Probabilities 101

SLIDES COURTESY OF ANDREW MOORE, TOM MITCHELL, AARTI SINGH

Terminology

- Event: Every possible outcome of an experiment
- Sample Space: The set of all possible outcomes of an experiment (the set of all events).
- Example: Rolling the dice
  - Every possible outcome is an event
  - Sample space: {1, 2, 3, 4, 5, 6}
- Example: Toss a coin
  - Every possible outcome is an event
  - Sample space: {heads, tails}.

Binary Random Variables

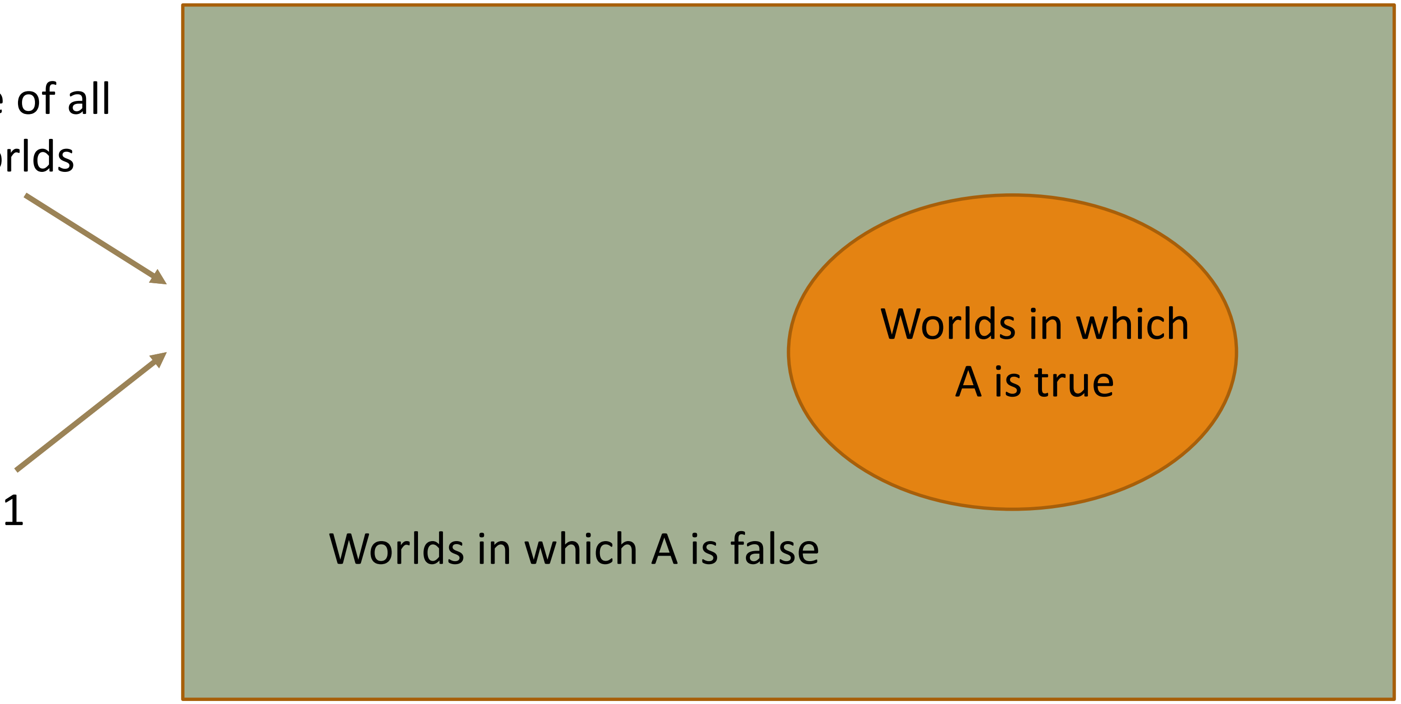
- A is a Boolean-random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs

- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You have Ebola

Visualizing A

Event space of all possible worlds

This area is 1



P(A) = Area of the reddish oval

Kolmogorov Axioms

Kolmogorov Axioms

Probability of an event A is a number assigned to this event such that:

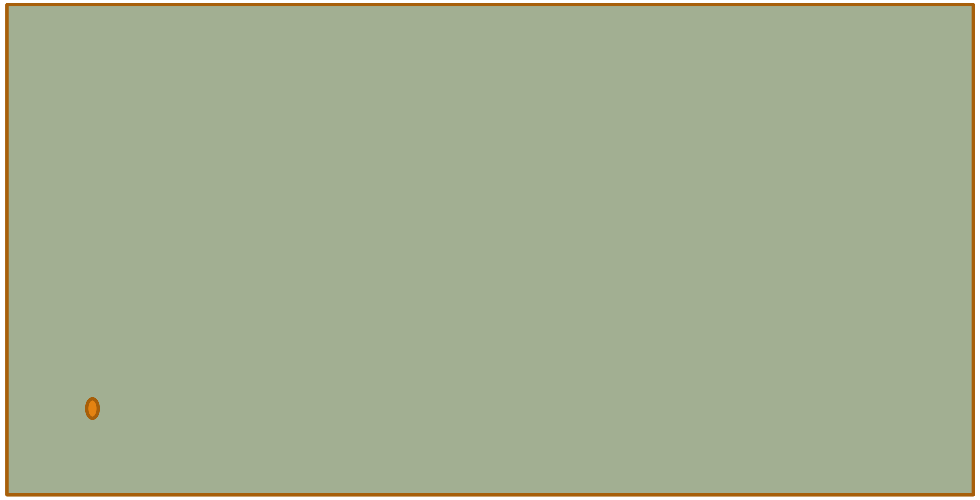
1. $0 \leq P(A) \leq 1$ –All probabilities are between 0 and 1
2. $P(\emptyset) = 0$ "no outcome" has zero probability
3. $P(S) = 1$ some outcome is bound to occur
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ probability of the union equals sum of probabilities minus probability of the intersection.

