ML & Opt Lecture 4

- Recall conditional distributions as capturing all uncertainties
- Point estimates
- Expectations & Variances revisited
- Conditional Expectations
- Regression Functions — best point estimate for least squares regression

Conditional Distributions

Given r.v.s Y and X,

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

captures the dependence of y on x

1) $p(y,z|x) = p(y|x)\,p(z|x)$

conditional independence $(y \perp\!\!\!\perp z \mid x)$

$$\iff \quad p(y|z,x) = p(y|x)$$

Show in participation 1

2) $p(x|z) = \int p(x,y|z)\,dy$

marginalizing joint conditional gives marginal conditional

Conditional distributions capture <u>everything</u> we know about y, given x.

But often we want an informative single number describing the behavior of y, conditioned on x

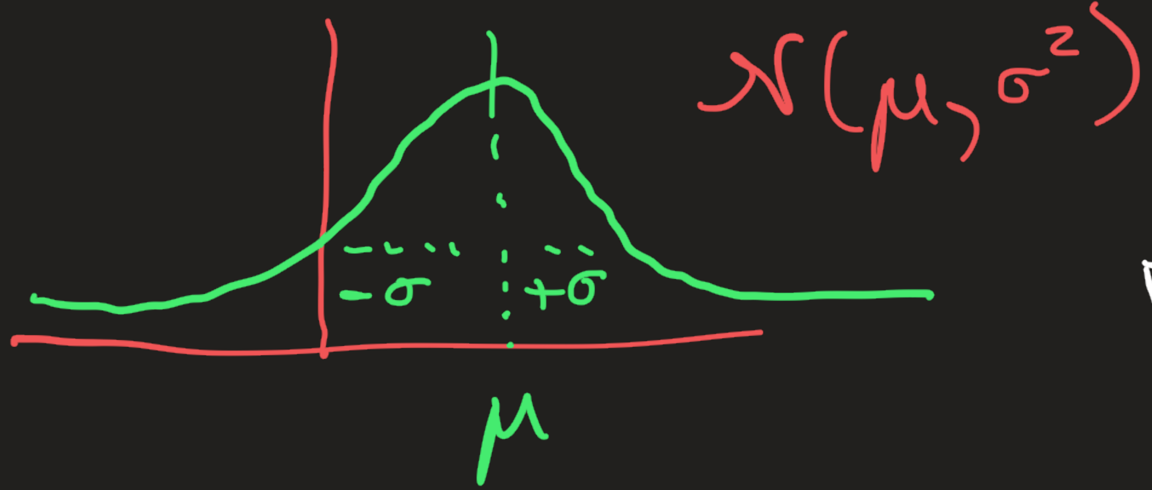
Call these point estimates.

Expected Values (point estimates of RVs)

