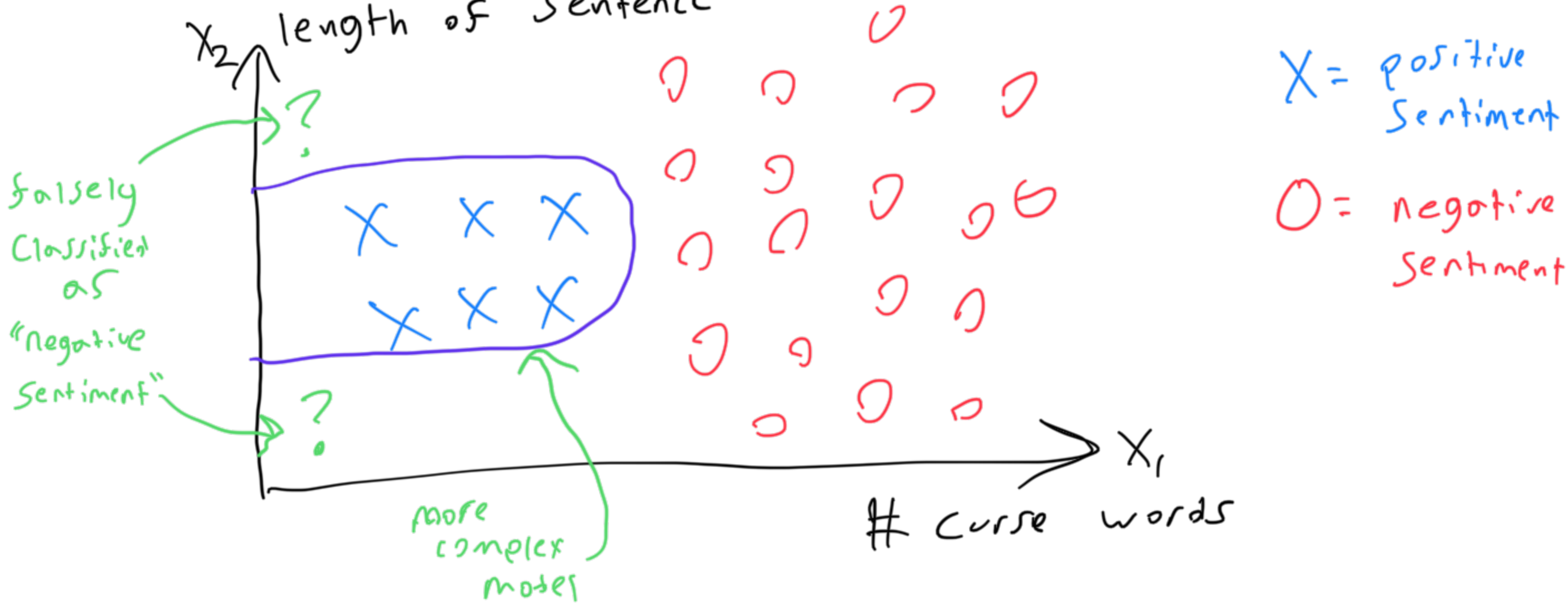
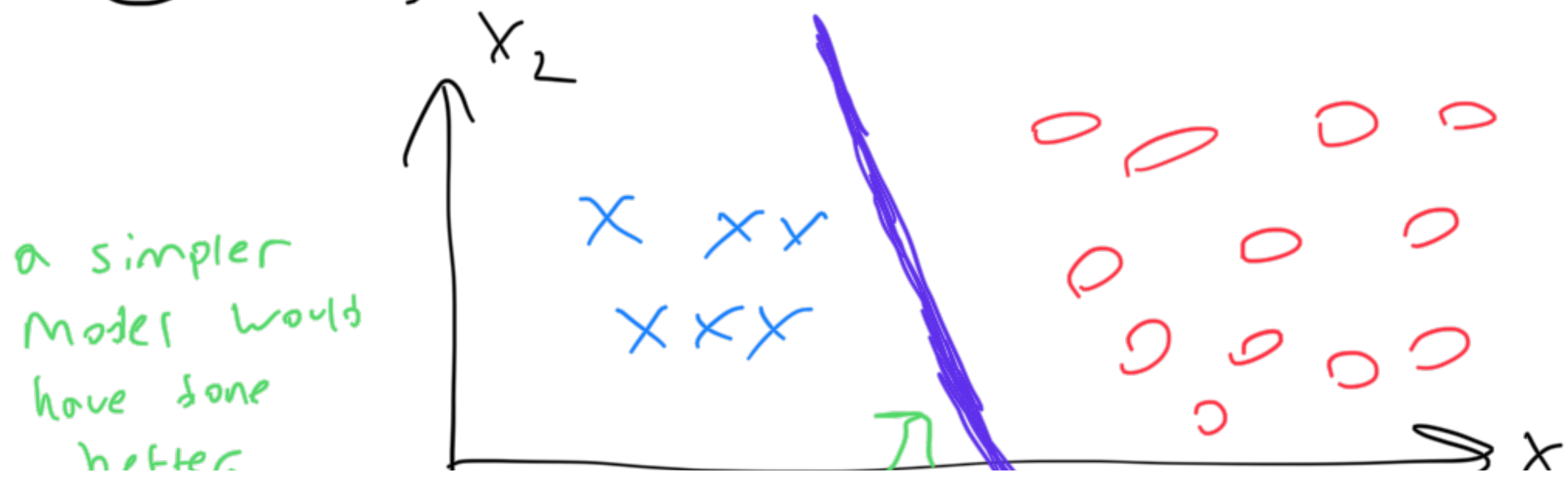
Day 9: Regularization and Probability

