# Machine Learning

Volker Roth

Department of Mathematics & Computer Science
University of Basel

# Chapter 2: Generative models for discrete data

- Foundations of Bayesian inference
- Bayesian concept learning: the number game
- The beta-binomial model: tossing coins
- The Dirichlet-multinomial model: rolling dice
- Example: Simple language models

# Bayesian concept learning

- Consider how a child learns the meaning of the word dog.
- Presumably from positive examples, like "look at the cute dog!"
- Negative examples much less likely, "look at that non-dog" (?)


Image by rawpixel.com on Freepik


Image by jcomp on Freepik


Image by wirestock on Freepik

- Psychological research has shown that people can learn concepts **from positive examples alone.**
- Learning meaning of a word = concept learning = binary classification: $f(x) = 1$ if $x$ is example of concept $C$, and 0 otherwise.
- Standard classification requires positive and negative examples...
  **Bayesian concept learning uses positive examples alone.**

# The number game (Tenenbaum 1999)

- I choose some arithmetical concept $C$, such as "prime number" or "powers of two". I give you a (random) series of positive examples $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn from $C$.
  **Question: does new $\tilde{x}$ belong to $C$?**
- Variation of a common typ of questions in elementary school:

Übungsblatt: Zahlenfolgen bis 100 - Bl. 1

87, 86, 85, 84, ________, ________, ________, ________,

Regel: ______________

20, 21, 22, 23, ________, ________, ________, ________,

Regel: ______________

http://aufgaben.schulkreis.de

# The number game

- Consider integers in $[1, 100]$. I tell you **16 is a positive example.** What are **other positive examples?** Difficult with only one example, predictions will be quite vague.
- **Intuition: numbers similar to 16 are more likely.**
- But what means similar? 17 (close by), 6 (one digit in common), 32 (also even and a power of 2), etc.
- Represent this as a **probability distribution:** $p(\tilde{x}|\mathcal{D})$: probability that $\tilde{x} \in C$ given $\mathcal{D}$. $\rightsquigarrow$ **posterior predictive distribution.**
- After seeing $\mathcal{D} = \{16, 8, 2, 64\}$, you may guess that the concept is "powers of two".
- ...if instead I tell you $\mathcal{D} = \{16, 23, 19, 20\}$...
- How can we explain this behavior and emulate it in a machine?
- Suppose we have a **hypothesis space of concepts, $\mathcal{H}$.**

## Examples

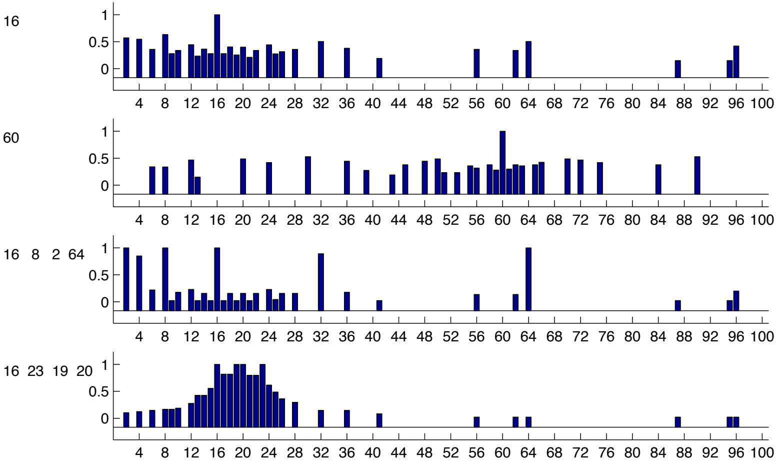



Figure 3.1 in K. Murphy, 2012. Empirical predictive distribution averaged over 8 humans in the number game. First two rows: after seeing $\mathcal{D} = \{16\}$ and $\mathcal{D} = \{60\}$. This illustrates diffuse similarity. Third row: after seeing $\mathcal{D} = \{16, 8, 2, 64\}$. This illustrates rule-like behavior (powers of 2). Bottom row: after seeing $\mathcal{D} = \{16, 23, 19, 20\} \rightsquigarrow$ focused similarity (numbers near 20)

# The number game

- **Version space:** subset of $\mathcal{H}$ that is consistent with $\mathcal{D}$.
- As we see more examples, the version space shrinks and we become increasingly certain about the concept.
  Example: $\mathcal{H} = \{$"even", "odd", "multiples of 4", "powers of two", "prime", "powers of 2 except for 32"$\}$
  $\mathcal{D} = \{16\}$: $\{$**"even"**, ~~"odd"~~, **"multiples of 4"**, **"powers of 2"**, ~~"prime"~~, **"powers of 2 except 32"**$\}$
  $\mathcal{D} = \{16, 8, 2\}$: $\{$**"even"**, ~~"odd"~~, ~~"multiples of 4"~~, **"powers of 2"**, ~~"prime"~~, **"powers of 2 except 32"**$\}$
- **But: version space is not the whole story:**
  - After seeing $\mathcal{D} = \{16\}$, there are many consistent rules; how do you combine them to predict if $\tilde{x} \in C$?
  - Also, after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, why did you choose the rule "powers of two" and not "all even numbers", or "powers of two except for 32", which are equally consistent with the evidence?
- **Bayesian explanation.**

# The number game: Likelihood

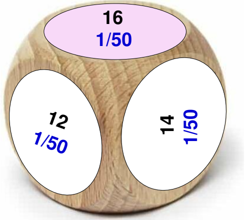
- Having seen $\mathcal{D} = \{16, 8, 2, 64\}$, we must explain why we chose $h_{\text{two}}$ = "powers of two", and not $h_{\text{even}}$ = "even numbers".
- Key intuition: want to avoid suspicious coincidences. If the true concept was $h_{\text{even}}$, how come we only saw powers of two?
- Formalization: assume that examples are **sampled uniformly at random** from the **extension of a concept** (= set of numbers that belong to it), e.g. $h_{\text{even}} = \{2, 4, 6, \ldots, 100\}$.

