ML and Optimization Lecture 3

- Recap of essentials from probability theory:
  - joint distributions $\mathbb{P}(X, Y)$
  - marginals $\mathbb{P}(X)$, $\mathbb{P}(Y)$
  - conditionals $\mathbb{P}(Y|X)$
  - expectations $\mathbb{E}[Y]$
  - conditional expectations $\mathbb{E}[Y|X]$
  - variance $\mathrm{Var}(Y)$
  - independence $X \perp\!\!\!\perp Y$
  - conditional independence $X \perp\!\!\!\perp Y | Z$
- Parameterized machine-learning models

Last class we saw that we model our ML problems as learning

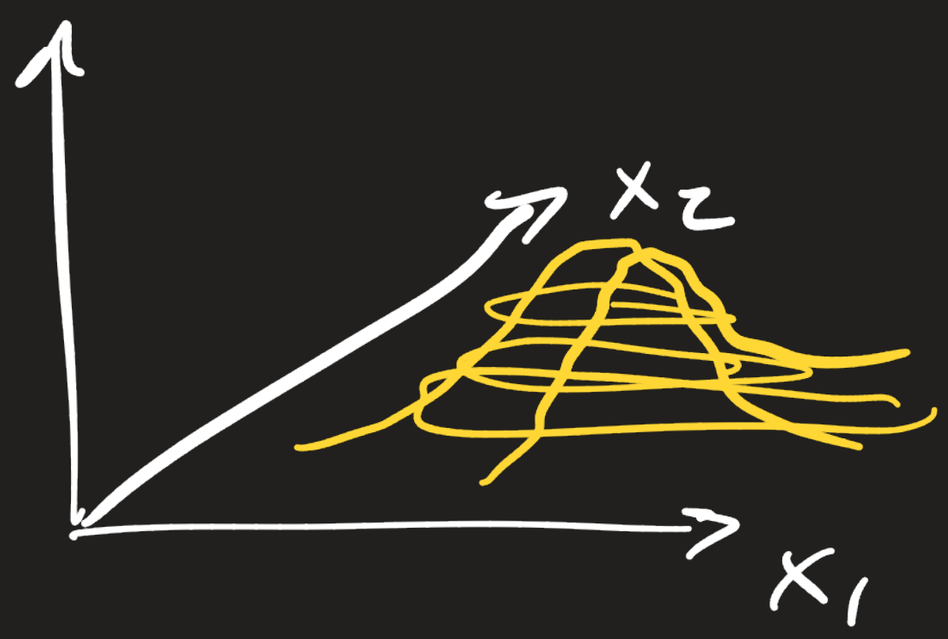
$$f : x \longmapsto y$$

x: covariates / independent features / predictors

y: target

by modeling $(x, y) \sim \mathbb{P}$ and minimizing the average loss of f w.r.t. $\mathbb{P}$

$$J(f) := \mathbb{E}_{(x,y)\sim\mathbb{P}} \ell(f(x), y)$$



Joint distributions

$X \sim \mathbb{P}$ means $\mathbb{P}(X \in A) \in [0,1]$ is the probability that X takes on values in the set A. We may also write this as $\mathbb{P}_X(A)$ or $\mathbb{P}(A)$

Two cases

- X is discrete

$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$ where p_X is the probability mass function. We have that $\sum_x p_X(x) = 1$ and $p_X(x) \geq 0$ everywhere.

- X is continuous

$\mathbb{P}(X \in A) = \int_A p_X(x)\,dx$ where p_X is the probability density function of X.

