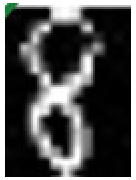
CS 189/289

Today's lecture:

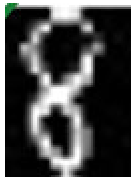
- Maximum likelihood estimation (MLE)

Recall from last class:

Problem of digit classification from handwriting: is  a "7", yes or no?

- 60K training examples of digits (6K per class)
- Each digit is a 28 x 28 pixel grey level image.

Recall from last class:

Problem of digit classification from handwriting: is  a "7", yes or no?

1	1	1	1
0	.5	1	0
.5	1	0	0
1	0	0	0

image patch

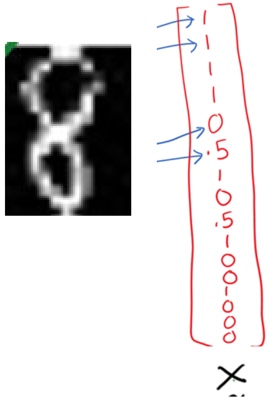
Note that there are several ways to construct a feature vector. This is one example.

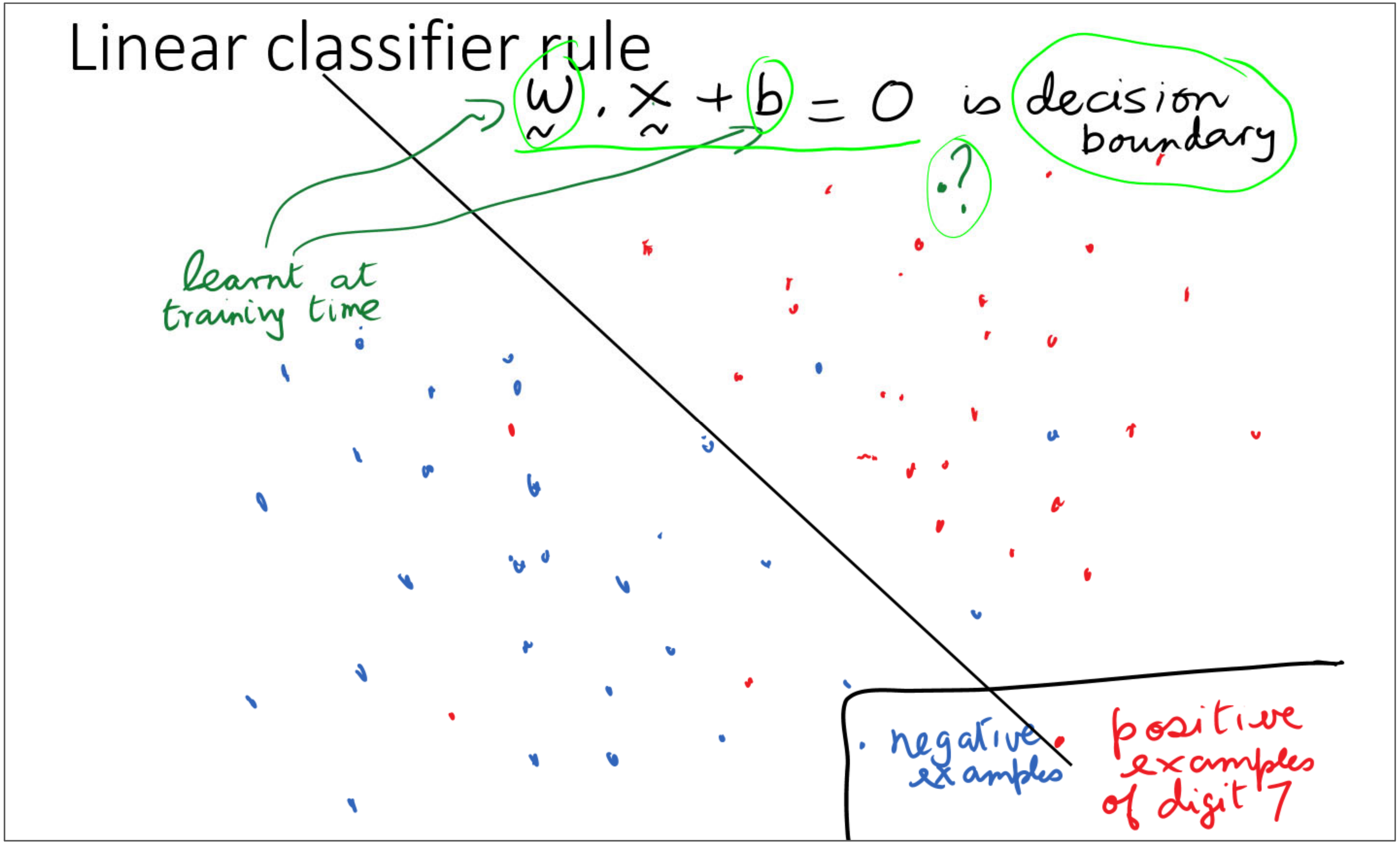
$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ .5 \\ 1 \\ 0 \\ .5 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Feature vector $\mathbb{R}^{16}$

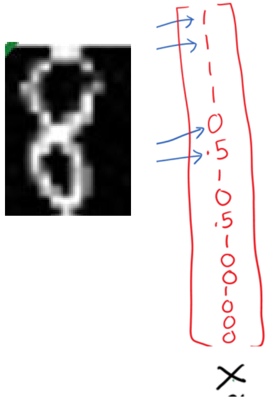
- 60K training examples of digits (6K per class)
- Each digit is a 28 x 28 pixel grey level image.

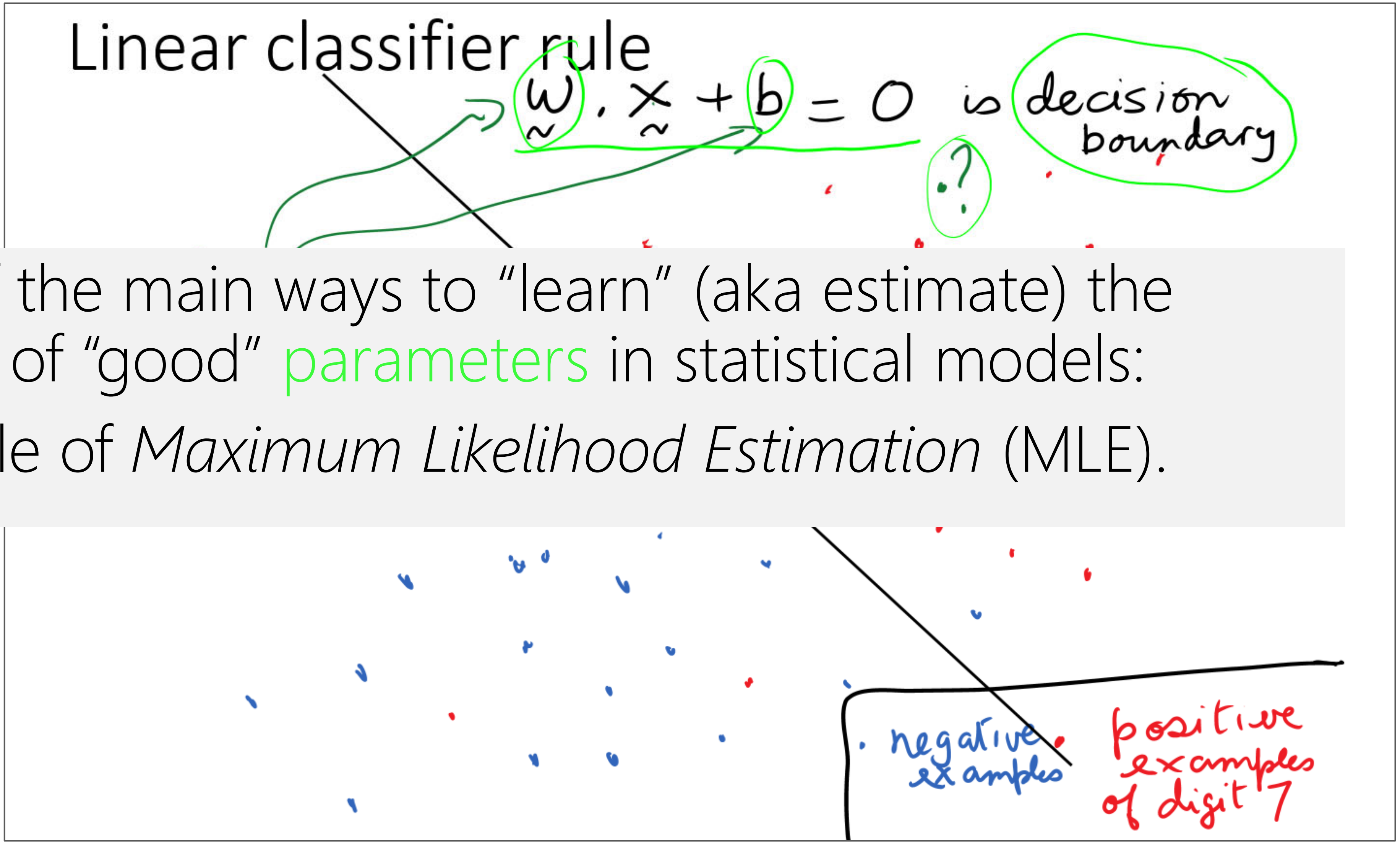
Recall from last class:





Recall from last class:





- One of the main ways to "learn" (aka estimate) the setting of "good" parameters in statistical models:
- Principle of *Maximum Likelihood Estimation* (MLE).

