CS540: Machine Learning I

Lecture 3: Bayesian parameter estimation and hypothesis testing for discrete data

Kevin Murphy

Monday September 19, 2005[1]

[1]Slides last updated on September 19, 2005

Administrivia

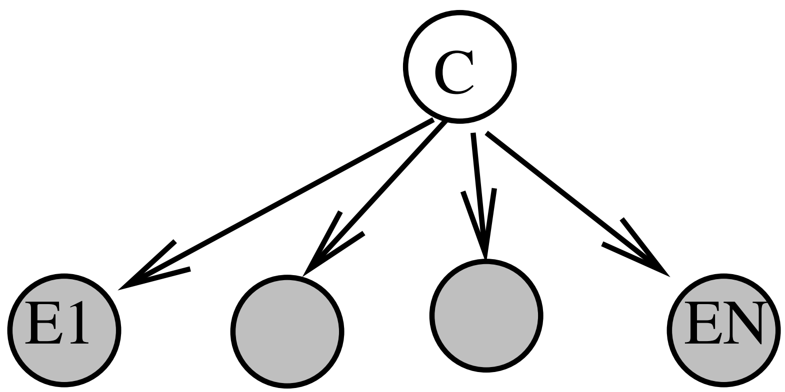
- Speed
- Reading
  - Lecture slides available at front
  - Chapter 2 and appendices at front
  - On web, reading for lecture K contains material related to lecture K; you should read this before hand!
- Homeworks
  - Easy/ hard?
  - Solutions to HW1 available
  - Hand in your HW1, pick up someone else's and grade it by next Monday (if enrolled for credit); put your name on it when you grade it!
  - HW2 now available online

Administrivia

- Auditors
  - Please sign the form (at front); I will give them to Joyce Poon
  - Please do *not* turn in your homeworks!
- Matlab
  - Everyone should have access; if not, see me.
  - Homeworks will *not* require stats toolbox etc.
- Discussion section
  - Useful?
  - Second discussion section Wednesday 5-6?

Naive Bayes classifier

- Let $C \in \{1, \dots, K\}$ represent the class of a document (e.g., $C =$spam or $C =$not spam).
- Let $W_i = 1$ if word i occurs in this document, otherwise $W_i = 0$.
- A naive Bayes classifier assumes the words (features) are conditionally independent given the class (written as $W_i \perp W_j | C$).
- This can be represented as a Bayes net (recall that a node is conditionally independent of its non-descendants given its parents).



Naive Bayes classifier: inference

- Since $W_i \perp W_j | C$, the joint is

$$P(C, W_{1:N}) = P(C) \left[\prod_{i=1}^{N} P(W_i | C) \right]$$

- Hence the posterior over class labels is given by

$$P(C = c | w_{1:N}) = \frac{P(C = c) \prod_{i=1}^{N} P(w_{1:N} | c)}{\sum_{c'} P(C = c') \prod_{i=1}^{N} P(w_{1:N} | c')}$$

Naive Bayes classifier: learning

- The root CPD $P(C = c)$ can be estimated by counting how many times each class occurs
(e.g., $P(C = \text{spam}) = 0.05$, $P(C = \text{non-spam} = 0.95)$).
- Each leaf CPD $P(w_i|c)$ can have a different kind of distribution, e.g., bernoulli, Gaussian, etc.
- For document classification, $P(W_i = 0/1|C = c)$ can be estimated by counting how many times word i occurs in documents of class c.
- For real-valued data, $p(W_i|C = c)$ can be estimated by fitting a Gaussian to all data points that are labeled as class c.
- If the class labels are not observed during training, this model can be used for clustering (see later).

Parameter learning

- We said that the root CPD $P(C = c)$ can be estimated by counting how many times each class occurs. Why?
- We said $P(W_i = 0/1|C = c)$ can be estimated by counting how many times word i occurs in documents of class c. Why? And what if the word never occurs?
- We now discuss these issues, which are equivalent to estimating the parameters of coins and dice.
- We will also discuss how to infer which words are useful for classification (feature selection) by computing the mutual information between two variables.
- You will implement this for homework 2.

Bernoulli distribution

- Let $X \in \{0, 1\}$ represent heads or tails.
- Suppose $P(X = 1) = \mu$. Then

$$P(x|\mu) = \mathrm{Be}(X|\mu) = \mu^x (1 - \mu)^{1-x}$$

- It is easy to show that

$$E[X] = \mu, \quad \mathrm{Var}[X] = \mu(1 - \mu)$$

MLE for a Bernoulli distribution

- Given $D = (x_1, \ldots, x_N)$, the likelihood is

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- The log-likelihood is

$$\begin{aligned} L(\mu) &= \log p(D|\mu) = \sum_n x_n \log \mu + (1-x_n)\log(1-\mu) \\ &= N_1 \log \mu + N_0 \log(1-\mu) \end{aligned}$$

where $N_1 = n = \sum_n x_n$ is the number of heads and $N_0 = m = \sum_n (1-x_n)$ is the number of tails (sufficient statistics).

- Solving for $\frac{dL}{d\mu} = 0$ yields

$$\mu_{ML} = \frac{n}{n+m}$$

Problems with the MLE

- Suppose we have seen 3 heads out of 3 trials. Then we predict that all future coins will land heads:

$$\mu_{ML} = \frac{n}{n+m} = \frac{3}{3+0}$$

- This is an example of the *sparse data problem*: if we fail to see something in the training set (e.g., an unknown word), we predict that it can never happen in the future.
- We will now see how to solve this pathology using Bayesian estimation.

Conjugate priors

- A Bayesian estimate of μ requires a prior $p(\mu)$.
- A prior is called conjugate if, when multiplied by the likelihood $p(D|\mu)$, the resulting posterior is in the same parametric family as the prior. (Closed under Bayesian updating.)
- The Beta prior is conjugate to the Bernoulli likelihood

$$\begin{aligned} P(\mu|D) &\propto P(D|\mu)P(\mu) \\ &\propto [\mu^n(1-\mu)^m][\mu^{a-1}\mu^{b-1}] \\ &= \mu^{n+a-1}(1-\mu)^{m+b-1} \end{aligned}$$

  where n is the number of heads and m is the number of tails.
- a, b are hyperparameters (parameters of the prior) and correspond to the number of "virtual" heads/tails (pseudo counts). $N_0 = a + b$ is called the effective sample size (strength) of the prior. $a = b = 1$ is a uniform prior (Laplace smoothing).

