Introduction to Bayesian Statistics: Sufficiency, Likelihood and Conditionality Principles & The Exponential Family of Distributions

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

September 4, 2017

Reverend Thomas Bayes
(ca. 1702–1761)

Sole probability paper,
"Essay Towards Solving a Problem in the Doctrine of Chances",
published posthumously in 1763

References

- C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
- A Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
- J M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
- D. Sivia and J Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2006.
- Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.
- Chris Bishops' PRML book, Chapter 2
- M. Jordan, An introduction to Probabilistic Graphical Models, Chapter 8 (pre-print)
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapters 2 and 4

Contents

Parametric Modeling

- **Statistical theory** derives from observations of a random phenomenon an inference about the probability distribution underlying this phenomenon.[a]

- We consider **parametric modeling**: The observations x are the realizations of a random variable X of known probability density function $f(x|\theta)$ where

  - θ is unknown and belongs to a space Θ of finite dimension.

  - The function $f(x \mid \theta)$ considered as a function of θ for a fixed realization of the observation $X=x$ is called **the likelihood function**.

$$\ell(\theta \mid x) = f(x \mid \theta)$$

[a] Here we follow closely:
- C. P. Robert, *The Bayesian Choice, Springer, 2nd edition,* chapter 1 (full text available)
- Brani Vidakovic, Bayesian Statistics for Engineers, online course.

Example of Parametric Modeling

- Consider the problem of forest fires. Ecological and meteorological factors influence their eruption. Determining the probability *p* of fire as a function of these factors can be useful in the prevention of forest fires.

- We assume a parametrized shape for the function *p.*

- Denoting by *h* the humidity rate, *t* the average temperature, *x* the degree of management of the forest, a logistic model (Bernoulli random variable of parameter *p*) could be proposed as:

$$p = exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)/ [1 + exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)]$$

- The statistical step is now dealing with the evaluation of $\alpha_1, \alpha_2, \alpha_3$.

Example of Parametric Modeling

- Price and salary variations are closely related.
- We can represent this dependence by assuming a linear relation:

$$\Delta P = a + b\Delta S + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where Δ P and ΔS are the price and salary variations, *a* and *b* are unknown coefficients and ε is the error factor.

- *We* assume that ε is normally distributed.
  - ✓ A justification can be given through the Central Limit Theorem (e.g. the additional influence of many small factors of similar magnitude). This model allows for a thorough statistical analysis, which remains valid even if the distribution of ε is not exactly normal.
- The inference problem is now the calculation of the parameters

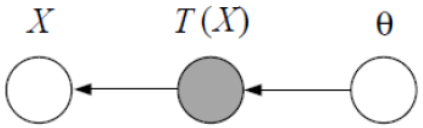
$$\theta = \left(a, b, \sigma^2\right)$$

Functional vs. Parametric Statistical Estimation

- Estimate the distribution underlying the phenomenon under minimal assumptions, e.g. using functional estimation:
  - Density estimation
  - Regression function estimation, etc.
- **The parametric approach is more pragmatic: a finite number of observations can efficiently estimate only a finite number of parameters.**
- Model assessment should be considered.

Sufficiency Principle

- Consider $X \sim f(x \mid \theta)$. **A function T of X (called a statistic of X) is said to be sufficient if the distribution of X conditional upon $T(X)$ is independent of θ.**



$$f(x \mid \theta) = h(T(x) \mid \theta)\, g(x \mid T(x))^*$$

- **A sufficient statistic $T(x)$ contains the whole information brought by x about θ.**

- Let us consider a simple example. Let $X=(X_1, X_2, \ldots, X_n)$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$. Then we can write:

$$f(x \mid \theta) = \prod_{j=1}^{N} \mathcal{N}(x_j \mid \theta) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_j - \mu)^2 \right) = \frac{1}{(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2} \sum_{j=1}^{N} (x_j - \theta_1)^2 \right) =$$

$$= \frac{1}{(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2} \sum_{j=1}^{N} x_j^2 + \frac{\theta_1}{\theta_2} \sum_{j=1}^{N} x_j - \frac{N\theta_1^2}{2\theta_2} \right)$$

* $f(x|\theta) = f(x, T(x)|\theta) = h(T(x)|\theta)\, g(x|T(x), \theta) = h(T(x)|\theta)\, g(x|T(x))$

Sufficiency Principle

$$f(x\,|\,\theta)=\frac{1}{\left(2\pi\theta_2\right)^{N/2}}\exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N}x_j^2+\frac{\theta_1}{\theta_2}\sum_{j=1}^{N}x_j-\frac{N\theta_1^2}{2\theta_2}\right)$$

- We can see that $f(x\,|\,\theta)$ depends only on $T(x)=\left(\sum_{j=1}^{N}x_j,\sum_{j=1}^{N}x_j^2\right)$ which is our set of sufficient statistics.
- Introducing $\bar{x}=\frac{\sum_{j=1}^{N}x_j}{N}$, $s^2=\sum_{j=1}^{N}(x_j-\bar{x})^2$, we can also re-write the above equation:

$$f(x|\theta)=\frac{1}{(2\pi\theta_2)^{N/2}}\exp\left(-\frac{s^2+N\bar{x}^2}{2\theta_2}+\frac{\theta_1 N\bar{x}}{\theta_2}-\frac{N\theta_1^2}{2\theta_2}\right)$$

So $\bar{x}, s^2$ is an alternative set of sufficient statistics.

Sufficiency Principle: Example

- Consider $x_1 \sim \mathcal{B}(n_1, p)$, $x_2 \sim \mathcal{B}(n_2, p)$, and $x_3 \sim \mathcal{B}(n_3, p)$, three binomial independent observations when the sample sizes n_1, n_2, *and* n_3 are known.

- The likelihood function is then

$$f(x_1, x_2, x_3 \mid p) = \binom{n_1}{x_1}\binom{n_2}{x_2}\binom{n_3}{x_3} p^{x_1+x_2+x_3} (1-p)^{n_1+n_2+n_3-(x_1+x_2+x_3)}$$

and the sufficient statistics is

$T_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$ *or alternatively*

$T_2(x_1, x_2, x_3) = (x_1 + x_2 + x_3) / (n_1 + n_2 + n_3)$

- Note that $x_1 / n_1 + x_2 / n_2 + x_3 / n_3$ is not a sufficient statistic!

Sufficiency Principle: Example

- Let $X=(X_1,X_2,\ldots,X_n)$ be i.i.d. from $\mathcal{U}(0,\theta)$ with density

$$f(x_i \mid \theta) = \frac{1}{\theta}\mathbb{I}_{[0,\theta]}(x_i)$$

- Then the likelihood can be written as:[a]

$$\ell(\theta \mid x) \equiv f(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{j=1}^{N} f(x_j \mid \theta) = \frac{1}{\theta^N}\mathbb{I}_{[\max\{x_i\},\infty]}(\theta)$$

- So the sufficient statistic is $T(X)=\max\{X_i\}$.

[a] Note the following:

$$0 \le x_1 \le \theta, \ldots, 0 \le x_N \le \theta \Rightarrow$$
$$\max_{i=1,\ldots,N}\{x_i\} \le \theta < \infty.$$

Sufficiency Principle: Example

- Let $X=(X_1,X_2,\ldots,X_n)$ be i.i.d. from the Poisson density

$$f(x_i \mid \theta) = e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

- Then the likelihood can be written as:

$$\ell(\theta \mid x) \equiv f(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{j=1}^{N} e^{-\theta} \frac{\theta^{x_j}}{x_j!} = e^{-N\theta} \frac{\theta^{\sum_{j=1}^{N} x_j}}{\prod_{j=1}^{N} x_j!}$$

- Thus the sufficient statistic is:

$$T(x) = \sum_{j=1}^{N} x_j$$

Sufficiency Principle and the MLE

- **Sufficiency principle:** Two observations *x* and y have the same values of *T (x) = T (y) of statistics sufficient for* $f(.|\theta)$. Then the inferences about θ based on x and y should be the same.

- Consider the model $X_i \sim \mathcal{N}(\mu, 1)$ and we want to estimate μ (our θ) based on *N* data. In this case, the sufficient statistic is

$$T(x_{1:N}) = \sum_{j=1}^{N} x_j$$

- Consider the (MLE) estimate of θ: $\hat{\mu}_1 = \sum_{j=1}^{N} x_j \Big/ N$. It satisfies the sufficiency principle because if we have another dataset $x'_{1:N}$ such that

$$T(x_{1:N}) = T(x'_{1:N}) \text{ then } \hat{\mu}_2 = \frac{1}{N}\sum_{j=1}^{N} x'_j = \frac{1}{N}\sum_{j=1}^{N} x_j = \hat{\mu}_1$$

- On the other hand, the estimate $\hat{\mu}_1 = x_1$ does not satisfy the sufficiency principle for $n > 1$ because if we have another dataset $x'_{1:N}$ such that $T(x_{1:N}) = T(x'_{1:N})$, then $\hat{\mu}_2 = x'_1 \neq \hat{\mu}_1, \text{ if } x'_1 \neq x_1$.

Sufficiency Principle: Rao-Blackwell theorem

- The Sufficiently principle is well accepted because of the Rao-Blackwell theorem.

- **A Rao–Blackwell estimator** $\delta_{RB}(X)$ of an unobservable quantity θ is the conditional expectation $\mathbb{E}(\delta(X) \mid T(X))$ of some estimator $\delta(X)$ given a sufficient statistic $T(X)$. Call $\delta(X)$ the "original estimator" and $\delta_{RB}(X)$ the "improved estimator".

- ***Rao-Blackwell theorem. Let $\delta(X)$ be an unbiased estimate of θ***

  and $\delta_{RB}(X) = \mathbb{E}[\delta(X) \mid T(X)]$, then $\delta_{RB}(X)$ is unbiased and

  $var[\delta_{RB}(X)] \leq var[\delta(X)]$

Proof: $var[\delta(X)] = \mathbb{E}[var[\delta(X) \mid T(X)]] + var[\mathbb{E}[\delta(X) \mid T(X)]]$

$= \mathbb{E}[var[\delta(X) \mid T(X)]] + var[\delta_{RB}(X)] \geq var[\delta_{RB}(X)]$.

