Lecture 5: Minima of Convex Functions; Algorithmic Setup

Yudong Chen

1 Minima of convex functions

Consider the constrained problem

$$\min_{x \in \mathcal{X}} f(x). \tag{P}$$

Recall definition of convex functions.

Theorem 1. *Consider the problem (P). Suppose f is convex, and $\mathcal{X}$ is convex, closed and non-empty. Then:*

1. *Any local solution to (P) is also a global solution.*
2. *The set of global solutions to (P) is convex.*

Proof. Part 1: Suppose f.p.o.c. that x^* is a local but not a global solution. Then there exists $\bar{x} \in \mathcal{X}$ such that $f(\bar{x}) < f(x^*)$. As $\mathcal{X}$ is convex, for all $\alpha \in (0,1)$,

$$(1-\alpha)x^* + \alpha\bar{x} \in \mathcal{X}.$$

As f is convex, for all $\alpha \in (0,1)$:

$$f\left((1-\alpha)x^* + \alpha\bar{x}\right) \le (1-\alpha)f(x^*) + \alpha f(\bar{x}) < f(x^*).$$

Hence every neighborhood of x^* must include a point $(1-\alpha)x^* + \alpha\bar{x}$ for some $\alpha > 0$ that will have a strictly lower function value. So x^* cannot be a local solution, a contradiction.

Part 2: Let $x^*, \bar{x} \in \mathcal{X}$ be any two global solutions.
$\mathcal{X}$ is convex $\implies$ $(1-\alpha)x^* + \alpha\bar{x} \in \mathcal{X}$.
f is convex $\implies$

$$f\left((1-\alpha)x^* + \alpha\bar{x}\right) \le (1-\alpha)f(x^*) + \alpha f(\bar{x}) = f(x^*) = f(\bar{x})$$

$\implies$ $f\left((1-\alpha)x^* + \alpha\bar{x}\right) = f(x^*)$, so $(1-\alpha)x^* + \alpha\bar{x}$ is also a global solution $\implies$ the set of global solution is convex. □

1.1 Differentiable convex functions

Theorem 2 (Equivalent characterization of convexity)**.** *The following are true.*

1. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be continuously differentiable. The function f is convex if and only if*

$$\forall x, y : f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle. \tag{1}$$

(A picture. From local to global.)

2. *Let* $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ *be twice continuously differentiable. The function* f *is convex if and only if*

$$\forall x : \nabla^2 f(x) \succcurlyeq 0.$$

Proof. Part 1, "only if": By convexity of f, for any $\alpha \in (0,1)$:

$$\begin{aligned} & f\left((1-\alpha)x + \alpha y\right) \le (1-\alpha) f(x) + \alpha f(y) \\ \overset{\text{rearranging}}{\Longrightarrow} & f(y) - (x) \ge \frac{f\left(x + \alpha(y-x)\right) - f(x)}{\alpha} \overset{\text{Taylor's}}{=} \frac{\langle \nabla f(x), \alpha(y-x)\rangle + o(\alpha)}{\alpha}. \end{aligned}$$

Taking $\alpha \to 0$ gives (1)

Part 1, "if": Take any x, y and $\alpha \in (0,1)$. Set $z = (1-\alpha)x + \alpha y$. Apply (1) to x, z and y, z:

$$f(x) \ge f(z) + \alpha \left\langle \nabla f(z), x - y \right\rangle, \tag{2}$$

$$f(y) \ge f(z) + (1-\alpha) \left\langle \nabla f(z), y - x \right\rangle. \tag{3}$$

(2)$\times(1-\alpha)$+(3)$\times\alpha$ gives

$$(1-\alpha) f(x) + \alpha f(y) \ge f(z),$$

which implies convexity of f.

Part 2: See Wright-Recht, Lemma 2.9. □

Theorem 3 (Sufficient condition for global optimality)**.** *Consider the problem (P), where* f *is continuously differentiable and convex. If* $x^* \in \mathcal{X}$ *and* $\nabla f(x^*) = 0$*, then* x^* *is a global minimizer of* f*.*

Proof. Use Part 1 of Theorem 2:

$$\forall x : f(x) \ge f(x^*) + \langle \cancel{\nabla f(x^*)}, x - x^* \rangle = f(x^*).$$

□

Remark 1. Theorem 3 holds for both unconstrained (i.e., $\mathcal{X} = \mathbb{R}^d$) and constrained problems. Using terminology from last time, x^* being a stationary point is sufficient for global optimality. For unconstrained problem, this is also necessary (Lecture 4, Theorem 1). For constrained problem, this may not be necessary (example).

