An overview of Bayesian inference Part II

Thomas Hamelryck

Bioinformatics center, University of Copenhagen

September, 2017

The choice of prior distributions

- Priors are supposed to represent our state of belief before the data was observed.
- How should priors be chosen?
  - Is it a purely subjective matter?
- In many cases, the choice of the prior is dictated by objective criteria [1, 8, 9, 7].
  - Maximum entropy and relative entropy considerations.
    - We will need some information theory.
  - Invariance considerations.
  - Computational efficiency.
  - Analytic convenience.

Some information theory

- Developed by Claude Shannon (1916-2001) at the end of the 1940s to study the transmission of signals over noisy channels. Quickly became of fundamental importance in science, computing and engineering.
- In statistics, information theory is important for constructing priors and approximate inference.

Information

- Information quantifies the "surprise" associated with gaining knowledge about the value of a variable.
  - For a discrete variable x the information gain is

$$I(x) = -\log p(x)$$

  - If the log is taken with base 2, the unit is the *bit*.
  - if the log is taken with base e, the unit is the *nat*.

Example

- On tossing a coin, the chance of 'tail' is 0.5.
- If we learn that 'tail' occurred, this amounts to $-\log(0.5) = 1$ bit of information.

Information example

Unbiased versus biased dice

- You learn the outcome is 2 for an unbiased dice.

$$I(x=2) = -\log\left(\frac{1}{6}\right) \approx 1.79$$

- You learn the outcome is 2 for a biased dice, with $p(2) = \frac{1}{4} > \frac{1}{6}$.

$$I(x=2) = -\log\left(\frac{1}{4}\right) \approx 1.38$$

Information entropy

- The information or Shannon entropy S_x is the expectation of the information of a discrete variable x,

$$S_x = -\sum_x p(x) \log p(x) = -\mathbb{E}\left[\log p(x)\right]$$

- S_x reaches a maximum for the uniform case.
- S_x becomes zero if $p(x) = 1$ for one of the values of x.
- Does not generalize to the continuous case, which would be $-\int p(x) \log p(x) \mathrm{d}x$.
  - The "entropy" can become negative.
  - Not invariant under a transformation of variables.
  - The discrete Shannon entropy for $N \to \infty$ does not have the continuous case as limit.

Kullback-Leibler divergence

- The Kullback-Leibler divergence provides a relative entropy for the continuous case.
- A natural measure of "distance" between two probability distributions.

$$\mathrm{KL}\left[p \parallel q\right] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\mathrm{KL}\left[p \parallel q\right] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

- Not a real "distance" (metric) as it is not symmetric and does not respect the triangle inequality.

KL divergence interpretation

- If we have a posterior distribution p and a prior distribution π, the quantity

$$I = \mathrm{KL}[p \parallel \pi]$$

  is a measure of the information gained when one revises one's beliefs from the prior probability distribution π to the posterior probability distribution p.
- In other words, it is the amount of information lost when π is used to approximate p.

Mutual information

- Measures the how dependent two random variables are.
  - How much information do you gain about x if y is given?
- The mutual information measures the KL divergence $\mathrm{KL}\left[p(x,y) \parallel p(x)p(y)\right]$ between the joint distribution and the product of the marginals .

$$I_{x,y} = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I_{x,y} = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}y\mathrm{d}x$$

- The measure is non-negative and, unlike the KL, symmetric.

Choice of prior distributions

- The prior distribution ideally reflects the state of belief before any data was observed.
- It is often said that the choice of priors is necessarily "subjective", but this is not the case.
  - Maximum Entropy priors for the finite, discrete case.
  - Jeffreys priors for the univariate, continuous case.
  - Reference priors for the general case (at least in theory).
- Sometimes, in practice, priors are chosen for computational or analytical convenience.

Principle of indifference

- The *principle of indifference* or *principle of insufficient reason* goes all the way back to Bayes and Laplace.
  - For a variable that adopts K values, which are indistinguishable except for their label, the prior belief is $\frac{1}{K}$ for each value (ie. the uniform distribution).
- This is a special case of a **maximum entropy prior**.

Maximum entropy priors

- Consider a discrete, finite random variable $k = \{1, 2, \ldots, N\}$. What should we pick as prior for k given its mean $\bar{k}$?

