Gradient Descent & its variations $\leftarrow$ Convex Optimization

First-order methods

Convex: Set, function

$$\min_{w} f(w)$$

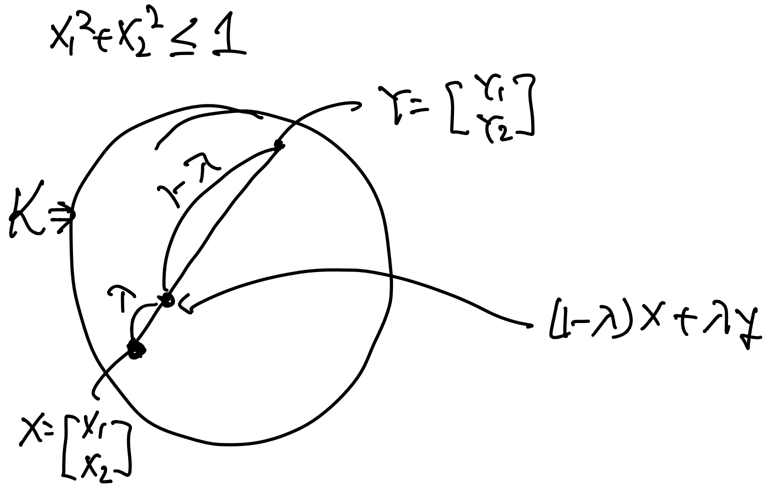
$$\text{s.t. } w \in K$$

Convex optimization if $\begin{cases} f: \text{convex} \\ K: \text{convex} \end{cases}$

Convexity

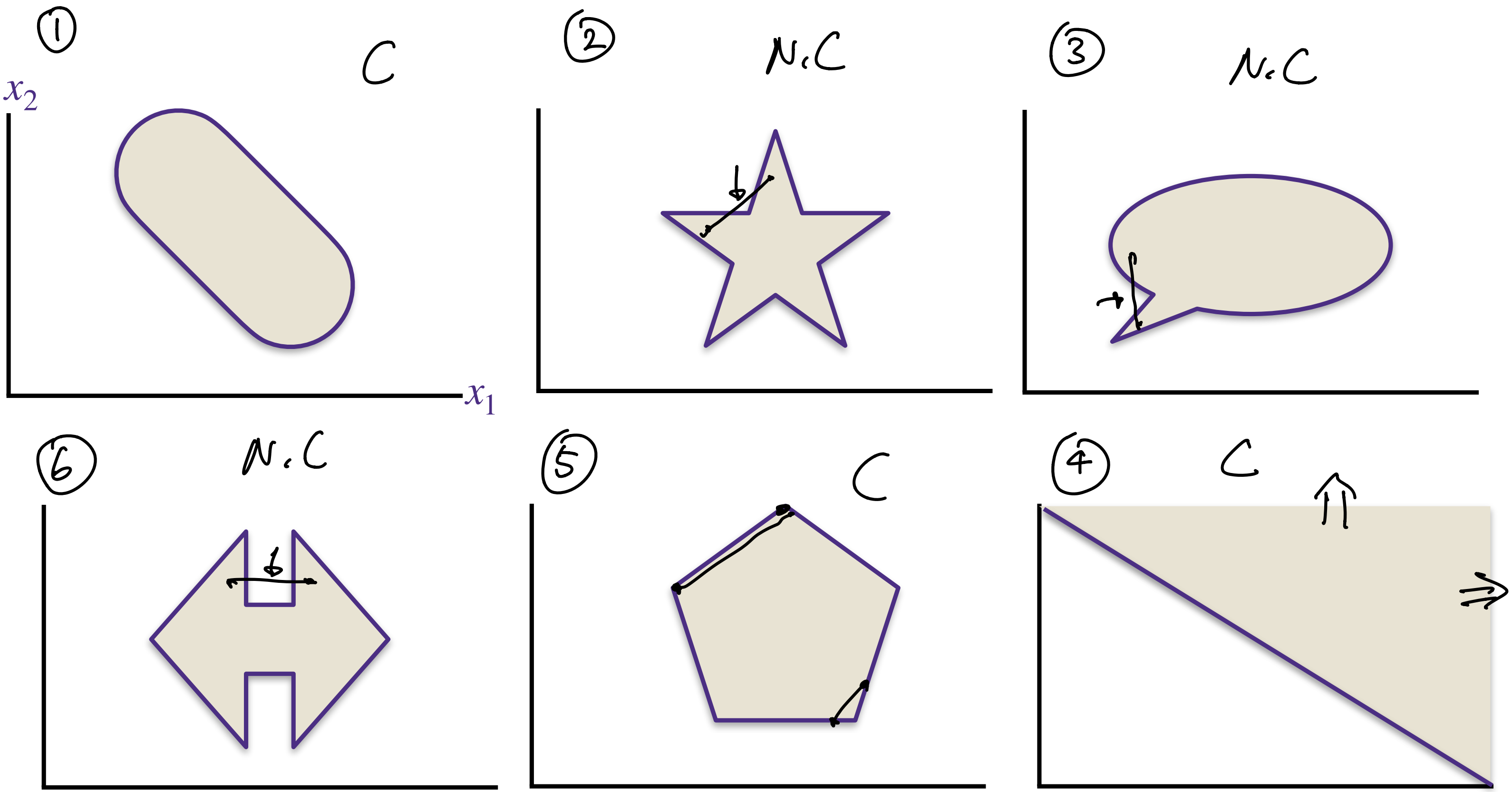
What is a convex set?

