

# Convexity

---

- When is an optimization (or learning) easy/fast to solve?

# Recap: Ridge vs. Lasso

---

- **Ridge**

$$\text{minimize}_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$$

- Very fast:
  - Closed form solution if used with linear models
  - Even with other loss functions, optimization is fast for squared  $\ell_2$  regularization, because  $\|w\|_2^2$  is **convex and smooth**

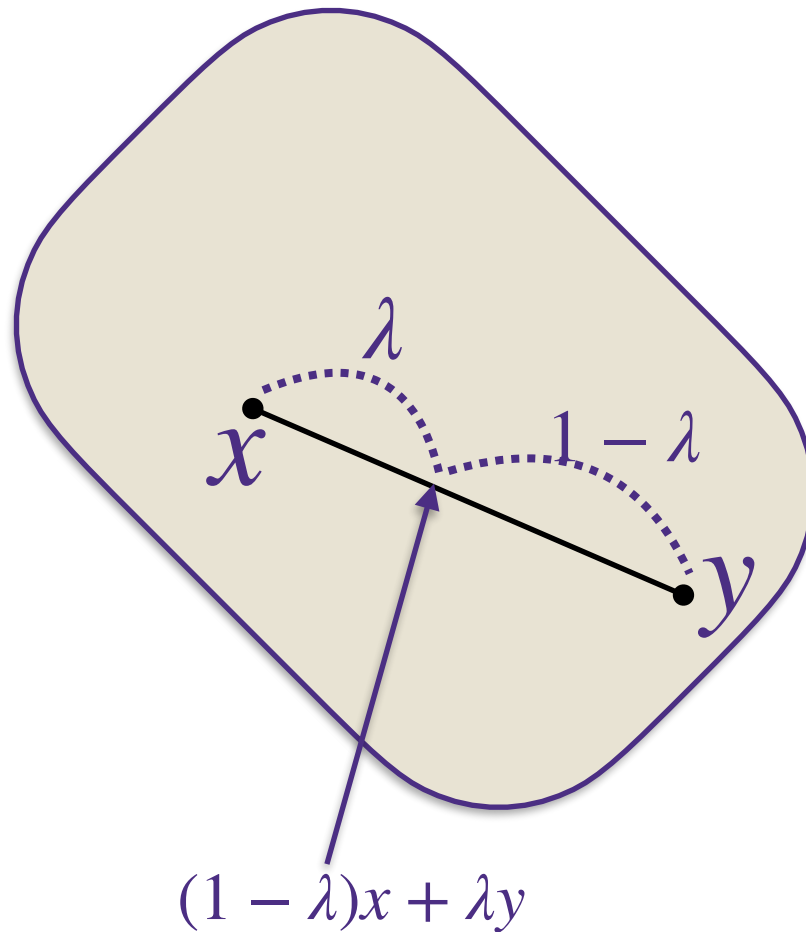
- **Lasso**

$$\text{minimize}_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_1$$

- Slower than Ridge:
  - Requires iterative optimization algorithm like sub-gradient descent
  - In particular, it is slower because  $\|w\|_1$  is **convex but non-smooth**

