Neural Networks

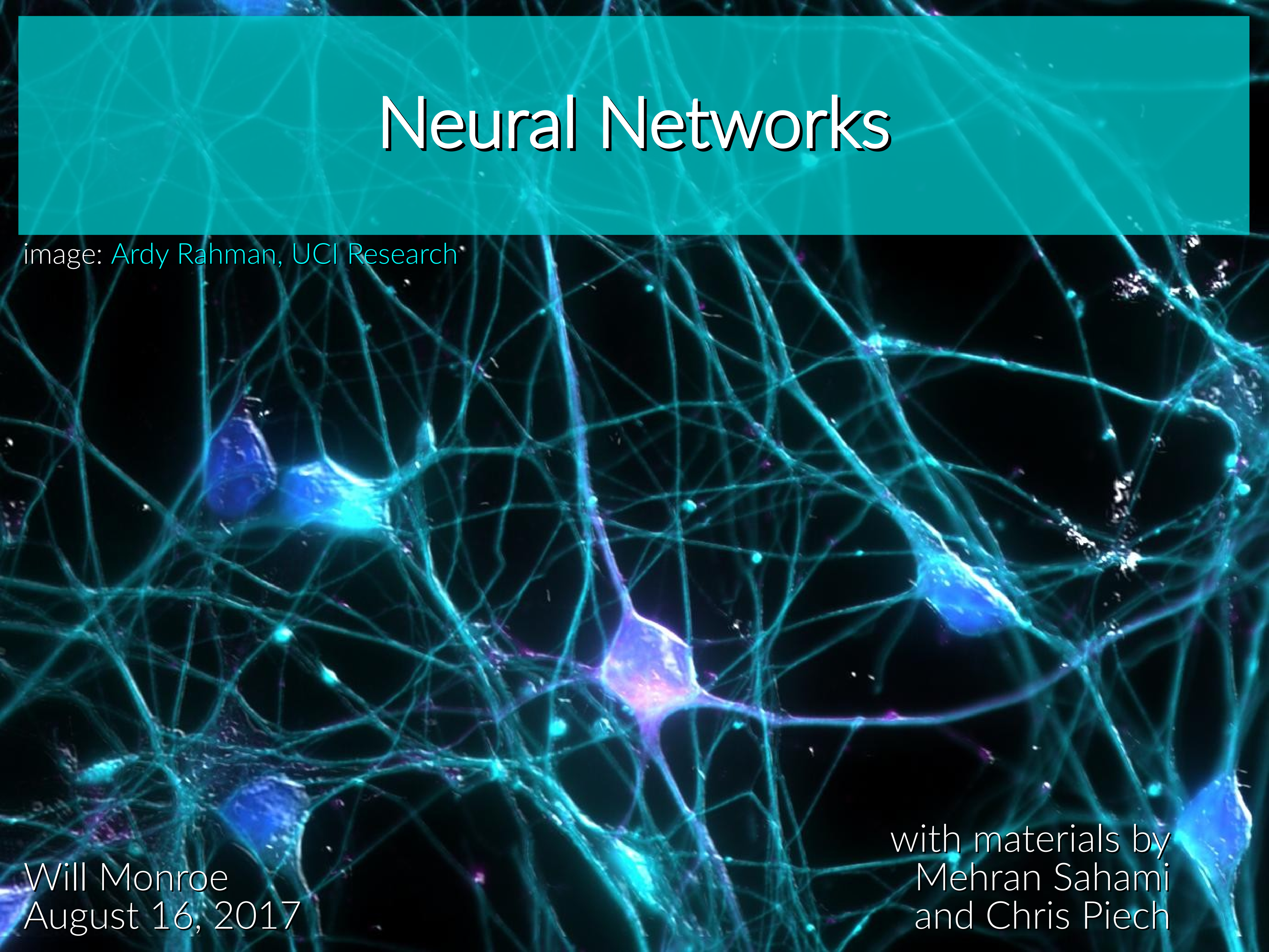


image: Ardy Rahman, UCI Research

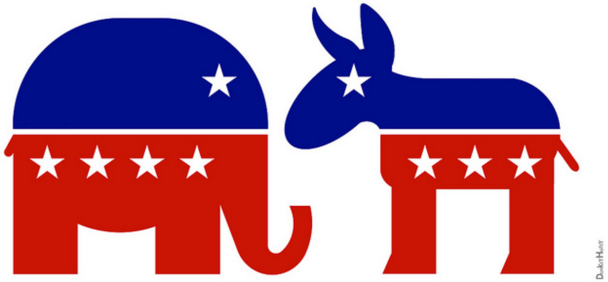
Will Monroe
August 16, 2017

with materials by
Mehran Sahami
and Chris Piech

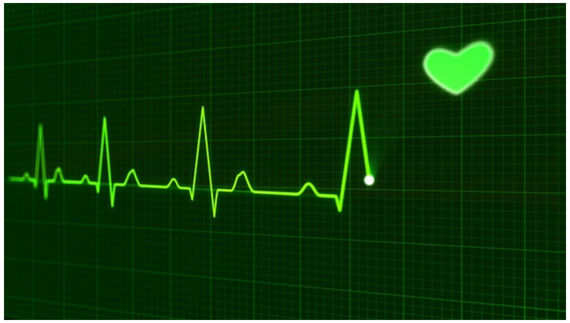
Announcement: Problem Set #6

Due **today!**

That's all, folks!



Congressional voting



Heart disease diagnosis

Announcements: Final exam

This Saturday, August 19, 12:15-3:15pm
in NVIDIA Auditorium
(pending maintenance)

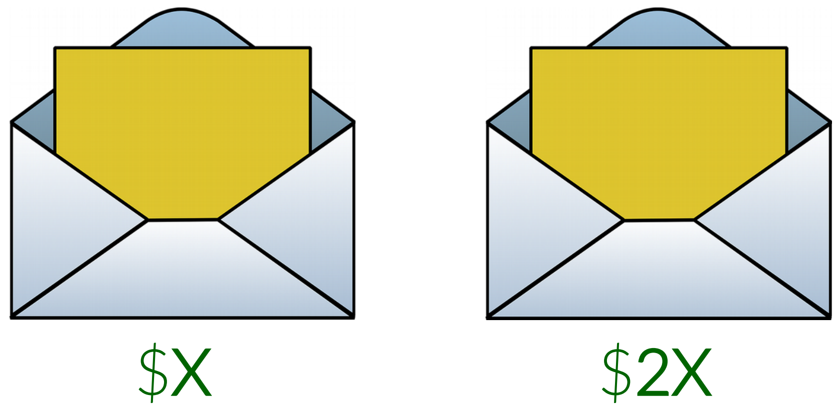
Two pages (both sides) of notes

All material in the class through Monday

Review session:
Today after lecture, 2:30-3:20 in **Huang 18**

Two envelopes: A resolution

"I'm trying to think: how likely is it that you would have put \$40 in an envelope?



Y = y: amount in envelope chosen

$$E[W|Y=y,\text{stay}]=y$$

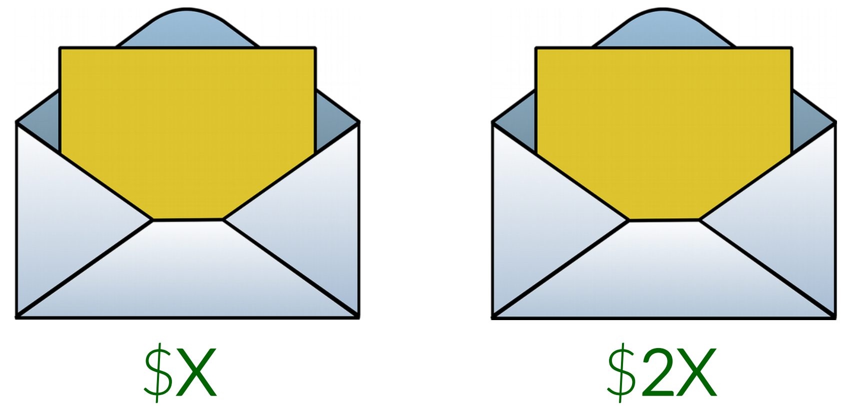
not necessarily 0.5!

$$E[W|Y=y,\text{switch}]=\frac{y}{2}\cdot\boxed{P\left(X=\frac{y}{2}|Y=y\right)}+2y\cdot\boxed{P(X=y|Y=y)}$$

$$P(X=y|Y=y)=\frac{P(X=y)}{P(X=y)+P(X=y/2)}$$

Two envelopes: A resolution

"I'm trying to think: how likely is it that you would have put \$40 in an envelope?



Y = y: amount in envelope chosen

$$E[W|Y=y,\text{stay}]=y$$

not necessarily 0.5!

$$E[W|Y=y,\text{switch}]=\frac{y}{2}\cdot\boxed{P\left(X=\frac{y}{2}|Y=y\right)}+2y\cdot\boxed{P(X=y|Y=y)}$$

prior: if all equally likely, then this will be 0.5

$$P(X=y|Y=y)=\frac{\boxed{P(X=y)}}{P(X=y)+P(X=y/2)}$$

$$P(X=y)=C?$$

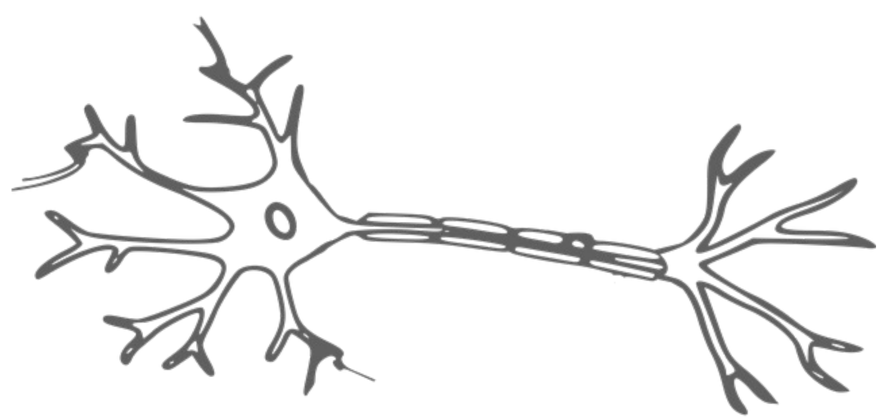
$$\sum_y P(X=y)=\sum_y C=1$$

$$\infty\cdot C=1???$$

Logistic regression

A classification algorithm using the assumption that **log odds** are a linear function of the features.

$$\hat{y} = \frac{1}{1+e^{-\vec{\theta}^T \vec{x}}}$$



Review: The logistic function

$$z = \log o_f$$

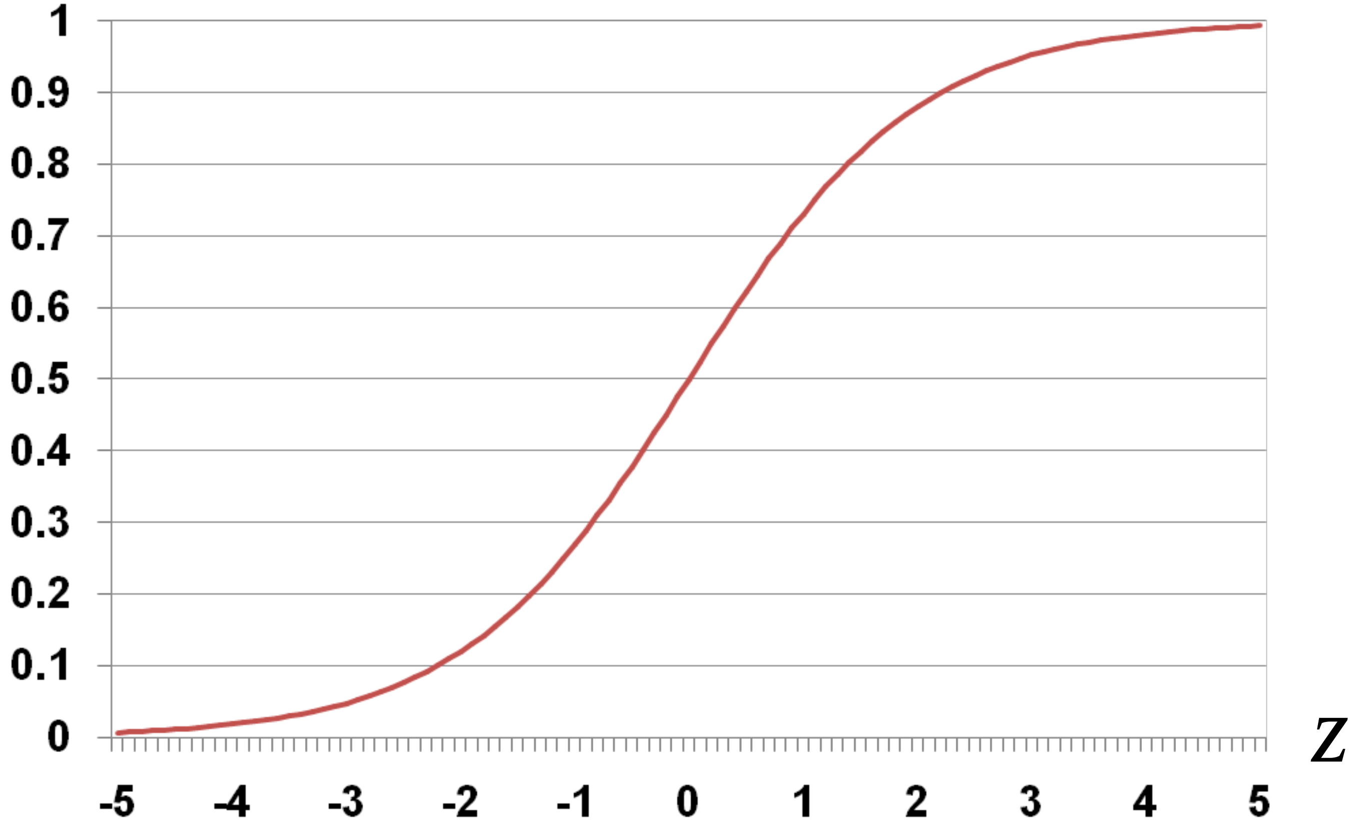
$$p = \frac{o_f}{o_f + 1} = \frac{1}{1 + \frac{1}{o_f}}$$

$$= \frac{1}{1 + e^{-\log(o_f)}}$$

$$= \frac{1}{1 + e^{-z}}$$

$$= \sigma(z)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Review: Logistic regression assumption

$$P(Y=1|\vec{X}=\vec{x})=\sigma(\vec{\theta}^T\vec{x})=\frac{1}{1+e^{-\vec{\theta}^T\vec{x}}}$$

or in other words:

$$p=\sigma(z)$$

$$z=\log o_f$$

$$\vec{\theta}^T\vec{x}=\log o_f(Y=1|\vec{X}=\vec{x})$$

$$\vec{\theta}^T\vec{x}=\vec{\theta}\cdot\vec{x}=\theta_0\cdot 1+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m$$

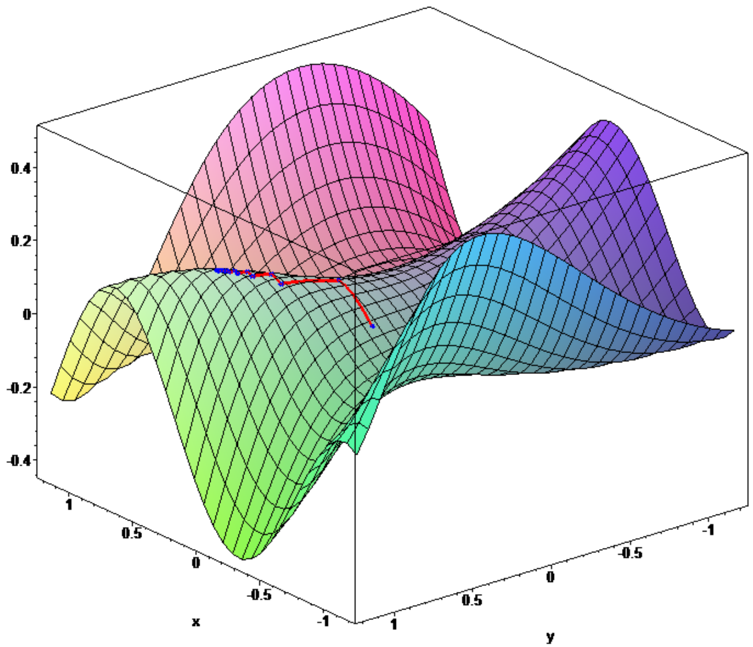
$$=\sum_{i=0}^{m}\theta_i x_i$$

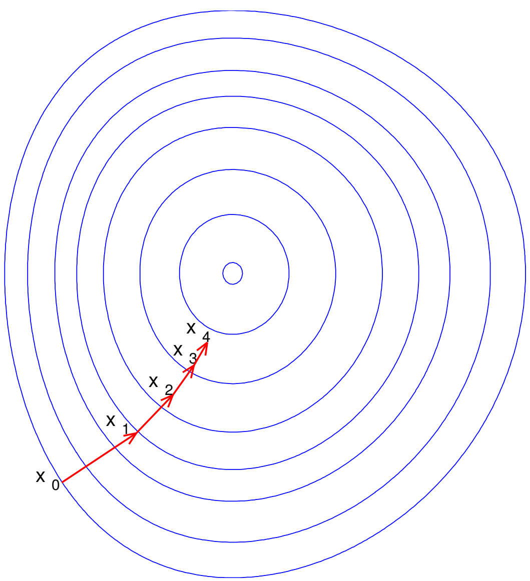
$$(x_0=1)$$

Review: Gradient ascent

An algorithm for computing an **arg max** by taking small steps **uphill** (i.e., in the direction of the **gradient** of the function).

$$\vec{\theta} \leftarrow \vec{\theta} + \eta \cdot \nabla_{\vec{\theta}} f(\vec{\theta})$$





Review: Logistic regression algorithm

```
initialize: θ = [0, 0, …, 0]   (m elements)

repeat many times:
    gradient = [0, 0, …, 0]   (m elements)

    for each training example (x^(i), y^(i)):
        for j = 0 to m:
```

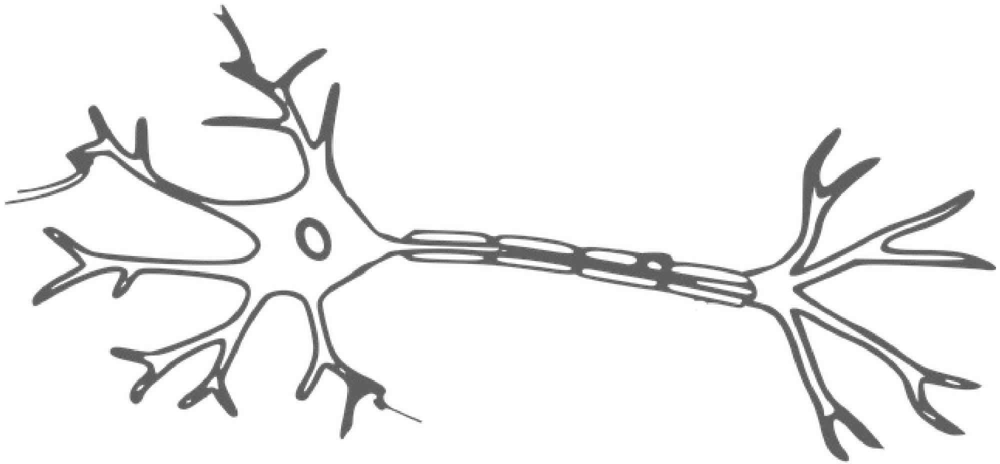
$$\text{gradient[j]} \mathrel{+}= [y^{(i)} - \sigma(\vec{\theta}^T \vec{x}^{(i)})]\vec{x}_j^{(i)}$$

```
    for j = 0 to m:
        θ[j] += η * gradient[j]

return θ
```

Your brain on logistic regression

$p = \sigma(z)$

$z = \vec{\theta}^T \vec{x}^{(i)}$



dendrites:
take a weighted sum
of incoming stimuli
with electric potential

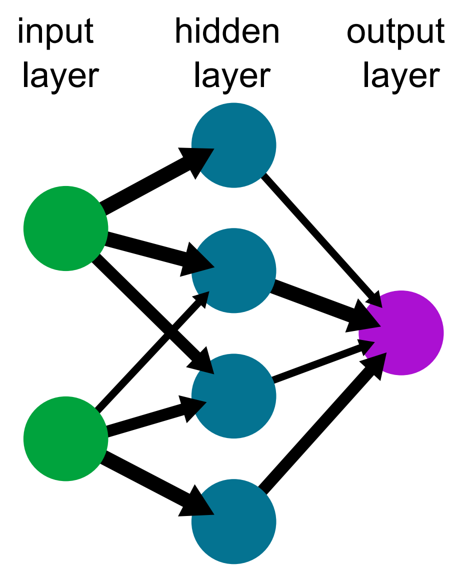
axon:
carries outgoing
pulse if potential
exceeds a threshold

Caution: Just a (greatly simplified) model! All models are wrong—but some are useful…

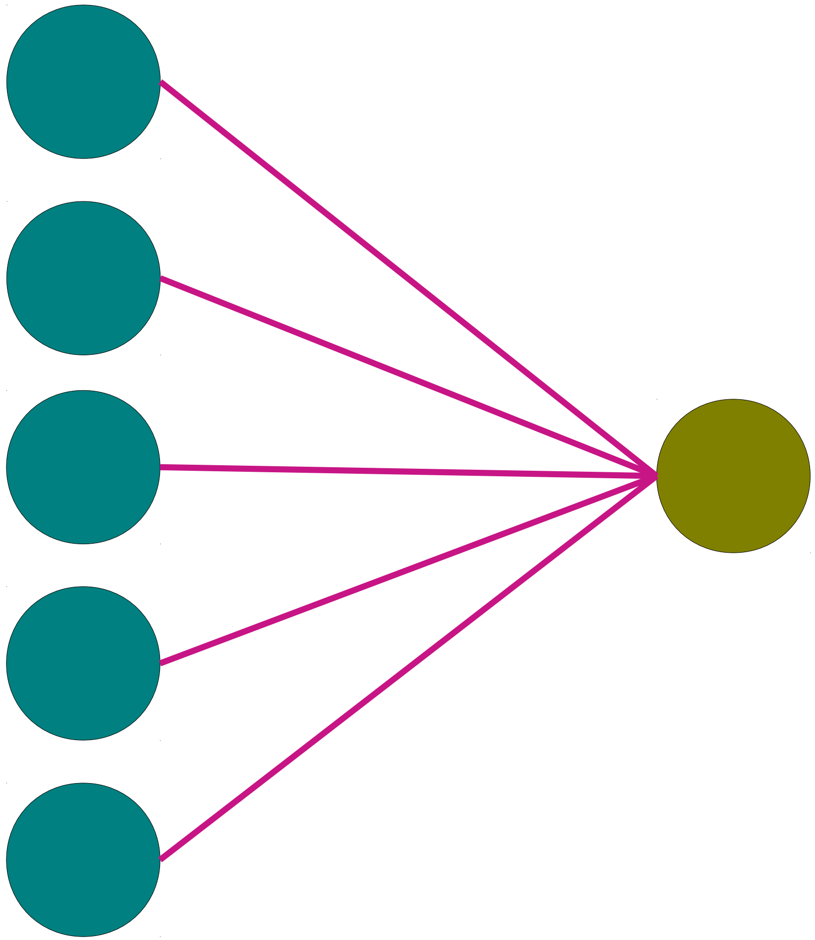
Feedforward neural network

An algorithm for classification or regression that uses **layers of logistic regressions** to **discover its own features.**

$$\hat{y} = \sigma\left(\theta^{(\hat{y})}\sigma\left(\theta^{(h)}\vec{x}\right)\right)$$