Probability of sampling $x$ randomly from $h$:

$$P(x|h) = \frac{1}{|h|} = \frac{1}{50} \quad \textbf{for } h = h_{\text{even}}$$

Probability of independently sampling $N$ items (with replacement): $p(\mathcal{D}|h) = \left[\frac{1}{|h|}\right]^N$.

"next number" dice




*uniform probabilities*
*concept specific faces*
*"even" = {2,4,6,8,...98,100}*

# The number game: Likelihood

- Let $\mathcal{D} = \{16\} \rightsquigarrow p(\mathcal{D}|h_{\text{two}}) = 1/6$, since there are 6 powers of two less than 100, but $p(\mathcal{D}|h_{\text{even}}) = 1/50$, since there are 50 even numbers.
- So the likelihood that $h = h_{\text{two}}$ is higher than if $h = h_{\text{even}}$.
- After 4 examples, $p(\mathcal{D}|h_{\text{two}}) = (1/6)^4$, $p(\mathcal{D}|h_{\text{even}}) = (1/50)^4$.
- This is a **likelihood ratio** of almost 5000:1 in favor of $h_{\text{two}}$.
- This quantifies our earlier intuition that $\mathcal{D} = \{16, 8, 2, 64\}$ would be a **very suspicious coincidence** if generated by $h_{\text{even}}$.
- **Size principle:** the model favors the "simplest" hypothesis consistent with the data. Known as **Occam's razor.**
- **William of Ockham** (1287-1347):
  When presented with **competing hypotheses that make the same predictions**, select the **simplest one.**

# The number game: Prior

- Given $\mathcal{D} = \{16, 8, 2, 64\}$, the concept

  $$h' = \text{“powers of two except 32”}$$

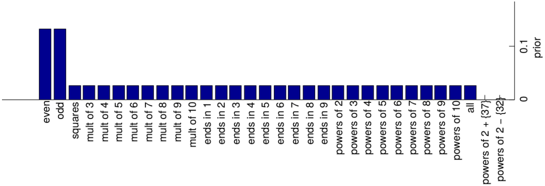
  is even more likely than

  $$h = \text{“powers of two”},$$

  since $h'$ does not need to explain the coincidence that 32 is missing.
- However, $h'$ seems “conceptually unnatural”.
- Capture such intuition by assigning **low prior probability** to “unnatural” concepts.
- Your prior might be different than mine, and this **subjective aspect** of Bayesian reasoning is a source of much controversy.
- But priors are actually quite useful:
  - If you are told the numbers are from some arithmetic rule, then given 1200, 1500, and 900, you may think 400 is likely but 1183 is unlikely.
  - But if you are told that the numbers are examples of healthy cholesterol levels, you would probably think 400 is unlikely and 1183 is likely.

# The number game: Prior

- The prior is the mechanism to formalize background knowledge. **Without this, rapid learning is impossible.**
- Example: use a simple prior which puts uniform probability on 30 simple arithmetical concepts.
- To make things more interesting, we make the concepts “even” and “odd” more likely a priori.
- We also include two “unnatural” concepts, namely “powers of 2, plus 37” and “powers of 2, except 32”, but give them low prior weight.




From Figure 3.2 in K. Murphy: “Machine Learning”, MIT Press, 2012. Prior.

# Bayes Formula

$$\underbrace{\textbf{a-posteriori} \quad P(h|\mathcal{D})}_{\text{after observing data}} \leftarrow \underbrace{\textbf{Likelihood} \;\; P(\mathcal{D}|h)}_{\text{How well does } h \text{ explain the data?}} \cdot \underbrace{\textbf{a-priori} \;\; P(h)}_{\text{prior to observing anything}}$$

Thomas Bayes (1701-61): English statistician, philosopher and Presbyterian minister.

# The number game: Posterior

- The posterior is simply the likelihood times the prior, normalized:
$$p(h|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\mathcal{D}|h)p(h) = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(\mathcal{D} \in h')/|h'|^N},$$
where $\mathbb{I}(\mathcal{D} \in h) = 1$ iff the data are in extension of hypothesis $h$.
- After seeing $\mathcal{D} = \{16, 8, 2, 64\}$, the likelihood is much more peaked on the *powers of two* concept, so this dominates the posterior.
- In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the **MAP estimate**
$$p(h|\mathcal{D}) \to \delta_{\hat{h}^{\text{MAP}}}(h),$$
where
$$\hat{h}^{\text{MAP}} = \arg\max_h p(h|\mathcal{D})$$
is the posterior mode, and $\delta$ is the Dirac measure
$$\delta_x(A) = \begin{cases} 1 & \text{, if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