$p_X(x) \geq 0$ and $\int p_X(x)\,dx = 1$

Examples

$X \sim \text{Bern}(p)$ — Bernoulli distribution w/ parameter p
$p \in [0,1]$

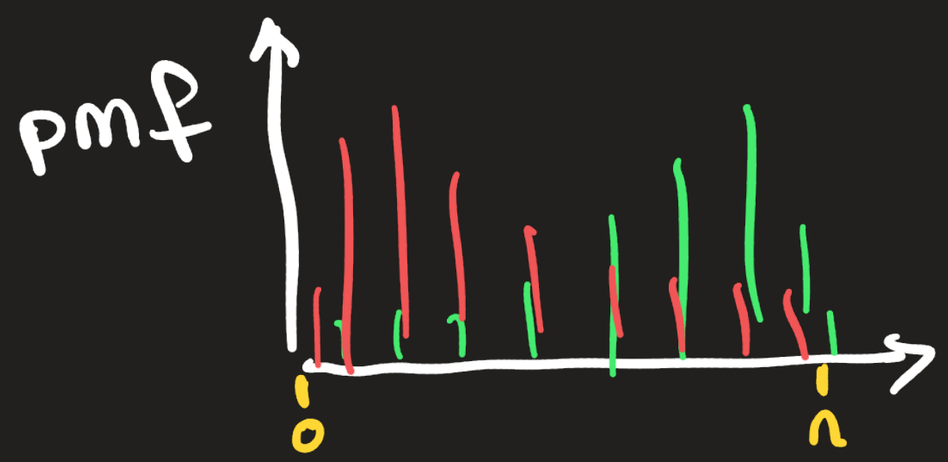
means

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

$X \sim \text{Binom}(n, p)$ — Binomial distribution w/ parameters
$n \in p$

means

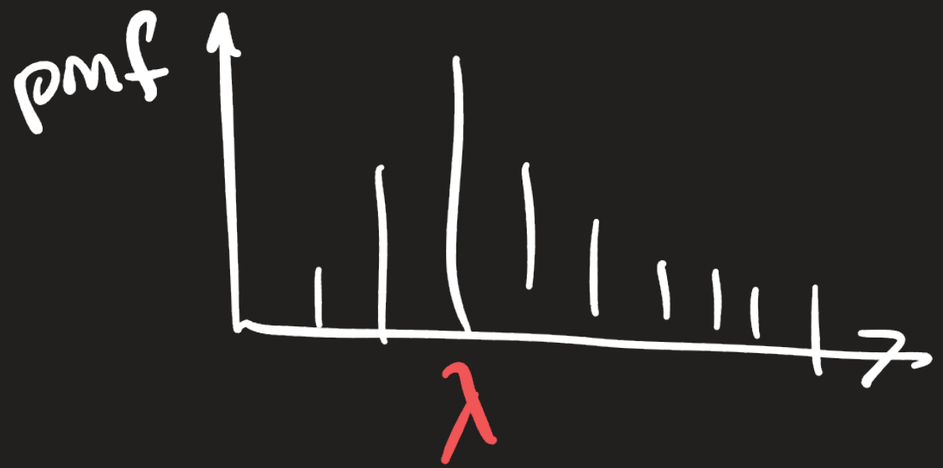
$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0,1,\dots,n\}$$



p close to 1
p close to 0

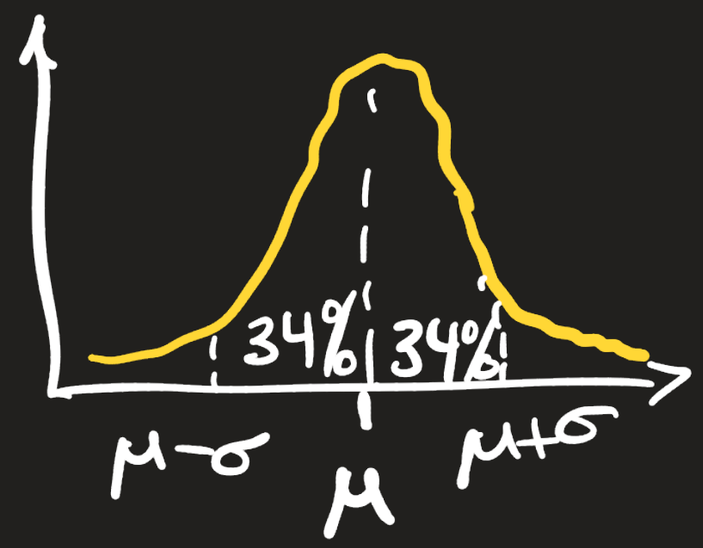
$X \sim \text{Poisson}(\lambda)$ — Poisson distribution w/ rate parameter λ

$$\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k=0,1,2,\dots$$



$X \sim \mathcal{N}(\mu, \sigma^2)$ — Normal or Gaussian distribution w/ mean μ and variance σ^2

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



To model dependencies b/w random variables, we use joint distributions

$\mathbb{P}(X \in A, Y \in B)$ written as $(X, Y) \sim \mathbb{P}$

Analogously we have joint pdfs and pmfs

$$\mathbb{P}(X \in A, Y \in B) = \int_{A \times B} p_{X,Y}(x, y)\, dx\, dy$$

where $p_{X,Y}(x, y) \geq 0$ and $\int_{\mathbb{R} \times \mathbb{R}} p_{X,Y}(x, y)\, dx\, dy = 1$

