

1 Strong convexity and smoothness constants

Let $f : \Omega \rightarrow \mathbb{R}$ be smooth on $\Omega \subseteq \mathbb{R}^n$. Recall the relationship between minimization and the infimum; the analogous concept for maximization is the supremum $\sup_{x \in \Omega} h(x) = -\inf_{x \in \Omega} (-h(x))$. Define

$$\mu^* = \inf_{x \in \Omega} \lambda_{\min}(\nabla^2 f(x)), \quad L^* = \sup_{x \in \Omega} \|\nabla^2 f(x)\|_{\text{op}}, \quad L_2^* = \sup_{u \neq w} \frac{\|\nabla^2 f(u) - \nabla^2 f(w)\|_{\text{op}}}{\|u - w\|}.$$

Then f is μ -strongly convex if and only if $0 < \mu \leq \mu^*$, ∇f is L -Lipschitz if and only if $L \geq L^*$, and $\nabla^2 f$ is L_2 -Lipschitz if and only if $L_2 \geq L_2^*$. When $n = 1$, $L_2^* = \sup_{x \in \Omega} |f'''(x)|$. For each function below:

- compute μ^* , L^* , and L_2^* ;
- determine whether f is strongly convex, whether ∇f is Lipschitz, and whether $\nabla^2 f$ is Lipschitz.

Rigorously justify each conclusion.

Worked example 1.1

Consider $f(x) = \frac{1}{2}x^2 + \cos x$ for $x \in \mathbb{R}$.

Solution. (a) Here $f''(x) = 1 - \cos x$ and $f'''(x) = \sin x$. Then $\mu^* = \inf_x (1 - \cos x) = 0$, attained at $x = 0$. Next, $L^* = \sup_x (1 - \cos x) = 2$, attained at $x = \pi$. Finally, $L_2^* = \sup_x |\sin x| = 1$. (b) Since $\mu^* = 0$, f is not strongly convex. Since L^* and L_2^* are finite, ∇f and $\nabla^2 f$ are Lipschitz.

Question 1

$f(x) = 2x_1^2 + 2x_1x_2 + 2x_2^2$ for $x \in \mathbb{R}^2$.

Solution

- (a) The Hessian is constant,

$$\nabla^2 f(x) = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix},$$

with eigenvalues 2 and 6. Thus $\mu^* = 2$, $L^* = 6$, and $L_2^* = 0$. (b) f is strongly convex, and ∇f and $\nabla^2 f$ are Lipschitz.

Question 2

$f(x) = \frac{1}{2}(x_1 - x_2)^2$ for $x \in \mathbb{R}^2$.

Solution

- (a) The Hessian is constant,

$$\nabla^2 f(x) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

with eigenvalues 0 and 2. Thus $\mu^* = 0$, $L^* = 2$, and $L_2^* = 0$. (b) f is not strongly convex, and ∇f and $\nabla^2 f$ are Lipschitz.

Question 3

$f(x) = x^4$ for $x \in \mathbb{R}$.

Solution

(a) Here $f''(x) = 12x^2$ and $f'''(x) = 24x$. Thus $\mu^* = \inf_x 12x^2 = 0$, attained at $x = 0$, while $L^* = \sup_x 12x^2 = \infty$ and $L_2^* = \sup_x |24x| = \infty$. (b) f is not strongly convex, and neither ∇f nor $\nabla^2 f$ is Lipschitz.

Question 4

$f(x) = e^x$ for $x \in (-1, 1)$.

Solution

(a) Here $f'' = f''' = e^x$, and $e^{-1} < e^x < e$ on $(-1, 1)$. Thus $\mu^* = e^{-1}$ and $L^* = L_2^* = e$, none attained. (b) f is strongly convex, and ∇f and $\nabla^2 f$ are Lipschitz.

Question 5

$f(x) = \sin x$ for $x \in \mathbb{R}$.

Solution

(a) Here $f''(x) = -\sin x$ and $f'''(x) = -\cos x$. Thus $\mu^* = \inf_x (-\sin x) = -1$, $L^* = \sup_x |\sin x| = 1$, and $L_2^* = \sup_x |\cos x| = 1$. (b) f is not strongly convex, since it is not even convex. Both ∇f and $\nabla^2 f$ are Lipschitz.

Question 6

$f(x) = x^2 + x^4$ for $x \in \mathbb{R}$.

Solution

(a) Here $f''(x) = 2 + 12x^2$ and $f'''(x) = 24x$. Thus $\mu^* = 2$, attained at $x = 0$, while $L^* = \infty$ and $L_2^* = \sup_x |24x| = \infty$. (b) f is strongly convex, and neither ∇f nor $\nabla^2 f$ is Lipschitz.

