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.*

pf: Part I: Suppose for contra, x^* is a local sol is not a global sol.

$\exists \bar{x}$ s.t. $f(\bar{x}) < f(x^*)$. $\mathcal{X}$ is cvx $\Rightarrow (1-\alpha)x^* + \alpha\bar{x} \in \mathcal{X}$.

$\because f$ is cvx $\therefore f((1-\alpha)x^* + \alpha\bar{x}) \le (1-\alpha)f(x^*) + \alpha f(\bar{x}) < (1-\alpha+\alpha)f(x^*) = f(x^*)$.

Set $\alpha \to 1$. $\Rightarrow \nexists B(x^*, \varepsilon)$ s.t. $\forall x \in B(x^*, \varepsilon)$, $f(x) \ge f(x^*)$.

Contradicts with x^* is a local min.

Part 2. $x^*, y^* \in \mathcal{X}$

Suppose $f(x^*) = f(y^*)$, both global sols. $\forall x \in \mathcal{X}$, $f(x) \ge f(x^*) = f(y^*)$.

Consider $(1-\alpha)x^* + \alpha y^*$. $\in \mathcal{X}$ Since f is cvx, $f((1-\alpha)x^* + \alpha y^*) \le (1-\alpha)f(x^*) + \alpha f(y^*)$

$= f(x^*)$

$\because x^*$ is a global min $\therefore f((1-\alpha)x^* + \alpha y^*) \ge f(x^*)$

$\therefore f((1-\alpha)x^* + \alpha y^*) = f(x^*) \Rightarrow (1-\alpha)x^* + \alpha y^*$ is a global sol.

$\therefore$ The set of all global sols is cvx.

1.1 Continuously 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*

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

Recall: Def of cvx function: epi(f) is a cvx set.

pf: part I: $\quad f(\alpha x+(1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y)$.

$(\Rightarrow)$ Suppose f is cvx. $\quad f((1-\alpha)x+\alpha y) \le (1-\alpha)f(x) + \alpha f(y)$.

$$f(y)-f(x) \ge \frac{f((1-\alpha)x+\alpha y) - f(x)}{\alpha} = \frac{\langle \nabla f(x), \alpha(y-x)\rangle + o(\|\alpha\|)}{\alpha}$$

$$= \langle \nabla f(x), y-x\rangle + \frac{o\|\alpha\|}{\alpha}$$

Set $\alpha \to 0$. $\square$.

$(\Leftarrow)$. Set $z=(1-\alpha)x+\alpha y$.

$$f(x) \ge f(z) + \langle \nabla f(z), \alpha(x-y)\rangle$$

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

Hence $\quad f(z) = (1-\alpha)f(z) + \alpha f(z) \le (1-\alpha)\left(f(x) - \langle \nabla f(z), \alpha(x-y)\rangle\right)$
$+\alpha\left(f(y) - \langle \nabla f(z), (1-\alpha)(y-x)\rangle\right)$

$$= (1-\alpha)f(x) + \alpha f(y).$$

$\therefore f((1-\alpha)x+\alpha y) \le (1-\alpha)f(x) + \alpha f(y)$.

Part 2.

By Taylor's thm, $\alpha>0$, $\quad \exists \gamma \in (0,1)$

$$f(x+\alpha u) = f(x) + \langle \nabla f(x), \alpha u\rangle + \frac{\alpha^2}{2} u^T \nabla^2 f(x+\gamma\alpha u) u.$$

$(\Rightarrow)$ $\nabla^2 f(x) \succeq 0 \Rightarrow u^T \nabla f(x+\gamma\alpha u) u \ge 0, \forall u \Rightarrow f(x+\alpha u) \ge f(x) + \langle \nabla f(x), \alpha u\rangle$

$\Rightarrow f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle$.

$(\Leftarrow)$ If f is cvx. $\quad f(x+\alpha u) \ge f(x) + \langle \nabla f(x), \alpha u\rangle$. $\Rightarrow u^T \nabla^2 f(x) u \ge 0$. $\forall u$.

