Lecture 14: Newton method and Hessian Geometry

Lecture 13 closed the first-order smooth convex story by reaching the optimal $O(LR^2/T^2)$ rate. Lecture 14 changes the local model. Instead of asking only for a linear approximation plus a fixed smoothness constant, we now use the Hessian itself as the local metric. Newton's method is the algorithmic form of this idea: at the current point, solve the quadratic Taylor model exactly, and use that model to choose both a direction and a stopping quantity.

A useful way to read this lecture is as the fixed-metric story from Lecture 10 taken to its logical endpoint. If a positive-definite matrix H is fixed, the preconditioned gradient step is

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

For a quadratic objective

$$f(x) = \frac{1}{2}x^\top Qx + b^\top x + c,$$

choosing $H = Q$ makes this one step exact. For a nonquadratic objective there is no single matrix H that represents the curvature everywhere, so Newton's method changes the preconditioner online by using $\nabla^2 f(x_t)$ at the current iterate. The price is that each iteration must solve a linear system. The gain is that the algorithm becomes insensitive to affine changes of coordinates and can enter a quadratic local convergence regime.

14.1 Hessian Geometry and Local Norms

Throughout the intrinsic statements in this lecture, E denotes a finite-dimensional real vector space. Before using coordinates, it is useful to keep the intrinsic type in mind. For such an E, the gradient $\nabla f(x)$ is an element of E^*, and the Hessian is a symmetric linear map

$$\nabla^2 f(x) : E \to E^*, \qquad \left\langle \nabla^2 f(x)u, v \right\rangle = D^2 f(x)[u, v].$$

This is the same type as a metric operator. For a symmetric linear map $A : E \to E^*$, meaning $\langle Au, v\rangle = \langle Av, u\rangle$ for all $u, v \in E$, write $A \succeq 0$ if $\langle Au, u\rangle \geq 0$ for all $u \in E$, and write $A \succ 0$ if $\langle Au, u\rangle > 0$ for all nonzero $u \in E$. After choosing coordinates and identifying $E^* \simeq E$ by the Euclidean pairing, the same object is represented by the usual Hessian matrix. In coordinate estimates below we use matrix notation; the intrinsic Newton equation is always

$$\nabla^2 f(x)d = -\nabla f(x).$$

Lemma 14.1 (Convexity and positive semidefinite Hessian)**.** *Let $U \subseteq E$ be open and convex, and let $f : U \to \mathbb{R}$ be twice differentiable. Then f is convex on U if and only if*

$$\nabla^2 f(x) \succeq 0 \qquad \forall x \in U.$$

In particular, the stronger condition $\nabla^2 f(x) \succ 0$ makes the local quadratic model strictly convex

and makes the Newton direction uniquely defined.

Proof of Lemma 14.1. Fix $x, y \in U$, and define $\phi(t) := f((1-t)x + ty)$ on $[0,1]$. If $\nabla^2 f(z) \succeq 0$ on U, then

$$\phi''(t) = \Big\langle \nabla^2 f((1-t)x + ty)(y-x), y-x \Big\rangle \geq 0.$$

Hence ϕ' is nondecreasing. For $0 < \theta < 1$,

$$\phi(\theta) - \phi(0) = \int_0^\theta \phi'(s)\, ds \leq \theta\phi'(\theta), \qquad \phi(1) - \phi(\theta) = \int_\theta^1 \phi'(s)\, ds \geq (1-\theta)\phi'(\theta).$$

Combining the two inequalities gives $\phi(\theta) \leq (1-\theta)\phi(0) + \theta\phi(1)$, so f is convex on U.
Conversely, assume f is convex. Fix $x \in U$ and $v \in E$. Since U is open, $x + tv \in U$ for all sufficiently small $|t|$. The function $\psi(t) := f(x + tv)$ is convex, so for small $h > 0$,

$$\psi(0) \leq \frac{\psi(h) + \psi(-h)}{2}.$$

Dividing by h^2 and letting $h \downarrow 0$ gives

$$\Big\langle \nabla^2 f(x)v, v \Big\rangle = \psi''(0) \geq 0.$$

Since this holds for every $v \in E$, $\nabla^2 f(x) \succeq 0$. □

Definition 14.1 (Hessian local norm and dual local norm)**.** Let $f : U \to \mathbb{R}$ be twice differentiable on an open set $U \subseteq E$, and let $x \in U$ satisfy

$$\nabla^2 f(x) \succ 0.$$

The Hessian local norm at x is

$$\|u\|_x := \sqrt{\langle \nabla^2 f(x)u, u\rangle}, \qquad u \in E,$$

and the corresponding dual local norm is

$$\|g\|_{x,*} := \sqrt{\langle g, (\nabla^2 f(x))^{-1} g\rangle}, \qquad g \in E^*,$$

where $(\nabla^2 f(x))^{-1} : E^* \to E$.

