

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

1. Use Lipschitz continuity of the Hessian to bound the error in first- and second-order Taylor models.
2. Analyze backtracking Newton using the Newton decrement, identifying a fixed-decrement phase followed by a quadratic-convergence phase.
3. Formulate trust-region methods from local Taylor models and explain why trust regions provide a natural extension of Newton's method to nonconvex problems.

1 Recap: local quadratic convergence

We are motivated by **quadratic convergence** via Newton's method.

Recall 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

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

Theorem 1. 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

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

Question 1. How many Newton iterations for $\|x_k - x_\star\| \leq \epsilon$ if:

(a) Initial point satisfies $\|x_0 - x_\star\| \leq \frac{\mu}{L_2}$?

(b) Function f is exactly quadratic?

Solution: (a) At most

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

(b) At most one iteration.

Question 2. What if $\|x_0 - x_\star\| > \frac{2\mu}{L_2}$ and f is *not* quadratic?

Solution: Then not guaranteed to converge; it may diverge or fail.

Question 3. What if f is not strongly convex, i.e. $\nabla^2 f(x) \not\succeq 0$?

Solution: If $\nabla^2 f(x)$ is not invertible, then Newton immediately fails.

If $\nabla^2 f(x) \not\succeq 0$ but is invertible, then Newton tries to converge towards a saddle point, potentially “up hill”.

2 Backtracking Newton

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

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

where $B = \nabla^2 f(x)^{-1}$. Today we formalize the intuitive discussion from last lecture.

Lemma 2. 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 has two phases:

(Fixed decrement) If $\delta(x) > \eta$, then $f(x_+) \leq f(x) - \gamma$.

(Quadratic convergence) Once $\delta(x) \leq \eta$, then $\delta(x_+) \leq \frac{1}{2\eta} \delta(x)^2$.

Theorem 3. Under the same conditions as above, backtracking Newton converges to an iterate satisfying $\delta(x_k) \leq \epsilon$ in at most

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

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

Con: Fundamentally unable to handle nonconvexity.

2.1 Quadratic convergence phase

Recall that a differentiable function f is said to have L_1 -Lipschitz gradient (i.e. L_1 -smooth) if

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

By analogy, a twice differentiable function is said to have L_2 -Lipschitz Hessian if

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

Lemma 4. If f has L_2 -Lipschitz Hessian, then

$$\begin{aligned} \left| f(x+v) - \left[f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \right] \right| &\leq \frac{L_2}{6} \|v\|_2^3, \\ \|\nabla f(x+v) - [\nabla f(x) + \nabla^2 f(x)v]\|_2 &\leq \frac{L_2}{2} \|v\|_2^2. \end{aligned}$$

Proof: By the fundamental theorem of calculus,

$$\begin{aligned} \nabla f(x+v) - \nabla f(x) &= \int_0^1 \nabla^2 f(x+tv) v \, dt. \\ &= \nabla^2 f(x) v + \int_0^1 [\nabla^2 f(x+tv) - \nabla^2 f(x)] v \, dt, \end{aligned}$$

and

$$\begin{aligned} \left\| \int_0^1 [\nabla^2 f(x+tv) - \nabla^2 f(x)] v \, dt \right\| &\leq \int_0^1 \|\nabla^2 f(x+tv) - \nabla^2 f(x)\| \|v\| \, dt \\ &\leq \int_0^1 L_2 t \|v\|^2 \, dt = L_2 \frac{\|v\|^2}{2} \end{aligned}$$

Similarly,

$$\begin{aligned} f(x+v) - f(x) &= \int_0^1 \nabla f(x+tv)^T v \, dt. \\ &= \nabla f(x)^T v + \int_0^1 [\nabla f(x+tv) - \nabla f(x)]^T v \, dt. \end{aligned}$$

and this error term is bounded

$$\begin{aligned} \int_0^1 v^T [\nabla f(x+tv) - \nabla f(x)] \, dt &= \int_0^1 \int_0^t v^T \nabla^2 f(x+\tau v) v \, d\tau \, dt \\ &= \frac{1}{2} v^T \nabla^2 f(x) v + \int_0^1 \int_0^t v^T [\nabla^2 f(x+\tau v) - \nabla^2 f(x)] v \, d\tau \, dt \end{aligned}$$

