# Accelerated Gradient Method

$$\min_x f(x) = g(x) + h(x)$$

- $g$ : convex and differentiable
- $h$ : convex

The algorithm:

$$y = (1-\theta_k)\, x_{k-1} + \theta_k u_{k-1}$$

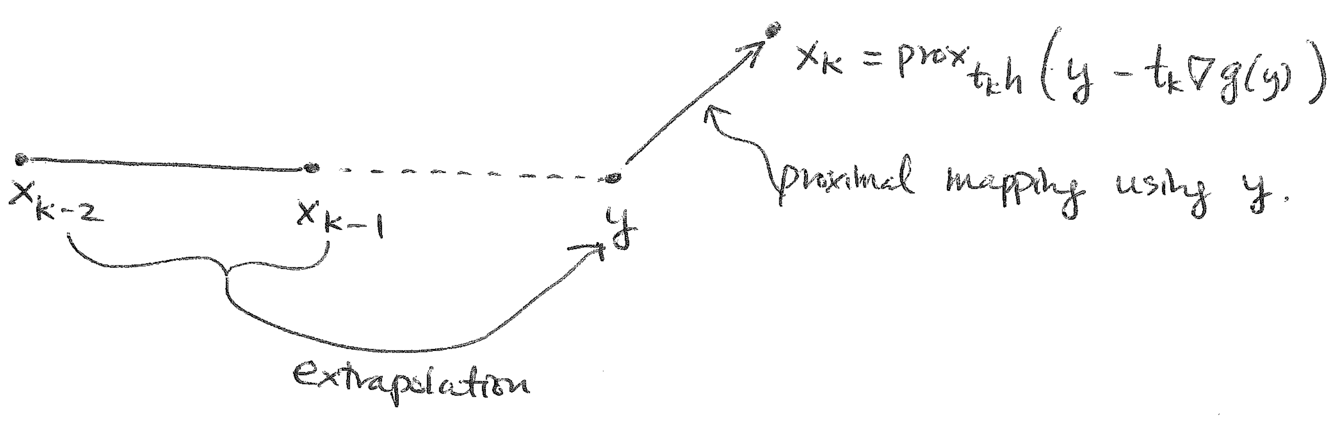
$$x_k = \text{prox}_{t_k h}\left(y - t_k \nabla g(y)\right)$$

$$u_k = x_{k-1} + \frac{1}{\theta_k}(x_k - x_{k-1})$$

What is it doing?

$$\begin{aligned} y &= (1-\theta_k)\, x_{k-1} + \theta_k u_{k-1} \\ &= (1-\theta_k)\, x_{k-1} + \theta_k \left[ x_{k-2} + \frac{1}{\theta_{k-1}}(x_{k-1} - x_{k-2}) \right] \\ &= x_{k-1} + \theta_k \left( \frac{1}{\theta_{k-1}} - 1 \right)(x_{k-1} - x_{k-2}). \end{aligned}$$

and $x_k = \text{prox}_{t_k h}\left(y - t_k \nabla g(y)\right)$.




$$y = x_{k-1} + \text{const}\,(x_{k-1} - x_{k-2})$$

<u>Example</u> ( logistic regression)