This notation is deliberately parallel to the norm-dual-norm language from Lectures 7–10. The difference is that the norm now changes with x. The rest of the lecture studies what this moving geometry buys us, and Lecture 15 will study when the moving geometry remains stable in its own local units.

14.2 Newton Step and Model Improvement

Definition 14.2 (Newton direction and Newton decrement)**.** Let $U \subseteq E$ be open, let $f : U \to \mathbb{R}$ be twice differentiable, and let $x \in U$ satisfy $\nabla^2 f(x) \succ 0$. The Newton direction at x is

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

The Newton decrement at x is

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

Definition 14.3 (Damped Newton step)**.** For $t \in (0, 1]$, the damped Newton step with step size t is

$$x^+ = x + t\, d_f(x).$$

The case $t = 1$ is called a full Newton step.

Lemma 14.2 (Quadratic model, Newton direction, and model improvement)**.** *Let $U \subseteq E$ be open, let $f : U \to \mathbb{R}$ be twice differentiable, and let $x \in U$ satisfy $\nabla^2 f(x) \succ 0$. Define the quadratic Taylor model*

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

Then $d_f(x)$ is the unique minimizer of m_x, equivalently

$$d_f(x) = \arg\min_d \left\{ \langle \nabla f(x), d \rangle + \frac{1}{2} \|d\|_x^2 \right\},$$

and

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

Moreover,

$$\|d_f(x)\|_x^2 = \lambda_f(x)^2, \qquad \langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2.$$

Proof of Lemma 14.2. The function $d \mapsto m_x(d)$ is strictly convex because $\nabla^2 f(x) \succ 0$. Its first-order optimality condition is

$$\nabla f(x) + \nabla^2 f(x) d = 0,$$

whose unique solution is

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

Thus $d_f(x)$ uniquely minimizes m_x. Moreover,

$$\langle \nabla f(x), d_f(x) \rangle = -\left\langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \right\rangle = -\lambda_f(x)^2$$

and

$$\left\langle \nabla^2 f(x) d_f(x), d_f(x) \right\rangle = \left\langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \right\rangle = \lambda_f(x)^2.$$

Substituting these identities into m_x gives

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

□

Thus Newton's method is steepest descent in the Hessian metric, but with the metric chosen from the objective rather than fixed in advance. The decrement $\lambda_f(x)$ measures the size of the Newton step in this same local metric and the amount of improvement predicted by the quadratic model.

Lemma 14.3 (Uniform Hessian bounds make the decrement comparable to the gradient norm)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be twice differentiable, and suppose that at $x \in U$*

$$\mu I \preceq \nabla^2 f(x) \preceq MI \qquad \text{for some } 0 < \mu \leq M.$$

Then

$$\frac{1}{\sqrt{M}} \|\nabla f(x)\|_2 \leq \lambda_f(x) \leq \frac{1}{\sqrt{\mu}} \|\nabla f(x)\|_2 ,$$

and the Newton direction satisfies

$$\frac{1}{M} \|\nabla f(x)\|_2 \leq \|d_f(x)\|_2 \leq \frac{1}{\mu} \|\nabla f(x)\|_2 .$$

In particular, $\lambda_f(x) = 0$ if and only if $\nabla f(x) = 0$.