where the factor of $1/2$ is because

$$\int_0^1 \int_0^t d\tau dt = \int_0^1 t dt = \frac{1}{2}.$$

Combined,

$$\begin{aligned} & \int_0^1 \int_0^t v^T [\nabla^2 f(x + \tau v) - \nabla^2 f(x)] v \, d\tau dt \\ & \leq \int_0^1 \int_0^t L_2 \tau \|v\|^3 \, d\tau dt = \frac{L_2 \|v\|^3}{6}. \end{aligned}$$

where the factor of $1/6$ is because

$$\int_0^1 \int_0^t \tau \, d\tau dt = \int_0^1 \frac{t^2}{2} dt = \frac{1}{6}.$$

□

Let f be L_2 -Lipschitz Hessian and μ -strongly convex. We define the threshold

$$\eta := \frac{\mu^{3/2}}{L_2} \min \{1, 3(1 - 2c_1)\}.$$

First, we verify that a full stepsize is taken at this threshold.

Lemma 5. Full Newton step. Let $\eta \leq 2\frac{\mu^{3/2}}{L_2}(1 - 2c_1)$. If $\delta(x) \leq \eta$, then the full Newton step $v = -\nabla^2 f(x)^{-1} \nabla f(x)$ with $\alpha = 1$ satisfies the Armijo condition

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

Proof: By $\delta(x) \leq \eta$, we have

$$\frac{L_2}{6\mu^{3/2}} \delta(x)^3 \leq \left(\frac{1}{2} - c_1\right) \delta(x)^2.$$

The second-order Taylor bound gives

$$\begin{aligned} f(x + v) &\leq f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \frac{L_2}{6} \|v\|_2^3 \\ &= f(x) - \frac{1}{2} \delta(x)^2 + \frac{L_2}{6} \|v\|_2^3 \\ &\leq f(x) - \frac{1}{2} \delta(x)^2 + \frac{L_2}{6\mu^{3/2}} \delta(x)^3 \\ &\leq f(x) - \frac{1}{2} \delta(x)^2 + \left(\frac{1}{2} - c_1\right) \delta(x)^2 \end{aligned}$$

which is exactly the Armijo condition with $\alpha = 1$. □

Once the full Newton step is accepted, the Newton decrement contracts quadratically.

Lemma 6. Quadratic convergence. Let $\eta \leq \frac{\mu^{3/2}}{L_2}$. The full Newton step $\alpha = 1$ satisfies

$$\delta(x_+) \leq \frac{1}{2\eta} \delta(x)^2.$$

Proof: By the definition of η , we have

$$\eta \leq \frac{\mu^{3/2}}{L_2}.$$

Strong convexity gives $\nabla^2 f(x)^{-1} \preceq \frac{1}{\mu} I$ and therefore

$$\begin{aligned} \|v\|_2^2 &= \|\nabla^2 f(x)^{-1} \nabla f(x)\|_2^2 = \nabla f(x)^T \nabla^2 f(x)^{-2} \nabla f(x) \\ &\leq \frac{1}{\mu} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = \frac{1}{\mu} \delta(x)^2. \end{aligned}$$

Since $\nabla f(x) + \nabla^2 f(x)v = 0$, the gradient Taylor bound gives

$$\|\nabla f(x_+)\|_2 = \|\nabla f(x+v) - [\nabla f(x) + \nabla^2 f(x)v]\|_2 \leq \frac{L_2}{2} \|v\|_2^2.$$

Strong convexity also gives $\nabla^2 f(x_+)^{-1} \preceq \frac{1}{\mu} I$. Therefore

$$\begin{aligned} \delta(x_+) &= \sqrt{\nabla f(x_+)^T \nabla^2 f(x_+)^{-1} \nabla f(x_+)} \\ &\leq \frac{1}{\sqrt{\mu}} \|\nabla f(x_+)\|_2 \leq \frac{1}{\sqrt{\mu}} \left(\frac{L_2}{2} \|v\|_2^2 \right) \\ &\leq \frac{L_2}{2\sqrt{\mu}} \left(\frac{1}{\mu} \delta(x)^2 \right) = \frac{1}{2} \left(\frac{L_2}{\mu^{3/2}} \right) \delta(x)^2 \leq \frac{1}{2\eta} \delta(x)^2. \end{aligned}$$