ML: main concepts

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$

$x_i \in R^D$

$y_i \in R$ or $y_i \in \{-1,1\}$

classification

regression

ML: main abstract ideas

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in R^D$, $y_i \in R$ or $y_i \in \{-1,1\}$

SUPERVISED

"label" provides supervision

$D = \{(x_i\ )\}_{i=1}^{N}$ $\quad x_i \in R^D$

UNSUPERVISED

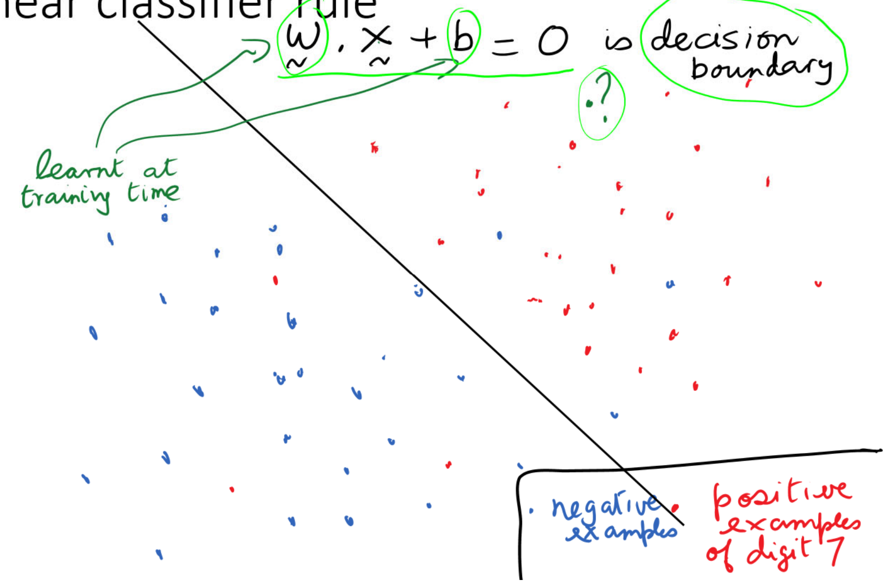
ML: main abstract ideas

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ $x_i \in R^D$ $y_i \in R$ or $y_i \in \{-1,1\}$

- Model class: $f(x|w, b) = w^T x + b$ **Linear Models**

aka hypothesis class

Linear classifier rule



ML: main abstract ideas

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in R^D$, $\; y_i \in R$ or $y_i \in \{-1,1\}$

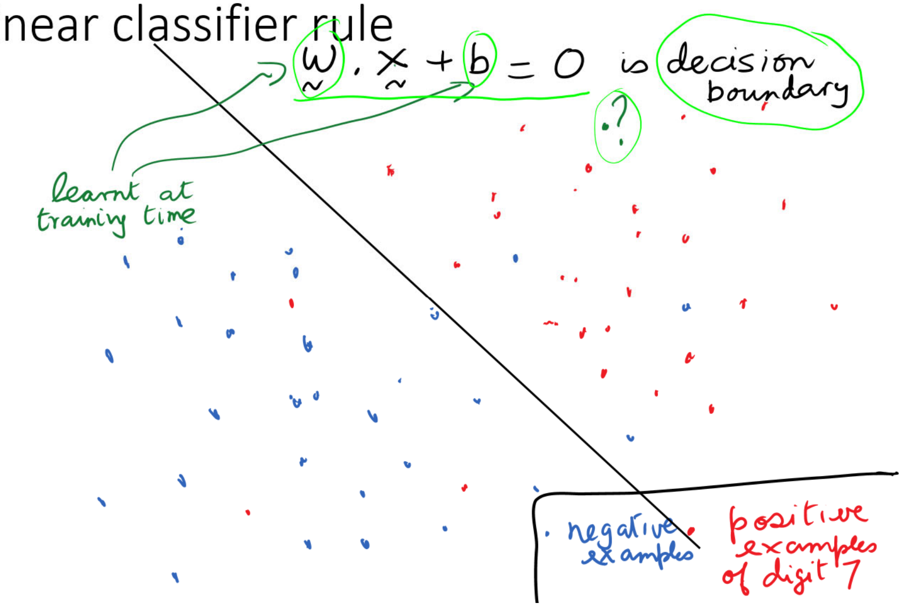
- Model class: $f(x|w, b) = w^T x + b$ **Linear Models**

aka hypothesis class

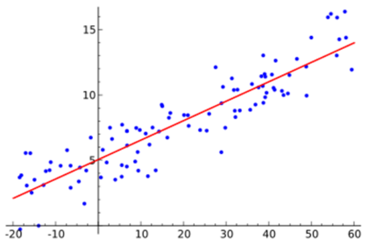
- **Optimization goal:** find "good" values of parameters (w, b).

  But was does "good" mean?

Linear classifier rule



ML: main abstract ideas



- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ $x_i \in R^D$ $y_i \in R$ or $y_i \in \{-1,1\}$

- Model class: $f(x|w,b) = w^T x + b$ **Linear Models**

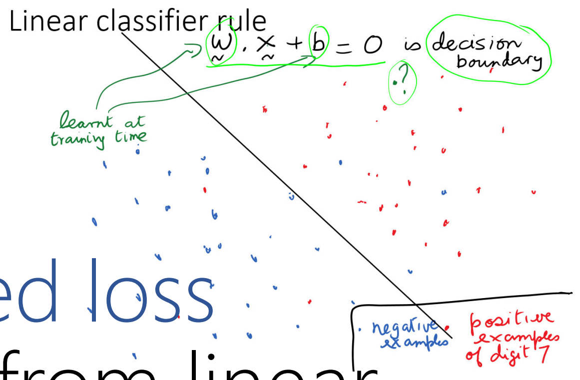
aka hypothesis class

- Loss Function: $L(a,b) = (a-b)^2$ **Squared Loss**

- Learning Objective: $\underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f(x_i \mid w,b))$

Optimization Problem

Maximum Likelihood Estimation (MLE)



This principle gives a useful, principled and widely-used loss function to estimate parameters of statistical models (from linear regression, to neural networks, and beyond).

RVs!

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ — $x_i \in R^D$, $y_i \in R$ or $y_i \in \{-1,1\}$
- Model class: $f(x|w,b) = w^T x + b$ — **Linear Models**
  aka hypothesis class
- Loss Function: $L(a,b) = (a-b)^2$ — **Squared Loss**
- Learning Objective: $\underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f(x_i \mid w,b))$

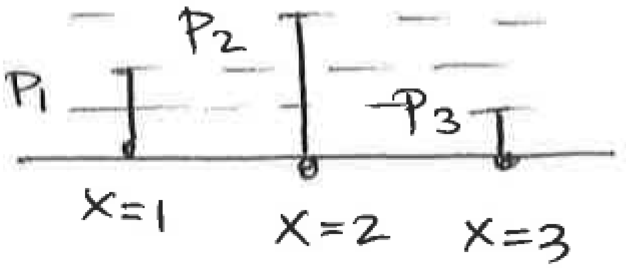
Optimization Problem

Reminder: probability distributions

Random variable (RV) is a function: $\mathbf{x} \rightarrow \mathbb{R}$ e.g. $p(\text{heads}) = 0.5$

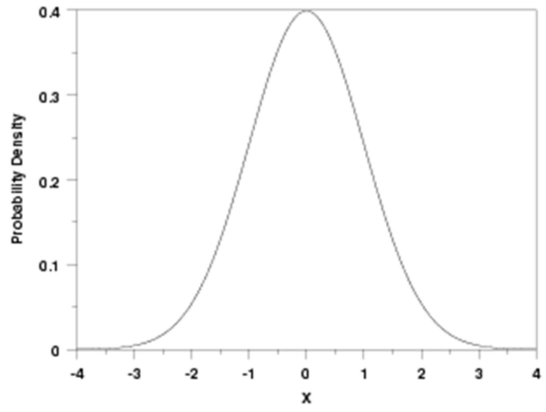
1. Discrete RV, e.g. coin toss heads/tails.
2. Continuous RV, e.g. height

Discrete RVs have a Probability Mass Function (PMF)



$P_1+P_2+P_3=1$

Continuous RVs have a Probability Density Function (PDF)



integrates to 1

e.g. distributions of discrete RVs

1. Bernouilli RV—model the toss of a coin that can be biased
   $P(heads) = p, \quad P(tails) = \mathbf{1 - p}$, parameter is $\boldsymbol{p}$.

e.g. distributions of discrete RVs

1. Bernouilli RV—model the toss of a coin that can be biased
   $P(heads) = p, \quad P(tails) = 1 - p$, parameter is p.
2. Binomial RV—model number of heads, $\mathbf{k}$, of n biased coin tosses.