Proof of Lemma 14.3. Write $g := \nabla f(x)$ and $H := \nabla^2 f(x)$. From $\mu I \preceq H \preceq MI$, we get

$$\frac{1}{M} I \preceq H^{-1} \preceq \frac{1}{\mu} I.$$

Therefore

$$\frac{1}{M} \|g\|_2^2 \leq g^\top H^{-1} g \leq \frac{1}{\mu} \|g\|_2^2 .$$

Taking square roots gives the bounds on $\lambda_f(x)$. Also $d_f(x) = -H^{-1}g$, and the eigenvalues of H^{-1} lie in $[1/M, 1/\mu]$, so

$$\frac{1}{M} \|g\|_2 \leq \left\|H^{-1} g\right\|_2 \leq \frac{1}{\mu} \|g\|_2 .$$

The final statement follows because H^{-1} is positive definite. □

This Euclidean coordinate lemma explains why the decrement is a reasonable stopping quantity in well-conditioned regions: it is the gradient norm measured in the inverse Hessian geometry. In Lecture 15, self-concordance will make this statement sharper by turning $\lambda_f(x)$ into an affine-invariant suboptimality certificate near the solution.

The practical stopping rule is usually stated in terms of $\lambda_f(x)^2/2$, because $\lambda_f(x)^2/2$ is exactly the predicted gap between the current value and the minimum of the local quadratic model in Lemma 14.2. Without additional assumptions this is only a model-based certificate, not a true global duality gap. The self-concordant analysis in Lecture 15 is the point where this model quantity becomes a controlled certificate for the original objective.

14.3 Invariance and Quadratic Exactness

Proposition 14.4 (Affine invariance of Newton trajectories)**.** *Let E, F be finite-dimensional real vector spaces. Let $U \subseteq E$ be open, and let $f : U \to \mathbb{R}$ be twice differentiable with $\nabla^2 f(x) \succ 0$ for every $x \in U$. Let $T : F \to E$ be an invertible affine map, write $Tz = Az + b$ with $A : F \to E$ a linear isomorphism, set*

$$V := T^{-1}(U), \qquad g : V \to \mathbb{R}, \qquad g(z) := f(Tz).$$

If $z \in V$ and $x = Tz$, then

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

Consequently, if $x_0 = Tz_0$, then full Newton iterates satisfy

$$x_t = Tz_t \qquad \forall t$$

for as long as both trajectories are defined. The same trajectory correspondence holds for backtracking damped Newton with the same Armijo parameters α, β: the two line searches accept the same step sizes, and therefore

$$x_t = Tz_t \qquad \forall t.$$

Proof of Proposition 14.4. Write $x = Tz = Az + b$, and let $A^* : E^* \to F^*$ denote the pullback $A^*\xi := \xi \circ A$. Then

$$\nabla g(z) = A^* \nabla f(x), \qquad \nabla^2 g(z) = A^* \nabla^2 f(x) A.$$

The Hessian of g is positive definite because A is a linear isomorphism. Moreover,

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

By uniqueness of the Newton direction for g, this gives $d_g(z) = A^{-1} d_f(x)$.
For the decrement,

$$(\nabla^2 g(z))^{-1} \nabla g(z) = A^{-1} (\nabla^2 f(x))^{-1} \nabla f(x),$$

because applying $A^* \nabla^2 f(x) A$ to the right-hand side gives $A^* \nabla f(x) = \nabla g(z)$. Hence

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

Taking square roots proves the decrement identity.
For full Newton, if $x_t = Tz_t$, then

$$T(z_t + d_g(z_t)) = Tz_t + A d_g(z_t) = x_t + d_f(x_t).$$

Induction proves the full-step trajectory claim. For damped Newton, the domain test is invariant because

$$T(z_t + \eta d_g(z_t)) = x_t + \eta d_f(x_t),$$

and the Armijo test is invariant because

$$\langle \nabla g(z_t), d_g(z_t) \rangle = \langle \nabla f(x_t), d_f(x_t) \rangle, \qquad g(z_t + \eta d_g(z_t)) = f(x_t + \eta d_f(x_t)).$$

Thus the two backtracking loops accept the same tested step sizes, and induction proves the damped-trajectory claim. □

Theorem 14.5 (Quadratic exactness)**.** *Let*

$$f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$$

with $Q \succ 0$*. Then for every* $x \in \mathbb{R}^n$*, one full Newton step sends* x *exactly to the unique minimizer*

$$x^\star = -Q^{-1}b.$$

Proof of Theorem 14.5. For the quadratic objective, $\nabla f(x) = Qx + b$ and $\nabla^2 f(x) = Q$. Hence

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

The full Newton step gives $x + d_f(x) = -Q^{-1}b = x^\star$. Since $Q \succ 0$, the quadratic is strictly convex, so $x^\star$ is the unique minimizer. □