# The number game: Posterior

- Note that the MAP estimate can be written as
$$\hat{h}^{\text{MAP}} = \arg\max_h p(h|\mathcal{D}) = \arg\max_h [\log p(\mathcal{D}|h) + \log p(h)]$$
- Likelihood depends exponentially on $N$, prior stays constant
$\rightsquigarrow$ as we get more data, the MAP estimate converges to the **maximum likelihood estimate** (MLE):
$$\hat{h}^{\text{MLE}} = \arg\max_h p(\mathcal{D}|h) = \arg\max_h \log p(\mathcal{D}|h).$$
$\rightsquigarrow$ **Enough data overwhelms the prior.**
- If the true hypothesis is in the hypothesis space, then the MAP/ ML estimate will converge upon this hypothesis. Thus Bayesian inference (and ML estimation) are **consistent estimators.**
- We also say that the hypothesis space is identifiable in the limit, meaning we can recover the truth in the limit of infinite data.

$$\underbrace{\textbf{a-posteriori}\quad P(h|\mathcal{D})}_{\text{after observing data}} \leftarrow \underbrace{\textbf{Likelihood}\ P(\mathcal{D}|h)}_{\text{How well does } h \text{ explain the data?}} \cdot \underbrace{\textbf{a-priori}\ P(h)}_{\text{prior to observing anything}}$$

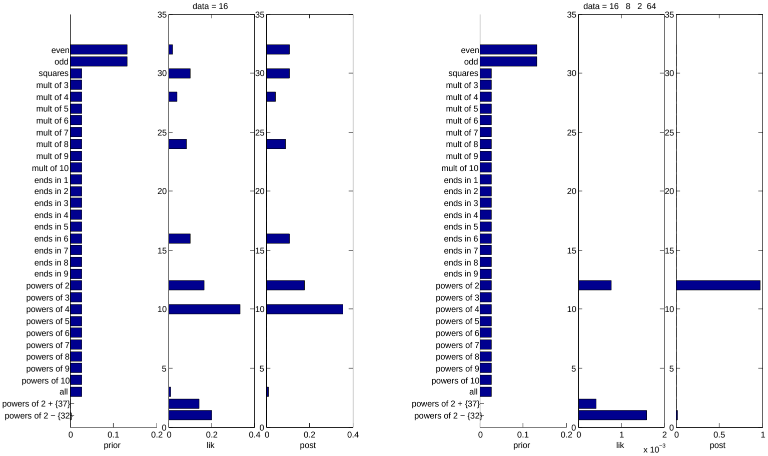
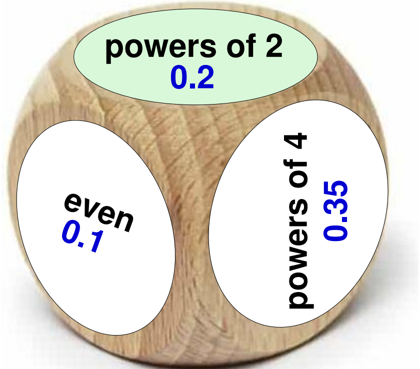



Figure 3.2 in K. Murphy: "Machine Learning", MIT Press, 2012. $\mathcal{D} = \{16\}$ (left) and $\mathcal{D} = \{16, 8, 2, 64\}$ (right)

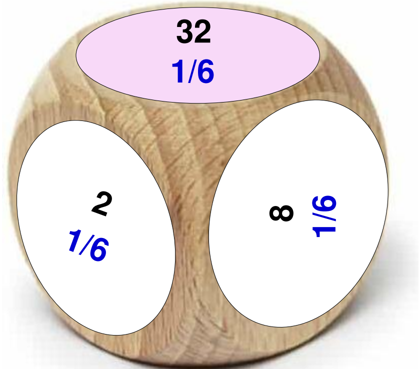
# Generating new numbers

**"arithmetic concept" dice**




***depends on observations***
***{16}***

**"next number" dice**




***uniform probabilities***
***concept specific faces***
***{2,4,8,16,32,64}***

# The number game: Posterior predictive distribution

- Posterior = internal belief state about the world.
  Test these beliefs by making predictions.
- The **posterior predictive distribution** is given by

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)p(h|\mathcal{D})$$

$\rightsquigarrow$ weighted average of the predictions of each hypothesis
$\rightsquigarrow$ **Bayes model averaging.**

- Small dataset $\rightsquigarrow$ vague posterior $p(h|\mathcal{D})$ $\rightsquigarrow$ broad predictive distribution.
- Once we have "figured things out", posterior becomes a delta function centered at the MAP estimate:

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}^{\mathrm{MAP}}}(h) = p(\tilde{x}|\hat{h})$$

$\rightsquigarrow$ **Plug-in approximation.** In general, under-represents uncertainty!

- Typically, predictions by plug-in and Bayesian approach quite different for small $N$ although they converge to same answer as $N \to \infty$.

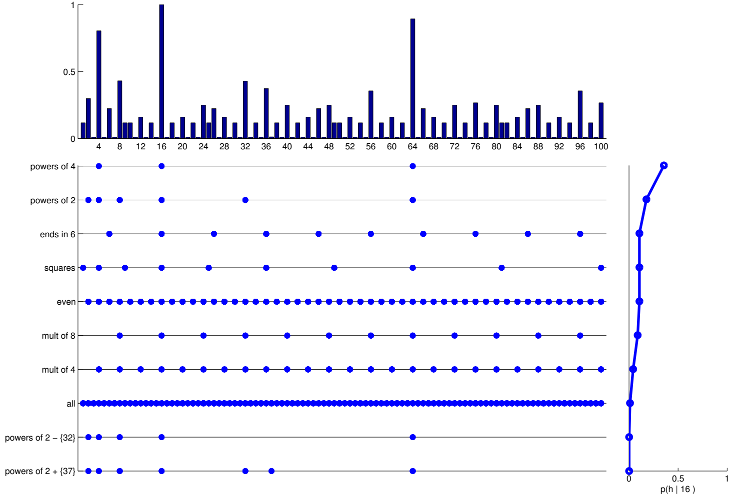



Figure 3.4 in K. Murphy: "Machine Learning", MIT Press, 2012. Posterior over hypotheses and predictive distribution after seeing $\mathcal{D} = \{16\}$. A dot means this number is consistent with $h$.