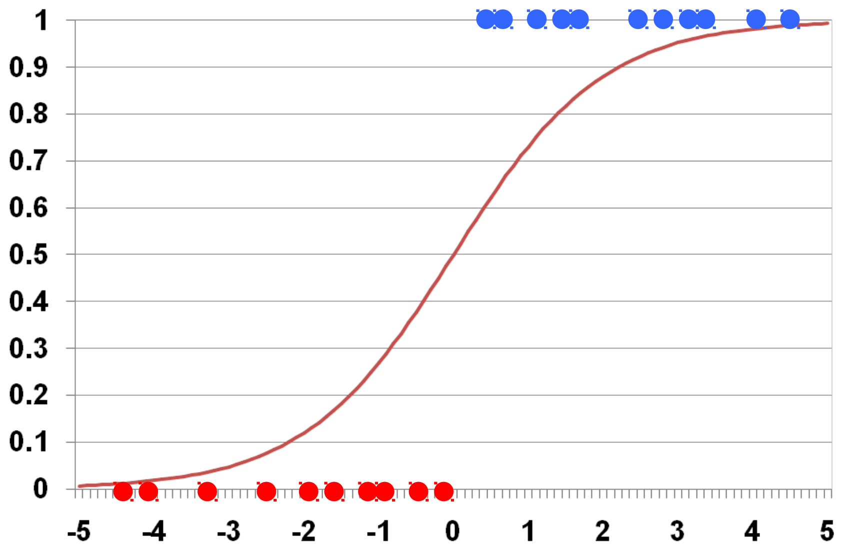
A cartoon of logistic regression

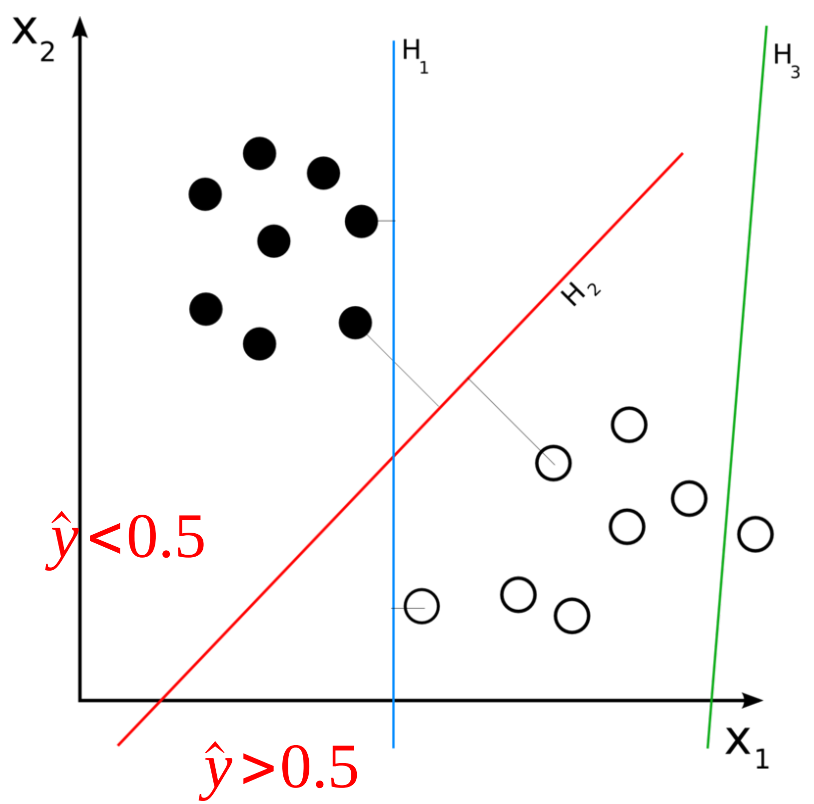


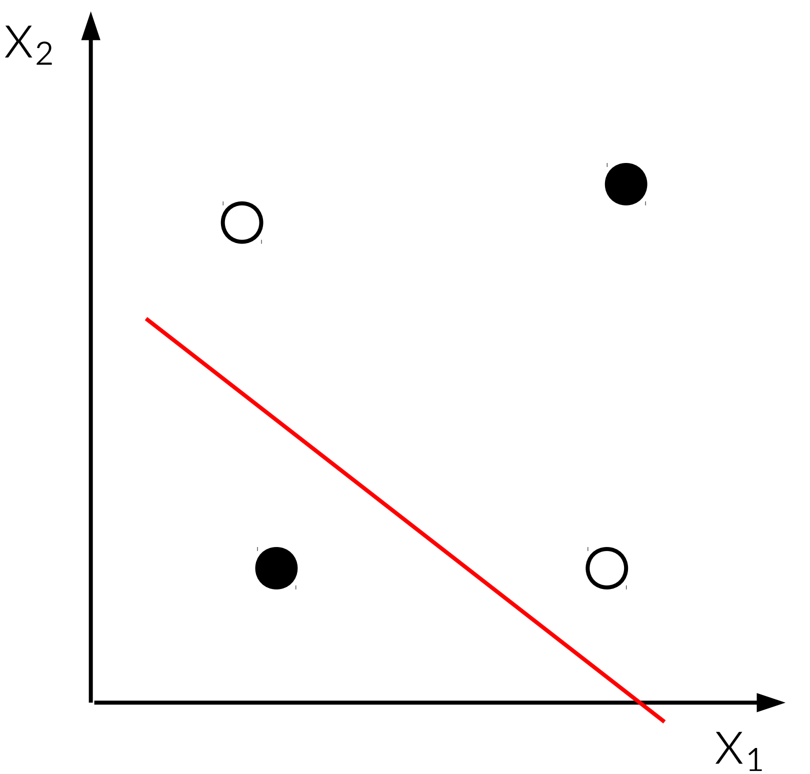
$\vec{x}$

$$\hat{y} = \sigma\left(\vec{\theta}^T \vec{x}\right)$$

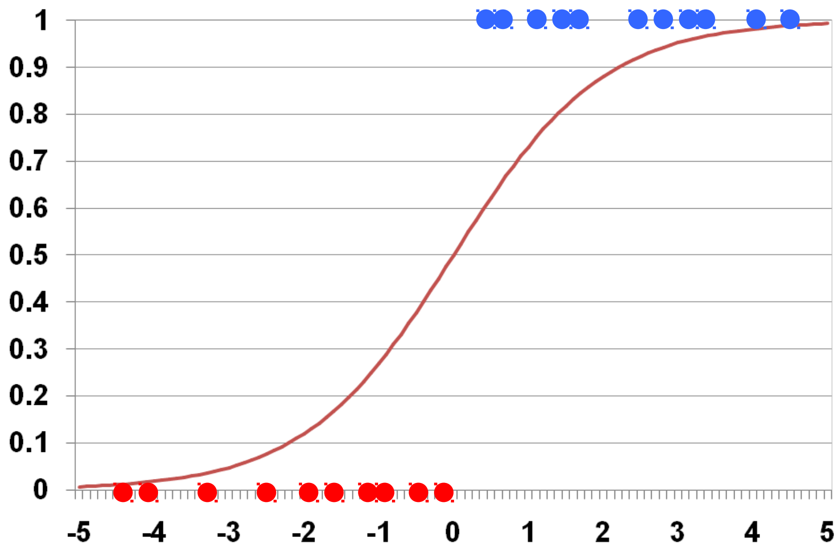
Logistic regression is linear

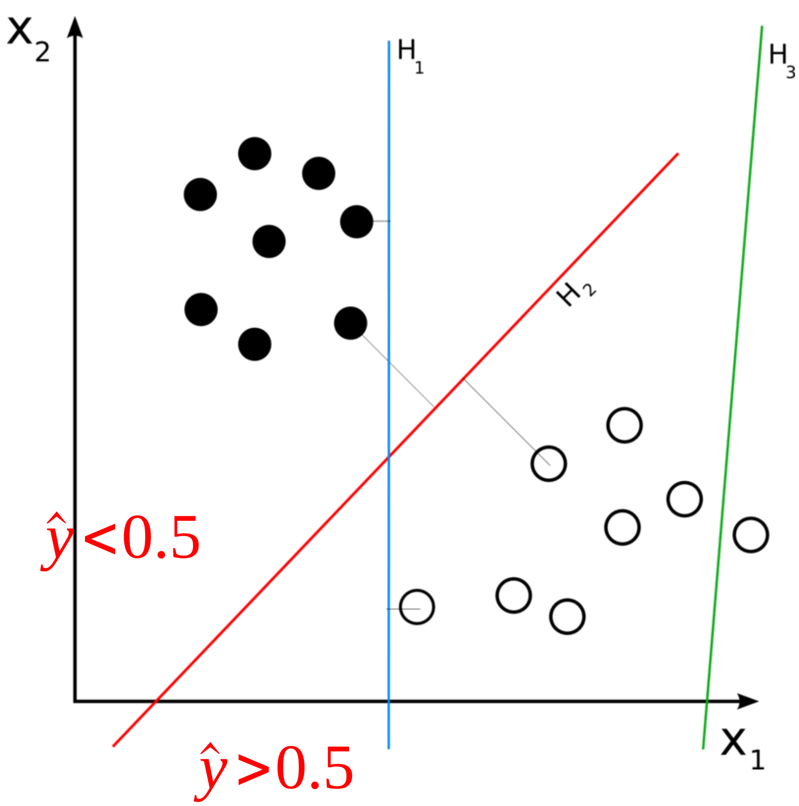


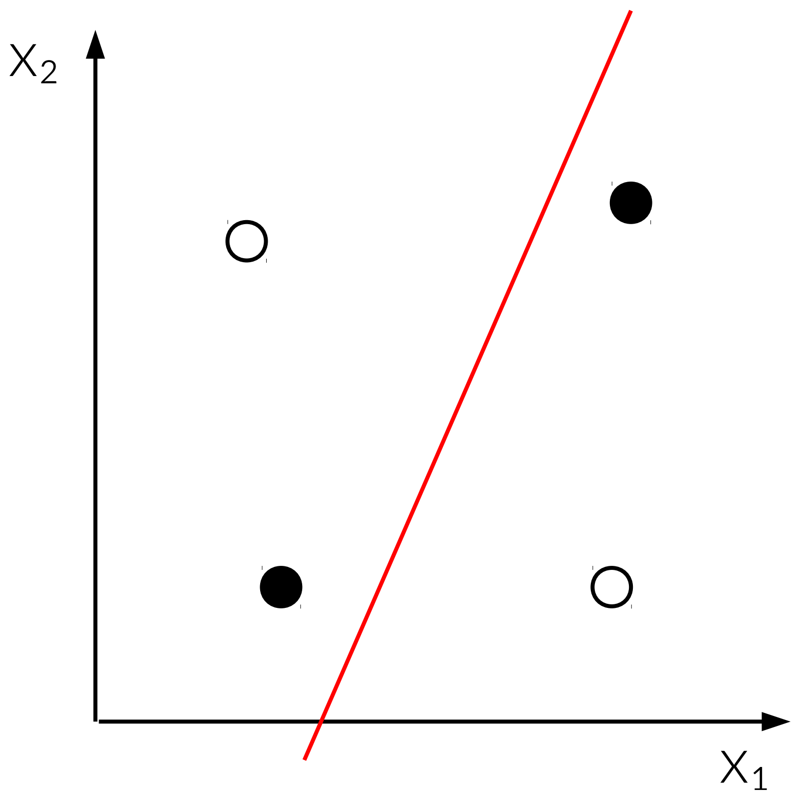




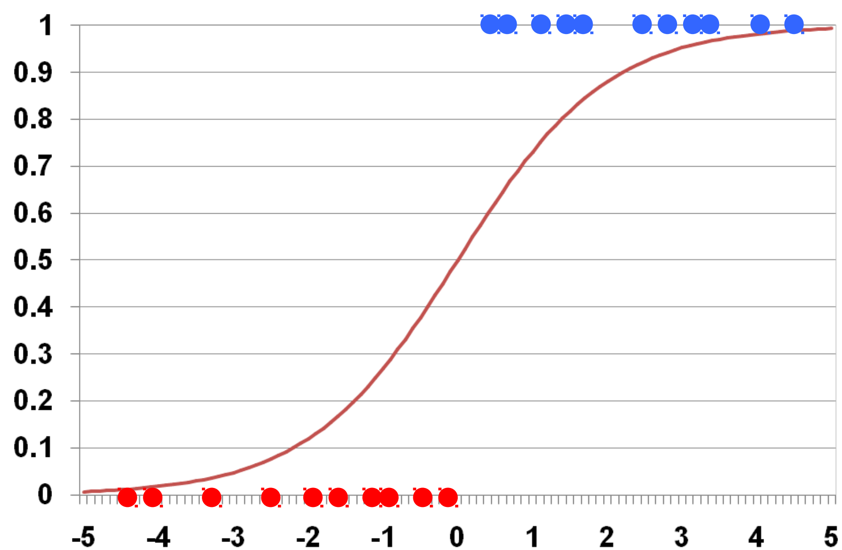
Logistic regression is linear







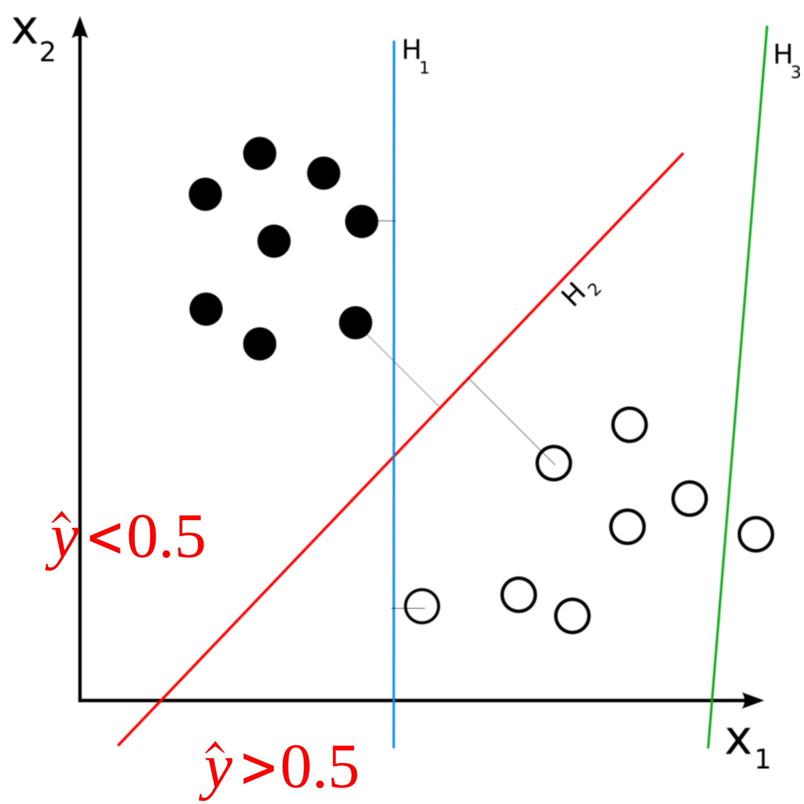
Logistic regression is linear

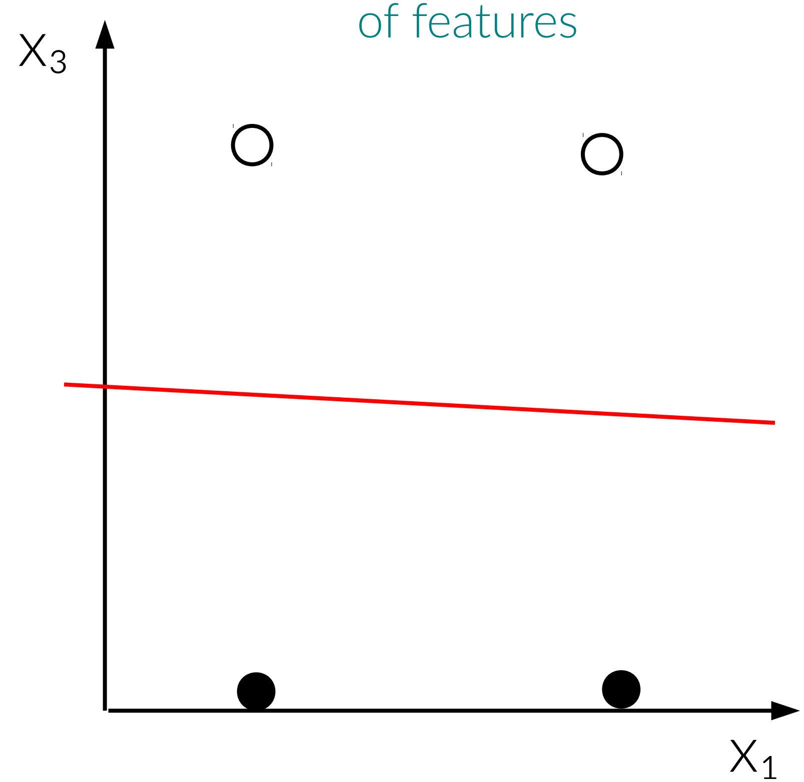


non-linear function

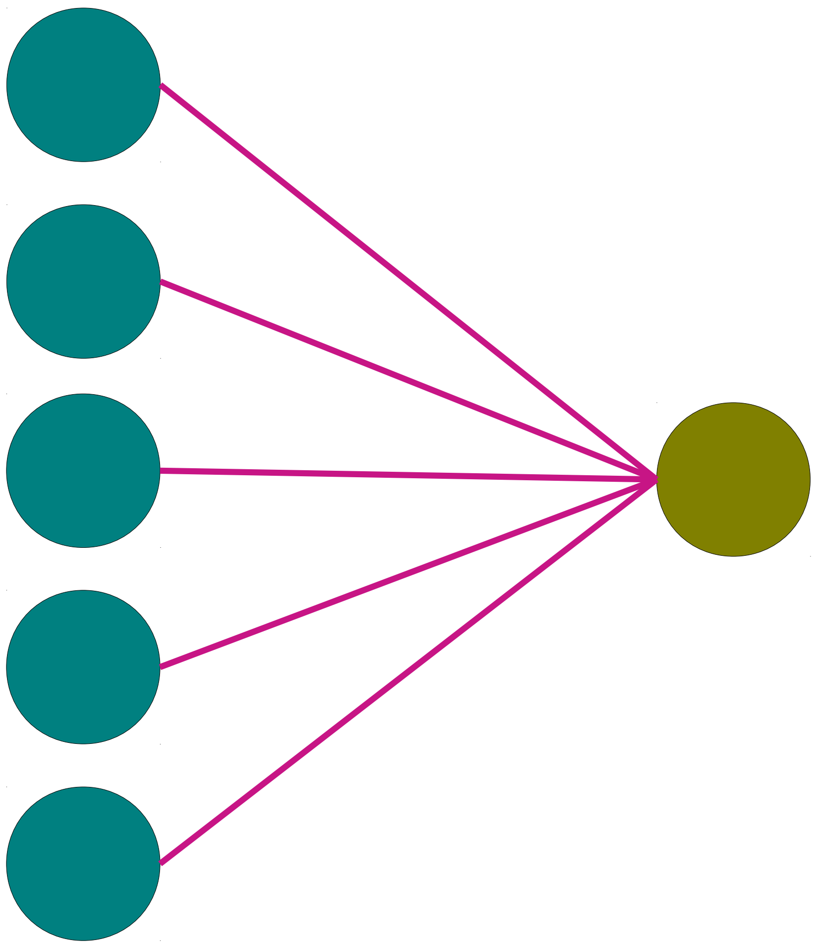
$$X_3 = \boxed{(\boxed{X_1 - X_2})^2}$$

linear combination of features





A cartoon of logistic regression



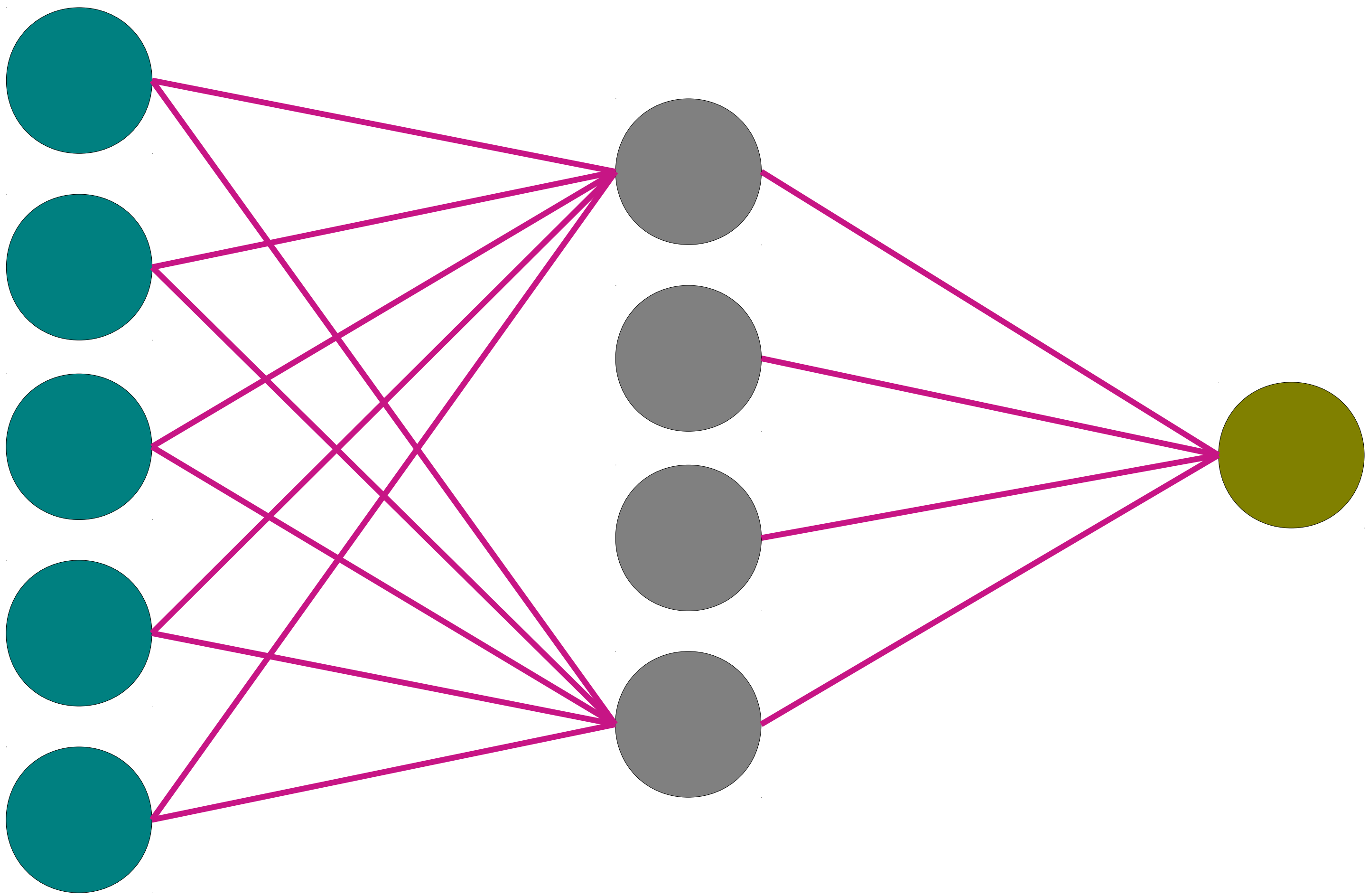
$\vec{x}$

non-linear function

$$\hat{y} = \boxed{\sigma\left(\boxed{\vec{\theta}^{T}\vec{x}}\right)}$$

linear
combination
of features

Stacking logistic regression

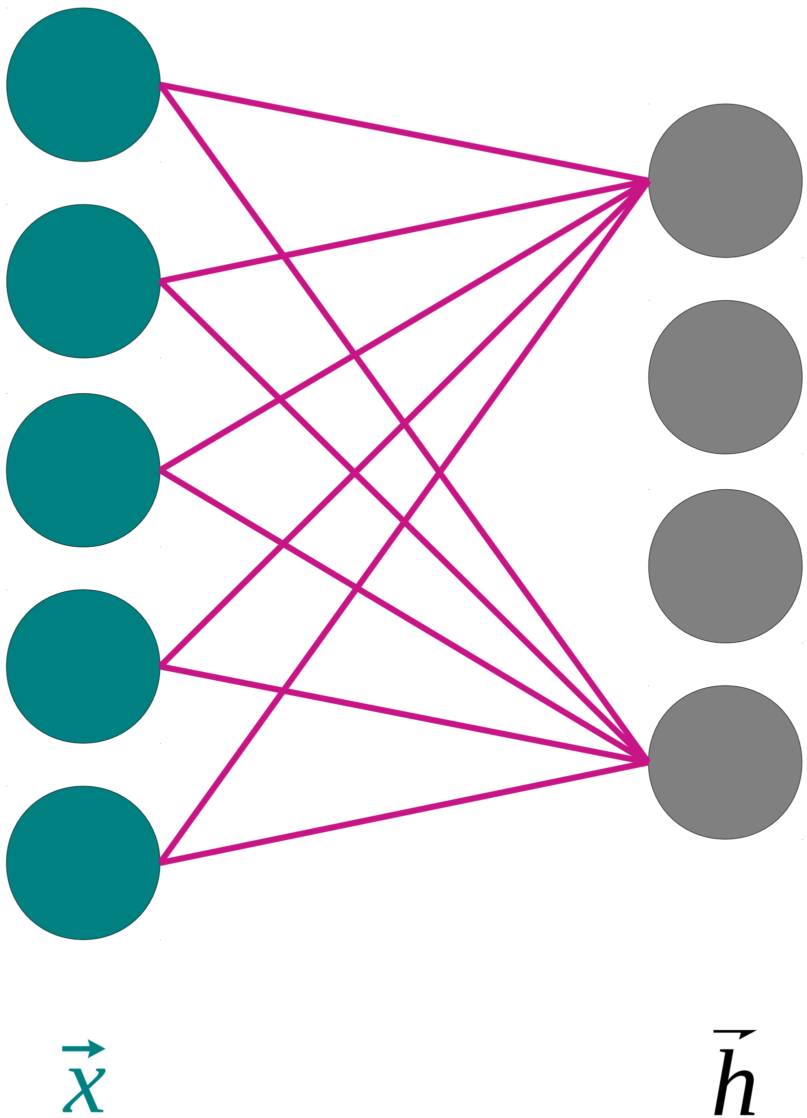


$\vec{x}$

$$\vec{h} = \sigma\left(\theta^{(h)} \vec{x}\right)$$

$$\hat{y} = \sigma\left(\vec{\theta}^{(\hat{y})T} \vec{h}\right)$$

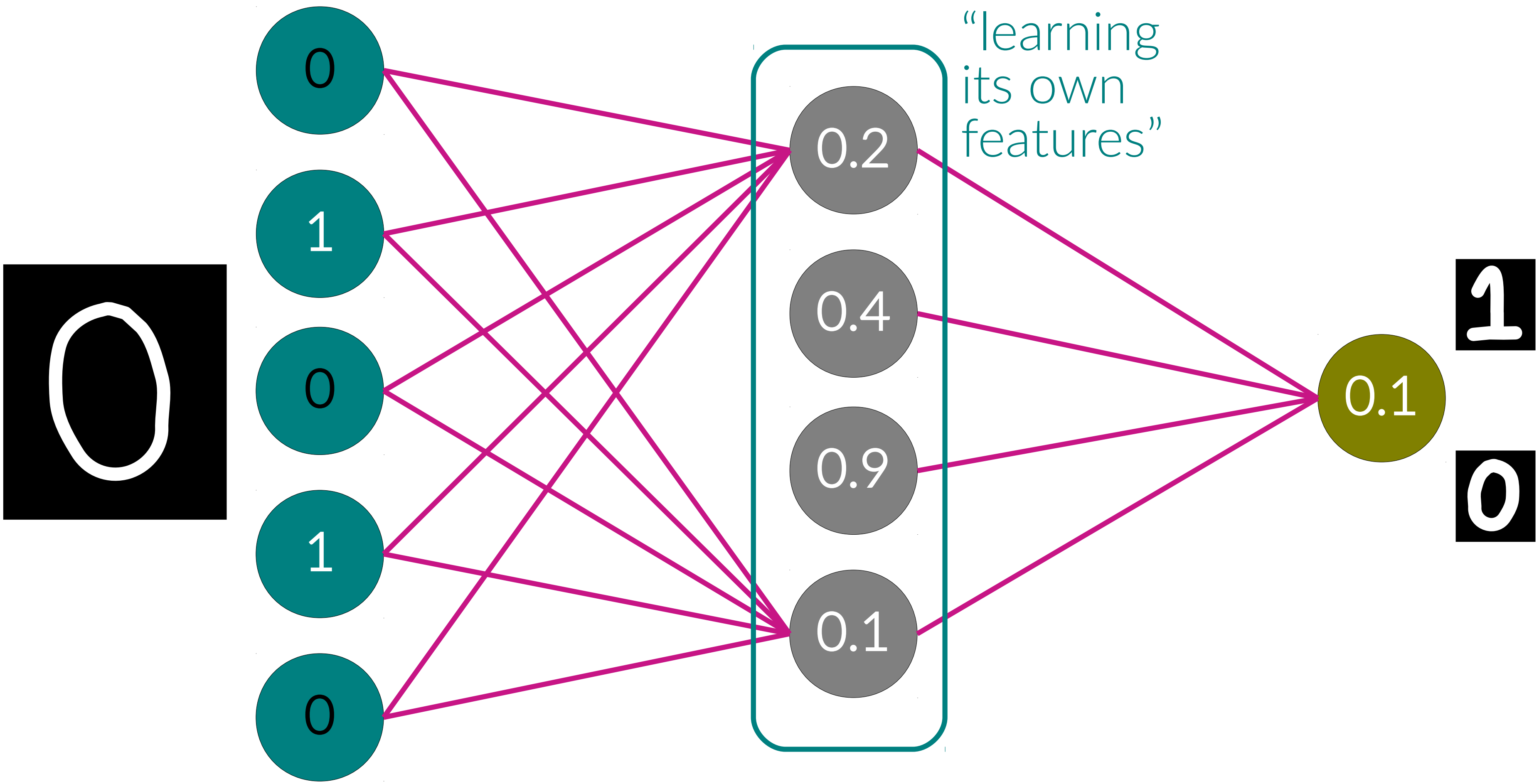
Unpacking the linear algebra



$$\vec{h} = \underbrace{\sigma\Big(\underbrace{\overset{\text{matrix}}{\theta^{(h)}}\,\overset{\text{vector}}{\vec{x}}}_{\text{vector}}\Big)}_{\text{vector}}$$