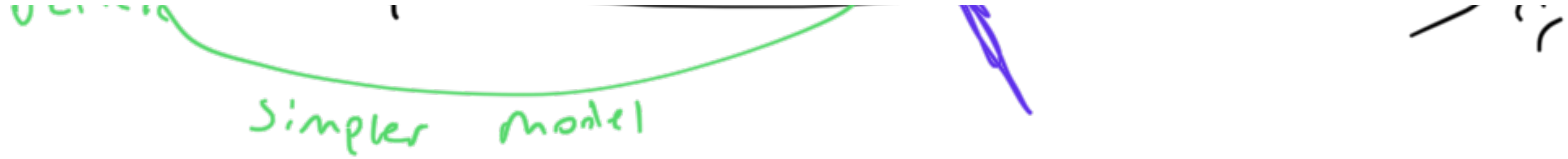
Overfitting Situations:

① using too many inputs

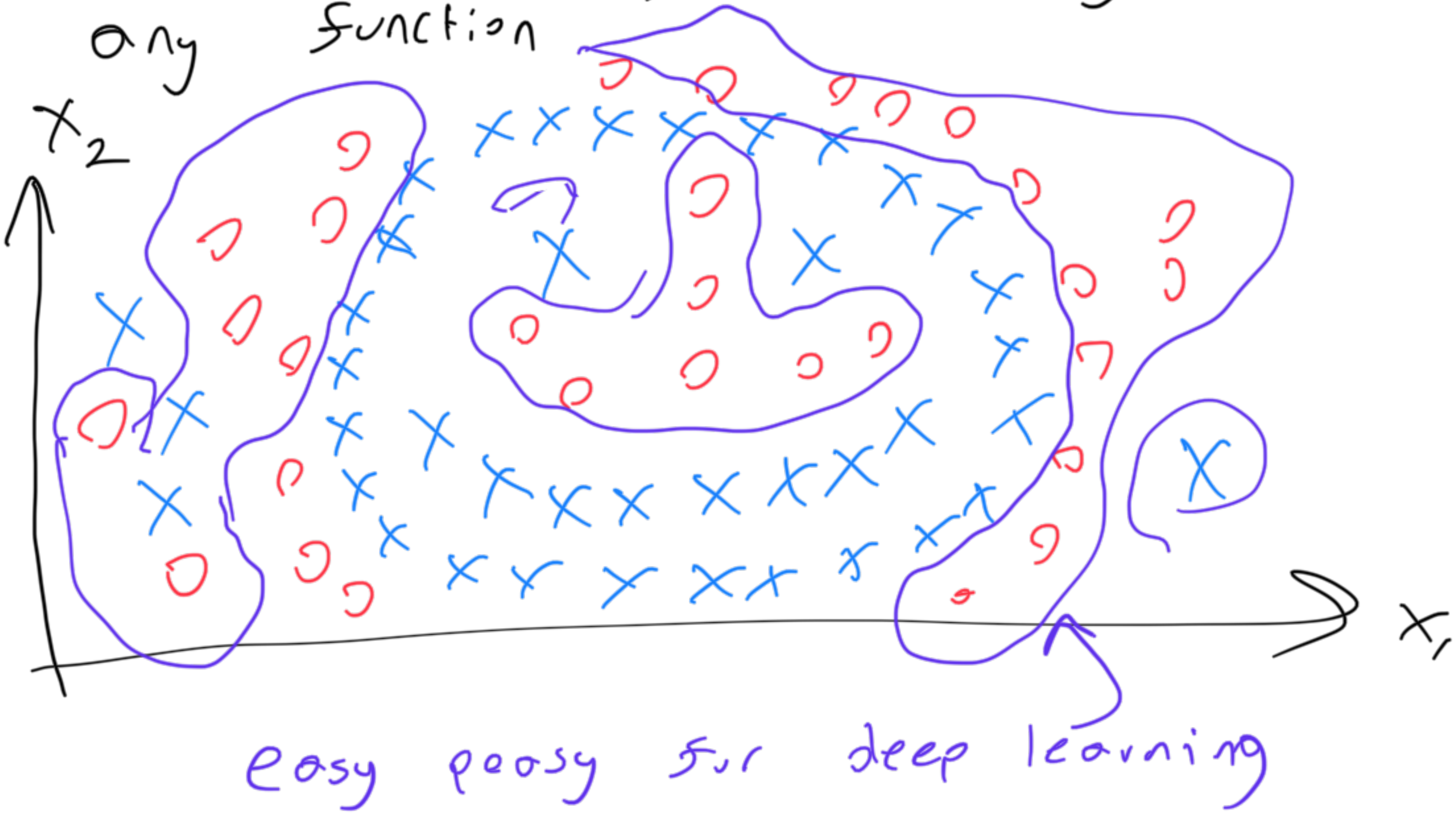


② using a too complex of a model





Modern ML (i.e, deep learning) can learn any function



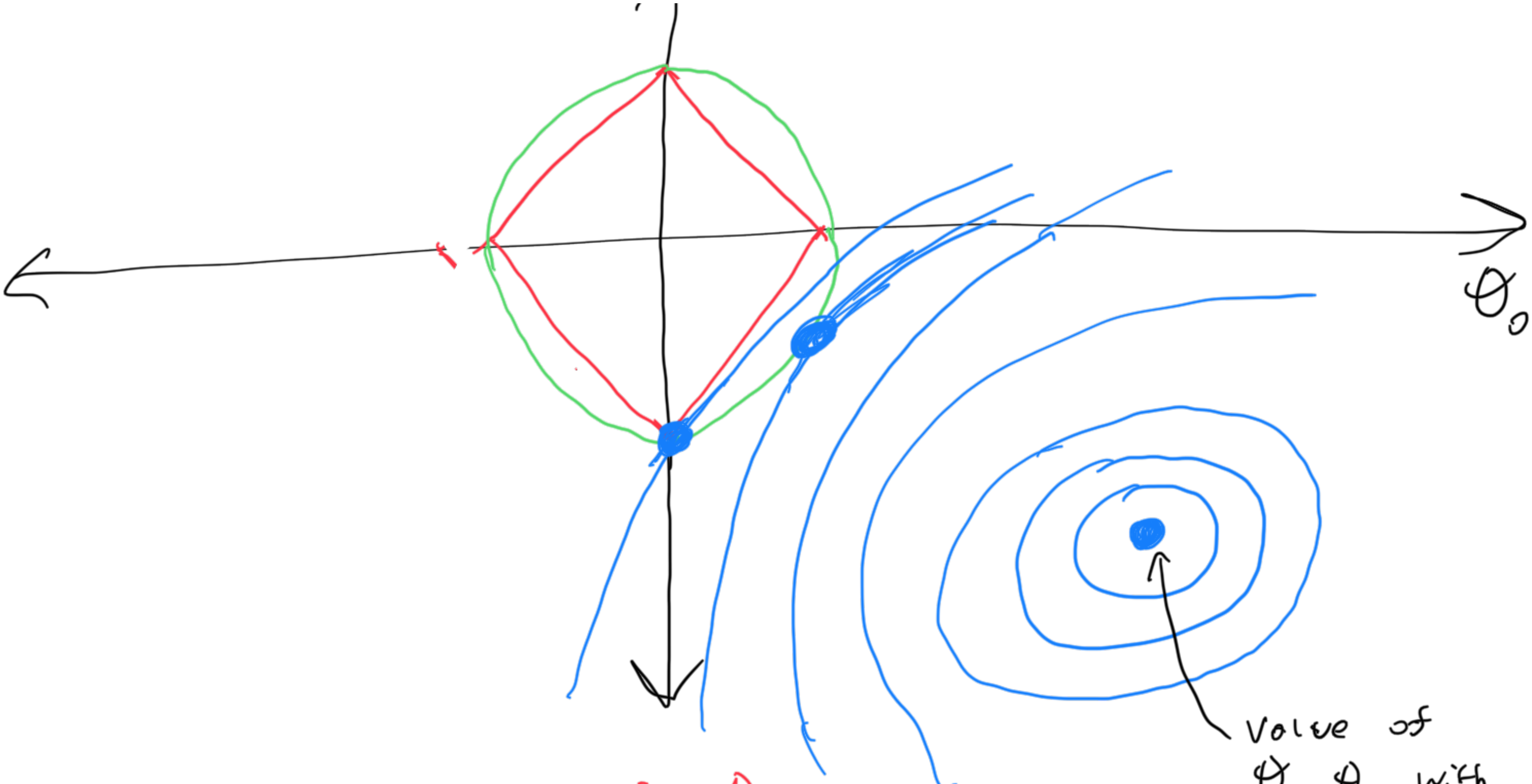
<u>Regularization</u> helps us constrain the model from doing "so well" that it fails to generalization

L1 Regularization: add $\lambda \sum_{i=1} |w_i|$ to loss function