Right: $p(h|\mathcal{D})$. Weighed sum of dots $\rightsquigarrow p(\tilde{x} \in C|\mathcal{D})$ (top).

# Machine predictions

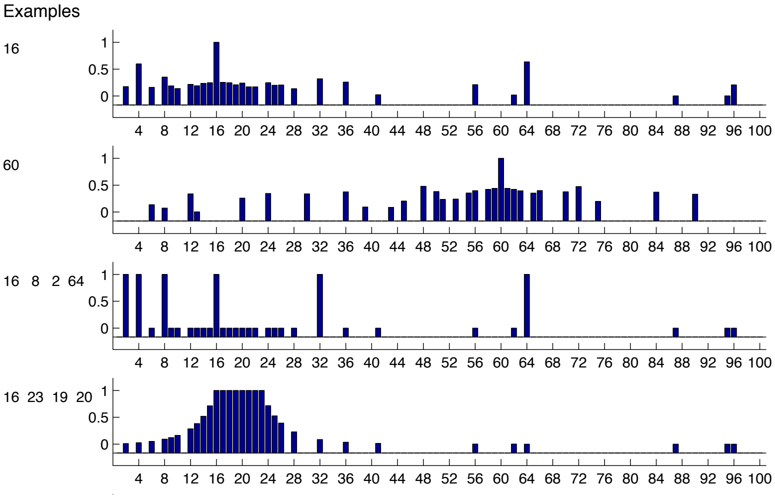



Figure 3.5 in K. Murphy: “Machine Learning”, MIT Press, 2012. Predictive distributions for the model using the full hypothesis space.

# Human predictions

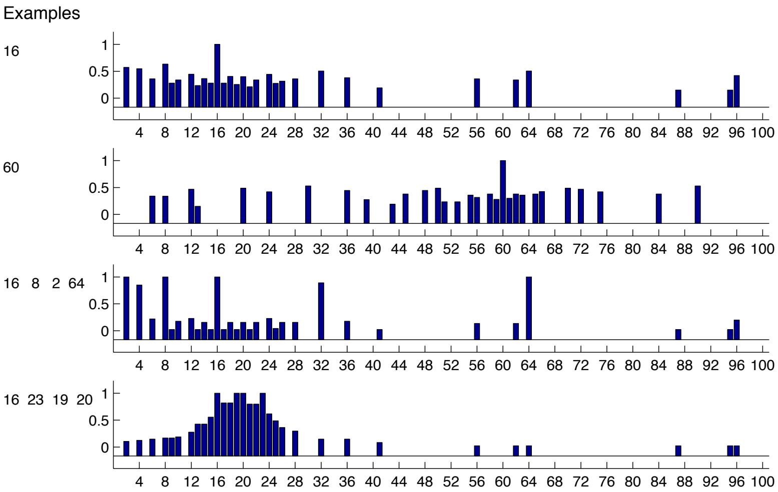



Figure 3.1 in K. Murphy: “Machine Learning”, MIT Press, 2012.

# The beta-binomial model

- Number game: inferring a distribution of a **discrete variable** drawn from a **finite hypothesis space**, $h \in \mathcal{H}$, given a **series of discrete observations.**
- This made the computations simple: just needed to sum, multiply and divide.
- Often, the $K$ unknown parameters are **continuous**, so the hypothesis space is (some subset) of $\mathbb{R}^K$.
- This complicates mathematics (replace sums with integrals), but the basic ideas are the same.
- Example: inferring the probability that a coin shows up heads, given a series of observed coin tosses.

# Common discrete distributions: Binomial and Bernoulli

- Toss a coin $n$ times. Let $X \in \{0, 1, \ldots, n\}$ be the number of heads.
- If the probability of heads is $\theta$, then we say the RV $X$ has a **binomial distribution**, $X \sim \text{Bin}(n, \theta)$:

$$\text{Bin}(X = k|n, \theta) = \binom{n}{k} \theta^{k} (1 - \theta)^{n-k}.$$

- Special case for $n = 1$: **Bernoulli distribution.**
  Let $X \in \{0, 1\}$ $\rightsquigarrow$ binary random variable.
  Let $\theta$ be the probability of **success**. We write $X \sim \text{Ber}(\theta)$.

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)} (1 - \theta)^{\mathbb{I}(x=0)};$$

where $\mathbb{I}(x)$ is the indicator function of a binary $x$:

$$\text{Ber}(x|\theta) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0. \end{cases}$$

# The beta-binomial model: Likelihood

- Suppose $X_i \sim \text{Ber}(\theta)$, where $X_i = 1$ represents "heads", and $\theta \in [0, 1]$ is the probability of heads.
- Assuming **i.i.d. data**, i.e. we observe a sequence of trials, $\mathcal{D} = \{x_1, \dots, x_N\},\ x_i \in \{\text{heads}, \text{tails}\}$, the **Bernoulli likelihood** is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \text{Ber}(x_i|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1) \text{ heads}, \quad N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0) \text{ tails.}$$

- $\{N_1, N_0\}$ are a **sufficient statistics of the data:** all we need to know to infer $\theta$.
- Formally: $s(\mathcal{D})$ is a sufficient statistic for $\mathcal{D}$ if $p(\theta|\mathcal{D}) = p(\theta|s(\mathcal{D}))$.
- Two datasets with the same sufficient statistics $\rightsquigarrow$ same estimated value for $\theta$.

