Prediction pitfalls

Matt Golub
Hunter Schafer

UNIVERSITY of WASHINGTON

Linear coefficients

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

$\langle a, b \rangle \equiv a^T b$

Claim: $|\widehat{\theta}_j| > |\widehat{\theta}_k|$ means feature j is more important than feature k to the model.

Linear coefficients

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

Claim: $|\widehat{\theta}_j| > |\widehat{\theta}_k|$ means feature j is more important than feature k to the model.

False: if I rescaled jth dimension of data by some large constant (i.e., instead of square-feet, I change it to square-inches so that $x_{i,j} \to 144 x_{i,j}$) the magnitude of $\widehat{\theta}_j \to \widehat{\theta}_j / 144$ but the predictive power did not change!

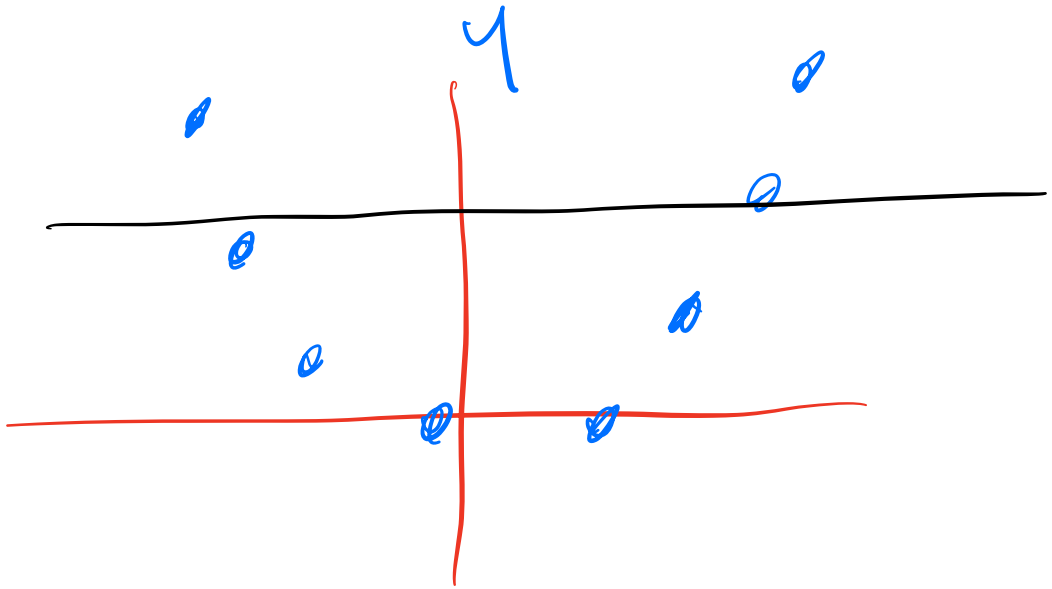
This is why we frequently **normalize** our data by mean and variance:

$x_{i,j} \to (x_{i,j} - \mu_j)/\sigma_j$ where $\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$ and $\sigma_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \mu_j)^2$

Linear coefficients

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

Claim: If $\widehat{\theta}_j = 0$ then the jth feature has no predictive power of target y



Linear coefficients

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

Claim: If $\widehat{\theta}_j = 0$ then the jth feature has no predictive power of target y

False: Suppose feature 1 is #bathrooms and feature 2 is #toilets. Reasonable to assume these features extremely **correlated** so any choice of following gives same answer: $(\widehat{\theta}_1, \widehat{\theta}_2) = (\alpha, 0)$, $(\widehat{\theta}_1, \widehat{\theta}_2) = (0, \alpha)$, $(\widehat{\theta}_1, \widehat{\theta}_2) = (\frac{1}{3}\alpha, \frac{2}{3}\alpha)$

Correlation is not causation

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

Claim: Suppose $\widehat{\theta}_j = 30{,}000$ and the jth feature represents #fireplaces. If I add m fireplaces in my house, I can expect to sell my house for $30{,}000m$ more!

Correlation is not causation

Given data $\{(x_i, y_i)\}_{i=1}^n$ consider a linear model: $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2$

Claim: Suppose $\widehat{\theta}_j = 30{,}000$ and the *j*th feature represents #fireplaces. If I add m fireplaces in my house, I can expect to sell my house for $30{,}000m$ more!

False:

1) Correlation is not causation. If $\widehat{\theta}_j$ is large, it is correlated with the house price, but it is not necessarily the *cause* for the house price. More expensive houses are typically bigger and size correlates with more fireplaces. Absurd example: fireman and fires seem to co-occur. Do fireman cause fire?

2) Train != Test dataset. A linear model may fit your data perfectly, but that doesn't mean it accurately extrapolates to data outside your training data. Moreover, this is impossible to know just using your training data.

How was the dataset collected?

Citizen reporting:
In the early 2010s, the city of Boston wanted to repair pot holes but wanted to allocate resources as efficiently as possible. So they released a smart phone app that automatically detects potholes via accelerometer data and sends back the GPS coordinates.

Claim: By fixing the potholes that are reported most frequently, resources are allocated to minimize the greatest number of total interactions with potholes

Crawford “The Hidden Biases in Big Data” https://hbr.org/2013/04/the-hidden-biases-in-big-data

How was the dataset collected?

Citizen reporting:
In the early 2010s, the city of Boston wanted to repair pot holes but wanted to allocate resources as efficiently as possible. So they released a smart phone app that automatically detects potholes via accelerometer data and sends back the GPS coordinates.

Claim: By fixing the potholes that are reported most frequently, resources are allocated to minimize the greatest number of total interactions with potholes

False. Counts are correlated with geographical areas with more smartphone ownership, which is correlated with wealthier areas versus those less wealthy or more elderly.

Crawford “The Hidden Biases in Big Data” https://hbr.org/2013/04/the-hidden-biases-in-big-data

ML automates bias of humans

Hiring example:
In an effort to avoid bias in its hiring practices, a company trains a machine learning model to predict Y from X where
X = resume
Y = whether applicant was hired, or job performance

Claim: By using a data-driven process, any bias of the human resume screener is avoided.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

ML automates bias of humans

Hiring example:
In an effort to avoid bias in its hiring practices, a company trains a machine learning model to predict Y from X where
X = resume
Y = whether applicant was hired, or job performance on the job

Claim: By using a data-driven process, any bias of the human resume screener is avoided.