L2 Regularization: add $\lambda \sum_{i=1}^{p} w_i^2$ to loss function

For example: minimize $\underbrace{MSE + \overbrace{\lambda \sum_{i=1}^{p} w_i^2}^{\text{regularization term}}}_{\text{full loss function with regularization}}$

L1 regularization leads to feature selection
↳ where we remove inputs from the model

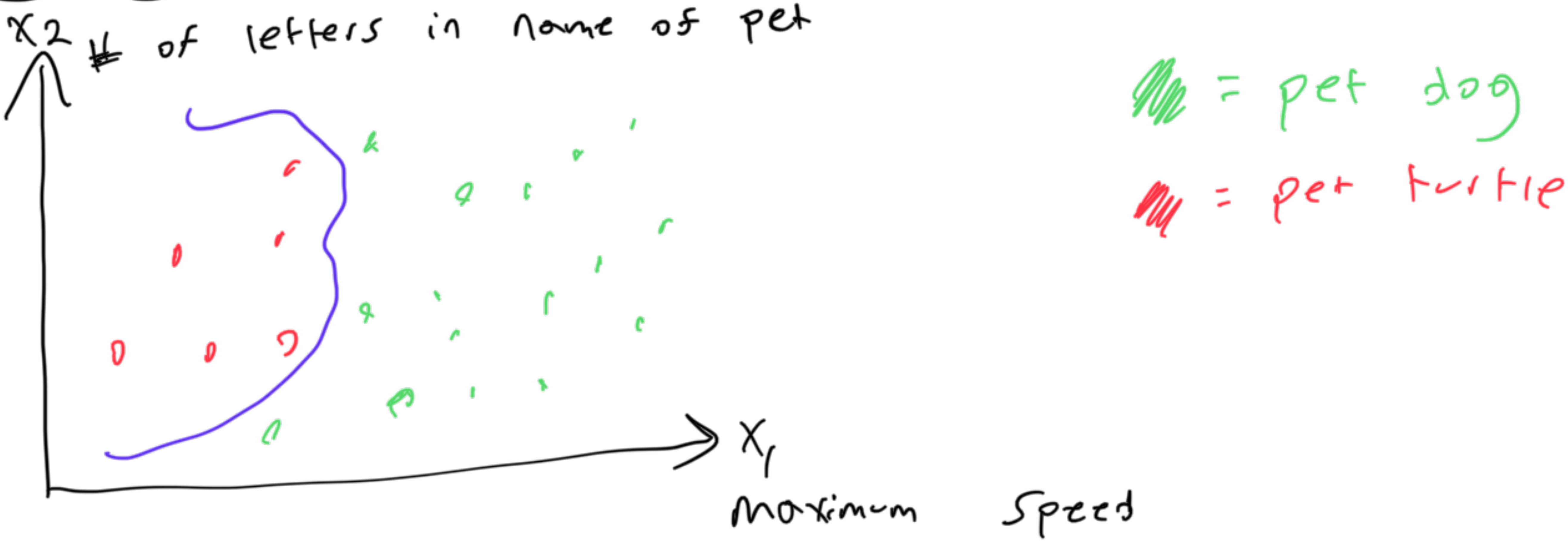


= combinations of θ_0 and θ_1 which form a particular value for the L1 regularization term (e.g., $|\theta_0| + |\theta_1| = 1$)