# The beta-binomial model: Likelihood

- **Binomial sampling model:** Suppose we observe the count of the number of heads $N_1$ in a fixed number $N = N_1 + N_0$ of trials, i.e. $\mathcal{D} = (N_1, N)$. Then, $N_1 \sim \text{Bin}(N_1|N, \theta)$, where

$$\text{Bin}(N_1|N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}.$$

- The factor $\binom{N}{N_1}$ is independent of $\theta$
  $\rightsquigarrow$ **likelihood for binomial sampling = Bernoulli likelihood.**
- Any inferences we make about $\theta$ will be the same whether we observe the counts, $\mathcal{D} = (N_1, N)$, or a sequence of trials, $\mathcal{D} = \{x_1, \ldots, x_N\}$.

# The beta-binomial model: Prior

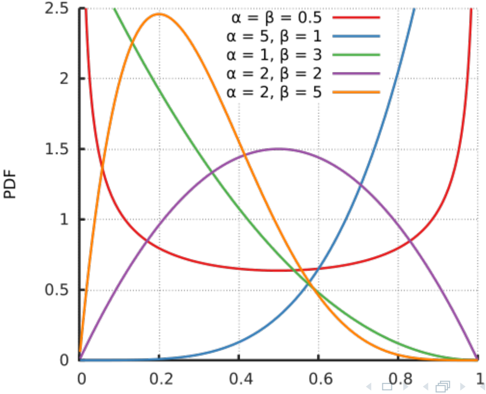
- Need a prior over the interval $[0, 1]$. Would be convenient if the prior had the same form as the likelihood: $p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}$.
- Then, the posterior would be
$$p(\theta|\mathcal{D}) \propto \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2}.$$
Prior and posterior have the same form $\rightsquigarrow$ **conjugate prior.**
- In the case of the Bernoulli likelihood, the conjugate prior is the **beta distribution:**
$$\text{Beta}(\theta|a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$
- The parameters of the prior are called **hyper-parameters.** We can set them to **encode our prior beliefs.**
- If we know "nothing" about $\theta$, we can use a uniform prior. Can be represented by a beta distribution with $a = b = 1$.

# Common continuous distributions: Beta

- The beta distribution is supported on the unit interval $[0, 1]$
- For $0 \leq x \leq 1$, and shape parameters $\alpha, \beta > 0$, the pdf is

$$p(x|\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

The **beta function**, $\mathrm{B}$, is a normalization constant to ensure that the total probability is 1. Note: $\mu[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta}$



# The beta-binomial model: Posterior

- Multiplying with the beta prior we get the following posterior:
$$p(\theta|\mathcal{D}) \propto \mathrm{Bin}(N_1|N,\theta)\mathrm{Beta}(\theta|a,b) \propto \mathrm{Beta}(\theta|N_1+a, N_0+b)$$
- Posterior is obtained by adding the prior hyper-parameters to the empirical counts
$\rightsquigarrow$ hyper-parameters are known as **pseudo counts.**
- The strength of the prior, also known as the **equivalent sample size**, is the sum of the pseudo counts, $\alpha_0 = a + b$.
- Plays a role analogous to the data set size, $N_1 + N_0 = N$.

# The beta-binomial model: Posterior predictive distribution

- So far: focus on **inference of unknown parameter(s).**
- Let us now turn our attention to **prediction of future observable data.**
- Consider predicting the probability of heads in a single future trial under a $\text{Beta}(N_1 + a, N_0 + b)$ posterior $\rightsquigarrow$ **posterior predictive distribution:**

$$\begin{aligned} p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(\tilde{x} = 1|\theta) p(\theta|\mathcal{D})\, d\theta \\ &= \int_0^1 \theta \underbrace{\text{Beta}(\theta|N_1 + a, N_0 + b)}_{p(\theta|\mathcal{D})}\, d\theta \\ &= E[\theta|\mathcal{D}] = \frac{N_1 + a}{N_1 + N_0 + a + b} \end{aligned}$$

$$\{Note : \mu[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}\}$$

# Overfitting and the black swan paradox

- Suppose that we plug-in the MLE, i.e., we use $p(\tilde{x}|\mathcal{D}) \approx \text{Ber}(\tilde{x}|\hat{\theta}_{\text{MLE}})$.
- Can perform quite poorly when the sample size is small: suppose we have seen $N = 3$ tails $\rightsquigarrow$ $\hat{\theta}_{\text{MLE}} = 0/3 = 0$
  $\rightsquigarrow$ **heads seem to be impossible.**
- This is called the **zero count problem** or sparse data problem.
- Even highly relevant in the era of "big data": think about partitioning (patient) data based on (personalized) criteria.
- Analogous to a problem in philosophy called **black swan paradox:** A black swan was a metaphor for something that could not exist.
- Bayesian solution: use a uniform prior: $a = b = 1$.
- Plugging in the posterior gives **Laplace's rule of succession**

$$p(\tilde{x} = 1|\mathcal{D}) = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

  Justifies common practice of adding 1 to empirical counts.