The beta distribution

- To ensure the prior is normalized, we define

$$P(\mu|a,b) = \text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

where the gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

Note that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Also, for integers, $\Gamma(x+1) = x!$.

- The normalization constant $1/Z(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ ensures

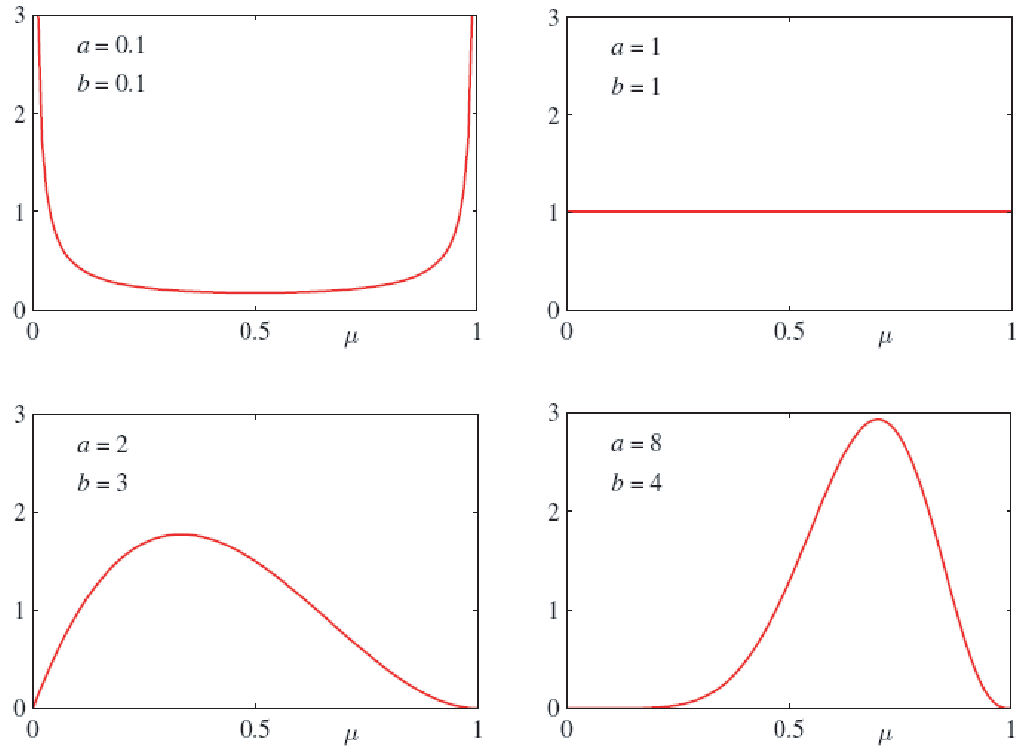
$$\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1$$

The beta distribution

If $\mu \sim Be(a, b)$, then

$$E\mu = \frac{a}{a+b}$$

$$\text{mode } \mu = \frac{a-1}{a+b-2}$$



Bayesian updating of a beta distribution

- If we start with a beta prior $Be(\mu|a,b)$ and see n heads and m tails, we end up with a beta posterior $Be(\mu|a+n, b+n)$:

$$
\begin{aligned}
P(\mu|D) &= \frac{1}{P(D)} P(D|\mu) P(\mu|a,b) \\
&= \frac{1}{P(D)} [\mu^n (1-\mu)^m] \frac{1}{Z(a,b)} [\mu^{a-1} \mu^{b-1}] \\
&= Be(\mu|n+a, m+b)
\end{aligned}
$$

- The marginal likelihood is the ratio of the normalizing constants:

$$
\begin{aligned}
P(D) &= \frac{Z(a+b, n+m)}{Z(a,b)} \\
&= \frac{\Gamma(a+n)\Gamma(b+m)}{\Gamma(a+n+b+m)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}
\end{aligned}
$$

Sequential Bayesian updating

- Start with beta prior $p(\theta|\alpha_h, \alpha_t) = \mathcal{B}(\theta; \alpha_h, \alpha_t)$.
- Observe N trials with N_h heads and N_t tails. Posterior becomes

$$p(\theta|\alpha_h, \alpha_t, N_h, N_t) = \mathcal{B}(\theta; \alpha_h + N_h, \alpha_t + N_t) = \mathcal{B}(\theta; \alpha'_h, \alpha'_t)$$

- Observe another N' trials with N'_h heads and N'_t tails. Posterior becomes

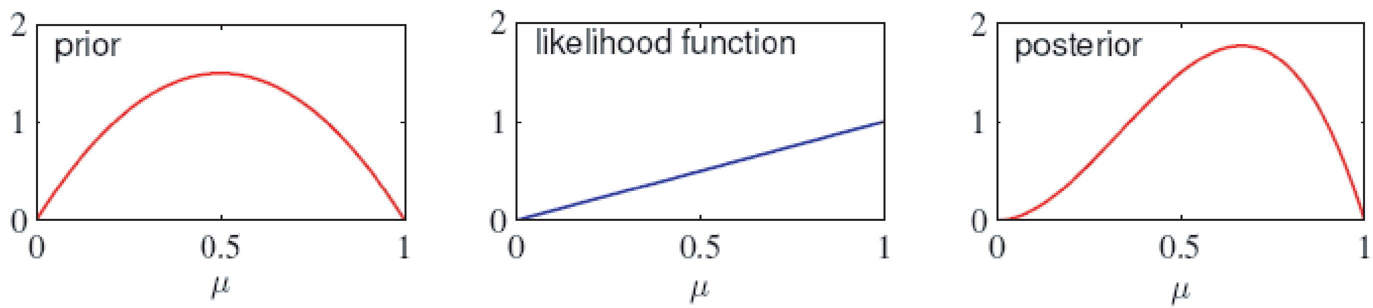
$$\begin{aligned} p(\theta|\alpha'_h, \alpha'_t, N'_h, N'_t) &= \mathcal{B}(\theta; \alpha'_h + N'_h, \alpha'_t + N'_t) \\ &= \mathcal{B}(\theta; \alpha_h + N_h + N'_h, \alpha_t + N_t + N'_t) \end{aligned}$$

- So sequentially absorbing data in any order is equivalent to batch update. (assuming iid data and exact Bayesian updating).
- This is useful for online learning and large datasets.