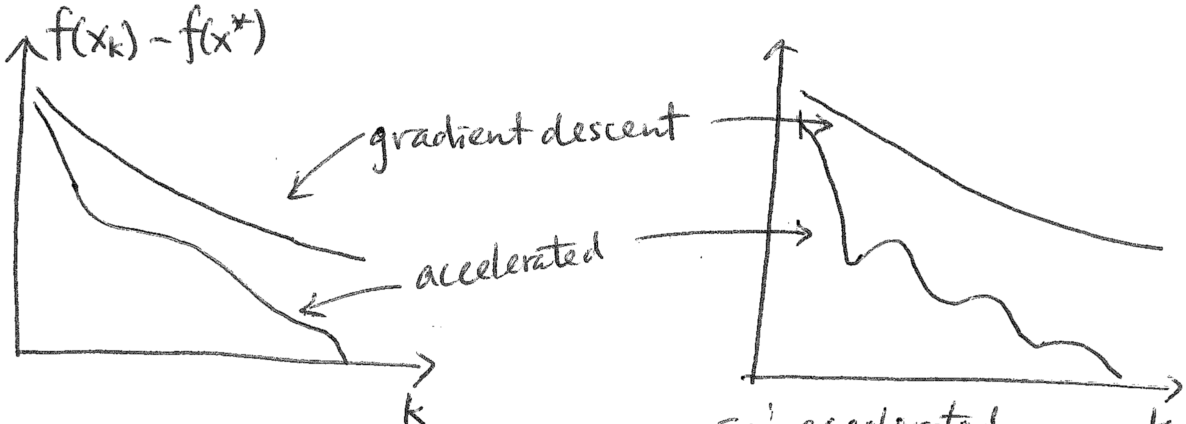
$$f(x) = \underbrace{\sum_{i=1}^{n}\left(-y_i\, a_i^T x + \log(1+\exp(a_i^T x))\right)}_{g(x)} + \underbrace{0}_{h(x)}$$

$$\nabla g(x) = -A^T\left(y - \cancel{\exp(a_i^T x)}\, p(x)\right),$$

where $\displaystyle P_i(x) = \frac{\exp(a_i^T x)}{1+\exp(a_i^T x)}$.

$\mathrm{prox}_h(x) = x$ because $h(x) = 0$.

Typical convergence plot




so accelerated is not a descent method.

<u>Choice of $\theta_k$:</u>

<u>Require</u>: $\displaystyle \frac{\theta_k^2}{t_k} \geq (1-\theta_k)\left(\frac{\theta_{k-1}^2}{t_{k-1}}\right).$

$t_k$ can be fixed, e.g. $t_k = \frac{1}{L}$, $L$ = Lipschitz constant of $\nabla g$.

then,

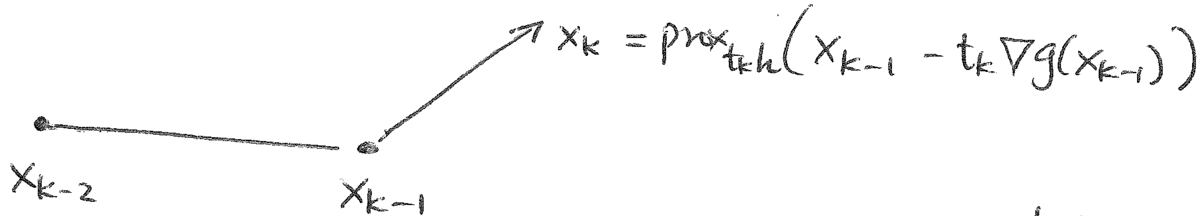
$$\theta_k^2 \geq (1-\theta_k)\,\theta_{k-1}^2$$

$$\Rightarrow \frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2}$$

if $\theta_k = 0$, then

$$y = x_{k-1}, \quad x_k = \text{prox}_{t_k h}\left(x_{k-1} - t_k \nabla g(x_{k-1})\right)$$

is the classic gradient projection




$$g(x) = \frac{1}{2}\|Ax - b\|^2$$

$$\nabla g(x) = A^T(Ax - b)$$

$$h(x) = \lambda \|x\|_1$$

<u>Example</u>

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1.$$

<u>algorithm</u>:

$$y = x_{k-1} + \theta_k\left(\frac{1}{\theta_{k-1}} - 1\right)(x_{k-1} - x_{k-2})$$

$$x_k = \text{prox}_{t_k h}\left(y - t_k \nabla g(x_k)\right)$$

$$= \operatorname*{argmin}_v \left\{ \lambda t_k \|v\|_1 + \frac{1}{2}\left\|v - \left(y - t_k \nabla g(x_k)\right)\right\|^2 \right\}$$

$$= S_{\lambda t_k}\left(y - t_k \nabla g(x_k)\right)$$

$$= \max\left\{ \left|y - t_k \nabla g(x_k)\right| - \lambda t_k, 0 \right\} \operatorname{sgn}\left(y - t_k \nabla g(x_k)\right).$$