# Common discrete distributions: Multinomial

- Tossing a $K$-sided die $\rightsquigarrow$ can use the **multinomial distribution**.
- Let $\boldsymbol{X} = (X_1, X_2, \ldots X_K)$ be a **random vector**.
  Let $x_j$ be the number of times side $j$ of the die occurs in $n$ trials.

  $$\mathrm{Mu}(\boldsymbol{X} = \boldsymbol{x} | n, \boldsymbol{\theta}) = \binom{n}{x_1 \cdots x_K} \prod_{j=1}^{K} \theta_j^{x_j},$$

  where $\theta_j$ is the probability that side $j$ shows up, and

  $$\binom{n}{x_1 \cdots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

  is the **multinomial coefficient** (the number of ways to divide a set of size $n = \sum_{k=1}^{K} x_k$ into subsets with sizes $x_1$ up to $x_K$).
- Special case for $n = 1$: **Mutinoulli distribution.**

# The Dirichlet-multinomial model

- So far: inferring the probability that a coin comes up heads.
- Generalization: probability that a die with $K$ sides comes up as face $k$.
- Multinomial sampling model: We observe counts, $\mathcal{D} = (N_1, \ldots, N_K)$, where $N_k$ is the number of times event $k$ occurred and $\sum_k N_k = N$:

$$p(\mathcal{D}|N, \boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{N_k},$$

  The counts are again the sufficient statistics. The normalization constant (multinomial coefficient) ist irrelevant for estimating $\boldsymbol{\theta}$.
- Prior: $\boldsymbol{\theta}$ lives in the probability simplex, i.e. $\sum_{k=1}^{K} \theta_k = 1$ and $\theta_k \geq 0$. Conjugate prior with this property: **Dirichlet distribution**

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}.$$

# Common continuous distributions: Dirichlet

- The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ is a multivariate generalization of the beta distribution.
- For the $K$-dimensional random vector $\boldsymbol{X}$, the distribution is supported on $\mathbb{R}^{K-1}$ and is defined as

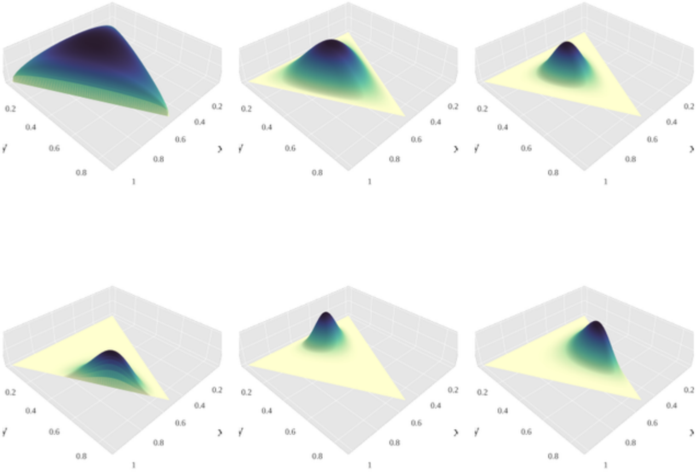
$$\text{Dir}\left(\boldsymbol{X} = (x_1, \ldots, x_K) | \alpha_1, \ldots, \alpha_K\right) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

  where the $\{x_1, x_2, \ldots, x_k\}$ belong to the standard $K-1$ simplex (a.k.a. the probability simplex), i.e.

$$\sum_{i=1}^{K} x_i = 1 \text{ and } x_i \geq 0.$$

  The vertices of this simplex are the $K$ standard unit vectors in $\mathbb{R}^K$.
- The normalizing constant is the **multivariate beta function.**
- The mean is $E[X_i] = \frac{\alpha_i}{\sum_k (\alpha_k)}$.

# Dirichlet distribution




wikimedia.org/w/index.php?curid=49908662

# The Dirichlet-multinomial model

- Posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \\
&= \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1 + N_1, \ldots, \alpha_K + N_K)
\end{aligned}
$$

- Note that we (again) add pseudo-counts $\alpha_k$ to empirical counts $N_k$.