$$\bar{k} = \sum_{k=1}^{N} p(k)k$$

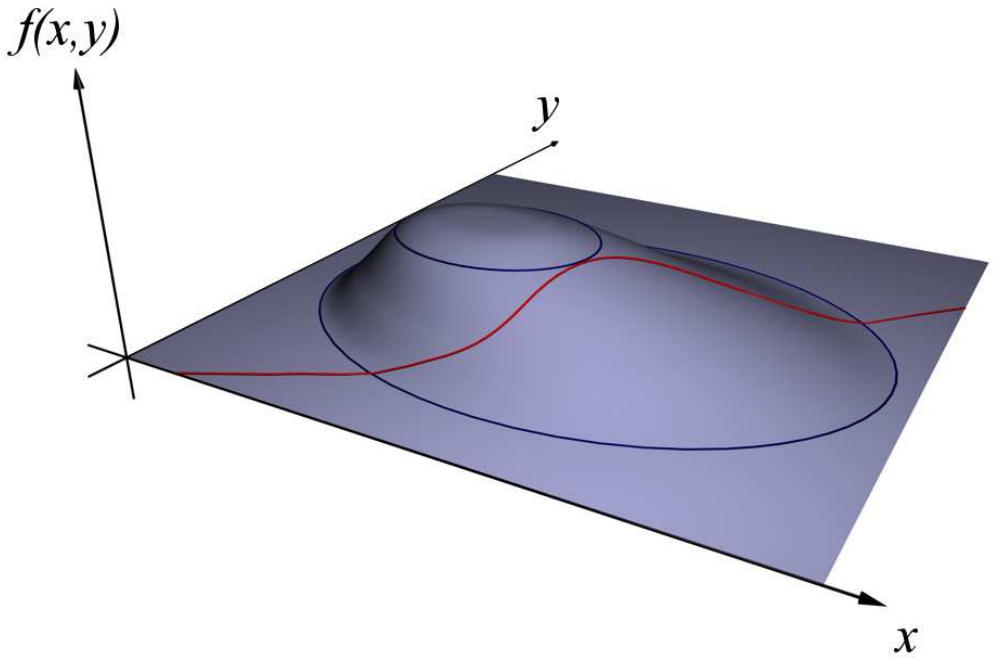
- The solution is to pick the distribution with maximum information entropy which is compatible with the mean [7].

Brandeis dice problem I

- The mean of an unbiased dice is 3.5, but we know the dice is biased with mean 4.5. What should be the prior over the 6 values of the dice?
- The method of **Lagrange multipliers** can be used to optimize a function under some constraints.

$$L = \overbrace{-\sum_{k=1}^{6} p_k \log p_k}^{\text{Entropy}} - \alpha \overbrace{\left(\sum_{k=1}^{6} p_k - 1\right)}^{\text{C1:Normalization}} - \beta \overbrace{\left(\sum_{k=1}^{6} p_k k - \bar{k}\right)}^{\text{C2:Mean}}$$

The method of Lagrange multipliers I



- Maximize $f(x, y)$ subject to the constraint $g(x, y) = c$ (shown in red).[1] The point where the red line tangentially touches the blue contour is the solution.

[1] Picture from Wikipedia.

The method of Lagrange multipliers II

- If $f(x_0, y_0)$ is a (constrained) maximum, then there exists α such that (x_0, y_0, α) is a stationary point for the Lagrange function,

$$L(x, y, \alpha) = f(x, y) - \alpha g(x, y).$$

  - Stationary points are those points where the partial derivatives of $L(x, y, \alpha)$ are zero.

$$\frac{\partial L(x, y, \alpha)}{\partial x} = \frac{\partial L(x, y, \alpha)}{\partial y} = 0$$

  - This is a *necessary (but not sufficient) condition* for optimality in constrained problems.

Brandeis dice problem II

- Solve the Lagrangian by setting the partial derivatives to zero.

$$\frac{\partial L(\mathbf{p}, \alpha, \beta)}{\partial p_k} = 0 = -\log(p_k) - 1 - \alpha - \beta k,$$

  which leads to $p_k = \exp(-1 - \alpha - \beta k)$.
- Eliminating α results in an exponential expression

$$p_k = \frac{1}{Z} \exp\left(-\beta k\right).$$

- β can be obtained numerically, which for $\bar{k} = 4.5$ results in

$$\mathbf{p} \approx \{0.05, 0.08, 0.11, 0.16, 0.24, 0.35\}$$

Jeffreys prior

- MaxEnt can break down in the continuous case.

Example: MaxEnt on the positive real line

- For $\theta \in \mathbb{R}^+$, the MaxEnt prior is the uniform distribution.
- If we reparametrize θ, the corresponding transformed uniform density will NOT necessarily be uniform.

- The Jeffreys prior is *invariant under reparametrization of* θ,

$$p(\theta) \propto \sqrt{\mathcal{I}(\theta)}$$

where $\mathcal{I}(\theta)$ is the Fisher information.

Why is the Jeffreys prior invariant?

- The Fisher information has the following property under a transformation of variables $y = f(x)$.

$$\mathcal{I}(y) = \mathcal{I}(x) \left(\frac{\mathrm{d}x}{\mathrm{d}y} \right)^2 \Rightarrow \sqrt{\mathcal{I}(y)} = \sqrt{\mathcal{I}(x)} \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right|$$

- Now let's examine what happens when we pick a prior $\pi(x) = \sqrt{\mathcal{I}(x)}$ under a change of variables $y = f(x)$

$$\begin{aligned} \pi(y) &= \pi(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \leftarrow \text{ (true for all priors)} \\ &\propto \sqrt{\mathcal{I}(x)} \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\ &= \sqrt{\mathcal{I}(y)} \end{aligned}$$

.