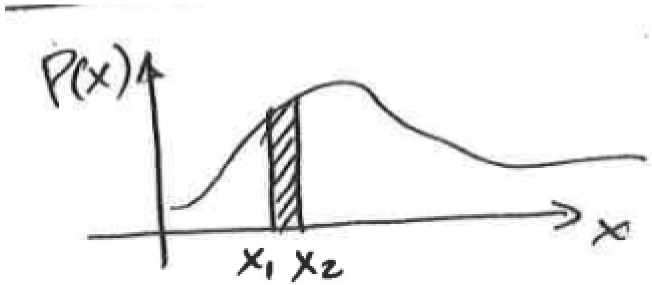
$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

e.g. distributions of discrete RVs

1. <u>Bernouilli RV</u>—model the toss of a coin that can be biased
   $P(heads) = p, \quad P(tails) = 1 - p$, parameter is p.
2. <u>Binomial RV</u>—model number of heads, $\mathbf{k}$, of n biased coin tosses.
   $$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$
3. <u>Poisson RV</u>– model number of mutations, k, occurring in a cell population with mean mutation rate, λ , over fixed time interval
   $$P(x=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Distributions of continuous RVs

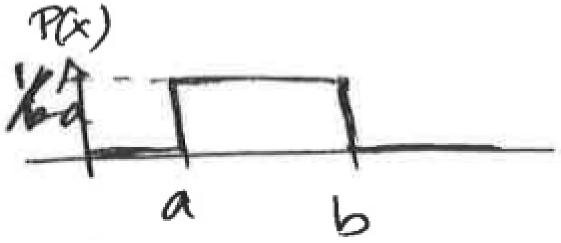
Continuous RVs have a Probability Density Function



$$\int_{x_1}^{x_2} P(x)\,dx$$ = Prob that the random variable X has a value in the interval $[x_1 \dots x_2]$

Examples:

1. Uniform -



Parameters: a, b

2. Gaussian

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$ Parameters: μ, σ

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

Multivariate distributions

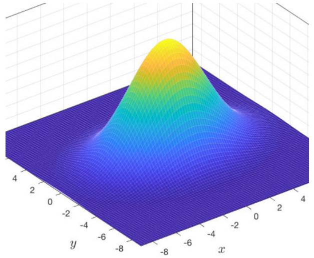
Space of outcomes is a vector instead of a scalar:

<u>Multinomial</u> (generalization from binomial):

- urn with balls of different colors.
- Pick a ball at random.
- $\boldsymbol{p_1}$ it is green, $\boldsymbol{p_2}$ it is blue and $\boldsymbol{p_3}$ it is red

<u>Multivariate Gaussian:</u>

- Mean is a vector, and variance becomes covariance.
- Will learn more about this next lecture.



The basic set-up of MLE

Same as $p(x|\theta)$

- Given data $D = \{x_i\}_{i=1}^{N}$ for $x_i \in R^d$
- Assume a set (family) of distributions on R^d, $\{p_\theta(x)|\theta \in \Theta\}$.

e.g. mean (μ) and variance (σ^2) for $x \in \mathbb{R}^1$

The basic set-up of MLE

Same as $p(x|\theta)$

- Given data $D = \{x_i\}_{i=1}^N$ for $x_i \in R^d$
- Assume a set (family) of distributions on R^d, $\{p_\theta(x)|\theta \in \Theta\}$.
- Assume D contains samples from one of these distributions:

$$x_i \sim p_{\hat{\theta}}(x)$$

- This assumes that each element of D is *identically and independently distributed* (iid).

The basic set-up of MLE

e.g mean (μ) and variance (σ^2) for $x \in \mathbb{R}^1$

- Given data $D = \{x_i\}_{i=1}^N$ for $x_i \in R^d$
- Assume a set (family) of distributions on R^d, $\{p_\theta(x) | \theta \in \Theta\}$.
- Assume D contains samples from one of these distributions:

$$x_i \sim p_{\hat{\theta}}(x)$$

- This assumes that each element of D is *identically and independently distributed* (iid).

"likelihood function"

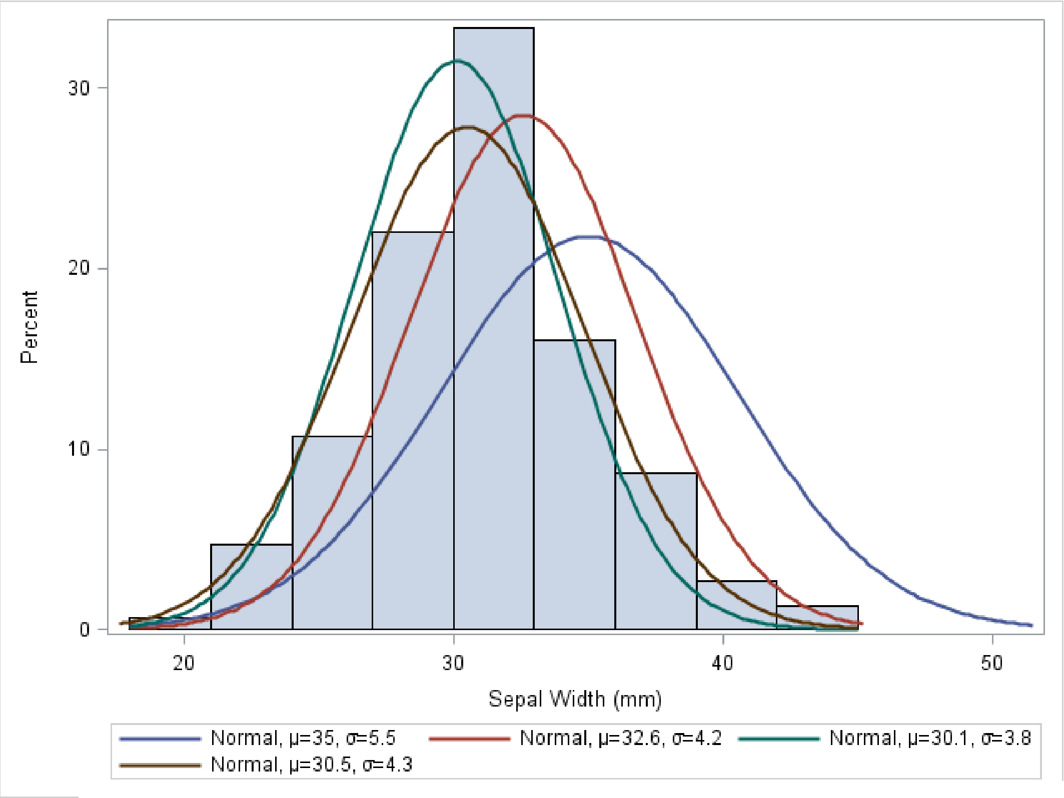
Goal of MLE: "learn"/estimate the value of $\boldsymbol{\theta}$ that defines the distribution from which the data came.

Definition: $\boldsymbol{\theta_{MLE}}$ is a MLE for $\boldsymbol{\theta}$ with respect to the data and set of distributions, if $\boldsymbol{\theta_{MLE}} = \underset{\theta \in \Theta}{\mathbf{argmax}}\, p(D|\theta)$.

function of θ.