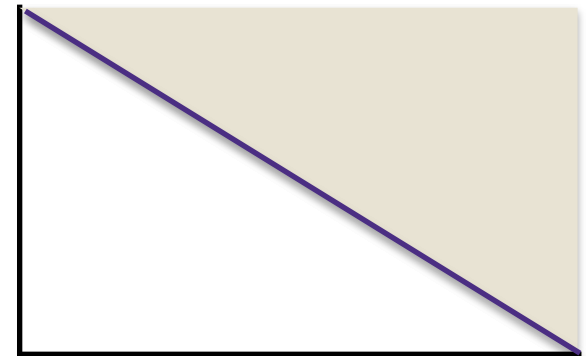
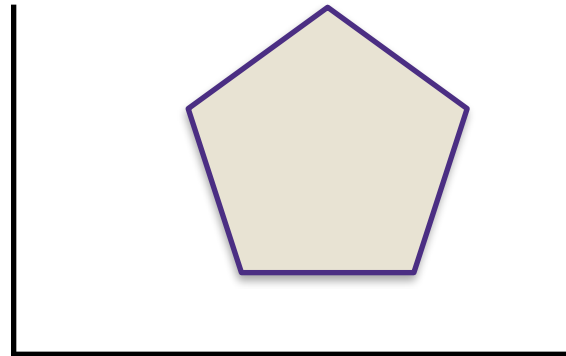
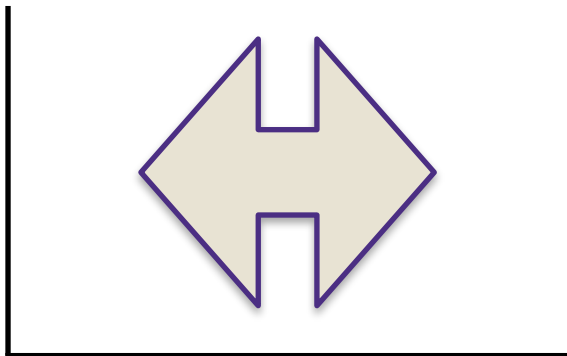
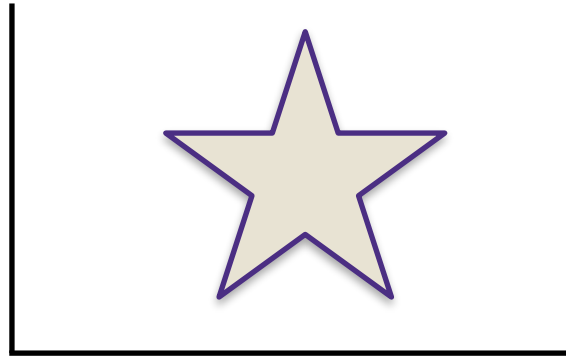
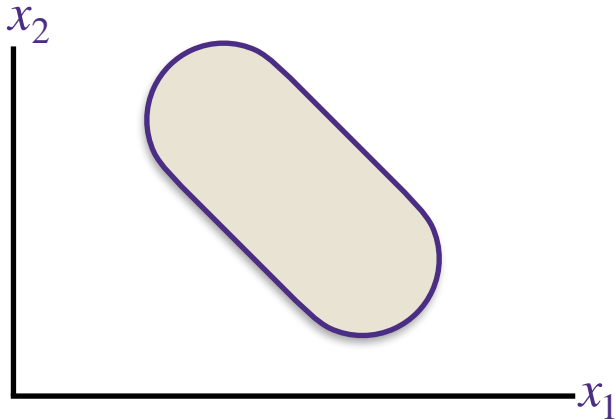
# What is a convex set?