Bayesian updating in pictures

- Start with $Be(\mu|a = 2, b = 2)$ and observe $x = 1$, so the posterior is $Be(\mu|a = 3, b = 2)$.

```
thetas = 0:0.01:1;
alphaH = 2; alphaT = 2; Nh=1; Nt=0; N = Nh+Nt;
prior = betapdf(thetas, alphaH, alphaT);
lik = choose(N,Nh) * thetas.^Nh .* (1-thetas).^Nt;
post = betapdf(thetas, alphaH+Nh, alphaT+Nt);
```



Posterior predictive distribution

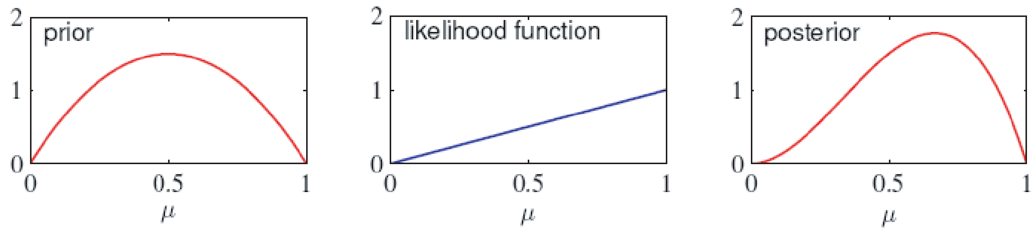
- The posterior predictive distribution is

$$p(X=1|D) = \int_0^1 p(X=1|\mu)p(\mu|D)d\mu$$
$$= \int_0^1 \mu p(\mu|D)d\mu = E[\mu|D] = \frac{n+a}{n+m+a+b}$$

- With a uniform prior $a=b=1$, we get Laplace's rule of succession

$$p(X=1|N_h,N_t) = \frac{N_h+1}{N_h+N_t+2}$$

- Start with $Be(\mu|a=2,b=2)$ and observe $x=1$ to get $Be(\mu|a=3,b=2)$, so the mean shifts from $E[\mu]=2/4$ to $E[\mu|D]=3/5$.



Effect of prior strength

- Let $N = N_h + N_t$ be number of samples (observations).
- Let N' be the number of pseudo observations (strength of prior) and define the prior means

$$\alpha_h = N'\alpha'_h, \quad \alpha_t = N'\alpha'_t, \quad \alpha'_h + \alpha'_t = 1$$

- Then posterior mean is a convex combination of the prior mean and the MLE (where $\lambda = N'/(N + N')$):

$$\begin{aligned}
P(X = h|\alpha_h, \alpha_t, N_h, N_t) &= \frac{\alpha_h + N_h}{\alpha_h + N_h + \alpha_t + N_t} \\
&= \frac{N'\alpha'_h + N_h}{N + N'} \\
&= \frac{N'}{N + N'}\alpha'_h + \frac{N}{N + N'}\frac{N_h}{N} \\
&= \lambda\alpha'_h + (1 - \lambda)\frac{N_h}{N}
\end{aligned}$$

Effect of prior strength

- Suppose we have a uniform prior $\alpha'_h = \alpha'_t = 0.5$, and we observe $N_h = 3$, $N_t = 7$.
- Weak prior $N' = 2$. Posterior prediction:

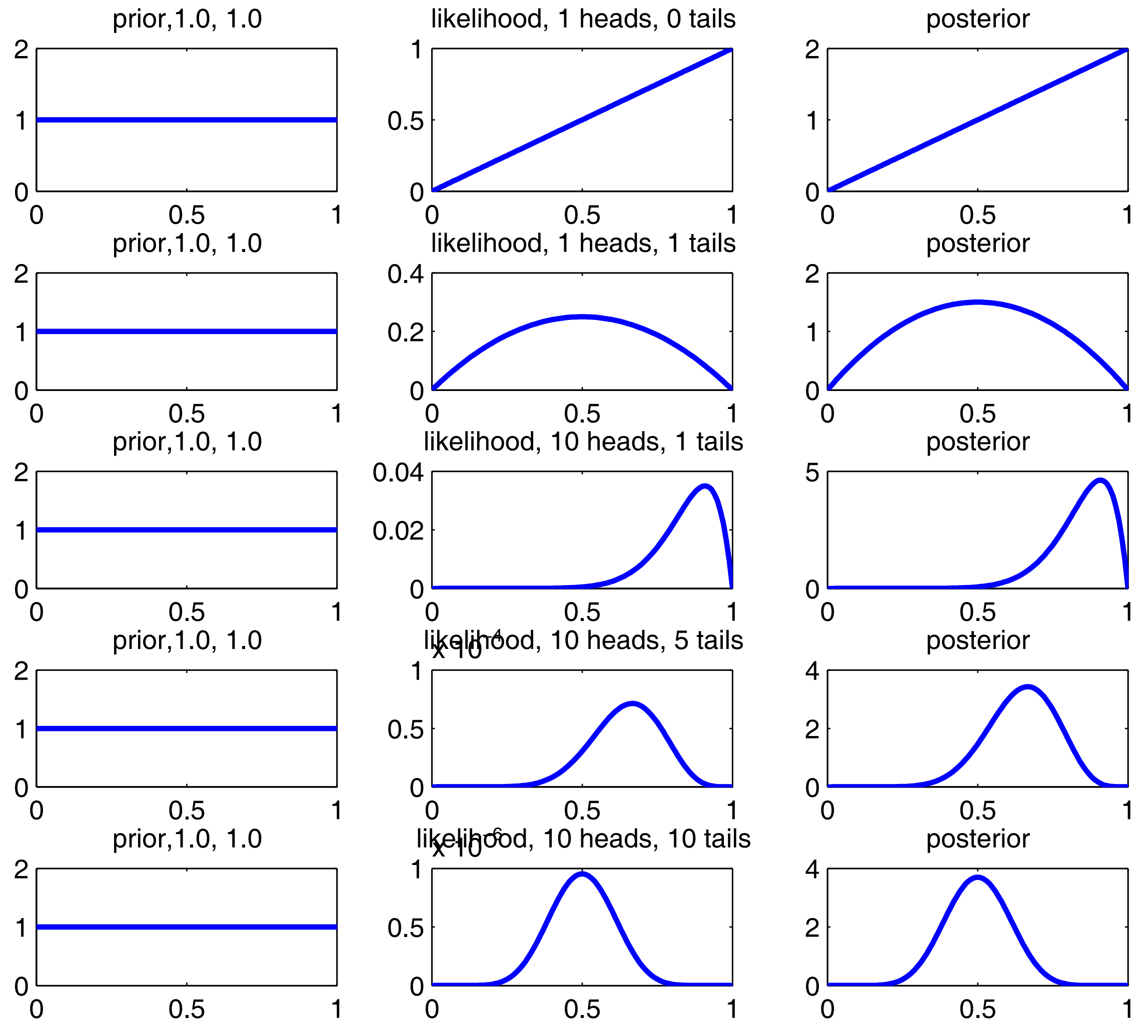
$$P(X = h|\alpha_h = 1, \alpha_t = 1, N_h = 3, N_t = 7) = \frac{3+1}{3+1+7+1} = \frac{1}{3} \approx 0.33$$