Variance Decomposition

- In the proof of the Rao-Blackwell theorem, we used the following identity for two random variables:

$$\text{var}(X) = \mathbb{E}(\text{var}(X|Y)) + \text{var}(\mathbb{E}(X|Y))$$

- This can be easily proved as follows (use the conditional expectation derivation introduced in an earlier lecture):

$$\begin{aligned}
\text{var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \\
&= \mathbb{E}\Big(\mathbb{E}(X^2|Y)\Big) - \Big(\mathbb{E}\Big(\mathbb{E}(X|Y)\Big)\Big)^2 = \\
&= \mathbb{E}\Big(\mathbb{E}(X^2|Y)\Big) - \mathbb{E}\Big(\big(\mathbb{E}(X|Y)\big)^2\Big) + \underbrace{\mathbb{E}\Big(\big(\mathbb{E}(X|Y)\big)^2\Big) - \Big(\mathbb{E}\Big(\mathbb{E}(X|Y)\Big)\Big)^2}_{\text{var}\left(E(X|Y)\right)} \\
&= \underbrace{\mathbb{E}\Big(\mathbb{E}(X^2|Y) - \big(\mathbb{E}(X|Y)\big)^2\Big)}_{\mathbb{E}\left(Var(X|Y)\right)} + \underbrace{\mathbb{E}\Big(\big(\mathbb{E}(X|Y)\big)^2\Big) - \Big(\mathbb{E}\Big(\mathbb{E}(X|Y)\Big)\Big)^2}_{\text{var}\left(\mathbb{E}(X|Y)\right)}
\end{aligned}$$

The Likelihood Principle

- **Likelihood Principle.** In the inference about θ, the information brought by an observation is entirely contained in the likelihood function $\ell(\theta \mid x) = f(x \mid \theta)$.

- Also, two likelihood functions contain the same information about θ if they are proportional to each other; i.e.

$$\ell_1(\theta \mid x) = c(x)\, \ell_2(\theta \mid x)$$

- It is straight forward to show that the MLE (maximum likelihood procedure) satisfies the likelihood principle

$$arg\ \max_{\theta} \ell_1(\theta \mid x) = arg\ \max_{\theta} \ell_2(\theta \mid x)$$

- Classical approaches do not necessarily satisfy the likelihood principle.

The Likelihood Principle

$$\ell_1(\theta \mid x) = c(x)\,\ell_2(\theta \mid x)$$

- You can have two different probabilistic models for the data. However, if $\ell_1(\theta \mid x) \propto \ell_2(\theta \mid x)$ then this should lead to the same inference.

- Some classical statistics procedures do not satisfy this principle because they rely on quantities such as

$$\Pr(X > \alpha) = \int_{\alpha}^{\infty} f(x \mid \theta)dx$$

whereas the likelihood principle does not bother about data you have not observed!

The Likelihood Principle

- We want to test *the unknown probability* θ of heads for a possibly biased coin. Consider the two hypotheses:[a]

$$H_0 : \theta = \frac{1}{2}, H_1 : \theta > \frac{1}{2}$$

- **Scenario 1:** We predetermine the number of flips $n = 12$ and then the number of heads $X \sim \mathcal{B}(n, \theta)$ (Binomial). If we collect $x = 9$ heads, then:

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^{12-9} = 220\theta^9(1-\theta)^3$$

For a frequentist, the p-value of the test is $P_\theta(X \geq 9|H_0) = 1\text{-}0.927 = 0.073$[b] *and* H_0 is not rejected at level $\alpha = 0.05$.[c]

[a] The Likelihood Principle, B. Vidakovic

[b] See binomial distribution tables. This is the probability to observe data with less compatibility with H_0 relative to the data collected. This is not $P(H_0)$ – In the Bayesian formalism, we can compute the posterior $P(H_0|\text{data})$.

[c] p-value ≡ Pr(data or data more extreme | H_0 true). The P value is NOT the probability that the hypothesis (or any other hypothesis) is right or wrong. In fact, it assumes the null hypothesis is right! The P value as it gets smaller-and smaller, provides stronger-and-stronger evidence against the null hypothesis. When the P value is small (less than the set level α), the evidence against the null hypothesis cannot easily be explained by chance ("statistical significance"). H_0 is then rejected (Mathematical Statistics and Data Analysis, J. A. Rice, Chapter 9)

The Likelihood Principle

- Suppose we want to test the unknown probability θ of heads for a possibly biased coin. Consider the two hypothesis:

$$H_0 : \theta = \frac{1}{2}, H_1 : \theta > \frac{1}{2}$$

- **Scenario 2:** We flip the coin until a predetermined number (3) of tails is observed (the flipping is continued until we observe 3 tails). Then $X \sim \mathcal{NB}(3, 1\text{-}\theta)$.[a] Let us assume that we collected $x = 9$ heads, then:

$$f(x|\theta) = \binom{3+x-1}{3-1}(1-\theta)^3\theta^x, \text{ for } x = 9, f(x|\theta) = 55\theta^9(1-\theta)^3$$

[a] Let p be the probability of success in a trial. The number of failures in a sequence of trials until r-th success is observed is a Negative Binomial $\mathcal{NB}$(r,p) with probability mass function

$$P(X = x) = \binom{r+x-1}{r-1}p^r(1-p)^x, \text{ for } x = 0,1,2,\ldots$$

For r = 1, the Negative Binomial distribution becomes the Geometric distribution, $\mathcal{NB}(1, p) = \mathcal{G}(p)$.

The Likelihood Principle

- For a frequentist, the *p-value of the test is* $P_\theta(X \geq 9|H_0) = 0.0327$[a]

$$P(X \geq 9 \mid H_0) = \sum_{x=9}^{\infty} \binom{3+x-1}{3-1} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^x = \left(\frac{1}{2}\right)^3 \frac{8+5\times 9+9^2}{2^9} = 0.0327$$

- Thus at scenario *2,* H_0 is rejected at level $\alpha = 0.05$ and this change in decision is not caused by observations! (still 9 heads and 3 tails)

- The likelihood principle is thus violated because the two scenarios lead to different hypotheses even though we have proportional likelihoods

$$f(x \mid \theta) \sim \theta^9 (1-\theta)^3$$

[a] Here, we used the identity:

$$\sum_{x=k}^{\infty} \binom{2+x}{2} \left(\frac{1}{2}\right)^x = \frac{8+5k+k^2}{2^k}$$

Frequentist Approach and the Likelihood Principle

- The frequentist approach focuses on the average behavior properties of procedures and thus justifies the use of an estimator for reasons that can contradict the Likelihood Principle.

- This theory (e.g. Neyman and Pearson, 1950) is called frequentist, because it evaluates statistical procedures according to their long-run performances (e.g. on the average or in frequency) rather than focusing on a procedure for the obtained data, as a conditional approach would do.

p-Values

- Consider the case $x \sim \mathcal{N}(\theta, 1)$ and the hypothesis to be tested: $H_0 : \theta = 0$

- The **Neyman-Pearson test procedure** at level 5% is to reject H_0 if

$$\frac{1}{N}\left|\sum_{i=1}^{N} x_i\right| > \frac{1.96}{\sqrt{N}}$$

on the basis that (see tables of normal distribution)

$$\Pr\left(\left|\sum_{i=1}^{N}\frac{1}{N}x_i - \theta\right| \geq \frac{1.96}{\sqrt{N}} \mid H_0\right) = \Pr\left(\left|\sum_{i=1}^{N}\frac{1}{N}x_i\right| \geq \frac{1.96}{\sqrt{N}} \mid H_0\right) = 2\times(1-0.975) = 0.05$$

- That is the decision is based on the event $\left|\sum_{i=1}^{N}\frac{1}{N}x_i\right| \geq \frac{1.96}{\sqrt{N}}$ rather than on the observations themselves. E.g., if we collect $x=1.96$, the hypothesis is rejected conditioning on $|x|\geq 1.96$ rather than $x=1.96$ (which is impossible using frequentist approach).

- Here $pvalue(1.96) = P(x \geq 1.96 | x \sim H_0) = P\left(\frac{x-0}{\frac{1}{\sqrt{1}}} \geq 1.96\right) = 0.05$

- The frequency argument is that in 5% of the cases when H_0 is true we reject wrongly the null hypothesis.

Stopping Rule Principle in Sequential Analysis

- Consider a sequence of experiments directed by a stopping rule which indicates when the experiments should stop. An implication of the likelihood principle is the stopping rule principle in sequential analysis.

- Consider a sequence of experiments that leads at time *i* to the observation $X_i \sim f(x_i \mid \theta)$ and we stop collecting data if at time *n* we have

$$(x_1, x_2, ..., x_n) \in A_n; e.g.\ A_n = \{x_1, x_2, ..., x_n : x_n > B\}.$$

In this case:

$$\ell(\theta \mid x_1, x_2, ..., x_n) = f(x_1 \mid \theta) f(x_2 \mid x_1, \theta) ... f(x_n \mid x_1, ..., x_{n-1}, \theta) \mathbb{I}_{A_n}(x_1, x_2, ..., x_n).$$

- **Stopping rule principle: Inference about θ must depend on the stopping rule only through the sample.**

Stopping Rule and p-Values

- The stopping rule principle is incompatible with frequentist modeling.
- Consider $X_i \sim \mathcal{N}(\theta, 1)$ and the hypothesis to be tested is $H_0 : \theta = 0$ and we stop collecting data at the first time N such that

$$\frac{1}{N}\left|\sum_{i=1}^{N} x_i\right| > \frac{1.96}{\sqrt{N}}$$

- The resulting sample will always reject $H_0 : \theta = 0$ at the level 5%.
- A Bayesian approach avoids this difficulty.

p-Values and the Likelihood Principle

- Consider the model X_1, X_2 *i.i.d.* $\sim \mathcal{N}(\theta, 1)$. The likelihood function is:

$$\ell(\theta \mid x_1, x_2) = f(x_1, x_2 \mid \theta) \propto \exp\left(-\left(\frac{x_1 + x_2}{2} - \theta\right)^2\right)$$

- Consider now the alternative distribution (with different tails):

$$g(x_1, x_2 \mid \theta) = \pi^{-3/2} \frac{\exp\left(-\left(\frac{x_1 + x_2}{2} - \theta\right)^2\right)}{1 + (x_1 - x_2)^2} \propto \ell(\theta \mid x_1, x_2)$$

- If computing p-values with the frequentist approach, then one will obtain different results for $f(x_1, x_2 \mid \theta)$ and $g(x_1, x_2 \mid \theta)$ *because* they have different tails. The likelihood principle is then violated since different inference results from proportional likelihood functions.

- **The likelihood principle does not bother about data you have not observed!**

Conditionality Principle

- Consider estimating θ in the model on basis of *2* observations, X_1 and X_2.

$$P_\theta(X = \theta - 1) = P_\theta(X = \theta + 1) = 1/2$$

- The procedure suggested is

$$\delta(X) = \begin{cases} \dfrac{X_1 + X_2}{2} & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2 \end{cases}$$

- For a frequentist, this procedure has confidence of 75%; i.e. $P(\delta(X) = \theta) = 0.75$ (note cases $(\theta - 1, \theta + 1)$, $(\theta + 1, \theta - 1)$, and $(\theta + 1, \theta + 1)$ result in the correct estimate).