## Example of $\theta_k$:

(i) $\theta_k = \frac{2}{k+1}$

$$\frac{1-\theta_k}{\theta_k^2} \overset{?}{\leq} \frac{1}{\theta_{k-1}^2} \iff \frac{1}{\theta_k^2} - \frac{1}{\theta_k} \overset{?}{\leq} \frac{1}{\theta_{k-1}^2}$$

Put $\theta_k = \frac{2}{k+1}$, i.e. $\frac{1}{\theta_k} = \frac{k+1}{2}$, then

$$\left(\frac{k+1}{2}\right)^2 - \frac{k+1}{2} = \frac{k^2+2k+1-2k-2}{4} = \frac{k^2-1}{4}$$

$$\leq \frac{k^2}{4} = \frac{1}{\theta_{k-1}^2}.$$

if we choose $\theta_k = \frac{2}{k+1}$, then the algorithm becomes

$$\begin{cases} y = x_{k-1} + \frac{k-2}{k+1}(x_{k-1} - x_{k-2}) \\ x_k = \mathrm{prox}_{th}\left(y - t\nabla g(y)\right) \end{cases}.$$

(ii) $\theta_k = \dfrac{2}{1+\sqrt{1+\frac{4}{\theta_{k-1}^2}}}$ (FISTA).

A constructive proof:

let $q_k = \frac{1}{\theta_k}$. Then

$$\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2} \iff q_k^2 - q_k = q_{k-1}^2$$

$$\Longrightarrow q_k = \frac{1+\sqrt{1+4q_{k-1}^2}}{2}.$$

# Convergence of Accelerated Gradient Method

Suppose we want to solve

$$\min_x f(x) = g(x) + h(x)$$

- $g$ is convex, differentiable, $\nabla g$ is Lipschitz with constant $L > 0$.
- $h$ is convex, and proximal function can be evaluated.

## Algorithm:

$$x_0 = u_0$$

$$y = (1-\theta_k) x_{k-1} + \theta_k u_{k-1} \qquad \theta_k = \frac{2}{k+1}$$

$$x_k = \mathrm{prox}_{th}(y - t\nabla g(y)) \qquad , \quad t \leq \frac{1}{L}.$$

$$u_k = x_{k-1} + \frac{1}{\theta_k}(x_k - x_{k-1})$$

## Theorem:

$$f(x_k) - f(x^*) \leq \frac{2\|x_0 - x^*\|^2}{t\,(k+1)^2}.$$

That means, the algorithm has $O(\frac{1}{k^2})$ rate of convergence, and this is a first-order method!

Lemma 1:

The function $g$ satisfies

$$g(x) \le g(y) + (x-y)^T \nabla g(y) + \frac{L}{2}\|x-y\|^2$$

Proof:

$$\begin{aligned}
g(x) &= g(y) + \int_0^1 (x-y)^T \nabla g(y + t(x-y))\, dt \\
&= g(y) + \int_0^1 \left[ (x-y)^T \nabla g(y + t(x-y)) - \nabla g(y) + \nabla g(y) \right] dt \\
&= g(y) + (x-y)^T \nabla g(y) + \int_0^1 (x-y)^T \left[ \nabla g(y + t(x-y)) - \nabla g(y) \right] dt \\
&\le g(y) + (x-y)^T \nabla g(y) + \int_0^1 \|x-y\| \, \|\nabla g(y + t(x-y)) - \nabla g(y)\| \, dt \\
&\le g(y) + (x-y)^T \nabla g(y) + \int_0^1 \|x-y\| \, L t \|x-y\| \, dt \\
&= g(y) + (x-y)^T \nabla g(y) + \frac{L}{2}\|x-y\|^2.
\end{aligned}$$

Lemma 2:

The function $h$ satisfies

$$h(x_{k+1}) \le h(z) + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}) + \nabla g(y)^T (z - x_{k+1})$$

for any $z$.

