Lecture-5

A crash course in unconstrained optimization

Ref: Bertsekas/Tsitsiklis "Neuro dynamic programming" Some book → Appendix B

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

θ^* is a local minima if

$$(*) \quad f(\theta^*) \le f(\theta) \quad \forall \theta \in N_\epsilon(\theta^*) \quad \text{for some } \epsilon > 0$$

$$N_\epsilon(\theta^*) = \{ \theta \,|\, \|\theta - \theta^*\| < \epsilon \}$$

θ^* is a global minima if

$$(**) \quad f(\theta^*) \le f(\theta) \quad \forall \theta \in \mathbb{R}^d$$

Strict local/global minima if the inequality in $(*)/(**)$ is strict

First-order necessary conditions:

Let θ^* be a local minima of $f: \mathbb{R}^d \to \mathbb{R}$.
Suppose that f is "continuously differentiable".

Then $\nabla f(\theta^*) = 0$

Also, if f is twice continuously differentiable, then

$\nabla^2 f(\theta^*)$ is positive semi-definite (p.s.d.)

Aside: $A \preceq B$ if $(B-A)$ is p.s.d e.g. $A \preceq \lambda I$

$x^T A x \geq 0 \quad \forall x$ p.s.d.

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x(1)} \\ \vdots \\ \frac{\partial f(x)}{\partial x(d)} \end{pmatrix}$$

f is differentiable if $\nabla f(x)$ exists $\forall x \in \mathbb{R}^d$.

If $\nabla f(\cdot)$ is continuous, then f is continuously differentiable. Such functions admit a Taylor series expansion, i.e.,

$$f(x+y) = f(x) + y^T \nabla f(x) + o(\|y\|)$$

$\hookrightarrow \lim_{\|y\| \to 0} \frac{o(\|y\|)}{\|y\|} = 0$

If $\nabla f(\cdot)$ is differentiable, then

Hessian matrix $\nabla^2 f(x) = \left[\left[\frac{\partial^2 f(x)}{\partial x(i) \partial x(j)}\right]\right]_{i,j=1\sim d}$

If f is twice continuously differentiable, (i.e., $\nabla^2 f$ exists & is continuous), then

$$f(x+y) = f(x) + y^T \nabla f(x) + \frac{1}{2} y^T \nabla^2 f(x) y + o(\|y\|^2)$$

P.T.O

Proof of first-order necessary conditions:

Recall θ^* is a local min.

(i) Fix $s \in \mathbb{R}^d$. Then,

$$s^T \nabla f(\theta^*) = \lim_{\delta \downarrow 0} \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta} \geq 0$$

Similarly $\quad -s^T \nabla f(\theta^*) \geq 0$

$$\Rightarrow \quad \nabla f(\theta^*) = 0$$

(ii) f is twice cont-diff'ble. Fix $s \in \mathbb{R}^n$

θ^* is a local min

$$f(\theta^* + \delta s) - f(\theta^*) \overset{\text{Taylor series expansion}}{=} \delta\, \underbrace{s^T \nabla f(\theta^*)}_{0} + \frac{\delta^2}{2} s^T \nabla^2 f(\theta^*) s + o(\delta^3)$$

$$0 \leq \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta^2} = \frac{1}{2} s^T \nabla^2 f(\theta^*) s + o(\delta)$$

So, as $\delta \to 0$, $\quad s^T \nabla^2 f(\theta^*) s \geq 0 \quad \forall s \in \mathbb{R}^d$

$$\Rightarrow \quad \nabla^2 f(\theta^*) \text{ is p.s.d.}$$

Second-order sufficient conditions:

Let f be twice continuously differentiable.

