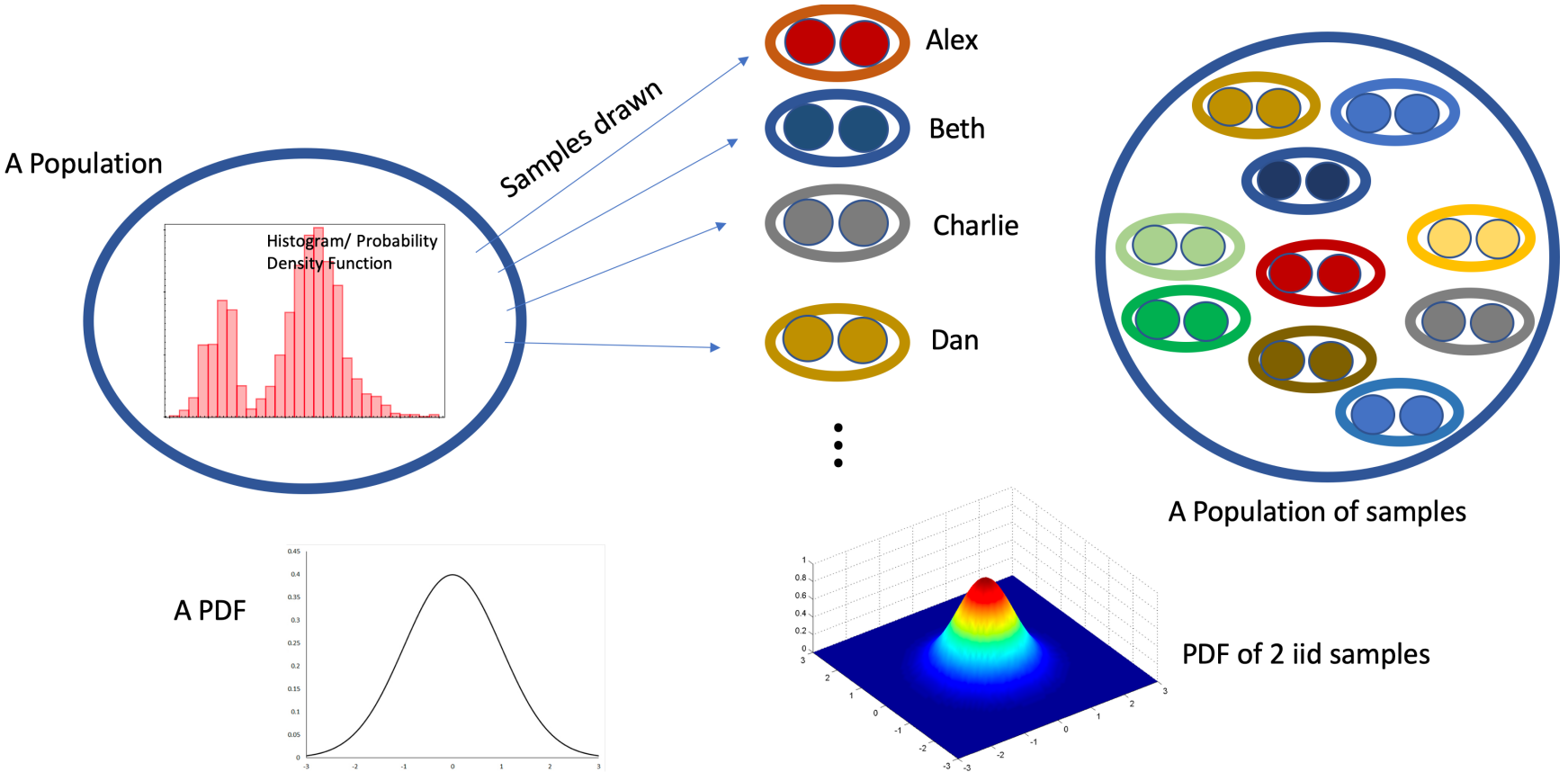
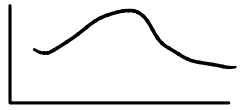
# Sampling Distribution

# Sampling Distribution




- We often have a reference population (with some PDF or histogram) and would like to draw multiple samples to discover something about the population

$f_X(x)$ :



$(x_1^{(1)}, x_2^{(1)})$

$(x_1^{(2)}, x_2^{(2)})$

$(x_1^{(3)}, x_2^{(3)})$

$$f_{X_1,X_2}(x_1,x_2) = f_X(x_1)\, f_X(x_2)$$

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \prod_{i=1}^{3} f_X(x_i)$$

# Sampling Distribution

- A useful fact to derive the distribution for multiple independent samples is: *if* $x_1 \sim f_1(x)$, $x_2 \sim f_2(x)$, ..., $x_N \sim f_N(x)$ *are independent random variables, their joint PDF is*

$$f(x_1, x_2, \ldots, x_N) = f_1(x_1) f_2(x_2) \ldots f_N(x_N) = \prod_{i=1}^{N} f_i(x_i)$$

- **Example:** We take two independent samples $x_1$ and $x_2$ from a standard normal distribution. What is the joint PDF of $x_1$ and $x_2$?