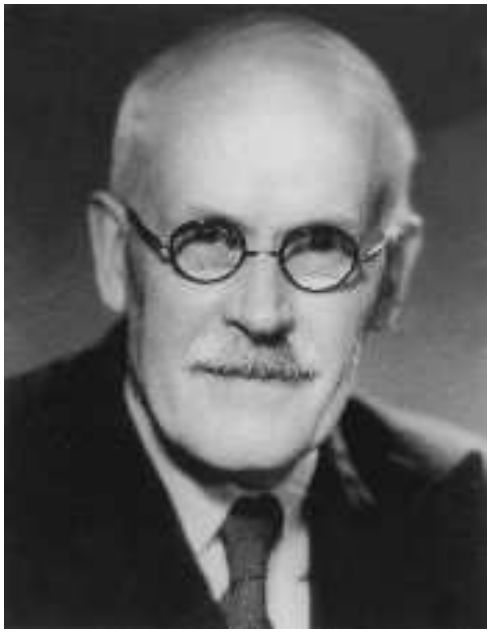
Jeffreys prior for a Gaussian with known μ

- The Jeffreys prior $\pi(\sigma) \propto \sqrt{\mathcal{I}(\sigma)}$ for a Gaussian with unknown σ and known μ is given by

$$\begin{aligned}
\pi(\sigma) &= \sqrt{\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2\right]} \\
&= \sqrt{\int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2) \underbrace{\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2}_{(\mathrm{d}\log p/\mathrm{d}\sigma)^2} \mathrm{d}x} \\
&= \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma}
\end{aligned}$$

- This is an *improper prior* - it's not a proper PDF.

Harold Jeffreys - Bayesian pioneer



- Sir Harold Jeffreys (1891-1989) was an English mathematician, statistician, geophysicist and astronomer.
- His book "Theory of Probability" (1939) played an important role in the revival of the Bayesian view of probability.

Reference priors

- Reference priors generalize MaxEnt and Jeffreys priors.
- Key idea: maximize the expected KL divergence between posterior and prior.

$$\mathbb{E}_{\mathbf{d}}\left[\int p(\boldsymbol{\theta} \mid \mathbf{d}) \log \frac{p(\boldsymbol{\theta} \mid \mathbf{d})}{\pi(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}\right]$$

- This maximizes the information brought in by the data. That is, the prior is maximally "vague".
- This expression is invariant under a change of variables.
- Often reference priors are improper (see below). This is fine as long as the posterior is a proper density.
- Can be seen as a sequential application of Jeffreys' rule for priors.

Empirical Bayes estimation of prior parameters

- In proper Bayesian inference, the parameters α of the prior are given and known.

$$p(h \mid d) \propto p(d \mid h)\pi(h \mid \alpha)$$

- If not, one can "cheat" and use a ML point estimate for the parameters of the prior.

$$\widehat{\alpha}_{\mathrm{EB}} = \arg\max_{\alpha} p(d \mid \alpha) = \int p(d \mid h)\pi(h \mid \alpha)\mathrm{d}h$$

- One obtains an approximation of a Bayesian posterior.

$$p_{\mathrm{EB}}(h \mid d) \propto p(d \mid h)\pi(h \mid \widehat{\alpha}_{\mathrm{EB}})$$

Conjugate priors

- A conjugate prior is a prior that has the convenient analytical property that the posterior has the same form as the prior.
  - Analytical convenience does not necessarily lead to a good choice for the prior!
- Some classic examples of likelihoods and their conjugate priors.
  - Binomial/Beta
  - Categorical/Dirichlet
  - Poisson/Gamma
  - Normal (μ unknown, σ^2 known)/Gamma
  - Exponential/Gamma

Example: the Binomial distribution with a Beta prior

- The binomial distribution gives the probability of k successes in n trials.

$$p(k \mid \theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- Posterior for a Beta prior

$$\begin{aligned} p(\theta \mid k, n) &\propto \overbrace{\theta^k (1-\theta)^{n-k}}^{\text{Likelihood}} \overbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{Beta prior}} \\ &= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \end{aligned}$$

- The posterior is also a Beta distribution. The parameters α and β of the Beta prior can be considered as added *pseudocounts* for success and failure.

Improper priors

- Priors that are not valid PDFs are called *improper priors*.
- This is acceptable, as long as the posterior is well-behaving.
- In practice, it is best to avoid such priors as it leads to non-generative models – it is formally not possible to generate samples from these models[2].