Suppose θ^* satisfies

$$\nabla f(\theta^*) = 0, \quad \nabla^2 f(\theta^*) \text{ is positive definite}$$

Then, θ^* is a strict local-min of f $\Leftrightarrow$

$\exists \delta > 0$, $\epsilon > 0$ s.t.

$$f(\theta) \geq f(\theta^*) + \delta \|\theta - \theta^*\|^2 \quad \forall \theta \text{ s.t. } \|\theta - \theta^*\| < \epsilon.$$

Example:

$$f(\theta) = \frac{1}{2}\theta^T A \theta - b^T \theta$$

If θ^* is a local min, then

necessary conditions → $\nabla f(\theta^*) = 0$ & $\nabla^2 f(\theta^*)$ is p.s.d.

i.e., $A\theta^* - b = 0$ & A is p.s.d.

Cases to consider:

① A is not p.s.d $\Rightarrow$ no local min

② A is p.s.d. Then, f is convex.

& so, any θ^* satisying $\nabla f(\theta^*) = A\theta^* - b = 0$

is a global min.

③ If A is p.d., then A^{-1} exists.

$A\theta^* - b = 0 \iff \theta^* = A^{-1}b$ is the unique global min.

(6) Think about the case when A is p.s.d. & singular.

Case (i) b is in the $\mathrm{Col}(A)$

Then, $\exists$ infinitely many solutions.

If $A\theta^* = b$ & $A\tilde{\theta} = 0$, then

$$A(\theta^* + \tilde{\theta}) = b, \quad \forall \tilde{\theta} \in N(A)$$

Case (ii) $b \notin \mathrm{Col}(A)$

Then, $\inf_{\theta} f(\theta) = -\infty$

why? Let $\tilde{\theta} \in N(A)$ and if $\tilde{\theta}^T b \neq 0$

unclear $\rightarrow$ (Such a $\tilde{\theta}$ exists since $b \notin \mathrm{Col}(A)$)

$A\tilde{\theta} = 0 \Rightarrow$

$$f(c\tilde{\theta}) = -cb^T\tilde{\theta}, \text{ for any } c \in \mathbb{R}$$

$$\Rightarrow \inf_{c\tilde{\theta}} f(c\tilde{\theta}) = -\infty$$

Taylor's theorem

Ref: Wright & Recht Ch. 2. "Opt for data analysis"

f continuously differentiable, $\theta, s \in \mathbb{R}^d$

$$f(\theta+s) = f(\theta) + \int_0^1 \nabla f(\theta+\lambda s)^T s \, d\lambda$$ → integral form

mean-value form → $$f(\theta+s) = f(\theta) + \nabla f(\theta+\tilde{\lambda} s)^T s, \text{ for some } \tilde{\lambda} \in (0,1)$$

If f twice cont-diff'ble

$$\nabla f(\theta+s) = \nabla f(\theta) + \int_0^1 \nabla^2 f(\theta+\lambda s) s \, d\lambda$$

$$f(\theta+s) = f(\theta) + \nabla f(\theta)^T s + \frac{1}{2} s^T \nabla^2 f(\theta+\bar{\lambda} s) s,$$

for some $\bar{\lambda} \in (0,1)$

Remark: f cont-diffble

$$f(\theta+s) = f(\theta) + \nabla f(\theta+\bar{\lambda} s)^T s$$

$$= f(\theta) + \nabla f(\theta)^T s + (\nabla f(\theta+\bar{\lambda} s) - \nabla f(\theta))^T s$$

why? → $$= f(\theta) + \nabla f(\theta)^T s + \underbrace{O(\|\nabla f(\theta+\bar{\lambda} s) - \nabla f(\theta)\| \|s\|)}_{\text{Big-Oh}}$$

why? → $$= f(\theta) + \nabla f(\theta)^T s + \underbrace{o(\|s\|)}_{\text{small-oh}}$$

$(\nabla f(\theta+\bar{\lambda} s) - \nabla f(\theta) \to 0 \text{ as } s \to 0, \forall \bar{\lambda})$

Def: A function $f: \mathbb{R}^d \to \mathbb{R}$ is L-smooth if $\exists L > 0$ s.t.