□

2.2 Fixed decrement phase

Let $B = B^T \succ 0$ and define $\|v\|_B^2 = v^T B v$ and $\|v\|_{B^{-1}}^2 = v^T B^{-1} v$. Assume throughout that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and satisfies the preconditioned smoothness condition

$$\|\nabla f(x) - \nabla f(y)\|_B \leq L\|x - y\|_{B^{-1}} \quad \text{for all } x, y \in \mathbb{R}^n.$$

Then, preconditioned gradient descent $x_{k+1} = x_k - \alpha B \nabla f(x_k)$ satisfies the descent lemma

$$f(x_{k+1}) \leq f(x_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x_k)\|_B^2. \quad (\text{DL})$$

Theorem 7. HW2 P4-6. Given parameters $c_1, c_2 \in (0, 1)$, we pick stepsize α by testing $\alpha \in \{1, c_2, c_2^2, c_2^3, \dots\}$ until the preconditioned Armijo condition

$$f(x_k - \alpha B \nabla f(x_k)) \leq f(x_k) - c_1 \alpha \|\nabla f(x_k)\|_B^2 \quad (\text{AC})$$

is satisfied. Then, the accepted step satisfies

$$f(x_{k+1}) \leq f(x_k) - c_1 \cdot \min \left\{ 1, \frac{2c_2(1 - c_1)}{L} \right\} \|\nabla f(x_k)\|_B^2.$$

Proof: By (DL), (AC) is guaranteed whenever $\alpha \leq 2(1 - c_1)/L$. If $\alpha = 1$ is not accepted, then the preceding trial α/c_2 was rejected. Because every step at most $2(1 - c_1)/L$ must be accepted, rejection implies

$$\frac{\alpha}{c_2} > \frac{2(1 - c_1)}{L}, \quad \text{so} \quad \alpha > \frac{2c_2(1 - c_1)}{L}.$$

Hence every accepted step obeys

$$\alpha \geq \underline{\alpha} := \min \left\{ 1, \frac{2c_2(1 - c_1)}{L} \right\}.$$

Substitute this lower-bound into (AC) completes the proof. $\square$

Let f be μ -strongly convex and L_1 -smooth. If $B = \nabla^2 f(x)^{-1}$, then preconditioned GD becomes Newton

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

We verify the preconditioned smoothness condition with $L = L_1/\mu$.

Lemma 8.

$$\|\nabla f(y) - \nabla f(z)\|_B \leq \frac{L_1}{\mu} \|y - z\|_{B^{-1}}.$$

Proof: Since $\nabla^2 f(x) \succeq \mu I$, we have $\|v\|_B \leq \frac{1}{\sqrt{\mu}} \|v\|_2$ and $\|v\|_2 \leq \frac{1}{\sqrt{\mu}} \|v\|_{B^{-1}}$. Therefore $\|\nabla f(y) - \nabla f(z)\|_B \leq \frac{1}{\sqrt{\mu}} \|\nabla f(y) - \nabla f(z)\|_2 \leq \frac{L_1}{\sqrt{\mu}} \|y - z\|_2 \leq \frac{L_1}{\mu} \|y - z\|_{B^{-1}}$.

□

Plugging in our homework solution then yields the following lemma.

Lemma 9. Fixed decrement. Let $\eta > 0$ be arbitrary. If $\delta(x) > \eta$, then Newton with backtracking makes the fixed amount of progress

$$f(x_+) \leq f(x) - c_1 \eta^2 \cdot \min \left\{ 1, \frac{2c_2(1 - c_1)\mu}{L_1} \right\}.$$

Proof: We already proved that backtracking line search achieves the following decrement

$$f(x_+) \leq f(x) - c_1 \min \left\{ 1, \frac{2c_2(1 - c_1)}{L} \right\} \|\nabla f(x)\|_B^2.$$

Now, plug in $\|\nabla f(x)\|_B^2 = \delta(x)^2 > \eta^2$.

□

3 Trust-region methods

If f is differentiable, then tangent line is *best* linear model

$$f(y) \approx \underbrace{f(x) + \nabla f(x)^T(y - x)}_{\ell_x(y)} + O(\|y - x\|^2),$$

the only one that matches function value and the first derivative

$$\ell_x(x) = f(x), \quad \nabla \ell_x(x) = \nabla f(x).$$

Question 4. What choice of y minimizes $\ell_x(y)$?