- The conditionalist would look at the data (after they are collected) and report 100% confidence if observed data are different or 50% if the observations coincide.

- Does it make sense to report the pre-experiment accuracy which is known to be misleading once we collect the data?

Conditionality Principle

- The conditional perspective concerns reporting data specific measures of accuracy.

- In contrast to the frequentist approach, performance of statistical procedures are judged looking at the observed data.

  - **Conditionality Principle:** If two experiments on θ are available and if one of these experiments is selected with probability p independently of θ, then the resulting inference depends only on the selected experiment.[a]

- The likelihood principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.[b]

[a] The Likelihood Principle, B. Vidakovic

[b] For a proof see, C. P. Robert, *The Bayesian Choice, Springer, 2nd edition,* chapter 1 (full text available)

MLE: Summary

- The likelihood principle is fairly vague since it does not lead to the selection of a particular procedure.

- Maximum likelihood estimation is one way to implement the sufficiency and likelihood principles

$$\hat{\theta} = \arg\sup_{\theta} \ell(\theta|x)$$

- Indeed:

$$\arg\sup_{\theta} \ell(\theta \mid x) = \arg\sup_{\theta} h(x) g(T(x) \mid \theta) = \arg\sup_{\theta} g(T(x) \mid \theta)$$

$$\ell_1(\theta \mid x) = c(x)\ell_2(\theta \mid x) \Rightarrow \arg\sup_{\theta} \ell_1(\theta \mid x) = \arg\sup_{\theta} \ell_2(\theta \mid x)$$

Maximum Likelihood Estimator

- Maximum likelihood estimation is just one way to implement the likelihood principle.

- Maximization can be difficult and lead to several global maxima. However, consistent and efficient in most cases (asymptotic properties).

- **Main problem of MLE: ML estimates can vary widely for small variations of the observations in particular for small sample sizes.**

- Example: If $X_i \sim \frac{1}{\theta}\mathbb{I}_{[0,\theta]}(x_i)$ then for *N* data we have seen earlier that:

$$\ell(\theta|x) = \prod_{i=1}^{N} f(x_i|\theta) = \frac{1}{\theta^N}\mathbb{I}_{[\max\{x_i\},\infty]}(\theta) \Rightarrow \hat{\theta} = \max_{i=1,\ldots,N}\{X_i\}$$

Alternative Approaches

- Tests require frequentists justifications.
- Many approaches with Bayesian flavor have been proposed: penalized likelihood
  - Akaike Information Criterion – AIC
  - Stochastic complexity theory, etc.

Bayesian Statistics

- A Bayesian model is made of a parametric statistical model ($\mathcal{X}$, f (x| θ)) and a prior distribution on the parameters (Θ, π (θ)).

- The unknown parameters are now considered as random.

  - Some statisticians question this approach but most accept the probabilistic modeling on the observations.

- Example: Assume you want to measure the speed of light given some observations. Why should you put a prior on a physical constant?

  - Due to the limited accuracy of the measurement, this constant will never be known exactly.

  - It is thus justified to put a (e.g. uniform) prior on this parameter reflecting this uncertainty.

Bayesian Vs Frequentist Approach

- **In the Bayesian approach, probability describes degrees of belief.**

- In the frequentist interpretation, you should repeat an infinite number of times an experiment and the probabilities corresponds to the limiting frequencies.

  - But how do you attribute a probability to the event “There will be a major crush in the Dow index tomorrow”?

- The prior has an obvious impact on the inference and Bayesian statisticians are honest about it.

Bayes rule: Simple Example Prosecutor's Fallacy

- For events A and B, the Bayes' rule is

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

- Classical simple example: Prosecutors Fallacy - A zealous prosecutor has collected an evidence (that the accused is guilty) and has an expert testify that the probability of finding this evidence if the accused were innocent is one-in-a-million.

- The prosecutor concludes that the probability of the accused being innocent is one-in-a-million. Why this is wrong thinking?

Bayes Theorem $\longleftrightarrow$ Inversion of Probabilities

Simple Example: Prosecutor's Fallacy

- Assume no other evidence is available and the population is of 10 million people
- Defining *A*=`The accused is guilty', then $P(A)=10^{-7}$.
- Defining *B*=`Finding this evidence'
- Then $P(B|\bar{A}) = 10^{-6}$ and $P(B|A)=1$.
- Bayes' formula

$$P(A|B) = P(The\ accused\ is\ guilty|Evidence\ Found) =$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{1 \times 10^{-7}}{1 \times 10^{-7} + 10^{-6} \times (1 - 10^{-7})} \approx 0.1$$

Another Example of Bayes' rule

- Coming back from a trip, you feel sick and your doctor thinks you might have contracted a rare disease (0.01% of the population has the disease).

- A test is available but not perfect.

  - If a tested patient has the disease, 100% of the time the test will be positive.

  - If a tested patient does not have the disease, 95% of the the time the test will be negative (5% false positive).

- Your test is positive, should you really care?

Simple Example

- Let A=`the patient has the disease'
- Let B=`the test returns a positive result'

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{1 \times 0.0001}{1 \times 0.0001 + 0.05 \times 0.9999} \approx 0.002$$

- Such a test would be a complete waste of money

A Simple Example of Bayes' Rule

- Consider two events *A, B (with A^c, B^c their complements)*

$$\Pr[A \mid B] = \frac{\Pr[B \mid A]}{\Pr[B]} \Pr[A]$$

- If *A =`it is rainy', B =`it is windy' and Pr [A] = 0.1, Pr [B] = 0.2, Pr [B | A] = 0.5, then:*

$$\Pr[rainy \mid windy] = \frac{\Pr[windy \mid rainy]}{\Pr[windy]} \Pr[rainy] = \frac{0.5}{0.2} 0.1 = 0.25$$

- The fact that we know it is windy (prior information) increased the probability of being rainy by 2.5 times e.g.

$$\frac{\Pr[rainy \mid windy]}{\Pr[rainy]} = \frac{0.25}{0.1} = 2.5$$

Prior, Likelihood and Posterior

In the previous example(s), we can identify the following:

- Data *x (e.g.` it is windy')*
- Hypothesis h (e.g. `does it rain or not?'). We want to make inferences about *h.*

In Bayesian settings all variables are random and all inferences are probabilistic.

We identify three key ingredients of a Bayesian inference approach:

- Prior $\pi(h)$*:* How likely is hypothesis h before looking at the data
- Likelihood $f(x \mid h)$*:* How likely is to observe *x* assuming *h* is true.
- Posterior $\pi(h \mid x)$*:* How likely is *h* after data *x* have been observed.

$$\pi(h \mid x) = \frac{f(x \mid h)\pi(h)}{m(x)}$$

Prior $\pi(\theta)$

- We use the prior to introduce quantitatively some insights on the parameters of interest.
- This can be as subjective or as objective as you want it to be – and that's why frequentists do not like Bayesian approaches!
- There is no such a thing as a true prior!
- Even when prior information is heavily subjective, the Bayesian inference model is honest.

Likelihood $f(x|\theta)$

- The likelihood encapsulates the mathematical model of the physical phenomena you are investigating.
- If you know the input $X=x$ to your problem, the likelihood can represent the computed output $y=f(x)$.
- It is the most computational expensive part of Bayesian approaches to engineering inference problems (inverse problems).

Posterior $\pi(\theta|x)$: Inference and Prediction

- It combines the prior and likelihood.
- It weights the data and the prior information in making probabilistic inferences
- The posterior distribution is also useful in estimating the probability of observing a future outcome (prediction)

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

Sufficiency and Likelihood Principles

- Bayesian statistics satisfy automatically the sufficiency principle and the likelihood principle:

- **Sufficiency principle:** If $f(x \mid \theta) = h(x) g(T(x) \mid \theta)$

$$\pi(\theta \mid x) = \frac{h(x) g(T(x) \mid \theta) \pi(\theta)}{h(x) \int g(T(x) \mid \theta) \pi(\theta) d\theta} = \frac{g(T(x) \mid \theta) \pi(\theta)}{\int g(T(x) \mid \theta) \pi(\theta) d\theta}$$

- **Likelihood principle:** Assume we have $f_1(x \mid \theta) = c(x) f_2(x \mid \theta)$. Then

$$\pi(\theta \mid x) = \frac{f_1(x \mid \theta) \pi(\theta)}{\int f_1(x \mid \theta) \pi(\theta) d\theta} = \frac{c(x) f_2(x \mid \theta) \pi(\theta)}{\int c(x) f_2(x \mid \theta) \pi(\theta) d\theta} = \frac{f_2(x \mid \theta) \pi(\theta)}{\int f_2(x \mid \theta) \pi(\theta) d\theta}$$

Bayesian Model

- A Bayesian statistical model is made of
  - a likelihood

$$f(x|\theta)$$

  - a prior distribution on the parameters, $\pi(\theta)$ on Θ

- Justifications
  - Move from an unknown (deterministic) θ to random θ
  - Extract information/knowledge on θ contained in the observation x
  - Allows incorporation of imperfect/imprecise information in the decision process
  - Unique mathematical way to condition upon the observations (conditional perspective)

Bayesian Inference and the Posterior Distribution

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

- Operates conditional upon the observations
- Integrates simultaneously prior information/knowledge and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected (sequential nature)

Posterior Inference: Point Estimates

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

Maximum A Posteriori estimate (MAP)

$$\theta^* = \arg\max_{\theta} \log\left(\pi(\theta \mid x)\right) = \arg\max_{\theta} \left(\log \pi(x \mid \theta) + \log \pi(\theta)\right)$$

Posterior Mean

$$\hat{\theta} = \mathbb{E}_{p(\theta|x)}[\theta] = \int \theta \pi(\theta|x) d\theta$$

Posterior Quantiles

$$\Pr\left[\theta > a\right] = \int_a^{\infty} \pi(\theta \mid x) d\theta$$

Bayes versus MLE

- Information provided by the Bayesian approach is much richer than the simple MLE estimate.