A set $K \subset \mathbb{R}^d$ is convex if $\underline{(1-\lambda)x + \lambda y \in K}$ for all $x, y \in K$ and $\lambda \in [0,1]$



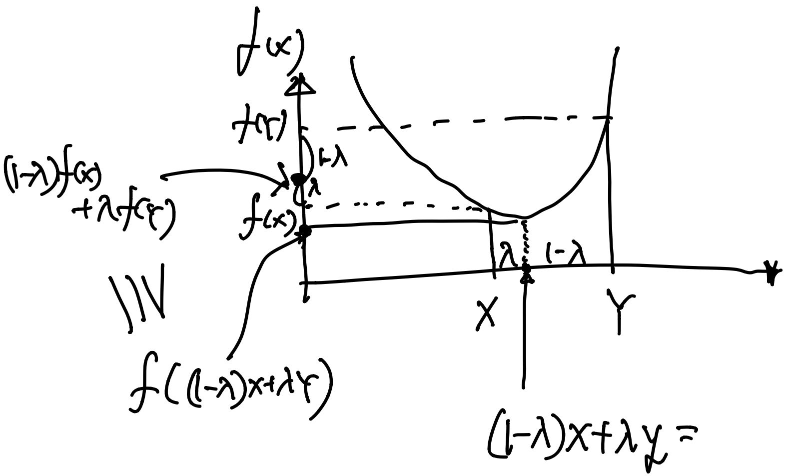
What is a convex set?

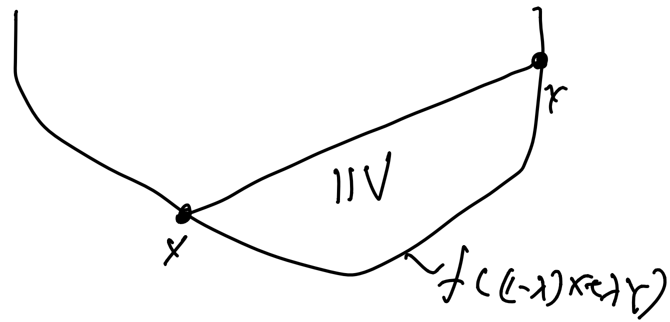
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0, 1]$



What is a convex function?