Marginals

Given $\mathbb{P}_{X,Y}$, we can extract the probability dist of X

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R}) = \int_{A \times \mathbb{R}} p_{XY}(x,y)\,dx\,dy$$

$$= \int_A \underbrace{\int_{\mathbb{R}} p_{XY}(x,y)\,dy}_{\text{"}p_X(x)\text{"}}\,dx$$

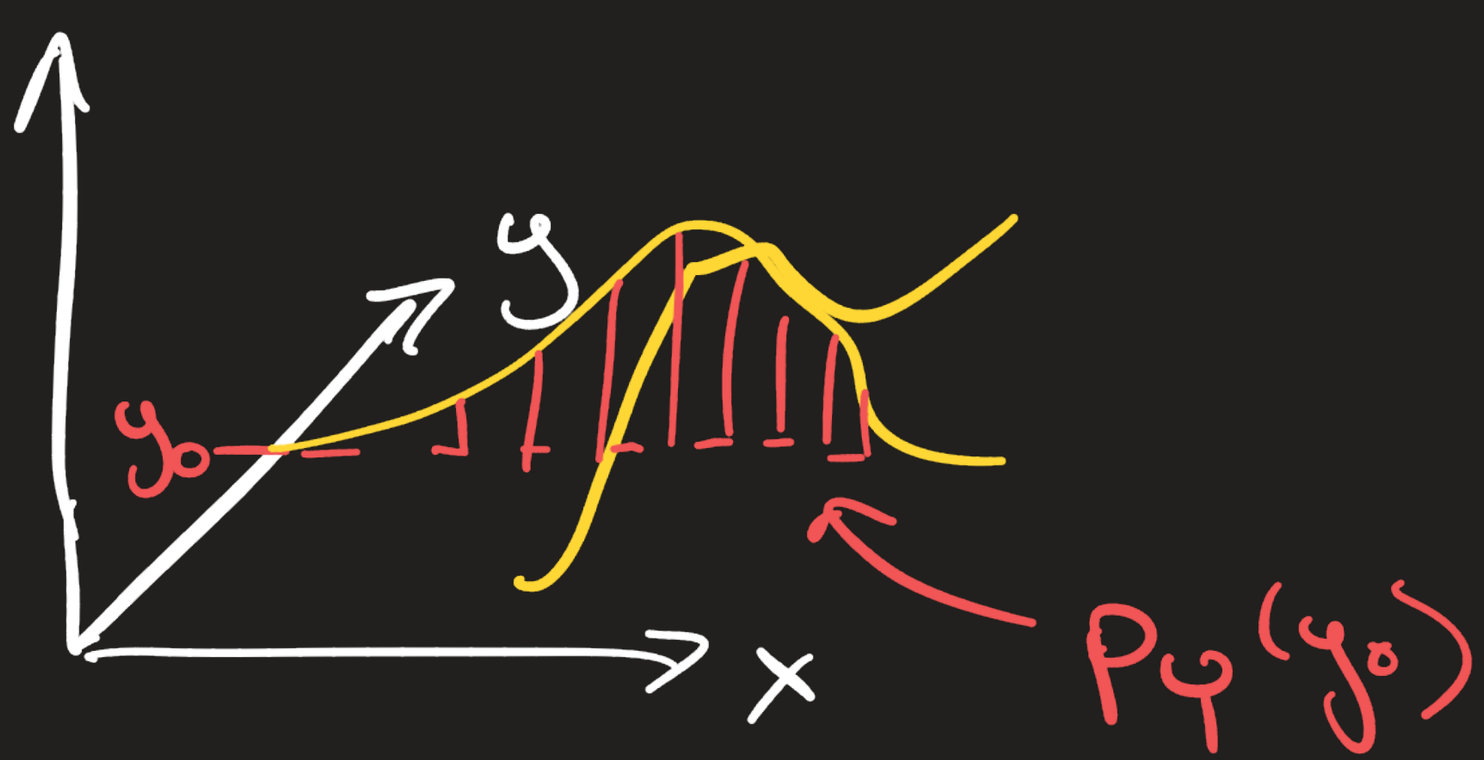
$$:= \int_A p_X(x)\,dx$$

where $p_X(x) = \int_{\mathbb{R}} p_{XY}(x,y)\,dy$ is the marginal pdf of x

note that $p_X(x) \geq 0$ and $\int_{\mathbb{R}} p_X(x)\,dx = 1$

Similarly we get p_y be marginalizing out x

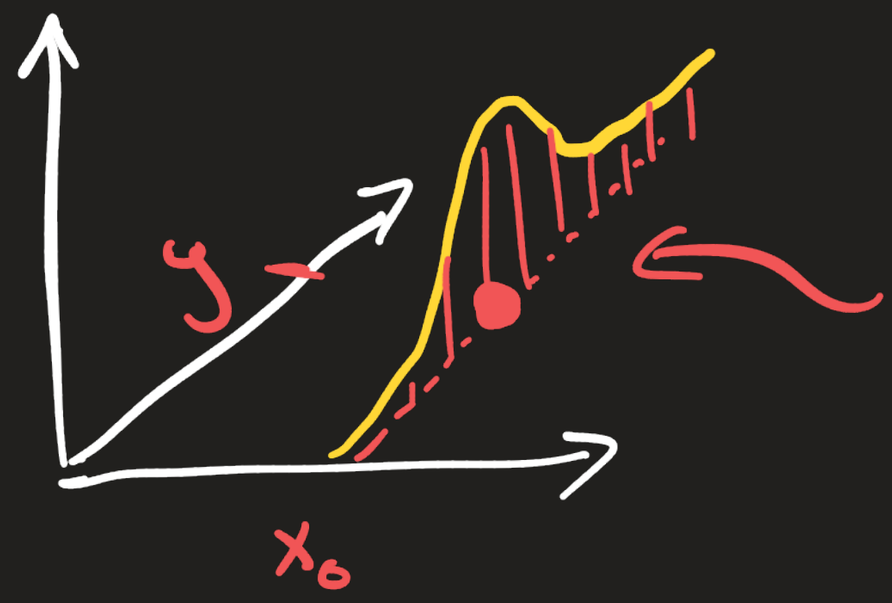
$$p_y(y) = \int_{\mathbb{R}} p_{x,y}(x, y)\, dx$$



Conditionals

If $(X, Y) \sim \mathbb{P}$, then knowledge of y conditioned on x is captured by the <u>conditional distribution</u> of Y given X, $\mathbb{P}(Y|X)$, whose <u>conditional pdf</u> is given by

$$p_{Y|X=x}(y) = p_Y(y|X=x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$



Expectation (Mean)