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\| \quad \forall \theta, \theta'$$

L-smooth $\Leftrightarrow$ gradient Lipschitz

Remark! For a L-smooth f, using Taylor's thm,

$$f(\theta) - f(\theta') - \nabla f(\theta')(\theta-\theta')$$
$$= \int_0^1 \left[\nabla f(\theta' + \lambda(\theta-\theta')) - \nabla f(\theta')\right](\theta-\theta')\,d\lambda$$

L-smoothness $\Rightarrow$

$$|(\nabla f(\theta' + \lambda(\theta-\theta')) - \nabla f(\theta'))(\theta-\theta')|$$
$$\leq \|\nabla f(\theta' + \lambda(\theta-\theta')) - \nabla f(\theta')\| \, \|\theta-\theta'\|$$
$$\leq L\lambda \|\theta-\theta'\|^2$$

$$\boxed{f(\theta) - f(\theta') - \nabla f(\theta')(\theta-\theta') \leq \frac{L}{2}\|\theta-\theta'\|^2} \quad \text{—— } (**)$$

Smoothness for twice cont diff'ble f

Claim: f twice cont-diff'ble + L-smooth

$$\Leftrightarrow \quad -LI \precsim \nabla^2 f(\theta) \precsim LI \quad \text{—— } (*)$$

psd sense

Pf: Suppose $(*)$ holds

$$\|\nabla f(\theta) - \nabla f(\theta')\| = \left\| \int_0^1 \nabla^2 f(\theta' + \lambda(\theta-\theta'))(\theta-\theta')\,d\lambda \right\|$$
$$\leq \int_0^1 \|\nabla^2 f(\theta' + \lambda(\theta-\theta'))\| \, \|\theta-\theta'\|\,d\lambda$$
$$\leq \int_0^1 L\|\theta-\theta'\|\,d\lambda = L\|\theta-\theta'\|$$

Converse:

$$f(\theta+\lambda s) - f(\theta) - \lambda \nabla f(\theta)^T s \le \frac{L}{2}\lambda^2 \|s\|^2$$

$$\text{LHS} = \frac{1}{2}\lambda^2 s^T \nabla^2 f(\theta + \alpha\lambda s) s \le \frac{L}{2}\lambda^2\|s\|^2$$

why?

$$s^T \nabla^2 f(\theta+\alpha\lambda s) s \le L\|s\|^2$$

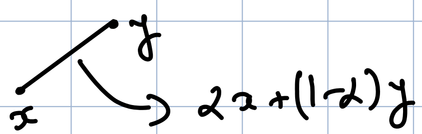
Letting $\alpha \downarrow 0$, $\quad s^T \nabla^2 f(\theta) s \le L\|s\|^2$

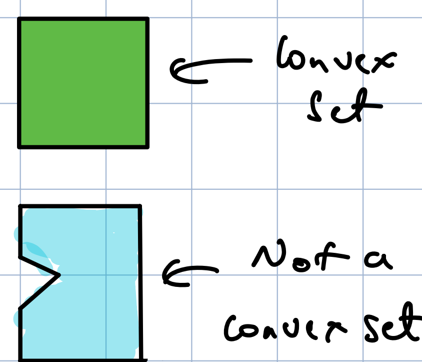
$$\Rightarrow \quad \nabla^2 f(\theta) \preceq L I$$

Tour of Convexity

Ref: Boyd & Vandenberghe "Convex Opt" Chapters 2 & 3

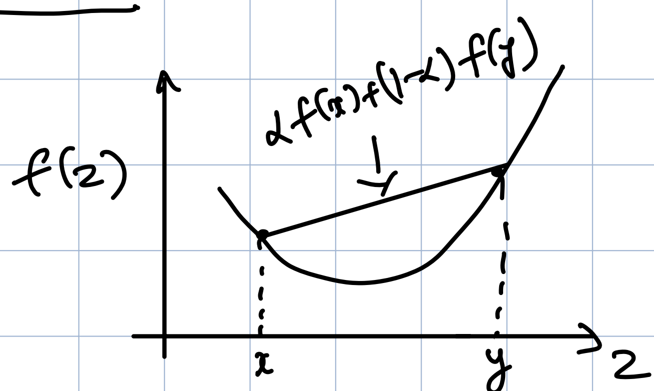
Convex sets:-





$C \subseteq \mathbb{R}^d$ is convex if $\alpha x + (1-\alpha) y \in C \quad \forall x, y \in C$ & $\forall \alpha \in [0,1]$