Example 14.1 (Ridge least squares: Newton solves the problem in one step)**.** This example is the constant-curvature benchmark: it shows what Newton should do when the quadratic model is the objective, not merely an approximation. Consider

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

Then

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

The Newton direction solves

$$(A^\top A + \sigma I)d_f(w) = -(A^\top(Aw - b) + \sigma w),$$

so the full step gives

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

This is exactly the ridge-regression minimizer. By contrast, gradient descent uses the same residual vector but only multiplies it by a scalar stepsize:

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

Newton replaces the scalar η by the curvature matrix $(A^\top A + \sigma I)^{-1}$. Since the Hessian is constant, its Lipschitz constant is 0; this is the degenerate case where the local quadratic model is globally exact.

14.4 Local Quadratic Convergence

The exact quadratic calculation is not just a special case. It predicts the local behavior of Newton's method: once the Hessian changes slowly on the scale of the current error, the next error is quadratic in the current error. Theorem 14.6 makes this local statement precise. Its hypothesis is a basin condition, not a global convergence theorem: it assumes the initialization is already close enough to a nondegenerate stationary point.

Theorem 14.6 (Local quadratic convergence of Newton's method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Assume that $x^\star \in \mathbb{R}^n$ satisfies*

$$\nabla f(x^\star) = 0, \qquad \nabla^2 f(x^\star) \succ 0.$$

Assume that there exist $r > 0$ and $\rho > 0$ such that

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \le \rho \left\|x - y\right\|_2 \qquad \forall x, y \in B(x^\star, r).$$

Then there exists $r_0 \in (0, r]$ such that, if $x_0 \in B(x^\star, r_0)$ and

$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t),$$

then every iterate is well defined, $x_t \to x^\star$, and there exists $C > 0$ such that

$$\left\|x_{t+1} - x^\star\right\|_2 \le C \left\|x_t - x^\star\right\|_2^2 \qquad \forall t \ge 0.$$

Proof of Theorem 14.6. The proof is the Hessian-Lipschitz perturbation of the exact quadratic case in Theorem 14.5. Let $H_\star := \nabla^2 f(x^\star)$. Since $H_\star \succ 0$, let $\mu := \lambda_{\min}(H_\star) > 0$. By continuity of $\nabla^2 f$, after shrinking r if necessary there exists $r_1 \in (0, r]$ such that

$$\nabla^2 f(x) \succeq \frac{\mu}{2} I, \qquad \left\|\nabla^2 f(x)^{-1}\right\|_2 \le \frac{2}{\mu} \qquad \forall x \in B(x^\star, r_1).$$

Fix $x \in B(x^\star, r_1)$, write $e := x - x^\star$, and let $x^+ = x + d_f(x)$. Since $\nabla f(x^\star) = 0$,

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

By the fundamental theorem of calculus,

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

so

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

Taking norms and using the Hessian-Lipschitz assumption gives

$$\left\|\nabla^2 f(x) e - \big(\nabla f(x) - \nabla f(x^\star)\big)\right\|_2 \le \int_0^1 \rho(1 - s) \left\|e\right\|_2^2 \, ds = \frac{\rho}{2} \left\|e\right\|_2^2.$$

Therefore

$$\left\|x^+ - x^\star\right\|_2 \le \frac{\rho}{\mu} \left\|x - x^\star\right\|_2^2.$$

Set $C := \rho/\mu$, and choose $r_0 \in (0, r_1]$ so small that $C r_0 \le 1/2$. If $x_t \in B(x^\star, r_0)$, then

$$\left\|x_{t+1} - x^\star\right\|_2 \le C \left\|x_t - x^\star\right\|_2^2 \le \frac{1}{2} \left\|x_t - x^\star\right\|_2,$$

so $x_{t+1} \in B(x^\star, r_0)$. Induction gives well-defined iterates inside the ball, the displayed quadratic recursion, and convergence to $x^\star$. □

Example 14.2 (Logistic regression as a Newton problem)**.** This example is the first nonquadratic test case: Newton is no longer exact in one step, but the Hessian has useful linear-algebra structure. For data $(a_i, b_i) \in \mathbb{R}^n \times \{\pm 1\}$, logistic regression minimizes

$$f(w) := \sum_{i=1}^{m} \log(1 + \exp(-b_i a_i^\top w)) + \frac{\sigma}{2} \|w\|_2^2, \qquad \sigma > 0.$$