Average value of a r.v. X

$$\mathbb{E}X = \int_{\mathbb{R}} x\, p(x)\, dx$$

Ex

$X \sim \text{Bern}(p) \Rightarrow \mathbb{E}X = 1 \cdot p + 0 \cdot (1-p) = p$

$X \sim \text{Binom}(n,p) \Rightarrow \mathbb{E}X = n \cdot p$

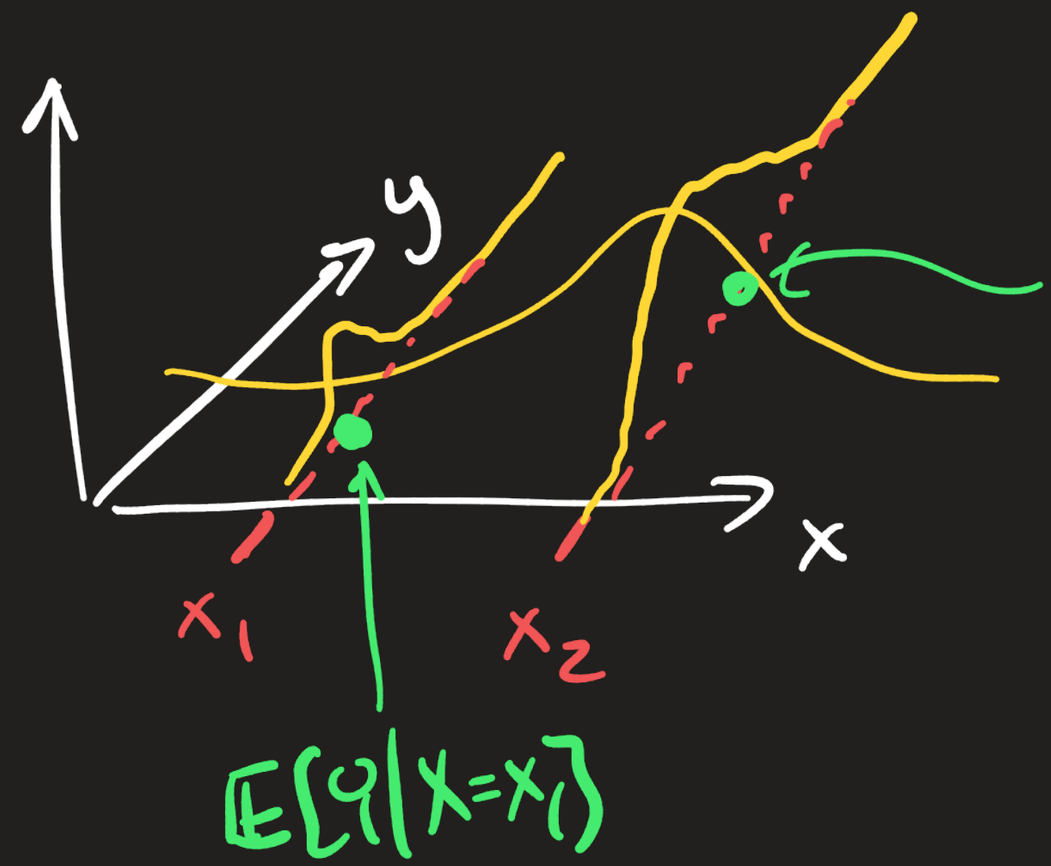
$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathbb{E}X = \mu$