Proof: (Requires sub-gradients)

## Proof of Theorem:

First of all, we observe these two equations using lemmas:

$$g(x_{k+1}) \leq g(y) + \nabla g(y)^T (x_{k+1} - y) + \frac{1}{2t}\|x_{k+1} - y\|^2$$

$$h(x_{k+1}) \leq h(z) + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}) + \nabla g(y)^T (z - x_{k+1})$$

Summing the two equations yields

$$\begin{aligned}
f(x_{k+1}) &= g(x_{k+1}) + h(x_{k+1}) \\
&\leq g(y) + \nabla g(y)^T \left[ (\cancel{x_{k+1}} - y) + (z - \cancel{x_{k+1}}) \right] + \frac{1}{2t}\|x_{k+1} - y\|^2 \\
&\qquad + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}) + h(z) \\
&= \underbrace{g(y) + \nabla g(y)^T [z - y]}_{g \text{ convex}} + \frac{1}{2t}\|x_{k+1} - y\|^2 + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}) \\
&\qquad + h(z) \\
&\leq g(z) + h(z) + \frac{1}{2t}\|x_{k+1} - y\|^2 + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}) \\
&= f(z) + \frac{1}{2t}\|x_{k+1} - y\|^2 + \frac{1}{t}(x_{k+1} - y)^T (z - x_{k+1}).
\end{aligned}$$

Since $z$ is arbitrary, we put $z = x_k$ and $z = x^*$:

$$\begin{cases} f(x_{k+1}) \leq f(x^*) + \frac{1}{t}(x_{k+1} - y)^T (x^* - x_{k+1}) + \frac{1}{2t}\|x_{k+1} - y\|^2 \\ f(x_{k+1}) \leq f(x_k) + \frac{1}{t}(x_{k+1} - y)^T (x_k - x_{k+1}) + \frac{1}{2t}\|x_{k+1} - y\|^2 \end{cases}$$

Subtracting the two equations yields

$$0 \leq f(x^*) - f(x_k) + \frac{1}{t}(x_{k+1} - y)^T (x^* - x_k)$$

$$\Longleftrightarrow \quad f(x_k) - f(x^*) \leq \frac{1}{t}(x_{k+1} - y)^T (x^* - x_k)$$

So if we consider

$$f(x_{k+1}) - f(x^*) - (1-\theta_k)\left[f(x_k) - f(x^*)\right]$$

$$\leq \frac{1}{t}(x_{k+1}-y)^T(x^* - x_{k+1}) + \frac{1}{2t}\|x_{k+1}-y\|^2 - (1-\theta_k)\frac{1}{t}(x_{k+1}-y)^T(x^* - x_k)$$

$$= \frac{1}{t}(x_{k+1}-y)^T\left[x^* - x_{k+1} - (1-\theta_k)(x^*-x_k)\right] + \frac{1}{2t}\|x_{k+1}-y\|^2$$

$$= \underbrace{\frac{1}{t}(x_{k+1}-y)^T\left[\theta_k x^* + (1-\theta_k)x_k - x_{k+1}\right]} + \frac{1}{2t}\|x_{k+1}-y\|^2$$

Recall that

$$\begin{cases} u_{k+1} = x_k + \frac{1}{\theta_k}(x_{k+1} - x_k) \Rightarrow x_{k+1} = \theta_k(u_{k+1} - x_k) + x_k \\ \text{and} \\ y = (1-\theta_k)\, x_k + \theta_k u_k \end{cases}$$