Convex function:



Let $C \subseteq \mathbb{R}^d$ be a convex set. A function $f: C \to \mathbb{R}$ is convex if

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y) \qquad (*)$$

Remark:

① f is concave if $(-f)$ is convex

② f is strictly convex if the inequality in $(\ast)$ is strict. $\forall x \neq y,\ \lambda \in (0,1)$

Examples:

① Linear functions are convex

② Any norm is convex

③ Weighted sum of convex functions is convex

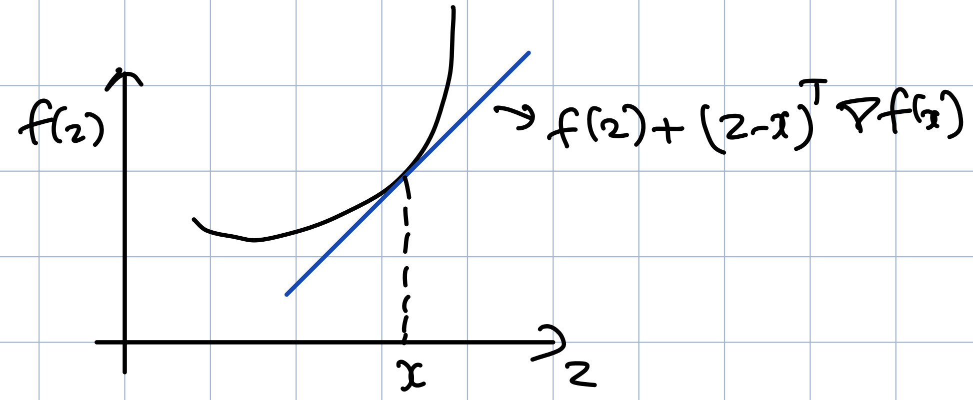
④ If $f_i : C \to \mathbb{R}$ is convex $\forall i \in I$, (index set)
then $h(x) = \sup_{i \in I} f_i(x)$ is convex.

Differentiable convex functions:

Let $C \subseteq \mathbb{R}^d$ be convex and $f : \mathbb{R}^d \to \mathbb{R}$ be a "differentiable" convex function. Then

(i) f is convex $\iff$ $f(z) \geq f(x) + (z-x)^T \nabla f(x)$, $\forall x, z \in C$. $(\ast\ast)$

(ii) If the inequality in $(\ast\ast)$ is strict, then f is strictly convex.



Lemma: Let $C \subseteq \mathbb{R}^d$ be convex and f be twice continuously differentiable. Then,

(i) If $\nabla^2 f(x)$ is positive semi-definite (p.s.d.) $\forall x \in C$, then f is convex over C.

(ii) If $\nabla^2 f(\cdot)$ is positive definite $\forall x \in C$, then f is strictly convex.

E.g. $f(x) = x^T A x$, A is symmetric $d \times d$

Then, f is convex $\Leftrightarrow$ A is p.s.d.

f is strictly convex $\Leftrightarrow$ A is p.d.

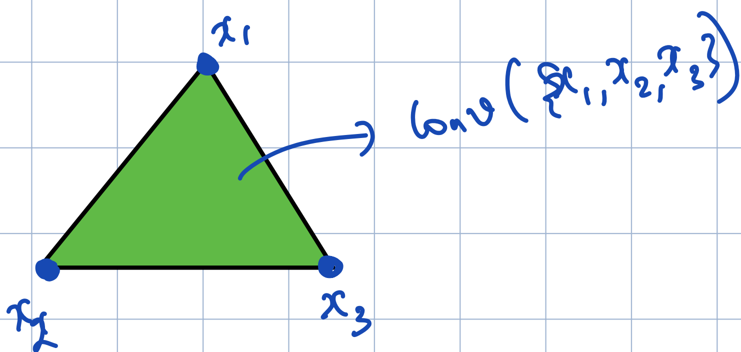
Theorem: Let $C \subseteq \mathbb{R}^d$ be convex & $f: C \to \mathbb{R}$ be convex. Then,

"a local minima of f is a global minima".