Conditional Expectation

$\mathbb{E}[Y|X]$ – average value of Y given knowledge of X

$$\mathbb{E}[Y|X=x] = \int y\, p_Y(y|X=x)\, dy$$

↑ incorporates knowledge of how y is affected by x

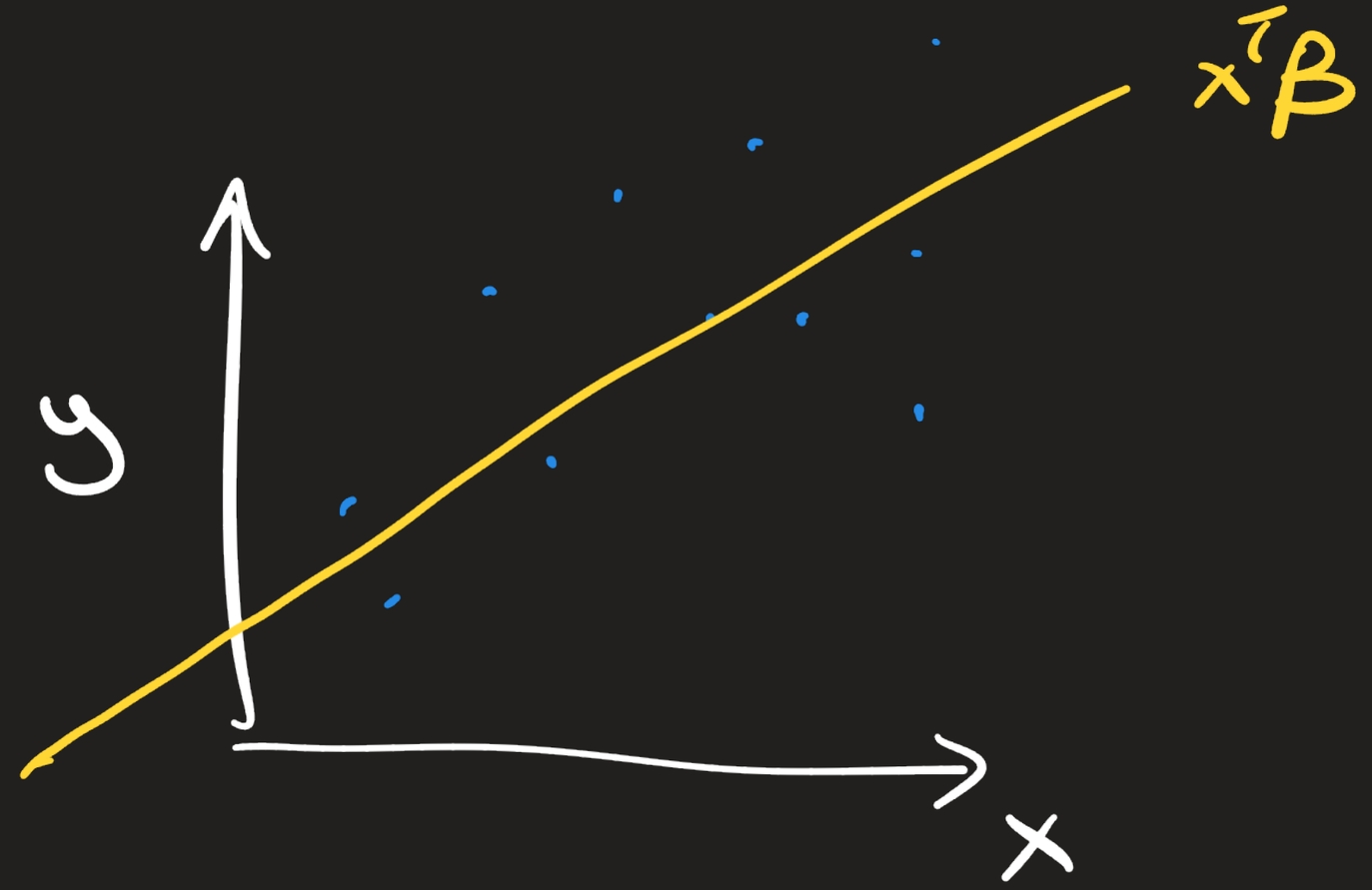


Aside Linear Regression

$$y = x^T\beta + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow y|x \sim \mathcal{N}(x^T\beta, \sigma^2)$$

$$\Rightarrow \mathbb{E}[y|x] = x^T\beta$$



Variance

Variance measures how close a random variable stays to its mean, on average