$\Rightarrow \nabla^2 f(x) \succeq 0$.

Convex function: 4 Equiv statement:

$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle, \forall x\, y.$

$\Leftrightarrow \nabla^2 f(x) \succeq 0, \forall x$

$\Leftrightarrow f((1-\alpha)x + \alpha y) \leq \alpha f(x) + (1-\alpha) f(y)$

$\Leftrightarrow$ epi(f) is a convex set.

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

Recall: For smooth only functions, check local min is NP-hard.

But for cvx function, it's easy to check global min.

Pf: $\nabla f(x^*) = 0$. $x = x^* + p$, arbitrary x. (p).

$f(x) - f(x^*) \geq \langle \nabla f(x^*), p \rangle = 0$. $\Rightarrow \forall x, f(x) \geq f(x^*)$. $\Rightarrow x^*$ is a global min.

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

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

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

$\Downarrow$

Verification. $g(x) = f(x) - \frac{m}{2}\|x\|_2^2$.

($\Rightarrow$) Assume $f(x)$ is m-strongly-cvx.

Then $g((1-\alpha)x + \alpha y) = f((1-\alpha)x + \alpha y) - \frac{m}{2}\|(1-\alpha)x + \alpha y\|_2^2$

$\leq (1-\alpha) f(x) + \alpha f(y) - \frac{m}{2}(1-\alpha)\alpha\|y-x\|_2^2 - \frac{m}{2}\|(1-\alpha)x + \alpha y\|_2^2$

$= (1-\alpha)\left(f(x) - \frac{m}{2}\|x\|^2\right) + \alpha\left(f(y) - \frac{m}{2}\|y\|^2\right) = (1-\alpha)\, g(x) + \alpha g(y).$

(⇐) Assume $g(x)$ is cvx.

$$f((1-\alpha)x+\alpha y) - \frac{m}{2}\|(1-\alpha)x+\alpha y\|_2^2 \le (1-\alpha)\left(f(x)-\frac{m}{2}\|x\|_2^2\right) + \alpha\left(f(y)-\frac{m}{2}\|y\|_2^2\right)$$

整理即得.

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

1. *Suppose f is continuously differentiable. Then f is m-strong convexity $\boxed{\text{if and only if}}$*

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

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

2. *Suppose f is twice continuously differentiable. Then f is m-strong convexity $\boxed{\text{if and only if}}$*

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

(Compare with L-smoothness)

Pf: Apply Thm 2 to $g(x) = f(x) - \frac{m}{2}\|x\|^2$.

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 unique global minimizer of f.*

强凸性保证 global min 的唯一性.

Proof. By Part 1 of Theorem 4:

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

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.*

$\langle \nabla f(x), p\rangle < 0.$

Pf: By Taylor's thm,

$$f(x+tp) = f(x) + \langle \nabla f(x), tp\rangle + o\|t\|.$$

Set $t \to 0$. For p s.t. $\langle \nabla f(x), p\rangle < 0$, $f(x+tp) < f(x)$. $\square$.

5 Gradient descent

Any p with $\langle -\nabla f(x), p\rangle > 0$ is a descent direction. What would be a good choice? One that maximizes $\boxed{\langle -\nabla f(x), p\rangle}$ over some set of p's.

$f(x+tp) \simeq f(x) - t\underline{\langle -\nabla f(x), p\rangle}$

$\Downarrow$ maximize this!

For example, look at all p with $\|p\|_2 = 1$. Then

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

attained for $p = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$. ✓. 与 $-\nabla f(x)$ 同方向的 unit vec.

That is, try to move in the direction of the negative gradient, $-\nabla f(x)$.

"Simplest" descent algorithm:

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

where α_k is the step size. Ideally, choose $\boxed{\alpha_k}$ small enough so that

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

For large α_k, $o(\|\alpha_k\|)$ need to be considered.

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

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