- You can compute for example posterior probabilities

$$\Pr\left(\theta \in A \,|\, x_1,...,x_N\right)$$

or $Var\left(\theta \,|\, x_1,...,x_N\right)$

- Also you can do prediction of future data

$$f\left(x \,|\, x_1,...,x_N\right)$$

Hypothesis Testing the Bayesian Way

Consider two hypotheses in coin flipping:

- Coin is fair with prior $\pi(h_1)$
- Coin always produces tails with prior $\pi(h_2)$

We flip the coin 5 times and obtain data $\boldsymbol{x} = \{HTHTT\}$

Inferencing: We want assess the validity of the two hypotheses

Likelihood $f(\boldsymbol{x} \mid \boldsymbol{h}_i)$:

$$f(\boldsymbol{x} \mid h_1) = \frac{1}{2^5}, \quad f(\boldsymbol{x} \mid h_2) = 0$$

Posterior:

$$\frac{\pi(h_1 \mid \boldsymbol{x})}{\pi(h_2 \mid \boldsymbol{x})} = \frac{f(\boldsymbol{x} \mid h_1)}{f(\boldsymbol{x} \mid h_2)} \frac{\pi(h_1)}{\pi(h_2)} \to \infty$$

✓ **Here there is no effect of the prior**

Hypothesis Testing the Bayesian Way

Consider two hypothesis in coin flipping:

- Coin is fair with prior $\pi(h_1)$
- Coin always produces tails with prior $\pi(h_2)$

We flip the coin 5 times and obtain data $\boldsymbol{x} = \{TTTTT\}$

Inferencing: We want to assess the validity of the two hypotheses

Likelihood $f(\boldsymbol{x} \mid \boldsymbol{h_i})$:

$$f(\boldsymbol{x} \mid h_1) = \frac{1}{2^5}, \quad f(\boldsymbol{x} \mid h_2) = 1$$

Posterior

$$\frac{\pi(h_1 \mid \boldsymbol{x})}{\pi(h_2 \mid \boldsymbol{x})} = \frac{f(\boldsymbol{x} \mid h_1)}{f(\boldsymbol{x} \mid h_2)} \frac{\pi(h_1)}{\pi(h_2)} = \frac{1}{32} \frac{\pi(h_1)}{\pi(h_2)}$$

✓ **The data (evidence) points to `tails' but the posterior inferences also depend strongly on the priors!**

Parametric Bayesian Models

Let θ the probability that the coin will draw heads

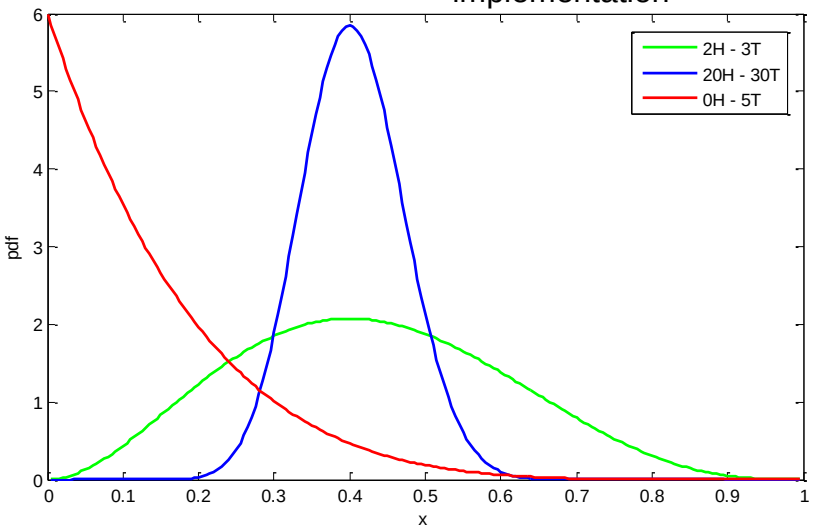
Let $\pi(\theta)$ *be the prior for* $\theta \in [0,1]$ *here taken as:*

$$\pi(\theta) = \mathbb{I}_{[0,1]}(\theta) \;\; (uniform)$$

Consider that we have the data ***x* = *{HTHTT}.*** We want to make an inference about θ?

Likelihood: $f(\boldsymbol{x} \mid \theta) = \theta^2 (1-\theta)^3 \;\; (binomial)$

Click here for a MatLab implementation



Posterior:

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{m(\boldsymbol{x})} =$$

$$= \frac{\theta^2 (1-\theta)^3}{beta(3,4)} \mathbb{I}_{[0,1]}(\theta) = \mathcal{B}(3,4)$$

Parametric Bayesian Models

Let θ the probability that the coin will draw heads

Suppose that $\mathcal{B}(a,b)$ is now our prior distribution:

$$\pi(\theta) = \frac{1}{beta(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

We are given data $\boldsymbol{x} = \{HTHTT\}$

Likelihood:

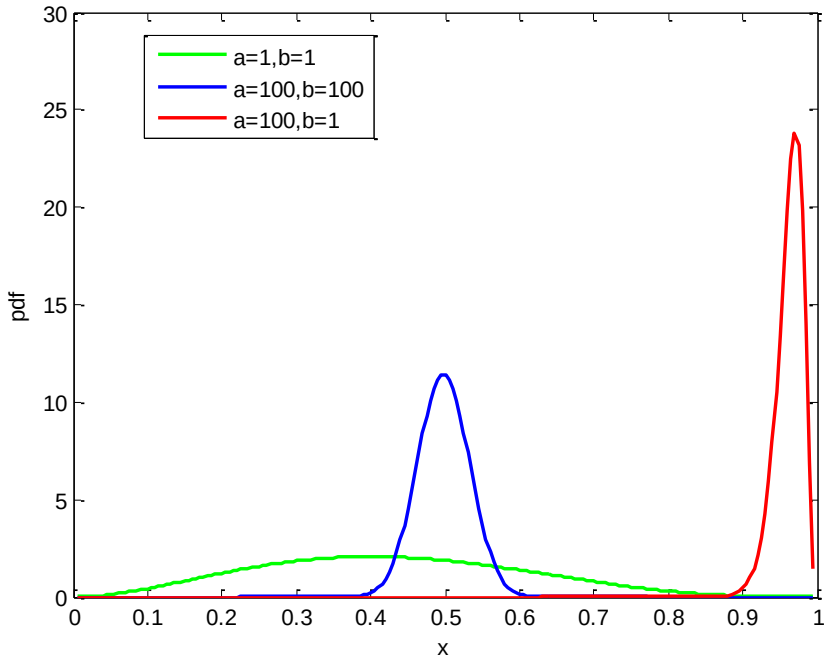
$$f(\boldsymbol{x} \mid \theta) = \theta^2(1-\theta)^3 \;\; (binomial)$$

Posterior:

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{m(\boldsymbol{x})} =$$

$$= \frac{\theta^{a+2-1}(1-\theta)^{b+3-1}}{beta(a+2,b+3)}\mathbb{I}_{[0,1]}(\theta) = \mathcal{B}(a+2,b+3)$$

Click here for a MatLab implementation



Parametric Bayesian Models

Data given: The coin is flipped n times and n_H of those came to be heads

Prior: We consider the Beta prior $\mathcal{B}(a,b)$ as in the earlier slide.

Posterior:

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{m(\boldsymbol{x})} = \frac{\theta^{a+n_H-1}(1-\theta)^{b+n-n_H-1}}{beta(a+n_H, b+n-n_H)} = \mathcal{B}(a+n_H, b+n-n_H)$$

Posterior mean: $\mathbb{E}[\theta \mid \boldsymbol{x}] = \frac{a+n_H}{a+b+n}$

Posterior variance: $Var[\theta \mid \boldsymbol{x}] = \frac{(a+n_H)(b+n-n_H)}{(a+b+n)^2(a+b+n+1)}$

Note that:

$$n \to \infty, \mathbb{E}[\theta \mid \boldsymbol{x}] \to \frac{n_H}{n}$$

$$Var[\theta \mid \boldsymbol{x}] \to 0 \; as \; \mathcal{O}(1/n)$$

Prediction

Suppose we have observed $\boldsymbol{x}$ and we want to make a prediction about (future) unknown observables: What is the probability of observing data $\hat{\boldsymbol{x}}$ If we already have observed data $\boldsymbol{x}$?

This means finding $g(\hat{\boldsymbol{x}}|\boldsymbol{x})$

We have:

$$g(\hat{\boldsymbol{x}}|\boldsymbol{x}) = \int g(\hat{\boldsymbol{x}}, \theta|\boldsymbol{x})d\theta = \int \frac{\pi(\hat{\boldsymbol{x}}, \theta, \boldsymbol{x})}{m(\boldsymbol{x})} d\theta = \int \frac{\pi(\hat{\boldsymbol{x}}, \theta, \boldsymbol{x})}{\phi(\theta, \boldsymbol{x})} \frac{\phi(\theta, \boldsymbol{x})}{m(\boldsymbol{x})} d\theta =$$

$$= \int f(\hat{\boldsymbol{x}}|\theta, \boldsymbol{x})\pi(\theta|\boldsymbol{x})d\theta = \int f(\hat{\boldsymbol{x}}|\theta)\pi(\theta|\boldsymbol{x})d\theta$$

Compare this with the normalizing factor:

$$m(\hat{\boldsymbol{x}}) = \int f(\hat{\boldsymbol{x}}|\theta)\pi(\theta)d\theta$$

Prediction: Example

Consider the coin flipping example

Let θ the probability that the coin draws heads

Consider $\pi(\theta)$ a uniform prior for θ

Given data **x = {**HTHTT**}**

We have seen that the posterior is $\mathcal{B}(3, 4)$

What is the probability that the next draw $\hat{\boldsymbol{x}}$ will be heads?

$$g(\hat{\boldsymbol{x}} = H|\boldsymbol{x}) = \int f(\hat{\boldsymbol{x}}|\theta)\pi(\theta|\boldsymbol{x})d\theta = \int \theta \frac{\theta^2(1-\theta)^3}{beta(3,4)} d\theta = \frac{3}{7}$$

Prediction

Consider the coin flipping example

Let θ the probability that the coin will draw heads

Consider $\pi(\theta)$ a uniform prior for θ

Given data $\boldsymbol{x} = \{HTHTT\}$

We have seen that the posterior is $\mathcal{B}(3, 4)$

What is the probability that the next 5 draws $\hat{\boldsymbol{x}}$ *will be all* heads?

$$g(\hat{\boldsymbol{x}} = HHHHH|\boldsymbol{x}) = \int g(\hat{\boldsymbol{x}}|\theta)\pi(\theta|\boldsymbol{x})d\theta =$$

$$= \int \theta^5 \frac{\theta^2(1-\theta)^3}{beta(3,4)} d\theta = \frac{1}{22}$$

Simple Binomial Example

- A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at θ (our prior knowledge).
- A second ball O is then rolled *n* times under the same assumptions and *X* denotes the number the ball O stopped on the left of *W* (our likelihood). Given *X*, what inference can we make about θ?
- We have $X|\,\theta \sim \mathcal{B}(n,\,\theta)$ binomial distribution and select $\theta \sim \mathcal{U}[0,1]$ and:

$$\Pr(X = x \mid \theta) = f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \Rightarrow \pi(\theta \mid x) = \frac{\theta^x(1-\theta)^{n-x}\mathbb{I}_{[0,1]}(\theta)}{\underbrace{\int \theta^x(1-\theta)^{n-x}\,d\theta}_{\mathcal{B}(x+1,n+1-x)}}$$

Simple Binomial Example

$$\pi(\theta \mid x) = \mathcal{B}(x+1, n+1-x) = \frac{\theta^x (1-\theta)^{n-x} \mathbb{I}_{[0,1]}(\theta)}{\int \theta^x (1-\theta)^{n-x} d\theta}$$

- Our distribution is a Beta one:

$$\pi(\theta \mid x) = \mathcal{B}(x+1, n+1-x)$$

- Prediction: Given $X=x$, you roll the ball once more and the $\Pr(Y=1|\theta)=\theta$ (ball stops to the left of θ) can be computed as:

$$\Pr(Y=1 \mid x) = \int \Pr\left(Y=1 \mid \theta, x\right) \pi(\theta \mid x) d\theta$$

$$= \int \theta \pi(\theta \mid x) d\theta = \mathbb{E}\left[\theta \mid x\right] = \frac{x+1}{n+2}$$

where we used the result for the mean of a Beta distribution.