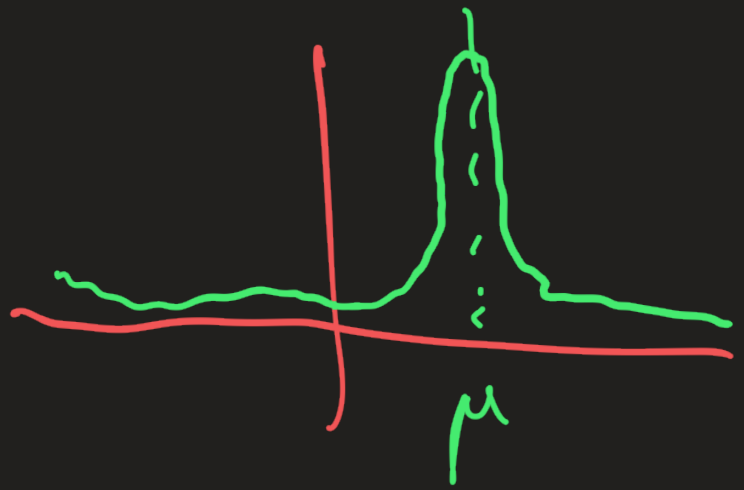
$$\mathbb{E}X = \sum_x x\, p(X=x)$$

weighted average of the possible values a r.v. takes on

$$= \int x\, p(x)\, dx$$



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

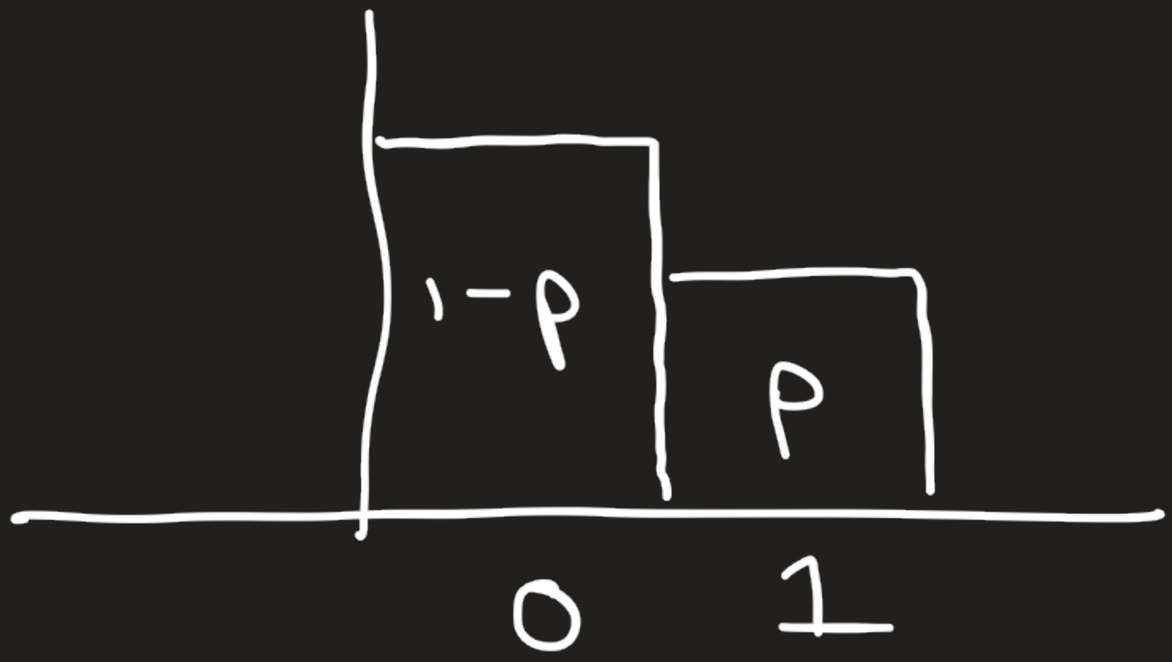


$$= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \mu$$

Bernoulli r.v.s

$$X = \begin{cases} 0 & 1-p \\ 1 & p \end{cases} \qquad \text{where } 0 \le p \le 1$$



pmf

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

Variance : how well does $\mathbb{E}X$ describe X, on average?

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

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X - \mathbb{E}X)^2 \\
&= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}X + (\mathbb{E}X)^2\right] \\
&= \mathbb{E}[X^2] - 2\,\mathbb{E}[X\,\mathbb{E}X] + (\mathbb{E}X)^2 \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 \\
&= \underbrace{\mathbb{E}[X^2]}_{\text{2nd moment}} - \underbrace{(\mathbb{E}X)^2}_{\text{square of the 1st moment}}
\end{aligned}
$$

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

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

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

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x^2 \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx - \mu^2$$