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \end{aligned}$$

Given $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

$$\rightarrow f_{X_1,X_2}(x_1, x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_1^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_2^2}{2}\right)$$

$$= \frac{1}{2\pi} \exp\left(-\frac{x_1+x_2}{2}\right)$$

$$f(x_1, \cdots, x_{10}) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_i-N)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{10} \exp\left(-\frac{\Sigma(x_i-N)^2}{2\sigma^2}\right)$$

# In-Class Exercise

- We take 10 samples from a normal distribution with mean $\mu$ and variance $\sigma^2$. What is the sampling distribution?

$$f(x_1, x_2, \ldots, x_N) = \ldots$$

# Functions of Samples

- Not only we can talk about the joint distribution of the samples, we can also talk about the distribution of a function applied to the samples.
- Finding the distribution for functions of random variables is not a straightforward task.
- **Example**: If $x_1$ and $x_2$ have the joint PDF $f_{X,Y}(x, y)$, what is the pdf for $z = 3x + y$?

  **Answer**. $F_Z(z) = \mathbb{P}(3X + Y \leq z) = \int\int_{3x+y \leq z} f_{X,Y}(x, y) dx\ dy$ and then taking a derivative of the CDF $F_Z(z)$ to acquire the pdf $f_Z(z)$.
- But sometimes there are shortcuts. For example when we know what is the distribution for the sum of two random variables, and all we need to do is estimating the distribution parameters.

If X, Y are normal → Z is normal

# In-Class Exercise

- **Example:** We have a normal distribution with mean 2 and variance 9. What is the distribution of of the sample mean acquired by averaging 4 independently drawn samples.
  **Hard Way.** We obtain the sampling distribution $f(x_1, x_2, x_3, x_4)$, which is the joint distribution, and then use the technique in the previous example to acquire the distribution of $z = (x_1 + x_2 + x_3 + x_4)/4$.
  **Easy Way.** Using the fact sheet from lecture 2, we know weighted sum of normal random variables is normally distributed, so all we need is to find the mean and the variance of $z = (x_1 + x_2 + x_3 + x_4)/4$.

$$\mathbb{E}(z) = 4 \times (2/4) = 2, \quad \text{var}(z) = 4 \times \frac{9}{16} = \frac{9}{4}.$$

  Therefore $z \sim \mathcal{N}(2, \frac{9}{4})$.
- See `https://en.wikipedia.org/wiki/Relationships_among_probability_distributions` for other shortcuts for the sum of independent random variables.

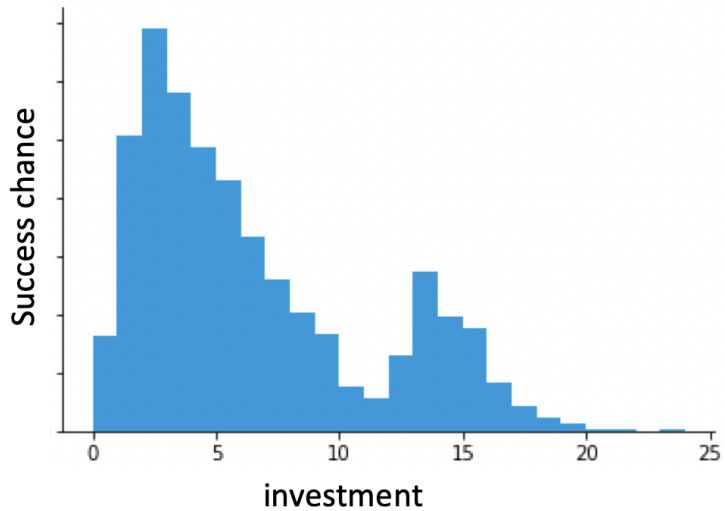
$$X_i \sim N(2, 9)$$

$$Z = \frac{X_1 + X_2 + X_3 + X_4}{4} \longrightarrow Z \sim N\left(\frac{2 \times 4}{4}, \frac{9 \times 4}{16}\right)$$

$$\Rightarrow Z \sim N\left(2, \frac{9}{4}\right)$$

# Brief Overview of Maximum Likelihood

– Say you are taking a risky investment and you know the following pdf corresponds to the succcess chance in terms of the investment. How much would you invest?