Probability of an event A is a number assigned to this event such that:

1. $0 \leq P(A) \leq 1$ –All probabilities are between 0 and 1
2. $P(\emptyset) = 0$ "no outcome" has zero probability
3. $P(S) = 1$ some outcome is bound to occur
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ probability of the union equals sum of probabilities minus probability of the intersection.

The area of A cannot be smaller than 0

And a zero area would mean no world could ever have A true.



Probability of an event A is a number assigned to this event such that:

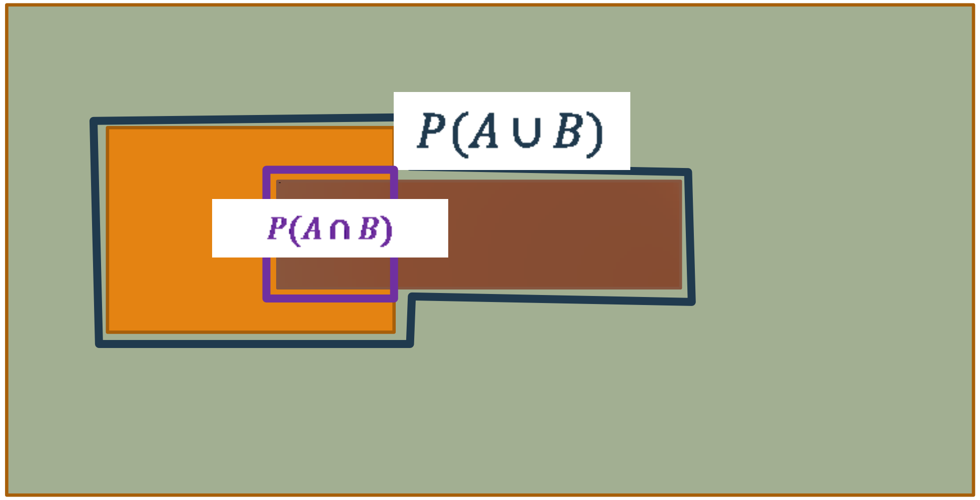
1. $0 \leq P(A) \leq 1$ –All probabilities are between 0 and 1
2. $P(\emptyset) = 0$ "no outcome" has zero probability
3. $P(S) = 1$ some outcome is bound to occur
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ probability of the union equals sum of probabilities minus probability of the intersection.

The area of A cannot be larger than 1

And an area of 1 would mean all worlds will have A true.

Probability of an event A is a number assigned to this event such that:

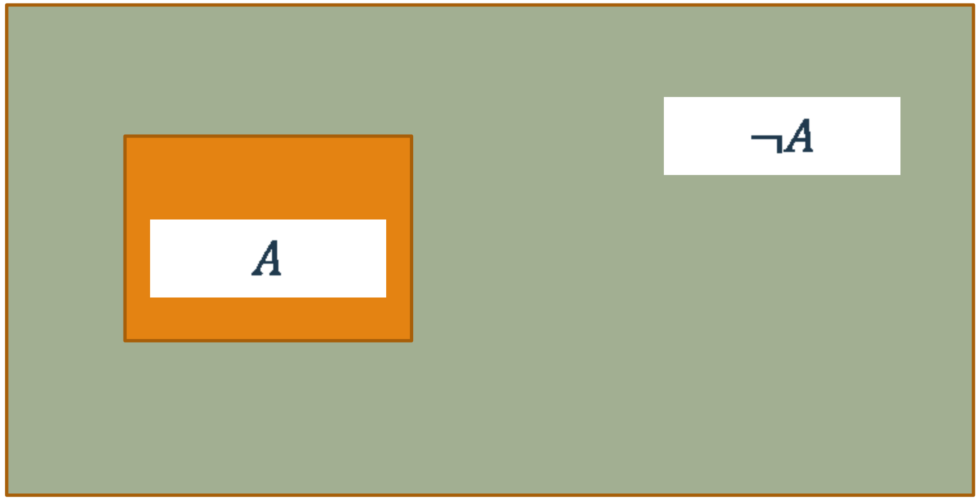
1. $0 \leq P(A) \leq 1$ All probabilities are between 0 and 1
2. $P(\emptyset) = 0$ "no outcome" has zero probability
3. $P(S) = 1$ some outcome is bound to occur
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ probability of the union equals sum of probabilities minus probability of the intersection.



Theorems from the axioms

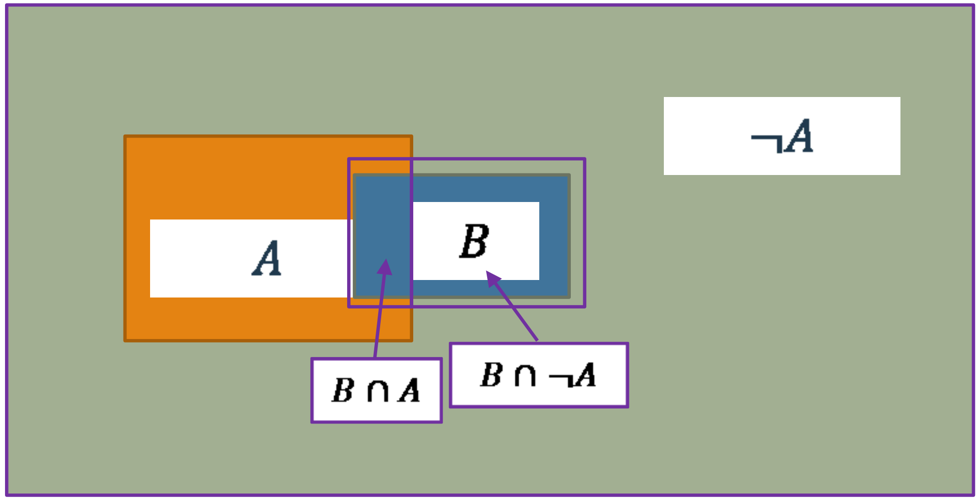
- $P(\neg A) = 1 - P(A)$
- How?
  - $P(A \cup \neg A) = P(S) = 1$
  - $P(A \cap \neg A) = P(\emptyset) = 0$
  - $P(A \cap A) = P(A) + P(\neg A) - P(A \cap \neg A)$