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



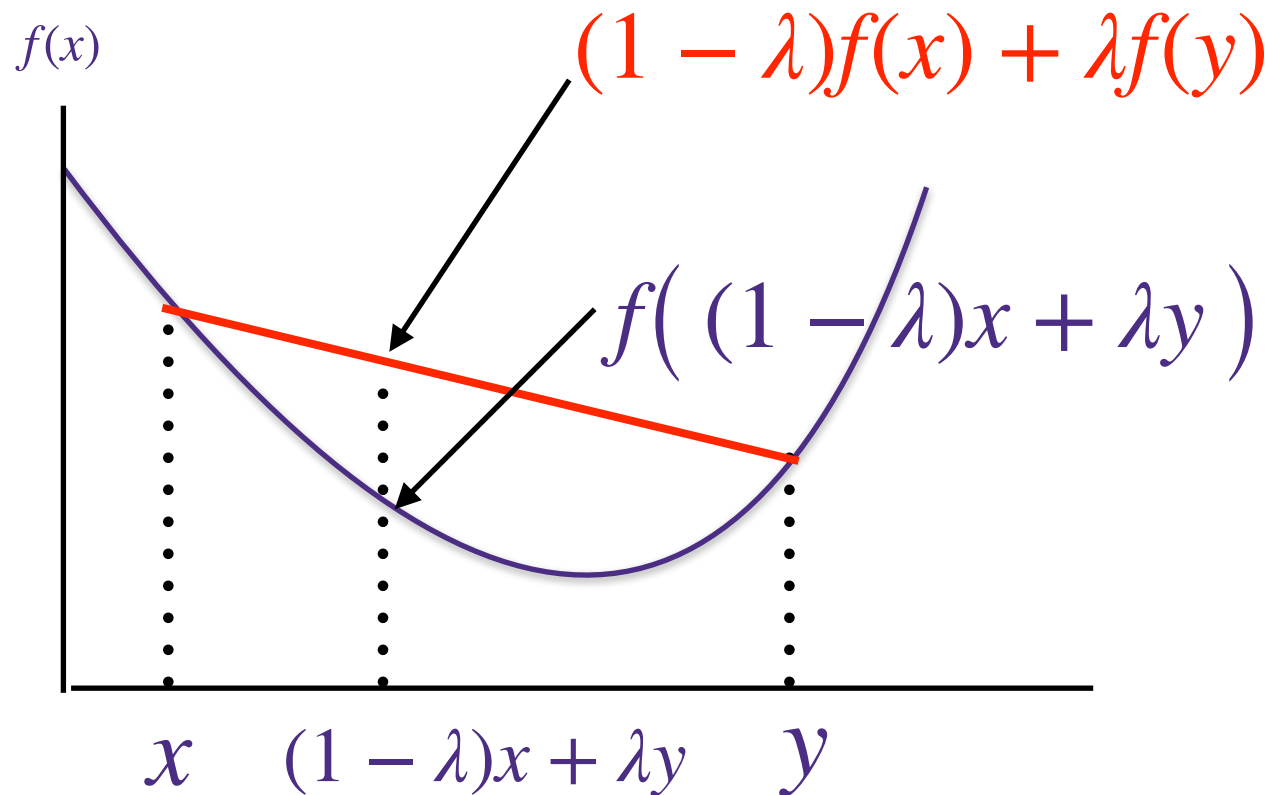
# What is a convex set?

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?