The basic set-up of MLE

$$\boxed{\theta_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(D|\theta)}$$



$$D = \{x_i\}_{i=1}^{N} = \{20.1, 33.8, 34.6, 36.2, \dots\}$$

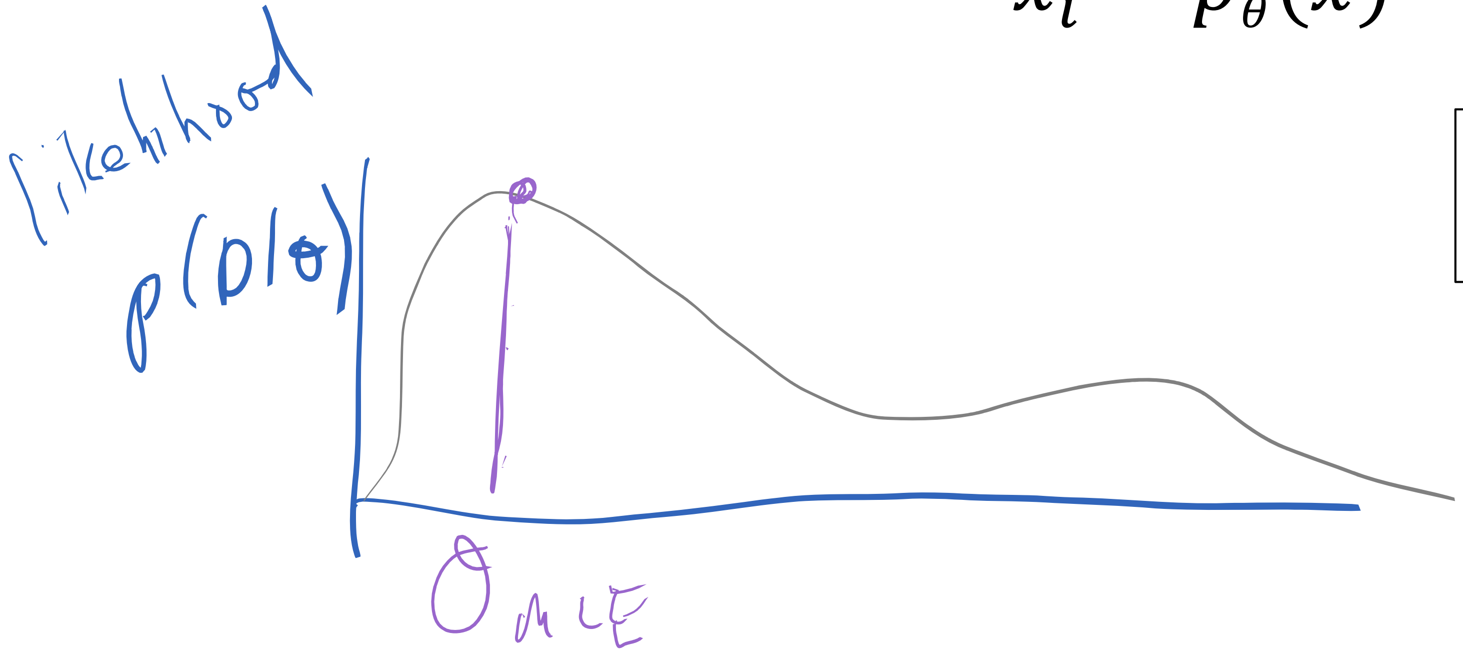
Note that $p(D|\theta) = p\left(\{x_i\}_{i=1}^{N}\middle|\theta\right) = \prod_{i=1}^{N} p(x_i|\theta)$

because iid

The basic set-up of MLE

- Given data $D = \{x_i\}_{i=1}^{N}$ for $x_i \in R^d$
- Assume a set (family) of distributions on R^d, $\{p_\theta(x) | \theta \in \Theta\}$.
- Assume D contains samples from one of these distributions:

$$x_i \sim p_{\hat{\theta}}(x)$$



$$\boxed{\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(D|\theta)}$$

Is there always one unique MLE parameter value?

e.g. $\mathcal{N}(x \mid \mu, \sigma)$

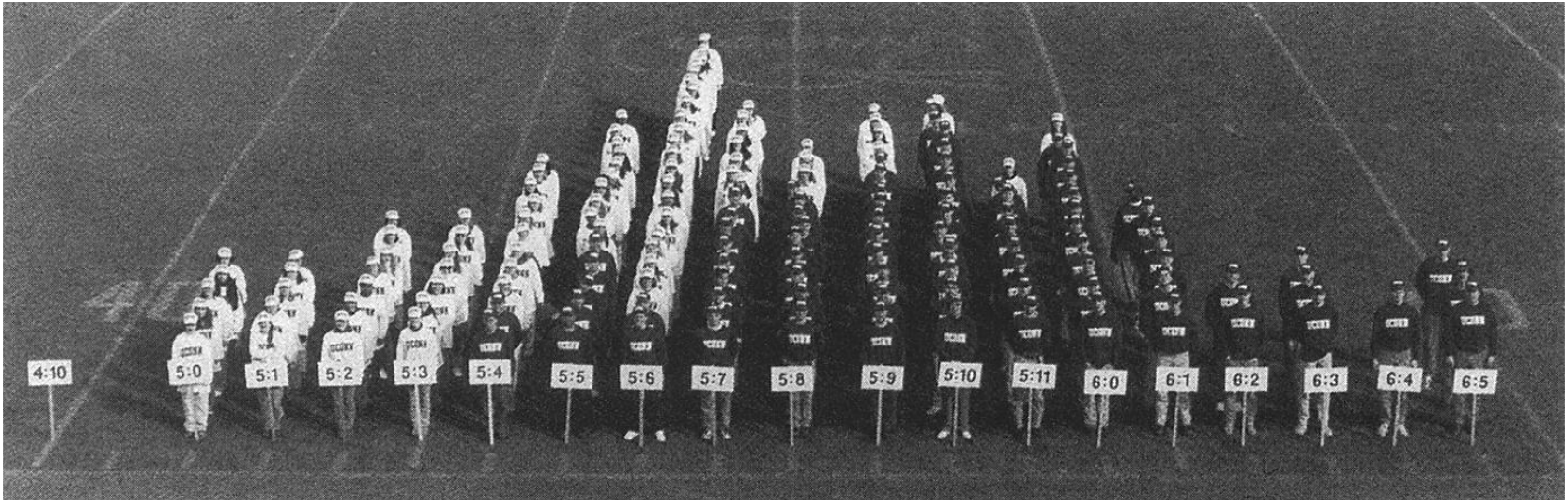
vs

$\mathcal{N}(x \mid 2+\mu, \sigma)$

Some properties of MLE

- The MLE is a *consistent* estimator: meaning that as we get more and more data (drawn from one distribution in our family), then we converge to estimating the true value of $\boldsymbol{\theta}$ for $\boldsymbol{D}$.
- The MLE is *statistically efficient*: it's making good use of the data available to it ("least variance" parameter estimates).
- The value of $p(D|\theta_{MLE})$ is invariant to re-parameterization.
- MLE can still yield a parameter estimate even when the data were not generated from that family (phew & caveat emptor).

e.g. MLE for univariate Gaussian



- Arguments can be made from the Central Limit Theorem that height is normally distributed.
- Suppose you were given a set if height measurements, $\{x_i\}$, how would you derive the estimate for the mean and variance, using MLE?

e.g. MLE for univariate Gaussian

Goal: $\theta_{MLE} = \underset{\theta \in \Theta}{\mathbf{argmax}}\, p(D|\theta)$ from set of data $D = \{x_i\}_{i=1}^{N}$

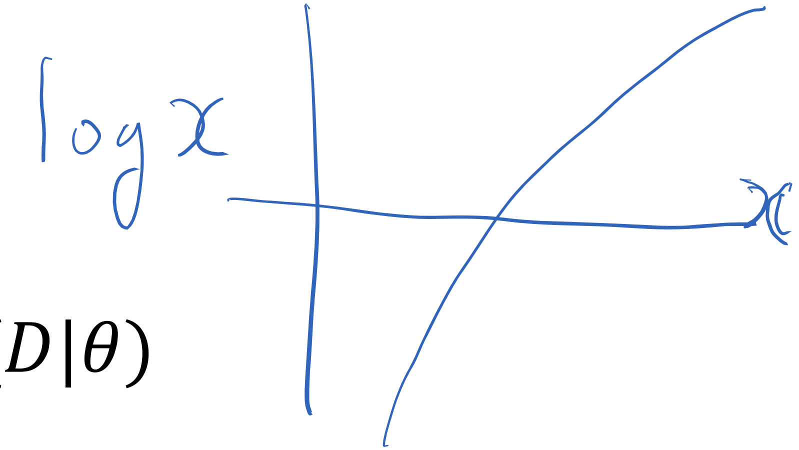
- Assume data are generated as $X \sim N\left(x\middle|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{(x-\mu)^2}{2\sigma^2}$
- So assume MLE family of distributions, $p(X = x|\theta) = N(X|\mu, \sigma^2)$.
- Now our goal is to find $\theta_{MLE} = \left(\mu_{MLE}, \sigma_{MLE}^2\right) = \underset{\theta \in \Theta}{\mathbf{argmax}}\, p\left(D\middle|\mu, \sigma^2\right)$.
- First step, write down the likelihood function:
  - $p(D|\theta) = p\left(x_1, x_2, \dots x_N\middle|\mu, \sigma^2\right) = \prod_{i=1}^{N} p\left(x_i\middle|\mu, \sigma^2\right)$.
- The product of the terms is a little inconvenient to work with.

e.g. MLE for univariate Gaussian

- Likelihood: $p\left(x_1, x_2, \ldots x_N \middle| \mu, \sigma^2\right) = \prod_{i=1}^{N} p\left(x_i \middle| \mu, \sigma^2\right)$.