- Strong prior $N' = 20$. Posterior prediction:

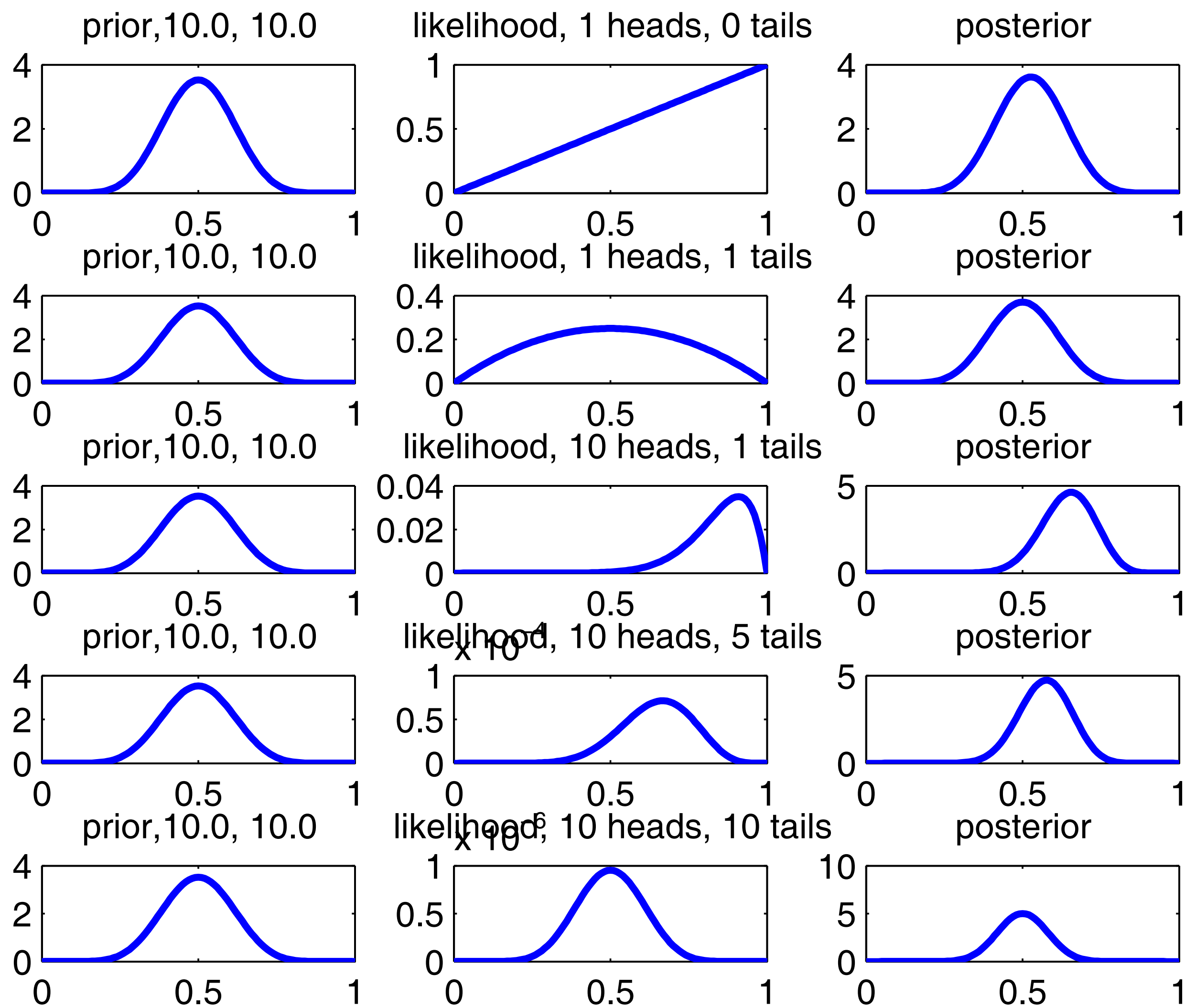
$$\frac{3+10}{3+10+7+10} = \frac{13}{30} \approx 0.43$$

- However, if we have enough data, it washes away the prior. e.g., $N_h = 300$, $N_t = 700$. Estimates are $\frac{300+1}{1000+2}$ and $\frac{300+10}{1000+20}$, both of which are close to 0.3
- As $N \to \infty$, $P(\theta|D) \to \delta(\theta, \hat{\theta}_{ML})$, so $E[\theta|D] \to \hat{\theta}_{ML}$.