$$h_i = \sigma\left(\sum_{j=0}^{m} \theta^{(h)}_{i,j}\, x_j\right)$$

Stacking logistic regression



$\vec{x}$

input features

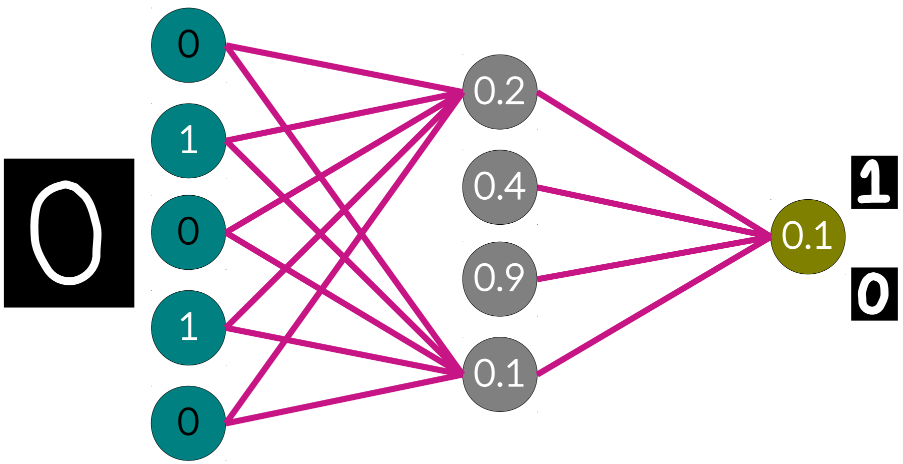
$\vec{h} = \sigma\left(\theta^{(h)}\vec{x}\right)$

hidden representation

$\hat{y} = \sigma\left(\vec{\theta}^{(\hat{y})T}\vec{h}\right)$

output class label

Maximum likelihood with neural nets



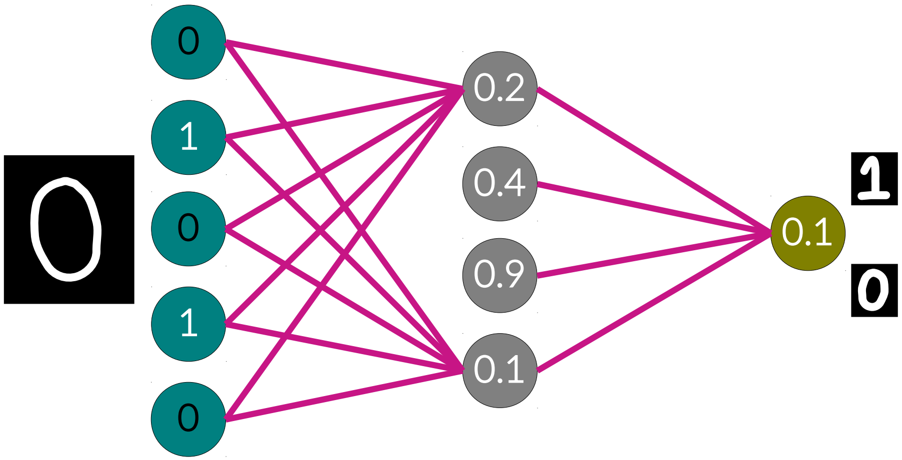
$\vec{x}$ $\qquad \vec{h} = \sigma\left(\theta^{(h)} \vec{x}\right) \qquad \hat{y} = \sigma\left(\vec{\theta}^{(\hat{y})T} \vec{h}\right)$

$$L(\theta) = P\left(X^{(1)}, \ldots, X^{(n)}, Y^{(1)}, \ldots, Y^{(n)} \middle| \theta\right)$$

$$= \prod_{i=1}^{n} P\left(X^{(i)}, Y^{(i)} \middle| \theta\right)$$

$$= \prod_{i=1}^{n} \hat{y}^{y^{(i)}} (1 - \hat{y})^{1 - y^{(i)}} P\left(X^{(i)}\right)$$

Maximum likelihood with neural nets



$\vec{x}$ $\qquad \vec{h} = \sigma\left(\theta^{(h)} \vec{x}\right) \qquad \hat{y} = \sigma\left(\vec{\theta}^{(\hat{y})T} \vec{h}\right)$

$$L(\theta) = \prod_{i=1}^{n} \hat{y}^{y^{(i)}} (1 - \hat{y})^{1 - y^{(i)}} P\left(X^{(i)}\right)$$

$$LL(\theta) = \sum_{i=1}^{n} \left[y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) + \log P\left(X^{(i)}\right) \right]$$

Maximum likelihood with neural nets

$$L(\theta)=\prod_{i=1}^{n}\hat{y}^{y^{(i)}}(1-\hat{y})^{1-y^{(i)}}P(X^{(i)})$$

$$LL(\theta)=\sum_{i=1}^{n}\left[y^{(i)}\log\hat{y}^{(i)}+(1-y^{(i)})\log(1-\hat{y}^{(i)})+\log P(X^{(i)})\right]$$

$$\frac{\partial}{\partial\theta_j^{(\hat{y})}}LL(\theta)=\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j^{(\hat{y})}}\left[y^{(i)}\log\boxed{\hat{y}^{(i)}}+(1-y^{(i)})\log(1-\boxed{\hat{y}^{(i)}})\right]$$

$$=\sum_{i=1}^{n}\left[\frac{y^{(i)}}{\hat{y}^{(i)}}-\frac{(1-y^{(i)})}{(1-\hat{y}^{(i)})}\right]\frac{\partial\hat{y}^{(i)}}{\partial\theta_j^{(\hat{y})}}$$

$$=\sum_{i=1}^{n}\left[\frac{y^{(i)}}{\hat{y}^{(i)}}-\frac{(1-y^{(i)})}{(1-\hat{y}^{(i)})}\right]\frac{\partial}{\partial\theta_j^{(\hat{y})}}\sigma(\hat{\theta}^{(\hat{y})T}\vec{h})$$

$$=\sum_{i=1}^{n}\left[\frac{y^{(i)}}{\hat{y}^{(i)}}-\frac{(1-y^{(i)})}{(1-\hat{y}^{(i)})}\right]\hat{y}^{(i)}(1-\hat{y}^{(i)})h_j$$

Maximum likelihood with neural nets

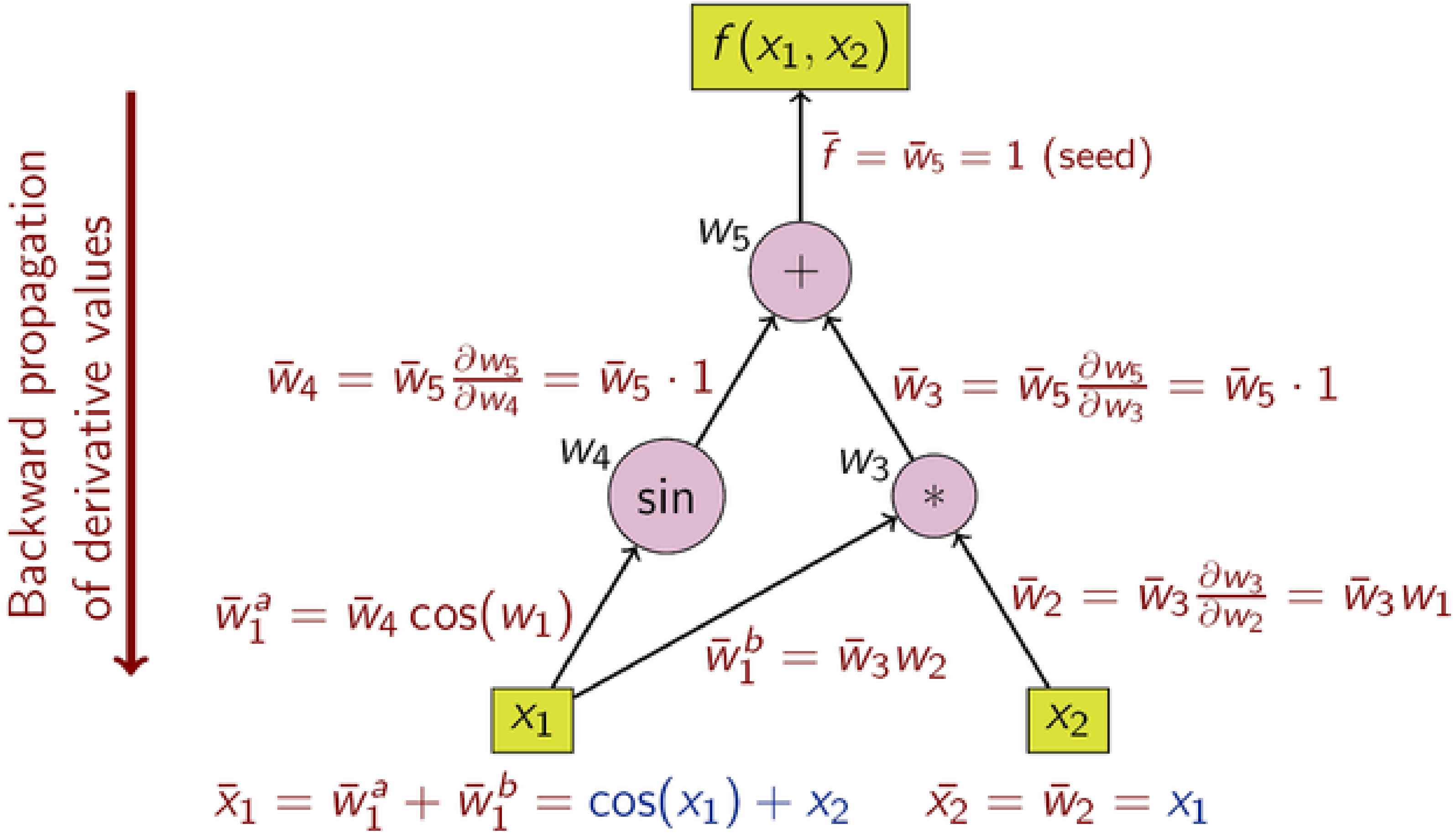
$$L(\theta)=\prod_{i=1}^{n}\hat{y}^{y^{(i)}}(1-\hat{y})^{1-y^{(i)}}P(X^{(i)})$$

$$LL(\theta)=\sum_{i=1}^{n}\left[y^{(i)}\log\hat{y}^{(i)}+(1-y^{(i)})\log(1-\hat{y}^{(i)})+\log P(X^{(i)})\right]$$

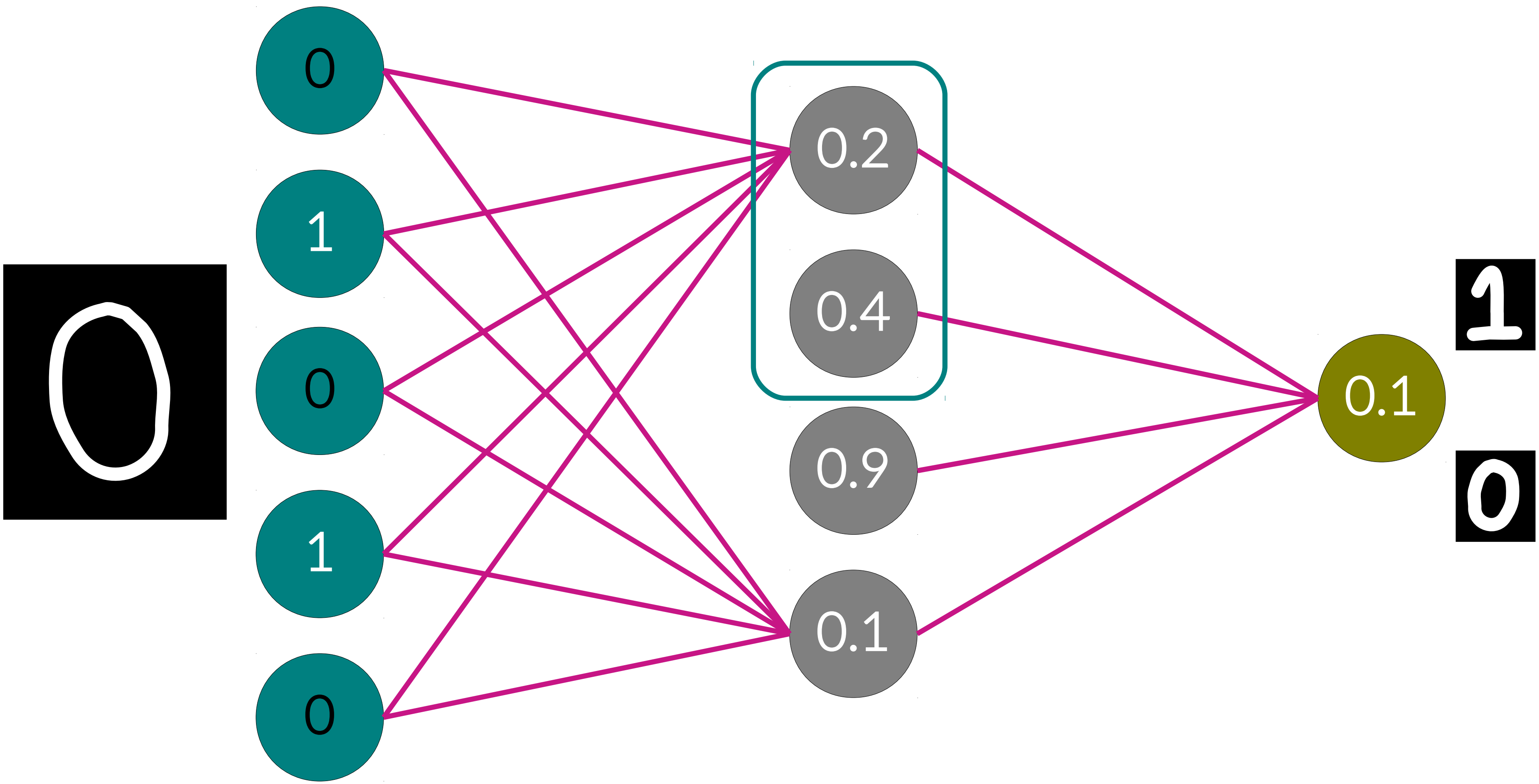
$$\frac{\partial}{\partial\theta_j^{(\hat{y})}}LL(\theta)=\sum_{i=1}^{n}\underbrace{\left[\frac{y^{(i)}}{\hat{y}^{(i)}}-\frac{(1-y^{(i)})}{(1-\hat{y}^{(i)})}\right]}_{\frac{\partial}{\partial\hat{y}^{(i)}}LL^{(i)}(\theta)}\underbrace{\hat{y}^{(i)}(1-\hat{y}^{(i)})}_{\frac{\partial\hat{y}^{(i)}}{\partial z^{(i)}}}\underbrace{h_j}_{\frac{\partial z^{(i)}}{\partial\theta_j^{(\hat{y})}}}$$

$$\frac{\partial}{\partial\theta_{j,k}^{(h)}}LL(\theta)=\sum_{i=1}^{n}\underbrace{\left[\frac{y^{(i)}}{\hat{y}^{(i)}}-\frac{(1-y^{(i)})}{(1-\hat{y}^{(i)})}\right]}_{\frac{\partial}{\partial\hat{y}^{(i)}}LL^{(i)}(\theta)}\underbrace{\hat{y}^{(i)}(1-\hat{y}^{(i)})}_{\frac{\partial\hat{y}^{(i)}}{\partial z^{(i)}}}\underbrace{\theta_j^{(\hat{y})}}_{\frac{\partial z^{(i)}}{d h_j^{(i)}}}\underbrace{h_j(1-h_j)x_k}_{\frac{d h_j^{(i)}}{\partial\theta_{j,k}^{(h)}}}$$

Automatic differentiation



Breaking the symmetry



$\vec{x}$

input features

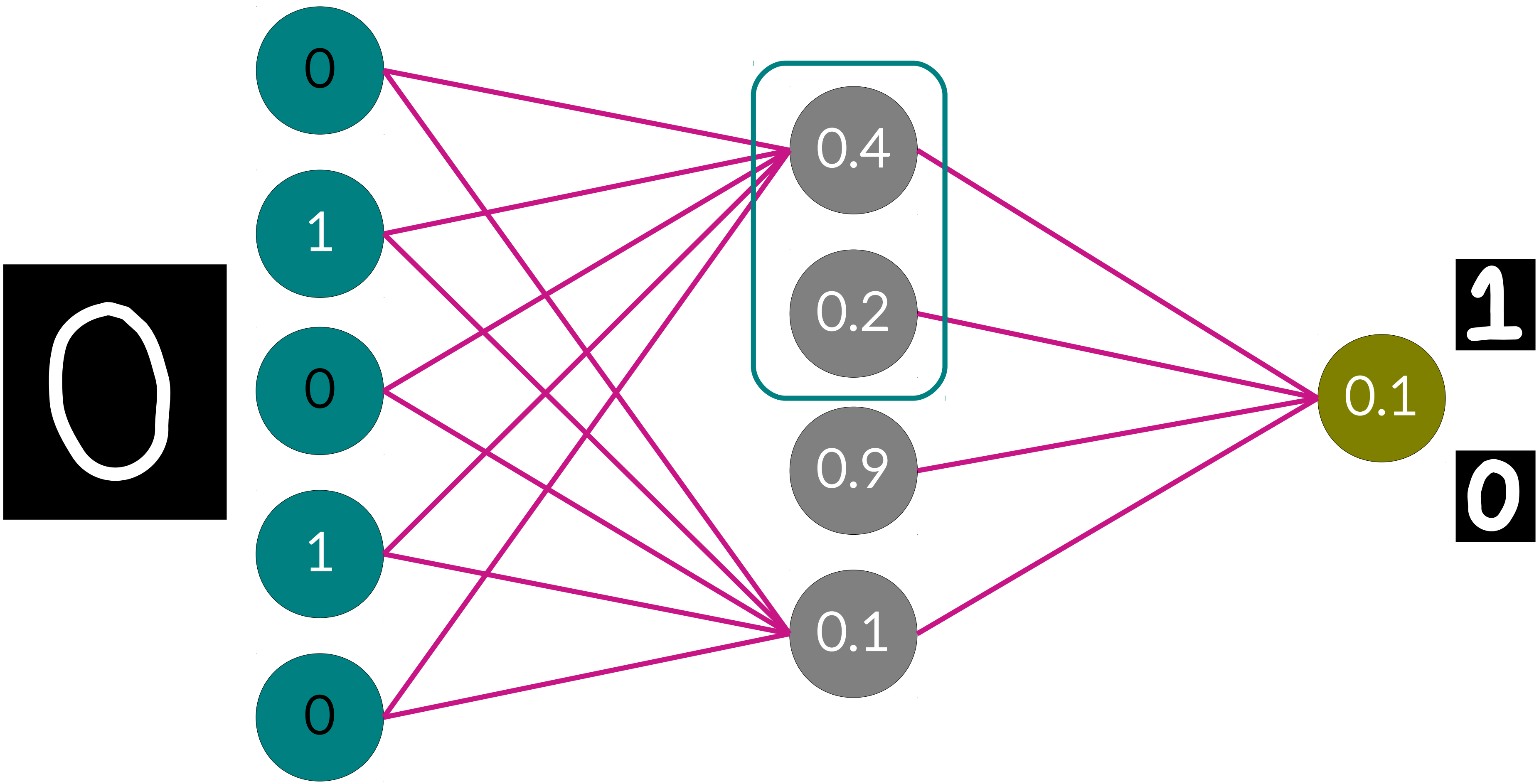
$$\vec{h}=\sigma\left(\theta^{(h)}\vec{x}\right)$$

hidden representation

$$\hat{y}=\sigma\left(\vec{\theta}^{(\hat{y})T}\vec{h}\right)$$

output class label

Breaking the symmetry



$\vec{x}$

input features

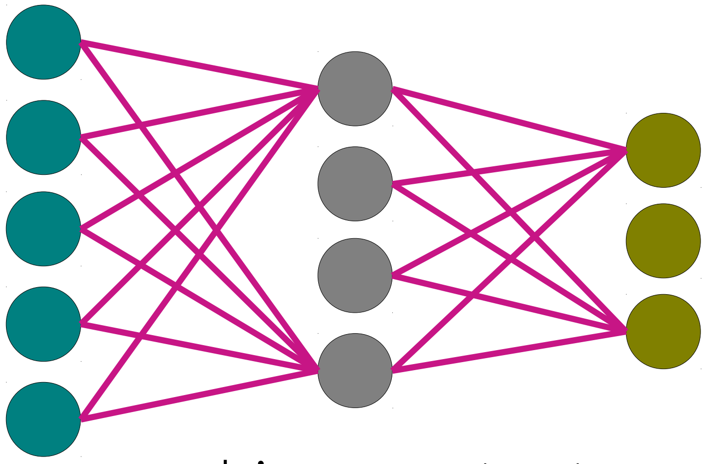
$$\vec{h} = \sigma\left(\theta^{(h)} \vec{x}\right)$$