$\tilde{\chi}_1 = 1.213$

$\tilde{\chi}_2 = -10.7$

$\tilde{\chi}_3 = 3.41$

$\vdots$

, so observation $= \begin{bmatrix} 1.213 \\ -10.7 \\ 3.41 \\ \vdots \end{bmatrix}$

# Brief Overview of Maximum Likelihood

- Maximum likelihood (ML) is a statistical estimation technique
- The main goal in ML is often estimating the parameters of the **reference** population from a set of samples
- Let $x_1, x_2, \cdots x_n$ be samples from a distribution with some unknown parameter $\theta$ and joint distribution

$$f(x_1, x_2, \cdots, x_n|\theta)$$

- The maximum likelihood estimate of $\theta$ based on the observations $\tilde{x}_1, \tilde{x}_2, \cdots \tilde{x}_n$ is

$$\theta_{ML} = \underset{\theta}{\text{argmax}} \; f(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n|\theta)$$

- When $x_1, x_2, \cdots x_n$ are i.i.d samples from a distribution $f(\cdot)$, then

$$f(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

$$= \prod_{i=1}^{n} f(x_i|\theta)$$

# Brief Overview of Maximum Likelihood

$E[X] = \mu$, $\sigma_X^2 = 1$

**Example 1.** We have a normal distribution $\mathcal{N}(\mu, 1)$ and we do not know $\mu$. We take 5 independent samples from this distribution and the values turn out to be

$$\tilde{x}_1 = 2.5377,\ \tilde{x}_2 = 3.8339,\ \tilde{x}_3 = -0.2588,\ \tilde{x}_4 = 2.8622,\ \tilde{x}_5 = 2.3188,$$

what is the ML estimate of $\mu$.

**Solution.** If we take 5 independent samples $x_1, x_2, \cdots x_5$ from a normal distribution $\mathcal{N}(\mu, 1)$, their joint distribution is

$$f(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4, \tilde{x}_5|\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - \mu)^2}{2}),$$

some basic calculus yields $\mu_{ML} = \frac{\tilde{x}_1 + \tilde{x}_2 + \cdots + \tilde{x}_5}{5} = 2.2587$ (why?)

$$N(\mu, 1) \rightarrow f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2}\right)$$

1st step: $f(x_1, \cdots, x_5) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i-\mu)^2}{2}\right)$

$$f(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_5 | \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^5 \exp\left(-\frac{\sum(x_i-\mu)^2}{2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^5 \exp\left(-\frac{(2.5377-\mu)^2}{2} - \frac{(3.8339-\mu)^2}{2} - \cdots\right)$$

$\mu$ is the only unknown,

$$l(\mu) = \log f(\tilde{x}_1, \cdots, \tilde{x}_5 | \mu) = 5 \cdot \log \frac{1}{\sqrt{2\pi}} - \frac{\sum(x_i-\mu)^2}{2}$$

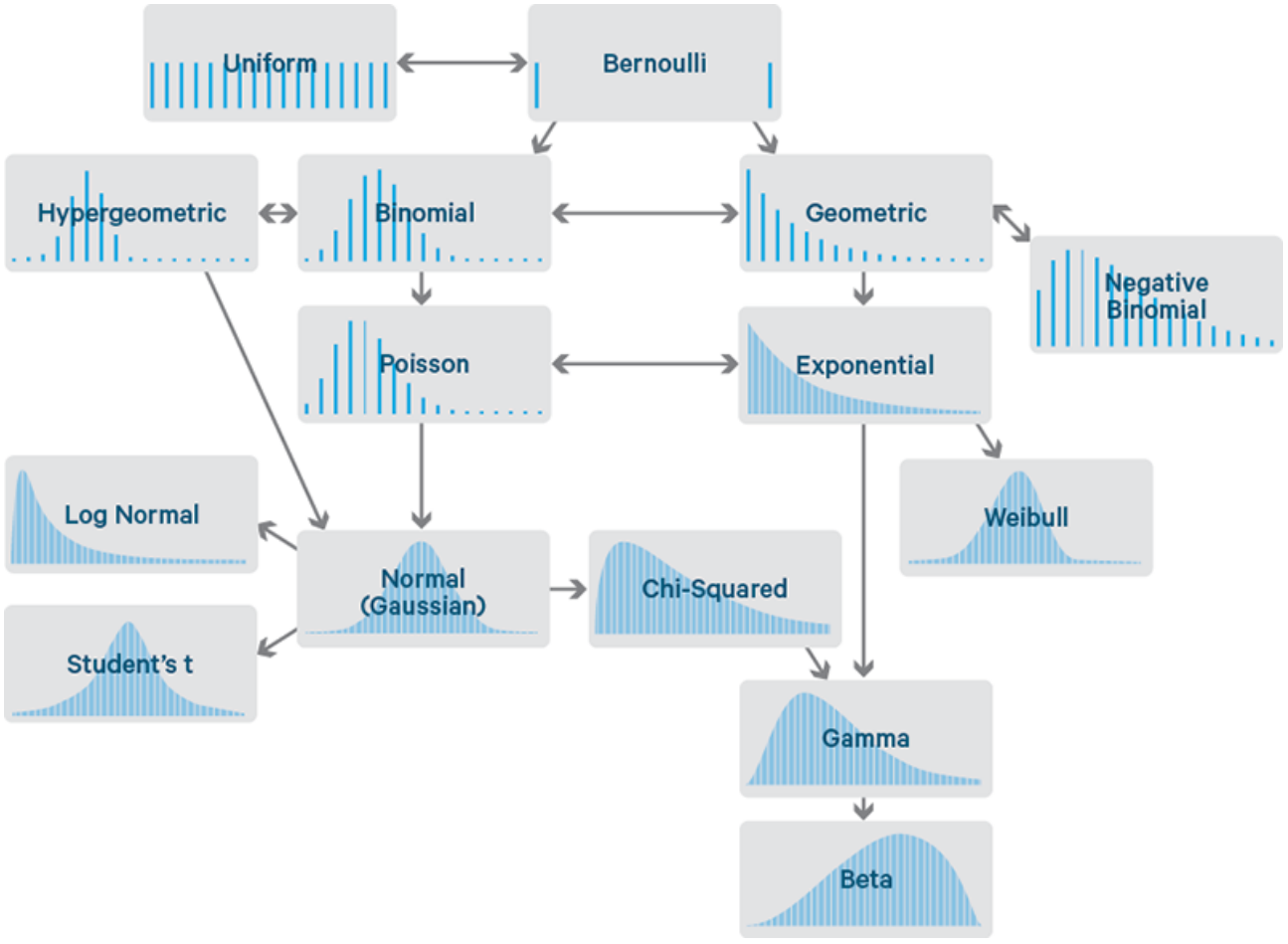
To find max of $\mu$, take partial derivative