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

$$= \int (x - \mathbb{E}X)^2 \, p(x)\, dx$$

$$= \int \left[x^2 - 2x\,\mathbb{E}X + (\mathbb{E}X)^2 \right] p(x)\, dx$$

$$= \int x^2 p(x)\, dx - 2\mathbb{E}X \int x\, p(x)\, dx + (\mathbb{E}X)^2 \int p(x)\, dx$$

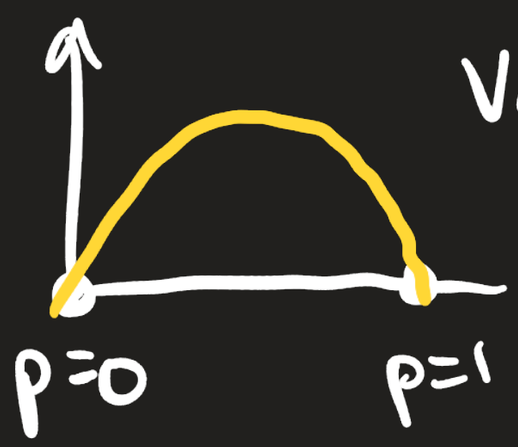
$$= \mathbb{E}[X^2] - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Ex of Variance

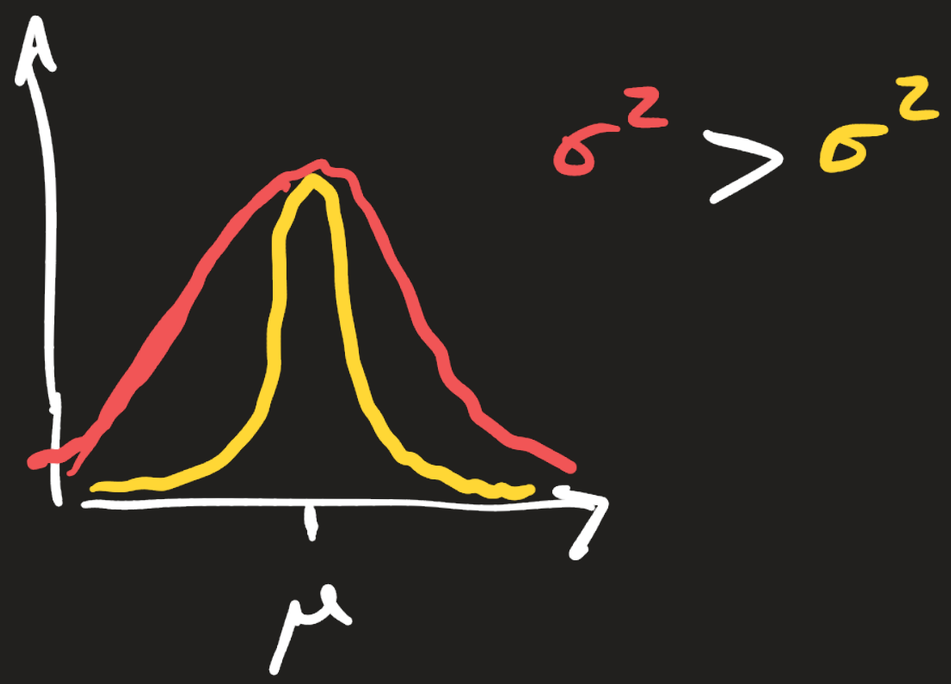
$X \sim \text{Bern}(p) \Rightarrow$ $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$