hidden representation

$$\hat{y} = \sigma\left(\vec{\theta}^{(\hat{y})T} \vec{h}\right)$$

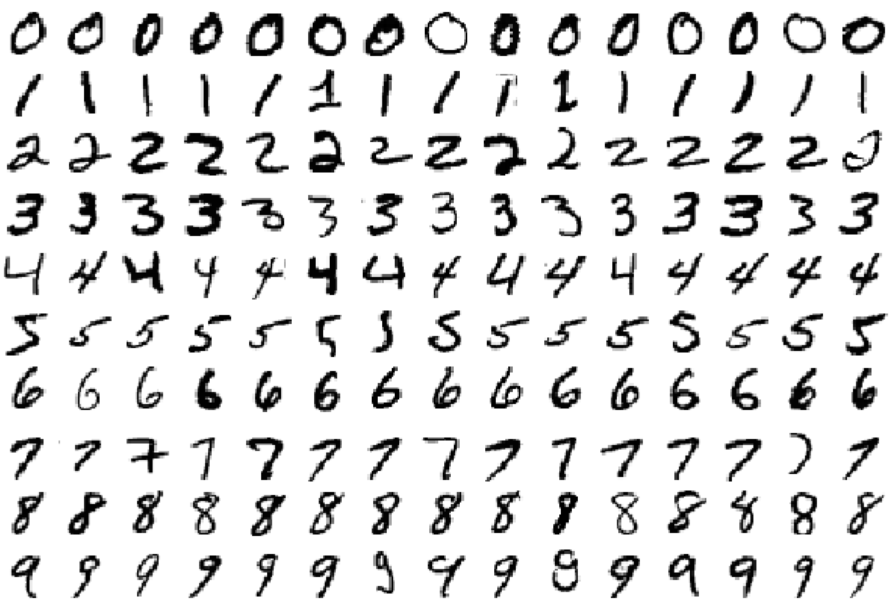
output class label

Expanding the toolbox

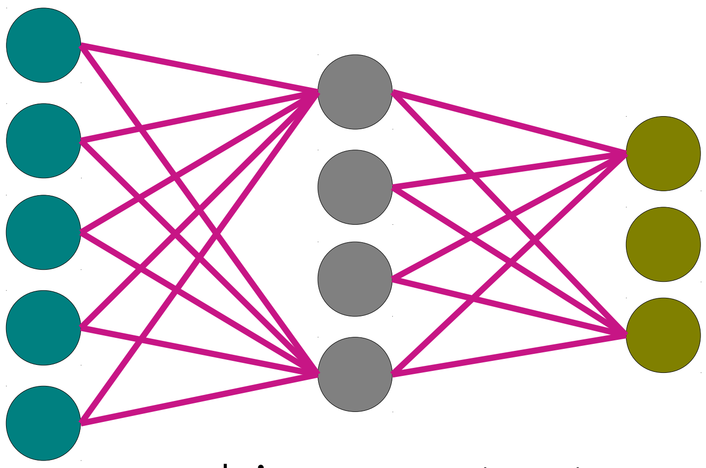


non-binary outputs

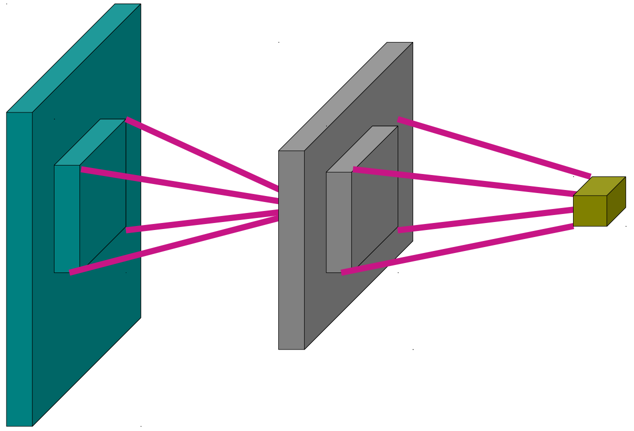
Applications: Image recognition



Expanding the toolbox

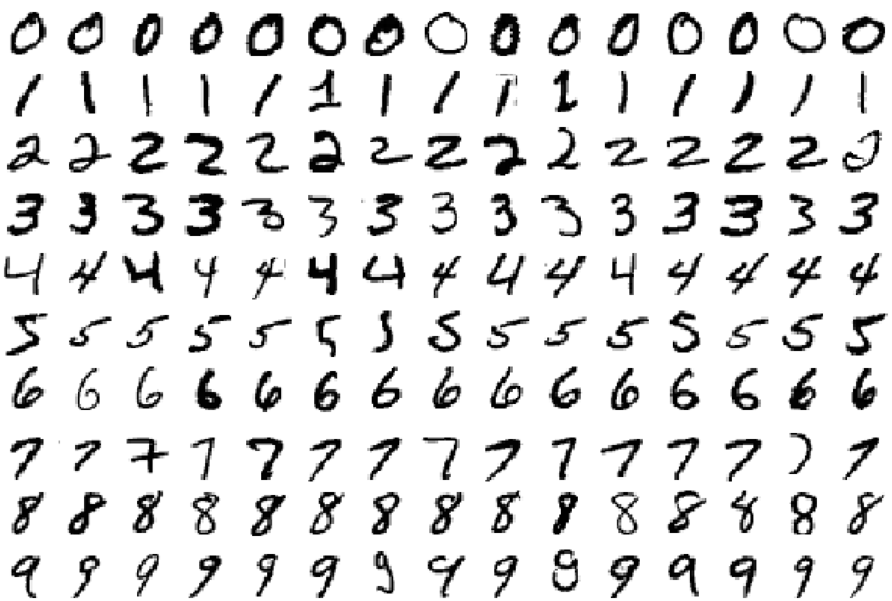


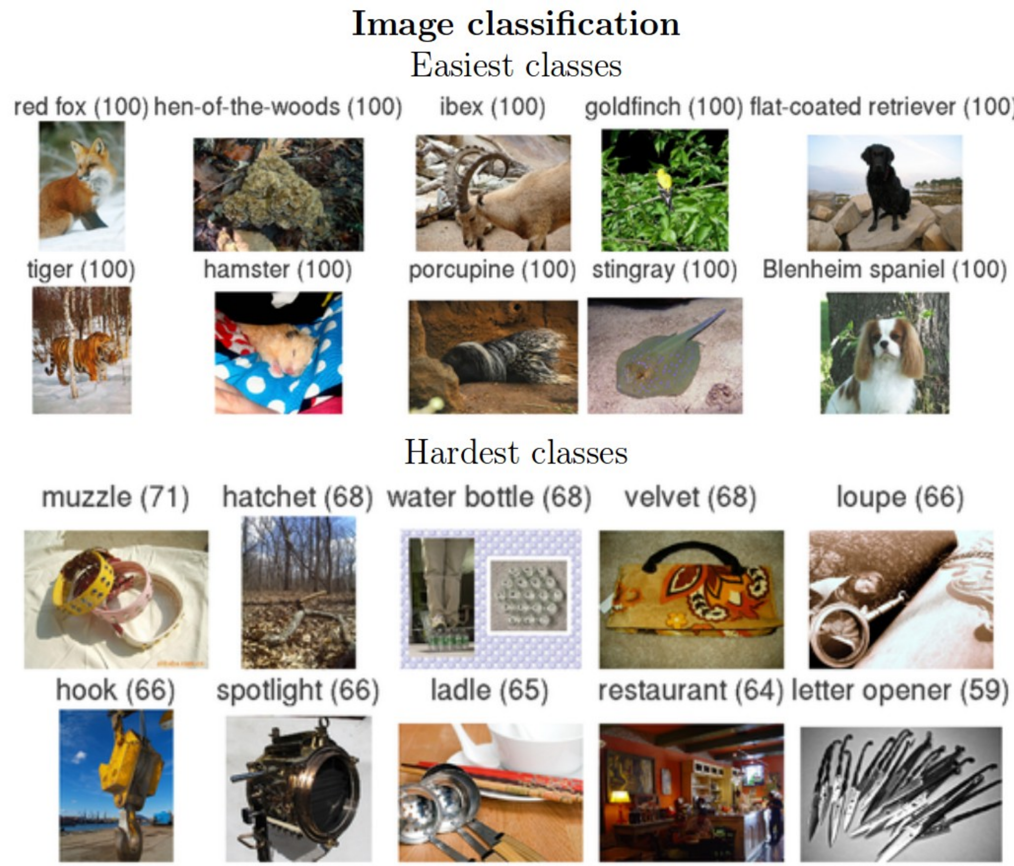
non-binary outputs



convolutional layers

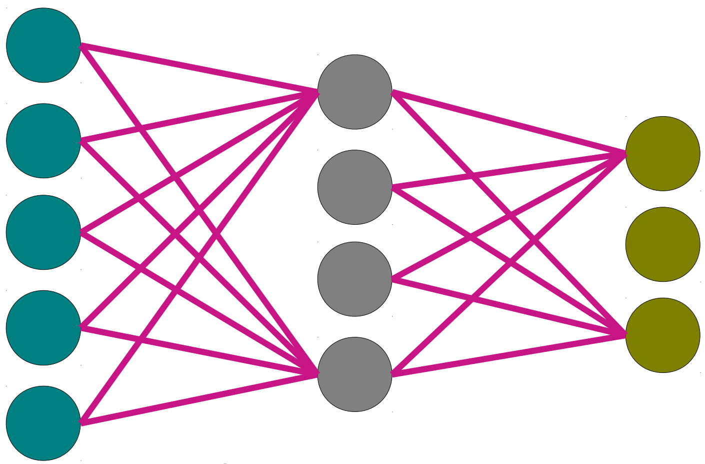
Applications: Image recognition



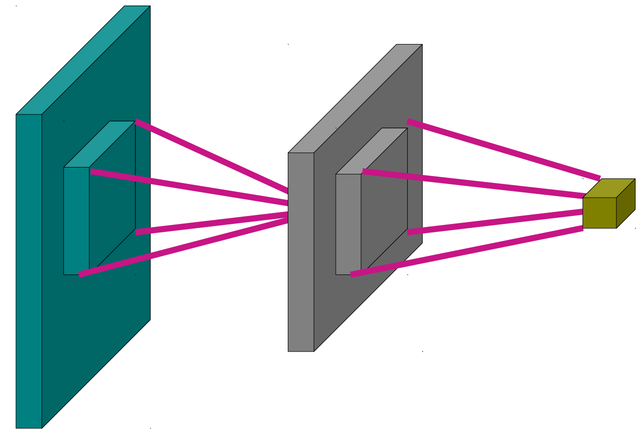


right: Russakovsky et al. (2015)

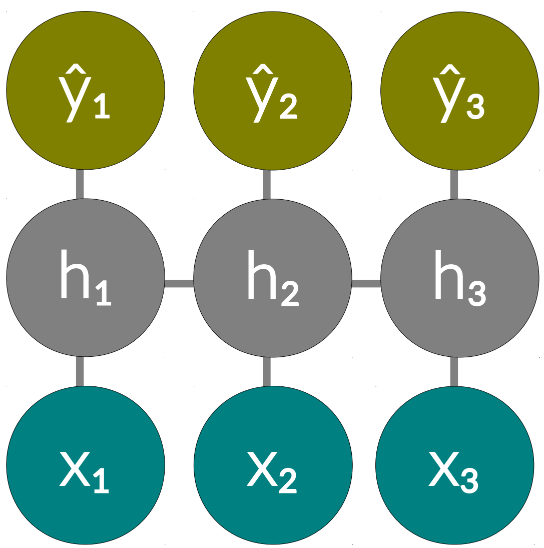
Expanding the toolbox



non-binary outputs



convolutional layers

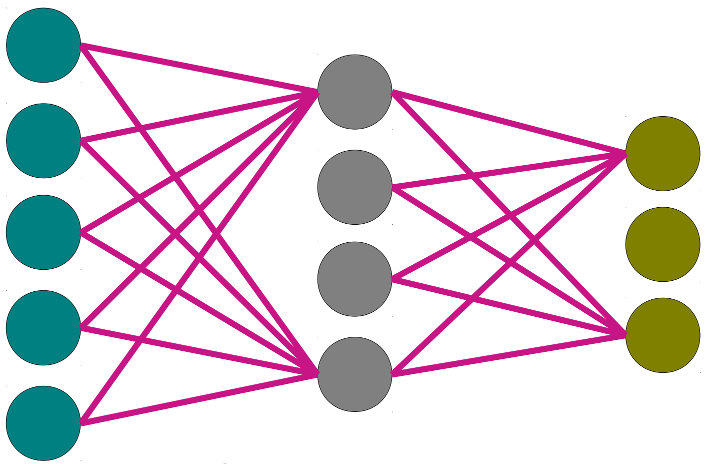


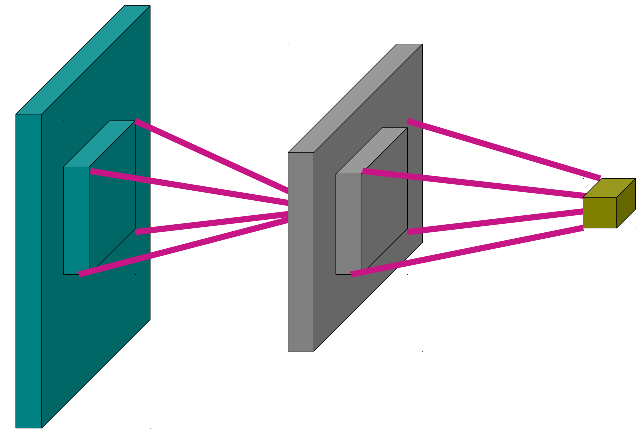
recurrent connections

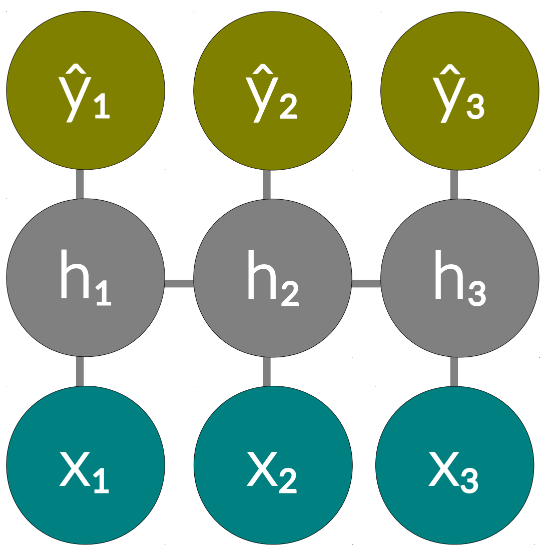
Applications: Speech recognition

who is the current president of France?

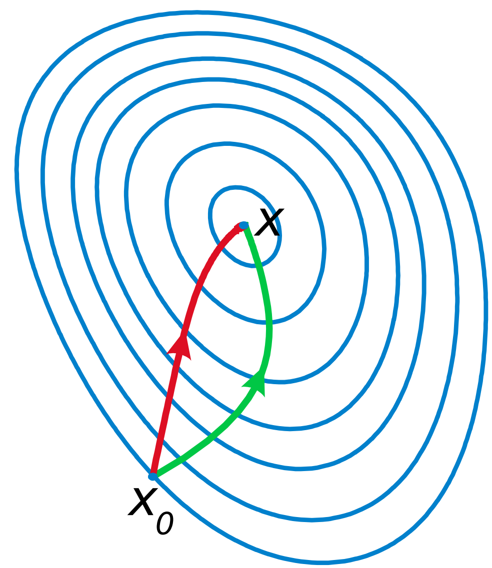
Expanding the toolbox


non-binary outputs


convolutional layers



recurrent connections



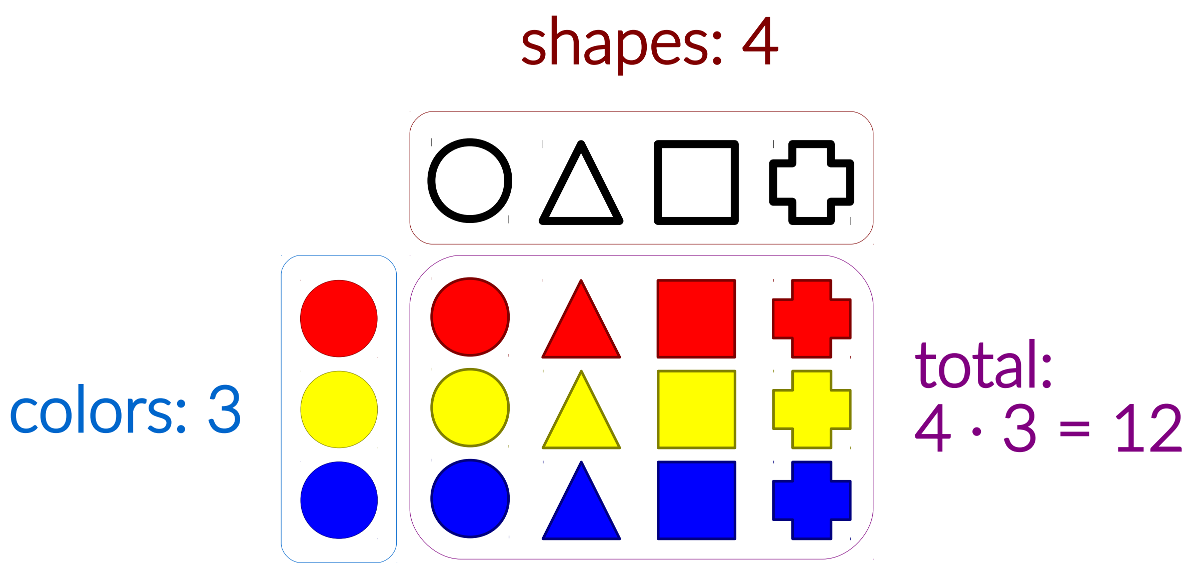
fancy maximization

Break time!

General principle of counting

An experiment consisting of two or more **separate parts** has a number of outcomes equal to the **product** of the number of outcomes of each part.

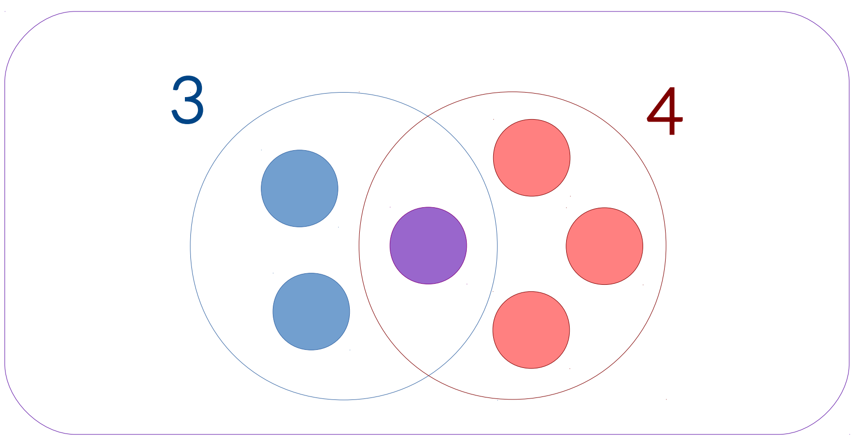
$$|A_1 \times A_2 \times \cdots \times A_n| = \prod_i |A_i|$$



Principle of Inclusion/Exclusion

The **total number of elements** in two sets is the sum of the number of elements of each set, **minus** the number of elements in both sets.

$$|A \cup B| = |A| + |B| - |A \cap B|$$



$$3+4-1 = 6$$

Inclusion/exclusion with more than two sets

size of the union

add or subtract (based on size)

size of intersections

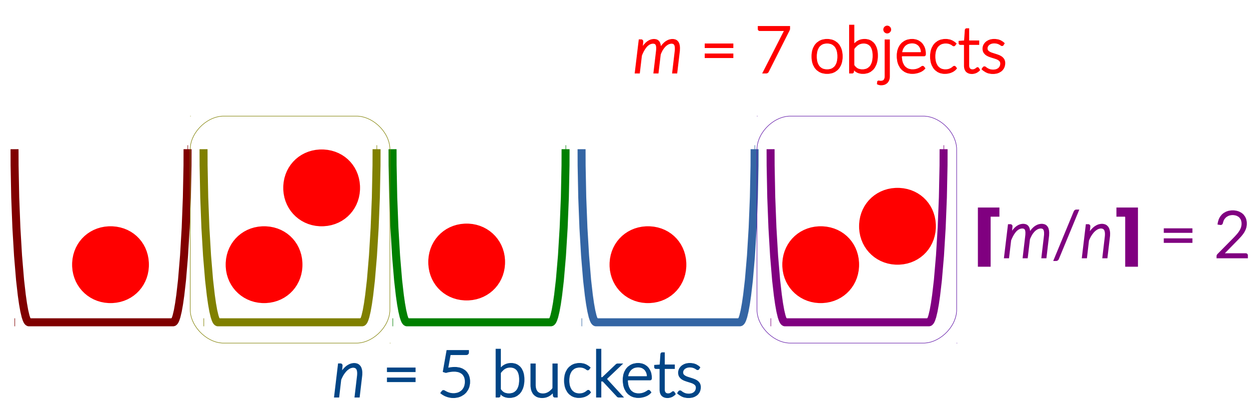
$$\left|\bigcup_{i=1}^{n} E_i\right| = \sum_{r=1}^{n} (-1)^{(r+1)} \sum_{i_1 < \cdots < i_r} \left|\bigcap_{j=1}^{r} E_{i_j}\right|$$

sum over subset sizes

sum over all **subsets** of that size

General Pigeonhole Principle

If *m* objects are placed in *n* buckets, then **at least one bucket** must contain at least $\lceil m/n \rceil$ objects.



Permutations

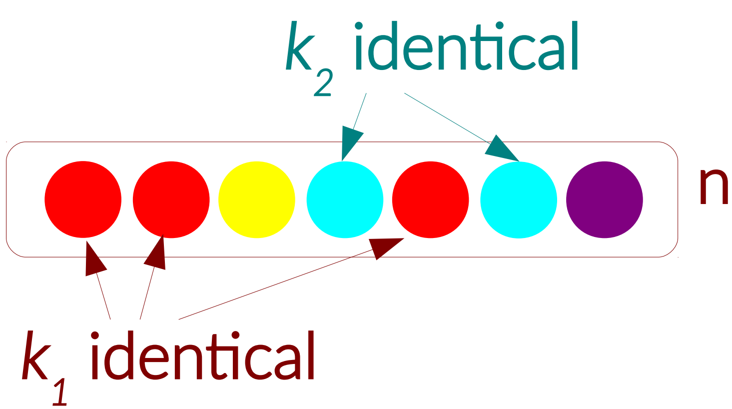
The number of **ways of ordering**
n **distinguishable** objects.

$$n! \quad = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n = \prod_{i=1}^{n} i$$

Permutations with indistinct elements

The number of **ways of ordering** *n*. objects, where some groups are **indistinguishable**.

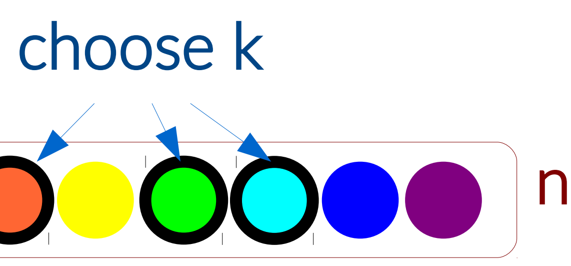
$$\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}$$



Combinations

The number of unique **subsets** of size *k* from a larger set of size *n*.
(objects are distinguishable, unordered)

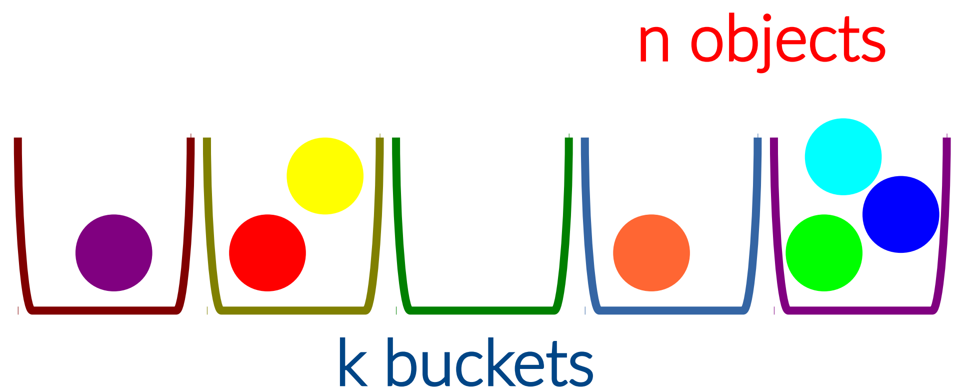
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$



Bucketing

The number of ways of assigning n **distinguishable objects** to a fixed set of k **buckets** or **labels**.

$$k^n$$

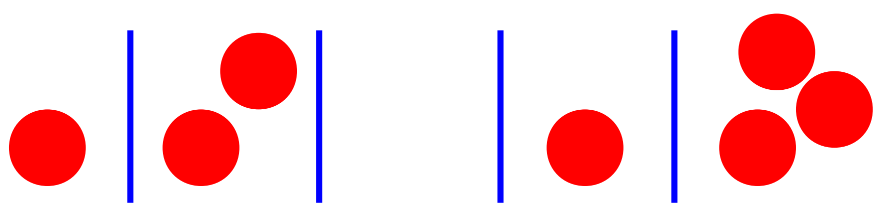


Divider method

The number of ways of assigning n <u>in</u>**distinguishable objects** to a fixed set of k **buckets** or **labels**.

$$\binom{n+(k-1)}{n}$$

n objects



k buckets

(k - 1 dividers)

A grid of ways of counting

	Ordering	Subsets	Bucketing
All distinct	$n!$	$\binom{n}{k}$	k^n
Some indistinct	$\frac{n!}{k_1!k_2!\ldots k_m!}$	Creativity! – Split into cases – Use inclusion/exclusion – Reframe the problem	
All indistinct	1	1	$\binom{n+(k-1)}{n}$

Axioms of probability

(1) $0 \le P(E) \le 1$

(2) $P(S) = 1$

(3) If $E \cap F = \emptyset$, then

$$P(E \cup F) = P(E) + P(F)$$

(Sum rule, but with probabilities!)

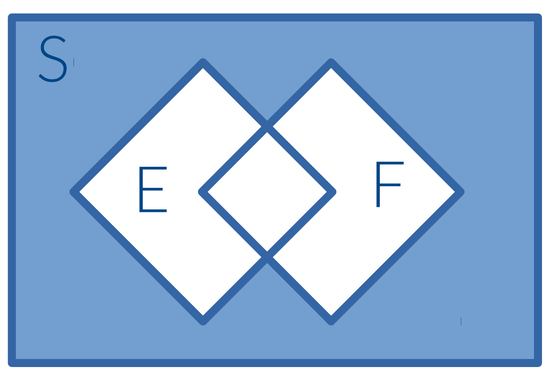
How do I get started?

For word problems involving probability,
start by defining **events**!

Getting rid of ORs

Finding the probability of an OR of events can be nasty. Try **using De Morgan's laws** to turn it into an AND!

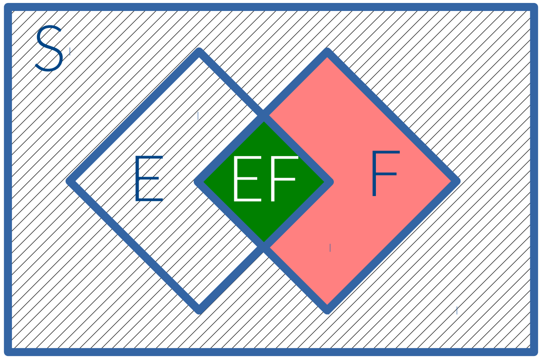
$$P\left(A \cup B \cup \cdots \cup Z\right) = 1 - P\left(A^c B^c \cdots Z^c\right)$$



Definition of conditional probability

The conditional probability P(E | F) is the probability that E happens, **given** that F has happened. F is the new sample space.

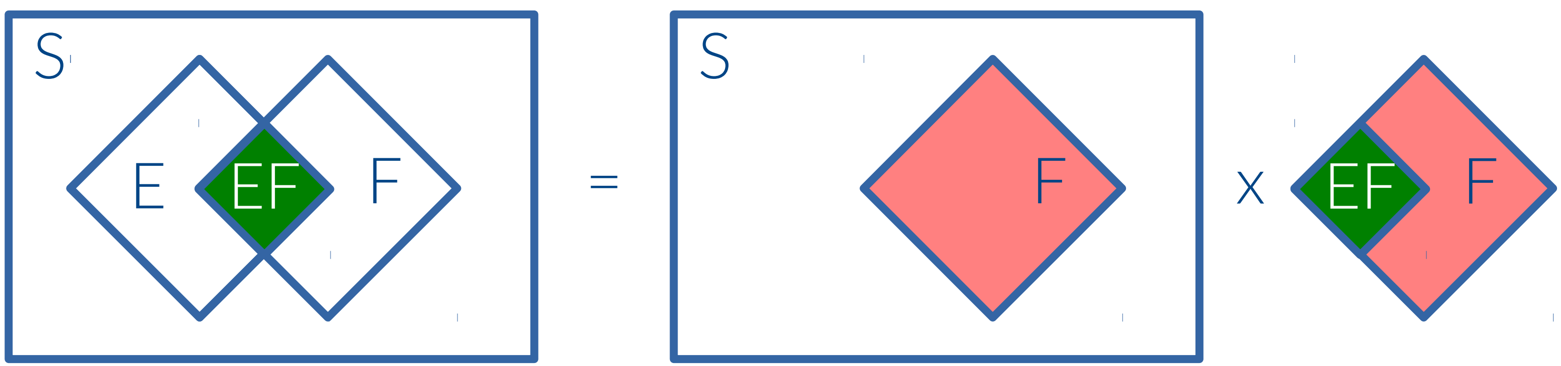
$$P(E|F) = \frac{P(EF)}{P(F)}$$



Chain rule of probability

The probability of **both** events happening is the probability of **one happening** times the probability of **the other happening given the first one**.

$$P(EF) = P(F)\,P(E|F)$$



General chain rule of probability

The probability of **all** events happening
is the probability of **the first** happening
times the prob. of **the second given the first**
times the prob. of **the third given the first two**
...etc.

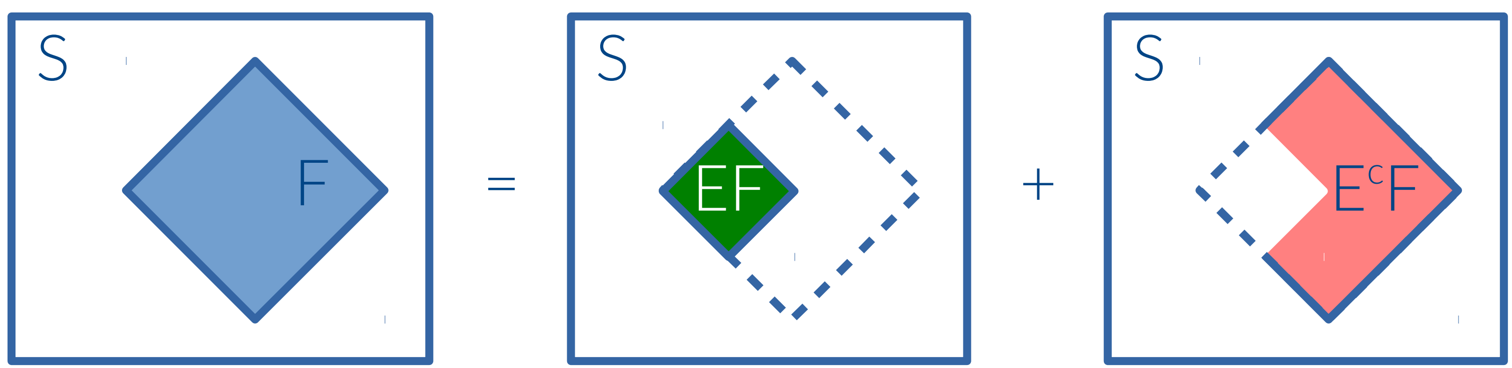
$$P(EFG\ldots) = P(E)\,P(F|E)\,P(G|EF)\ldots$$

Law of total probability

You can compute an overall probability by adding up the case when an event **happens** and when it **doesn't happen**.

$$P(F) = P(EF) + P(E^C F)$$

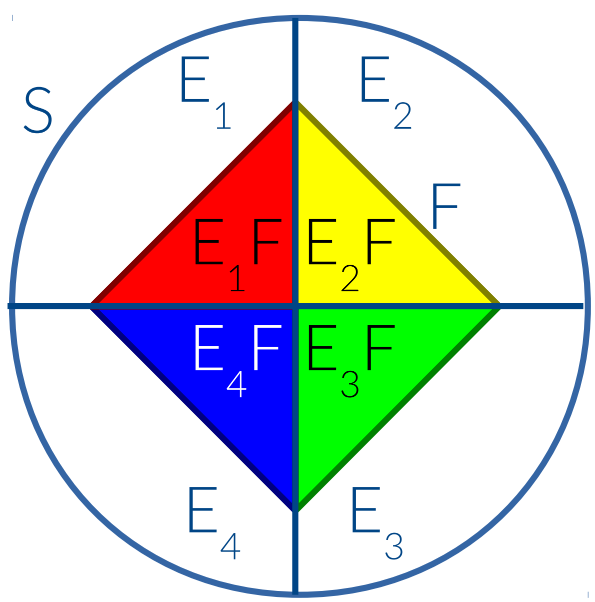
$$= P(E)P(F|E) + P(E^C)P(F|E^C)$$



General law of total probability

You can compute an overall probability by summing over **mutually exclusive** and **exhaustive** sub-cases.