- The *log likelihood ("LL")* is a monotonically increasing function of the likelihood.

$$\log p(D|\theta) = \sum_{i=1}^{N} \log p\left(x_i \middle| \mu, \sigma^2\right)$$



- Therefore $\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(D|\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta)$

e.g. MLE for univariate Gaussian

- Now we have a concrete optimization problem to work with:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log p\left(x_i \middle| \mu, \sigma^2\right)$$

- How will we solve this optimization problem?
- Find a setting of the parameters for which the partial derivatives are 0 (*i.e.*, a stationary point).
- Then check whether the setting is a maximum (negative second derivative), a minimum, etc. (first year calculus).
- (if #params>1, check if Hessian is negative definite; for 1D Gaussian, Hessian is diagonal, so can check each separately).

e.g. MLE for univariate Gaussian

- Find the setting of the parameters that set the partial derivatives to zero:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\mu, \sigma^2)$$

- Lets expand out so we can take the derivative:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log p(x_i|\mu,\sigma^2) = \sum_i \frac{\partial}{\partial \mu} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \right]$$

$e.g.$ MLE for univariate Gaussian

- Find the setting of the parameters that set the partial derivatives to zero:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\mu, \sigma^2)$$

- Lets expand out so we can take the derivative:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log p(x_i | \mu, \sigma^2) = \sum_i \frac{\partial}{\partial \mu} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right]$$

$$= \sum_i \frac{\partial}{\partial \mu} \left[-\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

$$= \sum_i \left[0 + \frac{1}{\sigma^2} (x_i - \mu) \right] \xRightarrow{\text{set to zero}} \sum_i x_i = \sum_i \mu$$

$$\sum_i x_i = N\mu$$

$$\Rightarrow \mu = \frac{\sum_i x_i}{N}$$

e.g. MLE for univariate Gaussian

$$\frac{\partial^2(LL)}{\partial\mu^2} =$$

- Find the setting of the parameters that set the partial derivatives to zero:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\mu,\sigma^2)$$

- Lets expand out so we can take the derivative:

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{N}\log p(x_i|\mu,\sigma^2) = \sum_i \frac{\partial}{\partial\mu}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right]$$

$$= \sum_i \frac{\partial}{\partial\mu}\left[-\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= \sum_i\left[0 + \frac{1}{\sigma^2}(x_i-\mu)\right] \Rightarrow$$

set to zero

$$\sum_i x_i = \sum_i \mu \Rightarrow \sum_i x_i = N\mu$$

$$\Rightarrow \mu = \frac{\sum_i x_i}{N}$$

e.g. MLE for univariate Gaussian

$$\frac{\partial^2 (LL)}{\partial \mu^2} = \sum_i \frac{1}{\sigma^2} \cdot (-1) = -\frac{N}{\sigma^2} < 0$$

$\Rightarrow$ min.

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log p(x_i \mid \mu, \sigma^2) = \sum_i \frac{\partial}{\partial \mu} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right]$$

MLE

$$= \sum_i \frac{\partial}{\partial \mu} \left[-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

$$= \sum_i \left[0 + \frac{1}{\sigma^2} (x_i - \mu) \right] \Rightarrow$$

set to zero

$$\sum_i x_i = \sum_i \mu \Rightarrow \sum_i x_i = N\mu$$

$$\Rightarrow \mu = \frac{\sum_i x_i}{N}$$

e.g. MLE for univariate Gaussian

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\mu_{MLE}, \sigma^2_{MLE} = argmax \sum_{i=1}^{N} \log N(x_i|\mu,\sigma^2)$$

- Again, but this time for $\boldsymbol{\sigma^2}$:

$$\frac{\partial \mathcal{LL}}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)$$

$$= -\frac{N}{2}\cdot\frac{\partial}{\partial\sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial\sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)$$

$$= -\frac{N}{2}\cdot\frac{1}{2\pi\sigma^2}\cdot 2\pi + \frac{\partial}{\partial\sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)$$

$$= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(-\frac{1}{2}\cdot -1\cdot(\sigma^2)^{-2}\cdot 1\cdot(x_n-\mu)^2\Big)$$

$$= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{1}{2\sigma^4}\cdot(x_n-\mu)^2\Big)$$

$$0 = \frac{1}{2\sigma^2}\Big(-N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)$$

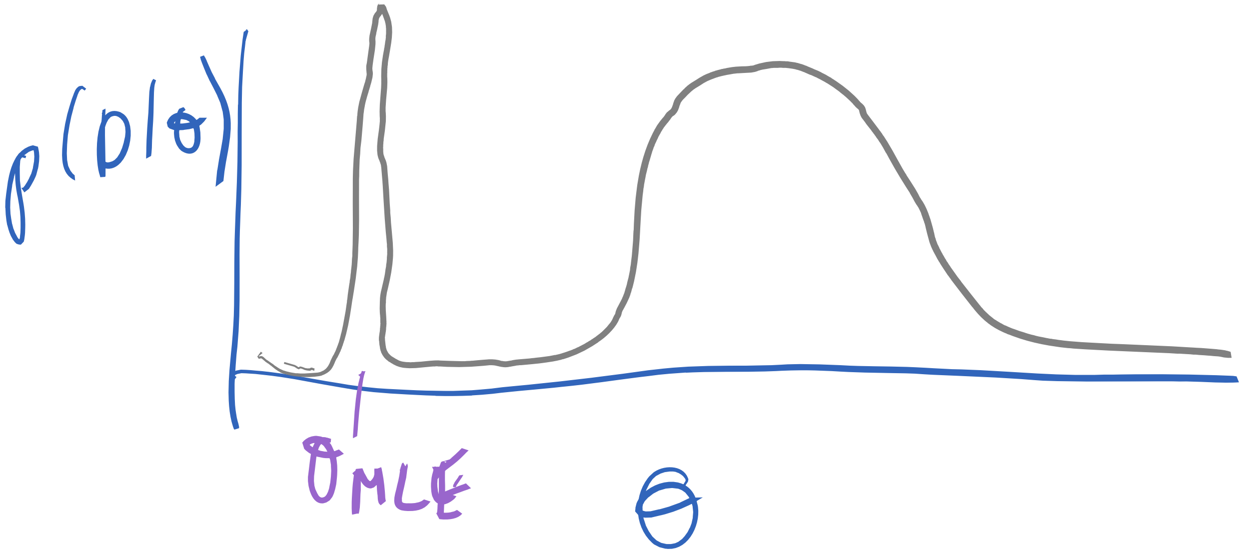
$$0 = -N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$\sigma^2 = \frac{1}{N}\sum^{N}(x_n-\mu)^2$$

σ^2_{MLE}

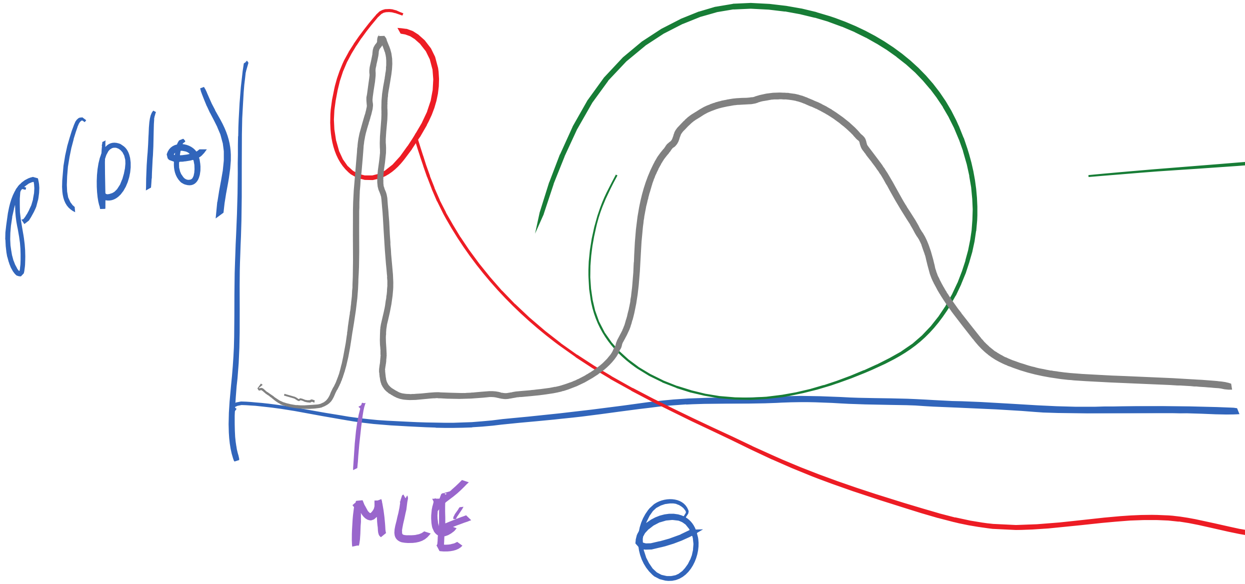
MLE yields a "point estimate" of our parameter

- When we perform MLE, we get just one single estimate of the parameter, $\boldsymbol{\theta}$, rather than a distribution over it which captures uncertainty.
- In Bayesian statistics, we obtain a (posterior) distribution over $\boldsymbol{\theta}$. We will touch more on this in a few lectures.