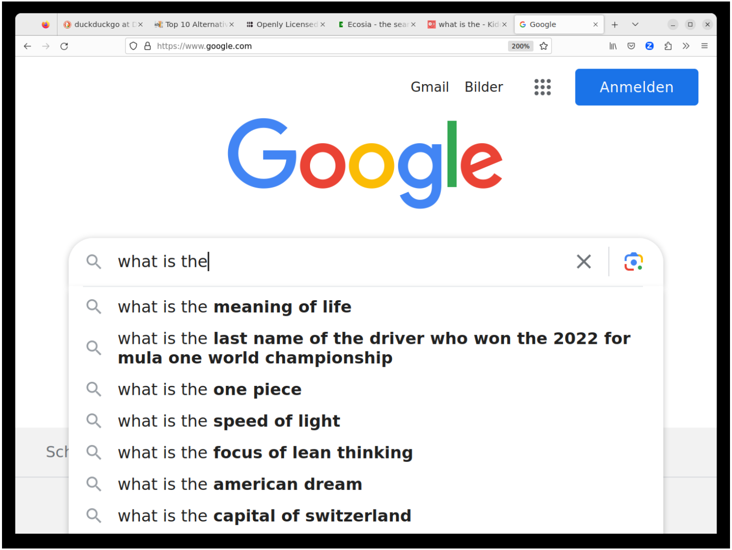
# The Dirichlet-multinomial model

- **Posterior predictive:**

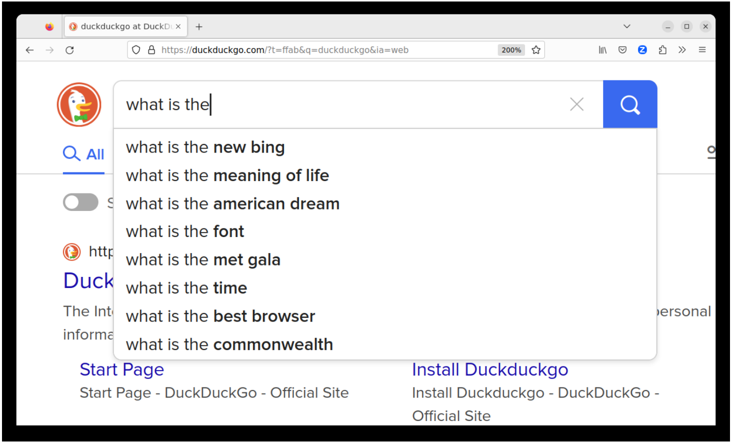
$$
\begin{aligned}
p(\tilde{X}=j|\mathcal{D}) &= \int p(\tilde{X}=j|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\,d\boldsymbol{\theta}, \quad \{\text{write } \boldsymbol{\theta}=(\boldsymbol{\theta}_{-j},\theta_j)^t\} \\
&= \int \underbrace{p(\tilde{X}=j|\theta_j)}_{\theta_j}\underbrace{\left[\int p(\boldsymbol{\theta}_{-j},\theta_j|\mathcal{D})\,d\boldsymbol{\theta}_{-j}\right]}_{p(\theta_j|\mathcal{D})} d\theta_j\} \\
&= E[\theta_j|\mathcal{D}] = \mu\left[\mathrm{Dir}(\boldsymbol{\theta}|\alpha_1+N_1,\ldots,\alpha_K+N_K)\right] \\
&= \frac{N_j+\alpha_j}{\sum_k(N_k+\alpha_k)}
\end{aligned}
$$

- Note: This **Bayesian smoothing** avoids the zero-count problem. Even more important in the multinomial case, since we partition the data into many categories.
- **Example: Simple language models** that predict the probability of the next word.

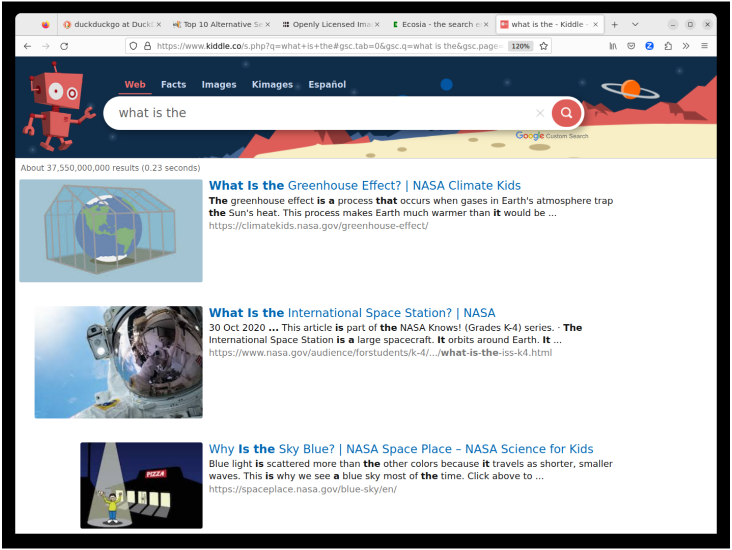
# Language Models



# Language Models



# Language Models



# Language Models

**Language Modeling** is the task of predicting what word comes next

*the students opened their* **......**

More formally: given a sequence of words $x^{(1)}, \ldots, x^{(t)}$, and a vocabulary $V$, compute the probability distribution of the next word $x^{(t+1)} \in V$:

$$P(x^{(t+1)} | x^{(t)}, \ldots, x^{(1)}).$$

How to implement a **simple** Language Model? **...with a *n*-gram model!**
***n*-gram: sequence of *n* consecutive words.**

**Mono-grams: "the", "students", "opened", "their"**
**Bi-grams: "the students", "students opened", "opened their"**
**Tri-grams: "the students opened", "students opened their"**
**4-grams: "the students opened their"**

# n-gram Language Models

**Idea:** observe the frequency of $(n-1)$-grams and estimate the probability of the next word. **Simplifying Markov assumption:**
Next word depends only on the preceding $n-1$ words.

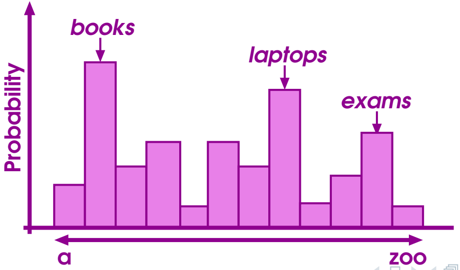
**Mono-gram model:** Choose words indpendently.

Example: **books** occurs 150 times (in a collection of 1000 words)
$\rightsquigarrow \hat{P}(\textbf{books}) = 0.15$

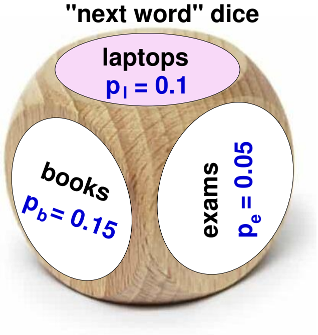
**laptops** occurs 100 times $\rightsquigarrow \hat{P}(\textbf{laptops}) = 0.1$

If needed: add **pseudocounts** to overcome **sparsity problem.**

$\rightsquigarrow$ **Bayesian inference for a Multinomial-Dirichlet model!**



# Mono-gram Model




*the students opened their* ......