Let

$$q_i(w) := \frac{1}{1 + \exp(b_i a_i^\top w)}.$$

Then

$$\nabla f(w) = -\sum_{i=1}^{m} q_i(w) b_i a_i + \sigma w.$$

Differentiating once more gives

$$\nabla^2 f(w) = \sum_{i=1}^{m} q_i(w)(1 - q_i(w))\, a_i a_i^\top + \sigma I.$$

If $A \in \mathbb{R}^{m \times n}$ has rows $a_i^\top$, then this can be written as

$$\nabla^2 f(w) = A^\top D(w) A + \sigma I,$$

where

$$D(w) := \operatorname{Diag}\big(q_i(w)(1 - q_i(w))\big)_{i=1}^{m}.$$

The regularization gives the uniform lower bound

$$\nabla^2 f(w) \succeq \sigma I.$$

It also has globally Lipschitz Hessian. If

$$\psi(z) := \frac{1}{1 + \exp z} \left(1 - \frac{1}{1 + \exp z}\right), \qquad C_\psi := \sup_z |\psi'(z)| = \frac{1}{6\sqrt{3}},$$

then, for every $u, v \in \mathbb{R}^n$,

$$\left\| \nabla^2 f(u) - \nabla^2 f(v) \right\|_2 \le C_\psi \left(\sum_{i=1}^{m} \|a_i\|_2^3 \right) \|u - v\|_2 .$$

The Newton step solves

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

This is an iteratively reweighted least-squares system. The weights $q_i(w)(1 - q_i(w))$ are largest near uncertain examples and small for examples with large margin under the current classifier. This example is typical: the expensive part of a Newton iteration is not evaluating the formula for d_f, but solving the structured linear system defined by the Hessian.

Remark 14.1 (Cost model). Newton's method should not be compared with gradient descent only by iteration count. A gradient step needs a first-order oracle and vector operations. A Newton step needs the Hessian or Hessian-vector products and a linear solve

$$\nabla^2 f(x_t) d_t = -\nabla f(x_t).$$

Dense direct linear algebra costs $O(n^3)$ per iteration, while structured or iterative solves may be much cheaper. This is why the logistic-regression example above emphasizes the matrix structure $A^\top D(w)A + \sigma I$, not only the formula $d_f(w) = -\nabla^2 f(w)^{-1}\nabla f(w)$.

Theorem 14.6 explains Newton's speed near the solution, but it says nothing about what happens from a remote starting point. A full Newton step can be too aggressive before the quadratic model is accurate. The next subsection adds a line search: keep the Newton direction, but shrink the step until the actual objective confirms enough descent.

14.5 Damping and Globalization

Algorithm 1 Backtracking damped Newton

Require: Initial point x_0, parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$.

for $t = 0, 1, 2, \ldots$ **do**
  Compute $d_t = d_f(x_t) = -(\nabla^2 f(x_t))^{-1}\nabla f(x_t)$.
  Compute $\lambda_t = \lambda_f(x_t)$.
  if $\lambda_t = 0$ **then**
    return x_t.
  end if
  Set $\eta = 1$.
  while $x_t + \eta d_t \notin \operatorname{dom} f$ or $f(x_t + \eta d_t) > f(x_t) + \alpha\eta \langle \nabla f(x_t), d_t \rangle$ **do**
    Set $\eta \leftarrow \beta\eta$.
  end while
  Set $x_{t+1} = x_t + \eta d_t$.
end for

Algorithm 1 separates two questions. First, is the Newton direction a descent direction? Second, if it is, does the Armijo loop eventually find an acceptable step size? Proposition 14.7 answers both questions under only local positive curvature at the current point.

Proposition 14.7 (Descent and well-defined Armijo backtracking)**.** *Let $U \subseteq E$ be open, let $f : U \to \mathbb{R}$ be differentiable on U, and assume that f is twice differentiable at $x \in U$ with $\nabla^2 f(x) \succ 0$. If $\nabla f(x) \neq 0$, then the Newton direction $d_f(x)$ is a strict descent direction:*

$$\langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2 < 0.$$

Moreover, for every Armijo parameter $\alpha \in (0, 1)$, there exists $\bar{\eta} > 0$ such that every $\eta \in (0, \bar{\eta}]$ satisfies

$$x + \eta d_f(x) \in U$$

and

$$f(x + \eta d_f(x)) \leq f(x) + \alpha\eta \langle \nabla f(x), d_f(x) \rangle .$$

Thus the Armijo loop used in Algorithm 1 terminates after finitely many reductions whenever $\nabla f(x) \neq 0$.