1 0



Theorems from the axioms

- $P(B) = P(B \cap A) + P(B \cap \neg A)$
- How?
  - Try it at home



Multivalued Random Variables

- Suppose A can take more than two values
- A is a random variable with arity k if it can take on exactly one value of $\{v_1, v_2, \dots, v_k\}$.
- Thus..

$$P(A = v_i \cap A = v_j) = 0, \qquad if\ i \neq \mathrm{j}$$
$$P(A = v_1 \cup A = v_2 \cup \cdots \cap A = v_k) = 1$$

Facts about multivalued random variables

- Using

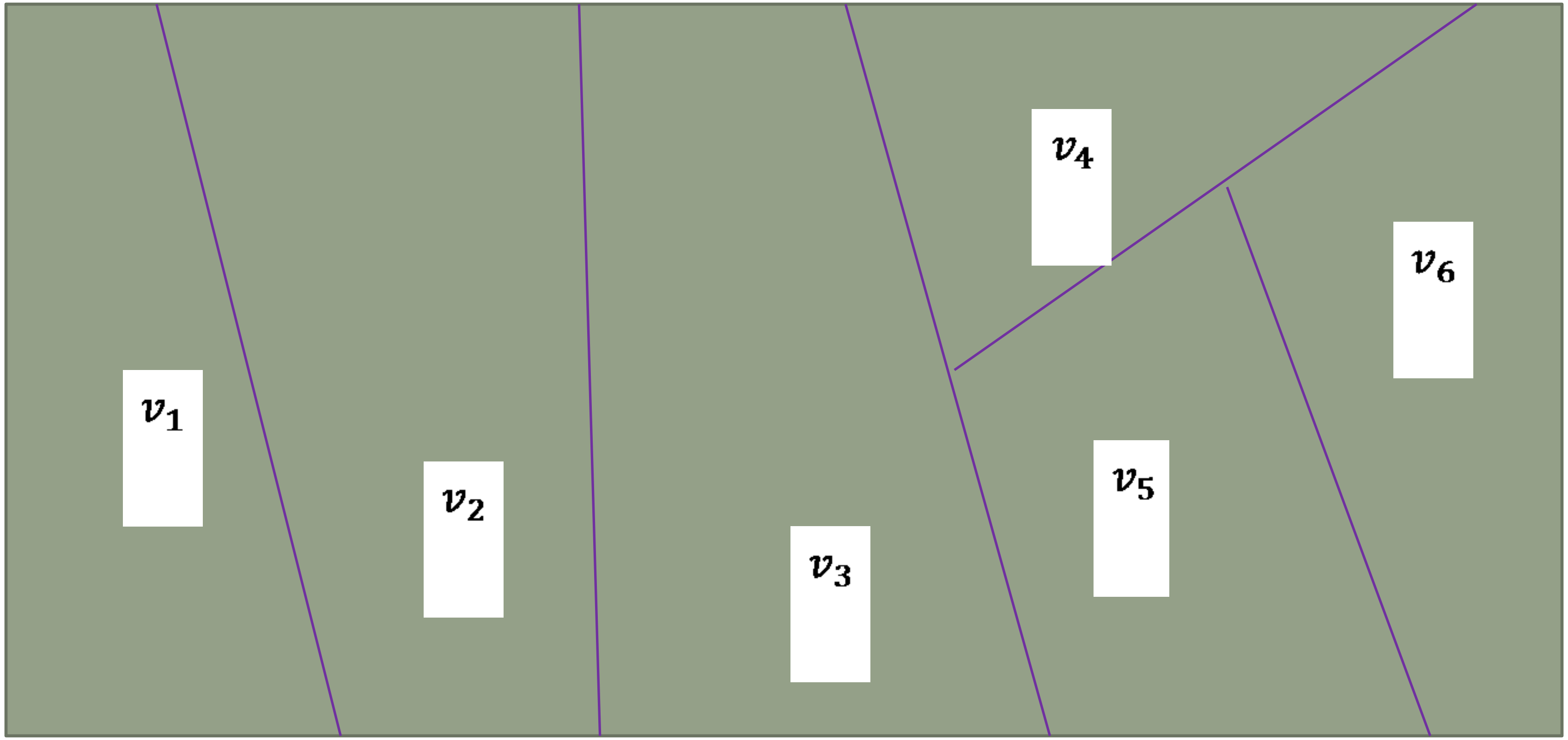
$$P(A = v_i \cap A = v_j) = 0, \qquad if\ i \neq j$$
$$P(A = v_1 \cup A = v_2 \cup \cdots \cap A = v_k) = 1$$

- And the axioms of probability we can prove:

  - $P(A = v_1 \cup A = v_2 \cup \cdots \cap A = v_k) = \sum_{i=1}^{k} P(A = v_i)$
  - Therefore $\sum_{i=1}^{k} P(A = v_i)$ =1
  - Also $P(B) = \sum_{i=1}^{k} P(B \cap A = v_i)$