So, $\theta_k x^* + (1-\theta_k)\, x_k - x_{k+1}$

$$= \theta_k x^* + (y - \theta_k u_k) - x_{k+1}$$

$$= \theta_k(x^* - u_k) + y - x_{k+1}$$

and hence

$$\frac{1}{t}(x_{k+1}-y)^T\left[\theta_k x^* + (1-\theta_k)x_k - x_{k+1}\right]$$

$$= \frac{1}{t}(x_{k+1}-y)^T\left[\theta_k(x^*-u_k) + y - x_{k+1}\right]$$

$$= \frac{\theta_k}{t}(x_{k+1}-y)^T(x^*-u_k) - \frac{1}{t}\|x_{k+1}-y\|^2.$$

Adding $\frac{1}{2t}\|x_{k+1}-y\|^2$, we have

$$\frac{\theta_k}{t}(x_{k+1}-y)^T(x^*-u_k) - \frac{1}{2t}\|x_{k+1}-y\|^2$$

$$= \frac{-1}{2t}\left\{\|x_{k+1}-y\|^2 - 2\theta_k(x_{k+1}-y)^T(x^*-u_k) + \theta_k^2\|x^*-u_k\|^2 - \theta_k^2\|x^*-u_k\|^2\right\}$$

$$= \frac{-1}{2t}\left\{\|\underset{\substack{\uparrow \\ \theta_k(u_{k+1}-x_k)+x_k}}{x_{k+1}}-y-\theta_k(x^*-u_k)\|^2-\theta_k^2\|x^*-u_k\|^2\right\}$$

So $x_{k+1}-y-\theta_k(x^*-u_k)$

$$= \theta_k(u_{k+1}-x_k)+x_k-y-\theta_k(x^*-u_k)$$

$$= \theta_k u_{k+1}+\underbrace{(1-\theta_k)x_k-y+\theta_k u_k}_{=0}-\theta_k x^*$$

$$= \theta_k(u_{k+1}-x^*).$$

$$= \frac{-\theta_k^2}{2t}\left\{\|u_{k+1}-x^*\|^2-\|u_k-x^*\|^2\right\}$$

$$= \frac{\theta_k^2}{2t}\left\{\|u_k-x^*\|^2-\|u_{k+1}-x^*\|^2\right\}$$

Thus, we have shown that

$$f(x_{k+1})-f(x^*)-(1-\theta_k)[f(x_k)-f(x^*)]$$

$$\leq \frac{\theta_k^2}{2t}\left\{\|u_k-x^*\|^2-\|u_{k+1}-x^*\|^2\right\}$$

$$\Leftrightarrow \frac{t}{\theta_k^2}[f(x_{k+1})-f(x^*)]+\frac{1}{2}\|u_{k+1}-x^*\|^2 \leq \frac{(1-\theta_k)t}{\theta_k^2}[f(x_k)-f(x^*)] + \frac{1}{2}\|u_k-x^*\|^2$$

$$\Rightarrow \frac{t}{\theta_k^2}[f(x_{k+1})-f(x^*)]+\frac{1}{2}\|u_{k+1}-x^*\|^2$$

$$\leq \frac{t}{\theta_{k-1}^2}[f(x_k)-f(x^*)]+\frac{1}{2}\|u_k-x^*\|^2, \quad \frac{1-\theta_k}{\theta_k^2}\leq\frac{1}{\theta_{k-1}^2}$$

$$\vdots$$

$u_0 = x_0$

$$\leq t\underbrace{\left(\frac{1-\theta_1}{\theta_1}\right)}_{=0 \text{ because } \theta_1=1.}[f(x_0)-f(x^*)]+\underbrace{\frac{1}{2}\|u_0-x^*\|^2}_{=\frac{1}{2}\|x_0-x^*\|^2}$$

$$\Rightarrow \frac{t}{\theta_k^2}\left[f(x_{k+1}) - f(x^*)\right] \le \frac{1}{2}\|x_0 - x^*\|^2 - \frac{1}{2}\|u_k - x^*\|^2$$

$$\le \frac{1}{2}\|x_0 - x^*\|^2$$

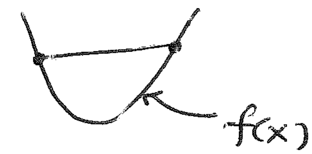
$$\Rightarrow f(x_{k+1}) - f(x^*) \le \frac{\theta_k^2}{2t}\|x_0 - x^*\|^2.$$