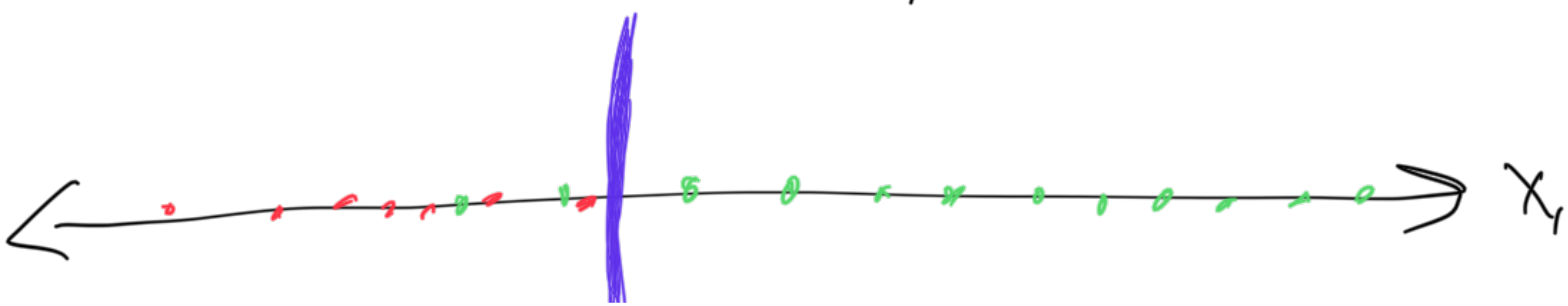
= same thing but for L2

L2 leads to "dampened" or smaller model weights for all features

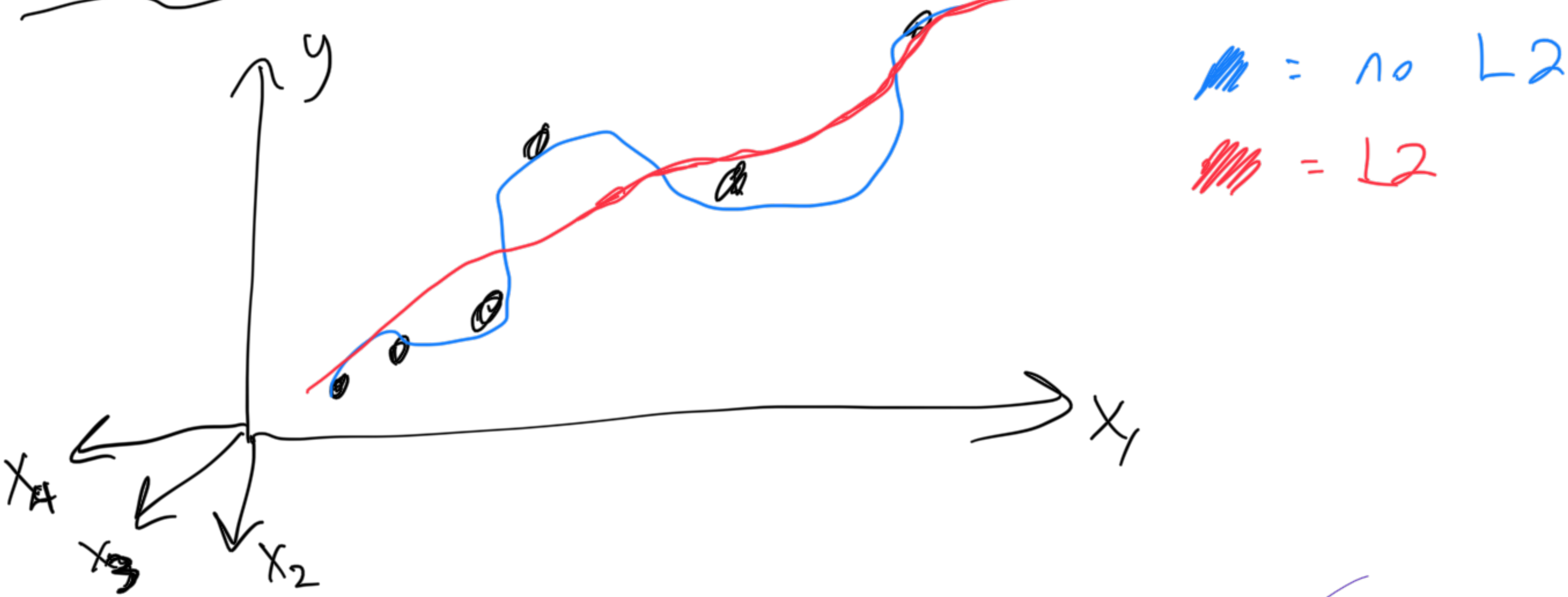
feature selection example



after feature selection, removing X_2
(e.g., after L1 regularization)



Smaller weights example



Probability

Notation:

$P(A)$ = "probability of event A"

$P(A|B)$ = "probability of event A given event B"

$P(\bar{A})$ = "probability of event A not occurring"

Basic Rules:

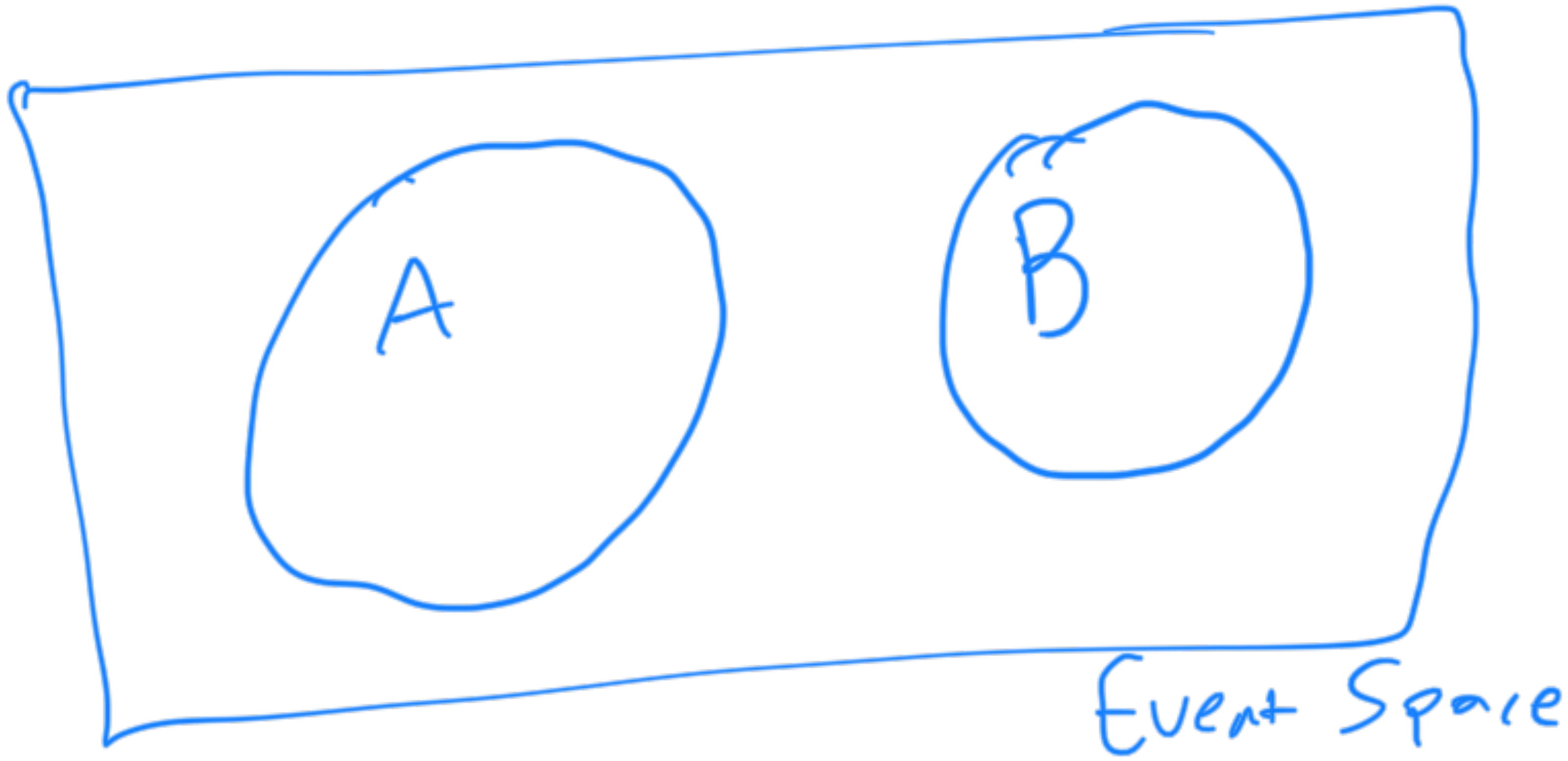
* sum of probabilities in an event space = 1
* $0 \leq P(X) \leq 1$
* $P(\bar{A}) = 1 - P(A)$

$P(A) = 1 - P(\bar{A})$

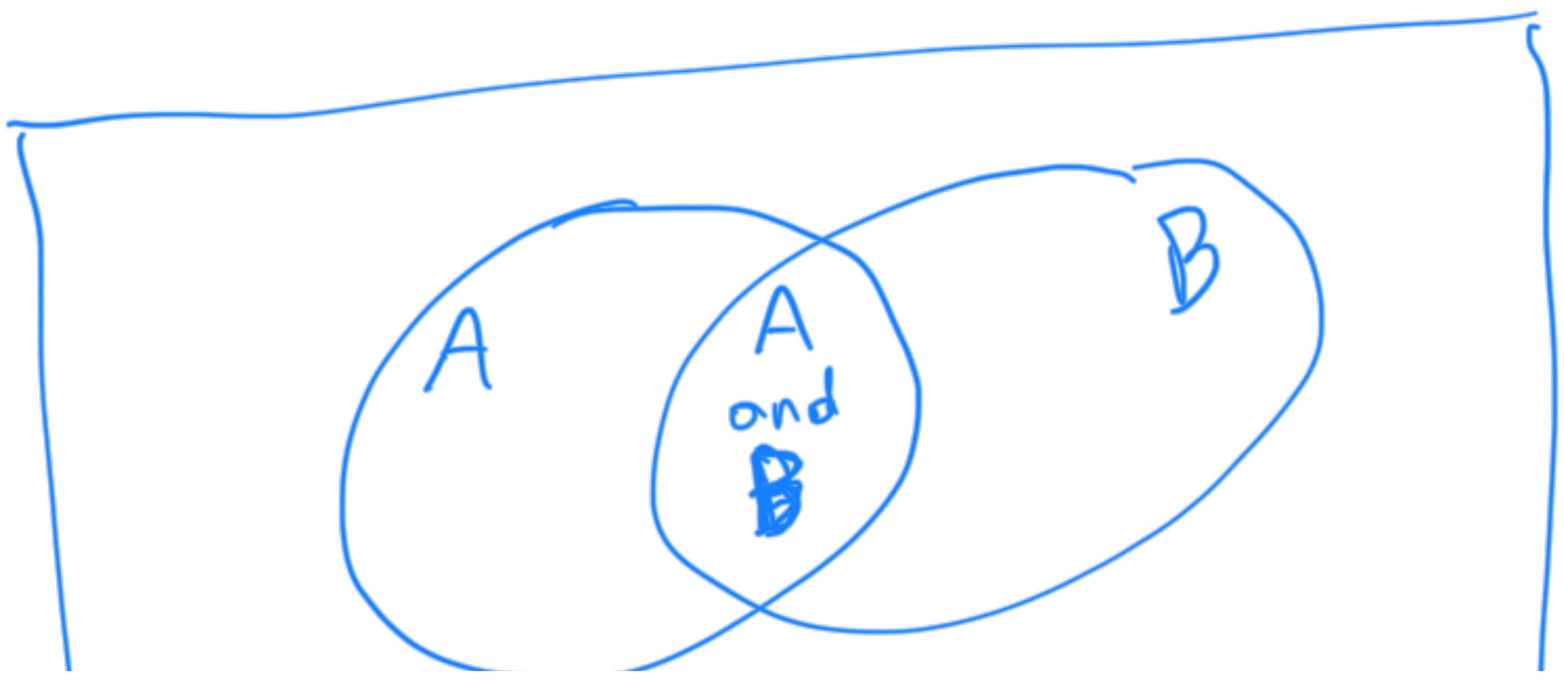
$P(A) = 1 - P(A)$

$* \ P(A \text{ or } B) = P(A) + P(B)$

if A and B are "mutually exclusive"



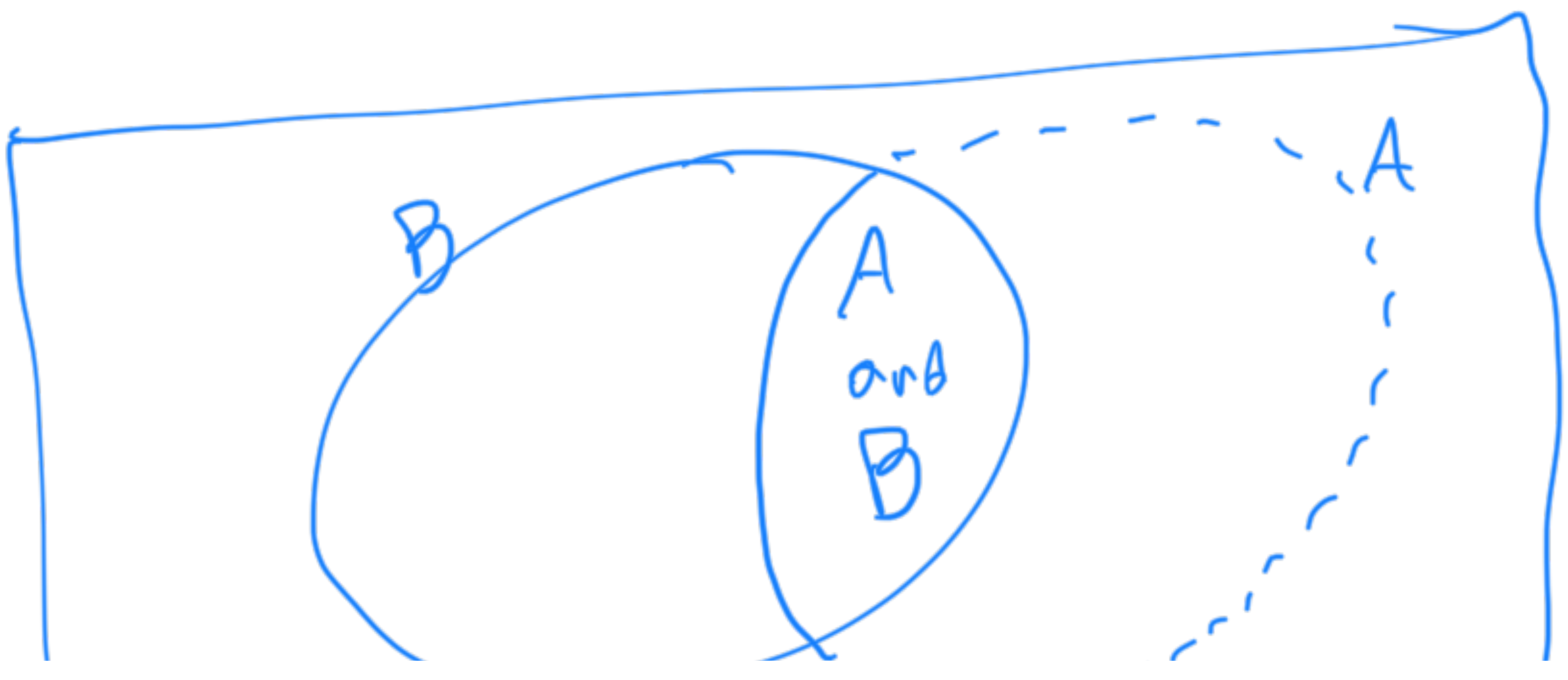
$* \ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



* $P(A \text{ and } B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$
* $P(A \text{ and } B) = P(A) \cdot P(B)$
  if A and B are independent

* Bayes' Rule:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$



Naive Bayes Classification:

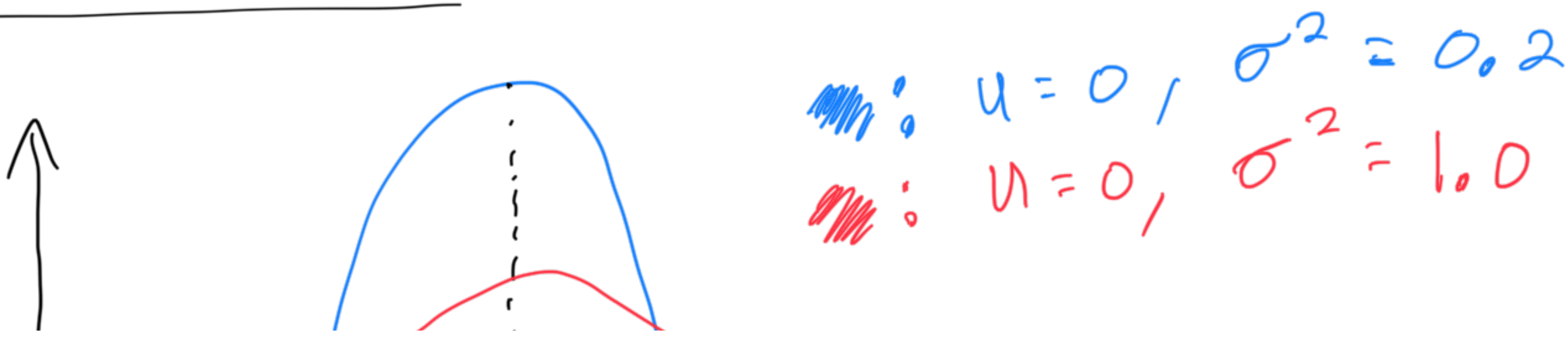
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

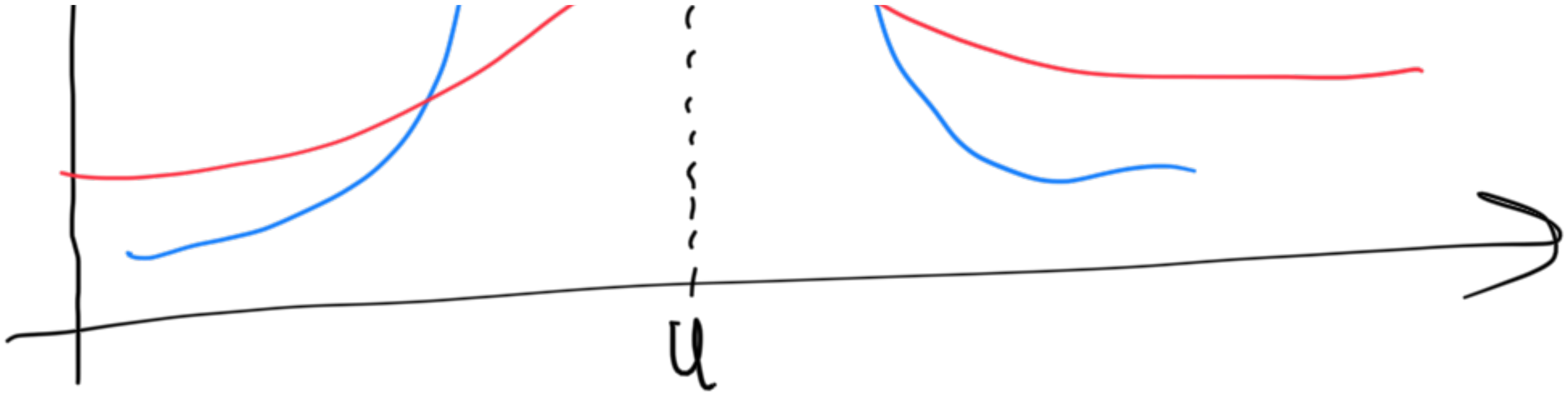
y: output classes

x: data

Probability Distributions

Normal Distribution ("Bell Curve")





Probability Density Function (PDF):

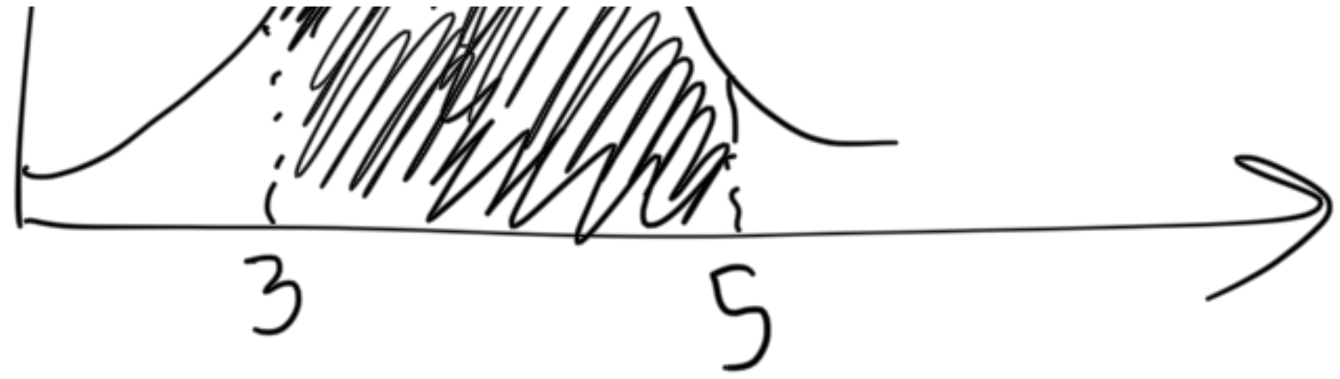
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(ugly equation, no need to memorize)

x: input

μ, σ: parameters

$$P(3 < X < 5) = \int_3^5 f(x)\,dx$$



= AUC of PDF from $x = 3$ to $x = 5$

Bernoulli Distribution

Probability of "flipping a coin" with probability of heads p

Probability Mass Functions (PMF):

$$f(x) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

for repeated coin tosses: $p^y (1-p)^{(1-y)}$

Example:

What's probability of H, H, T, H, T when flipping a coin with 60% probability of heads?

$$= 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.6 \cdot 0.4$$

$$= 0.6^3 \cdot 0.4^2$$

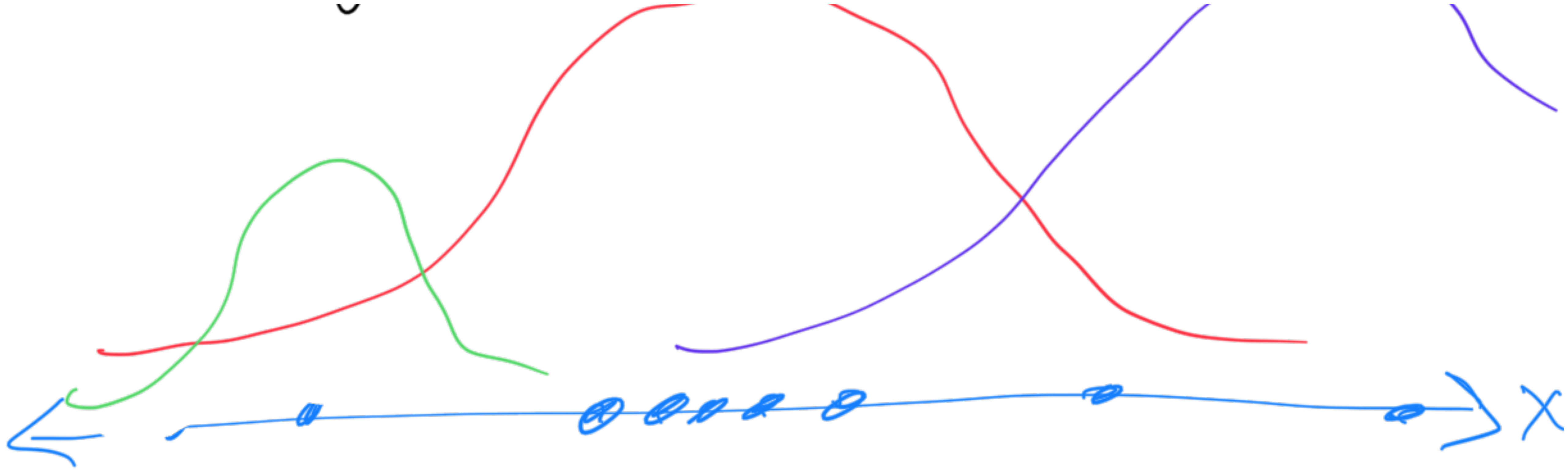
Maximum Likelihood Estimation (MLE)

Likelihood:

$L(\theta | X) =$ "likelihood of model parameters θ given the data X"

Example:

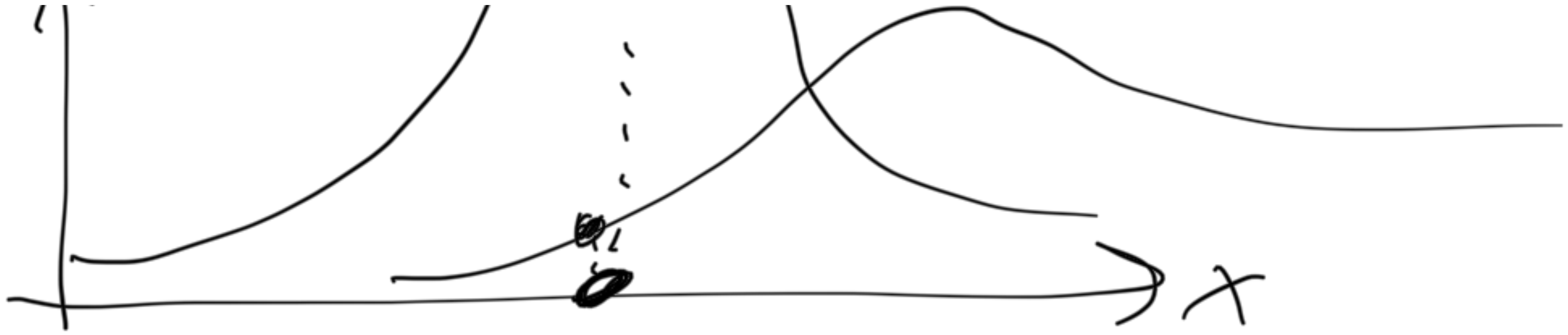
data:



, , are different parameterizations of the normal distribution (i.e., are different μ and σ values)

Assuming each data point is independent from each other (the usual case), then:

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1 \mid \theta) \cdot f(x_2 \mid \theta) \cdot \ldots \cdot f(x_n \mid \theta)$$



If you have many data points, then multiplying fractions repeatedly will underflow

$0.001 \times 0.001 \times \ldots$

Therefore, we take the log:

[recall $\log(ab) = \log(a) + \log(b)$]

$$\log L(\theta | X) = \log L(\theta | X_1) + \ldots + \log L(\theta | X_n)$$

$\log(f(x))$ increases/decreases the same as $f(x)$

Maximum Likelihood Estimation:

$$\frac{\partial L(\theta | X)}{\partial \theta} = 0$$

solve for θ

regular calculus optimization problem

Big Takeaway for Linear Regression:

Maximizing Likelihood = Minimizing MSE

for linear regression