Parameter posterior - small sample, uniform prior

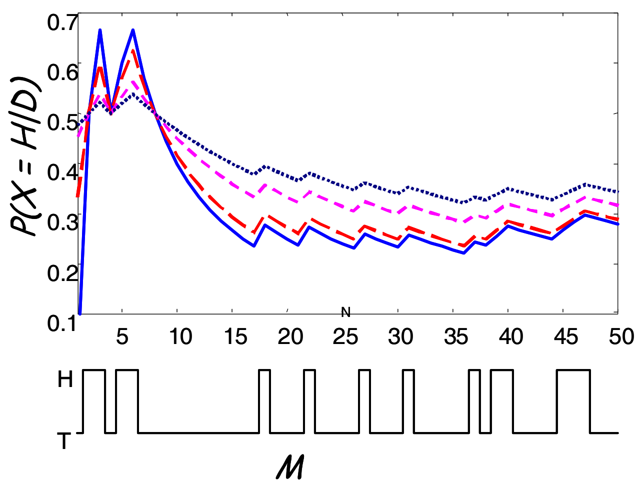


Parameter posterior - small sample, strong prior



Prior smooths parameter estimates

- The MLE can change dramatically with small sample sizes.
- The Bayesian estimate changes much more smoothly (depending on the strength of the prior).
- Lower blue=MLE, red = beta(1,1), pink = beta(5,5), upper blue = beta(10,10)



Maximum a posteriori (MAP) estimation

- MAP estimation picks the mode of the posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta)$$

- If $\theta \sim Be(a, b)$, this is just

$$\hat{\theta}_{MAP} = (a-1)/(a+b-2)$$

- MAP is equivalent to maximizing the penalized maximum log-likelihood

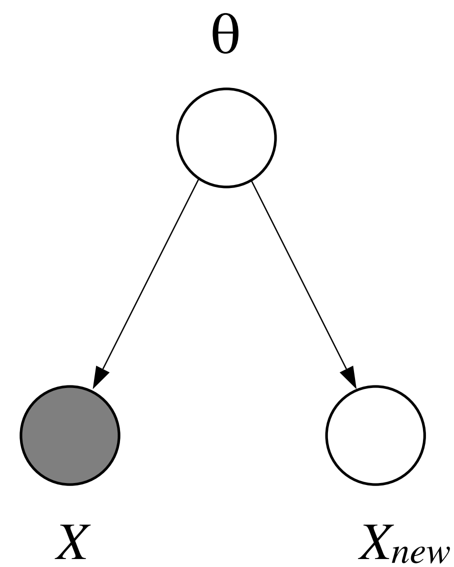
$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(D|\theta) - \lambda c(\theta)$$

where $c(\theta) = -\log p(\theta)$ is called a *regularization term*. λ is related to the strength of the prior.

Integrate out or Optimize?

- $\hat{\theta}_{MAP}$ is not Bayesian (even though it uses a prior) since it is a point estimate.
- Consider predicting the future. A Bayesian will integrate out all uncertainty:

$$\begin{aligned} p(x_{new}|X) &= \int p(x_{new}, \theta|X) d\theta \\ &= \int p(x_{new}|\theta, X) p(\theta|X) d\theta \\ &\propto \int p(x_{new}|\theta) p(X|\theta) p(\theta) d\theta \end{aligned}$$



- A frequentist will use a "plug-in" estimator eg ML/MAP:

$$p(x_{new}|X) = p(x_{new}|\hat{\theta}), \quad \hat{\theta} = \arg\max_{\theta} p(X|\theta)$$

From coins to dice

- Suppose we observe N iid die rolls (K-sided): D=3,1,K,2,...
- Let $[x] \in \{0,1\}^K$ be a one-of-K encoding of x eg. if $x = 3$ and $K = 6$, then $[x] = (0,0,1,0,0,0)^T$.
- Multinomial distribution: $p(X = k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood

$$\ell(\theta; D) = \log p(D|\theta) = \sum_m \log \prod_k \theta_k^{[x^m = k]}$$

$$= \sum_m \sum_k [x^m = k] \log \theta_k = \sum_k N_k \log \theta_k$$

- We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a Lagrange multiplier.

MLE for multinomial

- Constrained cost function:

$$\tilde{l} = \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k \right)$$

- Take derivatives wrt θ_k:

$$\begin{aligned}
\frac{\partial \tilde{l}}{\partial \theta_k} &= \frac{N_k}{\theta_k} - \lambda = 0 \\
N_k &= \lambda \theta_k \\
\sum_k N_k &= N = \lambda \sum_k \theta_k = \lambda \\
\hat{\theta}_{k,ML} &= \frac{N_k}{N}
\end{aligned}$$

- $\hat{\theta}_{k,ML}$ is the fraction of times k occurs.

Dirichlet priors

- Let $X \in \{1, \ldots, K\}$ have a multinomial distribution

$$P(X|\theta) = \theta_1^{I(X=1)} \theta_2^{I(X=2)} \cdots \theta_K^{I(X=k)}$$

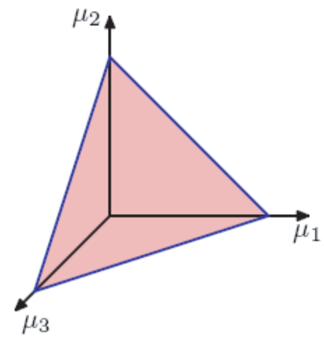
- For a set of data $X^1, \ldots, X^N$, the sufficient statistics are the counts $N_i = \sum_n I(X_n = i)$.
- Consider a Dirichlet prior with hyperparameters α

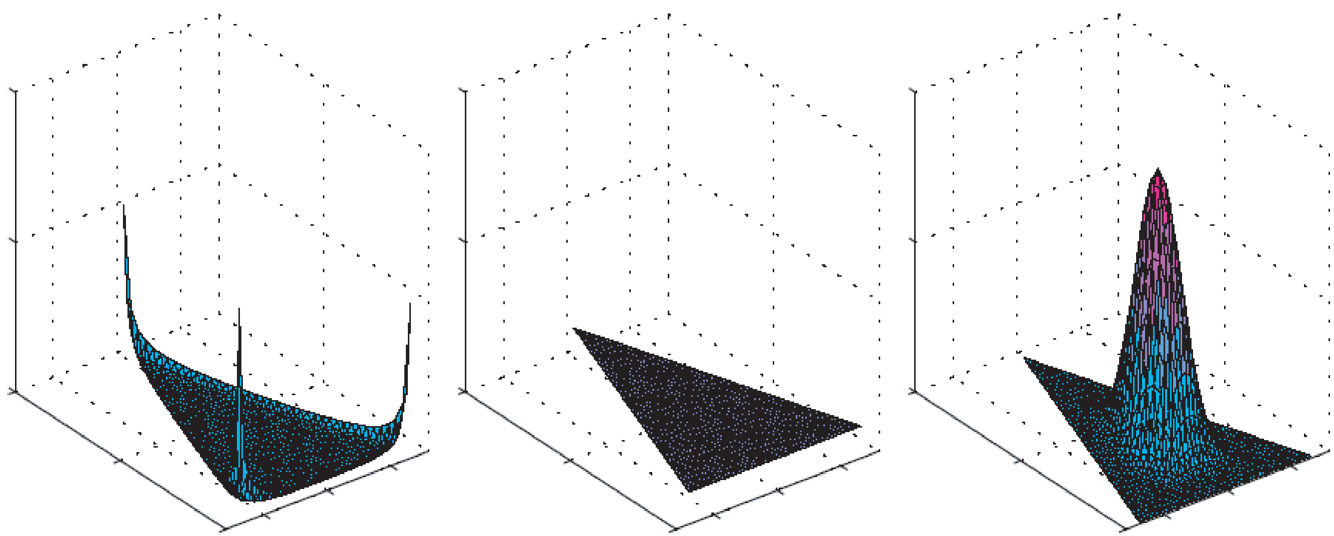
$$p(\theta|\alpha) = \mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}$$