False. The label Y (e.g., job performance) was evaluated by a human. If bias existed in these labels then all you've done is automated a system that is just as biased.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Generally good practices

- Normalize your data.
- Check for correlated features.
- Check for features with zero variance.
- Understand your data.
  - How was it collected?
  - What was experimental design?
  - What is the “right” experimental design to answer your question?
  - Visualize your data (before and after preprocessing).
- Check the quality of your model.
  - Quantify performance.
  - How does performance compare to that of simpler models?
  - Visualize model predictions. Visually compare to ground truth labels.

Further reading

Excellent textbook:
"*Fairness and Machine learning*" Solon Barocas, Moritz Hardt, Arvind Narayanan https://fairmlbook.org/

Relevant conferences:
ACM Conference on Fairness, Accountability, and Transparency https://facctconference.org/

Convexity

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

Recap: Ridge vs. Lasso

- **Ridge**

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

  - Very fast:
    - Closed form solution if used with linear models
    - Even with other loss functions, optimization is fast for squared ℓ_2 regularization, because $\|w\|_2^2$ is **convex and smooth**

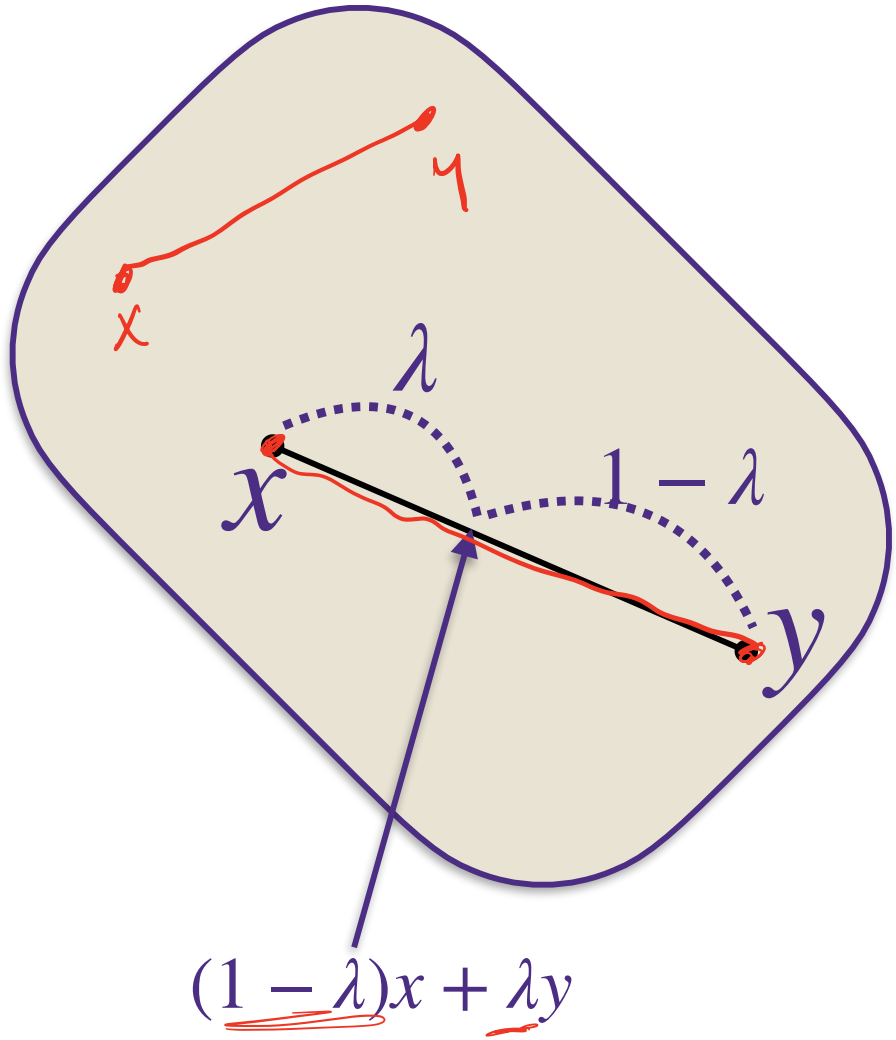
- **Lasso**

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

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

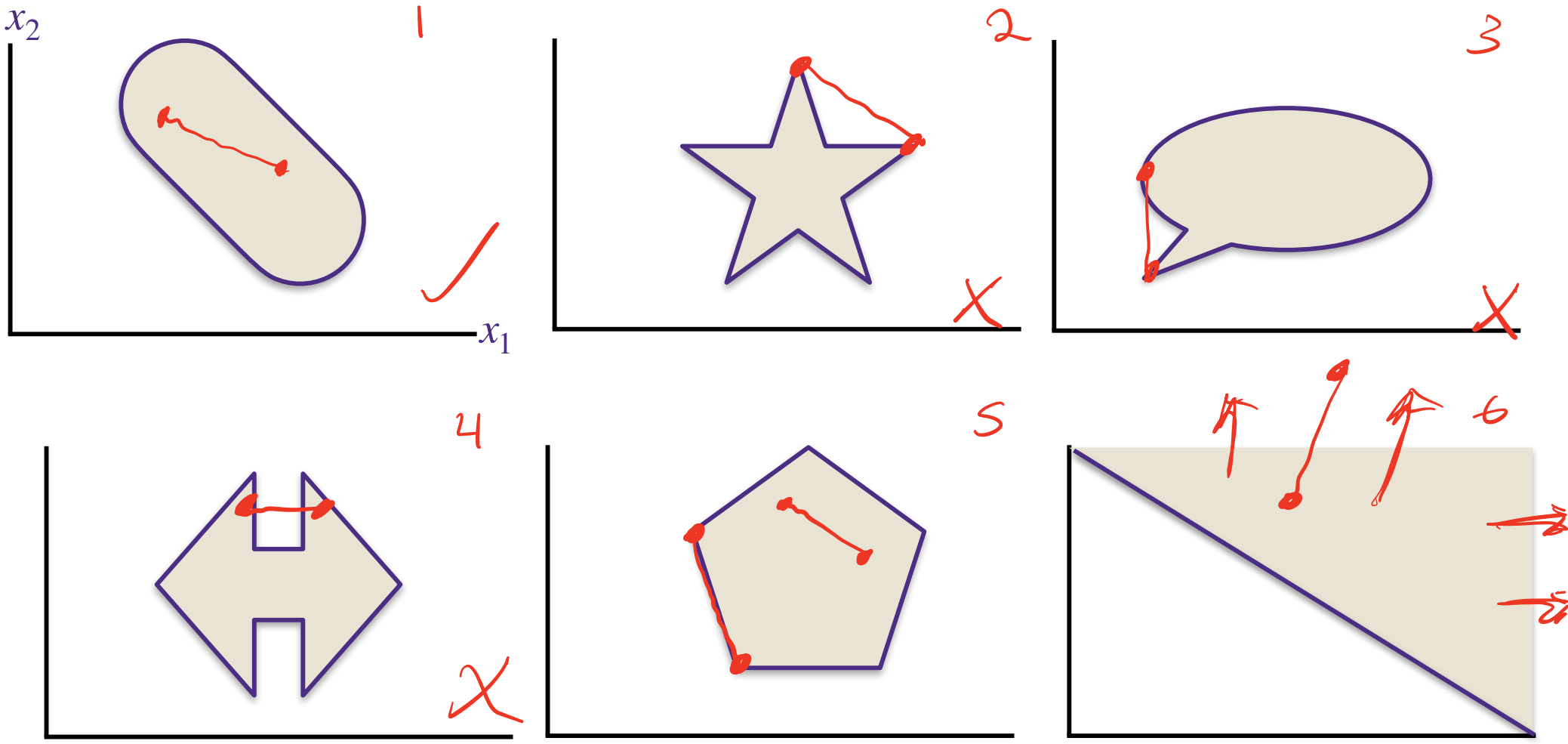
What is a convex set?

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



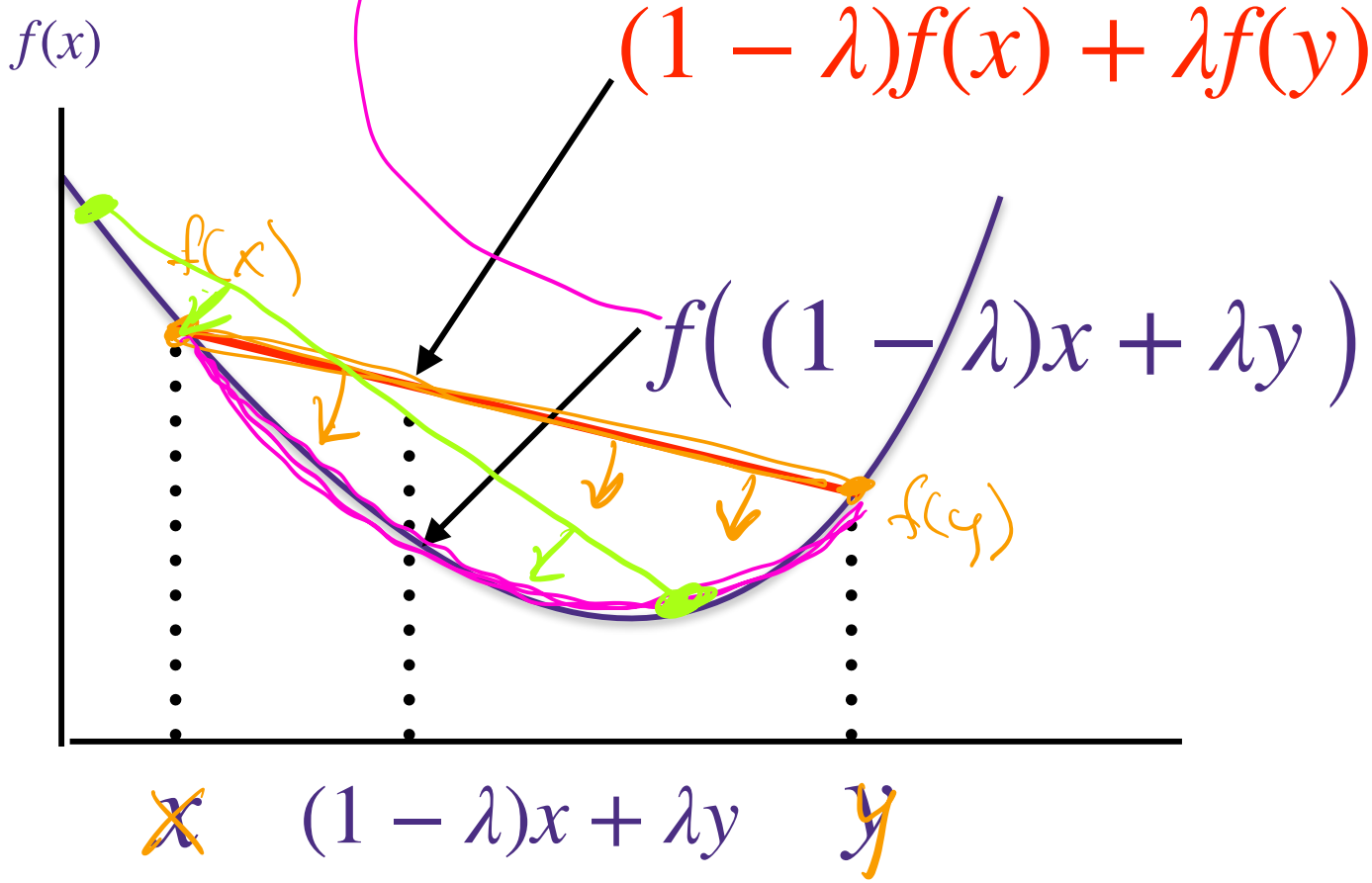
What is a convex set?