Simple Binomial Example

- Laplace developed independently such a model.
- From 1745 to 1770, 241945 girls and 251527 boys were born in Paris.
- Let θ be the probability that any birth is female, then n = 251527 + 241945

$$\Pr\left(\theta \geq 0.5 \,|\, x = 241{,}945\right) = \int_{0.5}^{\infty} \pi(\theta \,|\, x)\, d\theta = \int_{0.5}^{\infty} \mathcal{B}(x+1, n+1-x)\, d\theta =$$

$$= \int_{0.5}^{\infty} \mathcal{B}(x+1, n+1-x)\, d\theta \approx 1.15 \times 10^{-42}$$

- **This is completely different from the p-value since we dont integrate over observations we have never seen. The integration is in the parameter space.**

Testing Hypothesis in a Bayesian Framework

- Consider the problem where we have $\pi(\theta) = \mathcal{U}[0,1]$ and

$$\Pr(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \Rightarrow \pi(\theta \mid x) = \mathcal{B}(x+1, n+1-x)$$

- To test $H_0 : \theta \geq \frac{1}{2}$ *vs* $H_1 : \theta < \frac{1}{2}$ using the posterior, we simply compute:

$$\pi(H_0 \mid x) = 1 - \pi(H_1 \mid x) = \int_{1/2}^{1} \pi(\theta \mid x) d\theta$$

- Note the integration is in parameter space.

In Bayesian statistics you never integrate with respect to observations

- Contrary to a frequentist approach, hypothesis testing in Bayesian is never based on data you don't observe!

Bayes and the Poisson model

- Assume you have some counting observations $X_i \overset{i.i.d.}{\sim} \mathcal{P}(\theta)$, i.e.

$$f(x_i \mid \theta) = e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

- Assume we adopt a Gamma prior for θ, i.e. $\theta \sim \mathcal{G}(\alpha, \beta)$

$$\pi(\theta) = \mathcal{G}(\theta; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

- You can easily show that:

$$\pi(\theta \mid x_1, \ldots, x_N) = \mathcal{G}(\theta; \alpha + \sum_{i=1}^{N} x_i, \beta + N)$$

The Bayesian Analysis is Sequential

Using some given data ***x***, we computed the posterior $\pi(\theta \mid \boldsymbol{x})$.

If new data ***x**** *arrives, how can we* update our inference?

We assume that ***x*** *and* ***x**** *are conditionally independent on* θ *i.e.:*

$$f(\boldsymbol{x}, \boldsymbol{x}^* \mid \theta) = f(\boldsymbol{x} \mid \theta) f(\boldsymbol{x}^* \mid \theta)$$

The augmented posterior (based on both ***x*** *and* ***x****) *is:*

$$\pi(\theta \mid \boldsymbol{x}, \boldsymbol{x}^*) = \frac{\pi(\boldsymbol{x}, \boldsymbol{x}^* \mid \theta)\pi(\theta)}{\phi(\boldsymbol{x}, \boldsymbol{x}^*)} = \frac{f(\boldsymbol{x}^* \mid \theta)}{m(\boldsymbol{x}^*)} \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}^* \mid \theta)\pi(\theta \mid \boldsymbol{x})}{m(\boldsymbol{x}^*)}$$

- Note that the prior now is our old posterior computed with data ***x***.
- Thus Bayesian analysis is sequential.

$$\pi(\theta \mid \boldsymbol{x}, \boldsymbol{x}^*) = \frac{f(\boldsymbol{x}^* \mid \theta)\pi(\theta \mid \boldsymbol{x})}{m(\boldsymbol{x}^*)}$$

Sequential Nature of Bayesian Inference

- Assume we have observed $x_1 \sim \mathcal{N}(\theta, \sigma^2)$ and computed the corresponding posterior. Now we observe a second realization x_2 *of* $X_2 \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$.

- We are interested to update our posterior:

$$\pi(\theta \mid x_1, x_2) \propto f(x_2 \mid \theta, x_1)\pi(\theta \mid x_1) \propto f(x_2 \mid \theta) f(x_1 \mid \theta)\pi(\theta)$$
$$\propto f(x_2 \mid \theta)\pi(\theta \mid x_1)$$
$$\propto f(x_1 \mid \theta)\pi(\theta \mid x_2)$$

- Updating the prior one observation at a time or all observations together does not matter

- The sequential approach is useful for massive data sets, e.g.

$$\pi(\theta \mid x_1, x_2, ..., x_n) \propto f(x_n \mid \theta)\pi(\theta \mid x_1, x_2, ..., x_{n-1})$$

i.e. the prior at time *n* is the posterior at time *n-1*.

A Gaussian Example

- Consider $X_1 \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$, *with prior* $\theta \sim \mathcal{N}(m_0, \sigma_0^2)$.
- Then we can derive the following:

$$\pi(\theta \mid x_1) \propto f(x_1 \mid \theta)\pi(\theta) \propto \exp\left(-\frac{(x_1-\theta)^2}{2\sigma^2} - \frac{(\theta-m_0)^2}{2\sigma_0^2}\right) \Rightarrow$$

$$\pi(\theta \mid x_1) \propto \exp\left(-\frac{\theta^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)+\theta\left(\frac{x_1}{\sigma^2}+\frac{m_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\theta-m_1)^2\right) \Rightarrow$$

$$\theta \mid x_1 \sim \mathcal{N}(m_1, \sigma_1^2) \text{ with}$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}, \text{ and}$$

$$m_1 = \sigma_1^2\left(\frac{x_1}{\sigma^2}+\frac{m_0}{\sigma_0^2}\right)$$

A Gaussian Example: Continued

- To predict the distribution of a new observation $X \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$ in light of x_1, we use the predictive distribution as follows:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \theta)}_{Likelihood} \underbrace{\pi(\theta \mid x_1)}_{Posterior} d\theta = \int e^{-\frac{(x-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-m_1)^2}{2\sigma_1^2}} d\theta = \int e^{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta-m_1)^2}{\sigma_1^2}\right)} d\theta$$

- We use the properties of the bivariate normal distribution. The product in the integrand is the exponential of a quadratic function in (x,θ); hence (x,θ) have a joint normal distribution. One can verify that:

$$e^{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta-m_1)^2}{\sigma_1^2}\right)} \sim e^{-\frac{z}{2(1-\rho^2)}}, \; z = \frac{(x-m_1)^2}{\sigma_1^2+\sigma^2} + \frac{(\theta-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(\theta-m_1)}{\sqrt{\sigma_1^2+\sigma^2}\,\sigma_1}, \; \rho = \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma^2}}$$

- Thus the marginal (integrating θ) is a Gaussian with:

$$X \mid x_1 \sim \mathcal{N}(m_1, \sigma_1^2 + \sigma^2)$$

Gaussian Example: Bayes' versus MLE

- We have seen that the ML estimate of θ at time *n* is simply:

$$\theta_{ML} = \arg\sup_{\theta} \prod_{i=1}^{N} f(x_i \mid \theta) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The posterior of θ at time *n* is (simply generalizing the earlier result):

$$\theta \mid x_1, \ldots, x_N \sim \mathcal{N}(m_N, \sigma_N^2)$$

where

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2} \underset{N\to\infty}{\sim} \frac{\sigma^2}{N}$$

$$m_N = \sigma_N^2 \left(\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m_0}{\sigma_0^2} \right) \underset{N\to\infty}{\sim} \frac{\sum_{i=1}^{N} x_i}{N}$$

- As N→∞, the prior is washed out by the data and the posterior mean is the MLE estimate: $\mathbb{E}[\theta | x_1, \ldots, x_N] = m_N \simeq \theta_{ML}$

Example: Normal Distribution ($\sigma^2=1$)

- Consider our earlier model of a Gaussian likelihood with given variance $\sigma^2=1$.
- Introducing $\bar{x}=\frac{\sum_{j=1}^{N} x_j}{N}$, $s^2=\sum_{j=1}^{N}(x_j-\bar{x})^2$, we can also re-write the likelihood as (here we take only the mean as a parameter. i.e. $\theta=\theta_1$):

$$f(x_1,x_2,\ldots,x_N|\theta) \propto \exp\left(\theta N\bar{x}-\frac{N\theta^2}{2}\right)$$

- Consider the following prior on θ:

$$\pi(\theta) \propto \exp(-\theta)\mathbb{I}_{\theta>0}$$

- Then the posterior becomes a truncated Gaussian:

$$\pi(\theta|x_1,x_2,\ldots,x_N) \propto \exp\left(\theta(N\bar{x}-1)-\frac{N\theta^2}{2}\right) \propto \exp\left(-N\left(\theta-\left(\bar{x}-\frac{1}{N}\right)\right)^2/2\right)\mathbb{I}_{\theta>0}$$

$$\pi(\theta|x_1,x_2,\ldots,x_N) \propto \mathcal{N}_+\left(\left(\bar{x}-\frac{1}{N}\right),\frac{1}{N}\right)$$

Example (Normal Likelihood-Normal Prior)

- Consider $x|\theta \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(a, 10)$.