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

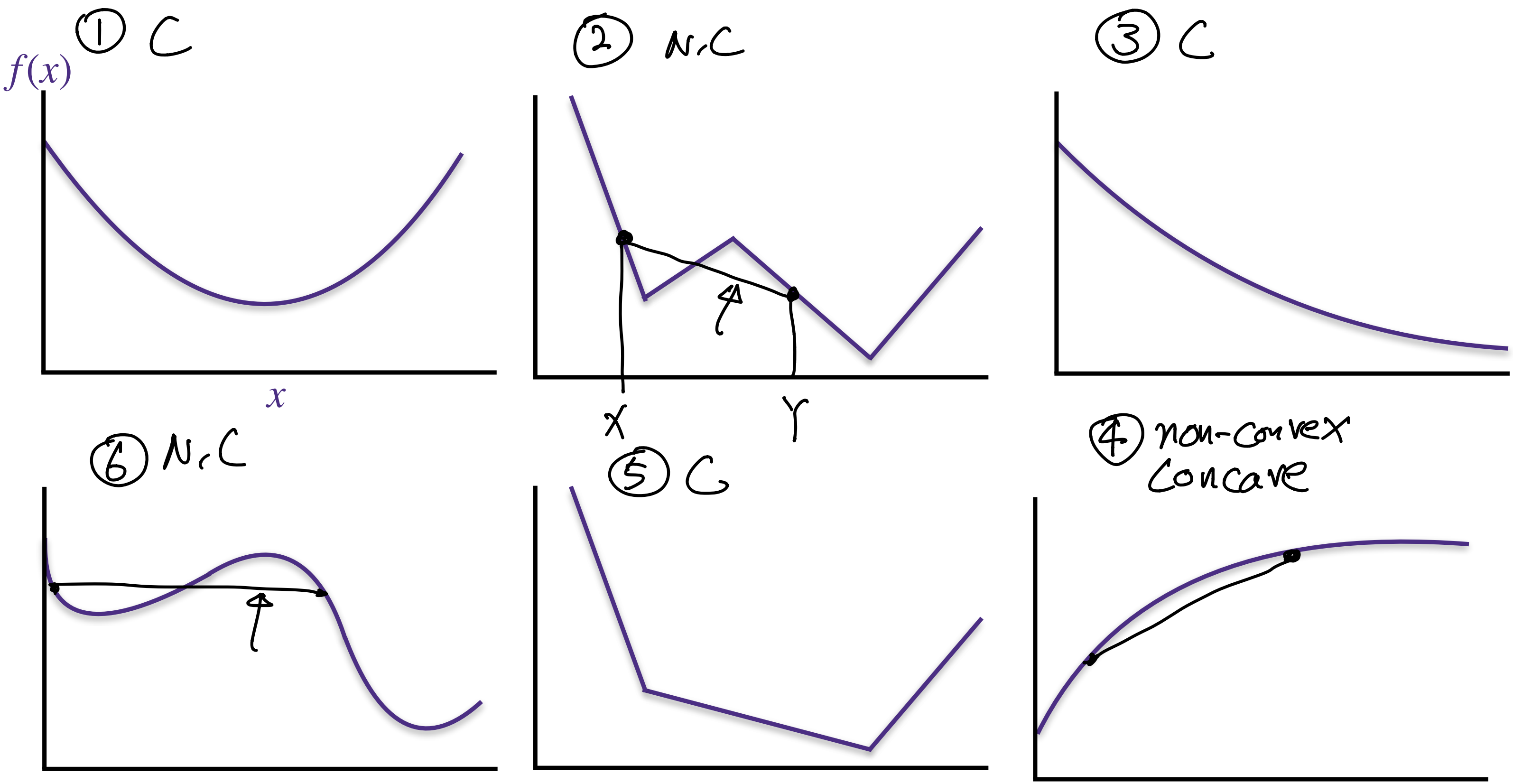




What is a convex function?

$\min\limits_{w} f(w)$ ← convex $\qquad \max\limits_{w} f(w)$ ← concave

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$



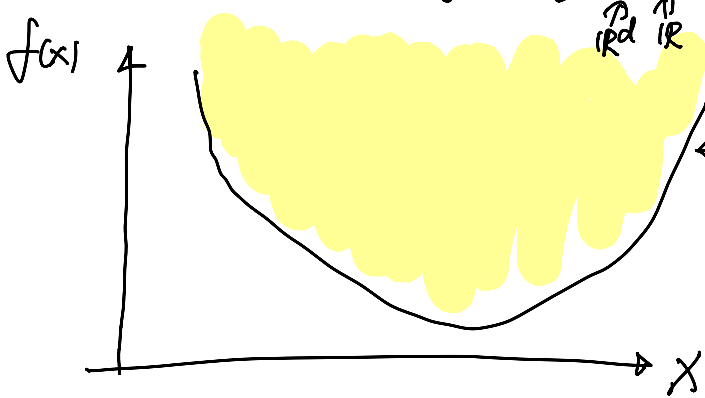
Convex functions and convex sets?