Example: Jeffreys prior for the Gaussian σ

- The Jeffreys prior $\pi(\sigma) \propto \sqrt{\mathcal{I}(\sigma)}$ for a Gaussian with unknown σ and known μ is given by $\frac{1}{\sigma}$.

[2] http://andrewgelman.com/2017/06/18/dont-say-improper-prior-say-non-generative-model/

The problem of model selection

- So far, we have assumed that the problem of inference is limited to the model's parameters. But often, we also have to decide on what model(s) to use.

Model selection examples

- How many mixture components should we use for a mixture model?
- For data on the real line, should we use a Gaussian or a Student-t distribution?
  - The latter has "fatter tails".
- The degree of the polynomial in polynomial curve fitting.

Bayes factor

- The Bayesian tool to select the best model is the Bayes factor – the ratio of the respective probabilities of the data $\mathbf{d}$ given the two different models M1 and M2.

$$\begin{aligned} B &= \frac{p(\mathbf{d} \mid M_1)}{p(\mathbf{d} \mid M_2)} \\ &= \frac{p(M_1 \mid \mathbf{d})p(\mathbf{d})}{p(M_1)} \times \frac{p(M_2)}{p(M_2 \mid \mathbf{d})p(\mathbf{d})} \\ &= \frac{p(M_1 \mid \mathbf{d})}{p(M_2 \mid \mathbf{d})} \Big/ \frac{p(M_1)}{p(M_2)} \end{aligned}$$

- Hence, the Bayes factor measures how the data affect the belief in M1 and M2 relative to the prior information on the models.

Calculation of the Bayes factor

- Calculation of the Bayes factor requires integrating out the model parameters.

$$B = \frac{p(\mathbf{d} \mid M_1)}{p(\mathbf{d} \mid M_2)} = \frac{\int p(\mathbf{d} \mid \theta_1, M_1)\pi(\theta_1 \mid M_1)\mathrm{d}\theta_1}{\int p(\mathbf{d} \mid \theta_2, M_2)\pi(\theta_2 \mid M_2)\mathrm{d}\theta_2}$$

- The frequentist way to model selection makes use of the **likelihood ratio**, which uses the maximum likelihood estimate $\widehat{\theta}_{\mathrm{ML}}$ instead of integrating out the unknown parameters.

$$L = \frac{p(\mathbf{d} \mid \widehat{\theta}_{\mathrm{ML},1}, M_1)}{p(\mathbf{d} \mid \widehat{\theta}_{\mathrm{ML},2}, M_2)}$$

Interpretation of the Bayes factor

- The Bayes factor is typically interpreted using a scale introduced by Jeffreys

Jeffreys' scale - If $\log(B)$...

- is below 0, the evidence is against M1 and in favor of M2.
- is between 0 and 0.5, the evidence in favor of M1 and against M2 is weak.
- is between 0.5 and 1, it is substantial.
- is between 1 and 1.5, it is strong.
- is between 1.5 and 2, it is very strong.
- is above 2, it is decisive.

Bayesian information criterion (BIC)

- The calculation of the Bayes factor is often intractable. In that case, one can approximate $\log p(\mathbf{d} \mid M)$ as follows

$$\log p(\mathbf{d} \mid M) \approx \mathrm{BIC} = \log p(\mathbf{d} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}, M) - \underbrace{\frac{1}{2} Q \log(R)}_{\text{Penalizes parameters}}$$

  where $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is the ML estimate, R is the number of data points and Q is the number of free parameters in $\widehat{\theta}_{\mathrm{ML}}$.
- This approximation is based on the Laplace approximation, which involves approximating a PDF as a Gaussian distribution centered at its mode.
- The model with the highest BIC is the best model.

(Akaike) An information criterion (AIC)

- Pick the model with highest AIC (Akaike,1974).

$$\mathrm{AIC} = 2 \log p(\mathbf{d} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}, M) - 2Q.$$

- The AIC and BIC look similar but are justified in different ways. The AIC aims for the model that minimizes the KL divergence with the "true", but unknown model $p_T(\mathbf{d})$.
- Since $p_T(\mathbf{d})$ is unknown, the AIC uses the *expected negative of* $\mathrm{KL}[p_T \parallel p]$,

$$\mathrm{AIC} \approx \underbrace{\mathbb{E}_{\mathbf{d_2}} \left[\underbrace{\mathbb{E}_{\mathbf{d_1}} \left[\log p(\mathbf{d}_1 \mid \widehat{\boldsymbol{\theta}}_{\mathrm{ML},\mathbf{d_2}}, M) \right]}_{-\mathrm{KL}[p_T \| p] + C} \right]}_{\text{Expectation over } p_T}.$$

- Both expectations are taken with respect to the true, but unknown, model $p_T(\mathbf{d})$.