Proof of Proposition 14.7. The strict descent identity is already part of Lemma 14.2:

$$\langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2.$$

If $\nabla f(x) \neq 0$, then $\lambda_f(x) > 0$, so this quantity is strictly negative.
Since U is open and $x \in U$, we have $x + \eta d_f(x) \in U$ for all sufficiently small $\eta > 0$. Let

$$c := -\langle \nabla f(x), d_f(x) \rangle > 0.$$

Differentiability of f at x gives

$$f(x + \eta d_f(x)) = f(x) + \eta \langle \nabla f(x), d_f(x) \rangle + o(\eta) = f(x) - \eta c + o(\eta).$$

Choose $\bar{\eta} > 0$ so small that, for all $\eta \in (0, \bar{\eta}]$, the domain condition holds and

$$o(\eta) \leq (1 - \alpha)\eta c.$$

Then

$$f(x + \eta d_f(x)) \leq f(x) - \eta c + (1 - \alpha)\eta c = f(x) - \alpha \eta c = f(x) + \alpha \eta \langle \nabla f(x), d_f(x) \rangle .$$

The backtracking loop tests the geometric sequence $1, \beta, \beta^2, \ldots$. Some tested value is eventually at most $\bar{\eta}$, so the loop terminates. □

Backtracking is not changing the Newton direction; it is only checking whether the local quadratic model is reliable at the proposed scale. The first test keeps the iterate inside the open domain, which will matter for log barriers. The second test is the usual sufficient-decrease test. Near a well-behaved minimizer both tests eventually accept $\eta = 1$, so damping is a globalization device rather than the local source of Newton's speed.

Remark 14.2 (Why positive curvature is part of the method). The formula $d_f(x) = -(\nabla^2 f(x))^{-1} \nabla f(x)$ by itself is not a descent method. If $\nabla^2 f(x)$ has a negative direction, then the quadratic model can be locally concave along that direction, and the Newton equation may point toward a saddle point or a local maximum of the model. The descent identity

$$\langle \nabla f(x), d_f(x) \rangle = -\lambda_f(x)^2 < 0$$

uses positive definiteness of $\nabla^2 f(x)$. This lecture therefore studies the convex / positive-curvature regime. Nonconvex Newton methods usually modify the Hessian, use trust regions, or otherwise regularize the linear system.

Proposition 14.7 is qualitative: it guarantees that the line search can take a small enough step. To recover the usual global-to-local picture, we need quantitative curvature control. The next theorem gives the classical two-phase conclusion under global strong convexity and Hessian-Lipschitz assumptions: a damped phase with guaranteed objective decrease, followed by full steps and quadratic convergence of the decrement.

Theorem 14.8 (Classical two-phase damped Newton theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Assume that $p^\star := \inf_x f(x) > -\infty$, and assume there exist constants $\mu > 0$ and $\rho > 0$ such that*

$$\nabla^2 f(x) \succeq \mu I \qquad \forall x \in \mathbb{R}^n,$$

and

$$\left\|\nabla^2 f(y) - \nabla^2 f(z)\right\|_2 \le \rho \|y - z\|_2 \qquad \forall y, z \in \mathbb{R}^n.$$

Run Algorithm 1 from any $x_0 \in \mathbb{R}^n$ with parameters $\alpha \in (0, 1/2)$ and $\beta \in (0, 1)$. Define

$$\lambda_{\mathrm{N}} := \min\{1,\, 3(1-2\alpha)\}\,\frac{\mu^{3/2}}{\rho}, \qquad \gamma_{\mathrm{N}} := \alpha\beta\lambda_{\mathrm{N}}^2.$$

Then every nonterminal iteration satisfies the following dichotomy.