Question 7

$f(x) = -\log x$ for $x \in (0, 1)$.

Solution

(a) Here $f''(x) = 1/x^2$ and $f'''(x) = -2/x^3$. On $(0, 1)$, $f''(x) > 1$ with $f''(x) \rightarrow 1$ as $x \uparrow 1$, so $\mu^* = 1$, not attained. As $x \downarrow 0$, $L^* = \sup_{(0,1)} 1/x^2 = \infty$ and $L_2^* = \sup_{(0,1)} 2/x^3 = \infty$. (b) f is strongly convex, and neither ∇f nor $\nabla^2 f$ is Lipschitz.

Question 8

(G) $f(x) = -\log(1 - \|x\|^2)$ for $x \in \mathbb{R}^2$ with $\|x\| < 1$.

Solution

(a) Write $\rho = \|x\|^2 < 1$. Then

$$\nabla^2 f(x) = \frac{2}{1-\rho} I + \frac{4}{(1-\rho)^2} x x^T,$$

with eigenvalue $\frac{2}{1-\rho}$ on the directions orthogonal to x and eigenvalue $\frac{2}{1-\rho} + \frac{4\rho}{(1-\rho)^2} = \frac{2(1+\rho)}{(1-\rho)^2}$ on the direction of x . Thus $\lambda_{\min}(\nabla^2 f(x)) = \frac{2}{1-\rho} \geq 2$, with equality at $x = 0$, so $\mu^* = 2$. The largest eigenvalue tends to ∞ as $\|x\| \rightarrow 1$, so $L^* = \infty$. If $\nabla^2 f$ were L_2 -Lipschitz, then every $x \in \Omega$ would satisfy

$$\|\nabla^2 f(x)\|_{\text{op}} \leq \|\nabla^2 f(0)\|_{\text{op}} + L_2 \|x\| \leq 2 + L_2,$$

because Ω is bounded. This contradicts $L^* = \infty$, so $L_2^* = \infty$. (b) f is strongly convex, and neither ∇f nor $\nabla^2 f$ is Lipschitz.

2 Minimizing quadratics

Let $H = H^T \in \mathbb{R}^{n \times n}$ and $g \in \mathbb{R}^n$, and consider

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

For each pair (g, H) below, either find all global minimizers of q , or prove that q is unbounded below by stating a path $x(t)$ such that $q(x(t)) \rightarrow -\infty$ as $t \rightarrow \infty$. Rigorously justify each conclusion.

Question 1

$$g = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, H = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Solution

For any x^* with $Hx^* = -g$ and any x , expanding gives

$$q(x) - q(x^*) = (g + Hx^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T H (x - x^*) = \frac{1}{2} (x - x^*)^T H (x - x^*).$$

So when $H \succeq 0$, every solution of $Hx = -g$ is a global minimizer. Conversely, every global minimizer satisfies $\nabla q(x) = g + Hx = 0$. Here H has eigenvalues 1 and 3, so $H \succ 0$ is invertible, and the unique global minimizer is $x^* = -H^{-1}g = (-1, 1)^T$.

Question 2

$$g = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, H = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}.$$

Solution

With $a = (2, -1)^T$, $H = aa^T \succeq 0$, with eigenvalues 0 and 5. As in the first item, the global minimizers are exactly the solutions of $Hx = -g$. Since $-g = a$, this reads $(a^T x)a = a$, that is, $2x_1 - x_2 = 1$. The global minimizers are exactly $x = (t, 2t - 1)^T$ for $t \in \mathbb{R}$.

Question 3

$$g = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, H = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}.$$

Solution

The eigenvalues of H are 0 and 4, so $H \succeq 0$, and $\text{null}(H) = \text{span}\{(1, 1)^T\}$. The rows of $Hx = -g$ read $2(x_1 - x_2) = -1$ and $-2(x_1 - x_2) = -1$, which is inconsistent, so q has no stationary point and hence no minimizer. Along $x(t) = -t(1, 1)^T$, $Hx(t) = 0$, so $q(x(t)) = g^T x(t) = -2t \rightarrow -\infty$.

Question 4

$$g = \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}, H = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Solution

The eigenvalues of H are 2, 2, and 0, so $H \succeq 0$. As in the first item, the global minimizers are exactly the solutions of $Hx = -g$, which reads $2x_1 = -2$ and $x_2 + x_3 = 1$ (twice). The global minimizers are exactly $x = (-1, t, 1 - t)^T$ for $t \in \mathbb{R}$.