Deviance information criterion (DIC)

- Often, the posterior is only available as samples. But we need the ML estimate for the BIC and the AIC. In that case, we can select the model with the smallest DIC.

$$\mathrm{DIC} = p_D + \overline{D}$$

  - D is the deviance. The larger D, the worse the fit.

$$D(\theta) = -2 \log(p(\mathbf{d} \mid \theta))$$

  - $\overline{D}$ is the **expected deviance**, calculated from the posterior samples.

$$\overline{D} = \mathbb{E}_{p(\theta|\mathbf{d})} [D(\theta)]$$

  - p_D estimates the **effective number of parameters**.

$$p_D = \overline{D} - D(\overline{\theta})$$

Why statistical physics?

- It's an old example of Bayesian reasoning – in disguise.
  - Uses (additive) energies instead of (multiplicative) probabilities.
- StatPhys methods are widely used in machine learning.
  - *Belief propagation methods* were first proposed by the physicist Hans Bethe in the 1930s and later rediscovered in statistics, computer science and communication engineering for inference of graphical models.
    - The Forward-backward and Viterbi algorithms for HMMs can be seen as examples of belief propagation.
  - StatPhys concepts such as phase transitions can be used to understand machine learning methods.
- Problems such as protein structure prediction require both StatPhys and Bayes [6].

On energies and probabilities

- We consider a physical system that can adopt a number of microstates $M = \{m_1, m_2, \ldots, m_N\}$ with probabilities $P = \{p_1, p_2, \ldots p_N\}$ and energies $E = \{e_1, e_2, \ldots, e_N\}$.
- Now, suppose we are given the average energy $\overline{e}$, also called the *internal energy* u.

$$u = \overline{e} = \sum_{n=1}^{N} e_n p_n$$

- Given only u, can we infer the probabilities of the individual states, that is, P?
- This can be solved using a Maximum Entropy approach, entirely equivalent to the approach we used for the Brandeis dice problem.

Boltzmann distribution I

- The resulting Lagrangian is

$$L = \overbrace{-\sum_{n=1}^{N} p_n \log p_n}^{\text{Entropy}} - \alpha \overbrace{\left(\sum_{n=1}^{N} p_n - 1\right)}^{\text{C1:Normalization}} - \beta \overbrace{\left(\sum_{n=1}^{N} p_n e_n - u\right)}^{\text{C2:Mean energy}}$$

- As with the dice problem, the solution is

$$p_n = \frac{1}{Z} \exp\left(-\beta e_n\right)$$

with $Z = \sum_{n=1}^{N} \exp\left(-\beta e_n\right)$. This is called the Boltzmann distribution.

- Z (for "Zustandssumme") is the partition factor.

Boltzmann distribution II

- The value of β is determined by the constraint

$$u = \overline{e} = \sum_{n=1}^{N} e_n p_n = \frac{1}{Z} \sum_{n=1}^{N} e_n \exp\left(-\beta e_n\right)$$

- In physics, β relates to the inverse of the temperature, $\frac{1}{T}$.
- Any probability distribution can be cast as a Boltzmann distribution, by choosing βe_n equal to $-\log(p_n)$.
  - The advantage is that energies can be added, while probabilities need to be multiplied.
- The result can be summarized as

$$\text{probability} = \frac{1}{\text{Partition factor}} \exp\left(-\frac{\text{Energy}}{\text{Temperature}}\right)$$

Microstates and macrostates

- Suppose we are interested in the probability of a set of *microstates*, which make up a *macrostate* M. What is the probability p_M of M?

$$p_M = \frac{1}{Z} \sum_{i:m_i \in M} \exp\left(-\beta e_i\right)$$

  where the sum runs over all microstates m_i in M.

- Now consider the ratio of probabilities of two macrostates M and N.

$$\frac{p_M}{p_N} = \frac{\sum_{i:m_i \in M} \exp\left(-\beta e_i\right)}{\sum_{j:n_j \in N} \exp\left(-\beta e_j\right)} = \frac{Z_M}{Z_N}$$

  Note that the overall Z cancels.

Free energy

- The Boltzmann factor $\exp(-\beta e_n)$ gives the relative probability of microstate m_n. In analogy, we define a Boltzmann factor for macrostates.

$$\frac{p_M}{p_N} = \frac{Z_M}{Z_N} = \frac{\exp(-\beta f_M)}{\exp(-\beta f_N)}$$

where f_M, f_N are the **free energies** of macrostates M, N.