1. *If $\lambda_f(x_t) \ge \lambda_{\mathrm{N}}$, then the accepted damped step decreases the objective by a definite amount:*

$$f(x_{t+1}) \le f(x_t) - \gamma_{\mathrm{N}}.$$

2. *If $\lambda_f(x_t) < \lambda_{\mathrm{N}}$, then backtracking accepts the full Newton step and*

$$\lambda_f(x_{t+1}) \le \frac{\rho}{2\mu^{3/2}}\lambda_f(x_t)^2 \le \frac{1}{2}\lambda_f(x_t).$$

Consequently, the number of damped-phase iterations $\lambda_f(x_t) \ge \lambda_{\mathrm{N}}$ is at most

$$\left\lceil \frac{f(x_0) - p^\star}{\gamma_{\mathrm{N}}} \right\rceil,$$

and after the method enters the regime $\lambda_f(x_t) < \lambda_{\mathrm{N}}$, it takes full Newton steps and the decrement converges quadratically to zero.

Proof of Theorem 14.8. Fix a nonterminal iterate x, and write

$$g := \nabla f(x), \qquad d := d_f(x), \qquad \lambda := \lambda_f(x).$$

By Lemma 14.2, the model-improvement identities give

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

and the lower Hessian bound gives $\|d\|_2 \le \lambda/\sqrt{\mu}$. Applying Lemma 14.9 with $s = td$, we get

$$f(x + td) \le f(x) - t\lambda^2 + \frac{t^2}{2}\lambda^2 + \frac{\rho t^3}{6\mu^{3/2}}\lambda^3, \qquad 0 \le t \le 1.$$

If $\lambda \ge \lambda_{\mathrm{N}}$, take $\widehat{t} = \lambda_{\mathrm{N}}/\lambda$. For any $0 < t \le \widehat{t}$, the Taylor estimate can be written as

$$f(x + td) \le f(x) - t\lambda^2\left(1 - \frac{t}{2} - \frac{\rho t\lambda}{6\mu^{3/2}}\right).$$

Since $t \le 1$ and $t\lambda \le \lambda_{\mathrm{N}}$, the definition

$$\lambda_{\mathrm{N}} \le 3(1-2\alpha)\frac{\mu^{3/2}}{\rho}$$

implies

$$\frac{t}{2} + \frac{\rho t \lambda}{6\mu^{3/2}} \le \frac{1}{2} + \frac{1-2\alpha}{2} = 1 - \alpha.$$

Therefore $f(x+td) \le f(x) - \alpha t\lambda^2 = f(x) + \alpha t g^\top d$, which is exactly the Armijo inequality. The backtracking grid tests $1, \beta, \beta^2, \ldots$; once it crosses $\widehat{t}$, the accepted step satisfies

$$t_{\mathrm{acc}} \ge \beta \widehat{t}.$$

This gives the damped-phase decrease

$$f(x) - f(x^+) \ge \alpha t_{\mathrm{acc}} \lambda^2 \ge \alpha\beta\lambda_{\mathrm{N}}^2 = \gamma_{\mathrm{N}}.$$

If $\lambda < \lambda_{\mathrm{N}}$, the same Taylor estimate with $t = 1$ gives the Armijo inequality, because now $1 \cdot \lambda < \lambda_{\mathrm{N}}$. Hence the full Newton step is accepted. Let $x^+ = x + d$. Since $g + \nabla^2 f(x) d = 0$, the fundamental theorem of calculus gives

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

and hence

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

Since $\nabla^2 f(x^+) \succeq \mu I$, the dual local norm at x^+ satisfies

$$\lambda_f(x^+) = \left\| \nabla f(x^+) \right\|_{x^+,*} \le \frac{1}{\sqrt{\mu}} \left\| \nabla f(x^+) \right\|_2.$$

Combining the last two displays gives

$$\lambda_f(x^+) \le \frac{\rho}{2\mu^{3/2}} \lambda^2 \le \frac{1}{2} \lambda.$$

The bound on the number of damped iterations follows by summing the uniform decrease γ_{N} and using $f \ge p^\star$. □

Remark 14.3 (Cubic regularization as a second-order upper model). If f has a ρ-Lipschitz Hessian, meaning

$$\left\| \nabla^2 f(x) - \nabla^2 f(y) \right\|_2 \le \rho \left\| x - y \right\|_2,$$

then Taylor's theorem gives the cubic upper model

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

Cubic regularization fixes a parameter $M \ge \rho$ and chooses s_t by minimizing the second-order model with a cubic safety term:

$$s_t \in \arg\min_s \left\{ \langle \nabla f(x_t), s \rangle + \frac{1}{2} s^\top \nabla^2 f(x_t) s + \frac{M}{6} \left\| s \right\|_2^3 \right\}, \qquad M \ge \rho.$$

This is the second-order analogue of replacing a function by the smoothness upper bound in gradient descent. The cubic term also makes the model coercive when the Hessian is indefinite, so it is a standard globalization device for Newton-type methods.