Question 5

$$g = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Solution

Here $q(x) = 2x_1 + x_2 + x_3 + \frac{1}{2}x_1^2 - \frac{1}{2}x_3^2$. Along $x(t) = (0, 0, t)^T$, $q(x(t)) = t - \frac{1}{2}t^2 \rightarrow -\infty$. The null direction also works, since $x(t) = (0, -t, 0)^T$ gives $q(x(t)) = -t$.

Question 6

(G) General $g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ and $H = \text{diag}(\lambda_1, \lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 < 0$.

Solution

Here $q(x) = g_1x_1 + \frac{1}{2}\lambda_1x_1^2 + g_2x_2 + \frac{1}{2}\lambda_2x_2^2$. For every g , the path $x(t) = (0, t)^T$ gives $q(x(t)) = g_2t + \frac{1}{2}\lambda_2t^2 \rightarrow -\infty$ because $\lambda_2 < 0$. So q is unbounded below for every g and every $\lambda_1 \geq 0$.

Question 7

(G) General $g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ and $H = \text{diag}(\lambda_1, 0)$ with $\lambda_1 > 0$.

Solution

Here $q(x) = g_1x_1 + \frac{1}{2}\lambda_1x_1^2 + g_2x_2$ and $H \succeq 0$. If $g_2 \neq 0$, the path $x(t) = (0, -g_2t)^T$ gives $q(x(t)) = -g_2^2t \rightarrow -\infty$. If $g_2 = 0$, then $Hx = -g$ reads $\lambda_1x_1 = -g_1$, and as in the first item the global minimizers are exactly $x = (-g_1/\lambda_1, t)^T$ for $t \in \mathbb{R}$.

3 Exact solution of the cubic-regularization subproblem

Let $g \in \mathbb{R}^n$, $M > 0$, and $d \in \mathbb{R}^n$ with $d_1 \leq \dots \leq d_n$, and write $D = \text{diag}(d)$. The cubic-regularization subproblem minimizes

$$c_x(v) = g^T v + \frac{1}{2} v^T D v + \frac{M}{6} \|v\|^3. \quad (\text{CR})$$

A vector v is globally optimal if and only if some $\lambda \geq 0$ satisfies

$$(D + \lambda I)v = -g, \quad D + \lambda I \succeq 0, \quad \lambda = \frac{M}{2} \|v\|. \quad (\text{C})$$

The subproblem is solved by bisection on the secular equation $\psi(\lambda) = 0$, where

$$\lambda_0 = \max\{0, -d_1\}, \quad \psi(\lambda) = \sum_{i=1}^n \frac{g_i^2}{(d_i + \lambda)^2} - \frac{4\lambda^2}{M^2} \quad \text{for } \lambda > \lambda_0.$$

A root $\lambda_\star > \lambda_0$ gives the global minimizer $v = -(D + \lambda_\star I)^{-1}g$.

Question 1

Reduction to diagonal form. Suppose an oracle $v = \text{SolveCR}(g, d, M)$ returns a global minimizer of (CR). Explain how a single call to $\text{SolveCR}(\cdot, \cdot, \cdot)$ globally solves

$$\min_{v \in \mathbb{R}^n} f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \frac{M}{6} \|v\|^3.$$

Solution

Eigendecompose $\nabla^2 f(x) = Q \text{diag}(d) Q^T$ with Q orthogonal and the eigenvalues sorted so that $d_1 \leq \dots \leq d_n$. Call $u = \text{SolveCR}(Q^T \nabla f(x), d, M)$ and return $v = Qu$.

To see why this works, substitute $v = Qu$. Using $Q^T Q = I$,

$$\nabla f(x)^T Qu = (Q^T \nabla f(x))^T u, \quad (Qu)^T Q \text{diag}(d) Q^T (Qu) = u^T \text{diag}(d) u, \quad \|Qu\| = \|u\|.$$

So the objective at $v = Qu$ equals $f(x)$ plus the (CR) objective with data $(Q^T \nabla f(x), d, M)$. The constant $f(x)$ does not change the minimizers, and $u \mapsto Qu$ is a bijection of $\mathbb{R}^n$. Hence u globally minimizes (CR) if and only if Qu globally minimizes the original objective.

Question 2

Monotonicity. Prove that ψ is strictly decreasing on (λ_0, ∞) .

Solution

On (λ_0, ∞) , every $d_i + \lambda > d_1 + \lambda_0 \geq 0$ and $\lambda > 0$, so

$$\psi'(\lambda) = -2 \sum_{i=1}^n \frac{g_i^2}{(d_i + \lambda)^3} - \frac{8\lambda}{M^2} < 0.$$

Question 3

Easy case. Prove that if $g_1 \neq 0$, then $\psi(\lambda_\star) = 0$ for some $\lambda_\star > \lambda_0$.