Elementary probability in pictures

$$\sum_{i=1}^{k} P(B \cap A = v_i)$$



Elementary probability in pictures

$$\sum_{i=1}^{k} P(B \cap A = v_i)$$

$B \cap A = v_1$

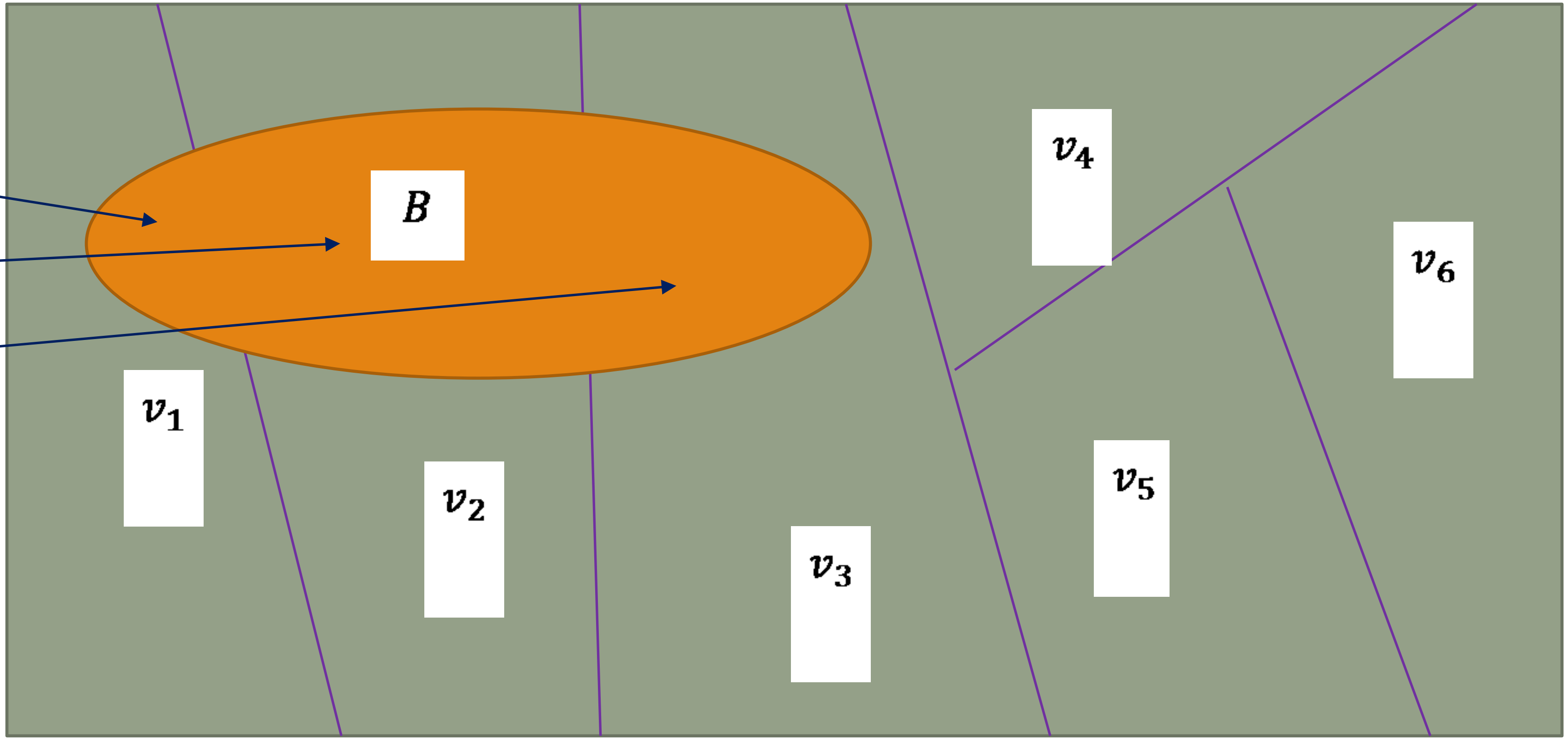
$B \cap A = v_2$

$B \cap A = v_3$

$B \cap A = v_4$

$B \cap A = v_5$

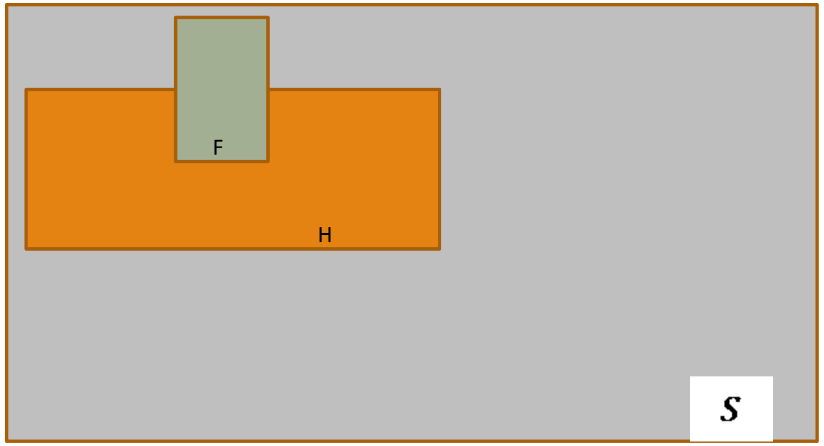
$B \cap A = v_6$



Independence

- A and B are independent events if $P(A|B) = P(A) \times P(B)$
- Outcome of A has no effect on the outcome of B (and vice versa)
- Examples
  - Possibility of tossing heads and then tails is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.
  - …

Conditional Probability

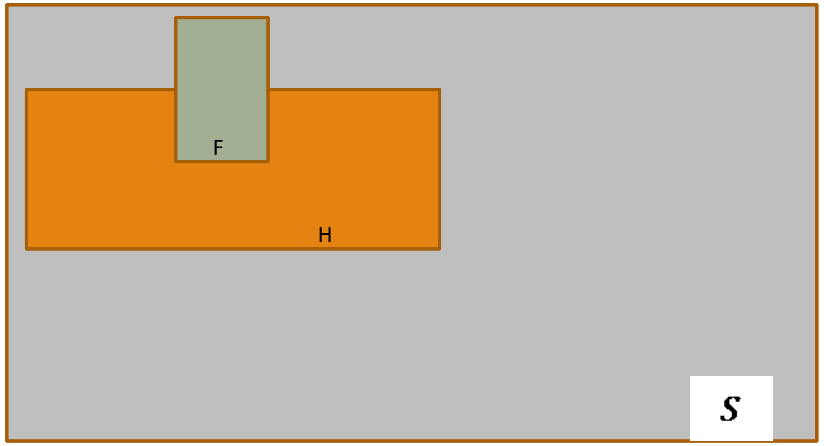


F = “Coming down with a flu”
H = “Have a headache”
P(H)= 1/10
P(F) = 1/40
P(H|F) =1/2

$P(A|B)$: Fraction of worlds where B is true where A is also true.

“Headaches are rare and flu is rarer, but if you’ re coming down with a flu there’ s a 50-50 chance you’ ll have a headache.”