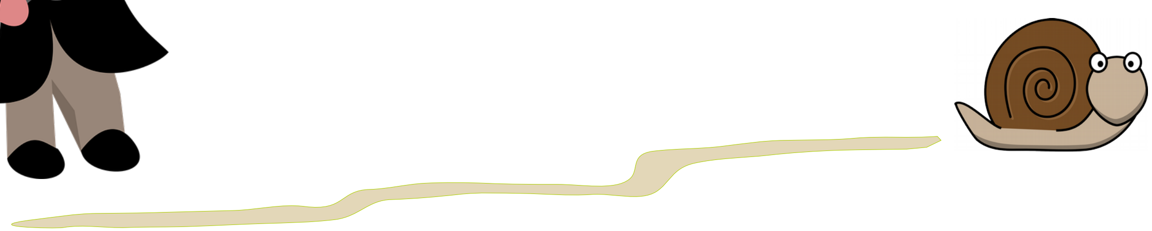
$$P(F) = \sum_i P(E_i F)$$
$$= \sum_i P(E_i) P(F|E_i)$$



Bayes' theorem

You can "flip" a conditional probability if you multiply by the probability of the **hypothesis** and divide by the probability of the **observation**.

$$P(E|F) = \frac{P(F|E)\,P(E)}{P(F)}$$



Finding the denominator

If you don't know P(F) on the bottom,
try using the **law of total probability**.

$$P(E|F) = \frac{P(F|E)P(E)}{\boxed{P(F|E)P(E) + P(F|E^c)P(E^c)}}$$

$$P(E|F) = \frac{P(F|E)P(E)}{\boxed{\sum_i P(F|E_i)P(E_i)}}$$

Independence

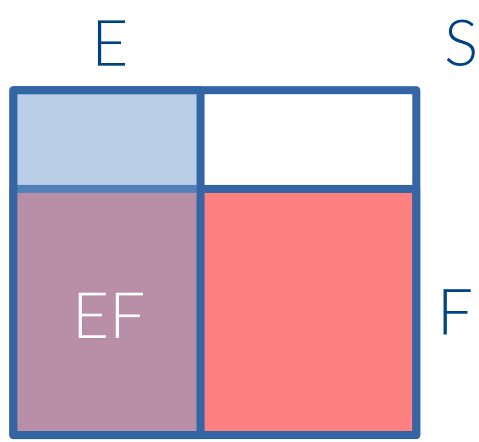
Two events are **independent** if you can **multiply** their probabilities to get the probability of **both** happening.

$$P(EF) = P(E)P(F)$$

$$\Leftrightarrow$$

$$E \perp F$$

(“independent of”)



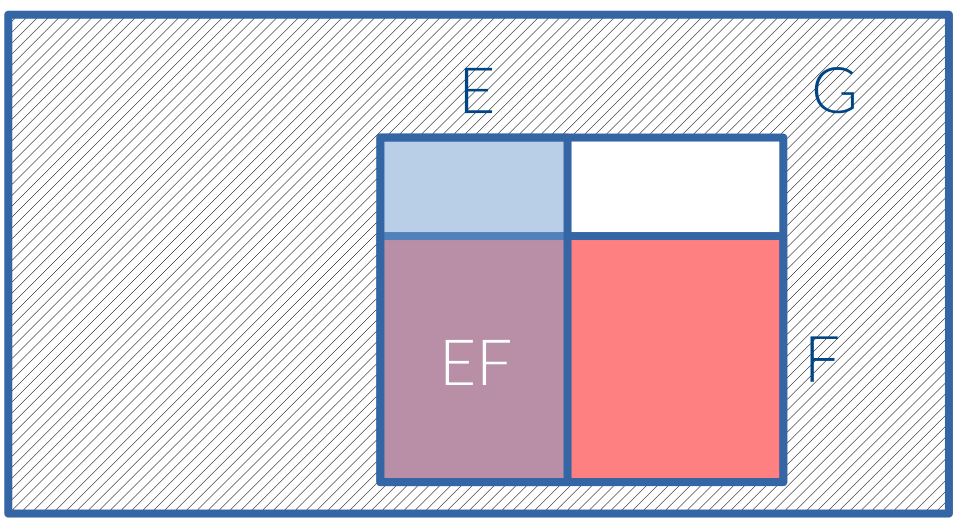
Conditional independence

Two events are **conditionally independent** if you can **multiply** their conditional probabilities to get the conditional probability of **both** happening.

$$P(EF|G) = P(E|G)\,P(F|G)$$

$$\Leftrightarrow$$

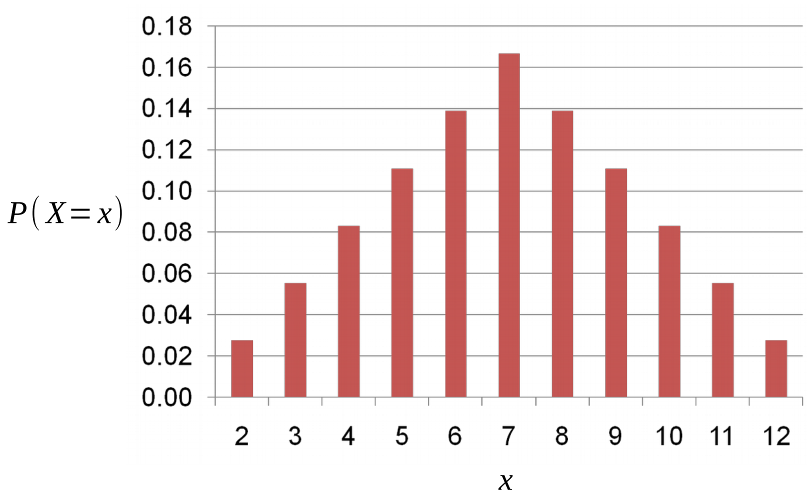
$$(E \perp F)\big|G$$



Random variables

A **random variable** takes on values probabilistically.

$$P(X=2)=\frac{1}{36}$$



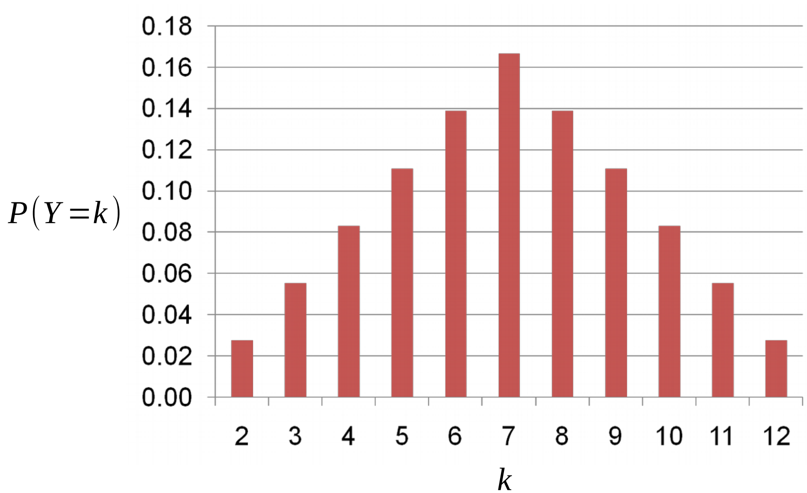
How do I get started?

For word problems involving probability,
start by defining **events** and **random variables**!

Probability mass function

The **probability mass function** (PMF) of a random variable is a function from values of the variable to probabilities.

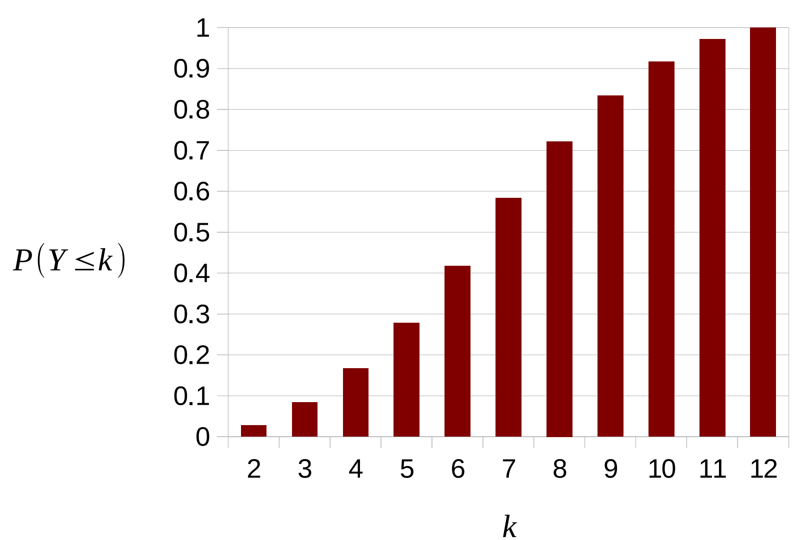
$$p_Y(k) = P(Y = k)$$



Cumulative distribution function

The **cumulative distribution function** (CDF) of a random variable is a function giving the probability that the random variable is **less than or equal to** a value.

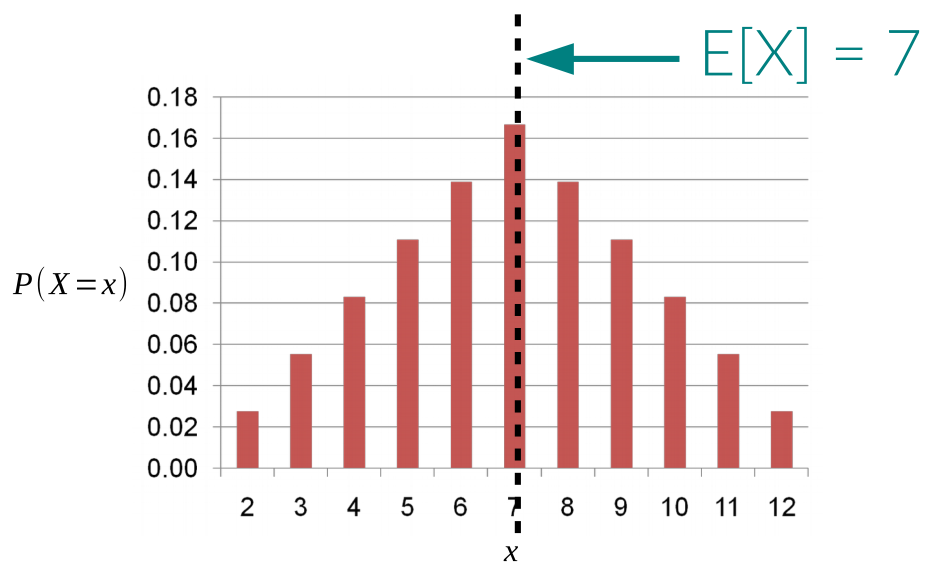
$$F_Y(k) = P(Y \le k)$$



Expectation

The **expectation** of a random variable is the "**average**" value of the variable (weighted by probability).

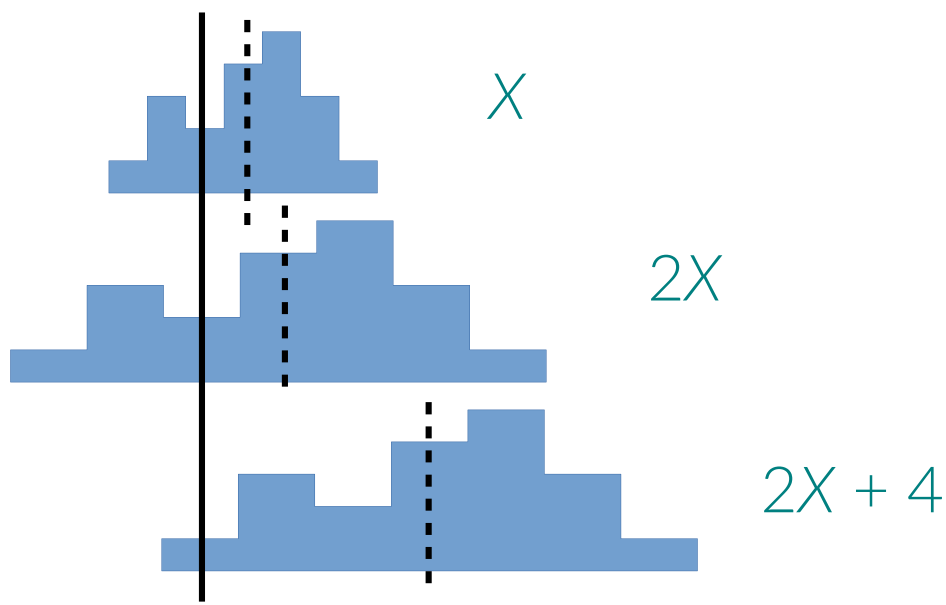
$$E[X] = \sum_{x:p(x)>0} p(x) \cdot x$$



Linearity of expectation

Adding random variables or constants? **Add** the expectations.
Multiplying by a <u>constant</u>? **Multiply** the expectation by the constant.

$$E[aX+bY+c]=aE[X]+bE[Y]+c$$



Indicator variable

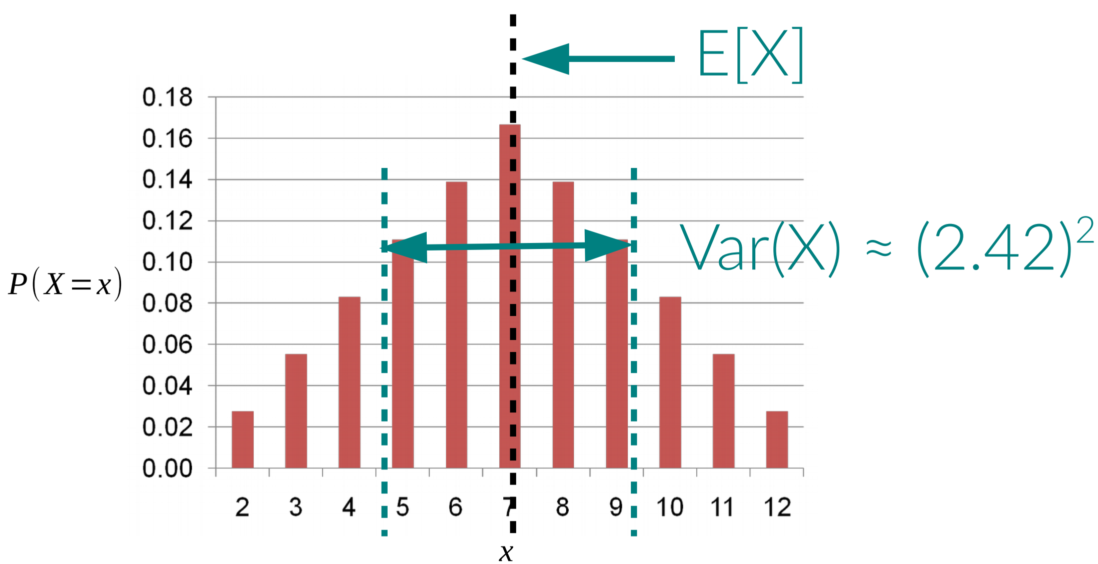
An **indicator variable** is a "Boolean" variable, which takes values 0 or 1 corresponding to whether an event takes place.

$$I = \mathbb{1}[A] = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Variance

Variance is the average **square** of the **distance** of a variable from the expectation. Variance measures the “**spread**” of the variable.

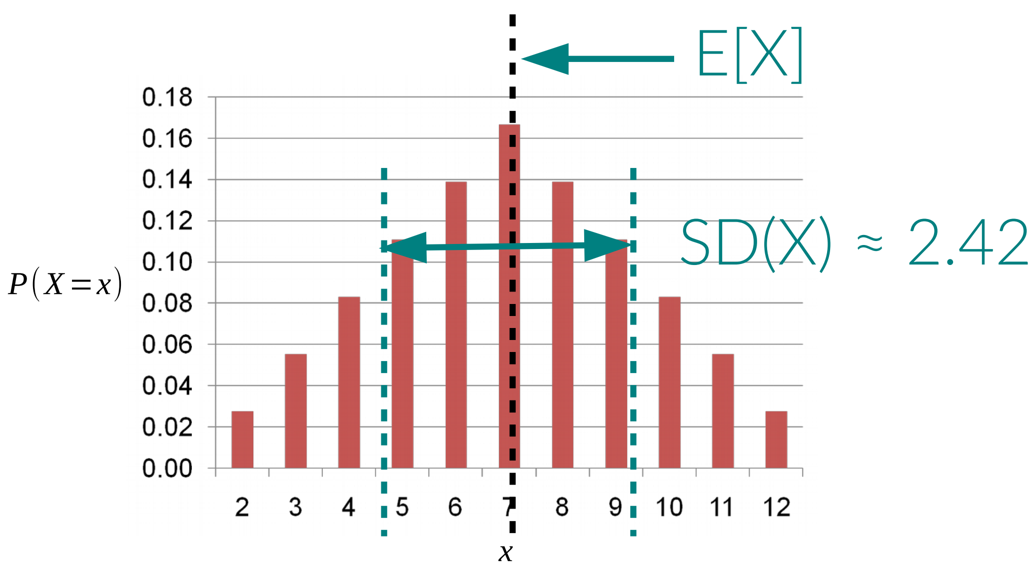
$$\mathrm{Var}(X) = E[(X - E[X])^2]$$
$$= E[X^2] - (E[X])^2$$



Standard deviation

Standard deviation is the ("root-mean-square") average of the **distance** of a variable from the expectation.

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{E[(X - E[X])^2]}$$



Variance of a linear function

Adding a <u>constant</u>? Variance **doesn't change**.
Multiplying by a <u>constant</u>? **Multiply** the variance by the **square** of the constant.

$$
\begin{aligned}
\mathrm{Var}(aX+b) &= E[(aX+b)^2]-(E[aX+b])^2 \\
&= E[a^2X^2+2abX+b^2]-(aE[X]+b)^2 \\
&= a^2E[X^2]+2abE[X]+b^2 \\
&\quad -[a^2(E[X])^2+2abE[X]+b^2] \\
&= a^2E[X^2]-a^2(E[X])^2 \\
&= a^2[E[X^2]-(E[X])^2] \\
&= a^2\mathrm{Var}(X)
\end{aligned}
$$

Basic distributions

Many types of random variables come up repeatedly. Known frequently-occurring **distributions** lets you do computations without deriving formulas from scratch.

$$\overset{\text{variable}}{X} \sim \overset{\text{family}}{\text{Bin}}\left(n, \overset{\text{parameters}}{p}\right)$$

We have _________ (INTEGER) independent __________ (PLURAL NOUN),
each of which _________ (VERB ENDING IN -S) with probability
_________ (REAL NUMBER). How many of the _________ (REPEAT PLURAL NOUN)
_________ (REPEAT VERB -S)?

Bernoulli random variable

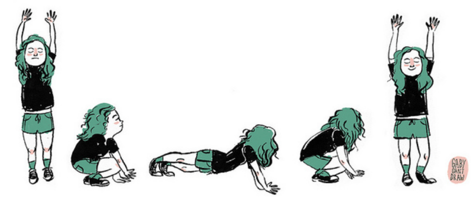
An indicator variable (a possibly biased coin flip) obeys a **Bernoulli distribution**. Bernoulli random variables can be 0 or 1.

$$X \sim \mathrm{Ber}(p)$$

$$p_X(1) = p$$
$$p_X(0) = 1 - p \qquad \text{(0 elsewhere)}$$

Bernoulli: Fact sheet

$$X \sim \mathrm{Ber}(p)$$



?

probability of “success” (heads, ad click, ...)

PMF:

$$p_X(1) = p$$
$$p_X(0) = 1 - p \qquad \text{(0 elsewhere)}$$

expectation: $E[X] = p$

variance: $\mathrm{Var}(X) = p(1-p)$

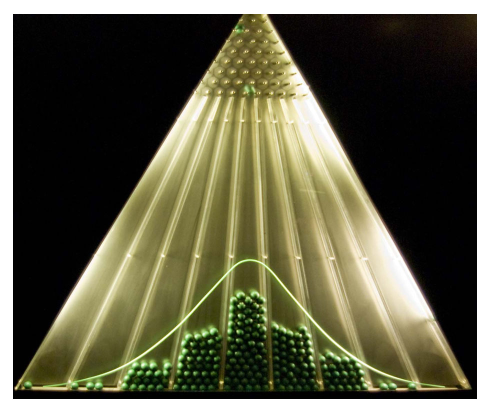
image (right): Gabriela Serrano

Binomial random variable

The **number of heads** on *n* (possibly biased) coin flips obeys a **binomial distribution**.

$$X \sim \mathrm{Bin}(n, p)$$

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \mathbb{N}, 0 \le k \le n \\ 0 & \text{otherwise} \end{cases}$$



Binomial: Fact sheet

number of trials (flips, program runs, ...)

$$X \sim \text{Bin}(n, p)$$

probability of “success” (heads, crash, ...)

PMF: $$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \mathbb{N}, 0 \le k \le n \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = np$$

variance: $$\text{Var}(X) = np(1-p)$$

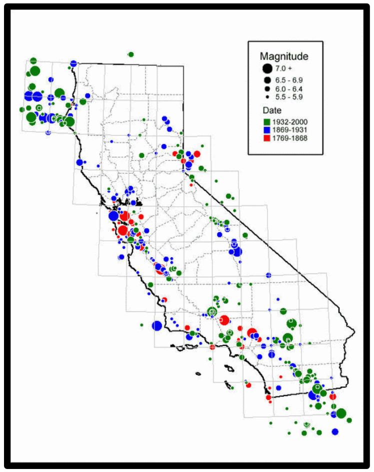
note: $\text{Ber}(p) = \text{Bin}(1, p)$

Poisson random variable

The **number of occurrences** of an event that occurs with **constant rate** λ (per unit time), in 1 unit of time, obeys a **Poisson distribution**.

$$X \sim \mathrm{Poi}(\lambda)$$

$$p_X(k) = \begin{cases} e^{-\lambda} \dfrac{\lambda^k}{k!} & \text{if } x \in \mathbb{Z}, x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Poisson: Fact sheet

$$X \sim \mathrm{Poi}(\lambda)$$

rate of events (requests, earthquakes, chocolate chips, …)
per unit time (hour, year, cookie, …)

PMF: $$p_X(k) = \begin{cases} e^{-\lambda} \dfrac{\lambda^k}{k!} & \text{if } k \in \mathbb{Z}, k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

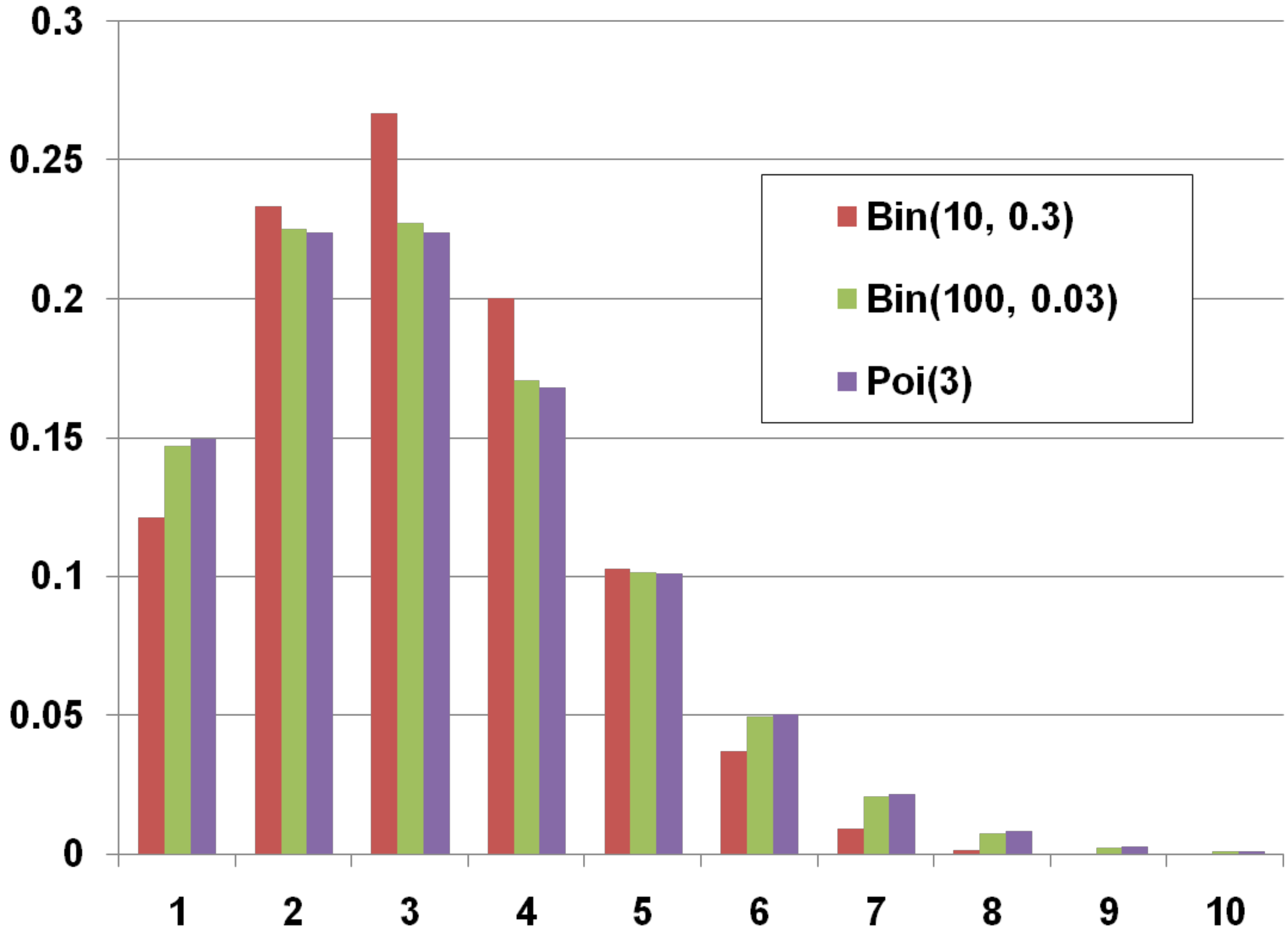
expectation: $$E[X] = \lambda$$

variance: $$\mathrm{Var}(X) = \lambda$$

Poisson approximation of binomial

$$\lambda = np$$

$P(X = k)$



k

Geometric random variable

The **number of trials** it takes to get **one success**, if successes occur independently with probability *p*, obeys a **geometric distribution**.

$$X \sim \mathrm{Geo}(p)$$

$$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Geometric: Fact sheet

$$X \sim \mathrm{Geo}(p)$$

probability of "success" (catch, heads, crash, ...)