In addition, if f is strictly convex, then $\exists$ at most one global minima.

Convex hull: The convex hull, denoted $\text{conv}(\mathcal{X})$, of a set of point $\mathcal{X} \subseteq \mathbb{R}^d$ is the smallest convex set containing $\mathcal{X}$.

$$\text{conv}(\mathcal{X}) = \left\{ \sum_{i=1}^{m} \alpha_i x_i \;\middle|\; m \geq 1, \forall i \in \{1 \ldots m\}, \; x_i \in \mathcal{X}, \; \alpha_i \geq 0, \; \sum_{i=1}^{m} \alpha_i = 1 \right\}$$



Lecture-6

Gradient descent:

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

$$\theta_{t+1} = \theta_t + \alpha_t s_t \quad \text{——} \quad (*)$$

α_t : learning rate / step size

$s_t \in$ descent direction i.e., $f(\theta_{t+1}) < f(\theta_t)$

or $\nabla f(\theta_t)^T s_t < 0$ (assuming $\nabla f(\theta_t) \neq 0$)

In a gradient descent (GD) algorithm,

$$s_t = -\nabla f(\theta_t)$$

$$\Rightarrow \quad \nabla f(\theta_t)^T s_t = -\|\nabla f(\theta_t)\|^2 < 0$$

A first-order approximation to f yields

$$f(\theta_{t+1}) \simeq f(\theta_t) - \alpha_t \|\nabla f(\theta_t)\|^2 < f(\theta_t)$$

If f is convex, GD converges to a global minima.

Else, convergence is to a stationary point θ^* i.e., $\nabla f(\theta^*) = 0$.

GD: $$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t)$$

Stochastic gradient algorithm:

$$\theta \longrightarrow \boxed{\text{Gradient oracle}} \longrightarrow \nabla f(\theta)$$

$$\theta \longrightarrow \boxed{\begin{array}{c}\text{(stochastic)}\\ \text{gradient}\\ \text{oracle}\end{array}} \longrightarrow \nabla f(\theta) + \underbrace{\omega}_{\substack{\text{zero-mean noise}\\ \text{(bounded variance)}}}$$

$$(1)\quad \theta_{t+1} = \theta_t - \underbrace{\alpha_t}_{\text{stepsize/learning rate}} \big(\underbrace{\nabla f(\theta_t) + w_t}_{\text{noisy observation of the gradient of } f \text{ at } \theta_t} \big)$$

$$S_t = -\nabla f(\theta_t) - w_t$$

Conditions for ensuring the update in (1) for θ_t guarantees convergence to θ^* (a local optima of f).

A1 Typical assumption for stepsizes

$$\sum \alpha_t = \infty, \qquad \sum \alpha_t^2 < \infty$$

e.g. $\alpha_t = \dfrac{c}{t}$.

(A2) Let $\mathcal{F}_t$ be the information available upto time t i.e., $\{\theta_0, \theta_1, \omega_1, ---, \theta_{t-1}\}$

Noise has "zero" conditional mean $\left\{ E(\omega_t \mid \mathcal{F}_t) = 0 \right.$, and

Bounded variance $\left\{ E(\|\omega_t\|^2 \mid \mathcal{F}_t) \leq A + B\|\nabla f(\theta_t)\|^2 \right.$

Simple case: $\{\omega_t\}$ is i.i.d.