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right)$$

$$\propto \exp\left(-\frac{11\theta^2}{20} + \theta\left(x + \frac{a}{10}\right)\right)$$

$$\propto \exp\left(-\frac{11}{20}\left\{\theta - \left(\frac{(10x+a)}{11}\right)\right\}^2\right)$$

- The posterior is now given as:

$$\theta|x \propto \mathcal{N}\left(\frac{(10x+a)}{11}, \frac{10}{11}\right)$$

Marginal Likelihood or Evidence

- In the Bayesian parametric model, we define the following:
  - The joint distribution of (θ, X)

$$\phi(\theta, x) = \pi(\theta) f(x \mid \theta)$$

  - The marginal distribution of X

$$m(x) = \int \phi(\theta, x) d\theta = \int \pi(\theta) f(x \mid \theta) d\theta$$

  - For a realization $X=x$, $m(x)$ is called marginal likelihood or evidence

Normalizing Factor in Bayes' Rule

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

- What about *m(x)* (the marginal distribution on the data *x*)

$$m(x) = \int \phi(x, \theta) d\theta = \int f(x \mid \theta)\pi(\theta) d\theta$$

- It is the normalizing *constant of the posterior* $\pi(\theta \mid x)$
- It is not needed for inference or prediction
- It is essential for model validation (we will discuss this later on)

Predictive Distribution and Approximations

- Given the prior $\pi(\theta)$ and the likelihood $\ell(\theta|x)=f(x|\theta)$, Bayes' formula yields:[a]

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta}$$

- This represents all the information on θ that can be extracted from *x*.

- Note the integral at the demoninator of the Bayes' rule.

- The predictive distribution of *Y* when *Y*~*g*(*y*|θ,*x*) is (we will revisit this in a forthcoming lecture)

$$g(y \mid x) = \int g(y, \theta \mid x)d\theta = \int \underbrace{g(y \mid \theta, x)}_{Likelihood} \underbrace{\pi(\theta \mid x)}_{Posterior} d\theta$$

- This is to distinguish from prediction using $\hat{\theta} = \theta^{MLE}$ or $\hat{\theta} = \theta^{MAP}$: $g(y|\hat{\theta}, x)$

[a] Probability, Conditional Probability and Bayes Formula, B. Vidakovic

Nuisance Parameters

- In case where $\theta = (\theta_1, \theta_2, ..., \theta_p)$ and one is only interested in the parameter θ_k then $\theta_{-k} = (\theta_1, \theta_2, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_p)$ are so called nuisance parameters.
- Bayesian inference tells us that all the information on θ_k that can be extracted from *x* is the marginal posterior distribution:

$$\pi(\theta_k \mid x) = \int ... \int \pi(\theta \mid x) d\theta_{-k}$$

- Computing $\pi(\theta_k \mid x)$ requires computing a high dimensional integral.
- Nuisance parameters can also be handled using profile likelihood technique in a maximum likelihood framework.

Exponential Family of Distributions

Exponential Family

- Large family of useful distributions with common properties
  - Bernoulli, beta, binomial, chi-square, Dirichlet, gamma, Gaussian, geometric, multinomial, Poisson, Weibull, . .
- Not in the family: Uniform, Student's T, Cauchy, Laplace, mixture of Gaussians, . . .
- Variable can be discrete/continuous (or vectors thereof)
- We will briefly introduce *the conditional setting in which we have a directed model X→Y with both X & Y observed and with p(Y|X) being an exponential family distribution parametrized using Generalized Linear Models (GLIM's)* (more on the topic on a forthcoming lecture).

Exponential Family

- The exponential family of distributions over $\boldsymbol{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \; or$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}, \; where: A(\boldsymbol{\eta}) = -\log g(\boldsymbol{\eta})$$

$\boldsymbol{x}$ is scalar/vector, discrete/continuous. ***η are the natural parameters and u(x) is referred to as a sufficient statistic.***

- *g(η) ensures that the distribution is normalized* and satisfies

$$g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} = 1$$

- The normalization factor Z and the log of it A are defined as:

$$Z(\boldsymbol{\eta}) = \frac{1}{g(\boldsymbol{\eta})}, \; A(\boldsymbol{\eta}) = \ln Z(\boldsymbol{\eta}) = -\ln g(\boldsymbol{\eta}) = \ln \int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x}$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}/Z(\boldsymbol{\eta})$$

- The space of η for which $\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} < \infty$ is the ***natural parameter space***.

Canonical or Natural Parameters

- When the parameter θ enters the exponential family as $\eta(\theta)$, we write the probability density of the exponential family as follows:

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = h(\boldsymbol{x})\, g\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right) \exp\left\{\boldsymbol{\eta}^T\left(\boldsymbol{\theta}\right) u(\boldsymbol{x})\right\} \; or$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T\left(\boldsymbol{\theta}\right) u(\boldsymbol{x}) - A\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right)\right\},$$

$$where: A\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right) = -\log g\left(\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\right)$$

- ***$\eta(\theta)$*** *are the canonical or natural parameters,*

- *θ is the parameter vector of some distribution that can be written in the exponential family format*

Joint Probability Distribution on Discrete RVs

- Any joint probability distribution on discrete random variables lies on the exponential family.* Indeed:

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{C\in\mathcal{C}}\log\Psi_C(\boldsymbol{x}_C)-\log Z(\Psi)\right)$$

$$\text{But for discrete rv's}:\Psi_C(\boldsymbol{x}_C)=\prod_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)^{\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)}$$

- Substitution to the 1st Eq. gives:

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{C\in\mathcal{C}}\sum_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)-\log Z(\Psi)\right)\Rightarrow$$

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\sum_{C\in\mathcal{C}}\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)-\log Z(\Psi)\right)$$

- This is in the exponential family with $\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)$ corresponding to each component of $\boldsymbol{\eta}$, and $\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_1,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)$ corresponding to components of the sufficient statistic $\mathbf{u}(x)$.

* We consider here the joint distribution of **x** written in terms of potentials Ψ's. We will see that this representation arises for rv's defined in undirected graphs where the potentials are defined over the random variables in each maximal clique c.

Exponential Family: The Bernoulli Distribution

- Consider the Bernoulli distribution:

$$p(x\,|\,\mu) = \mathcal{Bern}(x\,|\,\mu) = \mu^x(1-\mu)^{1-x} = \exp\left\{x\ln\mu + (1-x)\ln(1-\mu)\right\} =$$

$$= \underbrace{(1-\mu)}_{g(\eta)} \exp\left\{\underbrace{\ln\left(\frac{\mu}{1-\mu}\right)}_{\eta} x\right\}$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$
$$= h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

- From this we see that (note that *the relation $\mu(\eta)$ is invertible*)

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \Rightarrow \qquad \mu = \sigma(\eta) = \frac{1}{1+e^{-\eta}}$$

Logistic sigmoid function

and

$$g(\eta) = 1-\mu = 1-\sigma(\eta) = \sigma(-\eta)$$

- *Finally:*

$$p(x\,|\,\eta) = g(\eta)\exp\{\eta x\},\ u(x) = x,\ h(x) = 1,\ g(\eta) = \sigma(-\eta),$$
$$A(\eta) = -\log(1-\mu) = \log(1+e^{\eta})$$

Exponential Family: The Poisson Distribution

- Consider the Poisson distribution with parameter λ:

$$p(x \mid \lambda) = \textit{Poisson}(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \underbrace{\frac{1}{x!}}_{h(x)} \exp\left\{ \underbrace{\ln \lambda}_{\eta} \underbrace{x}_{u(x)} - \underbrace{\lambda}_{A(\eta)} \right\}$$

- Recall that λ is the mean of the distribution and observe once more that *the relation $\lambda(\eta)$ is invertible*:

$$\eta = \ln(\lambda) \Rightarrow \lambda = e^{\eta}$$

Exponential Family: The Multinoulli Distribution

- Consider the multinomial distribution:

$$p(\boldsymbol{x}\,|\,\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = \exp\left(\boldsymbol{\eta}^T \boldsymbol{x}\right),$$

$$\boldsymbol{x} = \left\{x_1, \ldots, x_M\right\}^T, \boldsymbol{\eta} = \left\{\eta_1, \ldots, \eta_M\right\}^T, \eta_k = \ln \mu_k$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$
$$= h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

- From this expression we see that $h(\boldsymbol{x})=1$, $u(\boldsymbol{x})=\boldsymbol{x}$, $g(\boldsymbol{\eta})=1$. It appears also that $A(\eta)=0$!

- We can resolve this problem by accounting for the dependence of μ_k, i.e. $\sum_{k=1}^{M} \mu_k = 1.$

Exponential Family: The Multinoulli Distribution

- We will express the distribution in terms of μ_k, k=1,…,M-1 subject to:

$$0 \le \mu_k \le 1, \quad \sum_{k=1}^{M-1} \mu_k \le 1$$

- The multinomial distribution becomes:

$$\exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} =$$

$$\exp\left\{\sum_{k=1}^{M-1} x_k \underbrace{\ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k}}_{\eta_k} + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\}$$

Exponential Family: The Multinoulli Distribution

- We identify

$$\eta_k = \ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} = \ln \frac{\mu_k}{\mu_M}, k = 1, \ldots, M-1$$

- Can also define:

$$\eta_M = \ln \frac{\mu_M}{\mu_M} = 0$$

- This equation can be inverted as:

$$\exp(\eta_k) = \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow \sum_{k=1}^{M-1} \exp(\eta_k) = \frac{\sum_{k=1}^{M-1} \mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow$$

$$1 + \sum_{k=1}^{M-1} \exp(\eta_k) = \frac{1}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow \sum_{k=1}^{M-1} \mu_k = \frac{\sum_{k=1}^{M-1} \exp(\eta_k)}{1 + \sum_{k} \exp(\eta_k)} \Rightarrow 1 - \sum_{k=1}^{M-1} \mu_k = \left(1 + \sum_{k} \exp(\eta_k)\right)^{-1}$$

- Substitution intro the expression on the top of the slide:

$$\eta_k = \ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} = \ln \left[\mu_k \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right) \right] \Rightarrow \mu_k = \frac{\exp(\eta_k)}{1 + \sum_{k=1}^{M-1} \exp(\eta_k)}$$

Exponential Family: The Multinoulli Distribution

- This is the so called the softmax function (note again *the relation $\mu(\eta)$ is invertible)*:

$$\mu_k = \frac{\exp(\eta_k)}{1+\sum_{k=1}^{M-1}\exp(\eta_k)}$$

Softmax function

- In this reduced representation, the distribution takes the form:

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = \exp\left\{\sum_{k=1}^{M-1} x_k\eta_k + \ln\left(1-\sum_{k=1}^{M-1}\mu_k\right)\right\} = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\boldsymbol{\eta}^T\boldsymbol{x})$$