$$\frac{\partial l(\mu)}{\partial \mu} = 0 \rightarrow \sum_{i=1}^{5}(\mu - x_i) = 0$$

$$\Rightarrow 5\mu = \sum_{i=1}^{5} x_i \Rightarrow \mu = \frac{\sum_{i=1}^{5} x_i}{5}$$

# Brief Overview of Maximum Likelihood

**Example 2.** We have a random generator which works as a black-box. Use the table below and a maximum likelihood approach to estimate the distribution and its parameters (see the MATLAB code).




[Figure Link]

Example 2:

$$X \sim \exp(\lambda), \quad \text{PDF: } f_X(x) = \lambda \cdot e^{-\lambda x} \quad (\lambda \geq 0,\ x \geq 0)$$

$$f(x_1, \cdots x_{1000} | \lambda) = \lambda^{1000} \cdot \exp\left(-\lambda \sum_{i=1}^{1000} x_i\right)$$

$$\mathcal{L}(\lambda) = \log f(\tilde{x}_1, \cdots \tilde{x}_{1000} | \lambda) = 1000 \log \lambda - \lambda \sum_{i=1}^{1000} \tilde{x}_i$$

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{1000}{\lambda} - \sum_{i=1}^{1000} \tilde{x}_i = 0 \Rightarrow \lambda = \frac{1000}{\sum_{i=1}^{1000} \tilde{x}_i}$$

# Brief Overview of Maximum Likelihood

More related to Linear Regression

**Example 3.** We have a simple linear model in the form of $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. We pass the points $x_1, \ldots, x_n$ to the model and receive the independent random responses $y_1, \ldots, y_n$. Based on the observed samples, what is the ML estimate for $\beta_0$ and $\beta_1$?

Hint:

$$\begin{aligned} &\arg\max_{\beta_0, \beta_1} \quad f(y_1, \cdots y_n | X_1 = x_1, \cdots, X_n = x_n, \beta_0, \beta_1) \\ &= \arg\max_{\beta_0, \beta_1} \quad \log\left(f(y_1, \cdots y_n | X_1 = x_1, \cdots, X_n = x_n, \beta_0, \beta_1)\right) \end{aligned}$$

**Important Note.** Similar to the way we treated the RSS minimization, in ML we also need to set the derivative with respect to all the variables to zero. You will do this in the homework.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \sim N(0,1)$$

random

$$\rightarrow \quad y_i \sim N(\beta_0 + \beta_1 x_i, 1)$$

$$f(y_1, y_2 \cdots y_n \mid \beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\sum (y_i - \beta_0 - \beta_1 x_i)^2}{2}\right)$$

$$L(\beta_0, \beta_1) = \log f(y_1, \cdots y_n \mid \beta_0, \beta_1)$$

$$= n \log \frac{1}{\sqrt{2\pi}} - \frac{\sum (\hat{y}_i - \beta_0 - \beta_1 x_i)^2}{2}$$

constant

instead of take derivative of whole equation

minimize $\frac{1}{2} \sum (\hat{y}_i - \beta_0 - \beta_1 x_i)^2$

**RSS**

# Brief Intro to Optimization

# Setting the Derivative to Zero?

- We have been setting the derivative both for minimization and maximization. How do we know if the resulting point is a minimizer or a maximizer?
- What happens if we cannot solve the equation resulted by setting the derivative to zero?
- Normally distinguishing a minimizer from a maximizer requires the second derivative information
- In this lecture we will introduce methods that are for general minimization
- Also would cover a class of functions that a stationary point corresponds to a minimizer

new methods !