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex set

Definition: Epigraph $f = \{(\underset{\mathbb{R}^d}{x}, \underset{\mathbb{R}}{t}) \mid f(x) \leq t\}$



graph $\{(\underset{\mathbb{R}^d}{x}, \underset{\mathbb{R}}{t}) \mid f(x) = t\}$

$d = 1$

Convex functions and convex sets?

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

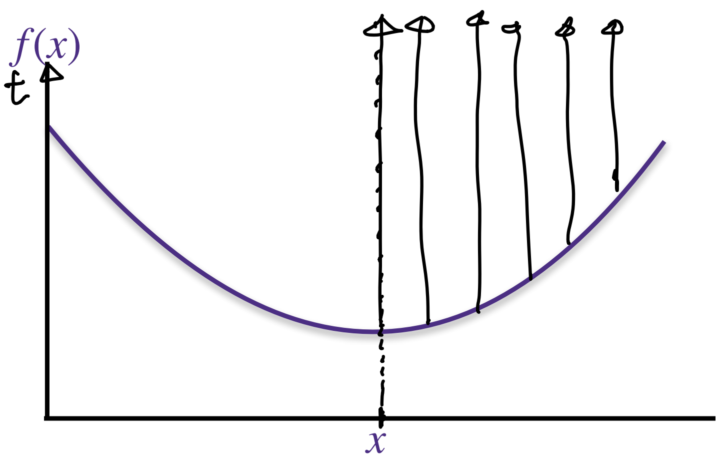
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

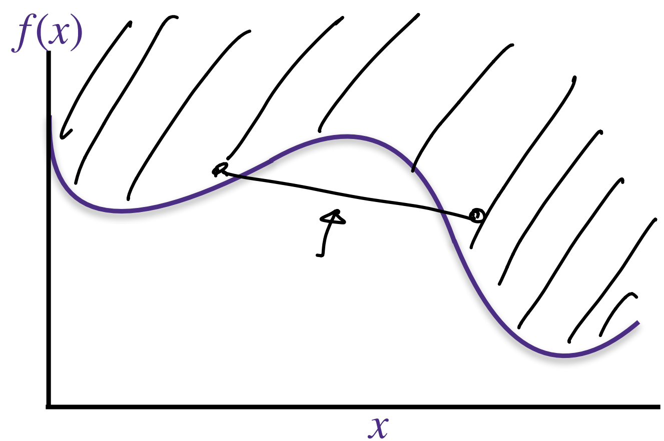
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

C

$\{t \mid t \geq f(x)\}$

N. C





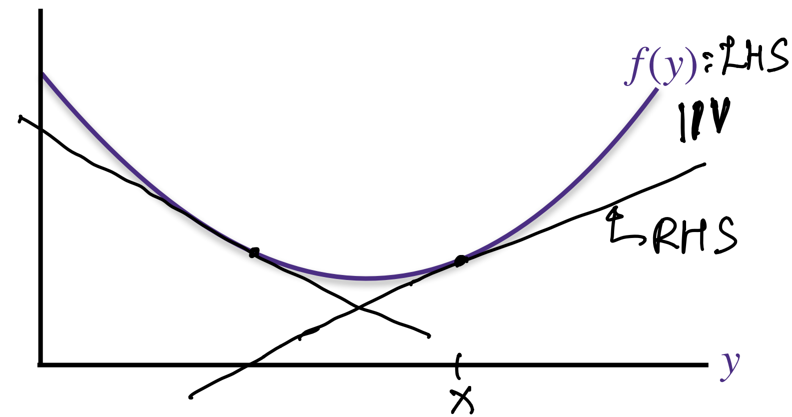
More definitions of convexity

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

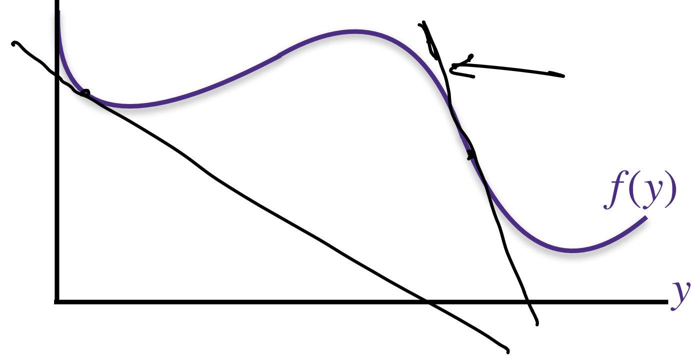
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