$$= \cdots = \sigma^2$$

$X \sim \text{Bern}(p)$

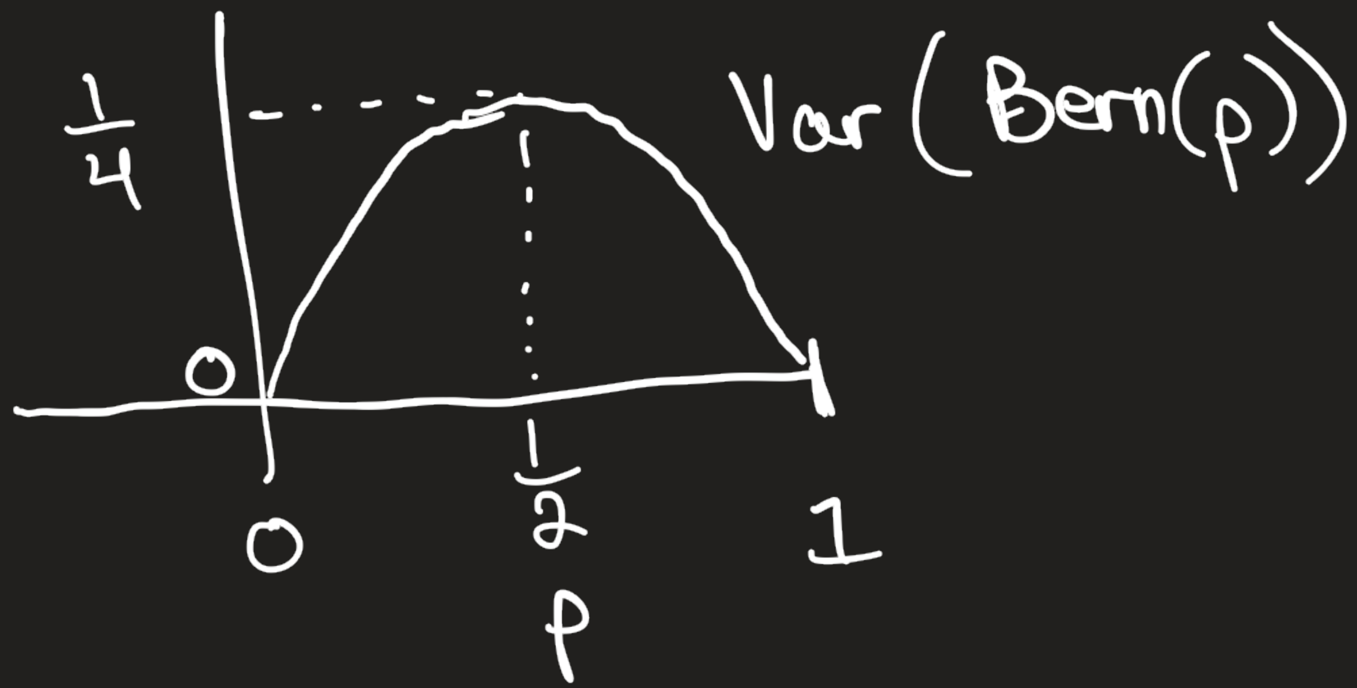
$\mathbb{E}X = p$

because $X = X^2$ since X is 0 or 1

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

$$= p(1-p)$$

so if $p \approx 1$ or $p \approx 0$ $\text{Var}(X) \approx 0$



Expectations & Variances & Independence

X & Y are independent then

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

$$\mathbb{E}XY = \int xy\, p(x,y)\, dx\, dy$$

$$\overset{\text{by } X \perp\!\!\!\perp Y}{=} \int xy\, p(x)\, p(y)\, dx\, dy$$

$$= \int x\, p(x)\, dx \int y\, p(y)\, dy$$

$$= (\mathbb{E}X)(\mathbb{E}Y)$$

If X & Y are independent

$$\mathrm{Var}(\alpha X + Y) = \alpha^2 \mathrm{Var}(X) + \mathrm{Var}(Y)$$

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

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

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

0 because $X \perp\!\!\!\perp Y$

$$\begin{aligned} = \;& \alpha^2 \mathbb{E}[X^2] + 2\alpha \mathbb{E}[XY] + \mathbb{E}[Y^2] \\ & - \alpha^2(\mathbb{E}X)^2 - 2\alpha \mathbb{E}X\,\mathbb{E}Y - (\mathbb{E}Y)^2 \end{aligned}$$

$$= \alpha^2 \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Q: if X & Y are not independent

what can we say about $\mathbb{E}[XY]$

and $Var(X+Y)$?

A: not anything useful, in general

e.g. $Y = -X$ then $Var(X+Y) = Var(0) = 0$

and if $Y = \frac{1}{X}$ then $\mathbb{E}[XY] = 1 \neq \mathbb{E}(X)\mathbb{E}\left(\frac{1}{X}\right)$

Consequence of Variance & Independence

Sometimes we want to estimate $\mathbb{E}X$ by sampling n independent copies of X, $X_1, \dots, X_n$ and claiming

$$\mathbb{E}X \approx \frac{1}{n}\sum_{i=1}^{n} X_i = S_n$$

Law of Large Numbers

Intuition for why this is true:

$$\mathbb{E}S_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n} n\,\mathbb{E}X = \mathbb{E}X$$

and

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\frac{1}{n}X_1 + \dots + \frac{1}{n}X_n\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} n\,\mathrm{Var}(X)$$

$$= \frac{1}{n}\mathrm{Var}(X)$$

$$\mathbb{E} S_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \int_{\mathbb{R}^n} \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) p(x_1, \dots, x_n)\, dx$$

$$= \int_{\mathbb{R}^n} \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) p(x_1) \cdots p(x_n)\, dx$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}^n} x_i\, p(x_1) \cdots p(x_n)\, dx$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} x_i\, p(x_i)\, dx_i = \frac{1}{n}\, n\, \mathbb{E} X = \mathbb{E} X$$