# Gradient Descent and I ts Variants

# In-Class Exercise

- Find the minimizer of the function

$$\mathcal{C}(p_1, p_2) = (p_1 - p_2 - 3)^2 + p_2^2$$

# Can We Always Find the Zero Easily?

- How about the minimizer of the function

$$\mathcal{C}(p_1, p_2) = (1 - p_1)^2 + (1 - p_2)^2 - 2\exp(-3p_1^2 - 3p_2^2)$$

Let's take a look at the function plot in Matlab.

# Gradient Descent for Minimization

- We saw that our fitting and ML problems ultimately can be reduced to a minimization

$$\min_{\boldsymbol{p}} \ \mathcal{C}(\boldsymbol{p})$$

  where $\boldsymbol{p}$ includes all the unknown variables.
- Assuming $\boldsymbol{p} \in \mathbb{R}^L$, a numerical way of minimization is to start from a point $\boldsymbol{p}^{(0)}$ and iteratively perform the following steps

$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^{(k)} - \eta \nabla \mathcal{C}\Big|_{\boldsymbol{p}=\boldsymbol{p}^{(k)}} \quad \text{where} \quad \nabla \mathcal{C} = \begin{pmatrix} \partial \mathcal{C}/\partial p_1 \\ \partial \mathcal{C}/\partial p_2 \\ \vdots \\ \partial \mathcal{C}/\partial p_L \end{pmatrix}$$

  parameter $\eta$ is called the **learning rate**
- Larger learning rate does not necessarily mean faster solve
- Let's go through a simple example to see how gradient descent works (see the MATLAB code and the next slide)

$$p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{bmatrix} 1.2 \\ 3.4 \end{bmatrix} \qquad \nabla C(p) = \begin{bmatrix} 3.1 \\ 4.2 \end{bmatrix}$$

$\eta = 0.01$

$$p^{(2)} = \begin{bmatrix} 1.2 \\ 3.4 \end{bmatrix} - 0.01 \begin{bmatrix} 3.1 \\ 4.2 \end{bmatrix}$$

# Gradient Descent for Minimization

- Please refer to the MATLAB `gradientDescent.m` script
- Lets consider the very simple objective

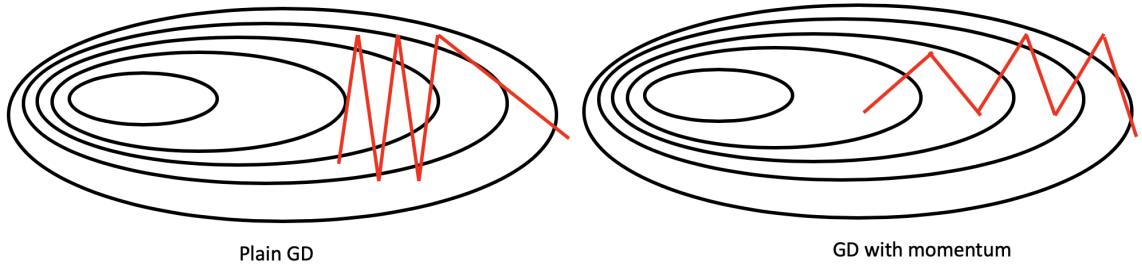
$$\mathcal{C}(p_1, p_2) = (1 - p_1)^2 + (1 - p_2)^2 - 2\exp(-3p_1^2 - 3p_2^2)$$

The gradient can be calculated as

$$\nabla\mathcal{C} = \begin{pmatrix} 2(p_1 - 1) + 12p_1\exp(-3p_1^2 - 3p_2^2) \\ 2(p_2 - 1) + 12p_2\exp(-3p_1^2 - 3p_2^2) \end{pmatrix}$$

- We can see that this objective has multiple local minimizers (two)
- Depending on where we start from we may land in either one
- A too small LR (learning rate) can make the minimization slow
- A too large LR can also make it slow or never converging!
- LR can affect which minimizer we converge to, but this is beyond our control

# Gradient Descent with Momentum Makes GD faster!




- Momentum is a method that can dampen the gradient descent oscillations and accelerate it
- It can even help skipping shallow minima and land into deeper minima
- Gradient descent with **learning rate** $\eta$ and **momentum** $\gamma$:

$$\boldsymbol{\theta}_{k+1} = \gamma \boldsymbol{\theta}_k + \eta \nabla \mathcal{C}(\boldsymbol{p}^k)$$
$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^k - \boldsymbol{\theta}_{k+1}$$