* A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

LHS RHS



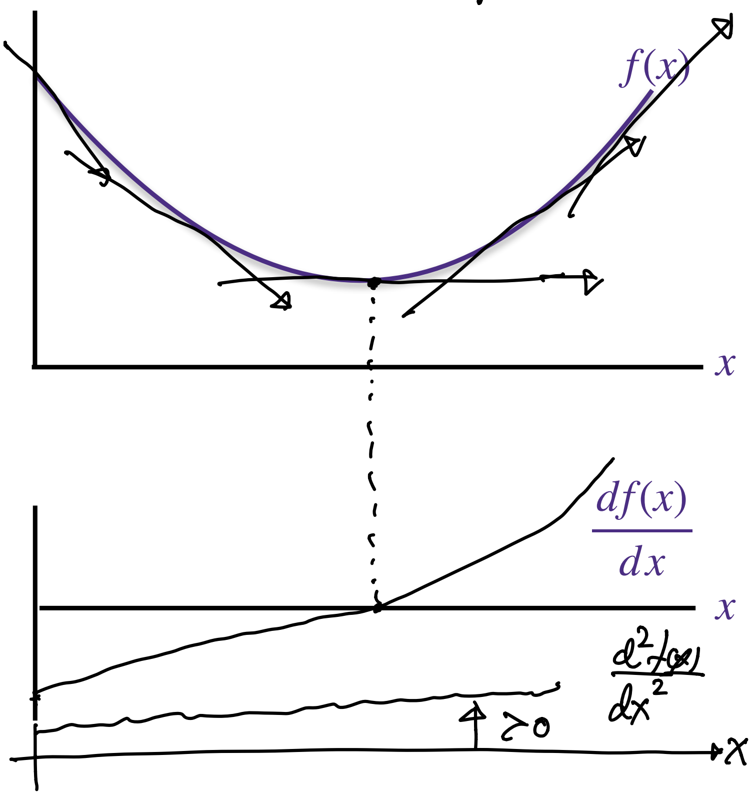
N.C.

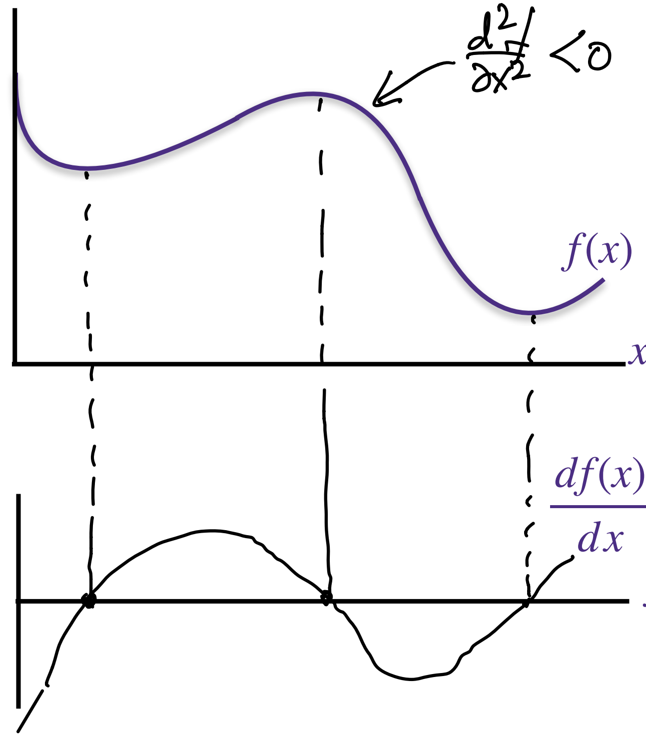


More definitions of convexity

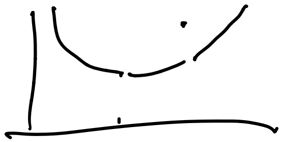
A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$

positive semidefiniteness.