Solution

The function ψ is continuous on (λ_0, ∞) , because every $d_i + \lambda > d_1 + \lambda_0 \geq 0$ there. As $\lambda \rightarrow \infty$, the sum tends to 0, so $\psi(\lambda) \rightarrow -\infty$. As $\lambda \downarrow \lambda_0$, ψ becomes positive. If $d_1 \leq 0$, then $d_1 + \lambda \downarrow 0$ and the term $g_1^2/(d_1 + \lambda)^2 \rightarrow +\infty$. If $d_1 > 0$, then $\lambda_0 = 0$ and $\psi(\lambda) \rightarrow \sum_i g_i^2/d_i^2 \geq g_1^2/d_1^2 > 0$. So ψ takes both signs on (λ_0, ∞) , and the intermediate value theorem gives a root $\lambda_\star > \lambda_0$.

Question 4

Hard case. Explain why the bisection method can fail when $\lambda_0 > 0$ and $g_1 = 0$.

Solution

Bisection needs a point $\lambda_2 > \lambda_0$ with $\psi(\lambda_2) > 0$, and none may exist. For example, take $d = (-1, 1)$, $g = (0, 1)^T$, and $M = 1$, so that $\lambda_0 = 1$. Then

$$\psi(\lambda) = \frac{1}{(1 + \lambda)^2} - 4\lambda^2 < \frac{1}{4} - 4 < 0 \quad \text{for all } \lambda > 1,$$

so the secular equation has no root at all.

In general, $\lambda_0 > 0$ means $d_1 = -\lambda_0$. The only terms of ψ that can blow up as $\lambda \downarrow \lambda_0$ are those with $d_i = d_1$, and $g_1 = 0$ removes the first of them. When $d_1 < d_2$, ψ therefore decreases from the finite value

$$\lim_{\lambda \downarrow \lambda_0} \psi(\lambda) = \sum_{i \geq 2} \frac{g_i^2}{(d_i - d_1)^2} - \frac{4\lambda_0^2}{M^2},$$

which is nonpositive whenever $\sum_{i \geq 2} g_i^2/(d_i - d_1)^2 \leq 4\lambda_0^2/M^2$, as in the example. Then $\psi < 0$ on (λ_0, ∞) . Any certificate in (C) with $\lambda > \lambda_0$ would give $v = -(D + \lambda I)^{-1}g$ with $\|v\|^2 = \sum_i g_i^2/(d_i + \lambda)^2 = 4\lambda^2/M^2$, that is, $\psi(\lambda) = 0$. So every certificate has $\lambda = \lambda_0$, where $D + \lambda_0 I$ is singular and the formula $v = -(D + \lambda_\star I)^{-1}g$ does not apply.

Question 5

(G) Bisection. It can be shown that ψ is strictly decreasing on (λ_0, ∞) . Suppose that $\psi(\lambda_\star) = 0$ for some $\lambda_\star > \lambda_0$. Find a choice of λ_1 and λ_2 for which $\psi(\lambda_1) < 0 < \psi(\lambda_2)$, and explain how to use bisection to find $\lambda_\star$.

Solution

Preliminaries. Since ψ is strictly decreasing, $\lambda_\star$ is its only root. Also $g \neq 0$, since $g = 0$ would give $\psi(\lambda_\star) = -4\lambda_\star^2/M^2 < 0$.

Choice of λ_1 . For $\lambda > \lambda_0$, each $d_i + \lambda \geq (d_1 + \lambda_0) + (\lambda - \lambda_0) \geq \lambda - \lambda_0$, so $\psi(\lambda) \leq \|g\|^2/(\lambda - \lambda_0)^2 - 4\lambda^2/M^2$. Take $\lambda_1 = \lambda_0 + \sqrt{M\|g\|}$. Since $\lambda_1^2 \geq M\|g\|$,

$$\psi(\lambda_1) \leq \frac{\|g\|}{M} - \frac{4\lambda_1^2}{M^2} \leq \frac{\|g\|}{M} - \frac{4\|g\|}{M} < 0.$$

Choice of λ_2 . Strict monotonicity gives $\psi(\lambda) > \psi(\lambda_\star) = 0$ for every $\lambda \in (\lambda_0, \lambda_\star)$. The points $\lambda_0 + 2^{-j}(\lambda_1 - \lambda_0)$, $j = 1, 2, \dots$, decrease to $\lambda_0 < \lambda_\star$, so after finitely many halvings one of them lies in $(\lambda_0, \lambda_\star)$. Take λ_2 to be the first such point with $\psi(\lambda_2) > 0$.