- Again refer to the MATLAB code

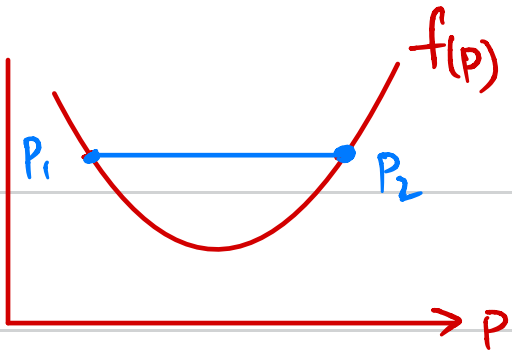
# Convex Functions

# What Are Convex Functions?

- Convex functions are a class of functions which have computationally attractive properties when it comes to minimization problems
- Suppose that $f(\boldsymbol{p}) : \mathbb{R}^n \to \mathbb{R}$, i.e., a function that operates on multiple variables (a vector) and produces a scalar as the output
- Function $f$ is convex if for all $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in the domain, and $0 \leq \theta \leq 1$:

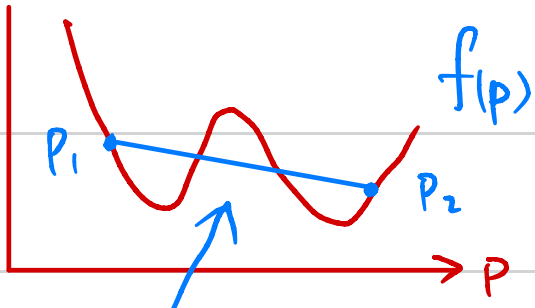
$$f(\theta \boldsymbol{p}_1 + (1-\theta)\boldsymbol{p}_2) \leq \theta f(\boldsymbol{p}_1) + (1-\theta) f(\boldsymbol{p}_2)$$

- Intuitively, a function is convex if the line segment between any two points on the graph of the function lies above (or just touching) the graph between the two points.