Example: Protein folding

- What is the ratio of folded versus unfolded proteins?

$$\frac{p_{\text{Folded}}}{p_{\text{Unfolded}}} = \frac{Z_F}{Z_U} = \frac{\exp(-\beta f_{\text{Folded}})}{\exp(-\beta f_{\text{Unfolded}})}$$

Entropy

- Suppose we know the average energy $\overline{e}_M$ for a macrostate M, $\overline{e}_M = \sum_{i:m_i \in M} p_i e_i$.
- We want to relate $\overline{e}_M$ to the relative probability $\exp(-\beta f_M)$ of macrostate M.

$$p_M \propto \exp(-\beta f_M) = C \times \exp(-\beta \overline{e}_M)$$

- The unknown factor C is related to the entropy s_M of M.

$$p_M \propto \exp(-\beta f_M) = \exp(s_M) \times \exp(-\beta \overline{e}_M)$$

- The factor $C = \exp(s_M)$ specifies how many times the Boltzmann factor of the average energy of M needs to be counted to arrive at the full relative probability of M.
- The entropy is a measure of the *extent* of M.

Potential of mean force

- Yet another important quantity in StatPhys.
- Suppose we assign an energy $e(x, y, z)$ to a positional coordinate (x, y, z),then

$$p(x, y, z) \propto \exp(-\beta e(x, y, z))$$

- Consider the marginal probability $p(x)$.

$$p(x) \propto \int_y \int_z \exp(-\beta e(x, y, z)) \mathrm{d}z \mathrm{d}y$$

- Now, per definition

$$p(x) \propto \exp(-\beta \mathrm{PMF}(x))$$

- Hence, the PMF above can be considered as a "free energy for marginals", as it relates to a marginal probability through yet another Boltzmann factor.

StatPhys summary

- StatPhys uses energies instead of probabilities.
- The Boltzmann factor $\exp(-\beta\mathcal{E})$ relates a (general) energy $\mathcal{E}$ to a relative probability.
  - An energy e_n relates to the probability p_n of an individual microstate m_n.
  - A free energy f_M relates to the probability p_M of a macrostate M.
    - A macrostate is a collection of microstates $\{m_1, \ldots, m_N\}$.
    - The entropy s_M is a measure of the extent of a macrostate M.

$$\exp(-\beta f_M) \propto \exp(s_M) \exp(-\beta u_M)$$

  - A potential of mean force $\mathrm{PMF}(x)$ relates to a marginal probability $p(x)$.

Limits of standard Bayesian updating

- Sometimes, the nature of new information does not quite fit within the usual Bayesian calculus [2, 6].

Example

- Suppose you have a probability distribution $p(x_1, \ldots, x_N)$ over an N-dimensional vector $\mathbf{x}$.
- You get new information about $\mathbf{x}$ in the form of $p(y)$, where y is a one-dimensional random variable that is a deterministic, many-to-one function $y = f(\mathbf{x})$ of $\mathbf{x}$.
- $p(y)$ brings new information about a *partition* of $\mathbf{x}$.
- How do you update $p(\mathbf{x})$?

Example: Whitworth's horses

- Three racing horses A, B and C have a probability of winning equal to $\frac{2}{11}, \frac{4}{11}$ and $\frac{5}{11}$.
- New information changes the probability of A winning to $\frac{1}{2}$. How do we update the probabilities of B and C winning?
- $p(A \text{ loses}) = (1 - \frac{2}{11}) = \frac{9}{11}$
- $p(A \text{ loses})$ was decreased by a factor $\frac{11}{18}$, as $\frac{9}{11} \times \frac{11}{18} = \frac{1}{2}$.
- A can lose in two ways – either B or C wins.
- We can thus postulate that $p(B \text{ wins})$ and $p(C \text{ wins})$ both decrease by the same factor [2].

$$\begin{cases} p(B \text{ wins}) & = \frac{4}{11} \times \frac{11}{18} = \frac{2}{9} \\ p(C \text{ wins}) & = \frac{5}{11} \times \frac{11}{18} = \frac{5}{18} \end{cases}$$

Jeffrey's conditioning or probability kinematics

- The solution the Whitworth's horses problem follows from **Jeffrey's conditioning** or **probability kinematics** [2].
- This form of Bayesian updating was proposed by the American philosopher of probability Richard C. Jeffrey[3] (1926-2002) in the 1950s.

[3] Not Harold Jeffreys (1891 – 1989).

Jeffrey's conditioning in action

- We have a PDF $p(\mathbf{x}) = p(x_1, \ldots, x_N)$.
- Consider a partition $\mathbf{E} = \{E_1, E_2, \ldots E_N\}$ with probabilities $p(E_1), p(E_2) \ldots$ Note that these follow from $p(\mathbf{x})$.
- Now we receive updated probabilities $p^*(E_1), p^*(E_2) \ldots$
- How do we update $p(\mathbf{x})$ given $p^*(E_1), p^*(E_2) \ldots$?

Solution

- We assume that the conditional probabilities within the partition's elements stay the same. Thus

$$p(\mathbf{x}) = p(\mathbf{x} \mid E_x)p(E_x) \Rightarrow p^*(\mathbf{x}) = p(\mathbf{x} \mid E_x)p^*(E_x)$$

where E_x is the partition element that contains $\mathbf{x}$.