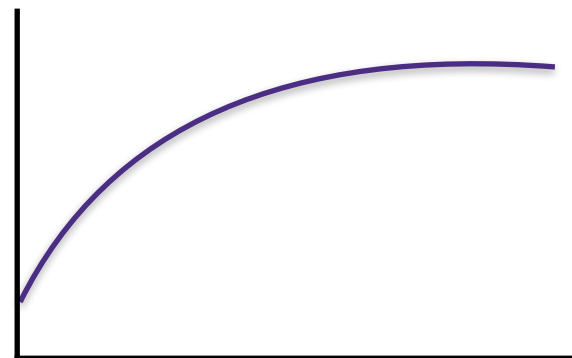
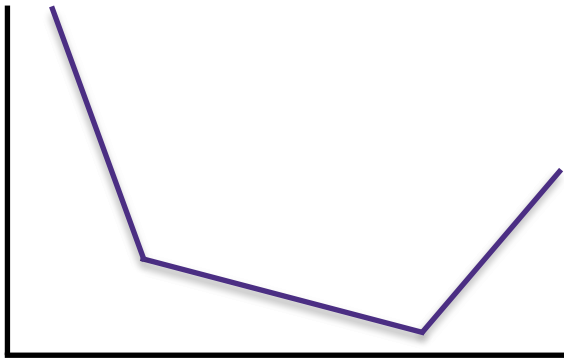
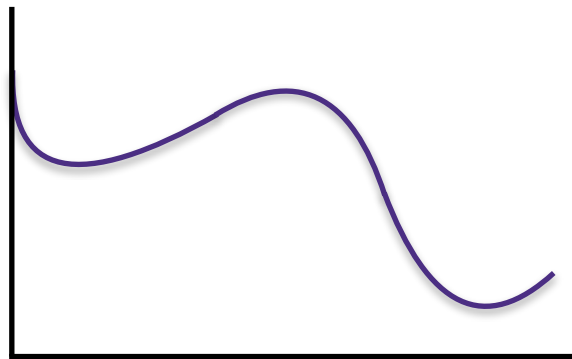
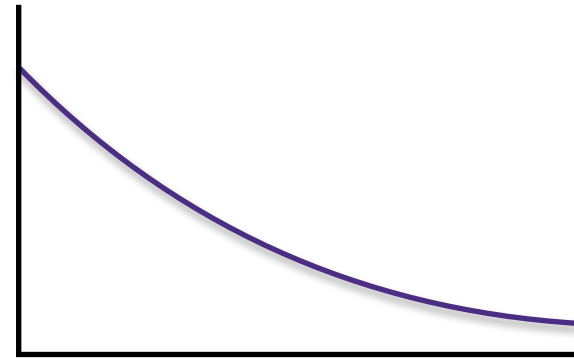
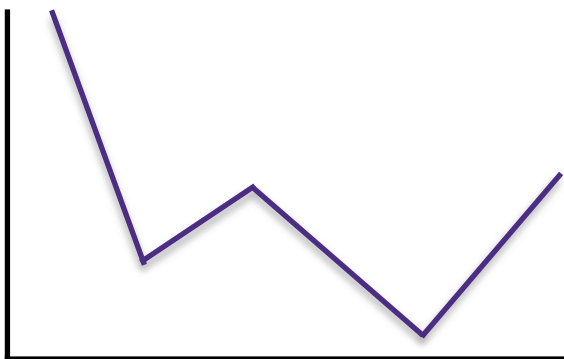
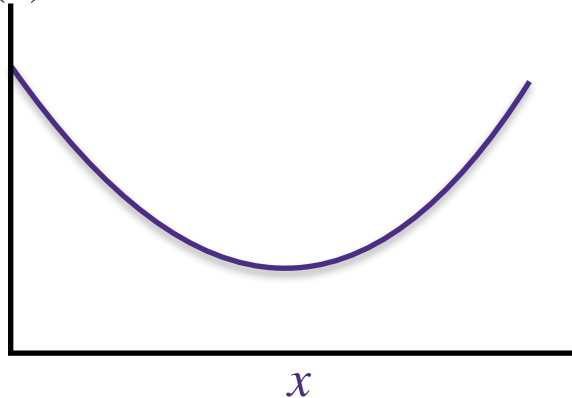
A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex if  $f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y)$  for all  $x, y \in \mathbb{R}^d$  and  $\lambda \in [0, 1]$



# What is a convex function?

A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $\lambda \in [0, 1]$

$f(x)$



# Convex functions and convex sets?

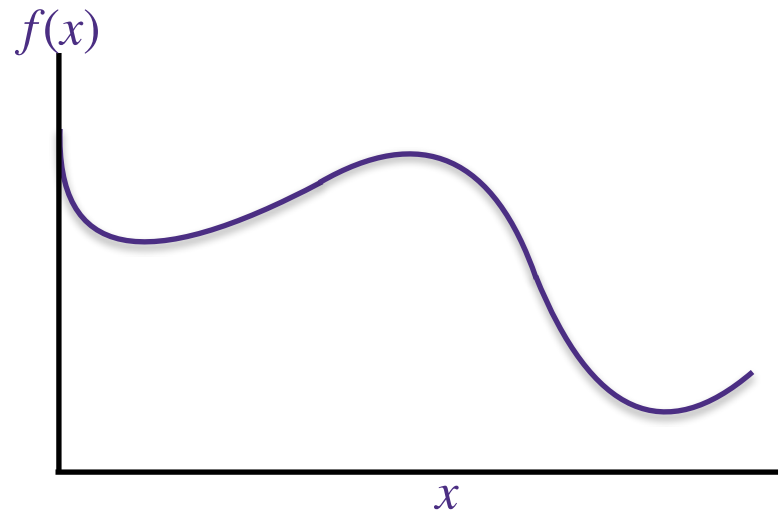
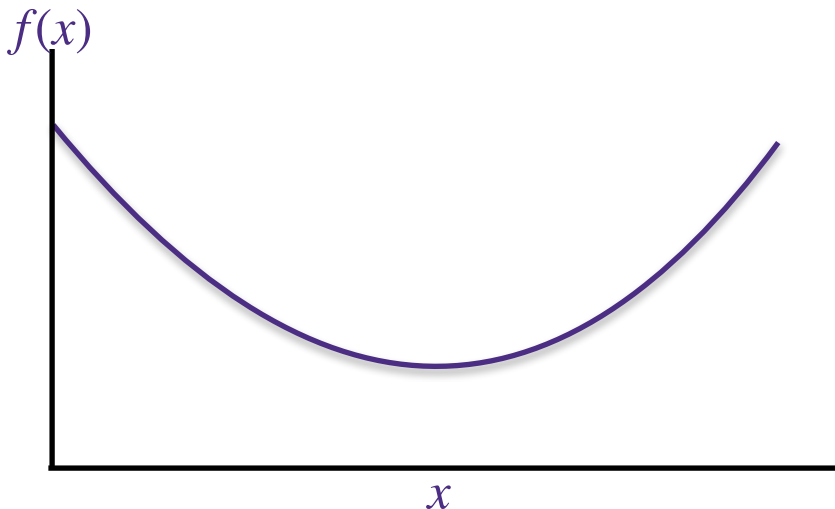
A set  $K \subset \mathbb{R}^d$  is convex if  $(1 - \lambda)x + \lambda y \in K$  for all  $x, y \in K$  and  $\lambda \in [0, 1]$