If $f(p)$ between $P_1 P_2$ is below line $P_1P_2$ or equal

→ Convex function



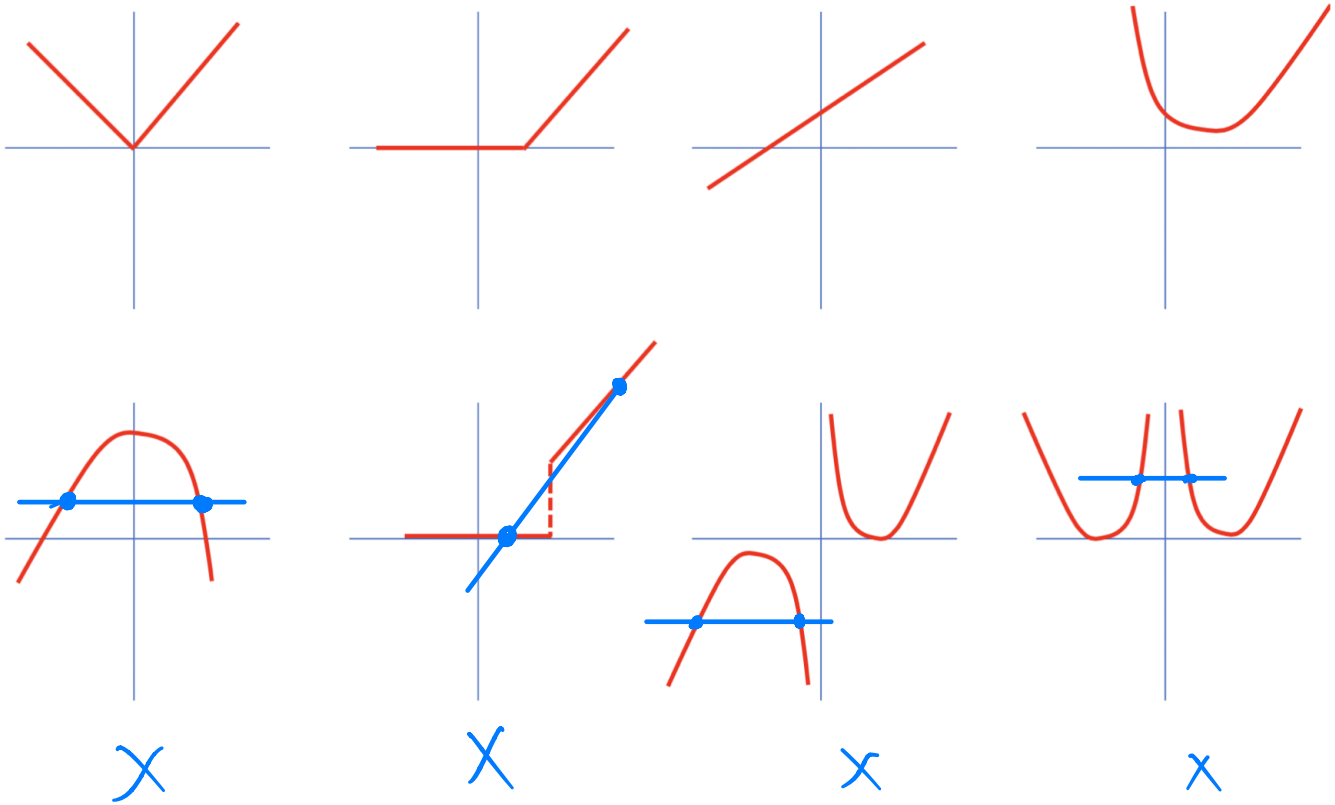

beyound the line $P_1P_2$

→ Not Convex Function

# In-Class Exercise

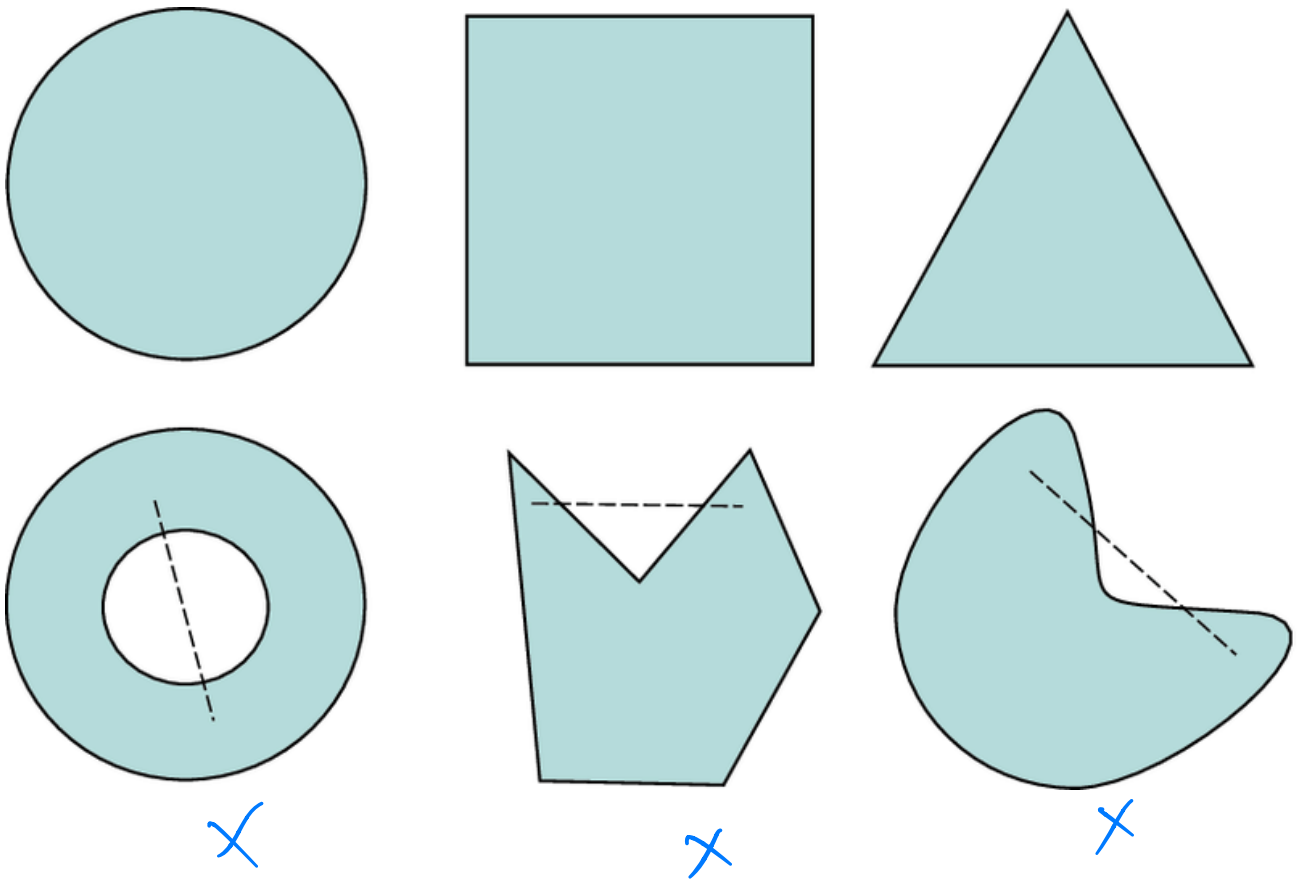
- Identify convex functions:

Any linear function is convex function



# Convex Sets

- A set is convex if the line connecting any two points of the set entirely stays within the set



is Convex

Convex Function is not necessarily Continuous Function

If constructing a Convex Function,

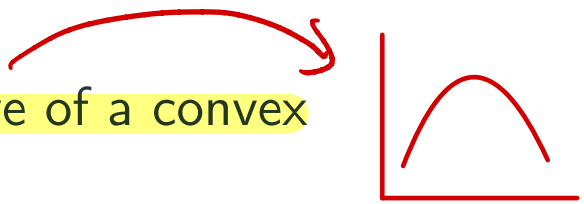
It should be Continuous

# More on Convexity

- In optimization theory, convex programs which are in the form

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad f(\boldsymbol{p}) \qquad \text{subject to:} \qquad \boldsymbol{p} \in \text{a convex set}$$

  have very desirable computational properties
- For differentiable convex functions gradient descent always lands to a **global minimizer**
- A function that can be represented as the negative of a convex function is called **concave** function.