More definitions of convexity



A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$

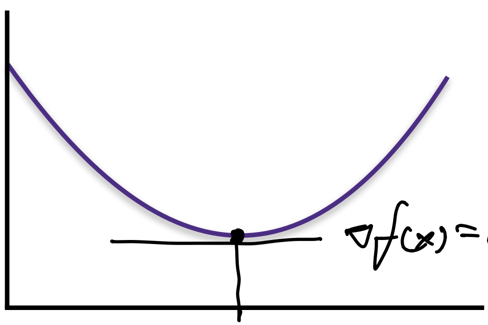
Why do we care about convexity?

Convex functions

- All local minima are global minima
- Efficient to optimize (e.g., gradient descent)

local = global minima

Convex Function

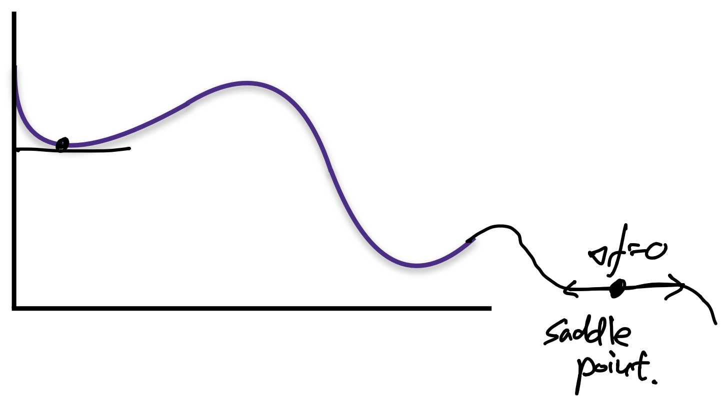


$\nabla f(x) = 0 \Longleftrightarrow$ local minima

$\Updownarrow$

global minima

Non-convex Function



Gradient Descent on $\min_w f(w)$

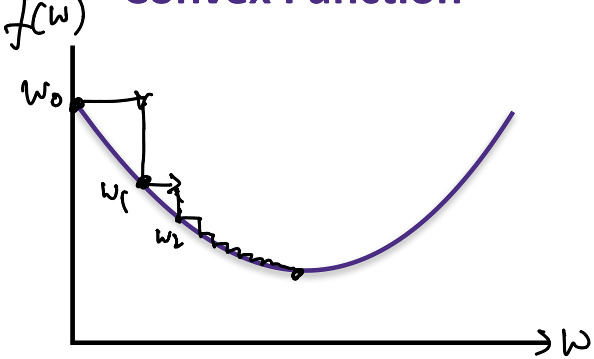
$$\text{Initialize: } w_0 = 0$$

$$\text{for } t = 1, 2, \ldots$$

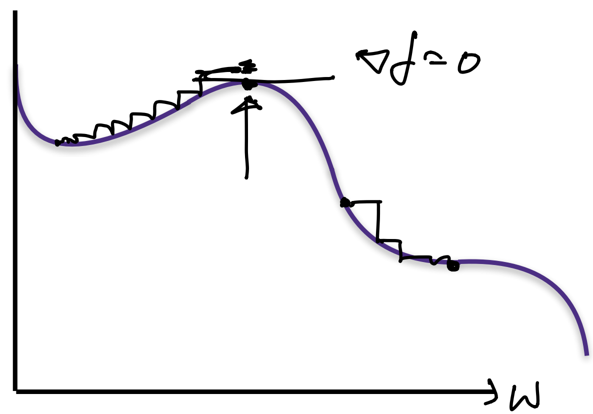
step size / learning rate

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Convex Function



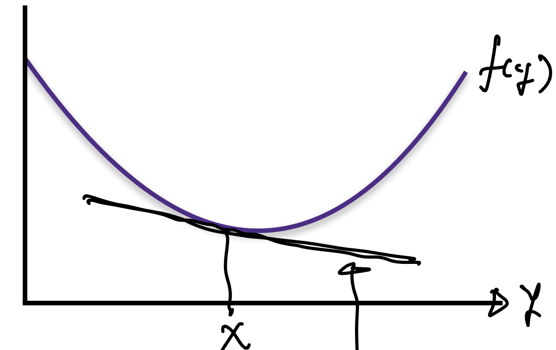
Non-convex Function



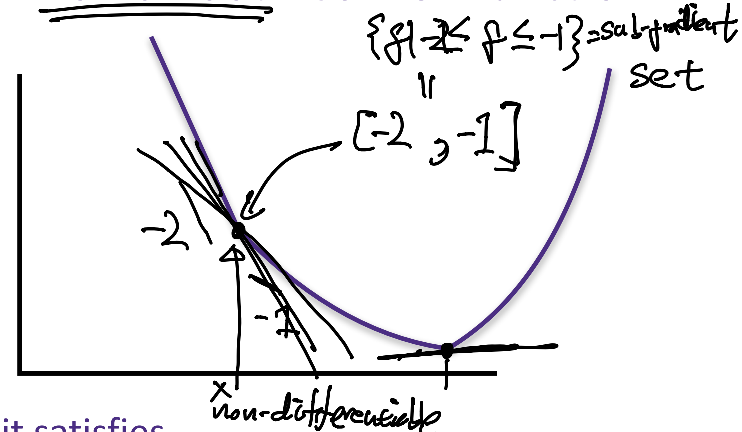
Sub-Gradient

Definition: a function is **non-smooth** if it is not differentiable everywhere

Smooth Convex Function



Non-smooth Convex Function



Definition: a vector $g \in \mathbb{R}^d$ is a **sub-gradient** at x if it satisfies

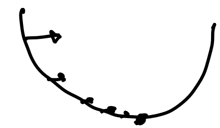
$$\left\{ g : f(y) \geq \boxed{f(x) + \underbrace{g}_{\nabla f(x)}{}^T (y - x)} \text{ for all } y \in \mathbb{R}^d \right\} = \text{sub-gradient set}$$

for smooth convex functions, the minimum is achieved at points where gradient is zero