where $Z(\alpha)$ is the normalizing constant

$$\begin{aligned} Z(\alpha) &= \int \cdots \int \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} d\theta_1 \cdots d\theta_K \\ &= \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \end{aligned}$$

Dirichlet priors





Properties of the Dirichlet distribution

- If $\theta \sim Dir(\theta|\alpha_1, \ldots, \alpha_K)$, then

$$E[\theta_k] = \frac{\alpha_k}{\alpha_0}$$
$$\text{mode}[\theta_k] = \frac{\alpha_k - 1}{\alpha_0 - K}$$

where $\alpha_0 \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k$ is the total strenght of the prior.

Likelihood, prior, posterior, evidence

- Likelihood, prior, posterior:

$$
\begin{aligned}
P(\vec{N}|\vec{\theta}) &= \prod_{i=1}^{K} \theta_i^{N_i} \\
p(\theta|\alpha) &= \mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1} \\
p(\theta|\vec{N}, \vec{\alpha}) &= \frac{1}{Z(\alpha) p(\vec{N}|\alpha)} \theta_1^{\alpha_1 + N_1} \cdots \theta_K^{\alpha_K + N_k} \\
&= \mathcal{D}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)
\end{aligned}
$$

- Marginal likelihood (evidence):

$$
P(\vec{N}|\vec{\alpha}) = \frac{Z(\vec{N} + \vec{\alpha})}{Z(\vec{\alpha})} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}
$$

Marginal likelihood ≈ negative entropy

- Marginal likelihood (evidence):

$$P(\vec{N}|\vec{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}$$

- If $\alpha_k = 1$, this becomes

$$P(\vec{N}|\vec{\alpha} = 1) = \frac{\Gamma(K)}{\Gamma(N + K)} \prod_k \frac{\Gamma(N_k + 1)}{\Gamma(1)}$$

- Using the fact that $\Gamma(1) = 1$ and Stirling's approximation $\log \Gamma(x+1) \approx x \log x - x$, we get

$$\begin{aligned} P(\vec{N}|\vec{\alpha} = 1) &\approx -N \log N + N + \sum_k (N_k \log N_k - N_k) \\ &= \sum_k N_k \log(N_k/N) = -N\mathcal{H}(\{N_k/N\}) \end{aligned}$$

where $\mathcal{H}(p_k) = -\sum_k p_k \log p_k$ is the entropy of a distribution.

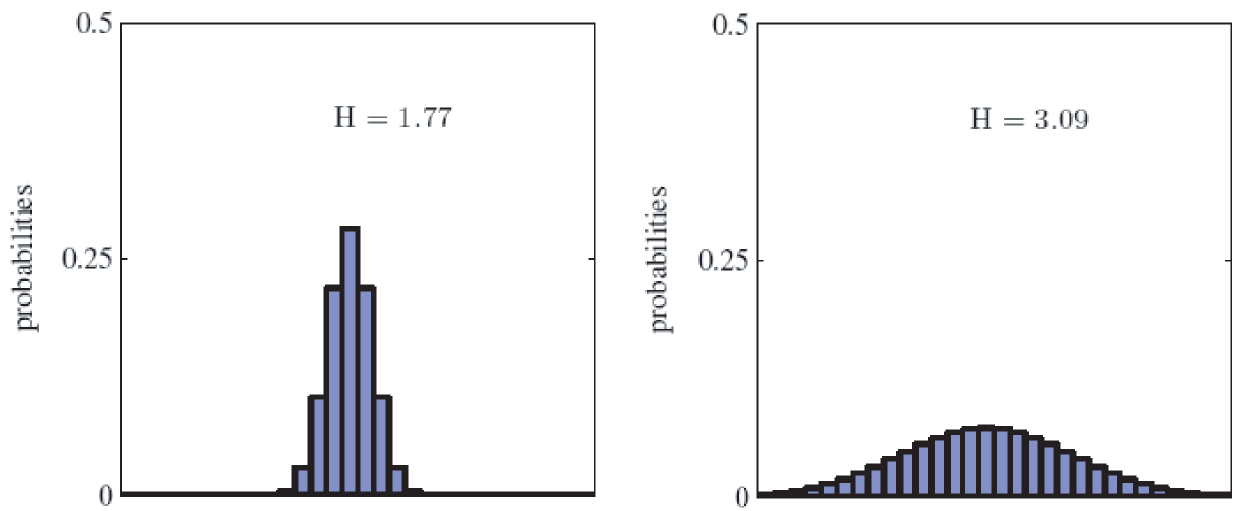
Entropy

- Suprising (unlikely) events convey more information, so we define the information content of an observation (in bits) to be

$$h(x) = \log_2 1/p(x)$$

- The average information content of a random variable X is

$$H(X) = -\sum_x p(x) \log_2 p(x)$$



Data compression

- There is a close link between density estimation and data comrpession.
- The noiseless coding theorem (Shannon 1948) says that the entropy is a lower bound on the number of bits needed to transmit the state of X.
- More likely states can be given shorter code words.
- There is also a close link between data compression and model selection/ hypothesis testing.

Hypothesis testing

- Consider this example from Mackay chapter 37 (on class web page).
- When spun on edge $N = 250$ times, a Belgian one-euro coin came up heads $Y = 141$ times and tails 109.
- "It looks very suspicious to me. We can reject the null hypothesis (that the coin is unbiased) with a signifcance level of 5%". — Barry Blight, LSE (modified from quote in *The Guardian*, 2002)
- Does this mean $P(H_0|D) < 0.05$? Let us compare classical hypothesis testing with a Bayesian approach (using marginal likelihood).