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

Graph of  $f$  is defined as  $\{(x, t) : f(x) = t\}$

Epigraph of  $f$  is defined as  $\{(x, t) : f(x) \leq t\}$

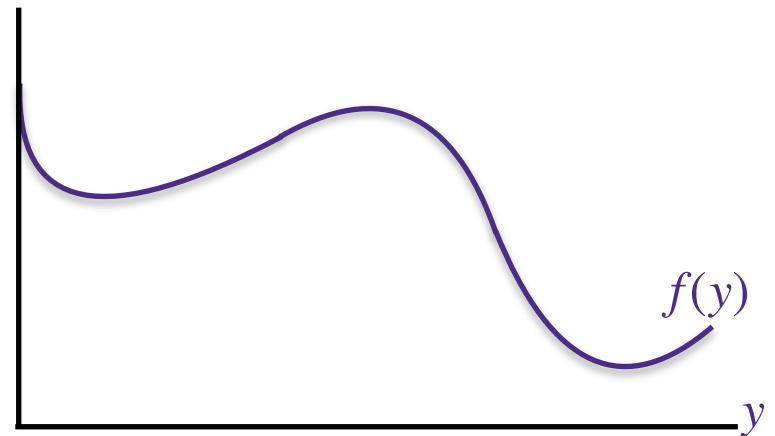
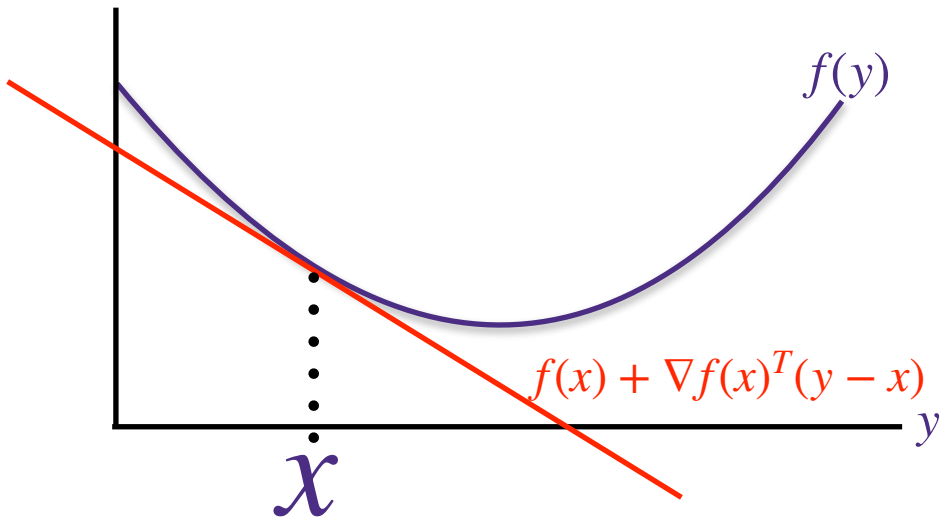