PMF: $$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

CDF: $$F_X(k) = \begin{cases} 1-(1-p)^{k} & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{1}{p}$$

variance: $$\mathrm{Var}(X) = \frac{1-p}{p^2}$$

Negative binomial random variable

The **number of trials** it takes to get *r* successes, if successes occur independently with probability *p*, obeys a **negative binomial distribution**.

$$X \sim \mathrm{NegBin}(r, p)$$

$$p_X(n) = \begin{cases} \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \in \mathbb{Z}, n \geq r \\ 0 & \text{otherwise} \end{cases}$$

Negative binomial: Fact sheet

number of **sucesses** (heads, crash, ...)

$$X \sim \mathrm{NegBin}(r, p)$$

number of **trials** (flips, program runs, ...)

probability of "success"

PMF: $$p_X(n) = \begin{cases} \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \in \mathbb{Z}, n \geq r \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{r}{p}$$

variance: $$\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

note:

$$\mathrm{Geo}(p) = \mathrm{NegBin}(1, p)$$

Continuous random variables

A **continuous** random variable has a value that's a **real number** (not necessarily an integer).

Replace sums with **integrals**!

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

$$F_X(a) = \int_{x=-\infty}^{a} dx \; f_X(x)$$

Probability density function

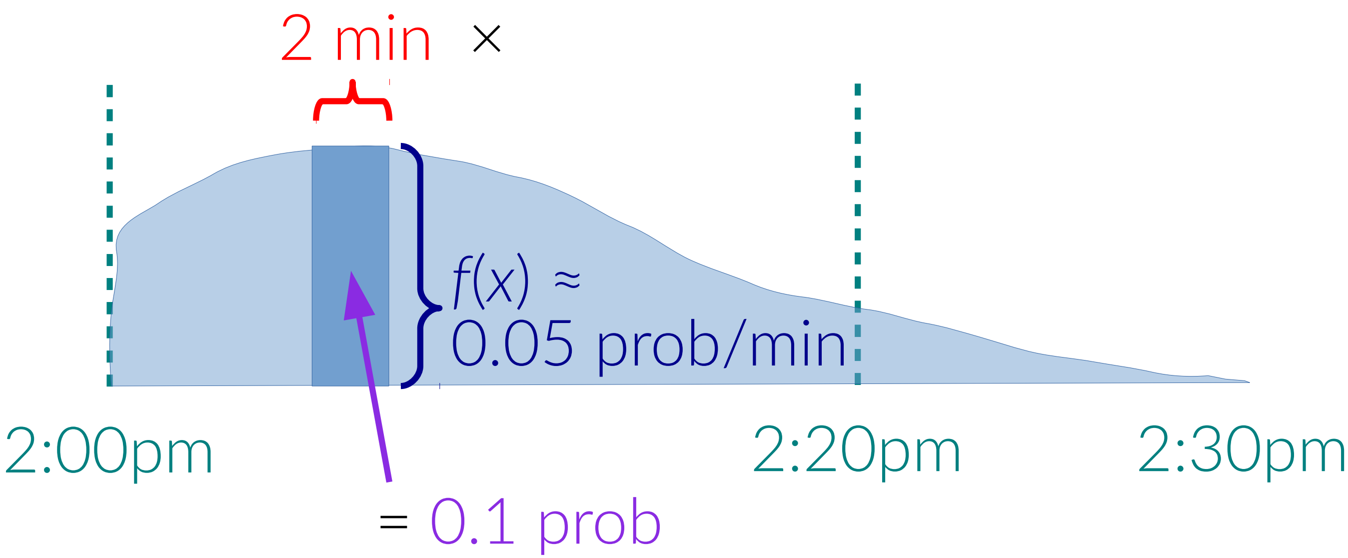
The **probability density function** (PDF) of a continuous random variable represents the relative likelihood of various values.

Units of probability *divided by units of X*. **Integrate it** to get probabilities!

$$P(a < X \leq b) = \int_{x=a}^{b} dx \boxed{f_X(x)}$$

$f(x)$ is not a probability

Rather, it has "units" of probability
divided by units of X.



Uniform random variable

A **uniform** random variable is **equally likely** to be any value in a single real number interval.

$$X \sim \mathrm{Uni}(\alpha, \beta)$$

$$f_X(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases}$$

Uniform: Fact sheet

$$X \sim \mathrm{Uni}(\overset{\text{minimum value}}{\downarrow \atop \alpha}, \underset{\text{maximum value}}{\uparrow \atop \beta})$$

PDF: $$f_X(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases}$$

CDF: $$F_X(x) = \begin{cases} \dfrac{x - \alpha}{\beta - \alpha} & \text{if } x \in [\alpha, \beta] \\ 1 & \text{if } x > \beta \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{\alpha + \beta}{2}$$

variance: $$\mathrm{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

image: Haha169

Exponential random variable

An **exponential** random variable is the **amount of time until the first event** when events occur as in the Poisson distribution.

$$X \sim \mathrm{Exp}(\lambda)$$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

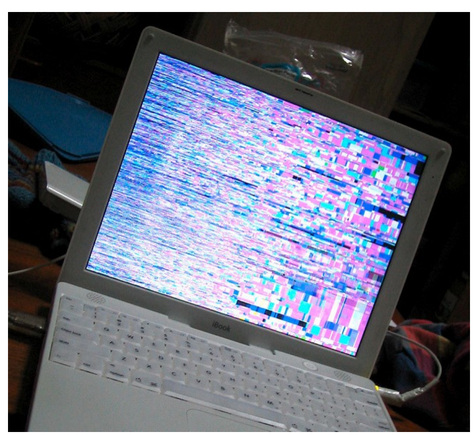


image: Adrian Sampson

Exponential: Fact sheet

rate of events per unit time

$$X \sim \mathrm{Exp}(\lambda)$$

time until first event

PDF: $$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

CDF: $$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{1}{\lambda}$$

variance: $$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

image: Adrian Sampson

A grid of random variables

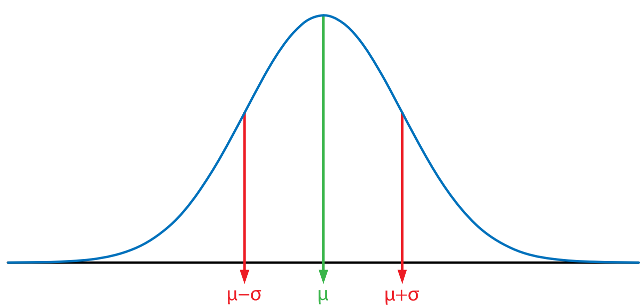
	number of **successes**	**time** to get successes	
One trial	$X \sim \mathrm{Ber}(p)$	$X \sim \mathrm{Geo}(p)$	One success
	$\uparrow$ $n = 1$	$\uparrow$ $r = 1$	
Several trials	$X \sim \mathrm{Bin}(n, p)$	$X \sim \mathrm{NegBin}(r, p)$	Several successes
Interval of time	$X \sim \mathrm{Poi}(\lambda)$	$X \sim \mathrm{Exp}(\lambda)$	One success after interval of time

Normal random variable

An **normal** (= **Gaussian**) random variable is a good approximation to many other distributions. It often results from **sums or averages** of independent random variables.

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

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



Normal: Fact sheet

mean

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

variance (σ = standard deviation)

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

CDF: $$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{x} dx\, f_X(x)$$

(no closed form)

expectation: $$E[X] = \mu$$

variance: $$\mathrm{Var}(X) = \sigma^2$$

The Standard Normal

$$Z \sim N(\underset{\uparrow \atop \mu}{0}, \underset{\uparrow \atop \sigma^2}{1})$$

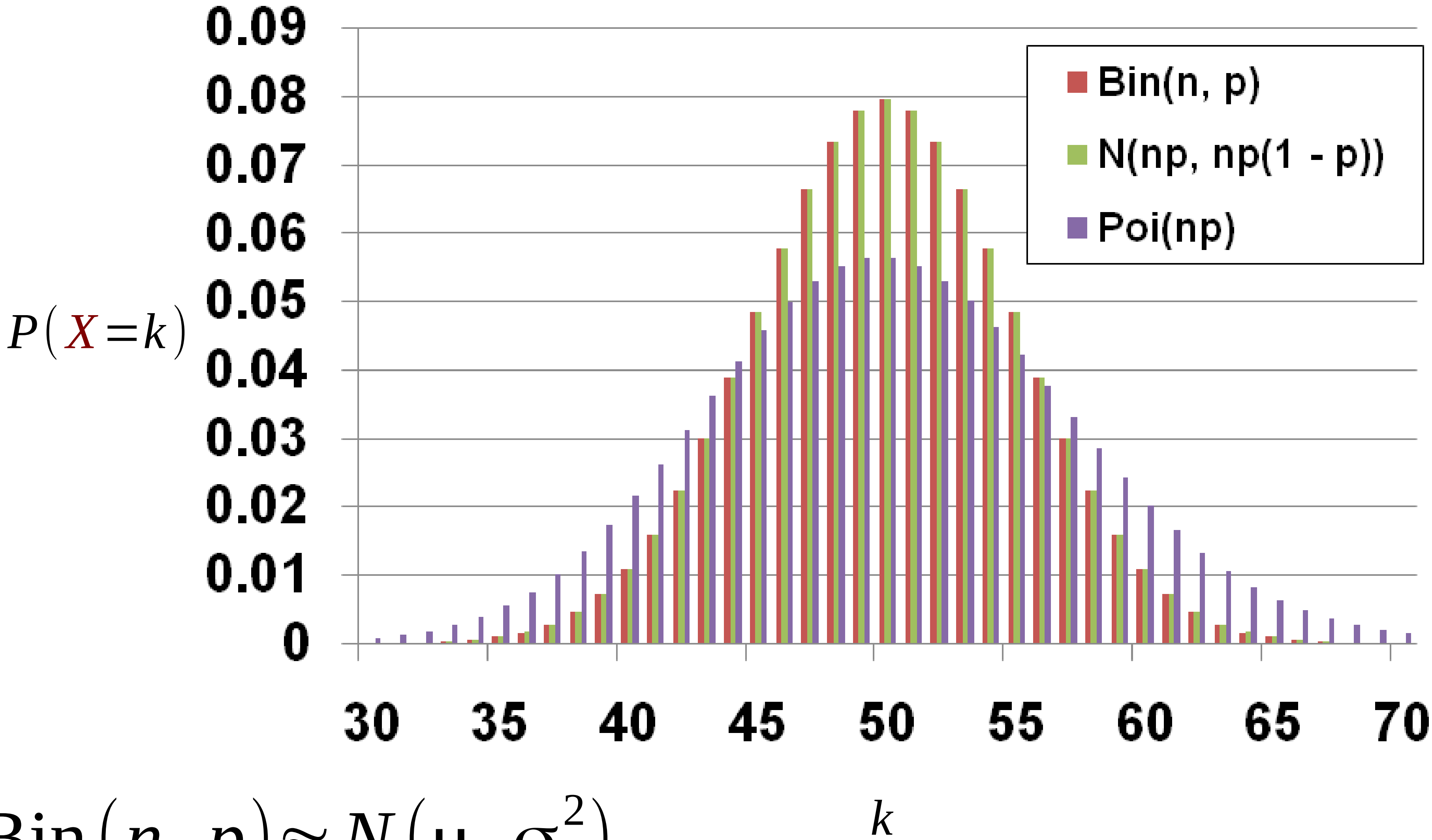
$$X \sim N(\mu, \sigma^2) \rightarrow X = \sigma Z + \mu$$

$$Z = \frac{X - \mu}{\sigma}$$

$$\Phi(z) = F_Z(z) = P(Z \le z)$$

Normal approximation to binomial

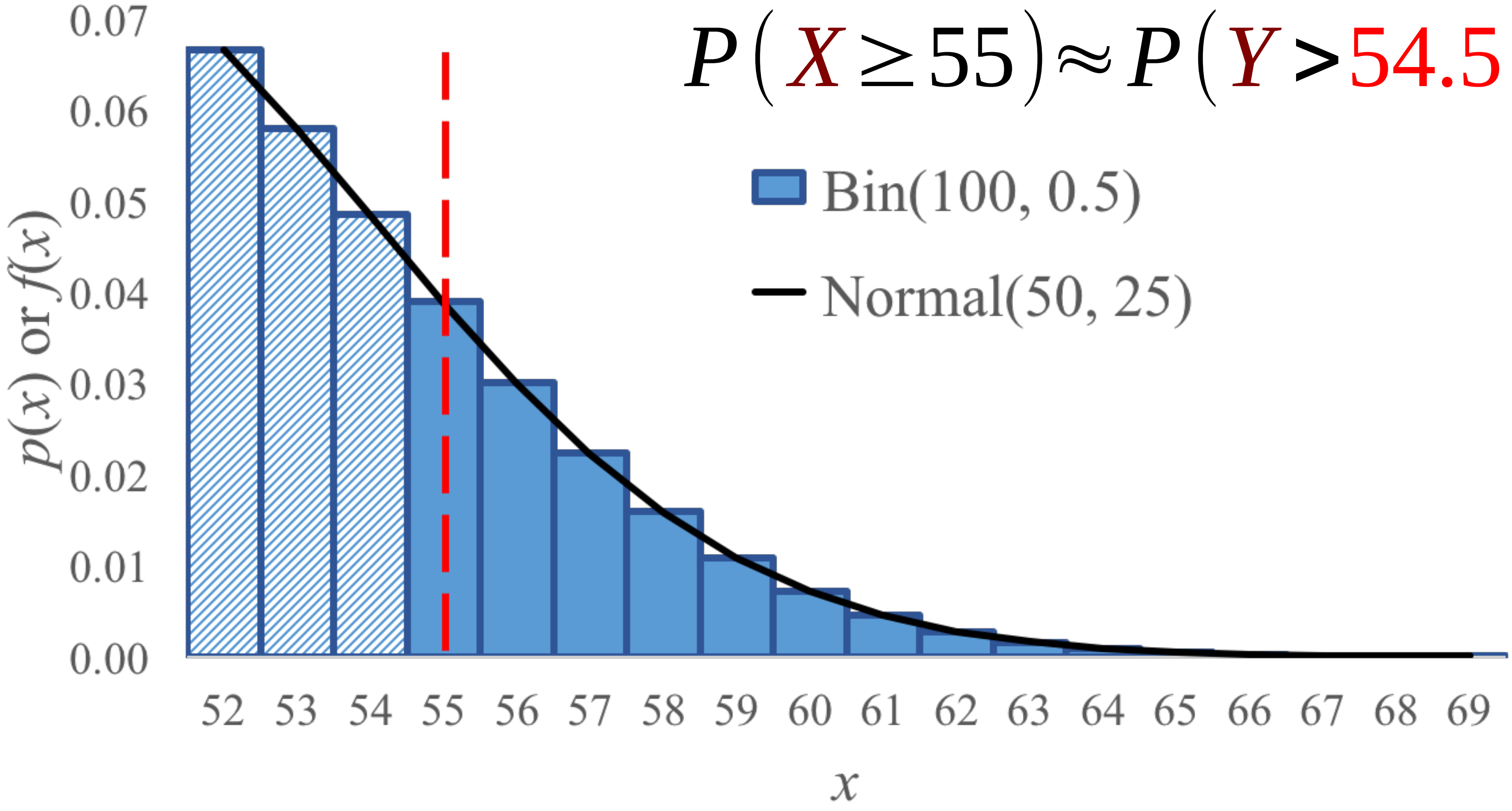
large n, **medium** p



$$\mathrm{Bin}(n,p) \approx N(\mu, \sigma^2)$$

Continuity correction

$$X \sim \text{Bin}(n, p)$$
$$Y \sim N(np, np(1-p))$$
$$P(X \geq 55) \approx P(Y > 54.5)$$



When approximating a **discrete** distribution with a **continuous** distribution, adjust the bounds by 0.5 to account for the missing half-bar.

Joint distributions

A **joint distribution** combines multiple random variables. Its PDF or PMF gives the probability or relative likelihood of **both** random variables taking on specific values.

$$p_{X,Y}(a,b)=P(X=a,Y=b)$$

Joint probability mass function

A **joint probability mass function** gives the probability of **more than one** discrete random variable **each** taking on a specific value (an AND of the 2+ values).

$$p_{X,Y}(a,b) = P(X=a, Y=b)$$

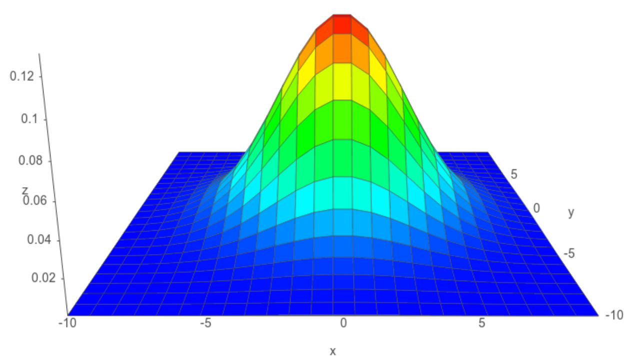
Y

		0	1	2
	0	0.05	0.20	0.10
X	1	0.10	0.10	0.10
	2	0.05	0.10	0.20

Joint probability density function

A **joint probability density function** gives the relative likelihood of **more than one** continuous random variable **each** taking on a specific value.

$$P\left(a_1 < X \leq a_2, b_1 < Y \leq b_2\right) = \int_{a_1}^{a_2} dx \int_{b_1}^{b_2} dy \boxed{f_{X,Y}(x,y)}$$



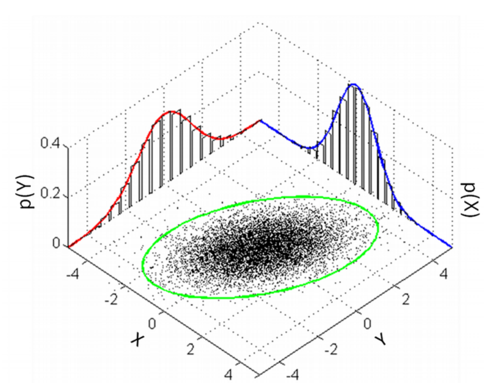
Marginalization

Marginal probabilities give the distribution of **a subset of the variables** (often, just one) of a joint distribution.

Sum/integrate over the variables you don't care about.

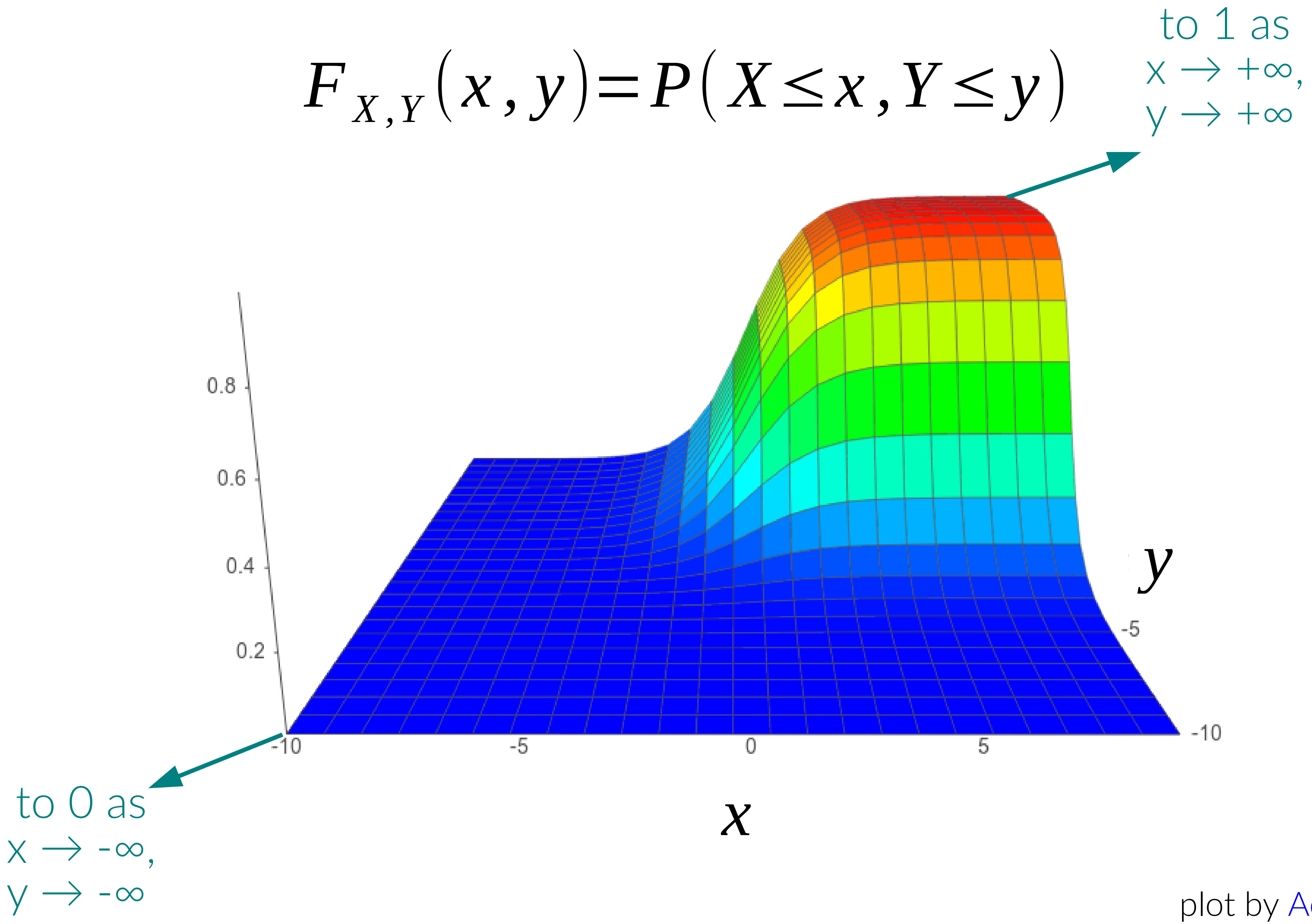
$$p_X(a) = \sum_y p_{X,Y}(a,y)$$

$$f_X(a) = \int_{-\infty}^{\infty} dy\, f_{X,Y}(a,y)$$



Joint cumulative distribution function

$$F_{X,Y}(x,y)=P(X\leq x, Y\leq y)$$



plot by Academo

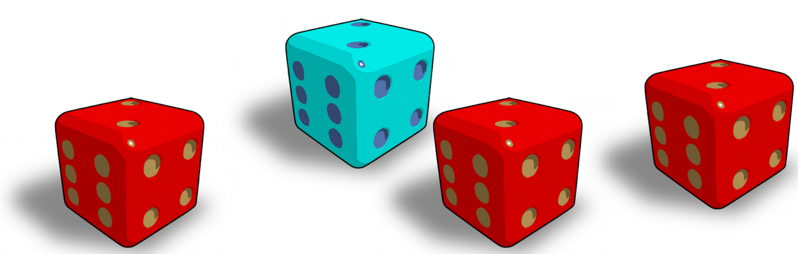
Multinomial random variable

An **multinomial** random variable records the number of times each outcome occurs, when an experiment with multiple outcomes (e.g. die roll) is run multiple times.

$$X_1, \ldots, X_m \sim \mathrm{MN}(n, p_1, p_2, \ldots, p_m)$$

vector!

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m)$$
$$= \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \ldots p_m^{c_m}$$



Independence of discrete random variables

Two **random variables** are **independent** if knowing the value of one tells you nothing about the value of the other (for **all** values!).

$$X \perp Y \text{ iff } \forall x, y:$$

$$P(X=x, Y=y) = P(X=x)\,P(Y=y)$$

- or -

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

Independence of continuous random variables

Two **random variables** are **independent** if knowing the value of one tells you nothing about the value of the other (for **all** values!).

$$X \perp Y \text{ iff } \forall x, y:$$

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

- or -

$$f_{X,Y}(x,y) = g(x) h(y)$$

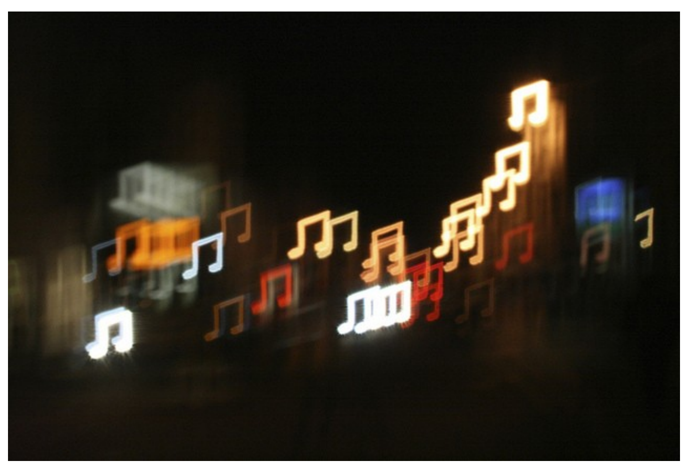
- or -

$$F_{X,Y}(x,y) = F_X(x) F_Y(y)$$

Convolution

A **convolution** is the distribution of the **sum** of two independent random variables.

$$f_{X+Y}(a)=\int_{-\infty}^{\infty} dy\, f_X(a-y)\, f_Y(y)$$



Sum of independent binomials

n flips

m flips

X: number of heads in first *n* flips

Y: number of heads in next *m* flips

$$X \sim \mathrm{Bin}(n, p) \qquad Y \sim \mathrm{Bin}(m, p)$$

$$X + Y \sim \mathrm{Bin}(n+m, p)$$

More generally:

$$\begin{array}{l} X_i \sim \mathrm{Bin}(n_i, p) \\ \text{all } X_i \text{ independent} \end{array} \Rightarrow \sum_{i=1}^{N} X_i \sim \mathrm{Bin}\left(\sum_{i=1}^{N} n_i, p\right)$$

Sum of independent Poissons

λ_1 chips/cookie

X: number of chips in first cookie

$$X \sim \text{Poi}(\lambda_1)$$

λ_2 chips/cookie

Y: number of chips in second cookie

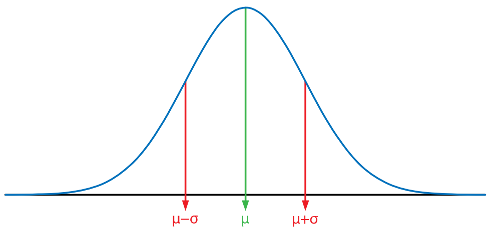
$$Y \sim \text{Poi}(\lambda_2)$$

$$X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$$

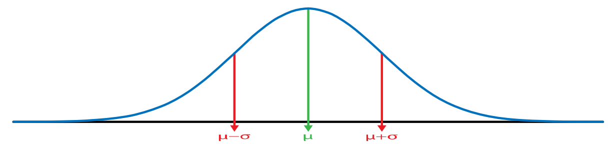
More generally:

$$\begin{array}{l} X_i \sim \text{Poi}(\lambda_i) \\ \text{all } X_i \text{ independent} \end{array} \Rightarrow \sum_{i=1}^{N} X_i \sim \text{Poi}\left(\sum_{i=1}^{N} \lambda_i\right)$$

Sum of independent normals



$X \sim N(\mu_1, \sigma_1^2)$



$Y \sim N(\mu_2, \sigma_2^2)$

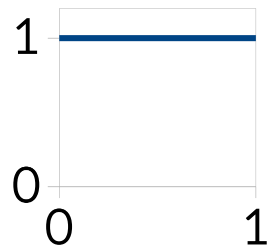
$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

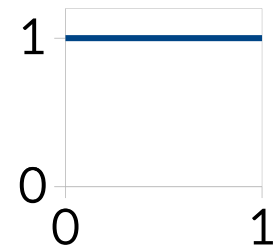
More generally:

$$X_i \sim N(\mu_i, \sigma_i^2) \quad \Rightarrow \quad \sum_{i=1}^{N} X_i \sim N\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$$

all X_i independent

Sum of independent uniforms





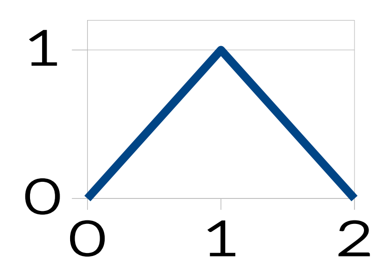
$X \sim \text{Uni}(0,1)$ $\qquad$ $Y \sim \text{Uni}(0,1)$

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} dy\, f_X(a-y) f_Y(y)$$

$$= \int_0^1 dy\, f_X(a-y) \cancelto{1}{f_Y(y)}$$

Case 1: if $0 \le a \le 1$, then we need $0 \le y \le a$ (for $a - y$ to be in [0, 1])
Case 2: if $1 \le a \le 2$, then we need $a - 1 \le y \le 1$

$$= \begin{cases} \int_0^a dy \cdot 1 = a & 0 \le a \le 1 \\ \int_{a-1}^1 dy \cdot 1 = 2 - a & 1 \le a \le 2 \\ 0 & \text{otherwise} \end{cases}$$



Discrete conditional distributions

The value of a random variable, **conditioned on the value of some other random variable**, has a probability distribution.

