IB Optimisation: Lecture 2

Mike Tehranchi

University of Cambridge

27 April 2020

Definition

Given a convex set $X \subseteq \mathbb{R}^n$, a function $f : X \to \mathbb{R}$ is *strictly convex* if for every $x, y \in X$ where $x \neq y$ and every number $0 < p < 1$ we have

$$f(px + (1-p)y) < pf(x) + (1-p)f(y)$$

Definition

An $n \times n$ matrix A is *positive definite* if

$$x^\top A x > 0$$

for all $x \in \mathbb{R}^n$ such that $x \neq 0$.

Theorem

Suppose $f : X \to \mathbb{R}$ is twice-differentiable. If $D^2 f(x)$ is positive definite for all $x \in X$, then f is strictly convex.

Proof. Example sheet.

Theorem (Uniqueness of optimal solutions)

Suppose x^* and y^* are optimal solutions to the problem

$$\text{minimise } f(x) \text{ subject to } x \in X.$$

If f is strictly convex, then $x^* = y^*$.

Proof. By definition

$$f(x^*) = f(y^*) \leq f(z)$$

for all $z \in X$. Suppose for the sake of finding a contradiction that $x^* \neq y^*$. Now $z = \frac{1}{2}(x^* + y^*)$. Note that $z \in X$ by the convexity of X, and by the strict convexity of f that

$$f(z) < \frac{1}{2}f(x^*) + \frac{1}{2}f(y^*) = f(x^*) = f(y^*),$$

a contradiction. □

Definition

A function $f : X \to \mathbb{R}$ is *strongly convex* if there exists a constant $m > 0$ such that the function $x \mapsto f(x) - \frac{m}{2}\|x\|^2$ is convex.

Notation. Here $\|z\| = \sqrt{z^\top z}$ is the usual Euclidean norm on $\mathbb{R}^n$. In the next slide we will let I be the $n \times n$ identity matrix.

Theorem

Suppose f is twice differentiable. Then f is strongly convex if there exists $m > 0$ such that for all $x \in X$ the matrix

$$D^2 f(x) - mI$$

is non-negative definite, or equivalently,

$$z^\top D^2 f(x) z \geq m\|z\|^2$$

for all $z \in \mathbb{R}^n$.

Notation. There is a natural partial order on the set of symmetric matrices. We write

$$B \succeq A$$

if $B - A$ is non-negative definite.
The hypothesis of the theorem can be rewritten $D^2 f(x) \succeq mI$.

Proof. Note $D^2\|x\|^2 = 2I$, and apply the definition. □

strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex

Theorem (Existence of an optimal solution)

Suppose $X \subseteq \mathbb{R}^n$ is closed and that f is continuous and strongly convex. Then there exists an optimal solution to the problem

$$\text{minimise } f(x) \text{ subject to } x \in X.$$

Proof. Let $g(x) = f(x) - \frac{m}{2}\|x\|^2$ where $m > 0$, and assume that g is convex. By the supporting hyperplane theorem, there is a vector $\lambda \in \mathbb{R}^n$ such that

$$\begin{aligned} g(x) &\geq g(0) + \lambda^\top x \\ &\geq g(0) - \|\lambda\|\|x\| \end{aligned}$$

by Cauchy–Schwarz. Hence for $\|x\| > R = 2\|\lambda\|/m$ we have

$$f(x) \geq f(0) - \|\lambda\|\|x\| + \frac{m}{2}\|x\|^2 > f(0)$$

Our problem becomes

$$\text{minimise } f(x) \text{ subject to } \; x \in X, \; \|x\| \leq R$$

From analysis, the continuous function f attains its minimum on the compact set $\{x \in X : \|x\| \leq R\}$, showing that there exists an optimal solution x^*. □

Theorem (Gradient lower bound)

Suppose $f : X \to \mathbb{R}$ is differentiable and strongly convex with constant $m > 0$. Then

$$\|Df(x)\|^2 \geq 2m(f(x) - f(y))$$

for any $x, y \in X$.

Proof. Applying the supporting hyperplane theorem from Lecture 1 to the convex function $g(x) = f(x) - \frac{m}{2}\|x\|^2$ yields

$$f(y) - f(x) \geq (y - x)^\top Df(x) + \frac{m}{2}\|y - x\|^2$$

Now, note that by completing the square, we have any $b, z \in \mathbb{R}^n$ that

$$b^\top z + \frac{m}{2}\|z\|^2 \geq -\frac{\|b\|^2}{2m}.$$

Combining these inequalities yields

$$f(y) - f(x) \geq -\frac{\|Df(x)\|^2}{2m}$$

The conclusion follows upon rearranging. □

For the rest of the lecture, we let $X = \mathbb{R}^n$, and focus on methods for computing an optimal solution.

Motivation.