e.g. $\omega_t \sim N(0, \sigma^2)$

(3) Descent direction

$$\nabla f(\theta_t)^T E(s_t \mid \mathcal{F}_t) < 0$$

Observe for a SG algorithm,

$$s_t = -\nabla f(\theta_t) - \omega_t$$

So, $$\nabla f(\theta_t)^T E(s_t | \mathcal{F}_t)$$

$$= \nabla f(\theta_t)^T \left(-\nabla f(\theta_t) - E(w_t | \mathcal{F}_t) \right)$$

$$= -\|\nabla f(\theta_t)\|^2 < 0$$

It can also be checked that

$$E\left(\|s_t\|^2 | \mathcal{F}_t\right)$$

$$= \|\nabla f(\theta_t)\|^2 + E\left(\|w_t\|^2 | \mathcal{F}_t\right) + 2 \nabla f(\theta_t)^T E(w_t | \mathcal{F}_t)$$

$$\leq \|\nabla f(\theta_t)\|^2 + A + B \|\nabla f(\theta_t)\|^2$$

$$= A + (B+1) \|\nabla f(\theta_t)\|^2$$

Under the preceding assumptions, i.e., under A1 & A2, and f is smooth,

$$\lim_{t\to\infty} \nabla f(\theta_t) = 0$$

and the limit point, say θ^*, of the sequence θ_t is a "stationary point" of f (or) $\nabla f(\theta^*) = 0$.

Stochastic gradient descent

Want to minimize a function f that is a finite sum of smooth functions, i.e.,

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

<u>Example</u>: Given $\{(x_i, y_i), i = 1 \ldots n\}$

$$f(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

Goal is to find a minimizer of f.

Deterministic GD or batch GD:

$$\theta_{t+1} = \theta_t - \alpha_t \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_t)$$

Batch GD is computationally intensive on large training datasets, i.e., $n >> 1$

Stochastic gradient descent (SGD):

Let $$i_n = \begin{cases} 1 & \text{w.p. } 1/n \\ \vdots & \text{——}''\text{——} \\ n & \text{——}''\text{——} \end{cases}$$

Pick one of the component functions using the distribution of i_n, i.e., each f_i has the same chance of being chosen.

or $f_{i_n} = f_i$ w.p. $\frac{1}{n}$ $\forall i$

$$\underbrace{\theta_{t+1}}_{\text{SGD update iteration}} = \theta_t - \alpha_t \underbrace{\nabla f_{i_n}(\theta_t)}_{\text{Only one gradient computed per-iteration.}}$$

Rewrite SGD update as

$$\theta_{t+1} = \theta_t - \frac{\alpha_t}{n} \sum_{i=1}^{n} \nabla f_i(\theta_t) - \alpha_t \Big[\underbrace{\nabla f_{i_n}(\theta_t) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_t)} \Big]$$

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t) - \alpha_t w_t$$

noise

Recall $f = \frac{1}{n} \sum_{i=1}^{n} f_i$

Choose α_t s.t. $\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$

Observe that

information upto time t

$$E\left(\nabla f_{i_n}(\theta_t) \,\middle|\, \mathcal{F}_t \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_t)$$

$$\Rightarrow \quad E(w_t \mid \mathcal{F}_t) = 0$$

$$E\left(\|w_t\|^2 \,\middle|\, \mathcal{F}_t \right) = E\left(\|\nabla f_{i_n}(\theta_t)\|^2 \,\middle|\, \mathcal{F}_t \right) - \left\| E\left[\nabla f_{i_n}(\theta_t) \,\middle|\, \mathcal{F}_t \right] \right\|^2$$

$$\leq E\left(\|\nabla f_{i_n}(\theta_t)\|^2 \,\middle|\, \mathcal{F}_t \right)$$

$$\leq \max_{i=1 \dots n} \|\nabla f_i(\theta_t)\|^2$$

If we assume $\|\nabla f_i(\theta)\|^2 \leq A + B\|\nabla f(\theta)\|^2 \quad \forall \theta$

then, $E(\|w_t\|^2 | \mathcal{F}_t) \leq C + D\|\nabla f(\theta_t)\|^2$

Hence, θ_t updated using SGD

converges w.p. 1 to a point θ^*

that satisfies $\nabla f(\theta^*) = 0$.