Classical hypothesis testing

- We would like to distinguish two models, or hypotheses: H_0 means the coin is unbiased (so $p = 0.5$); H_1 means the coin is biased (has probability of heads $p \neq 0.5$).
- We need a decision rule that maps data to accept/reject.
- We will do this by computing a scalar quantity of our data called the deviance, $d(D)$, and comparing its observed value with what we would expect if H_0 were true.
- We declare "H_1" if $d(D) > t$ for some threshold t (to be determined).
- In our case, we will use $d(D) = N_h$, the number of heads.

P-values

- The p-value of a threshold t is the probability of falsely rejecting the null hypothesis:

$$p(t) = P(\{D' : d(D') > t\}|H_0, N)$$

- Intuitively, the p-value is the probability of getting data *at least that extreme* given H0.
- Since computing the p-value requires summing over all possible datasets of size N, a standard approximation is consider the expected distribution of $d(D')$, assuming $D' \sim P(\cdot|H_0)$, as $N \to \infty$.

Significance levels

- The p-value of a threshold t is the probability of falsely rejecting the null hypothesis:

$$pval(t) = P(\{D' : d(D') > t\}|H_0, N)$$

- We usually choose a threshold t so that the probability of a false rejection is below some significance level $\alpha = 0.05$ (i.e., choose t s.t., $pval(t) \leq \alpha$).
- This means that on average we will "only" be wrong $1/20$ times (!).

Classical analysis of the euro-coin data

- Blight used a two-sided test and found a p-value of 0.0497, so he said "we can reject the null hypothesis at significance level 0.05".

$$\begin{aligned} pval &= P(Y \geq 141|H_0) + P(Y \leq 109|H_0) \\ &= (1 - P(Y < 141|H_0)) + P(Y \leq 109|H_0) \\ &= (1 - P(Y \leq 140|H_0)) + P(Y \leq 109|H_0) \\ &= 0.0497 \end{aligned}$$

```
n=250; p = 0.5;
p1 = 1-binocdf(140,n,p);
p2 = binocdf(109,n,p);
pval = p1 + p2
```

Classical analysis violates the likelihood principle

- Why do we care about tail probabilities, such as

$$P(Y \geq 141|H_0) = P(Y = 141|H_0) + P(Y = 142|H_0) + \cdots$$

  when the number of heads we observed was 141, not 142 or larger?

- P-values (and therefore all classical hypothesis tests) violate the likelihood principle, which says

  > In order to choose between hypotheses H0 and H1 given observed data D, one should ask how likely the observed data are under each hypothesis; do not ask questions about data that we might have observed but did not.

- For more examples, see "What is Bayesian statistics and why everything else is wrong", Michael Lavine (2000), on web page.

Bayesian approach

- We want to compute the posterior ratio of the 2 hypotheses:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)}$$

- Let us assume a uniform prior $P(H_0) = P(H_1) = 0.5$.
- Then we just focus on the ratio of the marginal likelihoods:

$$P(D|H_1) = \int_0^1 d\theta \;\; P(D|\theta, H_1)P(\theta|H_1)$$

- For H_0, there is no free parameter, so

$$P(D|H_0) = 0.5^N$$

where N is the number of coin tosses in D.

Ratio of evidences (Bayes factor)

- We compute the ratio of marginal likelihoods (evidence):

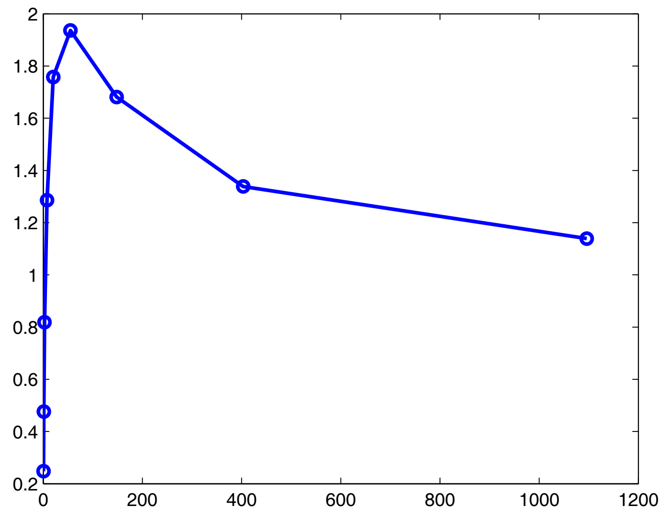
$$
\begin{aligned}
BF(1,0) &= \frac{P(D|H_1)}{P(D|H_0)} = \frac{Z(\alpha_h + N_h, \alpha_t + N_t)}{Z(\alpha_h, \alpha_t)} \frac{1}{0.5^N} \\
&= \frac{\Gamma(140+\alpha)\Gamma(110+\alpha)}{\Gamma(250+2\alpha)} \times \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} \times 2^{250}
\end{aligned}
$$

- We compute $BF(1,0)$ for a range of prior strengths $\alpha_t = \alpha_h = \alpha$. Must work in log domain to avoid underflow!

```
alphas = [0.37 1 2.7 7.4 20 55 148 403 1096];
Nh = 140; Nt = 110; N = Nh+Nt;
numer = gammaln(Nh+alphas) + gammaln(Nt+alphas) + ...
      gammaln(2*alphas)  + 250*log(2);
denom = gammaln(N+2*alphas) + 2*gammaln(alphas);
r = exp(numer ./ denom);
```

So, is the coin biased or not?

- We plot the likelihood ratio vs hyperparameter α:



- For a uniform prior, $\frac{P(H_1|D)}{P(H_0|D)} = 0.48$, (weakly) favoring the fair coin hypothesis H_0!
- At best, for $\alpha = 50$, we can make the biased hypothesis twice as likely.
- Not as dramatic as saying "we reject the null hypothesis (fair coin) with significance 5%".

Summary: Bayesian vs classical hypothesis testing

- The Bayesian approach is simpler and more natural (no need for "p-values", "significance tests", etc.)
- The Bayesian approach does not violate the likelihood principle.
- The Bayesian approach allows the use of prior knowledge to prevent us from jumping to conclusions too hastily.
- See the excellent tutorials on P-values and Bayes factors by Steven Goodman on the web page.