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



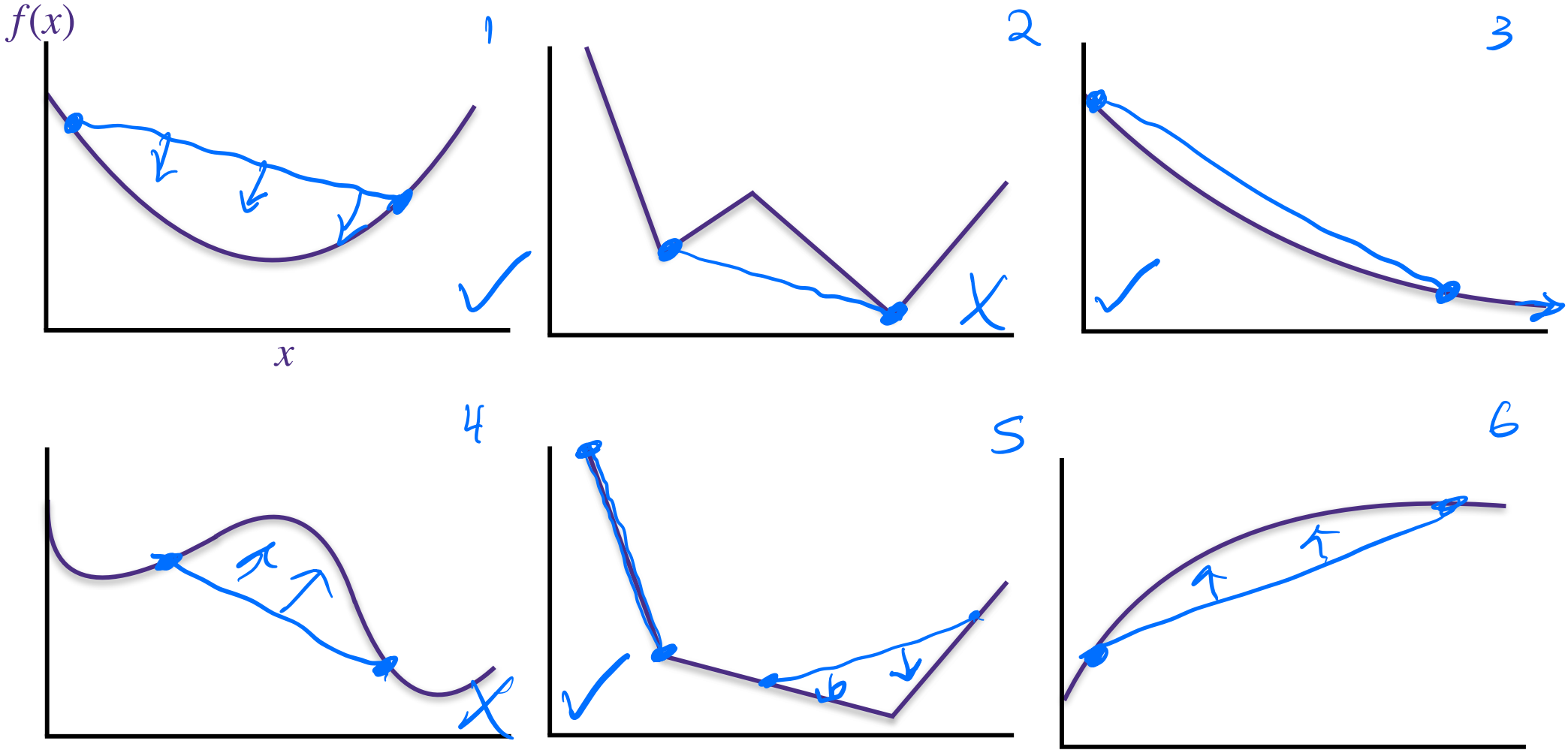
What is a convex function?

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



Convex functions and convex sets?

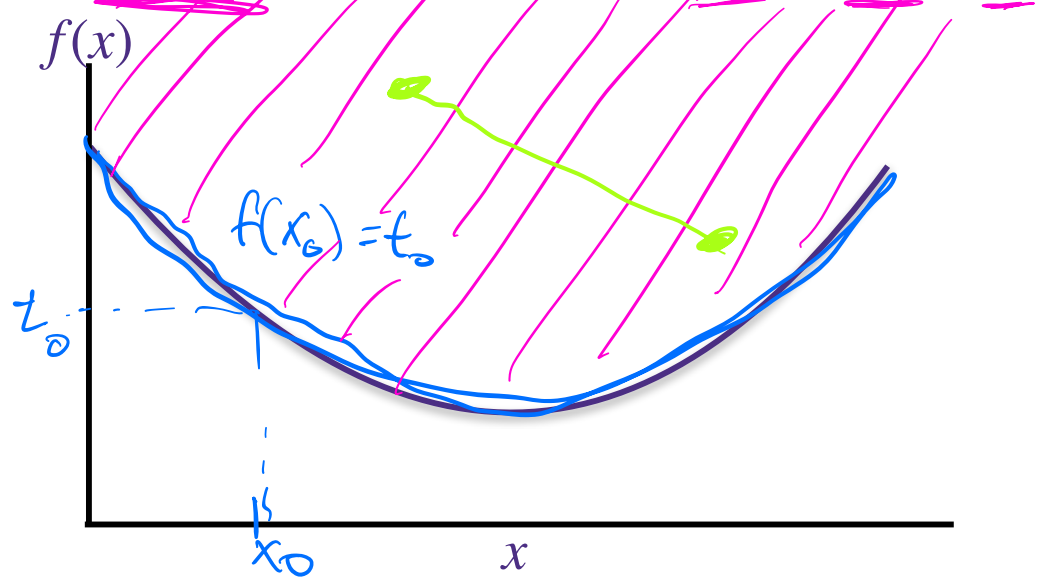
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