The reference ratio method

- Often, the conditional $p(\mathbf{x} \mid E_x)$ is not available. In that case, the reference ratio formulation of Jeffrey's conditioning applies [6, 5].

$$\begin{aligned} p^*(\mathbf{x}) &= p(\mathbf{x} \mid E_x) p^*(E_x) \\ &= \frac{p(E_x \mid \mathbf{x}) p(\mathbf{x})}{p(E_x)} p^*(E_x) \\ &= \frac{p^*(E_x)}{p(E_x)} p(\mathbf{x}) \end{aligned}$$

- This form of JC has been used heuristically in protein structure prediction for 20 years under the (mistaken) designation "potential of mean force" [5, 6].

Application to Whitworth's horses

- We know

$$\begin{aligned} p(A \text{ loses}) &= 1 - \frac{2}{11} = \frac{9}{11} \\ p^*(A \text{ loses}) &= \frac{1}{2} \\ p(B \text{ wins}) &= \frac{4}{11} \end{aligned}$$

- Following the reference ratio method, we obtain

$$\begin{aligned} p^*(B \text{ wins}) &= p(B \text{ wins} \mid A \text{ loses}) p^*(A \text{ loses}) \\ &= \frac{p^*(A \text{ loses})}{p(A \text{ loses})} p(B \text{ wins}) \\ &= \left(\frac{1}{2} \times \frac{11}{9} \right) \frac{4}{11} = \frac{2}{9} \end{aligned}$$

Why Bayes?

- There are good reasons to prefer the Bayesian interpretation of probability over its alternatives. The three most common justifications are:
  - If one adopts a small set of requirements (formulated as axioms) regarding beliefs including respecting the rules of logic, the rules of probability theory necessarily follow.
  - de Finetti's theorem states that if a data set follows certain common conditions, an appropriate probabilistic model for the data necessarily consists of a likelihood and a prior.
  - In gambling, the use of beliefs that follow the Bayesian calculus avoids situations of certain loss for a bookmaker.

The Cox axioms I

- The axioms were formulated by Richard T. Cox in 1946.
- The axioms emerge from a small set of requirements; properties that clearly need to be part of any consistent calculus involving degrees of belief. Informally, the Cox axioms state:
  - Consistency with logic, when beliefs are absolute (True or False).
  - Different ways of reasoning lead to the same result.
  - Identical states of knowledge, differing by labelling only, lead to the assignment of identical degrees of belief.
- From these axioms, the rules of probability follow, including the sum and product rule and Bayes' theorem.

The Cox axioms II

- A degree of belief in a is expressed as a real number $\mathcal{B}(a)$.
- Degrees of belief are ordered.
  - If $\mathcal{B}(a) > \mathcal{B}(b)$ and $\mathcal{B}(b) > \mathcal{B}(c)$ then $\mathcal{B}(a) > \mathcal{B}(c)$.
- There is a function $\mathcal{F}$ that connects the beliefs in a proposition a and its negation $\sim a$:

$$\mathcal{B}(a) = \mathcal{F}\left[\mathcal{B}(\sim a)\right]$$

- If we want to calculate the belief that a and b are true, we can first calculate the belief that b is true, and then the belief that b is true given that a is true. Since the labelling is arbitrary, we can switch a and b around, which leads to the existence of a function $\mathcal{G}$:

$$\mathcal{B}(a, b) = \mathcal{G}\left[\mathcal{B}(a \mid b)\mathcal{B}(b)\right] = \mathcal{G}\left[\mathcal{B}(b \mid a)\mathcal{B}(a)\right]$$

The Cox axioms III

- Surprisingly, this simple set of axioms is sufficient to pinpoint the rules of probabilistic inference completely.
- As expected, the functions $\mathcal{F}$ and $\mathcal{G}$ turn out to be

$$\begin{aligned} \mathcal{F}(x) &= 1 - x \\ \mathcal{G}(x, y) &= xy \end{aligned}$$

- In particular, the axioms lead to the two central rules of probability theory. To recall, these rules are the product rule

$$p(a, b) = p(a \mid b)p(b) = p(b \mid a)p(a)$$

which directly leads to Bayes' theorem, and the sum rule

$$p(a) = \sum_b p(a, b)$$

The exchangeability argument

- **de Finetti's representation theorem** essentially guarantees a likelihood and a prior for exchangeable data.
  - For exchangeable data, any permutation of the data does not alter the joint probability distribution.
- Let us consider the case of an exchangeable series of N Bernoulli random variables[4], consisting of zeros and ones. Then, de Finetti's theorem guarantees that the joint probability distribution of the data can be written as:

$$p(x_1, \ldots, x_N) = \int_0^1 \underbrace{\left\{ \prod_{n=1}^{N} \theta^{x_n} (1-\theta)^{x_n} \right\}}_{\text{Likelihood}} \underbrace{\pi(\theta)}_{\text{Prior}} \mathrm{d}\theta$$

[4] A binomial distribution with $n = 1$.