MLE yields a “point estimate” of our parameter

- When we perform MLE, we get just one single estimate of the parameter, $\boldsymbol{\theta}$, rather than a distribution over it which captures uncertainty.
- In Bayesian statistics, we obtain a (posterior) distribution over $\boldsymbol{\theta}$. We will touch more on this in a few lectures.

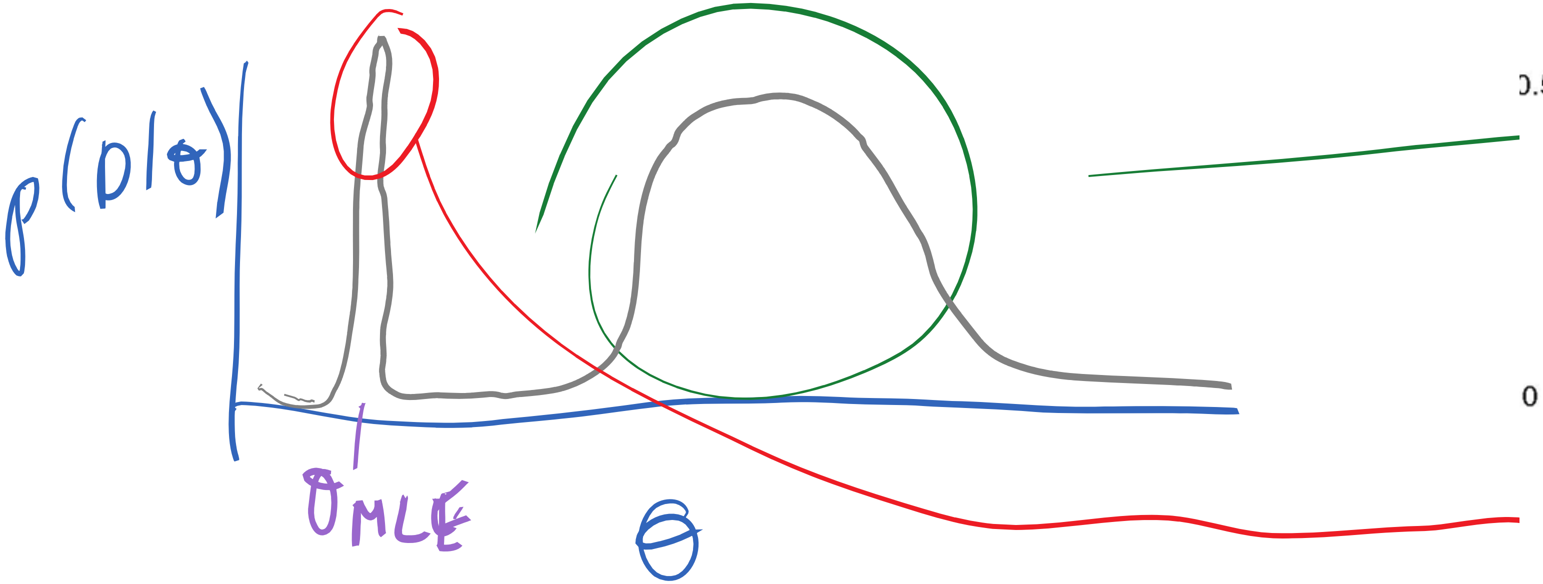


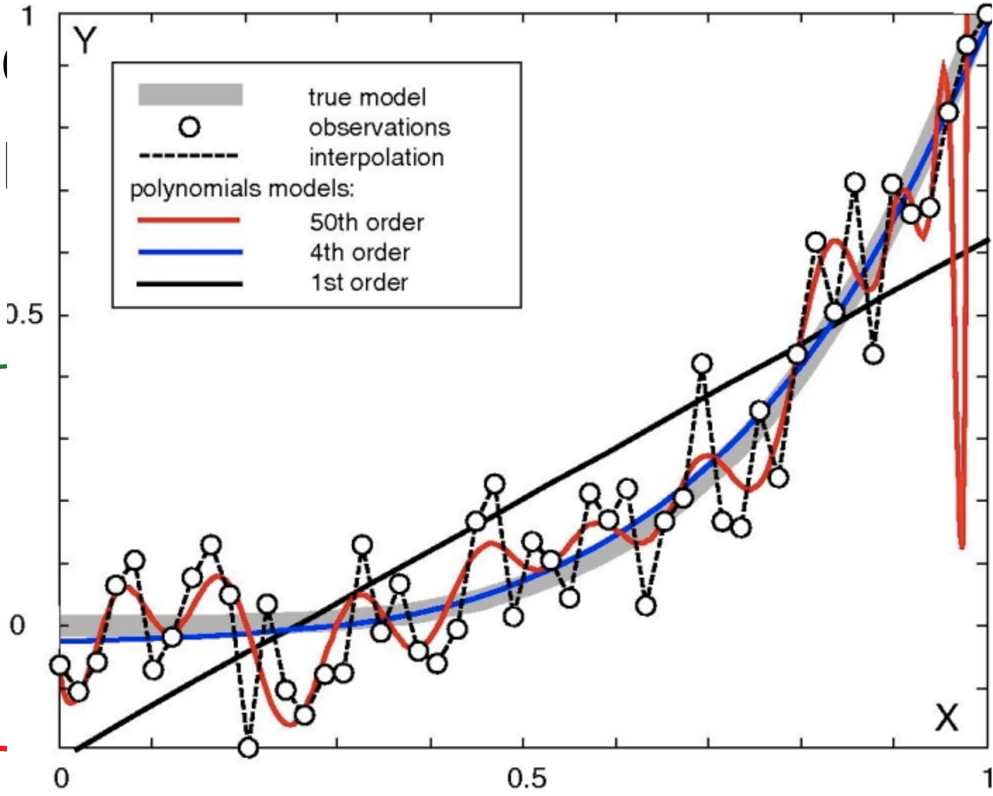
"contains" large set of "good" parameters in "robust" region of θ

"freakish" spike

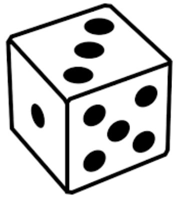
MLE yields a "point estimate" of our parameter

- When we perform MLE, we get just one single estimate of the parameter, $\boldsymbol{\theta}$, rather than a distribution over it which captures uncertainty.
- In Bayesian statistics, we obtain a (p over $\boldsymbol{\theta}$. We will touch more on this i





e.g. MLE for the multinomial distribution



- Consider a six-sided die that we will roll, and we want to know the probability of each side of the die turning up ($\theta = \theta_1 \dots \theta_6$).
- Assume we have observed N rolls, with RV, $X \sim p_\theta(X)$.
- We write that $P(X = k|\theta) = \theta_k$ (when k^{th} side faced up).
- Lets use MLE to estimate these parameters.
- First, since one side must always face up, we know that $1 = \sum_k \theta_k$.
- Second, we can write $P(X = x|\theta) = \theta_x$ (pick off the right parameter).
- Now we write the likelihood:

$n_k \equiv |\{i \mid x_i = k\}|$

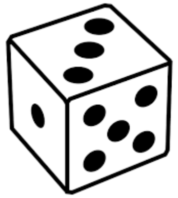
$$P(D|\theta) = p(x_1, \dots x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N}\prod_{k=1}^{6} \theta_k^{I[x_i=k]} = \prod_{k=1}^{6} \theta_k^{\sum_i^N I[x_i=k]} = \prod_{k=1}^{6} \theta_k^{n_k}$$

Now our MLE problem becomes:

constrained optimization

$$\theta_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\theta\in\{\Theta|1=\sum_k \theta_k\}}{\operatorname{argmax}} \sum_{k=1}^{6} \log \theta_k^{n_k}$$

e.g. MLE for the multinomial distribution



Have a constrained optimization problem:

constrained optimization

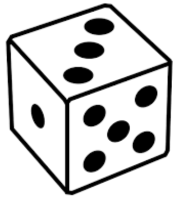
$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\theta \in \{\Theta | 1 = \sum_k \theta_k\}}{\operatorname{argmax}} \sum_{k=1}^{6} \log \theta_k^{n_k}$$

What is one technique you should have learned in first year calculus to solve this?