Conditional Probability



F = “Coming down with a flu”
H = “Have a headache”
P(H)= 1/10
P(F) = 1/40
P(H|F) =1/2

$P(Headache|Flu)$: Fraction of flu -infectd worlds where you also have a headache =

$$= \frac{\#worlds\ with\ flu\ and\ headache}{\#worlds\ with\ u} =$$

$$=$$

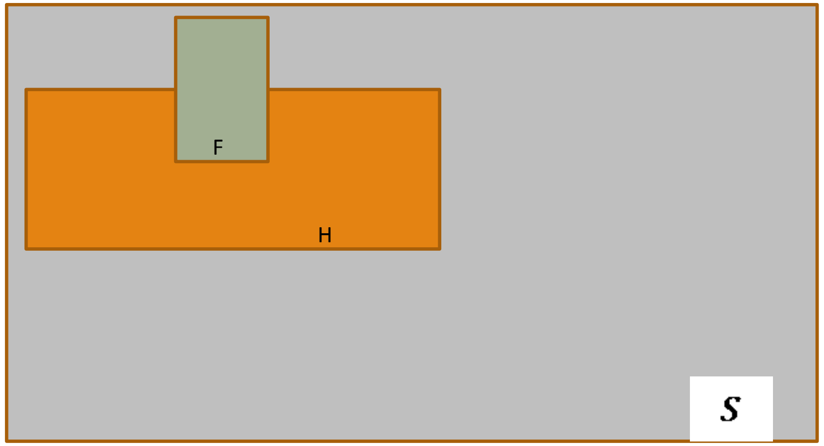
$$= \frac{area\ of\ F\ and\ H\ region}{area\ of H\ region} =$$

$$= \frac{P(H \cap F)}{P(H)}$$

Conditional Probability

- Definition of conditional Probability
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Corollary: The Chain Rule
  - $P(A \mid B) = P(A|B)P(B)$
- if A, B independent:
  - $P(A \mid B) = P(A) \times P(B) \Rightarrow P(A|B) = P(A)$

Conditional Probability



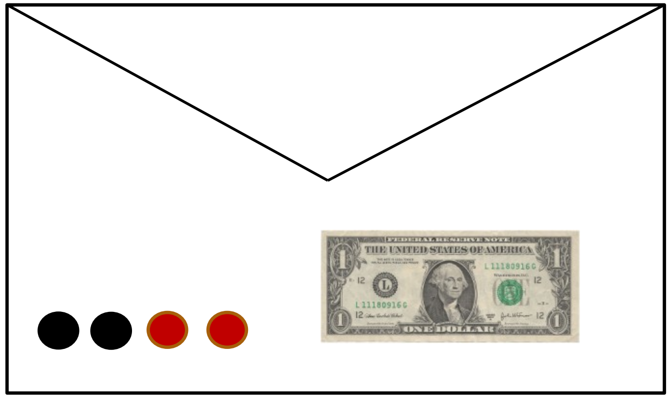
F = “Coming down with a flu”
H = “Have a headache”
P(H)= 1/10
P(F) = 1/40
P(H|F) =1/2

One day you wake up with a headache. You think: “Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with a flu”

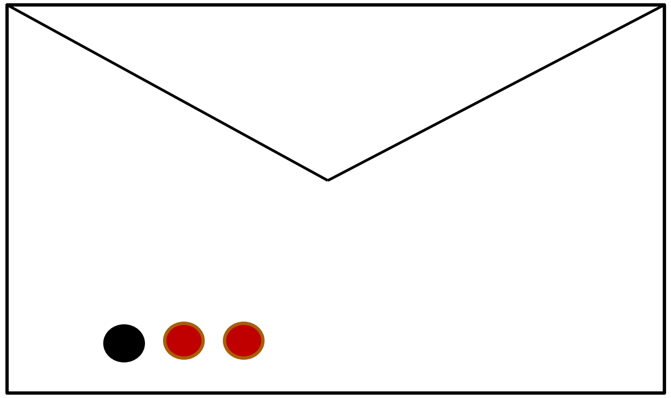
Bayes Rule

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

Using Bayes Rule to Gamble



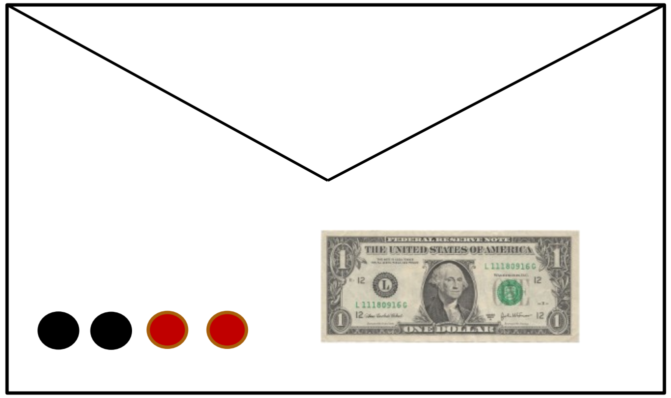
The win envelope has one dollar and four beads



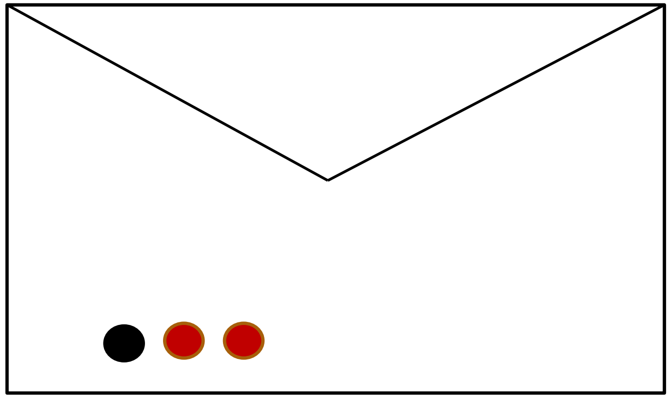
The lose envelope has no dollar and three beads

Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay?

Using Bayes Rule to Gamble



The win envelope has one dollar and four beads



The lose envelope has no dollar and three beads

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.
Suppose its black: How much should you pay?
Suppose its red: How much should you pay?