- Comparing with the generic form of the exponential family:

$$\boldsymbol{\eta} = (\eta_1,\ldots,\eta_{M-1},0)^T,\ u(\boldsymbol{x}) = \boldsymbol{x},\ h(\boldsymbol{x}) = 1,\ g(\boldsymbol{\eta}) = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}$$

$$A = -\ln g(\boldsymbol{\eta}) = \ln\left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right) = \ln\left(\sum_{k=1}^{M}\exp(\eta_k)\right)$$

Exponential Family: The Beta Distribution

- Consider the Beta distribution

$$\mathcal{Beta}(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left[(a-1)\ln\mu + (b-1)\ln(1-\mu)\right]$$

- Comparing this with our exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$
$$= h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

we can easily indentify:

$$u(\mu) = (\ln\mu, \ln(1-\mu))^T,\ \boldsymbol{\eta} = (a-1, b-1)^T,\ h(\mu) = 1,\ g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},$$

$$A(a,b) = \ln\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Exponential Family: Gamma Distribution

- Consider the Gamma distribution

$$\mathcal{Gamma}(\lambda \mid a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} = \frac{b^a}{\Gamma(a)}\exp\left[(a-1)\ln\lambda - b\lambda\right]$$

- Comparing this with our exponential family:

$$\begin{aligned} p(\boldsymbol{x}\mid\boldsymbol{\eta}) &= h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \\ &= h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\} \end{aligned}$$

we can easily indentify:

$$u(\lambda) = (\lambda, \ln\lambda)^T,\ \boldsymbol{\eta} = (-b, a-1)^T,\ h(\lambda) = 1,\ g(a,b) = \frac{b^a}{\Gamma(a)},\ A(a,b) = \ln\frac{\Gamma(a)}{b^a}$$

Exponential Family: The Gaussian

- Consider the univariate Gaussian

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}x\right\}$$

- Comparing this with our exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

we can indentify (this is a two parameter distribution):

$$u(x) = (x, x^2)^T, \boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, h(x) = \frac{1}{\sqrt{2\pi}}, g(\boldsymbol{\eta}) = \left(-2\eta_2\right)^{1/2} \exp\frac{\eta_1^2}{4\eta_2}$$

$$A(\boldsymbol{\eta}) = -\frac{1}{2}\ln\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}$$

The Multivariate Gaussian

- The exponent in the multivariate Gaussian is:

$$-\frac{1}{2}tr\left(\Lambda xx^T\right)+\mu^T\Lambda x-\frac{1}{2}\mu^T\Lambda\mu=-\frac{1}{2}tr\left(\Lambda xx^T\right)+\xi^T x-\frac{1}{2}\xi^T\Lambda^{-1}\xi,\ where:\ \Lambda=\Sigma^{-1},\ \xi=\Lambda\mu$$

- We need to put this in the form $p(\boldsymbol{x}\,|\,\boldsymbol{\eta})=h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$
- The 3nd term contributes to g(η) whereas the 2nd term is directly an inner product between $\boldsymbol{x}$ and $\boldsymbol{\xi}=\boldsymbol{\Lambda\mu}$.
- For the 1st term, define two D^2-dimensional vector vec(Λ) and vec($\mathbf{xx}^T$) that consist of the columns of Λ and $\mathbf{xx}^T$, respectively. Then the 1st term has the form of an inner product between these two vectors. In summary:

$$\boldsymbol{\eta}=\begin{pmatrix}\boldsymbol{\xi}\\ vec(\Lambda)\end{pmatrix},\ u(\boldsymbol{x})=\begin{pmatrix}\boldsymbol{x}\\ -\frac{1}{2}vec(\boldsymbol{xx}^T)\end{pmatrix},\ g(\boldsymbol{\eta})=|\Lambda|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{\xi}^T\Lambda^{-1}\boldsymbol{\xi}\right),\ h(\boldsymbol{x})=(2\pi)^{-D/2},\ \boldsymbol{\xi}=\boldsymbol{\Lambda\mu}$$

$$A=-\ln g(\boldsymbol{\eta})=-\frac{1}{2}\ln|\Lambda|+\frac{1}{2}\boldsymbol{\xi}^T\Lambda^{-1}\boldsymbol{\xi}$$

Exponential Family: von Mises Distribution

- Consider the von Mises distribution

$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta_0\right)$$

- Comparing this with our exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

we can easily identify that:

$$u(\theta) = (\cos\theta, \sin\theta)^T, \boldsymbol{\eta} = (m\cos\theta_0, m\sin\theta_0)^T, h(\theta) = 1, g(m, \theta_0) = \frac{1}{2\pi I_0(m)},$$

$$A(m, \theta_0) = \ln\left(2\pi I_0(m)\right)$$

Computing Moments of Sufficient Statistics u(x)

- Differentiate wrt η the $\int p(\boldsymbol{x}|\boldsymbol{\eta})d\boldsymbol{x}=1$ for the exponential family:

$$p(\boldsymbol{x}|\boldsymbol{\eta})=h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$

$$\nabla g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right\}d\boldsymbol{x}+g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=0\Rightarrow$$

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}=g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- The above equation can be further simplified if written in terms of the partition function $Z=1/g(\eta)$ or $A=\log Z=-\log g(\eta)$:

$$\nabla A(\boldsymbol{\eta})=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- Let us re-write explicitly the above equation as:

$$\nabla A(\boldsymbol{\eta})=g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}$$

- We can compute the variance of $\mathbf{u}(\mathbf{x})$ by differentiating the Eq. above with respect to $\boldsymbol{\eta}$.

Computing Moments of Sufficient Statistics u(x)

$$\nabla A(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x}$$

$$\nabla^2 A(\boldsymbol{\eta}) = \underbrace{\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x}}_{-\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})^T\right]} + \underbrace{g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x}) \boldsymbol{u}(\boldsymbol{x})^T d\boldsymbol{x}}_{\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x}) \boldsymbol{u}(\boldsymbol{x})^T\right]}$$

- Thus the covariance of u(**x**) can be expressed in terms of the 2nd derivatives of $A(\boldsymbol{\eta})$ and similarly for higher order moments.

$$\nabla^2 A(\boldsymbol{\eta}) = \mathrm{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \; so \; A(\boldsymbol{\eta}) \; is \; convex$$

- Provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

Computing Moments of Sufficient Statistics u(x)

$$\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

$$\nabla^2 A(\boldsymbol{\eta}) = \mathrm{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \; so \; A(\boldsymbol{\eta}) \; is \; convex$$

- Let us check these relations for the Univariate Gaussian:

$$A(\boldsymbol{\eta}) = -\frac{1}{2}\ln\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}, \boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, u(x) = (x, x^2)^T$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}\left[X\right], \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \mu^2 + \sigma^2 = \mathbb{E}\left[X^2\right]$$

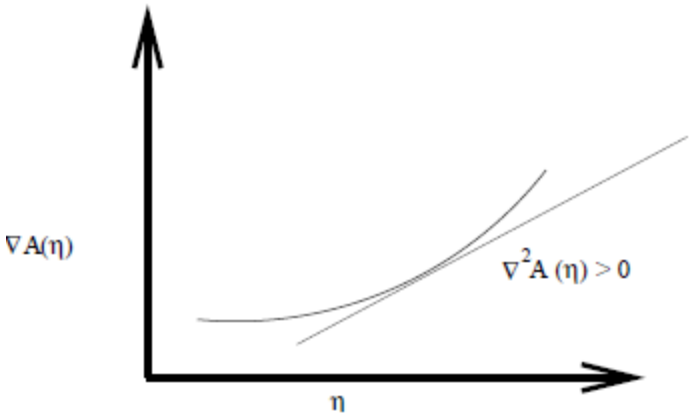
$$\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} = -\frac{1}{2\eta_2} = \sigma^2 = \mathrm{var}\left[X\right], etc.$$

Moment Parametrization

- We have shown that we can compute the mean of the distribution $\mu=\mathbb{E}[u(\mathbf{x})]$ in terms of the canonical parameter η:

$$\mu = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \nabla A(\boldsymbol{\eta})$$

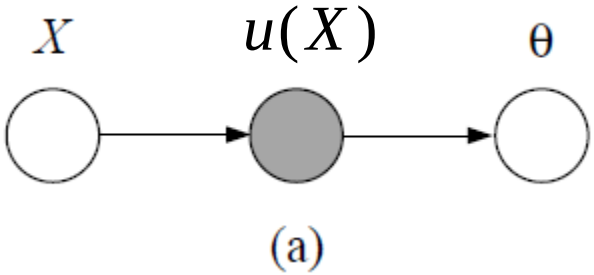
- We have also shown that $A(\eta)$ is a convex function. Since for a convex function there is one-to-one relation between the argument of the function and its derivative, the mapping $\mu(\eta)$ is invertable.



- Thus the exponential family of distributions can also be parameterized in terms of μ (*moment parametrization*) exactly as we started this course.

Sufficiency

- *u(X) is sufficient for θ if there is no information in X regarding θ beyond that in u(X).* Having observed u(X), we can throw away X for the purposes of inference with respect to θ.
- In the Bayesian approach in the Fig shown, we treat θ as an rv and say that u(X) is sufficient for θ if the following CI statement holds:



(a)

$$\theta \perp X \mid u(X)$$

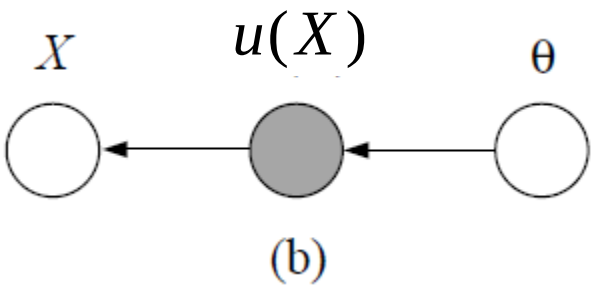
$$p\left(\theta \mid u(x), x\right) = p\left(\theta \mid u(x)\right)$$

- Thus, *u(X) contains all the needed information in X about θ.*

Frequentist Definition: Sufficiency

- The model in Fig b shown asserts the same CI relations as shown in Fig a earlier but has different parametrization.

$$p\left(x \mid u(x), \theta\right) = p\left(x \mid u(x)\right)$$



(b)

- Treating θ as a label, we can see the above CI statement as a frequentist definition of sufficiency.
- *u(X) is sufficient for θ if the p(x|u(x)) is not a function of θ.*
- The two approaches discussed imply a particular factorization of p(x|θ).