The technique of Lagrange multipliers:

$J(\theta, \lambda) = \log p(D|\theta) + \lambda(1 - \sum_k \theta_k)$ (look for stationary points wrt θ, λ)

e.g. MLE for the multinomial distribution



$$J(\theta,\lambda) = log\, p(D|\theta) + \lambda(1 - \sum_k \theta_k) = \sum_{k=1}^{6} \log \theta_k^{n_k} + \lambda(1 - \sum_k \theta_k)$$

1. $\frac{\partial J}{\partial \lambda} = 0 \Rightarrow 1 = \sum_k \theta_k$ (we just get the constraint back)
2. $\frac{\partial J}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{k=1}^{6} \log \theta_k^{n_k} - \frac{\partial}{\partial \theta_k} \lambda \theta_k = \frac{n_k}{\theta_k} - \lambda = 0 \Rightarrow \theta_k = \frac{n_k}{\lambda}$.
3. Lets plug this into 1), $1 = \sum_k \theta_k = \sum_k \frac{n_k}{\lambda} \Rightarrow \lambda = \sum_k n_k = N$.
4. All together then, $\theta_k = \frac{n_k}{N}$.

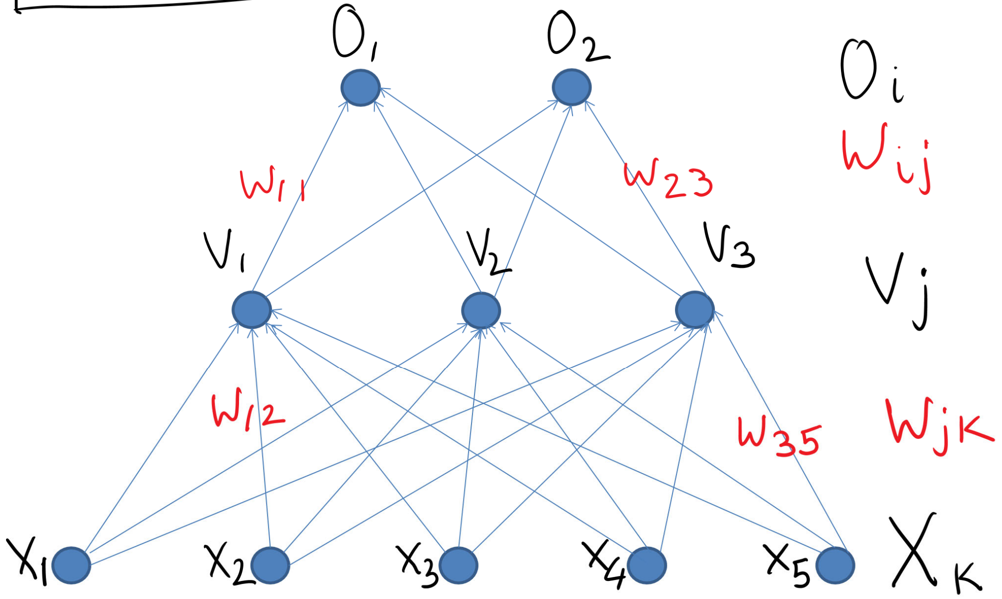
Doing MLE requires optimization

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta)$$

- For Gaussian, multinomial (and more), the MLE can be obtained in closed form by setting the derivative to zero.
- What if we had a model such as Prof. Malik mentioned in the first lecture?

$$\boxed{V_j = g\left(\sum_k w_{jk} x_k\right) \; ; \; O_i = g\left(\sum w_{ij} V_j\right)}$$

- Here, we need *iterative optimization* (can take entire classes on special cases of this (e.g. Convex Optimization). More later.



Prof. Malik in first lecture:

- Mentioned that a good loss to estimate parameters is the *cross-entropy* (rather than the likelihood).
- So why are we teaching you MLE?! They are equivalent.

Training a single layer neural network

- A good choice of loss function is the cross entropy

$$L = -\sum_{\text{input data}} \left(y_i \ln O_i + (1-y_i) \ln(1-O_i) \right)$$

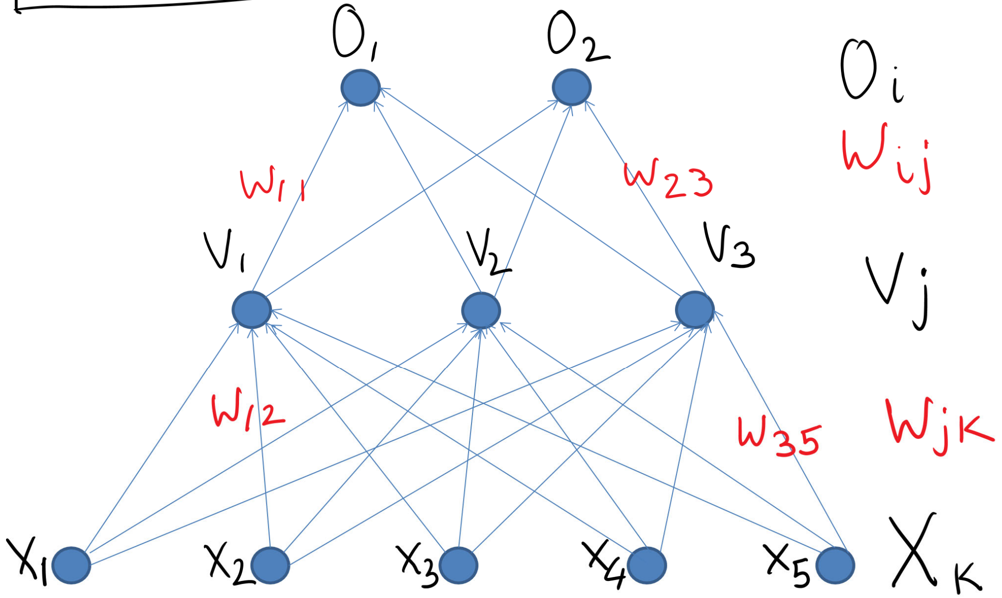
- We model the activation function *g* as a sigmoid

$$g(z) = \frac{1}{1+\exp(-z)}$$

- Finding w reduces to logistic regression!

We can use STOCHASTIC GRADIENT DESCENT.

$$\boxed{V_j = g\left(\sum_K w_{jK} x_K\right) ; \quad O_i = g\left(\sum w_{ij} V_j\right)}$$



Relationship between likelihood, cross-entropy, *etc.*

- The *cross-entropy* is a term from *information* theory.
- To understand the connection between MLE and maximizing the cross-entropy, we need to know some concepts from information theory:
  1. Entropy
  2. Cross-entropy
  3. KL-divergence (relative entropy).

Entropy: a measure of expected surprise

Think about a flipping a coin once, and how surprised you would be at observing a head.

$p(head) = 0$

$p(head) = 0.5$

$p(head) = 0.01$

$p(head) = 1$

Entropy: a measure of expected surprise

- The "surprise" of observing that a discrete random variable $\boldsymbol{Y}$ takes on value $\boldsymbol{k}$ is:

$$log\,\frac{1}{P(Y=k)} = -\log(P(Y=k))$$

- As $P(Y=k) \rightarrow \mathbf{0}$, the surprise of observing $\boldsymbol{k}$ approaches ∞.
- As $P(Y=k) \rightarrow \mathbf{1}$, the surprise of observing $\boldsymbol{k}$ approaches $\mathbf{0}$.
- The entropy of the distribution of $\boldsymbol{Y}$ is the *expected surprise*:

$$H(Y) \equiv E_Y[-\log P(Y=k)] = \sum_k P(Y=k)\log P(Y=k)$$

Entropy example: flipping a coin

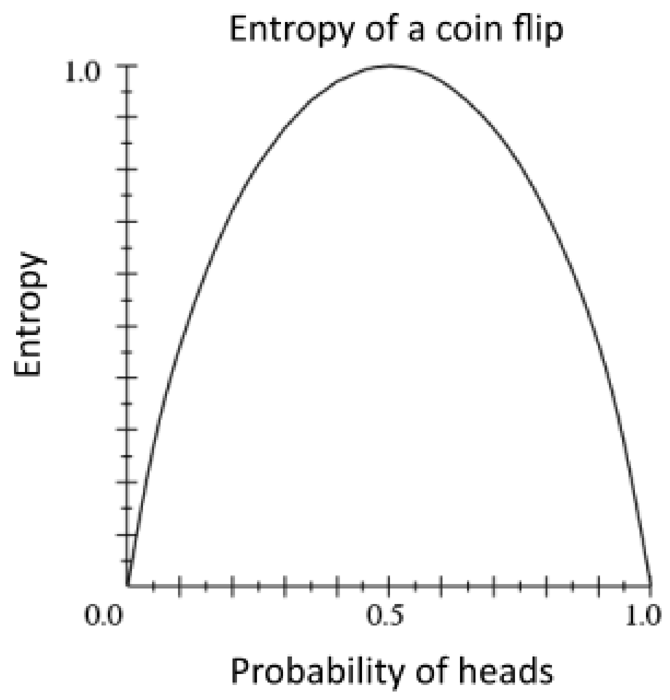
$$H(Y) = -\sum_{i=1}^{\kappa} P(Y = y_i) \log_2 P(Y = y_i)$$