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

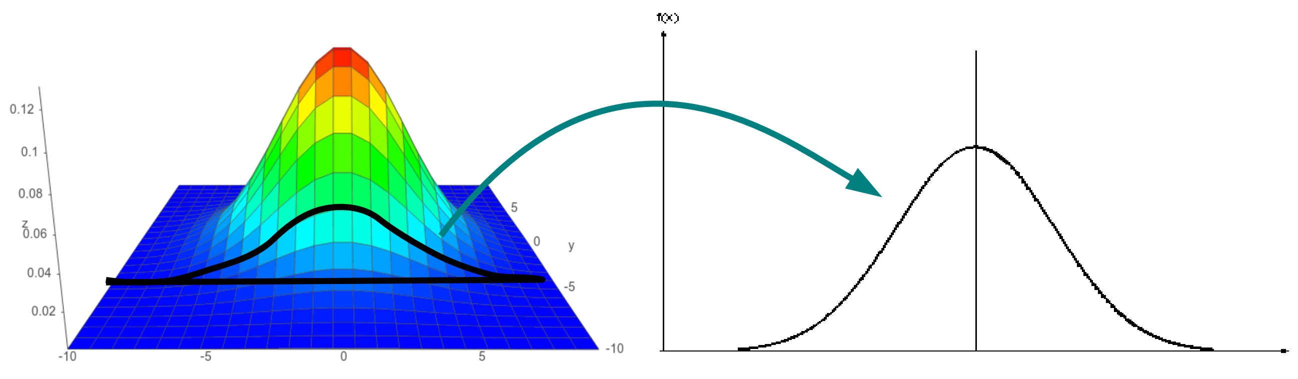
$$= \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

PDF	Single	In a relationship	It's complicated / Other	TOTALS
Freshman	0.00	0.00	0.00	0.00
Sophomore	0.06	0.00	0.00	0.06
Junior	0.19	0.19	0.13	0.50
Senior	0.00	0.00	0.00	0.00
Grad student / Other	0.38	0.06	0.00	0.44
TOTALS	0.63	0.25	0.13	1.00

Continuous conditional distributions

The value of a random variable, **conditioned on the value of some other random variable**, has a probability distribution.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$



Ratios of continuous probabilities

The probability of an exact value for a continuous random variable is 0.

But ratios of these probabilities are still well-defined!

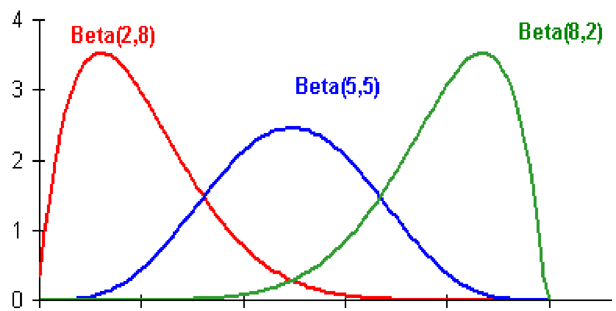
$$\frac{P(X=a)}{P(X=b)}=\frac{f_X(a)}{f_X(b)}$$

Beta random variable

An **beta** random variable models the **probability** of a trial's success, given previous trials. The PDF/CDF let you compute **probabilities of probabilities**!

$$X \sim \text{Beta}(a, b)$$

$$f_X(x) = \begin{cases} C\,x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



Beta: Fact sheet

number of **successes** + 1

$$X \sim \text{Beta}(a, b)$$

probability of success

number of **failures** + 1

PDF: $$f_X(x) = \begin{cases} C\,x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{a}{a+b}$$

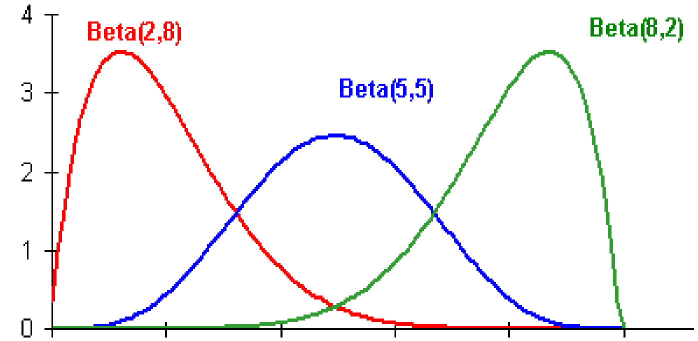
variance: $$\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Subjective priors

X | A ~ Beta(a + 1, N − a + 1)
"posterior"

X ~ Beta(1, 1)
"prior"

$$f_{X|A}(x|a) = \frac{P(A=a|X=x)\, f_X(x)}{P(A=a)}$$



How did we decide on
Beta(1, 1) for the prior?

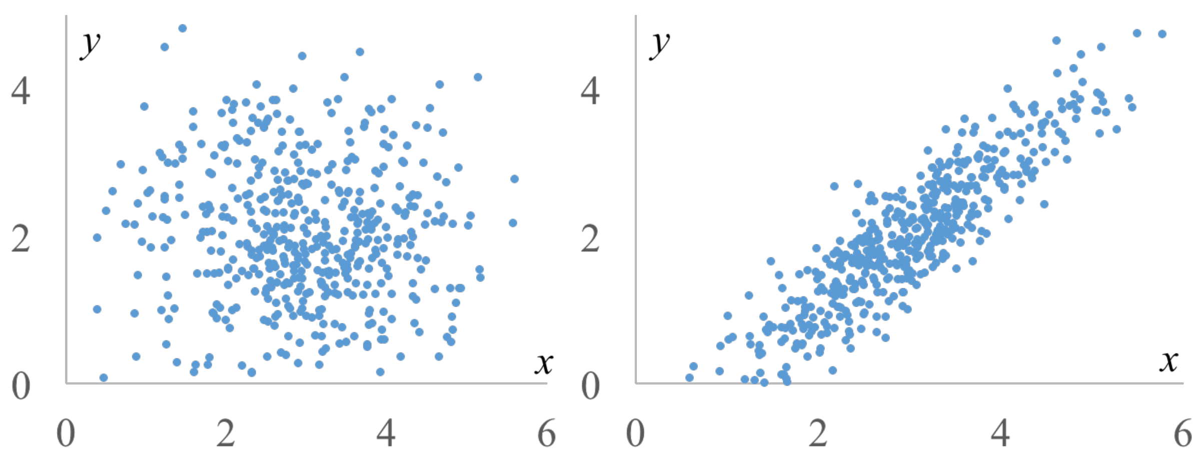
Beta(1, 1): "we haven't seen any rolls yet."
Beta(4, 1): "we've seen 3 sixes and 0 non-sixes."
Beta(2, 6): "we've seen 1 six and 5 non-sixes."

Beta prior = "imaginary" previous trials

Covariance

The **covariance** of two variables is a measure of how much they **vary together**.

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$



Expectation of a product

If two random variables are **independent**, then the **expectation of their product** equals the **product of their expectations.**

$$X \perp Y \quad \Rightarrow$$

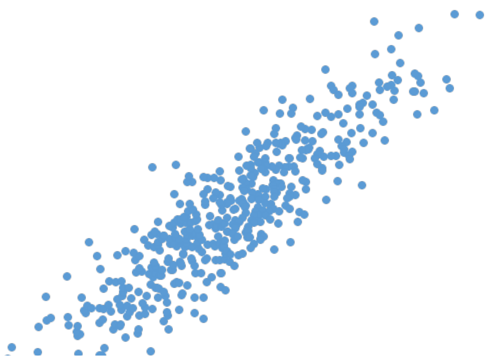
$$E[XY] = E[X]E[Y]$$

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Correlation

The **correlation** of two variables is a measure of the **linear dependence** between them, scaled to always take on values between -1 and 1.

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$



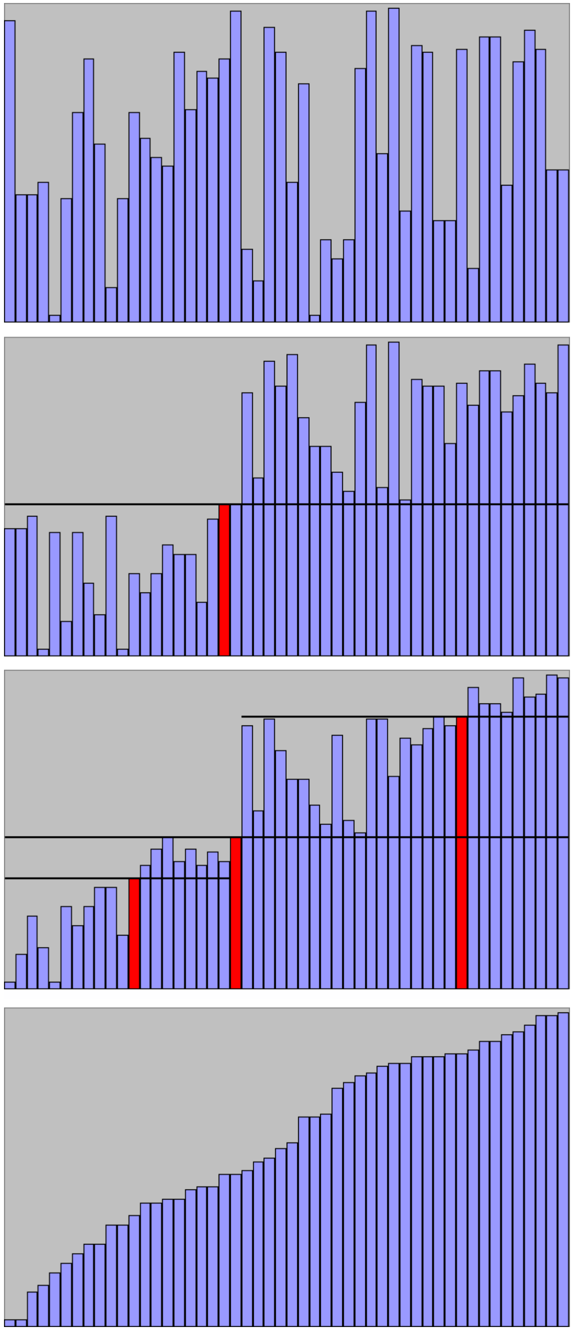
Conditional expectation

One can compute the **expectation** of a random variable while **conditioning** on the values of other random variables.

$$E[X|Y=y]=\sum_x x\, p_{X|Y}(x|y)$$

$$E[X|Y=y]=\int_{-\infty}^{\infty} dx\, x\, f_{X|Y}(x|y)$$

Quicksort



You’ve been told Quicksort is O(n log n), “average case”.

Now you get to find out why!

Quicksort's ordinary life

Let X = number of comparisons to the pivot.
What is E[X]? expected number of events = indicator variables!

1	2	3	4	5	6	7	8

$Y_1 \quad Y_2 \qquad \ldots \qquad Y_n$

Define $Y_1 \ldots Y_n$ = elements in sorted order.

Indicator variables I_{ab} = 1 if Y_a and Y_b are ever compared.

$$E[X] = E\left[\boxed{\sum_{a=1}^{n-1} \sum_{b=a+1}^{n}} I_{ab}\right] = \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} E[I_{ab}]$$

unique pairs

$$= \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} P\left(Y_a \text{ and } Y_b \text{ ever compared}\right)$$

Shall I compare thee to…

P(Y_a and Y_b ever compared) = ?

Case 1:

		Y_a				Y_b	
1	2	3	4	5	6	7	8

yes!

Case 2:

		Y_a				Y_b	
1	2	3	4	5	6	7	8

no!

Case 3:

		Y_a				Y_b	
1	2	3	4	5	6	7	8

recursive call

$$\therefore \quad P\left(Y_a \text{ and } Y_b \text{ ever compared}\right) = \frac{2}{b-a+1}$$

The home stretch

$$E[X]=\sum_{a=1}^{n-1}\sum_{b=a+1}^{n} P(Y_a \text{ and } Y_b \text{ ever compared})$$

$$=\sum_{a=1}^{n-1}\sum_{b=a+1}^{n}\frac{2}{b-a+1}$$

$$\sum_{b=a+1}^{n}\frac{2}{b-a+1}\approx\int_{b=a+1}^{n} db\,\frac{2}{b-a+1}$$

$$=\left[2\ln(b-a+1)\right]_{b=a+1}^{n}$$

$$=2\ln(n-a+1)-2\ln 2$$

$$\approx 2\ln(n-a+1)$$

for large n

$$\approx\sum_{a=1}^{n-1}2\ln(n-a+1)$$

$$\approx\int_{a=1}^{n-1} da\,2\ln(n-a+1)$$

$$=-2\int_{y=n}^{2} dy\ln y$$

$$u=\ln y \qquad v=y$$

$$du=\frac{1}{y}dy \qquad dv=dy$$

$$\int u\,dv=u\,v-\int v\,du$$

$$\int \ln y\,dy=y\ln y-\int \cancel{y}\frac{1}{\cancel{y}}dy$$

$$=y\ln y-y+C$$

$$=-2\left[y\ln y-y\right]_{y=n}^{2}$$

$$=-2[\cancel{(2\ln 2-2)}-(n\ln n-\cancel{n})]$$

constants

lower-order term

$$=O(n\ln n)$$

Variance of a sum

The **variance of a sum** of random variables is equal to the **sum of pairwise covariances** (*including* variances and double-counted pairs).

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)$$

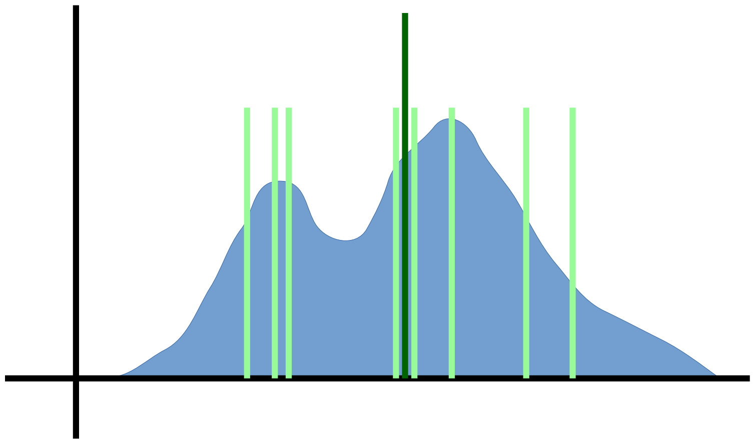
$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \boxed{2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)}$$

note: independent ⇒ Cov = 0

Sample mean

A **sample mean** is an **average** of random variables drawn (usually independently) from the same **distribution**.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

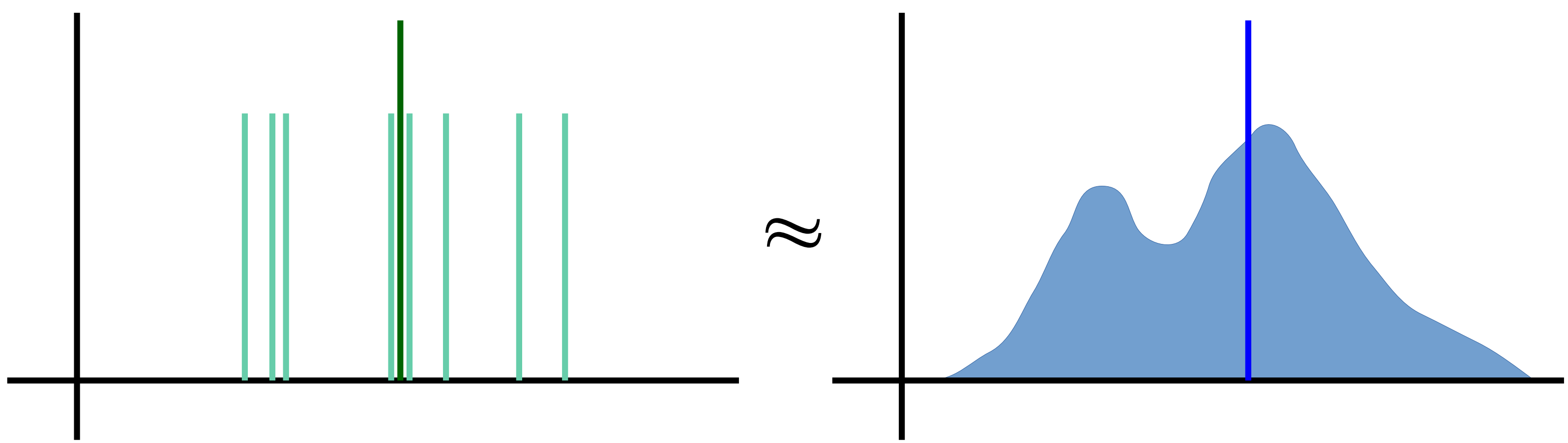


Parameter estimation

Sometimes we **don’t know** things like the expectation and variance of a distribution; we have to **estimate** them from incomplete information.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

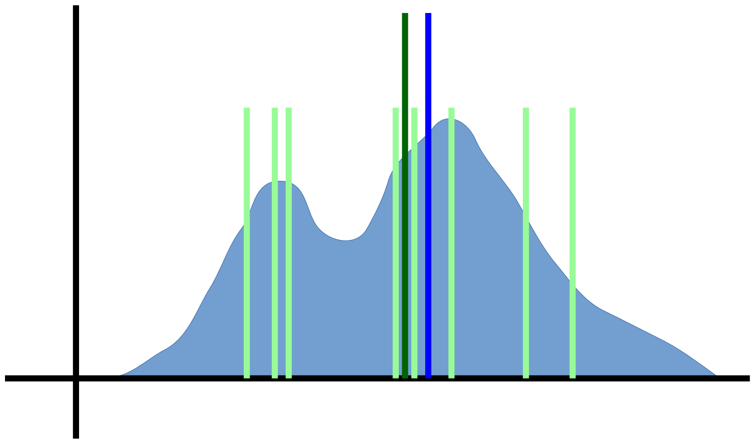
$$\hat{\theta} = \arg\max_{\theta} \mathrm{LL}(\theta)$$



Unbiased estimator

An **unbiased estimator** is a random variable that has **expectation** equal to the quantity you are estimating.

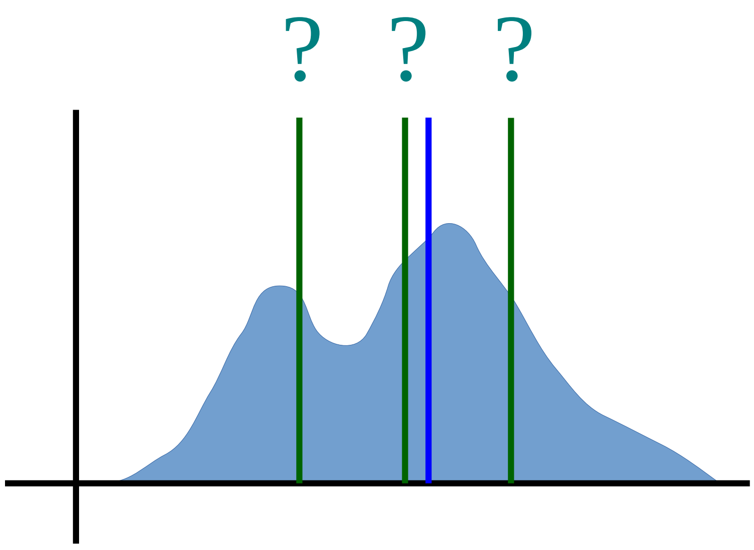
$$E[\bar{X}] = \mu = E[X_i]$$



Variance of the sample mean

The **sample mean** is a random variable; it can differ among samples. That means it has a **variance**.

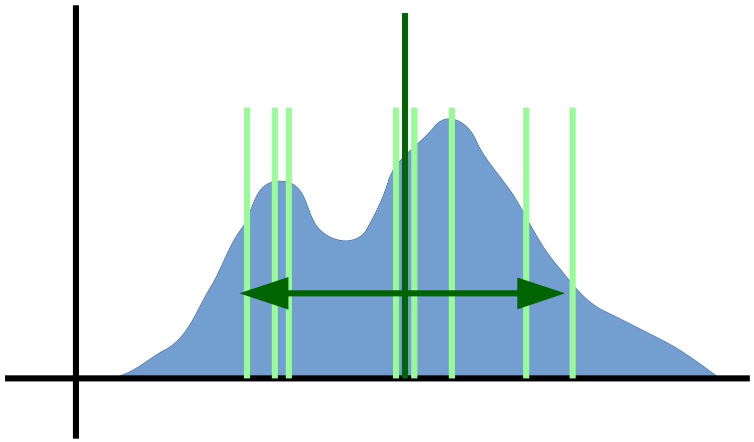
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$



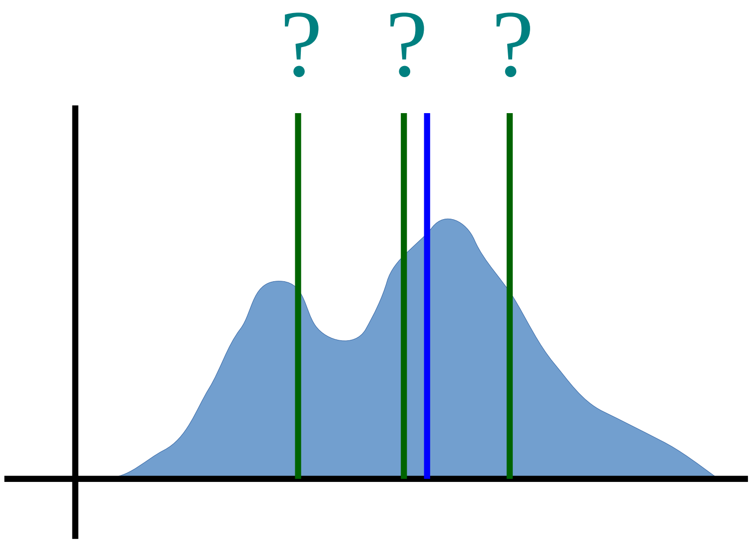
Sample variance

Samples can be used to **estimate the variance** of the <u>original</u> distribution.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$



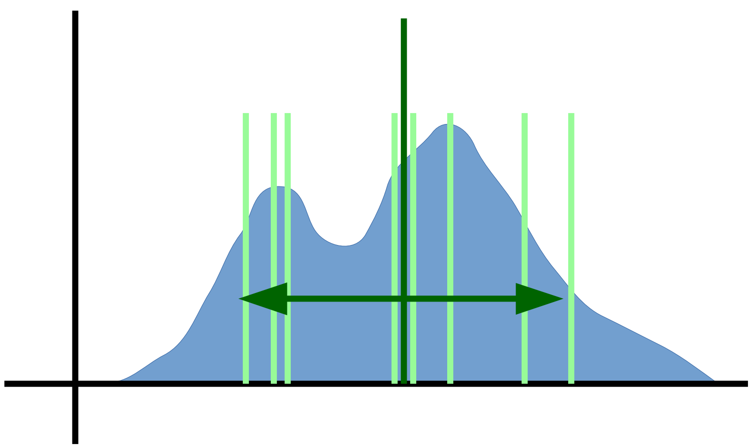
Variance of the sample mean



- Is a single number
- Shrinks with number of samples $\left(=\frac{\sigma^2}{n}\right)$
- Measures the stability of an estimate

vs.

Sample variance

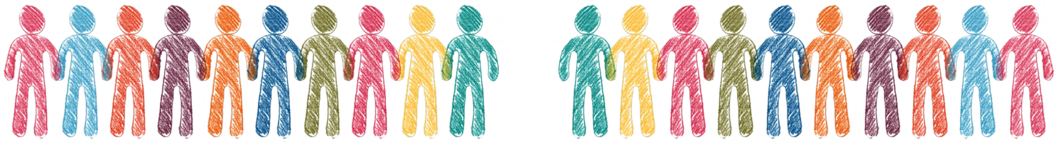


- Is a random variable
- Constant with number of samples $\left(\approx\sigma^2\right)$
- Is an estimate (of a variance) itself

p-values

A ***p*-value** gives the probability of an extreme result, assuming that any extremeness is due to chance.

$$p = P\left(|\bar{X} - \mu| > d \mid H_0\right)$$



Bootstrapping

Bootstrapping allows you to compute complicated statistics from samples using simulation.