Bisection. Now $\lambda_2 < \lambda_\star < \lambda_1$. Evaluate ψ at the midpoint $m = (\lambda_1 + \lambda_2)/2$. If $\psi(m) > 0$, replace λ_2 by m . Otherwise replace λ_1 by m . By monotonicity, each step keeps $\psi(\lambda_1) \leq 0 < \psi(\lambda_2)$, hence $\lambda_\star \in (\lambda_2, \lambda_1]$, and halves $\lambda_1 - \lambda_2$. After k steps, the bracket containing $\lambda_\star$ has width 2^{-k} times its initial width.

Question 6

(G) Hard-case minimizers. Suppose that

$$d_1 < \min\{0, d_2\}, \quad g_1 = 0, \quad \bar{v} = - \left(0, \frac{g_2}{d_2 - d_1}, \dots, \frac{g_n}{d_n - d_1} \right)^T, \quad \|\bar{v}\| \leq -\frac{2d_1}{M}.$$

Prove that $\bar{v} \pm \tau e_1$ with $\tau = \sqrt{4d_1^2/M^2 - \|\bar{v}\|^2}$ are global minimizers of (CR). The boundary case $\|\bar{v}\| = 2|d_1|/M$ gives $\tau = 0$.

Solution

Let $\lambda_0 = -d_1 > 0$ and $v = \bar{v} \pm \tau e_1$. We verify (C) with $\lambda = \lambda_0$. First, $D + \lambda_0 I = \text{diag}(0, d_2 - d_1, \dots, d_n - d_1) \succeq 0$. Second, the first row of $(D + \lambda_0 I)v$ is $0 = -g_1$, and row $i \geq 2$ is $(d_i - d_1)\bar{v}_i = -g_i$. Third, $\bar{v}_1 = 0$ makes $\bar{v} \perp e_1$, so $\|v\|^2 = \|\bar{v}\|^2 + \tau^2 = 4d_1^2/M^2$ and $\|v\| = 2\lambda_0/M$. By (C), v is a global minimizer.

4 Newton convergence analysis

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, and assume that for all $x, u, w \in \mathbb{R}^n$,

$$\mu I \preceq \nabla^2 f(x) \preceq L_1 I, \quad \|\nabla^2 f(u) - \nabla^2 f(w)\|_{\text{op}} \leq L_2 \|u - w\|, \quad (\text{A})$$

with $0 < \mu \leq L_1$ and $L_2 > 0$. Then f has a unique minimizer x^* , and $f^* = f(x^*)$. At an iterate x , define the Newton direction and the Newton decrement

$$v = -\nabla^2 f(x)^{-1} \nabla f(x), \quad \delta^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = -\nabla f(x)^T v = v^T \nabla^2 f(x) v. \quad (\text{N})$$

Backtracking Newton sets $x_+ = x + \alpha v$, where α is the first element of $\{1, c_2, c_2^2, \dots\}$ satisfying

$$f(x + \alpha v) \leq f(x) + c_1 \alpha \nabla f(x)^T v, \quad (\text{Armijo})$$

with $0 < c_1 < 1/2$ and $0 < c_2 < 1$.

Question 1

Gradient remainder. Starting from

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

prove that for all $x, d \in \mathbb{R}^n$,

$$\|\nabla f(x + d) - \nabla f(x) - \nabla^2 f(x) d\| \leq \frac{L_2}{2} \|d\|^2.$$

Solution

Subtracting $\nabla^2 f(x) d = \int_0^1 \nabla^2 f(x) d \, dt$ gives

$$\nabla f(x + d) - \nabla f(x) - \nabla^2 f(x) d = \int_0^1 [\nabla^2 f(x + td) - \nabla^2 f(x)] d \, dt.$$

By (A), the integrand has norm at most $L_2 t \|d\|^2$, and $\int_0^1 t \, dt = \frac{1}{2}$.

Question 2

Function remainder. Derive the identity

$$f(x + d) = f(x) + \nabla f(x)^T d + \int_0^1 (1 - t) d^T \nabla^2 f(x + td) d \, dt,$$

and use it to prove that for all $x, d \in \mathbb{R}^n$,

$$\left| f(x + d) - f(x) - \nabla f(x)^T d - \frac{1}{2} d^T \nabla^2 f(x) d \right| \leq \frac{L_2}{6} \|d\|^3.$$

Solution

Let $\phi(t) = f(x + td)$, so $\phi'(t) = \nabla f(x + td)^T d$ and $\phi''(t) = d^T \nabla^2 f(x + td) d$. Integration by parts gives

$$\phi(1) - \phi(0) = \int_0^1 \phi'(t) \, dt = [(t - 1)\phi'(t)]_0^1 + \int_0^1 (1 - t)\phi''(t) \, dt = \phi'(0) + \int_0^1 (1 - t)\phi''(t) \, dt,$$

which is the identity. Since $\int_0^1 (1-t) dt = \frac{1}{2}$, subtracting the quadratic model leaves $\int_0^1 (1-t) d^T [\nabla^2 f(x+td) - \nabla^2 f(x)] d dt$. By (A), its absolute value is at most $L_2 \|d\|^3 \int_0^1 t(1-t) dt = \frac{L_2}{6} \|d\|^3$.