P(Y=t) = 5/6

P(Y=f) = 1/6

$H(Y) = -\ 5/6 \log_2 5/6 - 1/6 \log_2 1/6$

$= 0.65$



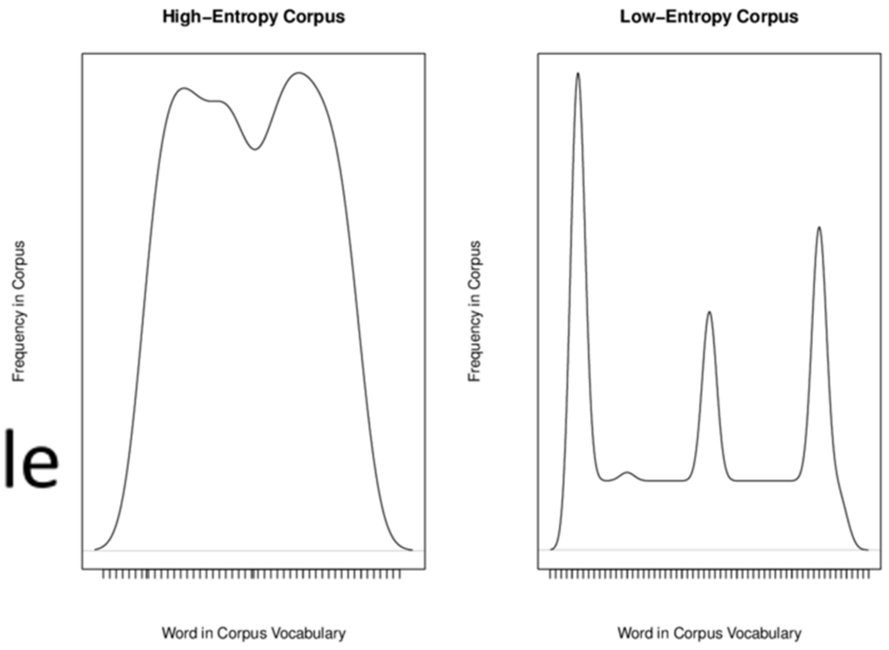
Entropy of a random variable Y:

"High Entropy"

- Y is from a uniform like distribution
- Flat histogram
- Values sampled from it are less predictable

"Low Entropy"

- Y is from a varied (peaks and valleys) distribution
- Histogram has many lows and highs
- Values sampled from it are more predictable



https://www.researchgate.net/figure/Hypothetical-distributions-of-term-frequency-in-high-and-low-entropy-corpora_fig1_305417514

(Slide from Vibhav Gogate)

From Entropy to *Relative Entropy*

- Also called the Kullback-Leibler (KL) Divergence.
- Measures how much one distribution diverges from another.
- For discrete probability distributions, P and Q, it is defined as:

$$D_{KL}(P||Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

- Not a true distance metric because not symmetric in P and Q:

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

Properties of KL Divergence

- $\mathrm{KL}(p||q) \geq 0$
- $\mathrm{KL}(p||q) = 0$ if and only if $p = q$

https://www.cs.ox.ac.uk/people/varun.kanade/teaching/ML-MT2016/slides/slides03.pdf

From Relative Entropy to Cross-Entropy (then to MLE!)

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

$$= E_{P(x)}\left[log \frac{1}{Q(X)}\right] - E_{P(x)}\left[log \frac{1}{P(X)}\right]$$

From Relative Entropy to Cross-Entropy (then to MLE!)

$$D_{KL}(P||Q) = \sum_x P(x)log\frac{P(x)}{Q(x)}$$

$$= E_{P(x)}\left[log\frac{1}{Q(X)}\right] - E_{P(x)}\left[log\frac{1}{P(X)}\right]$$

$$= \underbrace{H(P,Q)}_{\text{cross-entropy}} - \underbrace{H(P)}_{\text{entropy}}$$

- Consider data, D where $x_{\mathrm{i}} \sim \hat{p}_{data}$ and a model with params $\boldsymbol{\theta}$, $\boldsymbol{p(x|\theta)}$.
- If minimizing the KL divergence (instead of MLE),

$argmin_{\theta} D_{KL}(\hat{p}_{data}||p(x|\theta)) =)) =$

From Relative Entropy to Cross-Entropy (then to MLE!)

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

$$= E_{P(x)}\left[log \frac{1}{Q(X)}\right] - E_{P(x)}\left[log \frac{1}{P(X)}\right]$$

$$= \underbrace{H(P,Q)}_{\text{cross-entropy}} - \underbrace{H(P)}_{\text{entropy}}$$

no dependence on θ

- Consider data, D where $x_i \sim \hat{p}_{data}$ and a model with params θ, $p(x|\theta)$.
- If minimizing the KL divergence (instead of MLE),

$$argmin_\theta D_{KL}(\hat{p}_{data}||p(x|\theta)) =)) = argmin_\theta\, H(\hat{p}_{data}, p(x|\theta)) + H(\hat{p}_{data})$$

$$= argmax E_{\hat{p}_{data}}[\log p(x|\theta))]$$

negative cross-entropy

From Relative Entropy to Cross-Entropy (then to MLE!)

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

$$= E_{P(x)}\left[log \frac{1}{Q(X)}\right] - E_{P(x)}\left[log \frac{1}{P(X)}\right]$$

$$= \underbrace{H(P,Q)}_{\text{cross-entropy}} - \underbrace{H(P)}_{\text{entropy}}$$

MLE problem

- Consider data, D where $x_i \sim \hat{p}_{data}$ and a model with params θ, $p(x|\theta)$.
- If minimizing the KL divergence (instead of MLE),

$$argmin_\theta D_{KL}(\hat{p}_{data}||p(x|\theta)) =)) = argmin_\theta \, H(\hat{p}_{data}, p(x|\theta)) + H(\hat{p}_{data})$$

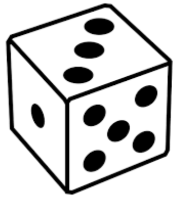
$$= argmax E_{\hat{p}_{data}}[\log p(x|\theta)] = argmax \sum_i^N \log p(x_i|\theta).$$

From Relative Entropy to Cross-Entropy (then to MLE!)

- Performing MLE maximizes the likelihood function.
- This is equivalent to maximizing the cross-entropy.
- And equivalent to minimizing the KL-divergence (aka relative entropy).

Extra

e.g. MLE for the multinomial distribution



$J(\theta,\lambda) = \log p(D|\theta) + \lambda(1 - \sum_k \theta_k)$ (look for stationary points wrt θ, λ)

1. $\frac{\partial J}{\partial \lambda} = 0 \to 1 = \sum_k \theta_k$ (we just get the constraint back)
2. $\frac{\partial J}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{k=1}^{6} \log \theta_k^{n_k} - \frac{\partial}{\partial \theta_k} \lambda \theta_k = \frac{n_k}{\theta_k} - \lambda = 0 \to \theta_k = \frac{n_k}{\lambda}$.
3. Lets plug this into 1: $\sum_k \theta_k = 1 = \sum_k \frac{n_k}{\lambda} \to \lambda = \sum_k n_k = N$.
4. All together then, $\theta_k = \frac{n_k}{N}$.

This is a stationary point. But is it a maximum? Could check Hessian, but lets instead consider our know equivalence

$$D_{KL}(p_{data}||p(x|\theta)) = \sum_{k=1}^{6} P_{data}(X = k) \log \frac{P_{data}(X=k)}{P(X=k|\theta)}$$

n_k/N

$\Rightarrow \theta_k = \frac{n_k}{N}$

$\log(1) = 0$

$\Rightarrow D_{KL} = 0$

but $D_{KL} \geq 0$

so we have minized the D_{KL} & maximized log likelihood!