Discrete Probability Distributions

- In the discrete case, a probability distribution P on S (and hence on the domain of X) is an assignment of a non-negative real number $P(x)$ to each $x \in X$ (or each valid value of x) such that:
  - $0 \leq P(X = x) \leq 1$
  - $\sum_x P(X = x)$
  - Example: The Bernoulli distribution with parameter θ:
  - $P(X = x) = \begin{cases} 1 - \theta, & x = 0 \\ \theta, & x = 1 \end{cases}$

Continuous Probability Distributions

- Sofar we have only mentioned disrete variables.
- A continuous random variable X can take any value in an interval on the real line or in a region in a high dimensional space
- X usually corresponds to a real-valued measurements of some property, e.g., length, position...
- It is not possible to talk about the probability of the random variable assuming a particular value:
  - $P(X = x) = 0$
- Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval:
  - $P(X \in [x1; x2])$
  - $P(X \leq x) = P(X \in (-\infty, x])$

Probability of a continuous random variable

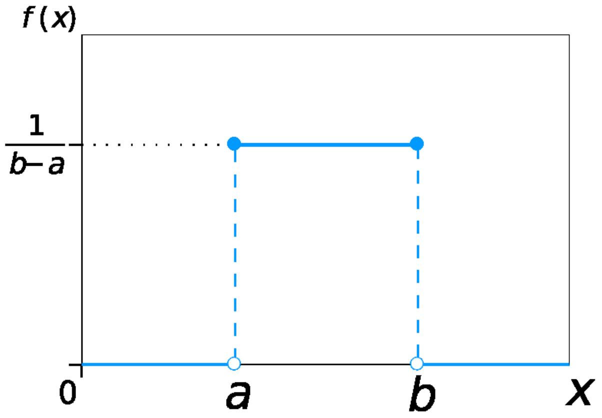
- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function, $p(x)$ between x_1 and x_2.
- Probability Density Function?
  - $\int_{x=-\infty}^{x=\infty} p(x)dx$
  - $P(X \in (x_1, x_2)) = \int_{x_2}^{x\_1} p(x)dx$
- It is NOT probability!

Probability Density Function

- What does $p(x_1) = a$ mean?
- What does $p(x_1) = a$ and $p(x_2) = b$ mean?
- When a value x is sampled from the distribution with density $p(x)$, you are $\frac{a}{b}$ times as likely to find that X is "very close to" x_1 than that X is "very close to" x_2.
- It's something like a histogram with innitely small bar widths.

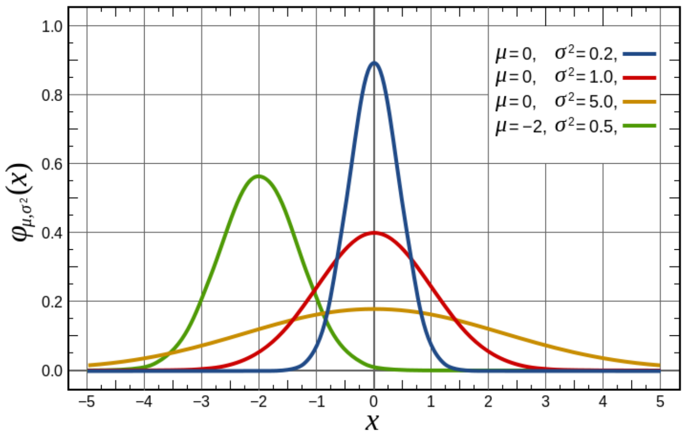
Famous continuous probability distributions

Uniform Distribution



$$p(x) = \begin{cases} \frac{1}{b-a}, & x = \in (a, b) \\ 0, & x = 1 \end{cases}$$

Gaussian Distribution



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

Expectations

- The expected value of some function $f(X)$ of a random variable X that follows a probability distribution is:
  - $E[f(X)] = \sum_{X=x} P(X = x)f(X = x)$, if the distribution is discrete
  - $E[f(X)] = \int_x p(x)f(x)dx$, if the distribution is continuous
  - What is the expected value of rolling a dice? (what is the expected value of function $f(x) = x$?

$$P(X = x) = \frac{1}{6}\ \forall x \in \{1,2,3,4,5,6\}$$

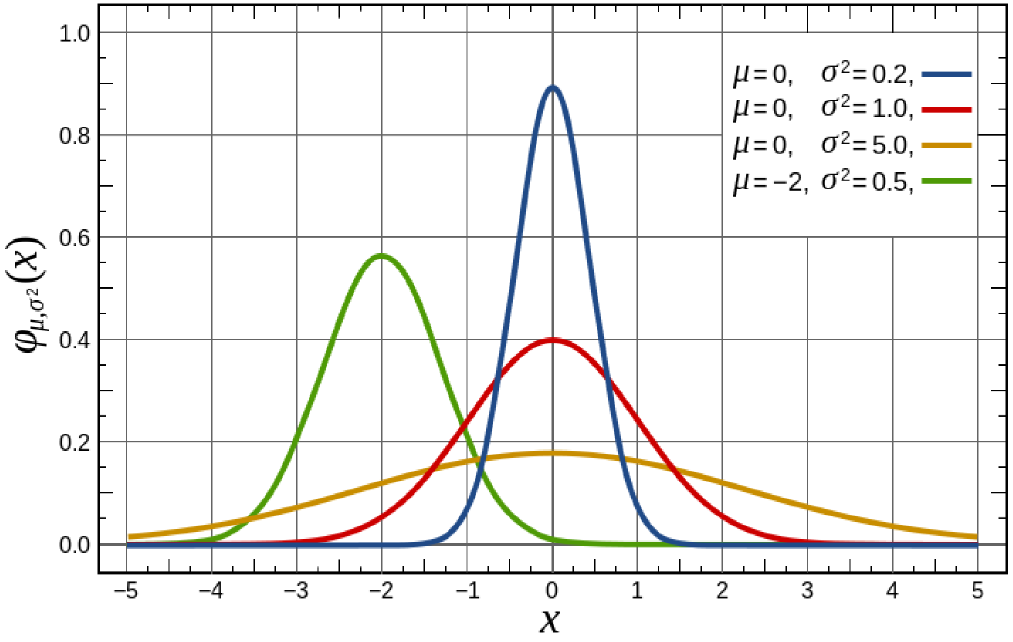
$$E[X] = \sum_{X=x} P(X = x)x = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 5 = 3.5$$

Expectations

- How far from the mean do you expect to be?
- What is the expected difference between a random quantity and its expected value (mean)?
- What is the expected value of $f(X) = X - \mu$?
- Differences from the mean can be either positive or negative, this can be confusing.
- What is the expected squared difference of a random quantity from its expected value?
  - $E[(X-\mu)^2]$
- This is called variance 2 of the distribution.