**Analogous to Number-Game:** Discrete observations:
$\mathcal{D}$ = collection of $n$ words $\rightsquigarrow$ Likelihood $P(\mathcal{D}|\text{hypothesis})$
**New: continous hypotheses** $\boldsymbol{h} = \{P_1, P_2, \ldots, P_n\}$.
**Conceptually the same model**, but more complicated mathematical formalism (Multinomial-Dirichlet).
**Bayesian updating:** **A-priori** $\rightsquigarrow$ **A-posteriori** word probabilities, given $\mathcal{D}$
**Generating new text:** draw from posterior predictive distribution

# n-gram Language Models

Simplifying assumption: next word depends only on the preceding $n-1$ words.

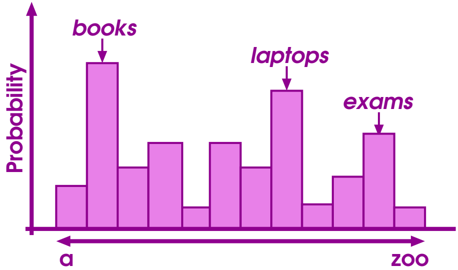
3-gram **students opened their** occurs 1000 times

4-gram extension **students opened their books** 400 times

$\rightsquigarrow \hat{P}($**books** $|$ **students opened their**$) = 0.4$

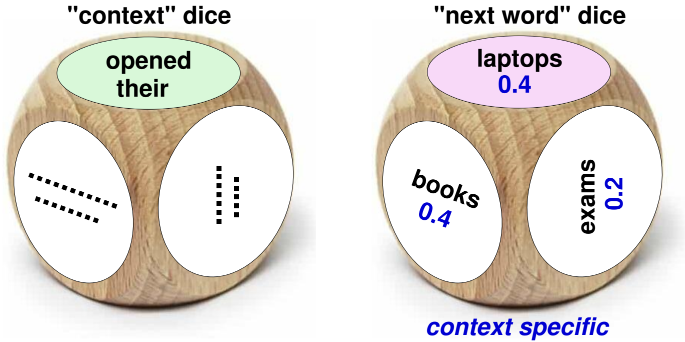
Extension **students opened their laptops** 300 times

$\rightsquigarrow \hat{P}($**laptops** $|$ **students opened their**$) = 0.3$



# 3-gram Model

*the students* ***opened their*** ......




Essentially, this is still a variant of the number game!
But there are two problems:
Sparsity (What if "students opened their books" never occurred?)
Storage (Need to store counts for all *n*-grams).

# n-gram Models: Bayesian interpretation

*the students* ***opened their*** **......**

**Bayesian interpretation:**
Compute posterior predictive, assuming pseudo-counts $\alpha_i = \alpha$:

$$P(X^{(t+1)} = j | \overbrace{x^{(t)}, \dots, x^{(t-n+2)}}^{(n-1)\text{words}}, \mathcal{D}) = \frac{N_j + \alpha}{\sum_k (N_k + \alpha)}$$

$$= \frac{\text{count}\{X^{(t+1)} = j, x^{(t)}, \dots, x^{(t-n+2)}\} + \alpha}{\text{count}\{x^{(t)}, \dots, x^{(t-n+2)}\}(1+\alpha)}.$$

Note: only well-defined if $\text{count}\{x^{(t)}, \dots, x^{(t-n+2)}\} > 0$ !
Example: "opened their" never occured. Possible work-around: just condition on "their" instead $\rightsquigarrow$ **"backoff".**

# Building a 3-gram model

You can build a simple tri-gram Language Model over a 1.7 million word corpus (Reuters) in a few seconds on your laptop.

https://alvinntnu.github.io/python-notes/nlp/language-model.html

```
# Count frequency of co-occurance
for sentence in reuters.sents():
for w1, w2, w3 in trigrams(sentence, pad_right=Tru
model[(w1, w2)][w3] += 1
# Transform the counts to probabilities
for w1_w2 in model:
total_count = float(sum(model[w1_w2].values()))
for w3 in model[w1_w2]:
model[w1_w2][w3] /= total_count
```

# Generative 3-gram Model

- Idea: Given 2 start words, choose the next word randomly from all words with 3-gram probabilty $> \epsilon$

| | |
|---|---|
| 'conference' | 0.25 |
| 'of' | 0.125 |
| '. ' | 0.125 |
| 'with' | 0.084 |
| 'agency' | 0.083 |
| 'that' | 0.083 |
| 'brought' | 0.042 |
| 'about' | 0.041 |
| 'broke' | 0.041 |

- Start word: **the news** **.....**

Note: **Severe sparsity problem: not much granularity!**

- 3rd word (random choice): **the news** **brought**

# Generative 3-gram Model

- New probability table:
  **news brought .....**

| | |
|---|---|
| 'by' | 0.99 |

- **brought by .....**

| | |
|---|---|
| 'the' | 0.27 |
| 'several' | 0.09 |
| 'British' | 0.09 |
| 'Pepsi' | 0.09 |
| 'tax' | 0.09 |
| 'groups' | 0.09 |

- *The news brought by Pepsi, which produced the reported negative inflation rates last year's Bureau of Statistics said.*
- Surprisingly grammatical ...but incoherent.
  More context information is necessary, but increasing $n$ worsens sparsity problem, and increases model size.
- We will discuss better models in the **Neural Networks** chapter!