So if $\theta_k = \frac{2}{k+1}$, then

$$f(x_{k+1}) - f(x^*) \le \frac{2}{t(k+1)^2}\|x_0 - x^*\|^2.$$

# Subgradients

## 1. Convex Review




A function is convex if

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y), \quad 0 \le t \le 1.$$

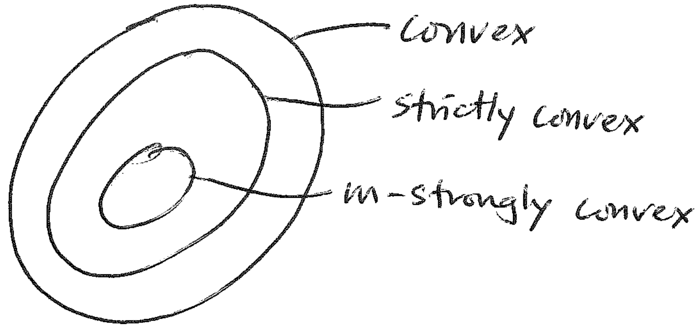
Strictly convex: the "$\le$" becomes "$<$".

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

m-Strongly convex:

$$f(x) - \frac{m}{2}\|x\|^2 \text{ is convex.}$$

i.e. $f(x)$ is "at least quadratic".




**Claim**: if $f$ is m-strongly convex, then $f$ is convex.

Proof:

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

Then,
$$g(tx + (1-t)y) = f(tx+(1-t)y) - \frac{m}{2}\|tx+(1-t)y\|^2$$
$$\le t g(x) + (1-t)g(y) = t\left\{f(x) - \frac{m}{2}\|x\|^2\right\} + (1-t)\left\{f(y) - \frac{m}{2}\|y\|^2\right\}$$
$$= \left[t f(x) + (1-t)f(y)\right] - \frac{m}{2}\left\{t\|x\|^2 + (1-t)\|y\|^2\right\}$$

So
$$f(tx+(1-t)y) \le t f(x) + (1-t)f(y) + \underbrace{\boxed{\frac{m}{2}\left\{\|tx+(1-t)y\|^2 - \left(t\|x\|^2 + (1-t)\|y\|^2\right)\right\}}}_{\le 0}$$

Note that
$$\|tx+(1-t)y\|^2 = t^2\|x\|^2 + 2t(1-t)x^Ty + (1-t)^2\|y\|^2$$
$$-\left(t\|x\|^2 + (1-t)\|y\|^2\right) \qquad - t\|x\|^2 - (1-t)\|y\|^2$$
$$= -t(1-t)\|x\|^2 + 2t(1-t)x^Ty - t(1-t)\|y\|^2$$
$$= -t(1-t)\|x-y\|^2 \le 0.$$

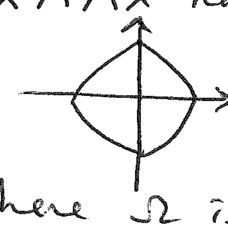
## Examples of Convex-Functions

(i) $f(x) = a^T x + b$ affine function

(ii) $f(x) = \frac{1}{2} x^T A x + b^T x + c, \quad A \geq 0.$

(iii) $f(x) = \frac{1}{2} \|Ax - b\|^2$, because $\frac{1}{2} x^T A^T A x$ has $A^T A \geq 0$.