- Suppose f is differentiable.
- The rate of change of f at the point $x \in \mathbb{R}^n$ in direction $u \in \mathbb{R}^n$ is

$$\lim_{t \to 0} \frac{f(x+tu) - f(x)}{t} = u^\top Df(x)$$

- By the Cauchy–Schwarz inequality, the rate of descent is steepest when u is pointing in the direction of $-Df(x)$.

Gradient descent algorithm

- Start with an initial guess $x_0 \in \mathbb{R}^n$
- Pick a step size $t > 0$
- For every $k \geq 0$, let

$$x_{k+1} = x_k - tDf(x_k)$$

Question: In what sense, if any, does the sequence $(x_k)_k$ converge to an optimal solution x^* ?

Theorem (Rate of convergence of gradient descent)

Suppose f is twice-differentiable and that there are constants $0 < m < M$ such that

$$mI \preceq D^2 f(x) \preceq MI$$

for all $x \in \mathbb{R}^n$. Applying the gradient descent algorithm with step size $t = 1/M$ we have

$$f(x_k) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^k (f(x_0) - f(x^*))$$

Proof. Fix $x, y \in \mathbb{R}^n$. By Taylor's theorem there is a $0 < p < 1$ such that for $\xi = px + (1-p)y$ we have

$$\begin{aligned} f(y) &= f(x) + (y-x)^\top Df(x) + \frac{1}{2}(y-x)^\top D^2 f(\xi)(y-x) \\ &\leq f(x) + (y-x)^\top Df(x) + \frac{M}{2}\|y-x\|^2. \end{aligned}$$

since $D^2 f(\xi) \preceq MI$.

Hence

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq (x_{k+1} - x_k)^\top Df(x_k) + \frac{M}{2}\|x_{k+1} - x_k\|^2 \\ &= \left(-t + \frac{M}{2}t^2\right)\|Df(x_k)\|^2 \\ &= -\frac{1}{2M}\|Df(x_k)\|^2 \\ &\leq -\frac{m}{M}(f(x_k) - f(x^*)). \end{aligned}$$

This shows

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{m}{M}\right)(f(x_k) - f(x^*)).$$

and the conclusion follows from induction. □

Motivation.

- Suppose f is twice continuously differentiable.
- Let x_0 be an intitial guess for the optimiser.
- By Taylor's theorem we have

$$f(x) \approx f(x_0) + (x - x_0)^\top Df(x_0) + \frac{1}{2}(x - x_0)^\top D^2 f(x_0)(x - x_0)$$

- Minimising the quadratic on the right (for instance, by completing the square) yields the approximation

$$x^* \approx x_0 - (D^2 f(x_0))^{-1} Df(x_0)$$

Newton's method

- Start with an initial guess $x_0 \in \mathbb{R}^n$
- For every $k \geq 0$, let

$$x_{k+1} = x_k - (D^2 f(x_k))^{-1} Df(x_k)$$

Notation. A matrix norm. If A is $n \times n$, let $\|A\|$ be the smallest constant $a \geq 0$ such that

$$\|Az\| \leq a\|z\|$$

for all $z \in \mathbb{R}^n$. If A is non-negative definite, then $\|A\|$ is the largest eigenvalue of A.

Theorem (Rate of convergence of Newton's method)

Suppose f is twice-differentiable and that there are constants $m, L > 0$ such that

$$D^2 f(x) \succeq mI$$

and

$$\|D^2 f(x) - D^2 f(y)\| \leq L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$. Applying Newton's method we have

$$f(x_k) - f(x^*) \leq \frac{2m^3}{L^2} \left(\frac{L}{2m^2} \|Df(x_0)\| \right)^{2^{k+1}}$$

Proof.[Non-examinable] Letting $\Delta x_k = x_{k+1} - x_k$ we have

$$\begin{aligned} Df(x_{k+1}) &= Df(x_{k+1}) - Df(x_k) - D^2 f(x_k)\Delta x_k \\ &= \int_0^1 [D^2 f(x_k + t\Delta x_k) - D^2(x_k)]\Delta x_k \, dt \end{aligned}$$

by the fundamental theorem of calculus.

By the triangle inequality applied to the integral, we have

$$\begin{aligned}\|Df(x_{k+1})\| &\leq \int_0^1 \|[D^2 f(x_k + t\Delta x_k) - D^2 f(x_k)]\Delta x_k\| dt \\ &\leq L\|\Delta x_k\|^2 \int_0^1 t \ dt \\ &= \frac{1}{2} L\|(D^2 f(x_k))^{-1} Df(x_k)\|^2 \\ &\leq \frac{L}{2m^2}\|Df(x_k)\|^2\end{aligned}$$

where we have used the fact that if $A \succeq mI$ then $\|A^{-1}\| \leq 1/m$.

By induction

$$\|Df(x_k)\| \leq \frac{2m^2}{L}\left(\frac{L}{2m^2}\|Df(x_0)\|\right)^{2^k}$$

By the conclusion follows from the lower bound on the gradient. □