The Dutch book argument

- If you have a belief $\mathcal{B}(x) = 0.8$ in an event x, then you should accept the following bet with odds:

$$\begin{cases} \text{if } x \text{ is true, you win} & > 2\$ \\ \text{if } x \text{ is false, you lose} & 8\$ \end{cases}$$

- Unless your beliefs satisfy the rules of probability theory, including Bayes' rule, there exists a set of bets (called a "Dutch Book") which you are willing to accept, and that will make you lose money, no matter what the outcome.
- The only way to avoid the possibility of a Dutch Book is to ensure that your beliefs satisfy the rules of probability.

Bayes and Maximum Relative Entropy I

- Updating a prior to a posterior distribution follows the general principle of Maximum Relative Entropy (MRE) [4].
  - Note that we now consider joint distributions $p(x, \theta)$ over the product of the data and the parameter space, $(x, \theta) \in \mathcal{X} \times \Theta$.
  - New information is brought in under the form of **constraints**.
- The selected posterior $p_{\text{new}}(x, \theta)$ maximizes the relative entropy

$$S\left[p, p_{\text{old}}\right] = -\int p(x, \theta) \log \frac{p(x, \theta)}{p_{\text{old}}(x, \theta)} \mathrm{d}x \mathrm{d}\theta$$

under suitable **constraints on the posterior** $p_{\text{new}}(x, \theta)$.

Bayes and Maximum Relative Entropy II

- Depending on the constraints, we recover different methods of belief updating.
- If the value of x is known to be x', we recover **classic Bayesian updating**.

$$p(x) = \int p(x, \theta) \mathrm{d}\theta = \delta(x - x')$$

  - From this constraint, the posterior becomes

$$p_{\mathrm{new}}(x, \theta) = \frac{p_{\mathrm{old}}(x, \theta)\delta(x - x')}{p_{\mathrm{old}}(x)} = \delta(x - x')p_{\mathrm{old}}(\theta \mid x)$$

    which corresponds to the familiar Bayesian updating,

$$p_{\mathrm{new}}(\theta) = p_{\mathrm{old}}(\theta \mid x').$$

Bayes and Maximum Relative Entropy III

- From the expected value of a function $f(\theta)$, we recover **Maximum Entropy updating**.

$$\int \int f(\theta) p(x, \theta) \mathrm{d}x \mathrm{d}\theta = \langle f(\theta) \rangle = F$$

$$\Rightarrow p_{\text{new}}(x, \theta) = p_{\text{old}}(x, \theta) \frac{\exp\left(\beta f(\theta)\right)}{Z}$$

- From the marginal distribution of x, $p(x)$, we recover **Jeffreys' conditioning**.

$$p(x) = \int p(x, \theta) \mathrm{d}\theta = p_D(x)$$

$$\Rightarrow p_{\text{new}}(x, \theta) = p_D(x) p_{\text{old}}(\theta \mid x)$$

Science as Bayesian inference

- Physicist Edwin Jaynes (1922 –1998) saw the scientific method as an application of Bayesian inference [7].
  - Update prior belief based on data. Predict, repeat.
  - Following the Cox axioms, Bayesian inference is seen as an extension of logic in the face of uncertainty.

The Bayesian brain

- The brain seems to have an underlying Bayesian model of reality [3], that is updated using sensory input.
  - Likelihood =probability of sensory data, given their causes.
  - Prior=the a priori probability of those causes.
  - Posterior=probability of the causes, given sensory data.

Summary of part II

- The choice of the prior distribution.
  - MaxEnt, Jeffreys and reference priors.
  - Empirical Bayes.
  - Improper priors and conjugate priors.
- Model selection.
  - Bayes factor, DIC, BIC and AIC.
- StatPhys.
  - Boltzmann distribution and energy.
  - Free energy (and entropy) and potentials of mean force.
- Jeffrey's conditioning and the reference ratio method.

References part II

Bernardo, JM., & Smith, AF. (2009). Bayesian theory. John Wiley & Sons.

Diaconis, P., Zabell, SL. (1982) Updating subjective probability. JASA, 77, 822-830.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11, 127-138.

Giffin, A. (2008) Maximum entropy: the universal method for inference. PhD thesis.

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W. Bottaro, S., Ferkinghoff-Borg, J. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS ONE 5(11): e13714.

Hamelryck, T., Mardia, KV., Ferkinghoff-Borg, J., Editors. (2012) Bayesian methods in structural bioinformatics. Book in the Springer series "Statistics for biology and health", Springer Verlag, 2012.

Jaynes, E. T. (2003). Probability theory: the logic of science. Cambridge university press.

Lee, PM. (2012) Bayesian statistics: an introduction. John Wiley & Sons.

Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer.