Gradient descent

Program for today:

1) Convex functions

2) Convex functions & Lipschitz continuity

3) GD: basic convergence thm

4) Final remarks.

Next Time: Stochastic Gradient Descent.

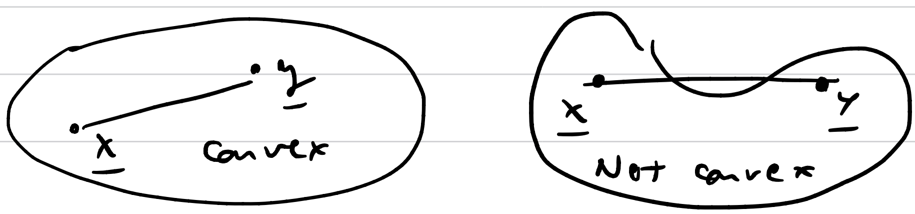
#.

[1] Convex functions.

Definition 1: convex sets.

A set S is convex if for all $\underline{x}, \underline{y} \in S$

then for all $\lambda \in [0,1]$: $\lambda \underline{x} + (1-\lambda)\underline{y} \in S$



Definition 2; convex functions.

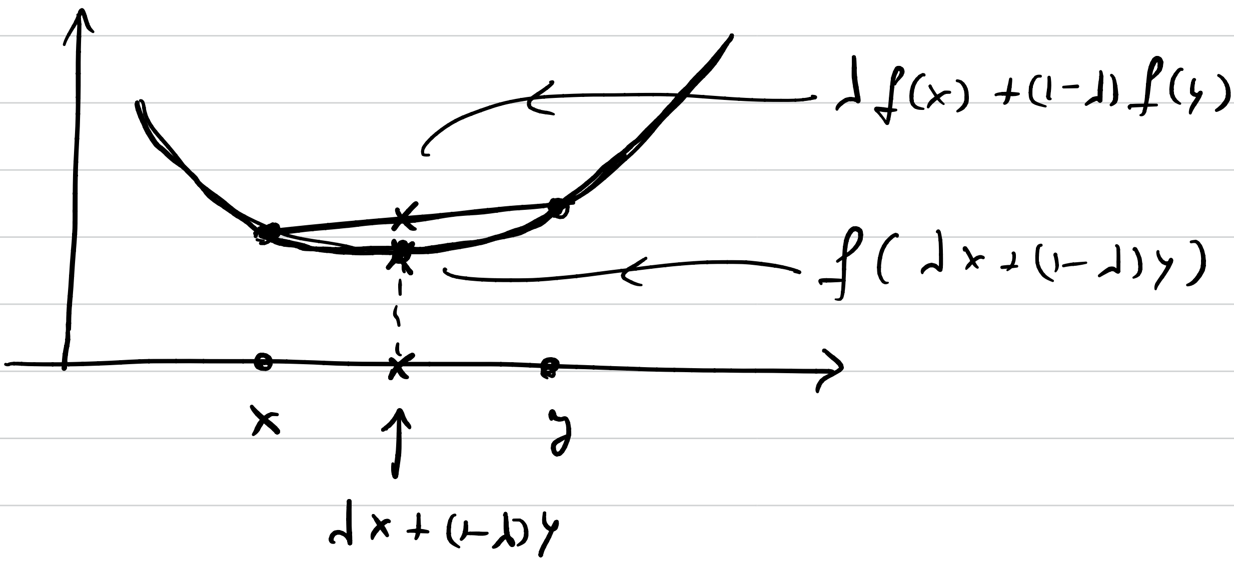
Let $f : S \to \mathbb{R}$, S an open convex set.

We say that f is convex if for all $\underline{x}, \underline{y} \in S$

and all $\lambda \in [0,1]$.

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(\underline{x}) + (1-\lambda) f(\underline{y})$$

Picture for $S = \mathbb{R}$ or $S =]a, b[$



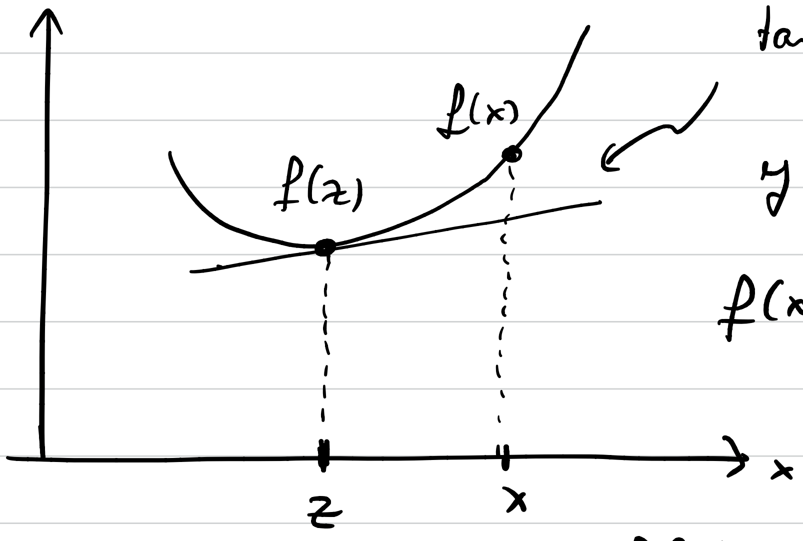
The chord is above the function.

Alternative characterisation of a convex function:

Lemma: Let S an open convex set and let $f : S \to \mathbb{R}$. Then f is convex if $\forall z \in S$ there exist $\underline{v}$ such that

$$f(\underline{x}) \geq f(\underline{z}) + \langle \underline{v}, \underline{x} - \underline{z} \rangle, \quad \forall x \in S$$

Intuition:



tangent line at p:

$$y = f(z) + f'(z)(x-z)$$

$$f(x) \geq f(z) + f'(z)(x-z)$$

i.e

$$f(x) - f(z) \geq \underbrace{f'(z)}_{v}(x-z)$$

If f is differentiable this intuition can be made rigorous and $\underline{v} = \underline{\nabla} f(z)$. Gradient at z.

Proof of Lemma.

Fix $z \in S$. Assume $\exists\, \underline{v}$ s.t $\forall\, x, y \in S$

$$f(\underline{x}) \geq f(z) + \langle \underline{v}, \underline{x} - \underline{z} \rangle$$

$$f(\underline{y}) \geq f(z) + \langle \underline{v}, \underline{y} - \underline{z} \rangle$$

Then $\forall\, \lambda \in [0,1]$ multiply first eqn by λ and second equation by $1-\lambda$, and finally sum them:

$$\lambda f(\underline{x}) + (1-\lambda) f(\underline{y}) \geq f(z) + \langle \underline{v}, \lambda x + (1-\lambda) y - z \rangle$$

Now set $z = \lambda x + (1-\lambda) y$. We get:

$$\lambda f(\underline{x}) + (1-\lambda) f(\underline{y}) \geq f(\lambda x + (1-\lambda) y)$$

which means that f is convex $\blacksquare$

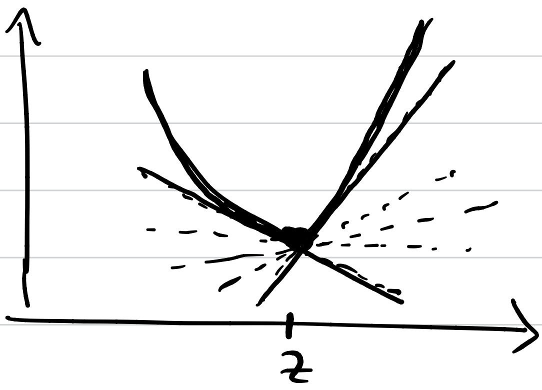
Remark: in the proof $\underline{v}$ depends on z only. eventually we choose $z = \lambda x + (1-\lambda) y$.

<u>Convexe Lemma</u>: Let $f : S \to \mathbb{R}$ and assume it is convex. Then given any $z \in S$ there exist $\underline{v}$ such that for all $x \in S$

$$f(x) \geq f(z) + \langle \underline{v}, (\underline{x} - \underline{z}) \rangle .$$

If f is differentiable at z, then $\underline{v}$ is unique (for this z) and $\underline{v} = \underline{\nabla} f(z)$.

<u>Remark</u>: When f is not differentiable at $\underline{z}$ $\underline{v}$ still exists as long as f is convex, however it is non-unique.



All slopes between the left and right derivatives can serve as $\underline{v}$ and we have

$f(x) \geq f(z) + v(x-z)$.

This leads us to the definition:

Definition 3. Subgradient

Any $\underline{v}_z$ that fullfills the condition

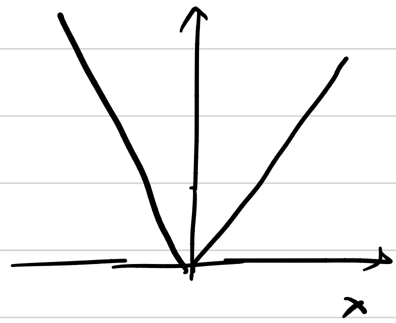
$$f(\underline{x}) \geq f(\underline{z}) + \langle \underline{v}_z, \underline{x} - \underline{z} \rangle$$

is called a subgradient. The set of all such $\underline{v}_z$'s is called the differential set. We denote this by $\underline{v}_z \in \underbrace{\partial f(z)}_{\text{notation for differential set}}$.

When f is differentiable at z we have

$$\partial f(z) = \{ \underline{\nabla} f(z) \}.$$

Example 1. $f(x) = \mathrm{Relu}(x)$ $\qquad |x|$

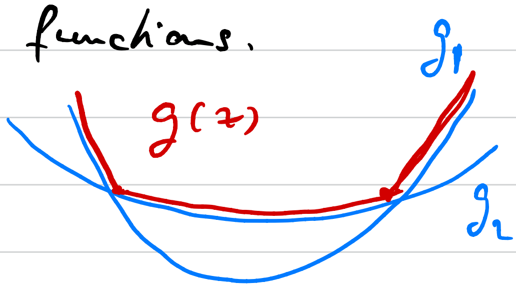


$$\partial f = \begin{cases} \{0\} & \text{if } x < 0 \\ \{+1\} & \text{if } x > 0 \\ [0,1] & \text{if } x = 0 \end{cases} \qquad \partial f = \begin{cases} -1, & x < 0 \\ +1, & x > 0 \\ [-1,+1] & x = 0 \end{cases}$$

Example 2 (important and classical).

$g_1, \dots, g_r$ convex differentiable functions.

Let $\quad g(z) = \max\limits_{i=1\dots r} g_i(z)$



$\boxed{\text{Then } g(z) \text{ is a convex function}}$

Proof: g_j are differentiable and convex so:

$$g(x) = \max_i g_i(x) \geq g_j(x) \geq g_j(z) + \langle \nabla g_j(z), x - z \rangle$$

Now take $j = \arg\max\limits_i g_i(z)$. Thus $g_j(z) = g(z)$

$$\Rightarrow \quad g(x) \geq g(z) + \langle \underbrace{\nabla g_j(z)}_{v_j}, x - z \rangle$$

with $j = \arg\max g_i(z)$ Not unique.

Proof of converse lemma in differentiable case.

We assume f is convex. Then

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$$

$$\Rightarrow \quad f(x) \geq \frac{1}{\lambda}\left[f(\lambda x + (1-\lambda) y) - (1-\lambda) f(y)\right]$$

$$f(x) \geq f(y) + \frac{1}{\lambda}\left[f(\lambda x + (1-\lambda) y) - f(y)\right]$$

$\forall \lambda \in [0,1]$.

$$\Rightarrow \quad f(x) \geq f(y) + \lim_{\lambda \to 0_+} \frac{f(y + \lambda(x-y)) - f(y)}{\lambda}$$

If f is differentiable the limit is $\langle \underline{\nabla} f(y), \underline{x}-\underline{y} \rangle$

(by definition of gradient or by Taylor expansion, ...)

Thus $\underline{v}_y = \underline{\nabla} f(y)$. ∎

2 Lipschitz functions (and convexity).

The Lipschitz condition provides a strong form of continuity. When combined with convexity nice properties emerge.

Definition 4. ρ-Lipschitz functions

f is ρ-Lipschitz if $\forall\, \underline{x}, \underline{y} \in S$ an open set we have

$$|f(\underline{x}) - f(\underline{y})| \leq \rho \, \| \underline{x} - \underline{y} \|.$$

Lemma Let S an open convex set and $f : S \to \mathbb{R}$ a convex function.

f is ρ-Lipschitz if and only if for all $z \in S$ and $\underline{v}_z \in \partial f(z)$ we have $\| v_z \| \leq \rho$.

For differentiable f in particular $\| \underline{\nabla} f(z) \| \leq \rho$.

Proof

First direction: assume $\|V_z\| \leq \rho$.

Since f is convex and V_z is a subgradient we have

$$f(x) \geq f(z) + \langle V_z, x-z \rangle$$

$$\Rightarrow \quad f(z) - f(x) \leq \langle V_z, z-x \rangle$$

$$\leq \|V_z\| \|z-x\| \text{ by Cauchy-Schwarz.}$$

$$\leq \rho \|z-x\|$$

by exchanging z & x: $f(x) - f(z) \leq \rho \|x-z\|$.

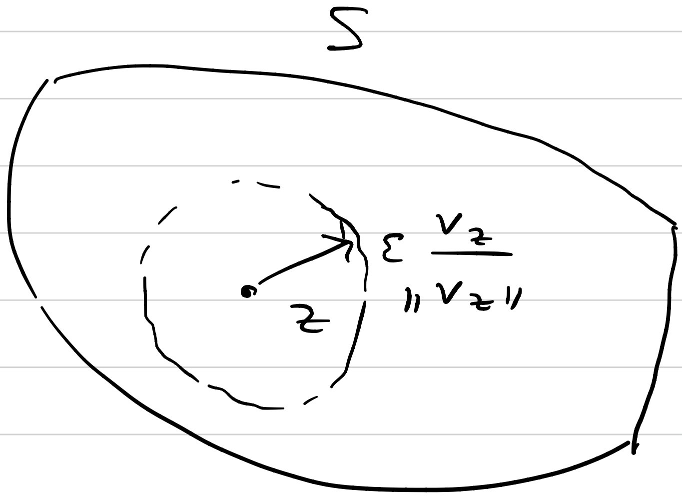
Thus $|f(x)-f(z)| \leq \rho \|x-z\|$ and f is ρ-Lipschitz.

Converse direction: assume f is ρ-Lipschitz.

Then $\forall x, z \in S$ $|f(x)-f(z)| \leq \rho \|x-z\|$

Take $x = z + \varepsilon \dfrac{V_z}{\|V_z\|}$, $V_z \in \partial f(z)$

(Recall f is convex so subgradients exist)



By convexity of f:

$$f(\underline{x}) \geq f(\underline{z}) + \langle \underline{v}_z, \underline{x} - \underline{z} \rangle$$

$$f(\underline{x}) - f(\underline{z}) \geq \underbrace{\left\langle v_z, \varepsilon \frac{v_z}{\|v_z\|} \right\rangle}_{\varepsilon \|v_z\|}$$

By ρ-Lipschitzness $\quad f(\underline{x}) - f(\underline{z}) \leq \rho \|\underline{x} - \underline{z}\|$

Thus $\quad \rho \underbrace{\|\underline{x} - \underline{z}\|}_{\varepsilon} \geq \varepsilon \|v_z\|$

$\Rightarrow \quad \|v_z\| \leq \rho . \quad \blacksquare$

3 Gradient Descent.

GD is a way to "walk" efficiently through S in discrete time steps in order to reach minimum of f in S.

GD algorithm. Let $S \ni 0$ open convex set and f convex fct $f: S \to \mathbb{R}$.

a) start at $w^1 = 0$

b) at steps $t = 1 \dots T$ do

$$\underline{w}^{t+1} = \underline{w}^t - \eta \underbrace{\underline{\nabla} f(\underline{w}^t)}$$

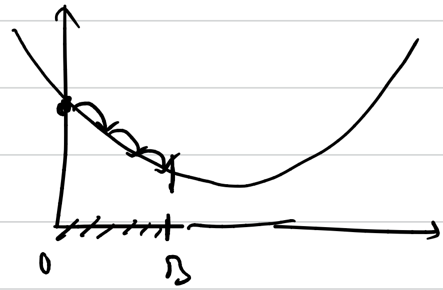
if f is differentiable and otherwise pick <u>any</u> subgradient. Not here we denote $\underline{\nabla} f(\underline{w}^t)$ by slight abuse of notation.

c) Output: $\overline{w} = \frac{1}{T} \sum_{t=1}^{T} w^t$.

Here η is called the "step size" or "rate".

Theorem: "GD approach to optimal values"

Let $\underline{w}^* = \underset{\|\underline{w}\| \leq B}{\arg\min} f(\underline{w})$



the minimizer of f in a ball $B \ni 0$.

Then after T steps and step size $\eta = \sqrt{\dfrac{B^2}{\rho^2 T}}$

we have:

$$0 \leq f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Remark:

If we ask for $0 \leq f(\bar{w}) - f(w^*) \leq \varepsilon$

it is enough to take $\dfrac{B\rho}{\sqrt{T}} \leq \varepsilon \Rightarrow T \geq \dfrac{B^2\rho^2}{\varepsilon^2}$

and $\eta = \sqrt{\dfrac{B^2}{\rho^2 T}} \leq \sqrt{\dfrac{B^2}{\rho^2 \dfrac{B^2\rho^2}{\varepsilon^2}}} = \dfrac{\varepsilon}{\rho^2}$.

$$\boxed{T \geq \frac{B^2\rho^2}{\varepsilon^2} \quad \& \quad \eta \leq \frac{\varepsilon}{\rho^2}. \quad \text{to get } \varepsilon \text{ close to Min.}}$$

Proof of Theorem.

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T}\sum_{t=1}^{T} w^t\right) - f(w^*)$$

$$\underset{\text{convexity}}{\leq} \frac{1}{T}\sum_{t=1}^{T} f(w^t) - f(w^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T} \left(f(w^t) - f(w^*)\right)$$

$$\underset{\text{convexity}}{\leq} \frac{1}{T}\sum_{t=1}^{T} \langle \underline{\nabla} f(\underline{w}^t), \underline{w}^t - \underline{w}^* \rangle$$

$$= \frac{1}{\eta T}\sum_{t=1}^{T} \langle \eta \underline{\nabla} f(\underline{w}^t), \underline{w}^t - \underline{w}^* \rangle$$

$$= \frac{1}{2\eta T}\sum_{t=1}^{T} \Big\{ -\|\underline{w}^t - \underline{w}^* - \eta \underline{\nabla} f(w^t)\|^2 + \|\underline{w}^t - \underline{w}^*\|^2 + \|\eta^2 \underline{\nabla} f(\underline{w}^t)\|^2 \Big\}$$

$$\underset{\text{use GD step}}{=} \frac{1}{2\eta T}\sum_{t=1}^{T} \Big\{ -\|\underline{w}^{t+1} - \underline{w}^*\|^2 + \|\underline{w}^t - \underline{w}^*\|^2 + \eta^2 \|\underline{\nabla} f(w^t)\|^2 \Big\}$$

$$= \frac{1}{2\eta T} \sum_{t=1}^{T} \left\{ - \| \underline{w}^{t+1} - \underline{w}^{*} \|^2 + \| \underline{w}^{t} - \underline{w}^{*} \|^2 \right\}$$

$$+ \frac{\eta}{2T} \sum_{t=1}^{T} \| \nabla f(\underline{w}^{t}) \|^2$$

The first sum is a "telescopic sum" and all successive terms cancel except for first and last one:

$$= \frac{1}{2\eta T} \left\{ \| w^{1} - w^{*} \|^2 - \| w^{T+1} - w^{*} \|^2 \right\}$$

$$+ \frac{\eta}{2T} \sum_{t=1}^{T} \| \underline{\nabla} f(\underline{w}^{t}) \|^2$$

Now use $\underline{w}^{1} = 0$ (initialization of GD):

$$\leq \frac{1}{2\eta T} \| w^{*} \|^2 + \frac{\eta}{2T} \sum_{t=1}^{T} \underbrace{\| \underline{\nabla} f(\underline{w}^{t}) \|^2}_{\leq \rho^2 \text{ by Lipschitz condition.}}$$

$$\leq \frac{B^2}{2\eta T} + \frac{\eta \rho^2}{2}.$$

Now we choose η to balance the two terms (best possible η):

$$\frac{B^2}{2\eta T} = \frac{\eta \rho^2}{2} \implies \eta = \sqrt{\frac{B^2}{\rho^2 T}} = \frac{B}{\rho\sqrt{T}}$$

Moreover with this choice the upper bound is

$$\frac{B^2}{2\eta T} + \frac{\eta \rho^2}{2} = \frac{B^2\sqrt{T}}{2\frac{B}{\rho}T} + \frac{B}{\rho}\frac{\rho^2}{2\sqrt{T}}$$

$$= \frac{B\rho}{2\sqrt{T}} + \frac{B\rho}{2\sqrt{T}} = \frac{B\rho}{\sqrt{T}}.$$

$\blacksquare$

Remarks.

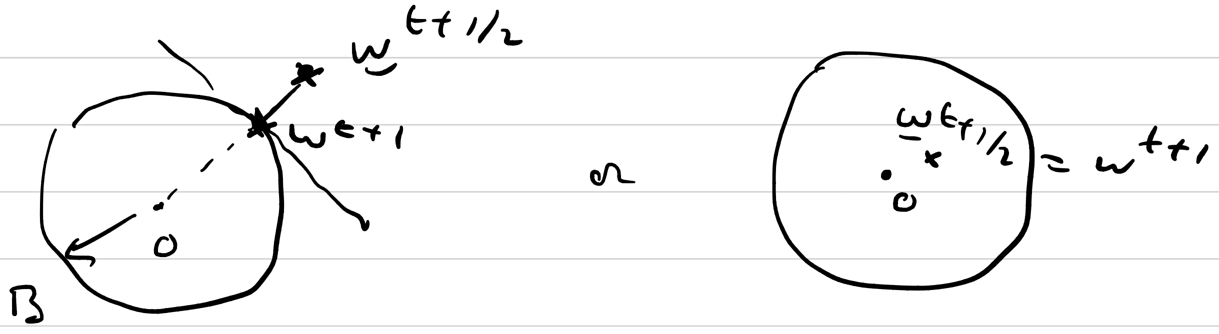
1) Instead of taking constant step size we can choose $\eta_t = \dfrac{B}{\rho\sqrt{t}}$. Steps are bigger at beginning and Thm basically unchanged.

2) In the thm $w^* \in \mathrm{Ball}(0, B)$. But the result does not garantee that $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^t$ is in this ball. If we want to make sure that it is we can modify algo by adding a projection step:

$$\begin{cases} \underline{w}^{t+\frac{1}{2}} = \underline{w}^t - \eta \underbrace{\underline{\nabla} f(\underline{w}^t)}_{\text{or } \underline{v}^t \in \partial f(\underline{w}^t)} \\ \underline{w}^{t+1} = \underset{\underline{w} \in \mathrm{Ball}(0,B)}{\arg\min} \| \underline{w} - \underline{w}^{t+1/2} \| \end{cases}$$

Because $B(0, B)$ is convex the minimizer

$\underline{w}^{t+1}$ in "projection step" is unique.



The proof of thm is almost same (see ULM).

3) One could consider other averagings instead of $\frac{1}{T}\sum_{t=1}^{T} w^t$, for example average over last αT steps.

4) <u>Notion of strong convexity:</u>

$$f(\underline{x}) \geq f(\underline{z}) + \langle \underline{\nabla} f(\underline{z}), \underline{z} - \underline{x} \rangle + \frac{\delta}{2} \| x - z \|^2$$

(excludes linear pieces and requires curvature.)

$$\Leftrightarrow \quad f(\lambda x + (1-\lambda) y) \leq \lambda f(\underline{x}) + (1-\lambda) f(y) - \frac{\delta}{2} \lambda(1-\lambda) \cdot \| x - y \|^2$$

yields better GD estimates.