Question 3

Local quadratic convergence. Starting from

$$e = x - x^*, \quad e_+ = x + v - x^*, \quad \nabla f(x) = \int_0^1 \nabla^2 f(x^* + te) e dt,$$

prove that $\|e_+\| \leq \frac{L_2}{2\mu} \|e\|^2$. Explain why one full Newton step minimizes a strictly convex quadratic exactly.

Solution

By (N), $\nabla^2 f(x)e_+ = \nabla^2 f(x)e + \nabla^2 f(x)v = \nabla^2 f(x)e - \nabla f(x)$, so

$$\nabla^2 f(x)e_+ = \int_0^1 [\nabla^2 f(x) - \nabla^2 f(x^* + te)] e dt.$$

By (A), $\|\nabla^2 f(x)e_+\| \geq \mu \|e_+\|$, and the integrand has norm at most $L_2 \|x - x^* - te\| \|e\| = L_2(1-t) \|e\|^2$. Hence $\mu \|e_+\| \leq L_2 \|e\|^2 \int_0^1 (1-t) dt = \frac{L_2}{2} \|e\|^2$. For a strictly convex quadratic, the Hessian is constant, so the integrand vanishes. Then $\nabla^2 f(x)e_+ = 0$ and $e_+ = 0$.

Question 4

Iteration complexity. Let δ_k denote the Newton decrement at x_k . Suppose that constants $\eta, \gamma > 0$ satisfy, for every k ,

- if $\delta_k > \eta$, then $f(x_{k+1}) \leq f(x_k) - \gamma$;
- if $\delta_k \leq \eta$, then $\delta_{k+1} \leq \delta_k^2/(2\eta)$.

Prove that for $0 < \epsilon < \eta$, backtracking Newton reaches $\delta_k \leq \epsilon$ within

$$\left\lceil \frac{f(x_0) - f^*}{\gamma} + \log_2 \log_2 \left(\frac{2\eta}{\epsilon} \right) \right\rceil$$

iterations.

Solution

Once $\delta_k \leq \eta$, the second case gives $\delta_{k+1} \leq \eta/2$, so every later iterate stays in the second case. The iterations with $\delta_k > \eta$ therefore come first. If there are m of them, the first case gives $f^* \leq f(x_m) \leq f(x_0) - m\gamma$, so $m \leq \lfloor A \rfloor$ with $A = (f(x_0) - f^*)/\gamma$.

Let k be the first index with $\delta_k \leq \eta$, and $a_j = \delta_{k+j}/(2\eta)$. Then $a_0 \leq \frac{1}{2}$ and $a_{j+1} \leq a_j^2$, so $a_j \leq (\frac{1}{2})^{2^j}$ by induction. This is at most $\epsilon/(2\eta)$ once $2^j \geq \log_2(2\eta/\epsilon)$, that is, once $j \geq B := \log_2 \log_2(2\eta/\epsilon)$, which is positive because $\epsilon < \eta$. So the second case needs at most $\lceil B \rceil$ iterations. Since $\lfloor A \rfloor + \lceil B \rceil$ is an integer smaller than $A + B + 1$, it is at most $\lceil A + B \rceil$.

The second case holds for $\eta = \frac{\mu^{3/2}}{L_2} \min\{1, 3(1 - 2c_1)\}$ by full-step acceptance and decrement contraction. The first case holds with $\gamma = c_1\eta^2 \min\{1, 2c_2(1 - c_1)\mu/L_1\}$ by the Homework 2 backtracking bound in the metric $\nabla^2 f(x)$.

Question 5

(G) Full-step acceptance. Starting from

$$f(x+v) \leq f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \frac{L_2}{6} \|v\|^3,$$

prove that

$$f(x+v) \leq f(x) - \frac{1}{2} \delta^2 + \frac{L_2}{6\mu^{3/2}} \delta^3,$$

and conclude that $\alpha = 1$ passes (Armijo) whenever $\delta \leq 3(1 - 2c_1)\mu^{3/2}/L_2$.