2 Strongly convex functions

We use Euclidean norm $\|\cdot\|_2$ in this section.

Definition 1 (Strong convexity)**.** Given $m > 0$, we say that $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is *strongly convex* with modulus/parameter m (or m*-strongly convex for short*), if

$$\forall x, y \in \mathbb{R}^d : f\left((1-\alpha)x + \alpha y\right) \le (1-\alpha) f(x) + \alpha f(y) - \frac{m}{2}(1-\alpha)\alpha \left\|y - x\right\|_2^2.$$

Remark 2. Verify yourself that the above is equivalent to convexity of the function $f(x) - \frac{m}{2}\|x\|_2^2$.

Theorem 4 (Equivalent characterization of strong convexity)**.** *The following hold.*

1. *Suppose f is continuously differentiable. Then f is m-strong convexity if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|_2^2.$$

(A picture. Compare with convexity only. Complements L-smoothness.)

2. *Suppose f is twice continuously differentiable. Then f is m-strong convexity if and only if*

$$\forall x : \nabla^2 f(x) \succcurlyeq mI.$$

(Compare with L-smoothness)

Proof. Apply Theorem 2 to the function $f(x) - \frac{m}{2} \|x\|_2^2$. □

Theorem 5. *Suppose that $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is continuously differentiable and m-strongly convex for some $m > 0$. If $x^* \in \mathcal{X}$ satisfies $\nabla f(x^*) = 0$, then x^* is the* <u>*unique*</u> *global minimizer of f.*

Proof. By Part 1 of Theorem 4:

$$f(x) \geq f(x^*) + \langle \cancel{\nabla f(x^*)}, x - x^* \rangle + \underbrace{\frac{m}{2} \|x - x^*\|_2^2}_{>0 \text{ unless } x = x^*}.$$

□

3 Algorithmic setup

1. First-order oracle:

$$x \longrightarrow \text{oracle} \longrightarrow f(x), \nabla f(x)$$

2. Second-order oracle:

$$x \longrightarrow \text{oracle} \longrightarrow f(x), \nabla f(x), \nabla^2 f(x)$$

All algorithms we consider in this course are iterative:

- start with some x_0
- at iteration $k = 0, 1, 2, \ldots$
  - get oracle answers for x_k, choose x_{k+1}

4 Basic descent methods

Take the form

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, \ldots$$

Definition 2. $p \in \mathbb{R}^d$ is a *descent direction* for f at x if

$$f(x + tp) < f(x)$$

for all sufficiently small $t > 0$.

Proposition 1. *If f is continuously differentiable (in a neighborhood of x), then any p such that $\langle -\nabla f(x), p\rangle > 0$ is a descent direction.*

Proof. By Taylor's theorem:

$$f(x+tp) = f(x) + t\,\langle \nabla f(x+\gamma tp), p\rangle$$

for some $\gamma \in (0,1)$. We know that $\langle \nabla f(x), p\rangle < 0$. As ∇f is continuous, for all sufficiently small $t > 0$,

$$\langle \nabla f(x+\gamma tp), p\rangle < 0,$$

hence $f(x+tp) < f(x)$. □

5 Gradient descent

What would be a good descent direction?

Could try to move in the direction of $-\nabla f(x)$. Justification: Look at all p with $\|p\|_2 = 1$. Then

$$\inf_{\|p\|_2=1} \langle \nabla f(x), p\rangle = -\|\nabla f(x)\|_2$$

attained for $p = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$.

"Simplest" descent algorithm:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where α_k is the step size. Ideally, choose α_k small enough so that

$$f(x_{k+1}) < f(x_k)$$

when $\nabla f(x_k) \neq 0$.

Known as "gradient method", "gradient descent", "steepest descent" (w.r.t. the ℓ_2 norm).