$= \mathbb{E}X - (\mathbb{E}X)^2$

$= p - p^2 = p(1-p)$



$X \sim \mathcal{N}(\mu, \sigma^2)$



Variance is a measure of uncertainty

Independence

We say two random variables are independent $(X \perp\!\!\!\perp Y)$ if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

$$\Updownarrow$$

$$p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$$

$$\Updownarrow$$

$$p_Y(y \mid X = x) = p_Y(y)$$

Ex

Y – # of divorces in Nebraska

X – dairy price in Brazil

$X \perp\!\!\!\perp Y$

Conditional dependence

We say $X \perp\!\!\!\perp Y \mid Z$, or X and Y are conditionally independent given Z if

$$P_{X,Y|Z}(x,y) = P_{X|Z}(x)\, P_{Y|Z}(y)$$

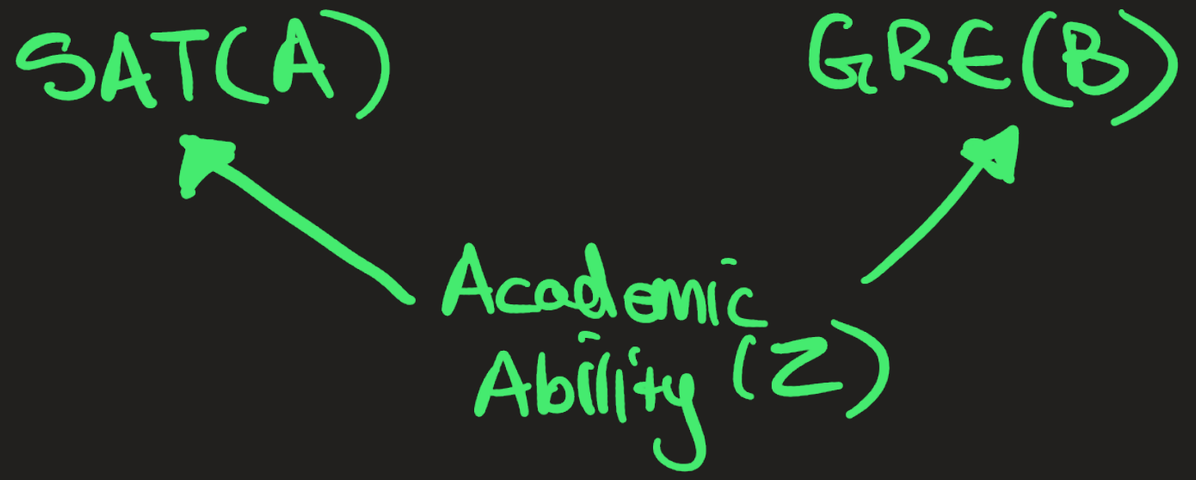
Note

$$P_{X,Y|Z}(x,y) = \frac{P_{XYZ}(x,y,z)}{P_Z(z)}$$

in general

Ex

$A \perp\!\!\!\perp B \mid Z$



Parameterized ML models

$P_Y(y|X=x)$ – captures everything about the possible values of Y given knowledge that $X=x$

$\mathbb{E}[Y|X=x]$ – captures a point estimate of Y given x

Issues with using either of these in ML: either can be an arbitrarily complicated unknown function of x

To deal with this, we use <u>parameterized models</u>, i.e. <u>assume</u>

$$P_Y(y|X=x) = f_\theta(y,x)$$

or

$$\mathbb{E}[Y|X=x) = f_\theta(x)$$

where $\theta \in \mathbb{R}^p$ is a parameter vector