for non-smooth convex functions, the minimum is achieved at points where sub-gradient set includes the zero vector

Sub-Gradient Descent

Gradient Descent converges with fixed η on smooth f



Sub-gradient Descent on non-smooth f.

Instead of last w_t, you keep best w_i

η_t to decrease with t

Initialize: $w_0 = 0$

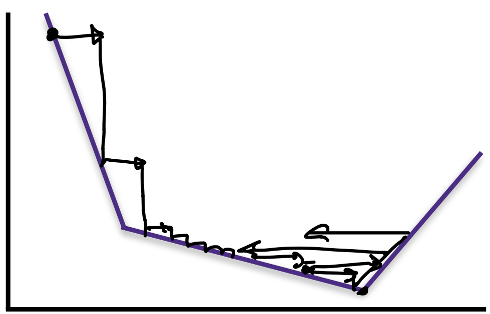
for $t = 1, 2, \ldots$

Find any g_t such that $f(y) \geq f(w_t) + g_t^\top (y - w_t)$

Current parameter

$$w_{t+1} = w_t - \overset{0.1\times}{\eta} \underset{\nabla f}{g_t}$$

Convex Function



Non-convex Function

Coordinate descent

Initialize: $w_0 = 0$

for $t = 1, 2, \ldots$

Let $i_t = t \ \% \ d$

$$w_{t+1}^{(i_t)} = w_t^{(i_t)} - \eta_t \frac{\partial f(w)}{\partial w^{(i_t)}}\Big|_{w=w_t}$$

Machine Learning Problems

- **Given data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Learning a model's parameters:** $\displaystyle\sum_{i=1}^n \ell_i(w)$

$$\text{Logistic Loss: } \ell_i(w) = \log(1 + \exp(-y_i\, x_i^T w))$$

$$\text{Squared error Loss: } \ell_i(w) = (y_i - x_i^T w)^2$$

Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \left(\frac{1}{n} \sum_{i=1}^n \ell_i(w) \right) \Big|_{w=w_t}$$

Optimization summary

- You can always run gradient descent whether f is convex or not. But you only have guarantees if f is convex
- Many bells and whistles can be added onto gradient descent such as momentum and dimension-specific step-sizes (Nesterov, Adagrad, ADAM, etc.)