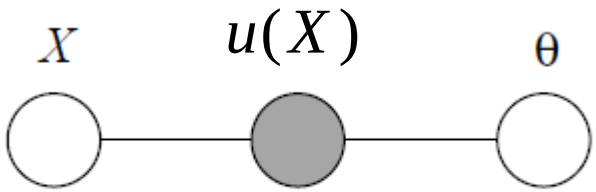
Neymann Factorization Theorem

- From the undirected graph (that expresses the same CI relations as the two earlier graphs), we can factorize as:

$$p\left(x,u(x),\theta\right)=g_1\left(u(x),\theta\right)g_2\left(x,u(x)\right)$$

- On the left, u(x) is a deterministic function of x and can be dropped as an argument:

$$p\left(x,\theta\right)=g_1\left(u(x),\theta\right)g_2\left(x,u(x)\right)$$



(c)

- One can derive for given ψ_1, ψ_2:

$$p\left(x\,|\,\theta\right)=p\left(x,\theta\right)/\,p\left(\theta\right)=\psi_1\left(u(x),\theta\right)\psi_2\left(x,u(x)\right)$$

- We can now see why u(x) was sufficient statistic for $\boldsymbol{\eta}$ in the exponential family:

$$p(\boldsymbol{x}\,|\,\theta)=\underbrace{h(\boldsymbol{x})}_{\psi_2(u(x),x)}\underbrace{\exp\left\{\boldsymbol{\eta}(\theta)^T u(\boldsymbol{x})-A\left(\boldsymbol{\eta}(\theta)\right)\right\}}_{\psi_1(u(x),\theta)}$$

MLE for the Exponential Family

- The joint density for a data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *is itself an exp. distribution with sufficient statistics* $\sum_{n=1}^{N} u(\boldsymbol{x}_n)$

$$p(\boldsymbol{X} \mid \boldsymbol{\eta}) = \prod_{n=1}^{N} \left(h(\boldsymbol{x}_n) g(\boldsymbol{\eta}) \exp\left\{ \boldsymbol{\eta}^T u(\boldsymbol{x}_n) \right\} \right) = \prod_{n=1}^{N} \left(h(\boldsymbol{x}_n) \right) g(\boldsymbol{\eta})^N \exp\left\{ \boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n) \right\} \Rightarrow$$

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\eta}) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) + N \ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) - NA(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- *The exponential family is the only family of distributions* ***with finite sufficient statistics*** *(size independent of the data set size).*
- *The log likelihood is concave (A convex) and has a unique maximum.*
- Maximizing wrt $\boldsymbol{\eta}$ gives: $\nabla A(\boldsymbol{\eta}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n) \Rightarrow \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$
- *At the MLE, the empirical average of the sufficient statistic is equal the model's theoretical expected sufficient statistics (moment matching).*
- *Thus to find the expected value of the sufficient statistics, one can use directly the data without having to estimate* $\boldsymbol{\eta}$. When $\mathbf{u}(\mathbf{x})=\mathbf{x}$, the above allows us to compute the expectation of x directly from the data.

MLE for the Exponential Family

$$\nabla A(\boldsymbol{\eta}_{ML}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N}\sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- Using the sufficient statistic, one can in principle invert the above equ. to compute $\boldsymbol{\eta}_{MLE}$. For example, for the Bernoulli distribution,

$$p(x\,|\,\eta) = g(\eta)\exp\{\eta x\},\ u(x) = x,\ h(x) = 1,$$

$$\mu = \frac{1}{1+e^{-\eta}},\ g(\eta) = \frac{1}{1+e^{\eta}},\ \eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

and thus:

$$\mathbb{E}[X] = p(X=1) = \bar{\mu} \equiv \mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(x_n = 1)$$

and

$$\eta_{MLE} = \ln\left(\frac{\bar{\mu}}{1-\bar{\mu}}\right)$$

MLE and Kullback-Leibler Distance

- A useful property for the MLE (and not just a property for the exponential family of distributions) is the following:
- Minimizing the KL distance to the empirical distribution is equivalent to maximizing the likelihood.
- Indeed, let us consider the model logp(x|θ) and the empirical distribution:

$$p_{emp}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta\left(x, x_n\right)$$

- We can then derive the following:

$$\sum_x p_{emp}(x)\log p(x\,|\,\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_x \delta\left(x, x_n\right)\log p(x\,|\,\theta) = \frac{1}{N}\sum_{n=1}^{N}\log p(x_n\,|\,\theta) = \frac{1}{N}\ell(\theta\,|\,\mathcal{D})$$

and from this:

$$KL\left(p_{emp}(x), p(x\,|\,\theta)\right) = \sum_x p_{emp}(x)\log\frac{p_{emp}(x)}{p(x\,|\,\theta)} = \sum_x p_{emp}(x)\log p_{emp}(x) - \sum_x p_{emp}(x)\log p(x\,|\,\theta)$$

$$= \sum_x p_{emp}(x)\log p_{emp}(x) - \frac{1}{N}\ell(\theta\,|\,\mathcal{D})$$

- Since the 1st term is independent of θ, the assertion is proved.

Conjugate Priors

- In general, for a given probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$, we can seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior.
  - For the Bernoulli, the conjugate prior is the Beta distribution
  - For the Gaussian, the conjugate prior for the mean is a Gaussian, and the conjugate prior for the precision is the Wishart distribution

Conjugate Priors

- For any member of the exponential family,

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = h(\boldsymbol{x})\,g(\boldsymbol{\eta}(\boldsymbol{\theta}))\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\,u(\boldsymbol{x})\right\}$$

there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) \propto g(\boldsymbol{\eta}(\boldsymbol{\theta}))^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\tau}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\},\ where:\ \boldsymbol{\tau}_0 \equiv \nu_0\bar{\boldsymbol{\tau}}_0$$

- In normalized form, we write:

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g(\boldsymbol{\eta}(\boldsymbol{\theta}))^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\}$$

$$where: Z(\nu_0,\boldsymbol{\tau}_0) = \int \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\}d\boldsymbol{\theta}$$

Conjugate Priors

$$p(\boldsymbol{X}\mid\boldsymbol{\theta})=\prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})u(\boldsymbol{x}_n)\right\}\right)=\prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)\right)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^N\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\sum_{n=1}^{N}u(\boldsymbol{x}_n)\right\}$$

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau_0})=\frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\}=\frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0-A\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\nu_0\}$$

- Using $\bar{\boldsymbol{u}}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)$, the posterior becomes (this form justifies $\bar{\boldsymbol{\tau}}_0$):

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{\chi},\nu)\propto g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)+\nu_0\bar{\boldsymbol{\tau}}_0\right)\right\}=g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\{\boldsymbol{\eta}^{\boldsymbol{T}}(\boldsymbol{\theta})(N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0)\}$$

- The parameter ν_0 can be interpreted as *effective number of fictitious observations* in the prior each of which has a value for the sufficient statistic equal to $\bar{\boldsymbol{\tau}}_0$.

$$p(\boldsymbol{\theta}|\boldsymbol{X},\nu_N,\boldsymbol{\tau_N})=\frac{1}{Z(\nu_N,\boldsymbol{\tau}_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\left\{(N+\nu_0)\boldsymbol{\eta}^T(\boldsymbol{\theta})\frac{N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0}{N+\nu_0}\right\}=\frac{1}{Z(\nu_N,\tau_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\{\nu_N\boldsymbol{\eta}^T(\theta)\bar{\boldsymbol{\tau}}_N\},$$

$$where\ \nu_N=\nu_0+N,\ \bar{\boldsymbol{\tau}}_N=\frac{N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0}{N+\nu_0},\boldsymbol{\tau}_N=\nu_N\bar{\boldsymbol{\tau}}_N=N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0=\sum_{i=1}^{N}\boldsymbol{u}(x_i)+\boldsymbol{\tau}_0$$

Posterior Predictive

- Let $\boldsymbol{u}(\boldsymbol{X}) = \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i), \boldsymbol{u}(\boldsymbol{X}') = \sum_{i=1}^{N'} \boldsymbol{u}(\boldsymbol{x}_i'),$ the posterior predictive is then:

$$p(\boldsymbol{X}' \mid \boldsymbol{X}) = \int p(\boldsymbol{X}' \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) d\boldsymbol{\theta}$$

$$= \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \int g(\boldsymbol{\eta})^{N'} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{X}')\right\} \frac{1}{Z(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0)} g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)\right\} d\boldsymbol{\theta}$$

- This is simplified as follows:

$$p(\boldsymbol{X}' \mid \boldsymbol{X}) = \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \frac{1}{Z\left(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)} \int g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{N' + \nu_N} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X}') + \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)\right\} d\boldsymbol{\theta}$$

$$= \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \frac{Z\left(\nu_0 + N + N', \boldsymbol{u}(\boldsymbol{X}') + \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)}{Z\left(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)}$$

- If *N=0*, this becomes the marginal likelihood of $\boldsymbol{X}'$, which reduces to the normalizer of the posterior divided by the normalizer of the prior multiplied by a constant.

Beta/Bernoulli: Posterior Predictive

- Consider a Bernoulli likelihood with a Beta prior. The likelihood takes the familiar exponential distribution form:

$$p(\mathcal{D}\mid\theta)=\theta^{\sum_i x_i}(1-\theta)^{N-\sum_i x_i}=(1-\theta)^N\exp\left(\log\frac{\theta}{1-\theta}\sum_i x_i\right)$$

- The conjugate prior is a Beta: $p(\theta\mid\nu_0,\tau_0)\propto(1-\theta)^{\nu_0}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_0\right)=\theta^{\tau_0}(1-\theta)^{\nu_0-\tau_0}$

$$p(\theta\mid\nu_0,\tau_0)=\mathcal{Beta}(\alpha,\beta),\alpha=\tau_0+1,\beta=\nu_0-\tau_0+1,$$

- Thus the posterior becomes: $p(\theta\mid\mathcal{D})\propto\theta^{\tau_0+s}(1-\theta)^{\nu_0-\tau_0+N-s}\Rightarrow$

$$p(\theta\mid\mathcal{D})=\mathcal{Beta}(\alpha_N,\beta_N),\alpha_N=\alpha+s,\beta_N=\beta+(N-s),s=\sum_i\mathbb{I}(x_i=1)$$

- Let s the number of heads in the past data. The probability of $s'=\sum_{i=1}^{m}\mathbb{I}(x_i'=1)$ future heads in m trials is then:

$$p(s'\mid\mathcal{D},m)=\int\theta^{s'}(1-\theta)^{m-s'}\frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\theta^{\alpha_N-1}(1-\theta)^{\beta_N-1}d\theta=\frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\frac{\Gamma(\alpha_{N+m})\Gamma(\beta_{N+m})}{\Gamma(\alpha_{N+m}+\beta_{N+m})}$$

$$\alpha_{N+m}=\alpha_N+s',\beta_{N+m}=\beta_N+(m-s')$$