Gaussians

- The distribution is symmetric, and is often illustrated as a bell-shaped curve.
- Two parameters, (mean) and (standard deviation), determine the location and shape of the distribution.
- Very important distribution.
- Why?

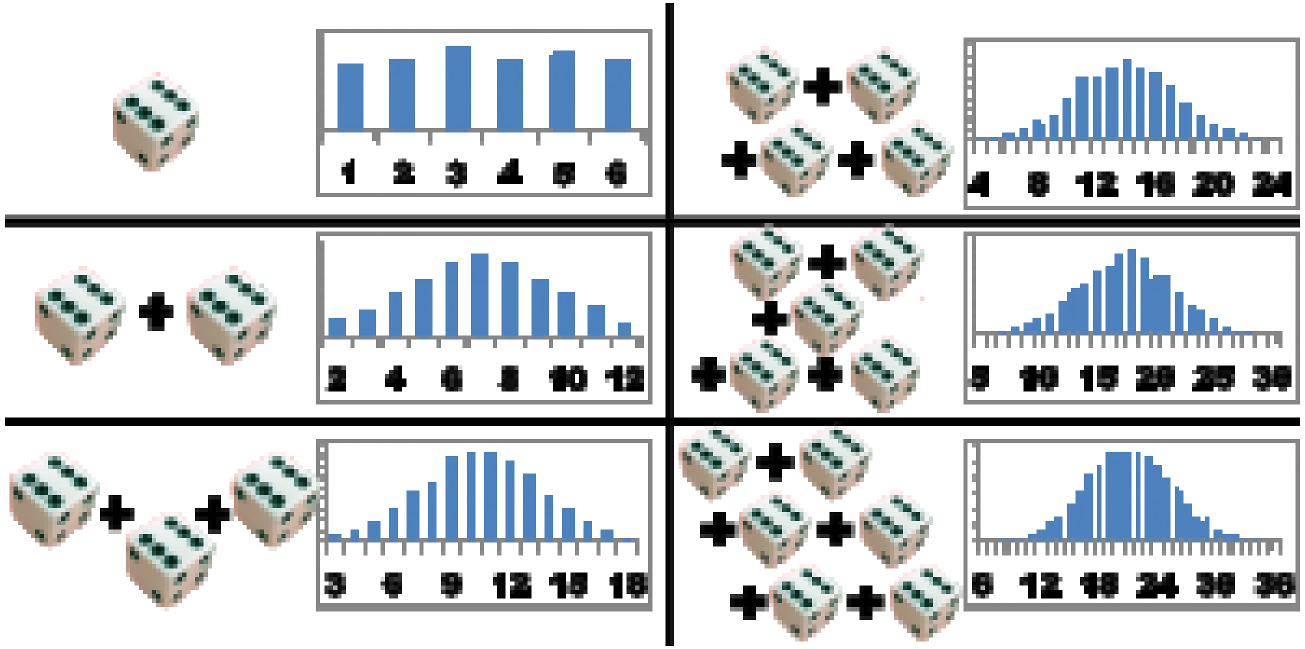


Central Limit Theorem

- If $(X_1, X_2, \ldots, X_n)$ are i.i.d. (independent and identically distributed) random variables
- Then define:

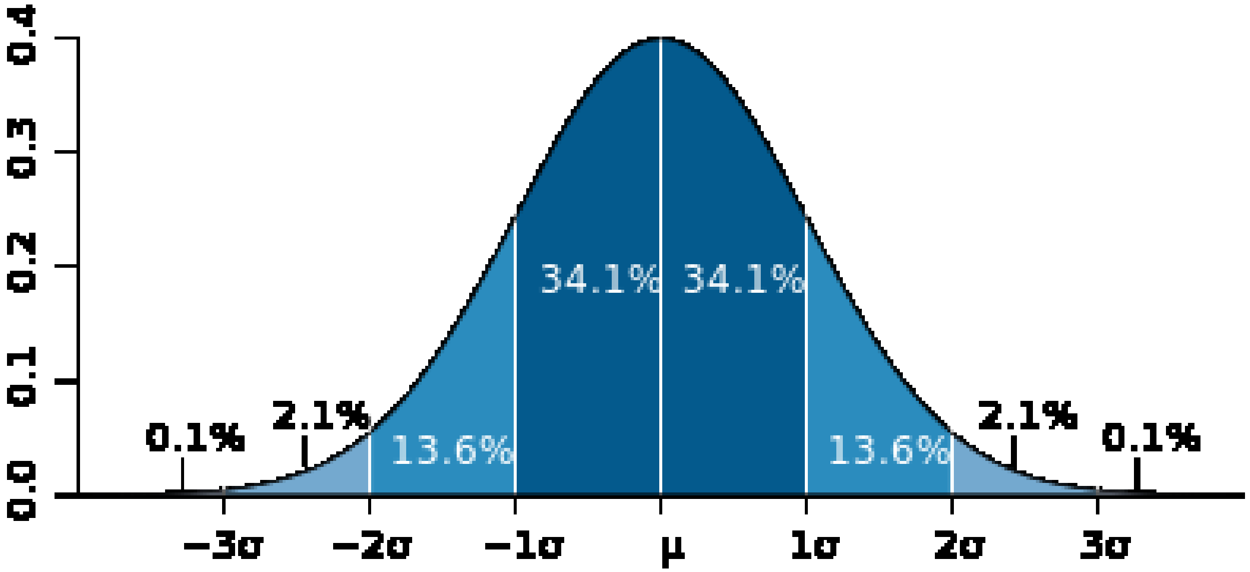
$$\bar{X} = \sum_{i=1}^{N} X_i$$

- As $n \to \infty$
- $P(\bar{X}) \to$ Gaussian with mean $E[(X_i)]$ and variance $Var[X_i]/n$
- Somewhat a justification for assuming Gaussian distribution for just about anything