In the context of linear regression,

$$f(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 \quad \text{--- (1)}$$

Batch GD: $\theta_{t+1} = \theta_t - \alpha_t \frac{1}{n} \sum_{i=1}^{n} x_i (y_i - x_i^T \theta_t)$

SGD: $\theta_{t+1} = \theta_t - \alpha_t \, x_{i_n} (y_{i_n} - x_{i_n}^T \theta_t)$

Computational advantages of SGD are clear

Further, SGD converges.

Strongly-convex functions

Def: f is μ-strongly convex if

$$f((1-\lambda)x+\lambda y) \le (1-\lambda)f(x)+\lambda f(y) - \frac{\mu}{2}(1-\lambda)\|y-x\|^2 \quad \forall x,y$$

For differentiable f:

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$$

For twice-diff'ble f:

$$\nabla^2 f(x) \succcurlyeq \mu I \quad \forall x$$

Example: $f(x) = x^T A x + b^T x + c$ (A symmetric)

$$\nabla^2 f(x) = A$$

f convex $\Leftrightarrow$ $A \succcurlyeq 0$

f strongly convex $\Leftrightarrow$ $A \succcurlyeq \mu I$

Remark: Strict convexity is $f((1-\lambda)x+\lambda y) < (1-\lambda)f(x)+\lambda f(y)$

f strictly convex $\Rightarrow$ f convex

f strongly convex $\Rightarrow$ f strictly convex

Strict convexity does not require $\nabla^2 f(x) \succ 0$

e.g. x^4

Other conditions for strong convexity:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2$$

$$\Downarrow$$

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \mu \|x-y\|^2$$

" Converse also true "

Proof hint: Show $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex &

use $(\nabla g(x) - \nabla g(y))^T (x-y) \geq 0$

Pb> f is μ-strongly convex

$\Rightarrow$ $f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$ $\rightarrow$ PL condition (Polyak-Lojasiewicz)

Pf: We know $f(y) - f(x) \geq \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2$

$$\underbrace{\min_y (f(y) - f(x))}_{= \, f(x^*) - f(x)} \geq \min_y \underbrace{\left(\nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2\right)}_{\text{differentiate}}$$

$$\nabla f(x) + \mu (y^* - x) = 0$$

$$y^* = -\frac{1}{\mu} \nabla f(x) + x$$

$$f(x^*) - f(x) \geq -\frac{1}{\mu} \|\nabla f(x)\|^2 + \frac{1}{2\mu} \|\nabla f(x)\|^2$$

(or) $f(x^*) - f(x) \geq -\frac{1}{2\mu} \|\nabla f(x)\|^2$ $\blacksquare$

1b) f strongly convex, g is convex

$\Rightarrow$ $f+g$ strongly convex

[redacted]

H.W.

① f diff'ble μ-strongly convex.

S.T.

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x-y\|$$

② f diff'ble μ-strongly convex.

Let $g_x(z) = f(z) - \nabla f(x)^T z$

Is $g_x(\cdot)$ strongly-convex with parameter μ?

③ f diff'ble μ-strongly convex.

S.T.

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)\|^2$$

Hint: PL inequality for g_x

④ $$\|x - x^*\| \leq \frac{2}{\mu} \|\nabla f(x)\|$$

Start with $f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|^2$

$$|\nabla f(x)^T (x^* - x)| \leq \|\nabla f(x)\| \, \|x^* - x\|$$

$\Rightarrow$ $$0 \geq f(x^*) - f(x) \geq -\|\nabla f(x)\| \, \|x^* - x\| + \frac{\mu}{2} \|x^* - x\|^2$$

$\Rightarrow$ $$\|x^* - x\| \leq \frac{2}{\mu} \|\nabla f(x)\|$$