Solution

By (N), $\nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v = -\delta^2 + \frac{1}{2} \delta^2 = -\frac{1}{2} \delta^2$. Also $\delta^2 = v^T \nabla^2 f(x) v \geq \mu \|v\|^2$, so $\|v\| \leq \delta/\sqrt{\mu}$. Substituting both proves the first claim. The cubic term is at most $(\frac{1}{2} - c_1)\delta^2$ exactly when $\delta \leq 6(\frac{1}{2} - c_1)\mu^{3/2}/L_2 = 3(1 - 2c_1)\mu^{3/2}/L_2$. Then $f(x+v) \leq f(x) - c_1 \delta^2 = f(x) + c_1 \nabla f(x)^T v$, which is (Armijo) with $\alpha = 1$.

Question 6

(G) Gradient after a full step. Starting from

$$x_+ = x + v, \quad \|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\| \leq \frac{L_2}{2} \|d\|^2 \quad \text{for all } d,$$

prove that $\|\nabla f(x_+)\| \leq \frac{L_2}{2\mu} \delta^2$.

Solution

Take $d = v$. Since $\nabla f(x) + \nabla^2 f(x)v = 0$ by (N), the given bound reads $\|\nabla f(x_+)\| \leq \frac{L_2}{2} \|v\|^2$. By (A) and (N), $\delta^2 = v^T \nabla^2 f(x) v \geq \mu \|v\|^2$, so $\|v\|^2 \leq \delta^2/\mu$.

Question 7

(G) Decrement contraction. Starting from

$$\delta_+^2 = \nabla f(x_+)^T \nabla^2 f(x_+)^{-1} \nabla f(x_+), \quad \|\nabla f(x_+)\| \leq \frac{L_2}{2\mu} \delta^2,$$

prove that $\delta_+ \leq \frac{L_2}{2\mu^{3/2}} \delta^2$, and conclude that $\delta_+ \leq \delta^2/(2\eta)$ for every $0 < \eta \leq \mu^{3/2}/L_2$.

Solution

By (A), $\nabla^2 f(x_+) \succeq \mu I$, so $\nabla^2 f(x_+)^{-1} \preceq I/\mu$ and $\delta_+^2 \leq \|\nabla f(x_+)\|^2/\mu$. Hence $\delta_+ \leq \|\nabla f(x_+)\|/\sqrt{\mu} \leq \frac{L_2}{2\mu^{3/2}} \delta^2$. If $\eta \leq \mu^{3/2}/L_2$, then $\frac{L_2}{2\mu^{3/2}} \leq \frac{1}{2\eta}$.

5 Cubic-regularized Newton convergence analysis

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable with $f_{\inf} = \inf_x f(x) > -\infty$, and assume that for all $u, w \in \mathbb{R}^n$,

$$\|\nabla^2 f(u) - \nabla^2 f(w)\|_{\text{op}} \leq L_2 \|u - w\|. \quad (\text{A})$$

At an iterate x , cubic-regularized Newton sets $x_+ = x + v$, where v globally minimizes the model

$$c_x(v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \frac{L_2}{6} \|v\|^3. \quad (\text{CN})$$

The minimizer satisfies the first- and second-order necessary conditions

$$\nabla f(x) + \nabla^2 f(x)v + \frac{L_2}{2} \|v\| v = 0, \quad \nabla^2 f(x) + \frac{L_2}{2} \|v\| \left(I + \frac{vv^T}{v^T v} \right) \succeq 0, \quad (\text{O})$$

where $vv^T/v^T v$ is read as 0 when $v = 0$.

Question 1

Decrease. Starting from

$$f(x_+) \leq c_x(v), \quad c_x(v) \leq c_x(-v),$$

prove that $\nabla f(x)^T v \leq 0$ and

$$f(x_+) \leq f(x) - \frac{L_2}{12} \|v\|^3.$$

Solution

The even-degree terms of c_x cancel in $c_x(-v) - c_x(v) = -2\nabla f(x)^T v$, so $\nabla f(x)^T v \leq 0$. Since $c_x(0) = f(x)$,

$$f(x) - f(x_+) \geq c_x(0) - c_x(v) = -\nabla f(x)^T v - \frac{1}{2} v^T \nabla^2 f(x) v - \frac{L_2}{6} \|v\|^3.$$

The inner product of the first condition in (O) with v gives $\frac{1}{2} v^T \nabla^2 f(x) v = -\frac{1}{2} \nabla f(x)^T v - \frac{L_2}{4} \|v\|^3$. Substituting,

$$f(x) - f(x_+) \geq -\frac{1}{2} \nabla f(x)^T v + \left(\frac{L_2}{4} - \frac{L_2}{6} \right) \|v\|^3 \geq \frac{L_2}{12} \|v\|^3.$$

Question 2

Gradient. Starting from

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\| \leq \frac{L_2}{2} \|v\|^2,$$

prove that $\|\nabla f(x_+)\| \leq L_2 \|v\|^2$.