# More definitions of convexity

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

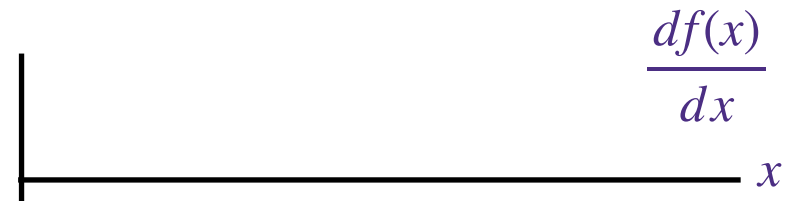
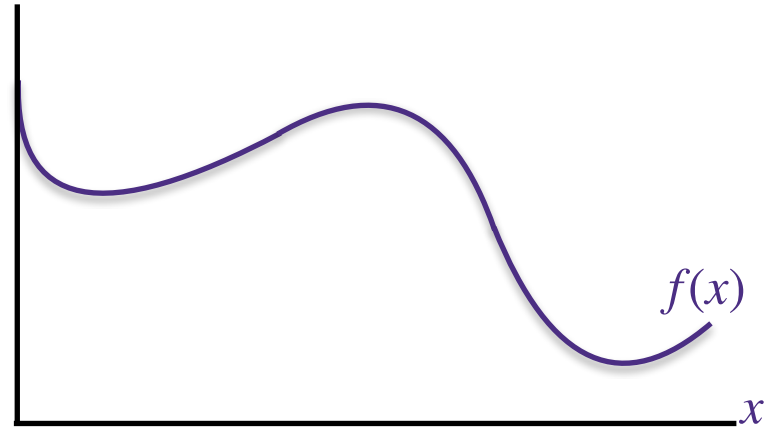
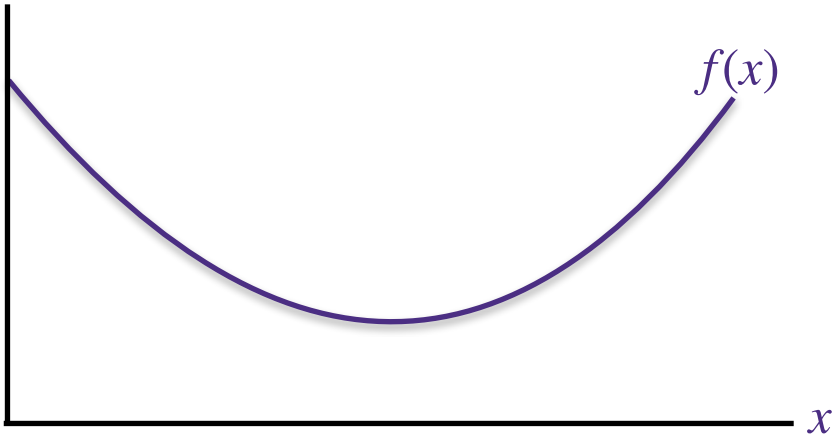
A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex if the set  $\{(x, t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$  is convex

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



# More definitions of convexity

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



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

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

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

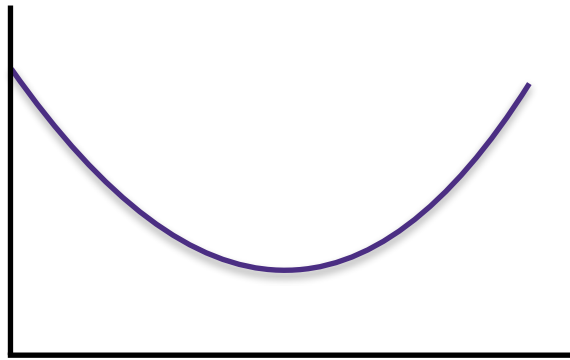
# Why do we care about convexity?

