A crash course in unconstrained optimization

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

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

θ^* is a local minima if

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

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

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.)

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

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

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

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

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

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

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

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

Second-order sufficient conditions:

Let f be twice continuously differentiable.

Suppose θ^* satisfies

$\nabla f(\theta^*) = 0$, $\nabla^2 f(\theta^*)$ is positive [redacted]te

Then, θ^* is a strict local-min of f Δ

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

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

<u>Example:</u> $f(\theta) = \frac{1}{2} \theta^T A \theta - b^T \theta$

If θ^* is a local min, then

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

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

<u>Cases to consider:</u>

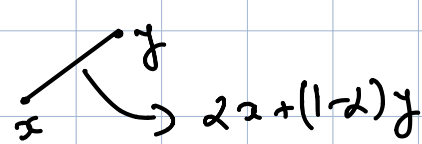
1. A is not p.s.d $\Rightarrow$ no local min
2. A is p.s.d. Then, f is convex.
   & so, any θ^* satisying $\nabla f(\theta^*) = A\theta^* - b = 0$ is a global min.
3. If A is p.d., then A^{-1} exists.

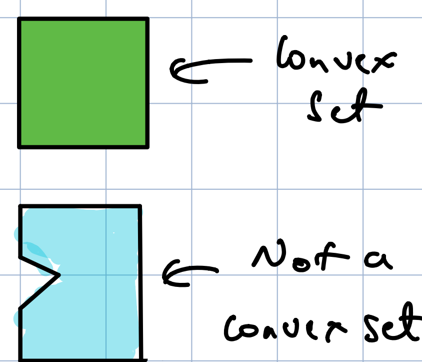
$A\theta^* - b = 0 \iff \theta^* = A^{-1}b$ is the unique global min.

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

Tour of Convexity

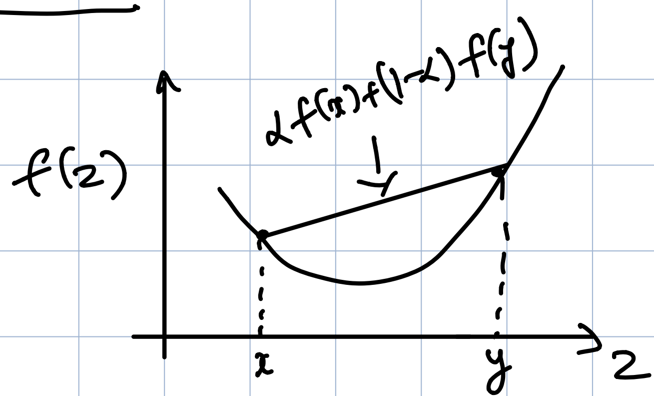
Convex sets:-





$C \subseteq \mathbb{R}^d$ 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

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

Remark:

1. f is concave if $(-f)$ is convex

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

Examples:

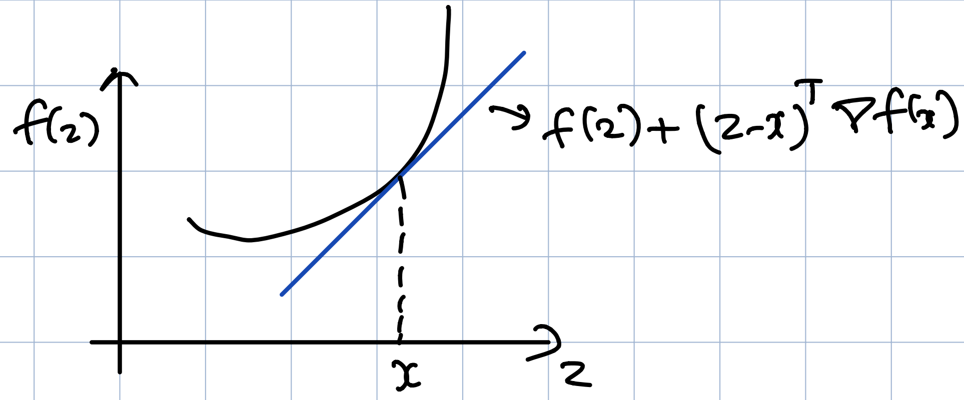
1. Linear functions are convex
2. Any norm is convex
3. Weighted sum of convex functions is convex
4. If $f_i : C \to \mathbb{R}$ is convex $\forall i \in I$, (index set)
then $h(x) = \sup_{i \in I} f_i(x)$ is convex.

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 $\Leftrightarrow$ $f(z) \geq f(x) + (z-x)^T \nabla f(x)$, $(**)$
$\forall x, z \in C$.

(ii) If the inequality in $(**)$ 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.

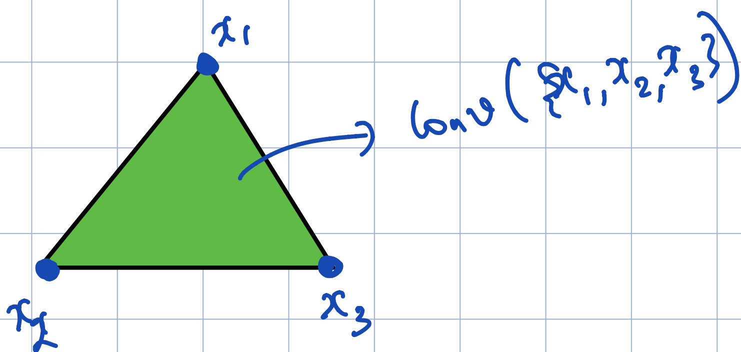
Theorem: Let $C \subseteq \mathbb{R}^d$ be convex & $f: C \to \mathbb{R}$ be convex. Then,

"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 $\operatorname{conv}(\mathcal{X})$, of a set of point $\mathcal{X} \subseteq \mathbb{R}^d$ is the smallest convex set containing $\mathcal{X}$.

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



Gradient descent:

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

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

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:

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

$$\theta \longrightarrow \boxed{\begin{matrix}\text{(stochastic)}\\ \text{gradient}\\ \text{oracle}\end{matrix}} \longrightarrow \nabla f(\theta) + \omega$$

zero-mean noise
(bounded variance)

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

stepsize / learning rate

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

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. $\alpha_t = \frac{c}{t}$.

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

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

$$E\left(\|s_t\|^2 | \mathcal{F}_t\right)$$
$$= \|\nabla f(\theta_t)\|^2 + E(\|w_t\|^2 | \mathcal{F}_t) + 2\nabla f(\theta_t)^T E(w_t | \mathcal{F}_t)$$
$$\leq \|\nabla f(\theta_t)\|^2 + A + B\|\nabla f(\theta_t)\|^2$$
$$= A + (B+1)\|\nabla f(\theta_t)\|^2$$

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.,

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

Example: Given $\{(x_i, y_i), i = 1 \dots n\}$

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

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

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

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

Rewrite SGD update as

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

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

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,

$$\mathbb{E}\left(\|\omega_t\|^2 \mid \mathcal{F}_t\right) \leq C + D\,\|\nabla f(\theta_t)\|^2$$

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,

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

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

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

Computational advantages of SGD are clear

Further, SGD converges.