Another example: testing for independence

- Suppose we are given N (x, y) pairs, where X has J possible values and Y has K. We want to know if X and Y are independent.
- eg. consider this contingency table

	$y = 1$	$y = 2$	$y = 3$
$x = 1$	15	29	14
$x = 2$	46	83	56

- Traditional approach: compare $H_0 = X \perp Y$ vs $H_1 = X \not\perp Y$. Compute p-value using χ^2 statistic

$$d_{\chi^2}(D) = \sum_{x,y} \frac{(O_{x,y} - E_{x,y})^2}{E_{x,y}} = \sum_{x,y} \frac{(N(x,y) - NP(x)P(y))^2}{NP(x)P(y)}$$

- Let us consider a Bayesian approach.

Bayesian test for independence

- If independent,

$$P(D|H_0) = p(X|\alpha_{j\cdot})p(Y|\alpha_{\cdot k})$$

where $\alpha_{j\cdot}$ and $\alpha_{\cdot k}$ are different prior vectors.

- If dependent,

$$P(D|H_1) = p(X,Y|\alpha_{jk})$$

- We want to compute

$$\begin{aligned} p(H_0|D) &= \frac{p(D|H_0)p(H_0)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)} \\ &= \frac{1}{1 + \frac{p(D|H_1)}{p(D|H_0)}\frac{p(H_1)}{p(H_0)}} \end{aligned}$$

- If we assume $p(H_0) = p(H_1)$, we can focus on the Bayes factor $B = \frac{p(D|H_0)}{p(D|H_1)}$.

Bayesian test for independence

- It is simple to show (homework!) that

$$B = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(X|\alpha_{j\cdot})p(X|\alpha_{\cdot k})}{p(X,Y|\alpha_{jk})} \;=\; \frac{\Gamma(\sum_{jk}\alpha_{jk})}{\Gamma(N+\sum_{jk}\alpha_{jk})}\prod_{j=1}^{J}\frac{\Gamma(N_{j\cdot}+\alpha_{j\cdot})}{\Gamma(\alpha_{j\cdot})}\prod_{k=1}^{K}\frac{\Gamma(N_{\cdot k}+\alpha_{\cdot k})}{\Gamma(\alpha_{\cdot k})}\prod_{j,k=1}\frac{\Gamma(\alpha_{jk})}{\Gamma(N_{jk}+\alpha_{jk})}$$

- Using the entropy approximation, we get

$$\begin{aligned}\log\frac{p(D|H_0)}{p(D|H_1)} &\approx -N\mathcal{H}(\frac{N_{j\cdot}}{N}) - N\mathcal{H}(\frac{N_{\cdot k}}{N}) + N\mathcal{H}(\frac{N_{jk}}{N}) \\ &= -N\mathcal{D}\left(\frac{N_{jk}}{N}||\frac{N_{j\cdot}}{N}\times\frac{N_{\cdot k}}{N}\right) \\ &= -N\mathcal{I}(X,Y)\end{aligned}$$

where

$$\mathcal{D}(p||q) \stackrel{\text{def}}{=} \sum_k p_k \log\frac{p_k}{q_k}$$

is the Kullback-Leibler divergence between distributions p, q, and

$$\mathcal{I}(X,Y) \stackrel{\text{def}}{=} \mathcal{D}(P(X,Y)||P(X)P(Y))$$

is the mutual information between X and Y.

KL divergence (relative entropy)

- $KL(p||q)$ is a "distance" measure of q from p

$$\mathcal{D}(p||q) \stackrel{\text{def}}{=} \sum_k p_k \log \frac{p_k}{q_k}$$

- It is not strictly a distance, since it is asymmetric.
- The KL can be rewritten as

$$\mathcal{D}(p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\sum_k p_k \log q_k - H(p_k)$$

  This makes it clear that the KL measures the extra number of bits we would need to use to encode X if we thought the distribution was q_k but it was actually q_k.
- KL satisfies $D(p||q) \geq 0$ with equality iff $p = q$.

Minimizing KL divergence is maximiming likelihood

- We would like to find $q(x|\theta)$ s.t. $D(p||q)$ is minimized, where $p(x)$ is the "true" distribution.
- Of course $p(x)$ is unknown but we can approximate by the empirical distribution given samples. Then

$$KL(p||q) \approx \frac{1}{N} \sum_n \log p(x_n) - \log q(x_n|\theta)$$

- Since $p(x)$ is independent of θ, we find that

$$\arg\min_q KL(p||q) = \arg\max_q \frac{1}{N} \sum_n \log q(x_n|\theta)$$

Mutual information

- The mutual information measures how close the joint and independent distributions are:

$$\mathcal{I}(X, Y) \stackrel{\text{def}}{=} \mathcal{D}(P(X, Y)||P(X)P(Y))$$

- It is easy to show that the mutual information between X and Y is how much our uncertainty about Y decreases when we observe X (or vice versa):

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- $I(X, Y) \geq 0$ with equality iff $X \perp Y$.

χ^2 IS AN APPROXIMATION TO THE MUTUAL INFORMATION

- We showed

$$\log \frac{p(D|H_0)}{p(D|H_1)} \approx -N\mathcal{D}\left(\frac{N_{jk}}{N}||\frac{N_{j\cdot}}{N} \times \frac{N_{\cdot k}}{N}\right)$$
$$= -N\mathcal{I}(X,Y)$$

- If we make the additional approximation

$$D(p||q) \approx \sum_k \frac{(p_k - q_k)^2}{2q_k}$$

then we recover the χ^2 statistic.

Are two histograms from the same distribution?

- To see if two samples X and Y come from the same multinomial distribution, create an indicator variable $C \in \{1, 2\}$ which specifies which data set each sample comes from.

	$z = 1$	$z = 2$	$\cdots$	$z = K$
$c = 1$	N_1	N_2	$\cdots$	N_K
$c = 2$	M_1	M_2	$\cdots$	M_K

- If the two histograms are from the same distribution, then C is independent of Z. So just compute $P(C \perp Z|D)$. As before, we get $\log \frac{P(D|\text{same})}{P(D|\text{diff})} \approx -N\mathcal{I}(X, Y)$, which can be further approximated using χ^2.