The classical reference is [NP06]. In the convex ρ-Hessian-smooth case, if a minimizer $x^\star$ exists and

$$R := \sup\{\|x - x^\star\|_2 : f(x) \le f(x_0)\} < \infty,$$

then the basic cubic regularization method with a fixed parameter $M \ge \rho$ satisfies

$$f(x_T) - f(x^\star) = O\left(\frac{MR^3}{T^2}\right).$$

We record this rate as context only; the proof is not part of this lecture.

Lemma 14.9 (Hessian-Lipschitz Taylor estimate)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, and assume that its Hessian is ρ-Lipschitz:*

$$\left\|\nabla^2 f(y) - \nabla^2 f(x)\right\|_2 \le \rho \|y - x\|_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Then for every $x, s \in \mathbb{R}^n$,

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

In particular, for every direction d and every $t \in [0,1]$,

$$f(x+td) \le f(x) + t\langle \nabla f(x), d\rangle + \frac{t^2}{2} d^\top \nabla^2 f(x) d + \frac{\rho t^3}{6}\|d\|_2^3.$$

Proof of Lemma 14.9. Define $\phi(\theta) := f(x + \theta s)$ for $\theta \in [0,1]$. Then

$$\phi'(0) = \langle \nabla f(x), s\rangle, \qquad \phi''(\theta) = s^\top \nabla^2 f(x+\theta s) s.$$

Using the integral form of Taylor's theorem,

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

Add and subtract $s^\top \nabla^2 f(x) s$ inside the integral:

$$f(x+s) = f(x) + \langle \nabla f(x), s\rangle + \frac{1}{2} s^\top \nabla^2 f(x) s + \int_0^1 (1-\theta) s^\top (\nabla^2 f(x+\theta s) - \nabla^2 f(x)) s \, d\theta.$$

By the Hessian-Lipschitz assumption,

$$s^\top (\nabla^2 f(x+\theta s) - \nabla^2 f(x)) s \le \left\|\nabla^2 f(x+\theta s) - \nabla^2 f(x)\right\|_2 \|s\|_2^2 \le \rho\theta \|s\|_2^3.$$

Therefore the remainder is at most

$$\rho \|s\|_2^3 \int_0^1 \theta(1-\theta)\, d\theta = \frac{\rho}{6}\|s\|_2^3.$$

This proves the first inequality. The second follows by substituting $s = td$. □

Dependency and proof sketch

1. Lemma 14.1 records the second-order test for convexity, and Definitions 14.1 and 14.2 introduce the moving metric used throughout the second-order block. The same notation is used again in Lecture 15 for self-concordant Newton calculus and in Lecture 16 for self-concordant barriers.

2. Lemma 14.2 is the core calculation: Newton is the minimizer of the quadratic Taylor model, and the decrement is exactly the model-improvement scale.

3. Lemma 14.3, Proposition 14.4, and Theorem 14.5 explain why Newton is a genuinely second-order method rather than gradient descent with a clever scalar stepsize: the decrement is the gradient measured in the inverse Hessian geometry, Newton is invariant under affine coordinate changes, and it is exact on positive-definite quadratics.

4. Examples 14.1 and 14.2 and Remark 14.1 translate the formula into linear algebra: a Newton iteration is a structured linear solve, and the structure of that solve matters for runtime.

5. Theorem 14.6 proves the local quadratic regime. The only additional ingredient beyond quadratic exactness is the Lipschitz control of the Hessian, which controls the Taylor remainder.

6. Proposition 14.7, Remarks 14.2 and 14.3, and Theorem 14.8 record the globalized picture and its boundary: positive curvature makes the Newton direction a descent direction, Hessian-Lipschitz control gives a quantitative two-phase theorem with a damped descent phase followed by full-step quadratic convergence of the decrement, and cubic regularization is the parallel model-based globalization route.

References

[NP06] Yurii Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.