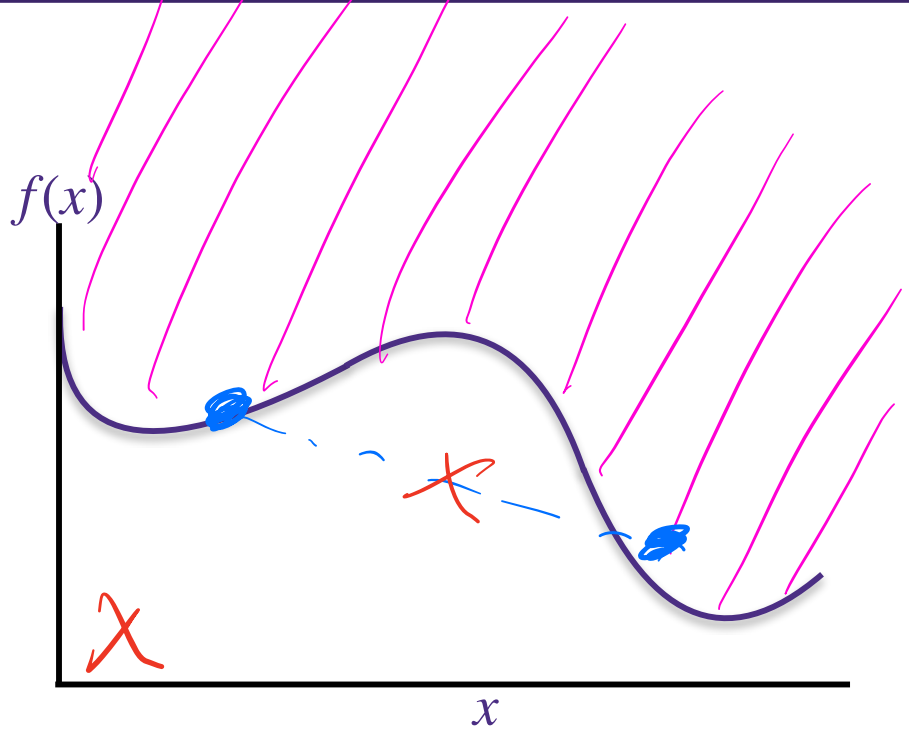
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