---

## Convex functions

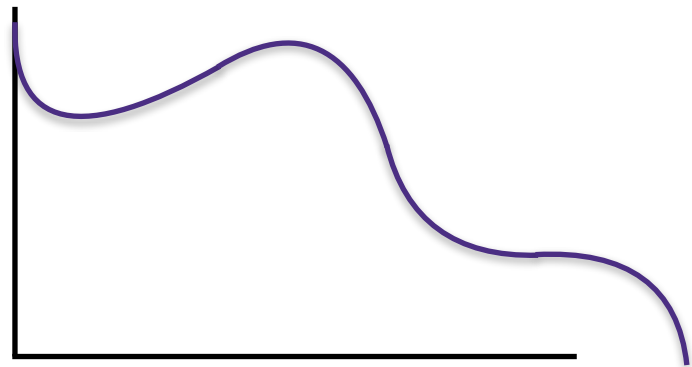
- All local minima are global minima
- Efficient to optimize (e.g., gradient descent)

**Convex Function**



We only need to find a point with  $\nabla f(x) = 0$ , which for convex functions implies that it is a local minima and a global minima

**Non-convex Function**



For non-convex functions, a stationary point with  $\nabla f(x) = 0$  could be a local minima, a local maxima, or a saddle point

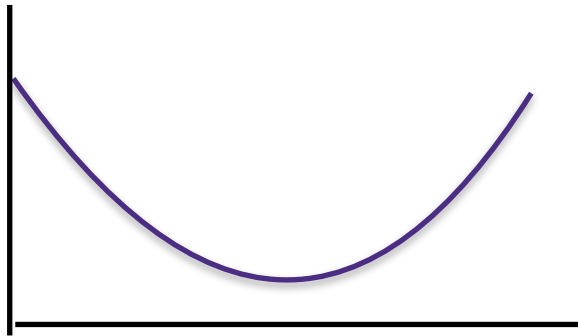
# Gradient Descent on $\min_w f(w)$

Initialize:  $w_0 = 0$

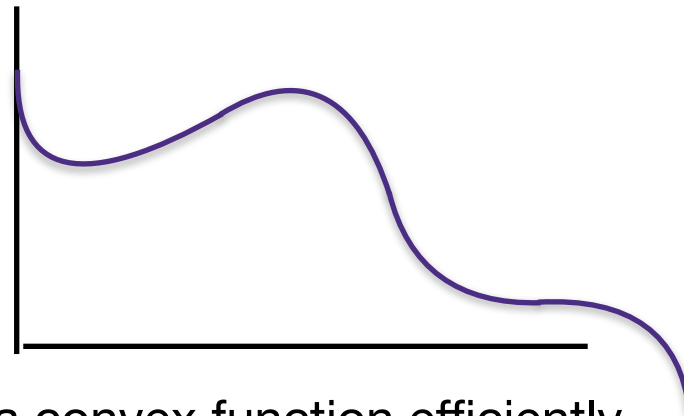
for  $t = 1, 2, \dots$

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

**Convex Function**



**Non-convex Function**



- Strength: Can find global minima of a convex function efficiently
- Weakness: Can only be applied to smooth functions
  - i.e., functions that is differentiable everywhere,
  - otherwise  $\nabla f(x)$  is not defined and gradient descent cannot be applied