(iv) $f(x) = \|x\|_p = \left(\sum x_i^p\right)^{1/p}$



(v) $f(x) = \mathbb{1}_\Omega(x) = \begin{cases} 0 & : x \in \Omega \\ +\infty, & x \notin \Omega, \end{cases}$ where $\Omega$ is convex

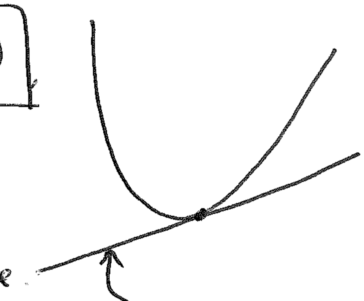
(vi) $f(x) = \max(x_1, x_2, \ldots, x_n)$.

## Properties of a convex function

(i) Suppose that $f$ is differentiable. Then $f$ is convex if and only if $\mathrm{dom} f$ is convex and

$$\boxed{f(y) \geq f(x) + \nabla f(x)^T (y - x)}$$

for all $x, y \in \mathrm{dom} f$.




(ii) Suppose that $f$ is twice differentiable. Then $f$ is convex if and only if $\mathrm{dom} f$ is convex and

$$\boxed{\nabla^2 f(x) \geq 0},$$

for all $x \in \mathrm{dom} f$.

(iii) Jensen's Inequality:
If $f$ is convex, and $X$ is a Random Variable, then

$$\boxed{f(\mathbb{E}[X]) \leq \mathbb{E}(f(X))}.$$

Proof: Consider the discrete case:

$$\mathbb{E}[X] = \sum_{i=1}^{n} \pi_i x_i , \quad \pi_i = \mathbb{P}(X = x_i)$$

So $$f(\mathbb{E}[X]) = f\left(\sum_{i=1}^{n} \pi_i x_i\right) \underbrace{\leq}_{\text{Convexity of } f.} \sum_{i=1}^{n} \pi_i f(x_i) = \mathbb{E}[f(X)].$$

## Operations Preserving Convexity

(i) Non-negative linear combination

if $f_1, f_2, \dots, f_m$ convex,

then $\sum_{i=1}^{m} \omega_i f_i$ is also convex. $\omega_i \geq 0$.

Example: $(a_i^T x - b_i)^2$ is convex, so

$\sum_{i=1}^{n} (a_i^T x - b_i)^2$ is also convex

(ii) Affine Composition

if $f$ is convex, then $f(Ax + b)$ is also convex

Example: $-\sum_{i=1}^{n} \log(x_i)$ is convex.

So $-\sum_{i=1}^{n} \log(a_i^T x + b_i)$ is also convex

(iii) log-sum-exp

The function

$$g(x) = \log\left(\sum_{i=1}^{n} e^{a_i^T x + b_i}\right) \text{ is convex.}$$

Proof: Just need to show $f(x) = \log\left(\sum_{i=1}^{n} e^{x_i}\right)$ is convex.

$$(\nabla f)_i = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}}, \quad (\nabla^2 f)_{ij} = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} \mathbb{1}\{i=j\} - \frac{e^{x_i} e^{x_j}}{\left(\sum_{i=1}^{n} e^{x_i}\right)^2}$$

We can show that

$$\left|(\nabla^2 f)_{ii}\right| = \left|\frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} - \frac{(e^{x_i})^2}{\left(\sum_{i=1}^{n} e^{x_i}\right)^2}\right|$$

$$\sum_{j\neq i}\left|(\nabla^2 f)_{ij}\right| = \sum_{j\neq i}\frac{e^{x_i}e^{x_j}}{\left(\sum e^{x_i}\right)^2} = \left|\sum_{j=1}^{n}\frac{e^{x_i}e^{x_j}}{\left(\sum e^{x_i}\right)^2} - \frac{(e^{x_i})^2}{\left(\sum e^{x_i}\right)^2}\right|$$

$$= \left|\frac{e^{x_i}}{\sum e^{x_i}} - \frac{(e^{x_i})^2}{\left(\sum e^{x_i}\right)^2}\right|.$$

So $\left|(\nabla^2 f)_{ii}\right| = \sum_{j\neq i}\left|(\nabla^2 f)_{ij}\right|$, and this implies $\nabla^2 f$ is diagonally dominant. Then by Gershgorin Disk Theorem, $\nabla^2 f$ is positive semi-definite.