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

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

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



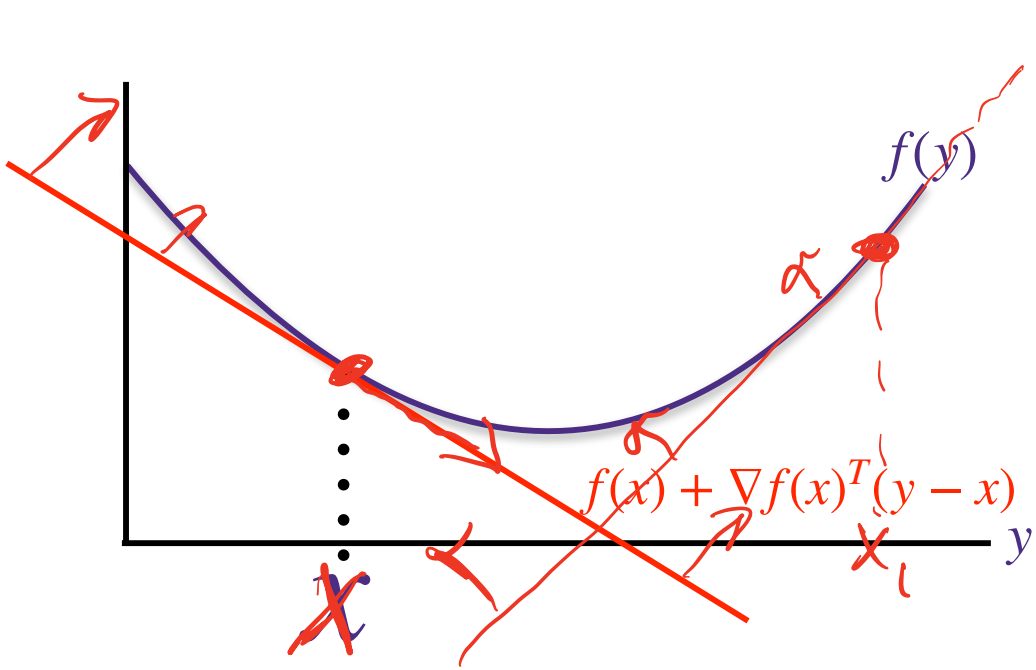


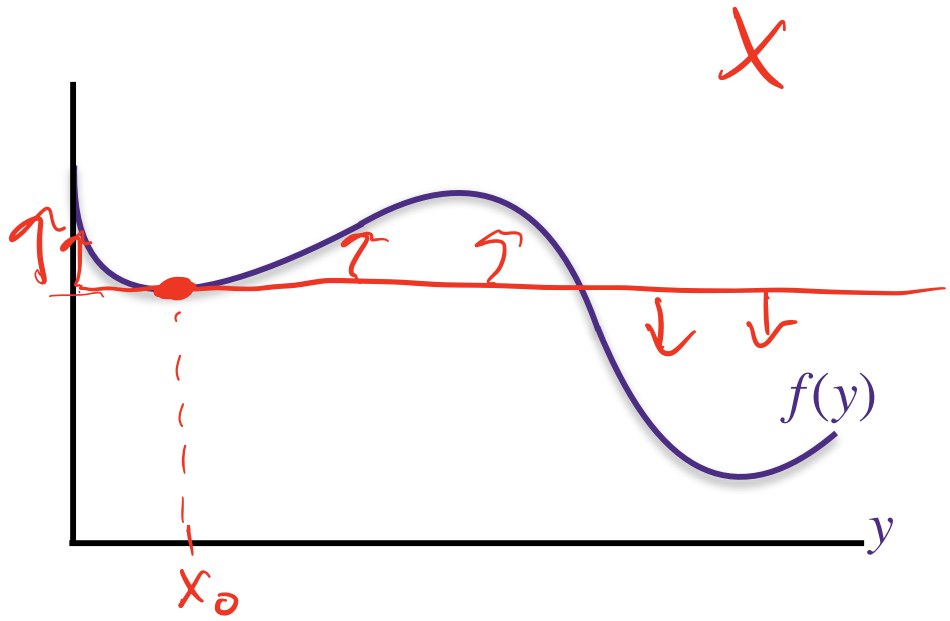
More definitions of convexity

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

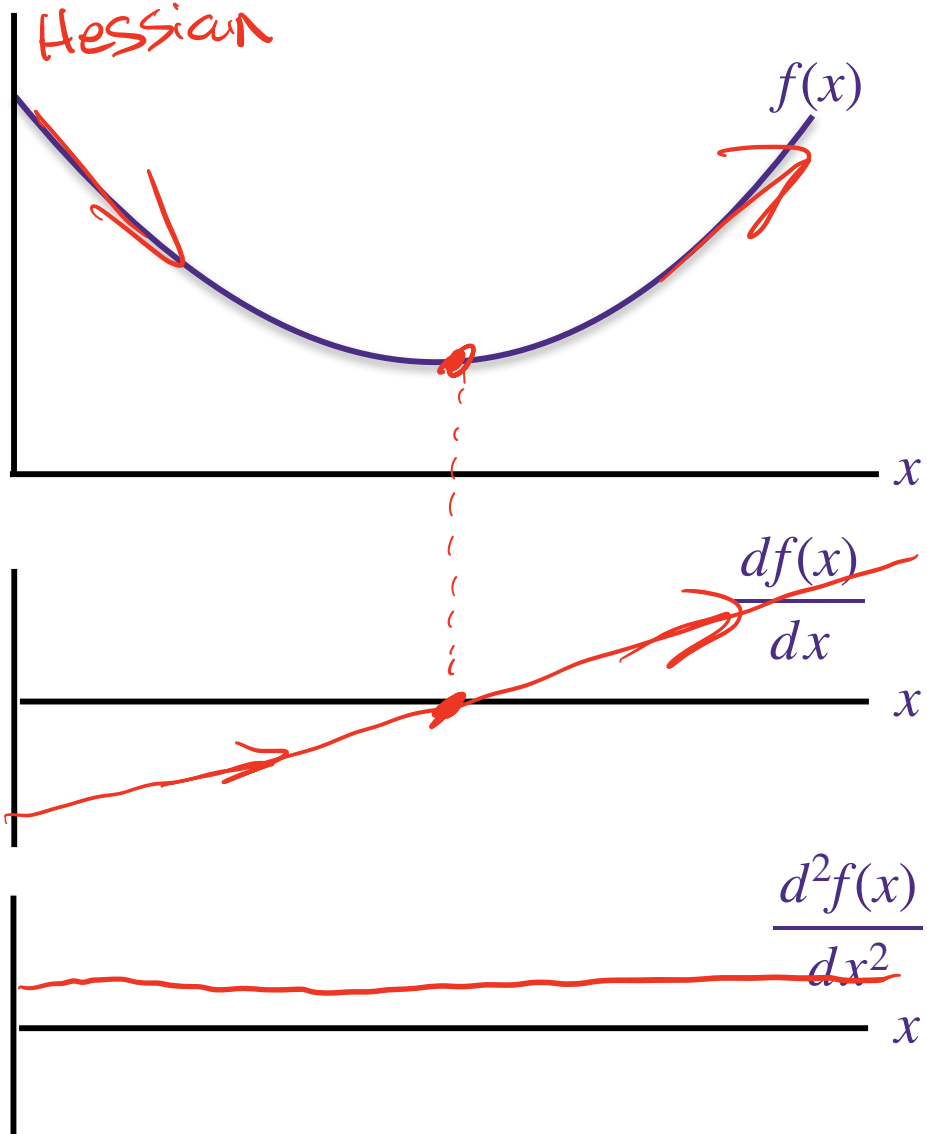
A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

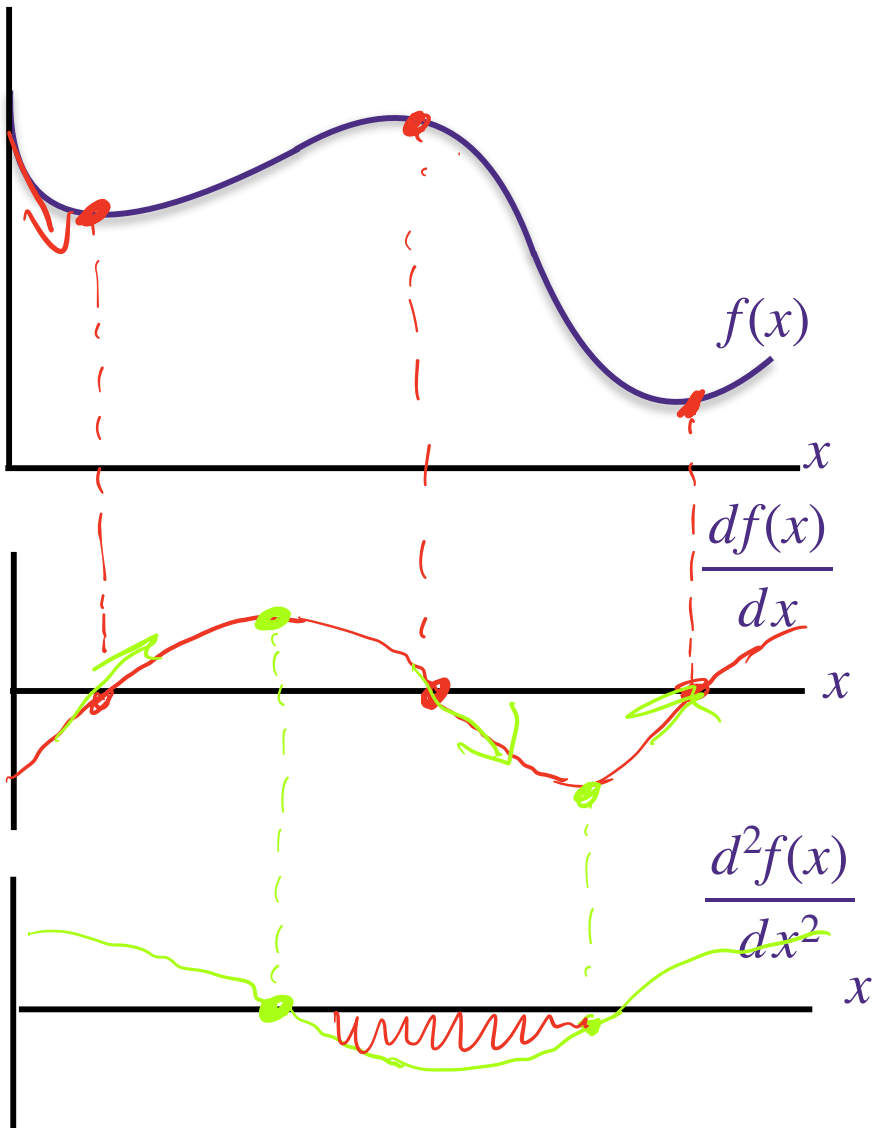




More definitions of convexity

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 $x \in dom(f)$





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 $x \in dom(f)$

$$\underbrace{\left[\nabla_x f(x)\right]}_{\text{Vector}}_i = \frac{\partial f(x)}{\partial x_i} \qquad \underbrace{\left[\nabla_x^2 f(x)\right]}_{\text{matrix } d \times d}_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

$$\nabla \left[\nabla f(x)\right]_i = \left[i^{th} \text{ row Hessian}\right] = \frac{\partial}{\partial x_j}\left[\nabla f(x)\right]_i$$