Recall Population Risk Minimization vs ERM

PRM: $$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \mathbb{E}\, \ell(f(x), y)$$

but in general we don't know $p(x, y)$

so sample (i.i.d.) copies of data (x_i, y_i)

and we approximate

$$\mathbb{E}\, \ell(f(x), y) \approx \frac{1}{n} \sum_{i=1}^{n} \underline{\ell(f(x_i), y_i)}$$

and take

ERM: $$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Conditional Expectation

point estimate for the r.v. $y \mid X = x$

$$\mathbb{E}[y \mid X = x] = \int y\, p(y \mid x)\, dy$$

Example

		Y = 0	1	2
X	0	2/6	1/12	1/12
	1	1/12	2/6	1/12

$p(y \mid x = 0)$

0	1	2
2/3	1/6	1/6

$p(y \mid x = 1)$

1/6	2/3	1/6

$$\mathbb{E}[Y \mid X = 0] = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} = \frac{1}{2}$$

$$\mathbb{E}[Y \mid X = 1] = 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{3} + 2 \cdot \frac{1}{6} = 1$$

$$\mathbb{E}Y = 0\cdot\left(\frac{2}{6}+\frac{1}{12}\right) + 1\cdot\left(\frac{1}{12}+\frac{2}{6}\right) + 2\cdot\left(\frac{1}{12}+\frac{1}{12}\right)$$

$$= \frac{5}{12} + \frac{1}{3} = \frac{5}{12} + \frac{4}{12} = \frac{9}{12} = \frac{3}{4}$$

By Law of Total Expectation

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}Y$$

Check:

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y|X=0]\,p(X=0) + \mathbb{E}[Y|X=1]\,p(X=1)$$

$$= \frac{1}{2}\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{4}+\frac{1}{2} = \frac{3}{4} = \mathbb{E}Y$$

Notice that $\mathbb{E}[Y|X]$ is a random variable

and

$\mathbb{E}[Y|X=x]$ is a number

Properties

I) if y is a function of x, $y=f(x)$

then $\mathbb{E}[y|x]=y$

II) if $y \perp\!\!\!\perp x$, $\mathbb{E}[y|x]=\mathbb{E}[y]$

$$\mathbb{E}[y|X=x]=\int y\,p(y|x)\,dy \overset{\text{b/c } y \perp\!\!\!\perp x}{=} \int y\,p(y)\,dy$$

$$=\mathbb{E}Y$$

III) For any function f

$$\mathbb{E}\{y f(x) \mid x\} = f(x)\,\mathbb{E}\{y \mid x\}$$

IV) Linearity

$$\mathbb{E}[\alpha Y + Z \mid X] = \alpha\,\mathbb{E}[Y \mid X] + \mathbb{E}[Z \mid X]$$

V) Total Expectation

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}\{Y\}$$

Point Estimates for least squares regression

$p(y|x)$ tells us everything about y given x

but we want a point estimate $y = f(x)$

$$f^* = \underset{f}{\operatorname{argmin}} \ \mathbb{E}\left(y - f(x)\right)^2$$

Note by the law of total expectation

$$\mathbb{E}\left(y - f(x)\right)^2 = \mathbb{E}\left[\mathbb{E}\left[\left(y - f(X)\right)^2 \mid X = x\right]\right]$$

so f^* can be obtained by choosing

$$f^*(x) = \underset{c}{\operatorname{argmin}} \ \mathbb{E}\left[(y - c)^2 \mid X = x\right]$$

$$f^*(x) = \underset{c}{\operatorname{argmin}} \ \mathbb{E}[(y-c)^2 \mid X=x]$$

notice

$$\mathbb{E}[(y-c)^2 \mid X=x]$$

$$= \mathbb{E}[y^2 - 2yc + c^2 \mid X=x]$$

$$= \mathbb{E}[y^2 \mid X=x] - 2c\,\mathbb{E}[y \mid X=x] + c^2$$

choose c^* by setting derivative to 0, get

$$-2\,\mathbb{E}[y \mid X=x] + 2c^* = 0$$

$$\Rightarrow c^* = \mathbb{E}[y \mid X=x]$$

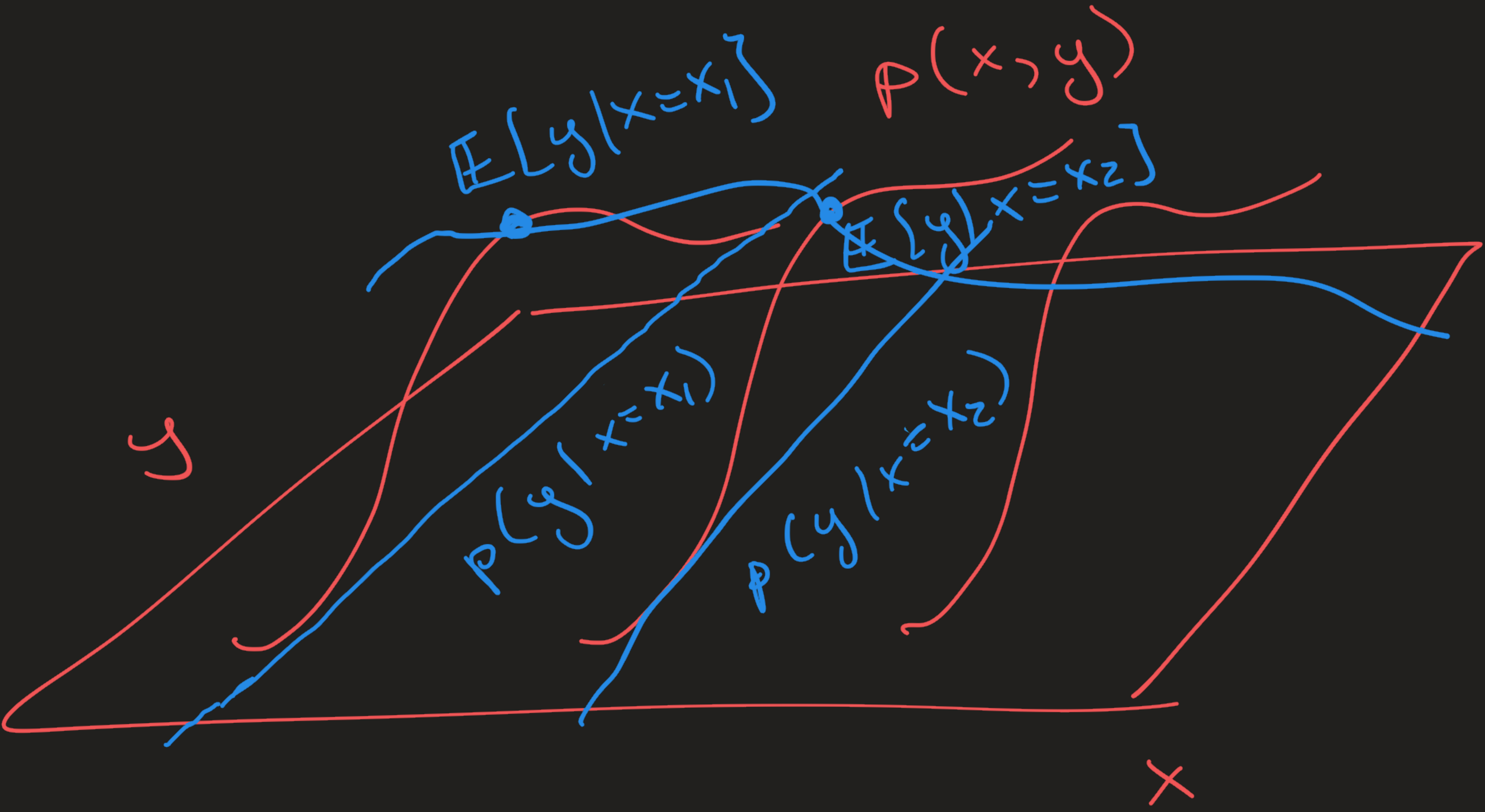
so $\quad f^*(x) = \mathbb{E}[y \mid X=x]$

Notice that this is the <u>BEST</u> possible estimate for y, given x, in the least squares sense

Call

$$\eta(x) = \mathbb{E}[y|x]$$

the <u>regression function</u>



$$f(x) = \mathbb{E}[y \mid x = x]$$