(iv) Composition:

$$f = h \circ g.$$

$$\Rightarrow f''(x) = h''(g(x))\,g'(x)^2 + h'(g(x))\,g''(x).$$

So

| f | h | | g |
|---|---|---|---|
| Convex | Convex | non-decreasing | Convex |
| Convex | Convex | non-increasing | Concave |
| Concave | Concave | non-decreasing | Concave |
| Concave | Concave | non-increasing | Convex |

# 2. Subgradients

The <u>subgradient</u> of a convex function $f$ is any $g \in \mathbb{R}^n$ s.t.

$$f(y) \geq f(x) + g^T(y-x).$$

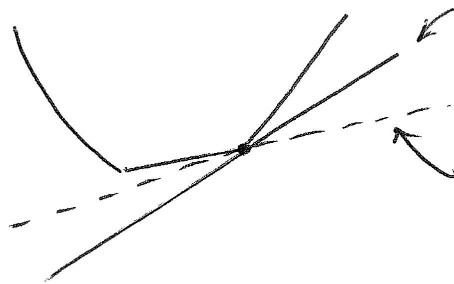
Why this definition?

Recall that a convex function $f$ always have

$$f(y) \geq f(x) + \nabla f(x)^T (y-x).$$

So a subgradient is a generalization for non-diff. functions.

Pictorial Illustration:



this is a subgradient $f(y) \geq f(x) + g_1^T(y-x)$

this is another subgradient $f(y) \geq f(x) + g_2^T(y-x)$.

<u>Sub-differential</u>:

$$\partial f(x^*) = \{ g \mid f(y) \geq f(x^*) + g^T(y-x^*) \}$$

the set containing all the subgradients at $x^*$.

- $\partial f(x^*)$ is closed and convex
- $\partial f(x^*)$ is non-empty
- if $f$ is differentiable at $x^*$, then $\partial f(x^*) = \{ \nabla f(x^*) \}$.

• For any $f$,

$$x^* = \underset{x}{\operatorname{argmin}}\ f(x) \iff 0 \in \partial f(x^*)$$

or, in other words,

$$f(y) \geq f(x^*) + 0^T(y - x^*) = f(x^*),\ \forall y \in \operatorname{dom} f.$$

(this of course implies $x^*$ is the minimizer.)

<u>Lemma 2</u> of Accelerated Gradient Method

$$h(x_{k+1}) \leq h(z) + \frac{1}{t}(x_{k+1} - y)^T(z - x_{k+1}) + \nabla g(y)^T(z - x_{k+1}).$$

<u>proof</u>: Note that

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ h(v) + \frac{1}{2t}\| v - (y - t\nabla g(y))\|^2 \right\}$$

$0 \in \partial(\cdot)$ implies that

$$\Rightarrow 0 \in \partial h(x_{k+1}) + \frac{1}{t}(x_{k+1} - (y - t\nabla g(y)))$$

$$\Rightarrow \frac{1}{t}(y - t\nabla g(y) - x_{k+1}) \in \partial h(x_{k+1})$$

def of subgrad

$$\Rightarrow h(z) \geq h(x_{k+1}) + \frac{1}{t}(y - t\nabla g(y) - x_{k+1})^T(z - x_{k+1})$$

$$\Rightarrow h(x_{k+1}) \leq h(z) + \frac{1}{t}(x_{k+1} - (y - t\nabla g(y)))^T(z - x_{k+1})$$

$$= h(z) + \frac{1}{t}(x_{k+1} - y)^T(z - x_{k+1})$$

$$+ \nabla g(y)^T(z - x_{k+1}).$$