Matrix $A \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if $\underbrace{z^T A z}_{\text{scalar}} \geq 0 \quad \forall z \in \mathbb{R}^d$.

Linear Regression: $f(w) = \|y - Xw\|_2^2$

$\nabla_w f(w) = 2X^T X w - 2X^T y$

$\nabla_w^2 f(w) = 2X^T X$

$z^T(X^T X)z = \|Xz\|_2^2 \geq 0 \Rightarrow f(w)$ is convex.

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

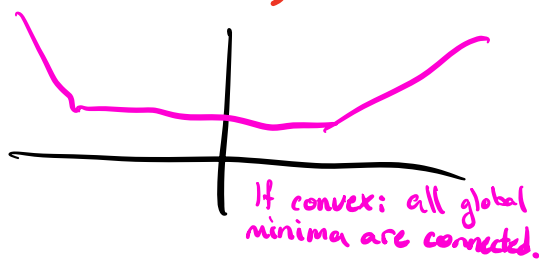
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

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 $x, y \in dom(f)$

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?

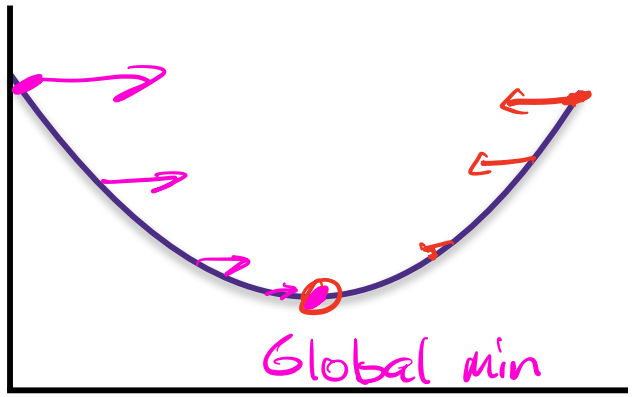
$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0)$



Convex functions

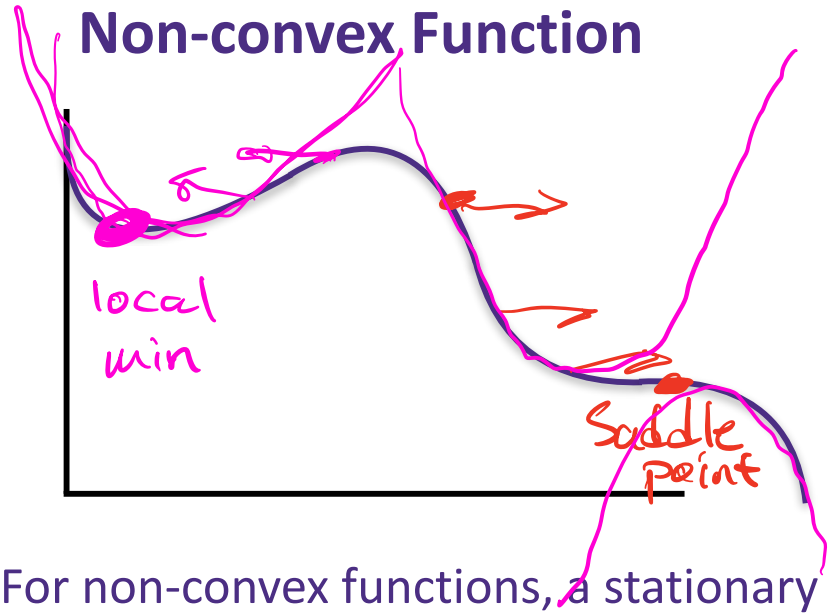
- All local minima are global minima
- Efficient to optimize (e.g., gradient descent)

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

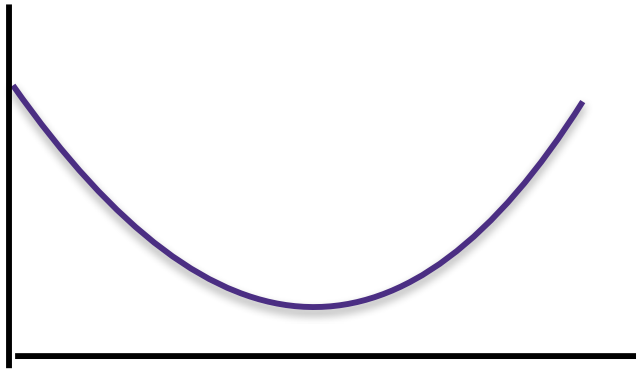
Gradient Descent on $\min_w f(w)$

Initialize: $w_0 = 0$

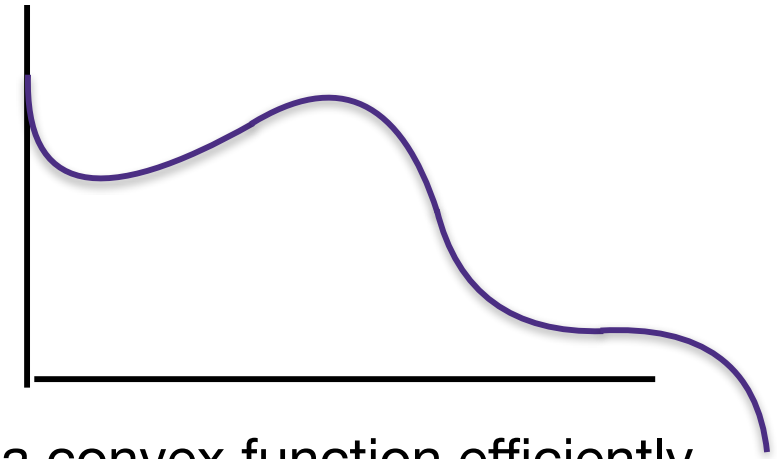
for $t = 1, 2, \ldots$

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

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 are differentiable everywhere,
  - otherwise $\nabla f(x)$ is not defined and gradient descent cannot be applied