Solution: We have $\arg \min \ell_x(y) = x - \alpha \nabla f(x)$ for $\alpha \rightarrow \infty$.

Problem is that f is not generally linear, so even the best linear model becomes untrustworthy as y gets too far from x .

Trust-region gradient method: Iterate

$$x_+ = \arg \min_{y \in \mathbb{R}^n} \ell_x(y) \quad \text{s.t.} \quad \|y - x\| \leq \Delta$$

where trust-region radius Δ is sized to the “big-O” term.

Question 5. What choice of y minimizes $\ell_x(y)$ within trust region $\|y - x\| \leq \Delta$?

Solution: Solution is just the gradient descent with a particular step-size

$$x_+ = x - \underbrace{\frac{\Delta}{\|\nabla f(x)\|}}_{\alpha} \nabla f(x).$$

So same thing as classic gradient descent.

If f is twice differentiable, then *best* second-order model

$$f(y) \approx f(x) + \nabla f(x)^T(y - x) + \underbrace{\frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)}_{q_x(y)} + O(\|y - x\|^3),$$

only one that matches function value and the two derivatives

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

Again, problem is that f is not generally quadratic, so even best quadratic model becomes untrustworthy as y gets too far from x .

Trust-region Newton method: Iterate

$$x_+ = \arg \min_{y \in \mathbb{R}^n} q_x(y) \quad \text{s.t.} \quad \|y - x\| \leq \Delta$$

where trust-region radius Δ is sized to the “big-O” term.

Lemma 10. Let f be L_2 -Lipschitz Hessian and μ -strongly convex. Define

$$\Delta = \frac{\mu}{L_2}, \quad \eta = \frac{\mu^{3/2}}{L_2} = \sqrt{\mu}\Delta$$

Trust-region Newton with fixed radius Δ has two phases:

(Fixed decrement) If $\delta(x) > \eta$, then $f(x_+) \leq f(x) - \frac{1}{3}\eta^2$.

(Quadratic convergence) Once $\delta(x) \leq \eta$, then $\delta(x_+) \leq \frac{1}{2\eta}\delta(x)^2$.

Theorem 11. Under the same conditions as above, trust-region Newton converges to an iterate satisfying $\delta(x_k) \leq \epsilon$ in at most

$$k \leq \frac{f(x_0) - f(x_*)}{\frac{1}{3}\eta^2} + \log_2 \log_2 \left(\frac{2\eta}{\epsilon} \right) \text{ iterations.}$$

Pro: Completely resolves μ -strongly convex case. Unlike line search, can easily handle nonconvexity.

Con: Trust-region subproblem significantly harder to solve.

4 Summary

Backtracking Newton combines the Newton direction

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

with Armijo backtracking, choosing $\alpha \in \{1, c_2, c_2^2, \dots\}$ until

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

For strongly convex problems, the analysis is naturally expressed using the Newton decrement

$$\delta(x) = \sqrt{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)}.$$

If f has L_1 -Lipschitz gradient, L_2 -Lipschitz Hessian, and is μ -strongly convex, then backtracking Newton has two phases.

When $\delta(x) > \eta$, every iteration makes a fixed amount of objective progress:

$$f(x_+) \leq f(x) - \gamma.$$

This can occur only finitely many times. Once $\delta(x) \leq \eta$, backtracking accepts the full Newton step $\alpha = 1$, and convergence is quadratic

$$\delta(x_+) \leq \frac{1}{2\eta} \delta(x)^2.$$

The total iteration count is

$$O\left(\frac{f(x_0) - f(x_*)}{\gamma} + \log \log \frac{1}{\epsilon}\right).$$

Trust-region methods instead restrict how far the quadratic model may be trusted:

$$x_+ = \arg \min_{\|y-x\| \leq \Delta} \left[f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \nabla^2 f(x) (y-x) \right].$$

For strongly convex problems, a suitable fixed trust-region radius gives the same fixed-decrement followed by quadratic-convergence

picture. Unlike line search, the trust-region subproblem also remains meaningful when the Hessian is indefinite, motivating its extension to nonconvex optimization.