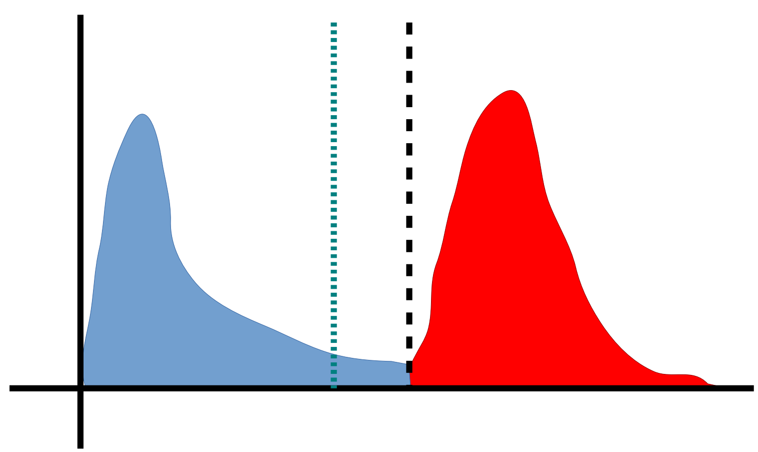
Bootstrap for p-values

```
def pvalue_bootstrap(sample1, sample2):
    n = len(sample1)
    m = len(sample2)
    observed_diff = abs(np.mean(sample2) -
                        np.mean(sample1))
    universal_pmf = sample1 + sample2
    count_extreme = 0
    for i in range(10000):
        resample1 = np.random.choice(universal_pmf, n)
        resample2 = np.random.choice(universal_pmf, m)
        new_diff = abs(np.mean(resample2) -
                       np.mean(resample1))
        if new_diff >= observed_diff:
            count_extreme += 1
    return count_extreme / 10000.
```

Markov's inequality

Knowing the **expectation** of a **non-negative** random variable lets you bound the probability of **high** values for that variable.

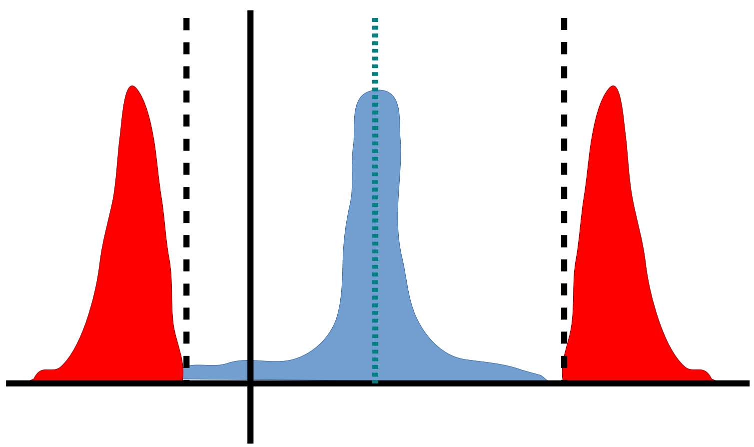
$$X \geq 0 \quad \Rightarrow \quad P(X \geq a) \leq \frac{E[X]}{a}$$



Chebyshev's inequality

Knowing the expectation **and variance** of a random variable lets you bound the probability of **extreme** values for that variable.

$$P\left(|X-\mu|\geq k\right)\leq\frac{\sigma^2}{k^2}$$



One-sided Chebyshev's inequality

$$P\left(X \geq \mu + a\right) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

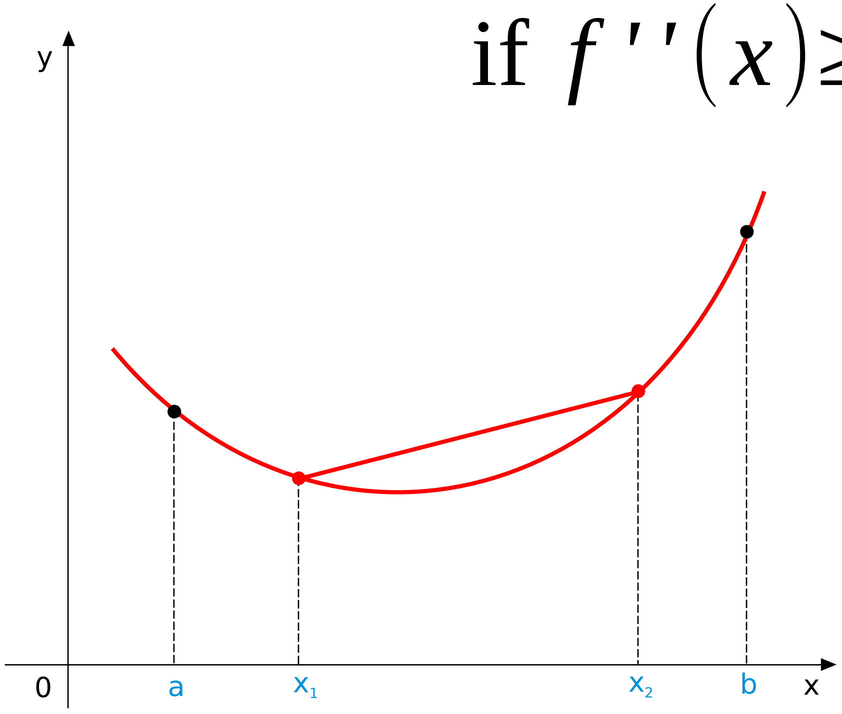
$$P\left(X \leq \mu - a\right) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Jensen's inequality

The expectation of a **convex function** of a random variable can't be less than the value of the function applied to the expectation.

$$E[f(X)] \geq f(E[X])$$

$$\text{if } f''(x) \geq 0$$



Law of large numbers

A **sample mean** will **converge to the true mean** if you take a large enough sample.

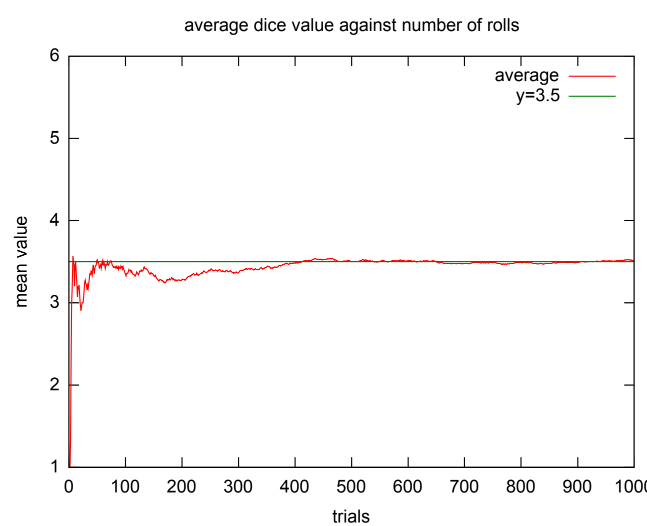
$$\lim_{n \to \infty} P\left(\left|\bar{X} - \mu\right| \geq \varepsilon\right) = 0$$

$$P\left(\lim_{n \to \infty}\left(\bar{X}\right) = \mu\right) = 1$$

Consistent estimator

An **consistent estimator** is a random variable that has a **limit** (as number of samples gets large) equal to the quantity you are estimating.

$$\lim_{n \to \infty} P\left(|\hat{\theta} - \theta| < \varepsilon\right) = 1$$



Review: Central limit theorem

Sums and **averages** of IID random variables are **normally distributed**.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Y = n\bar{X} = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right)$$

Easily-confused principles

Constant multiple of a normal	Sum of identical normals	CLT
$X \sim N(\mu, \sigma^2)$	$X_i \sim N(\mu, \sigma^2)$	$X_i \sim ???$
	(independent & identical)	(independent & identical)
⬇	⬇	⬇
$nX \sim N(n\mu, \boxed{n^2}\sigma^2)$	$\sum_{i=1}^{n} X_i \sim N(n\mu, \boxed{n}\sigma^2)$	$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$
	(exactly)	(approximately, for large *n*)

Parameters

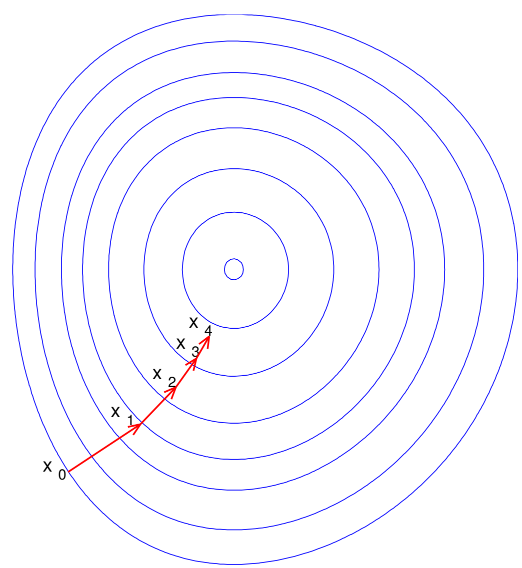
$$\boldsymbol{\theta}$$

$$X \sim \begin{cases} \text{Ber}(p) & \theta = p \\ \text{Poi}(\lambda) & \theta = \lambda \\ \text{Uni}(a, b) & \theta = [a, b] \\ \text{N}(\mu, \sigma^2) & \theta = [\mu, \sigma^2] \end{cases}$$

Maximum likelihood estimation

Choose parameters that **maximize** the likelihood (**joint probability given parameters**) of the example data.

$$\hat{\theta} = \arg\max_{\theta} LL(\theta)$$



How to: MLE

1. Compute the likelihood.

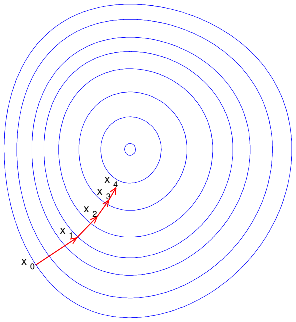
$$L(\theta) = P(X_1, \ldots, X_m | \theta)$$

2. Take its log.

$$LL(\theta) = \log L(\theta)$$

3. Maximize this as a function of the parameters.

$$\frac{d}{d\theta} LL(\theta) = 0$$



Maximum likelihood for Bernoulli

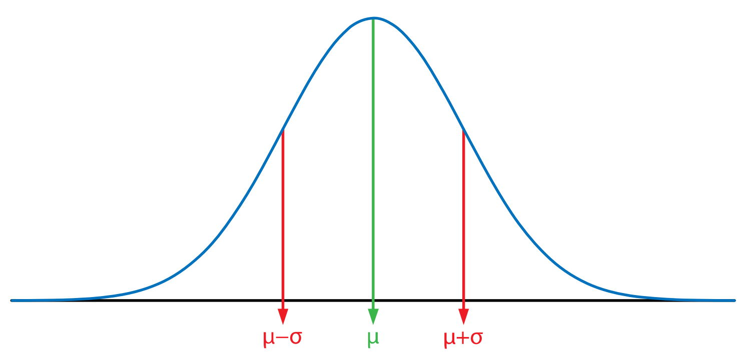
The maximum likelihood *p* for **Bernoulli** random variables is the **sample mean**.

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$$

Maximum likelihood for normal

The maximum likelihood μ for **normal** random variables is the **sample mean**, and the maximum likelihood σ^2 is the "uncorrected" **mean square deviation**.

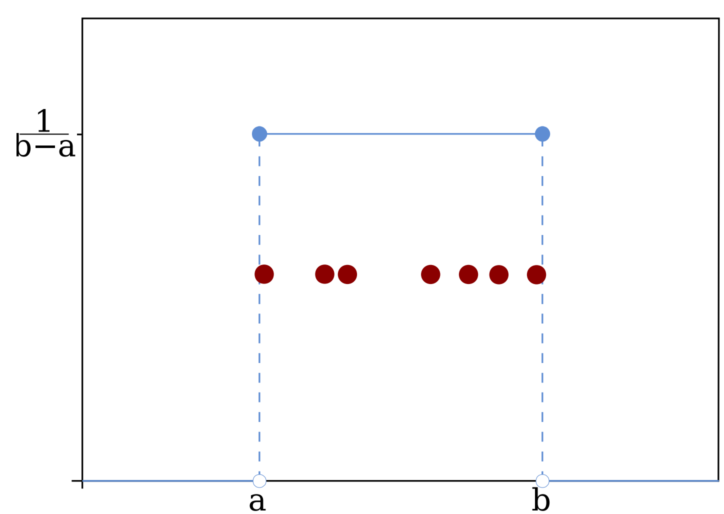
$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i \qquad \hat{\sigma}^2 = \frac{1}{\boxed{m}}\sum_{i=1}^{m} (X_i - \hat{\mu})^2$$



Maximum likelihood for uniform

The maximum likelihood *a* and *b* for **uniform** random variables are the **minimum and maximum** of the data.

$$\hat{a} = \min_i X_i \qquad \hat{b} = \max_i X_i$$



Maximum a posteriori estimation

Choose the **most likely** parameters **given the example data**. You'll need a **prior probability** over the parameters.

$$\hat{\theta} = \arg\max_{\theta} P(\theta | X_1, \ldots, X_n)$$
$$= \arg\max_{\theta} \left[LL(\theta) + \log P(\theta) \right]$$

Laplace smoothing

Also known as **add-one** smoothing: assume you've seen one "imaginary" occurrence of each possible outcome.

$$p_i = \frac{\#(X=i)+1}{n+m}$$

or: "add-*k*" smoothing (if you believe equally likely is more plausible)

$$p_i = \frac{\#(X=i)+k}{n+mk}$$

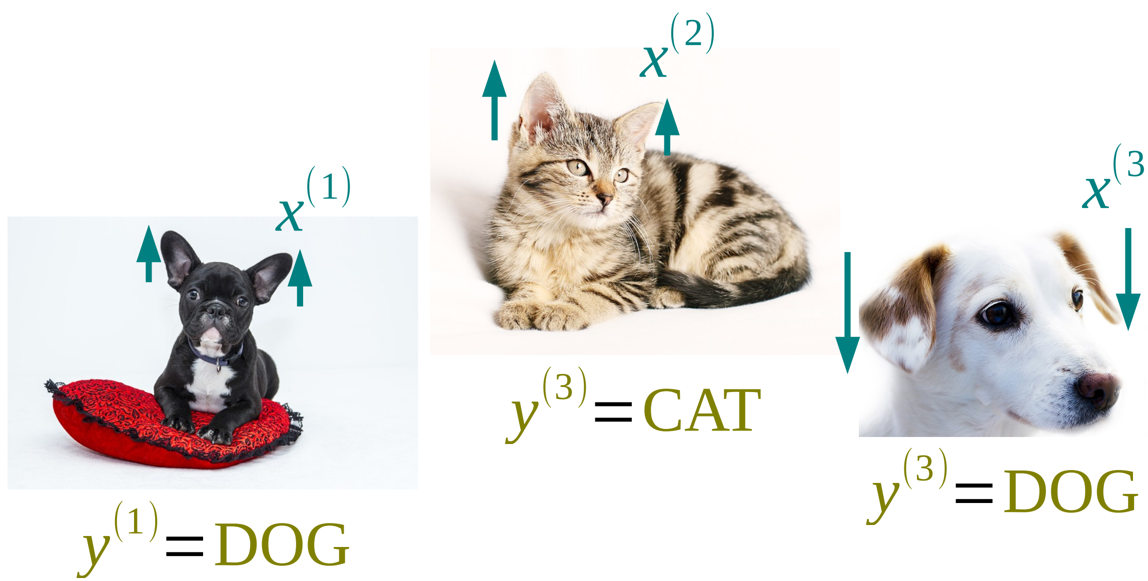
Parameter priors

$$X \sim \begin{cases} \mathrm{Ber}(p) & p \sim \mathrm{Beta}(a, b) \\ \mathrm{Bin}(n, p) & p \sim \mathrm{Beta}(a, b) \\ \mathrm{MN}(\boldsymbol{p}) & \boldsymbol{p} \sim \mathrm{Dir}(\boldsymbol{a}) \\ \mathrm{Poi}(\lambda) & \lambda \sim \mathrm{Gamma}(k, \theta) \\ \mathrm{Exp}(\lambda) & \lambda \sim \mathrm{Gamma}(k, \theta) \\ \mathrm{N}(\mu, \sigma^2) & \mu \sim \mathrm{N}(\mu', \sigma'^2) \\ & \sigma^2 \sim \mathrm{InvGamma}(\alpha, \beta) \end{cases}$$

Classification

The most basic machine learning task:
predict a **label** from a vector of **features**.

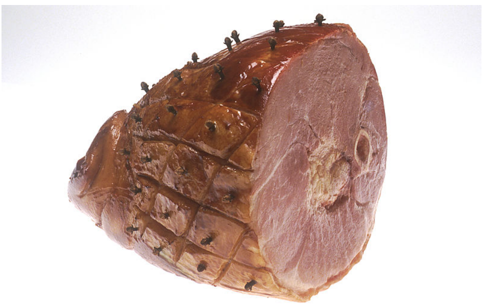
$$\hat{y} = \arg\max_{y} P\left(Y = y \,|\, \vec{X} = \vec{x}\right)$$



Naïve Bayes

A classification algorithm using the assumption that features are **conditionally independent** given the label.

$$\hat{y} = \arg\max_{y} \hat{P}(Y = y) \prod_{j} \hat{P}(X_j = x_j | Y = y)$$



images: (left) Virginia State Parks; (right) Renee Comet

Three secret ingredients

1. Maximum likelihood or maximum a posteriori for conditional probabilities.

$$\hat{P}(X_j = x_j | Y = y) = \frac{\#(X_j = x_j, Y = y)[+1]}{\#(Y = y)[+2]}$$

2. "Naïve Bayes assumption": features are independent conditioned on the label.

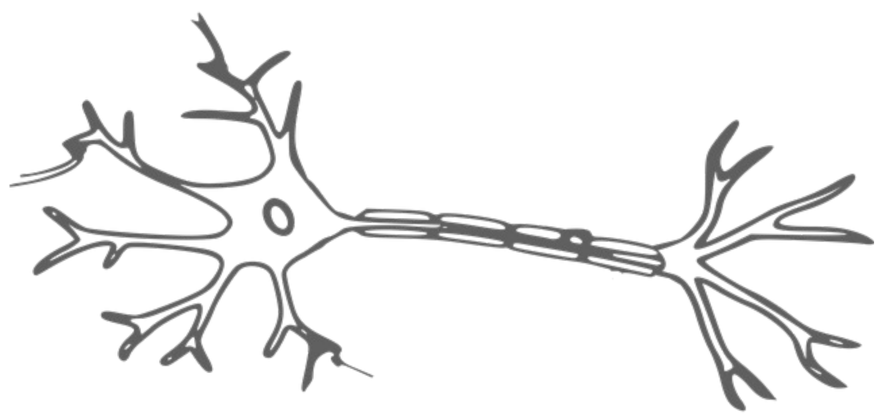
$$\hat{P}(\vec{X} = \vec{x} | Y = y) = \prod_j \hat{P}(X_j = x_j | Y = y)$$

3. (Take logs for numerical stability.)

Logistic regression

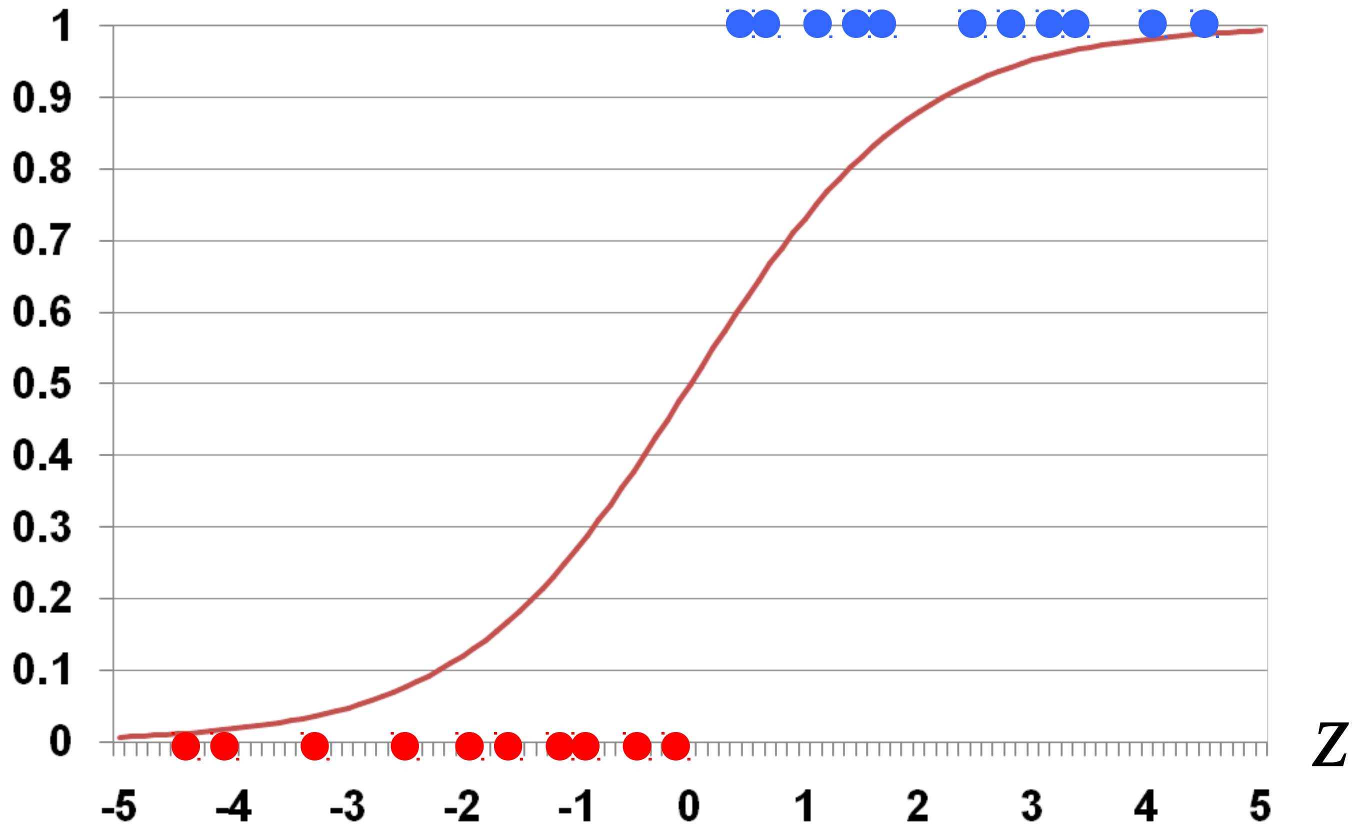
A classification algorithm using the assumption that **log odds** are a linear function of the features.

$$\hat{y} = \arg\max_{y} \frac{1}{1 + e^{-\vec{\theta}^T \vec{x}}}$$



Predicting 0/1 with the logistic

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Logistic regression: Pseudocode

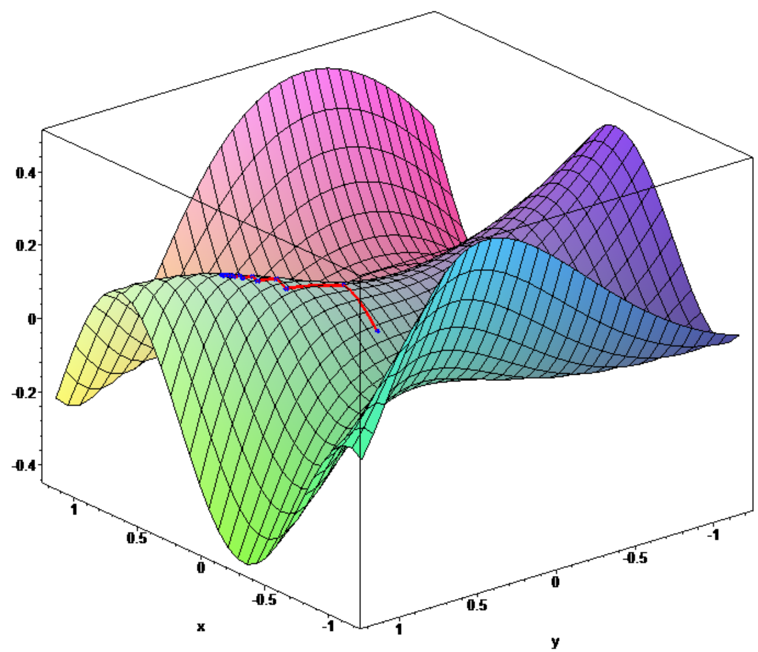
```
initialize: θ = [0, 0, …, 0]   (m elements)

repeat many times:
    gradient = [0, 0, …, 0]   (m elements)

    for each training example (x^(i), y^(i)):
        for j = 0 to m:

            gradient[j] += [y^(i) − σ(θ^T x^(i))] x_j^(i)

    for j = 0 to m:
        θ[j] += η * gradient[j]

return θ
```

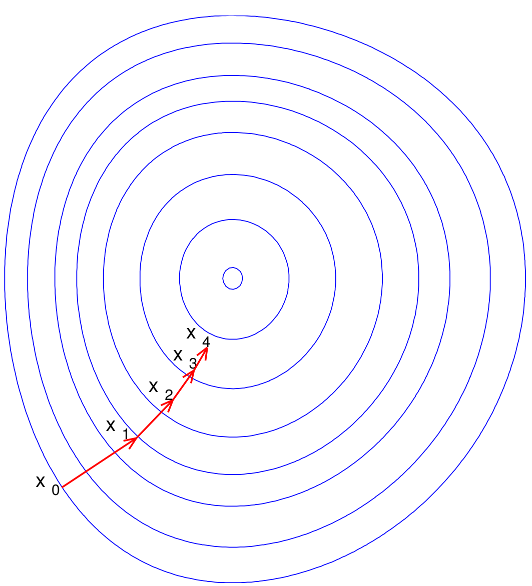
$$\text{gradient}[j] \mathrel{+}= \left[y^{(i)} - \sigma\left(\vec{\theta}^T \vec{x}^{(i)}\right)\right] x_j^{(i)}$$

Gradient ascent

An algorithm for computing an **arg max** by taking small steps **uphill** (i.e., in the direction of the **gradient** of the function).

$$\vec{\theta} \leftarrow \vec{\theta} + \eta \cdot \nabla_{\vec{\theta}} f(\vec{\theta})$$

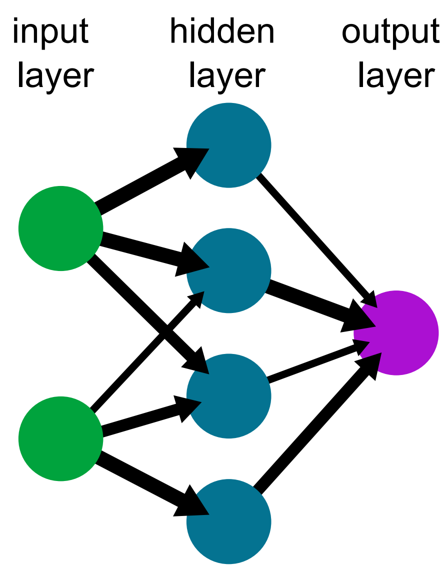




Feedforward neural network

An algorithm for classification or regression that uses **layers of logistic regressions** to **discover its own features.**

$$\hat{y} = \sigma\left(\theta^{(\hat{y})}\sigma\left(\theta^{(h)}\vec{x}\right)\right)$$




Keep in touch!