, the fundamental theorem of calculus gives

$$\nabla f(x^+) = \int_0^1 \big(\nabla^2 f(x + sd) - \nabla^2 f(x)\big) d \, ds.$$

Bound the residual. Hessian-Lipschitz gives $\|\nabla^2 f(x + sd) - \nabla^2 f(x)\|_2 \le \rho s \|d\|_2$. So

$$\|\nabla f(x^+)\|_2 \;\le\; \int_0^1 \rho s \|d\|_2^2 \, ds \;=\; \tfrac{\rho}{2}\|d\|_2^2 \;\le\; \tfrac{\rho}{2\mu}\lambda^2.$$

Convert to decrement. $\nabla^2 f(x^+) \succeq \mu I \;\Rightarrow\; \lambda_f(x^+) \le \frac{1}{\sqrt{\mu}}\|\nabla f(x^+)\|_2$. Combine:

$$\boxed{\lambda_f(x^+) \;\le\; \tfrac{\rho}{2\mu^{3/2}}\lambda^2 \;\le\; \tfrac{1}{2}\lambda \;\text{ (since } \lambda < \lambda_N).}$$

□

Aside: Cubic Regularization

Yurii Nesterov

Boris T. Polyak

Yu. Nesterov & B. T. Polyak, "Cubic regularization of Newton method and its global performance," Math. Program. 108 (2006).

Idea. If f has ρ-Lipschitz Hessian, Taylor gives the cubic upper model

$$f(x+s) \le f(x) + \langle \nabla f(x), s \rangle + \tfrac{1}{2} s^\top \nabla^2 f(x) s + \tfrac{\rho}{6} \|s\|_2^3.$$

Method. Pick $M \ge \rho$. At each step, choose s_t minimizing the cubic upper model:

$$s_t \in \operatorname*{argmin}_s \left\{ \langle \nabla f(x_t), s \rangle + \tfrac{1}{2} s^\top \nabla^2 f(x_t) s + \tfrac{M}{6} \|s\|_2^3 \right\}.$$

Rate (convex case). If $\|x - x^\star\|_2 \le R$ on the level set $\{f \le f(x_0)\}$:

$$f(x_T) - f^\star = O\big(MR^3/T^2\big).$$

Bonuses. Cubic term makes the model coercive even with indefinite Hessian — handles nonconvex Newton too.

Summary

Newton = exact minimizer of the local quadratic Taylor model.

- Hessian local norm $\|u\|_x$ + dual local norm $\|g\|_{x,*}$ define a metric that moves with x.
- Newton direction $d_f = -(\nabla^2 f)^{-1}\nabla f$; decrement $\lambda_f = \|\nabla f\|_{x,*}$.
- Affine-invariant under invertible affine changes of variables.

Behavior.

- Quadratic objective: *exact* in one step.
- Locally near a nondegenerate min: *quadratic* convergence (ρ/μ rate).
- Globally with $\mu I \preceq \nabla^2 f$ and $\nabla^2 f$ ρ-Lipschitz: damped phase + quadratic phase (Theorem 14.8).

Cost. Linear solve per iteration. Worth it for high accuracy and structured Hessians.

Next. Lecture 15: self-concordant Newton calculus — replace the Euclidean Hessian-Lipschitz constant by an affine-invariant local condition that survives near boundaries (e.g., $-\log t$). This is the right setting for interior-point methods in Lecture 16.

Bibliographic Notes

- **S. Boyd & L. Vandenberghe, Convex Optimization, Cambridge (2004).** Chapter 9 covers Newton's method, the damped variant, and the two-phase analysis used today.
- **Yu. Nesterov, Introductory Lectures on Convex Optimization, Kluwer (2004).** Chapter 1.2: classical Newton; Chapter 4: the self-concordant generalization that Lecture 16 will pick up.

Cubic regularization (Nesterov & Polyak 2006) is cited inline on the cubic-regularization slide.