Solution

By the first condition in (O), $\nabla f(x) + \nabla^2 f(x)v = -\frac{L_2}{2} \|v\| v$. The triangle inequality and the given bound give

$$\|\nabla f(x+v)\| \leq \|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\| + \|\nabla f(x) + \nabla^2 f(x)v\| \leq \frac{L_2}{2} \|v\|^2 + \frac{L_2}{2} \|v\|^2.$$

Question 3

Curvature. Using (A) and (O), prove that $\nabla^2 f(x) + L_2 \|v\| I \succeq 0$ and

$$\nabla^2 f(x_+) \succeq -2L_2 \|v\| I.$$

Solution

Since $vv^T/v^T v \preceq I$, the second condition in (O) gives

$$0 \preceq \nabla^2 f(x) + \frac{L_2}{2} \|v\| \left(I + \frac{vv^T}{v^T v} \right) \preceq \nabla^2 f(x) + L_2 \|v\| I.$$

By (A), every u satisfies

$$u^T [\nabla^2 f(x+v) - \nabla^2 f(x)] u \geq -\|\nabla^2 f(x+v) - \nabla^2 f(x)\|_{\text{op}} \|u\|^2 \geq -L_2 \|v\| \|u\|^2.$$

Hence $\nabla^2 f(x_+) \succeq \nabla^2 f(x) - L_2 \|v\| I \succeq -2L_2 \|v\| I$.

Question 4

Escaping saddle points. Starting from

$$\nabla f(x) = 0, \quad \lambda_{\min}(\nabla^2 f(x)) < 0, \quad f(x_+) \leq f(x) - \frac{L_2}{12} \|v\|^3,$$

prove that $v \neq 0$ and $f(x_+) < f(x)$. Explain why backtracking Newton cannot make progress from the same point.

Solution

If $v = 0$, the second condition in (O) reads $\nabla^2 f(x) \succeq 0$, which contradicts $\lambda_{\min}(\nabla^2 f(x)) < 0$. Hence $v \neq 0$, and the given bound yields $f(x_+) \leq f(x) - \frac{L_2}{12} \|v\|^3 < f(x)$. The Newton direction $-\nabla^2 f(x)^{-1} \nabla f(x)$ is 0 when $\nabla^2 f(x)$ is invertible and undefined otherwise. Either way, backtracking Newton stays at x .

Question 5

(G) Iteration complexity. Suppose that the iterates $x_{i+1} = x_i + v_i$ satisfy, for every $i \geq 0$,

$$f(x_{i+1}) \leq f(x_i) - \frac{L_2}{12} \|v_i\|^3, \quad \|\nabla f(x_{i+1})\| \leq L_2 \|v_i\|^2, \quad \nabla^2 f(x_{i+1}) \succeq -2L_2 \|v_i\| I.$$

Let $D = f(x_0) - f_{\text{inf}}$. Prove that for every $k \geq 1$, some $x_{\text{best}} \in \{x_1, \dots, x_k\}$ satisfies

$$\|\nabla f(x_{\text{best}})\| \leq \left(\frac{12L_2^{1/2} D}{k} \right)^{2/3}, \quad \lambda_{\min}(\nabla^2 f(x_{\text{best}})) \geq -2 \left(\frac{12L_2^2 D}{k} \right)^{1/3}.$$

Solution

Summing the first inequality over $i = 0, \dots, k-1$ telescopes:

$$\frac{L_2}{12} \sum_{i=0}^{k-1} \|v_i\|^3 \leq f(x_0) - f(x_k) \leq D.$$

The shortest step v_j is no longer than the average, so $\|v_j\|^3 \leq \frac{1}{k} \sum_i \|v_i\|^3 \leq \frac{12D}{L_2 k}$. Take $x_{\text{best}} = x_{j+1}$. By the second inequality,

$$\|\nabla f(x_{\text{best}})\| \leq L_2 \|v_j\|^2 \leq L_2 \left(\frac{12D}{L_2 k} \right)^{2/3} = \left(\frac{12L_2^{1/2} D}{k} \right)^{2/3}.$$

By the third inequality,

$$\lambda_{\min}(\nabla^2 f(x_{\text{best}})) \geq -2L_2 \|v_j\| \geq -2L_2 \left(\frac{12D}{L_2 k} \right)^{1/3} = -2 \left(\frac{12L_2^2 D}{k} \right)^{1/3}.$$