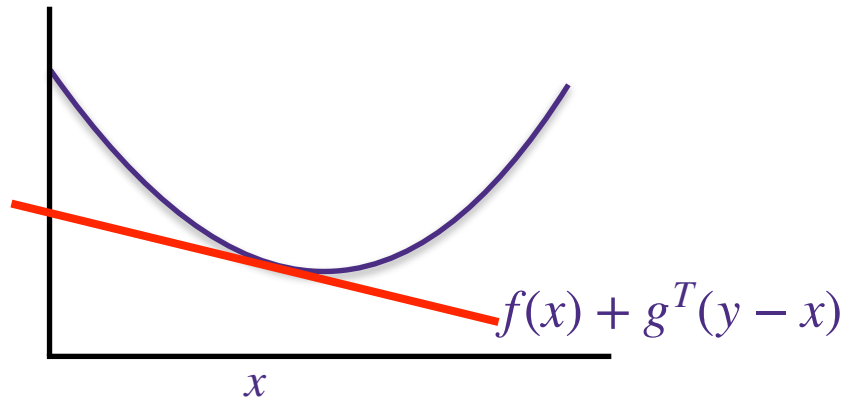
# Sub-Gradient

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

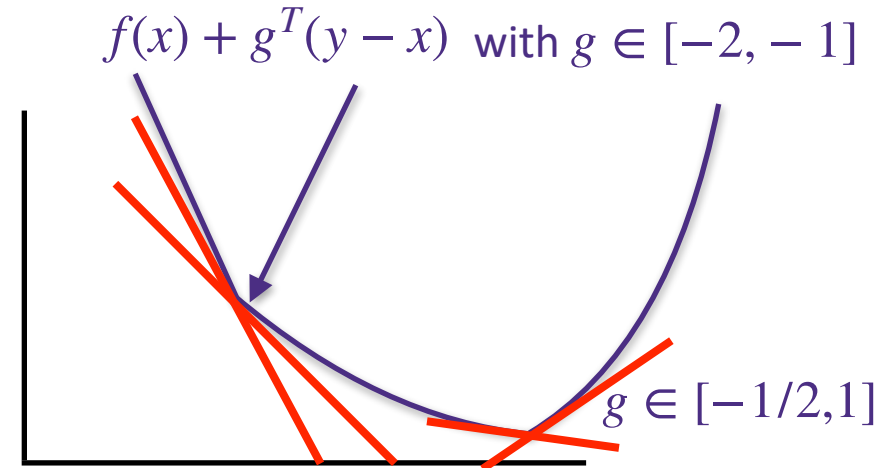
Definition: a vector  $g \in \mathbb{R}^d$  is a **sub-gradient** at  $x$  if it satisfies

$$f(y) \geq f(x) + g^T(y - x) \text{ for all } y \in \mathbb{R}^d$$

## Smooth Convex Function



## Non-smooth Convex Function



- for smooth convex functions,

- gradient is the unique sub-gradient, and
- the global 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 for non-smooth functions

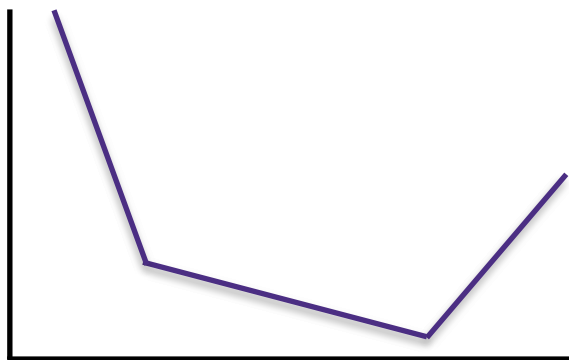
Initialize:  $w_0 = 0$

for  $t = 1, 2, \dots$

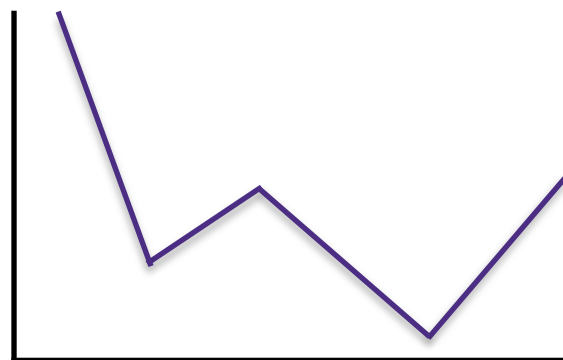
Find any  $g_t$  such that  $f(y) \geq f(w_t) + g_t^\top (y - w_t)$

$$w_{t+1} \leftarrow w_t - \eta_t g_t$$

**Convex Function**



**Non-convex Function**



- Strength: finds global minima for **non-smooth convex functions**
- Weakness: it is slower than gradient descent on convex smooth functions, because the gradient do not get smaller near the global minima
  - Instead of last iterate  $w_t$ , we use the best one we saw in all iterates
  - The stepsize needs to decrease with  $t$

# Optimization

---

- 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.)

# Questions?

---