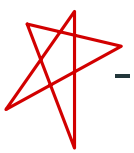
- Verifying the convexity in low dimensions, like 1 or 2, can be done visually. But for high dimensions we need to use the properties and definitions to show the convexity

# More Properties and Examples

- One way to show the convexity is using the definition and showing that for all $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in the domain, and $0 \leq \theta \leq 1$:

$$f(\theta \boldsymbol{p}_1 + (1-\theta)\boldsymbol{p}_2) \leq \theta f(\boldsymbol{p}_1) + (1-\theta) f(\boldsymbol{p}_2).$$

- **Example.** Show that the function $f(x, y) = x + y$ is convex.



- Generally, all linear functions of the form

$$f(x_1, x_2, \ldots, x_n) = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$$

are convex.

# More Properties and Examples

- Another way to show the convexity is using the following theorem: *A differentiable function is convex if for all $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in the domain:*

$$f(\boldsymbol{p}_2) - f(\boldsymbol{p}_1) - \nabla f(\boldsymbol{p}_1)^\top (\boldsymbol{p}_2 - \boldsymbol{p}_1) \geq 0.$$

(we got rid of $\theta$)!

- **Example.** Show that the function $f(x, y) = (x + y)^2$ is convex.

- If $g(z)$ is convex, then $g$ applied to a linear function is also convex, i.e., $g(\alpha_1 x_1 + \ldots + \alpha_n x_n)$ is convex.
- In other words, to show the convexity of $f(x, y) = (x + y)^2$ we only need to show that $f(z) = z^2$ is convex.

$$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = (x+y)^2 = f(x,y)$$

$$\nabla f(x,y) = \begin{bmatrix} 2(x+y) \\ 2(x+y) \end{bmatrix}$$

$$P_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad i = 1, 2$$

$$\nabla f(P_1) = 2\begin{bmatrix} x_1+y_1 \\ x_1+y_1 \end{bmatrix}$$

We need to show that

$$P_2 - P_1 = \begin{bmatrix} x_2-x_1 \\ y_2-y_1 \end{bmatrix}$$

$$\nabla f(P_1)^T$$

$$(x_2+y_2)^2 - (x_1+y_1)^2 - 2(x_1+y_1)(x_2-x_1) - 2(x_1+y_1)(y_2-y_1) \geq 0$$

$$\Rightarrow (x_2+y_2)^2 - (x_1+y_1)^2 - 2(x_1+y_1)\left(x_2+y_2-x_1-y_1\right) \geq 0$$

$$\begin{cases} u = x_2+y_2 \\ v = x_1+y_1 \end{cases}$$

$$\Rightarrow u^2 - v^2 - 2v(u-v) \geq 0$$

$$\Rightarrow u^2 + v^2 - 2uv \geq 0 \Rightarrow (u-v)^2 \geq 0 \quad \#$$

Suppose $f(z)$ is convex is $z \in \mathbb{R}$

$\rightarrow f(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p)$ is convex in $x_1, x_2, \cdots, x_p$

$f(z) = z^2$ is convex $\rightarrow$ $f(x, y)$ is convex if $f(x, y)$ is linear.

$f(z)$ is convex is $z \in \mathbb{R}$ if $f''(z) \geq 0 \quad \forall z \in \mathbb{R}$

$\Downarrow$

ex2. $f(z) = z^2$ is convex, since

$f'(z) = 2z$

$f''(z) = 2 \geq 0 \Rightarrow f(z)$ is convex

# More Properties and Examples

- Yet, another way to show the convexity is using the following theorem: *A twice differentiable function is convex if at any point in the domain all the eigenvalues of the Hessian matrix are non-negative.* For $f(x_1, x_2, \ldots, x_n)$ the Hessian matrix is defined as

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

- In the single variable case, a twice differentiable function $f(x)$ is convex if $f''(x)$ is non-negative for all the points in the domain.
- **Example.** Use this result to show that the function $f(x, y) = (x + y)^2$ is convex.

Prove $f(x,y) = (x+y)^2$ is convex

$$\nabla f = \begin{bmatrix} 2(x+y) \\ 2(x+y) \end{bmatrix}$$

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial f}{\partial y \partial x} & \frac{\partial^2 f}{(\partial y)^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \geq 0 \quad \#$$