Mean and variance in Gaussians

- The distribution is symmetric, and is often illustrated as a bell-shaped curve.
- μ shows the center of the bell.
- σ^2 shows the width of the curve.
- $N(0;\ 1)$: The normal distribution with mean 0 and variance 1.
- If $X \sim N(0;\ 1)$, 95% of the the value of X will be within $\pm 2\sigma$



Learning Gaussians from data

- Suppose we have a series of N i.i.d. observations of the scalar variable X, $x = \{x_1, x_2, \dots, x_N\}$
- We know X follows a Gaussian distribution.
- We do not know μ or σ^2

- Which are the most likely values for μ, σ^2 given the data
- Which μ, σ^2 maximizes $\mathrm{P}(\mu, \sigma^2 | x_1, x_2, \dots, x_N)$?
- For which μ, σ^2 are the data more likely?
- Which μ, σ^2 maximizes $\mathrm{P}(x_1, x_2, \dots, x_N | \mu, \sigma^2)$?
- Which sounds better? Which sounds easier?

Likelihood

- We have $X \sim \{x_1, x_2, \dots, x_N\}$
- For which $\theta = \mu, \sigma^2$ is $\{x_1, x_2, \dots, x_N\}$ most likely?
- $P(Data|\mu, \sigma^2)$ is called **Likelihood**
- "Find μ, σ^2 s.t $P(x_1, x_2, \dots, x_N|\mu, \sigma^2)$ is maximum" , aka maximize the likelihood
- If have $X \sim N(\mu, \sigma^2)$ then $L(x_1) = p(x_1|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$

$$\mu_{mle} = argmax_{\mu} P(x_1, x_2, \dots, x_N|\mu, \sigma^2)$$

$$\sigma^2_{mle} = argmax_{\sigma^2} P(x_1, x_2, \dots, x_N|\mu, \sigma^2)$$

Learning MLE

$$\mu_{mle} = argmax_{\mu} P(x_1, x_2, \dots, x_N | \mu, \sigma^2) =$$

$$= argmax_{\mu} \prod_{i=1}^{N} P(x_1, x_2, \dots, x_N | \mu, \sigma^2) =$$

$$= argmax_{\mu} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}.$$

Too hard!!

Find μ s.t $\frac{\partial L}{\partial \theta} = 0$

Learning MLE

Instead, minimize –log Likelihood

$\mu_{mle} = argmax_{\mu} P(x_1, x_2, \ldots, x_N | \mu, \sigma^2) =$

$= argmax_{\mu} - \log\left(\prod_{i=1}^{N} P(x_1, x_2, \ldots, x_N | \mu, \sigma^2)\right) =$

$= argmax_{\mu} - \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\right) = argmax_{\mu} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{i=1}^{N} \log\left(e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\right) =$

$= -argmax_{\mu} [N \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i}^{N} \left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)]$

Easy!!

Find μ s.t $\frac{\partial LL}{\partial \theta} = 0$

Learning MLE

Find μ such that $\frac{\partial\left[-N\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\sum_i^N\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]}{\partial\mu}=0$

...

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

Easy!!

Find μ s.t $\frac{\partial LL}{\partial\theta} = 0$

Learning MLE

Find μ such that $\frac{\partial\left[-\mathrm{N}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\sum_i^N\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]}{\partial\mu}$ =0

$\mu_{mle} = \mu$ s.t. $\frac{\partial LL}{\partial\mu} = 0$

$$\mu_{mle} = \frac{\sum_i^N x_i}{N}$$

$\sigma_{mle}^2 = \sigma^2$ s.t. $\frac{\partial LL}{\partial\sigma^2} = 0$

$$\sigma_{mle}^2 = \frac{\sum_i^N (x_i - \mu_{mle})^2}{N}$$

Learning MAP

- Likelihood : $P(Data|\theta)$
- Bayes Rule : $P(\theta|Data) = \frac{P(Data|\theta) \times P(\theta)}{P(Data)}$
- $P(\theta|Data) \propto P(Data|\theta) \times P(\theta)$
- Posterior ∝ Likelihood × Prior
- $\theta_{map} = \text{argmax}_\theta \, \text{P}(\theta|Data) = \text{argmax}_\theta \, \text{P}(Data|\text{\textbackslash theta}) \times P(\theta)$
- Bad news: You have to chose a prior

Learning MAP

- You have to chose a prior
- e.g., assume any value between x_{min} and x_{max} is equally possible for μ.
- $\mu_{map} = argmax_{\mu}\, P(Data|\theta) \times P(\theta) = \; argmax_{\mu}\, P(Data|\theta) \times \frac{1}{x_{min} - x_{max}}$

$$\mu_{map} = \frac{\sum_i^N x_i}{N} \;!$$

- But: If you assume $\mu \sim N(\mu_0, \sigma_0^2)$ then $\mu_{map} = argmax_{\mu}\, P(Data|\theta) \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$

$$\mu_{map} = \frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}} \frac{\sum_i^N x_i}{N} + \frac{\frac{1}{\sigma^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}} \mu_0$$

Sample size increases the denominator and makes prior less significant

MLE vs. MAP

- MLE : $\theta_{mle} = argmax_{\theta} P(Data|\theta)$
- Choose a value that maximizes the probability of the observed data.
- Easy to overt if dataset is too small.
- MAP : $\theta_{mle} = argmax_{\theta} P(\theta|Data)$
- Choose a value that is most probable given observed data and prior belief.
- People with different priors end up with different estimators.
- With uniform prior MAP = MLE.
- When sample is large, prior is forgotten.