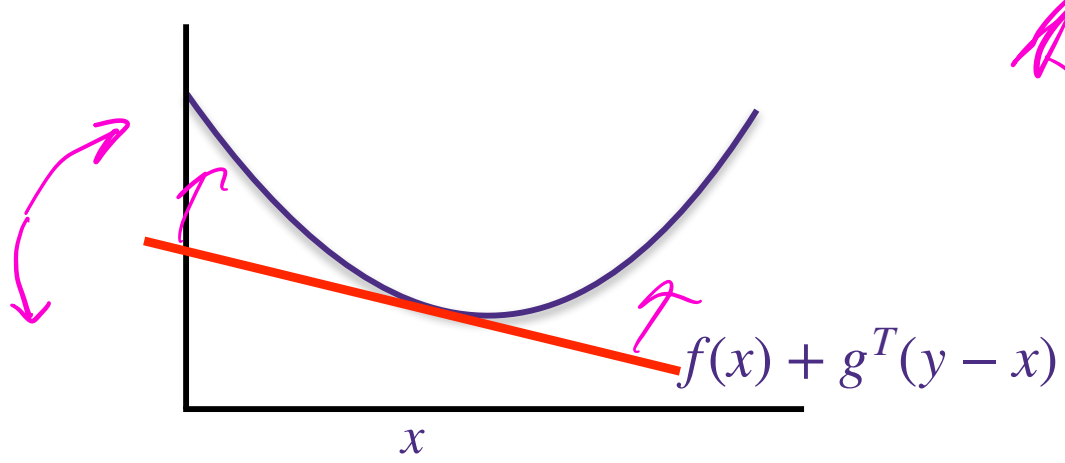
Sub-Gradient

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

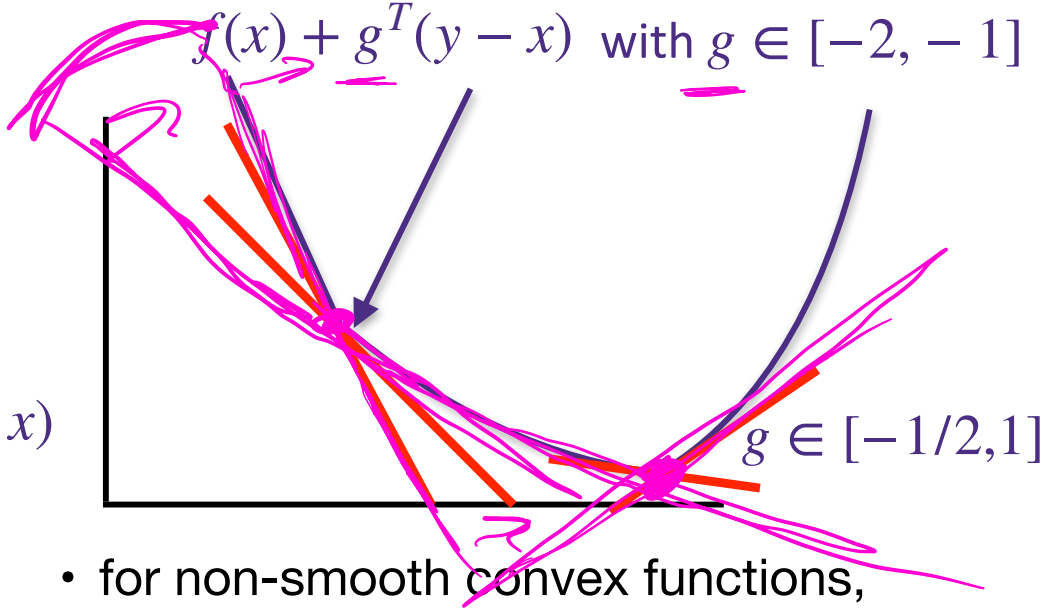
$$f(y) \geq f(x) + g^T(y - x) \text{ for all } y \in \mathbb{R}^d$$

Smooth Convex Function



- for smooth convex functions,
  - gradient is the unique sub-gradient, and
  - the global minimum is achieved at points where gradient is zero

Non-smooth Convex Function



- 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

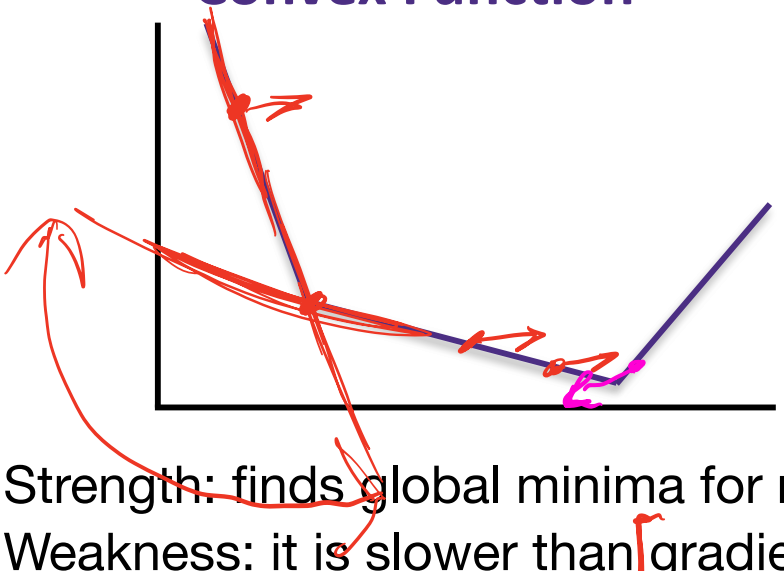
Initialize: $w_0 = 0$

for $t = 1, 2, \ldots$

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 does 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?