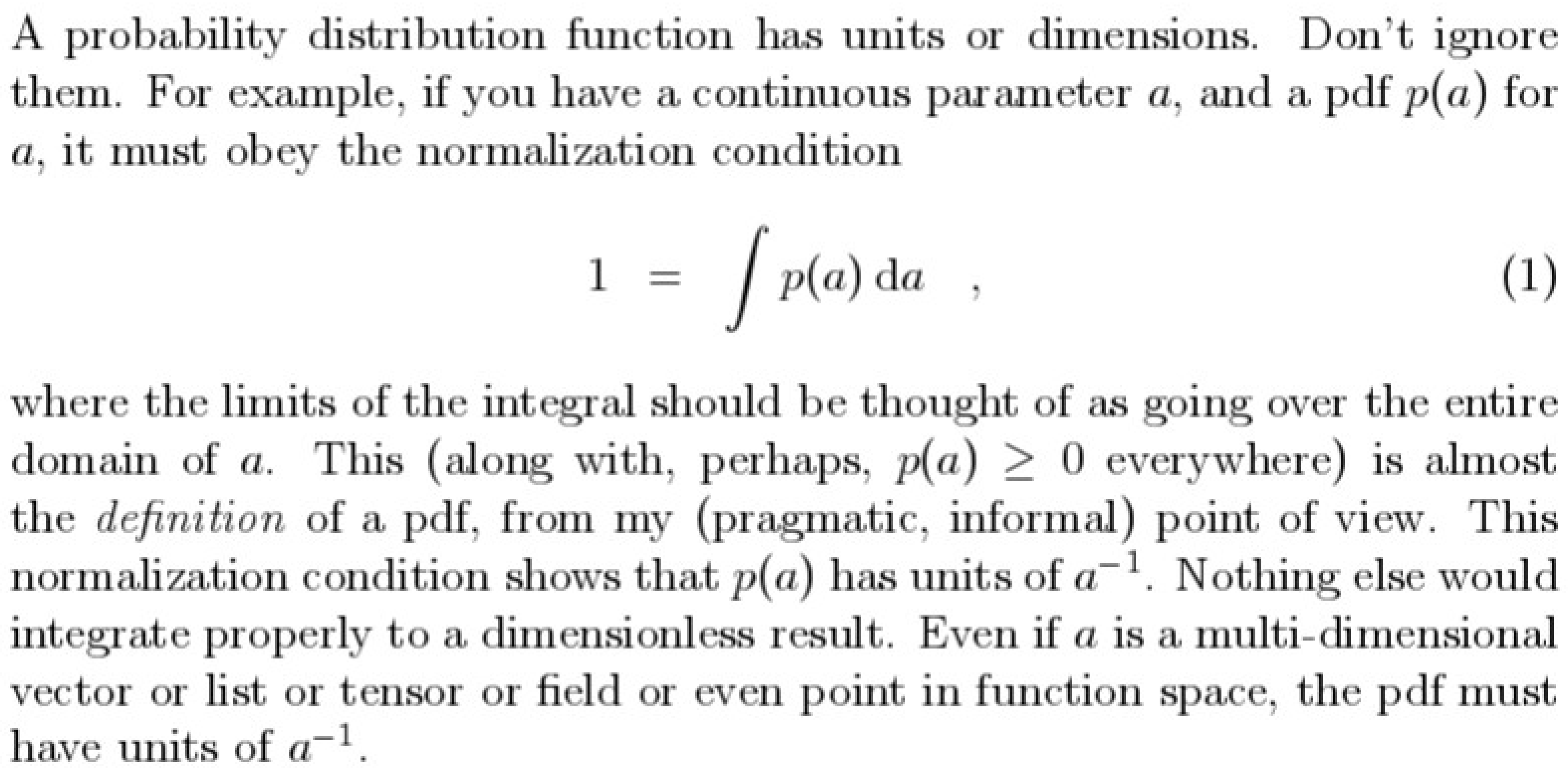
Session 1: A biased primer on statistics

Probability distributions

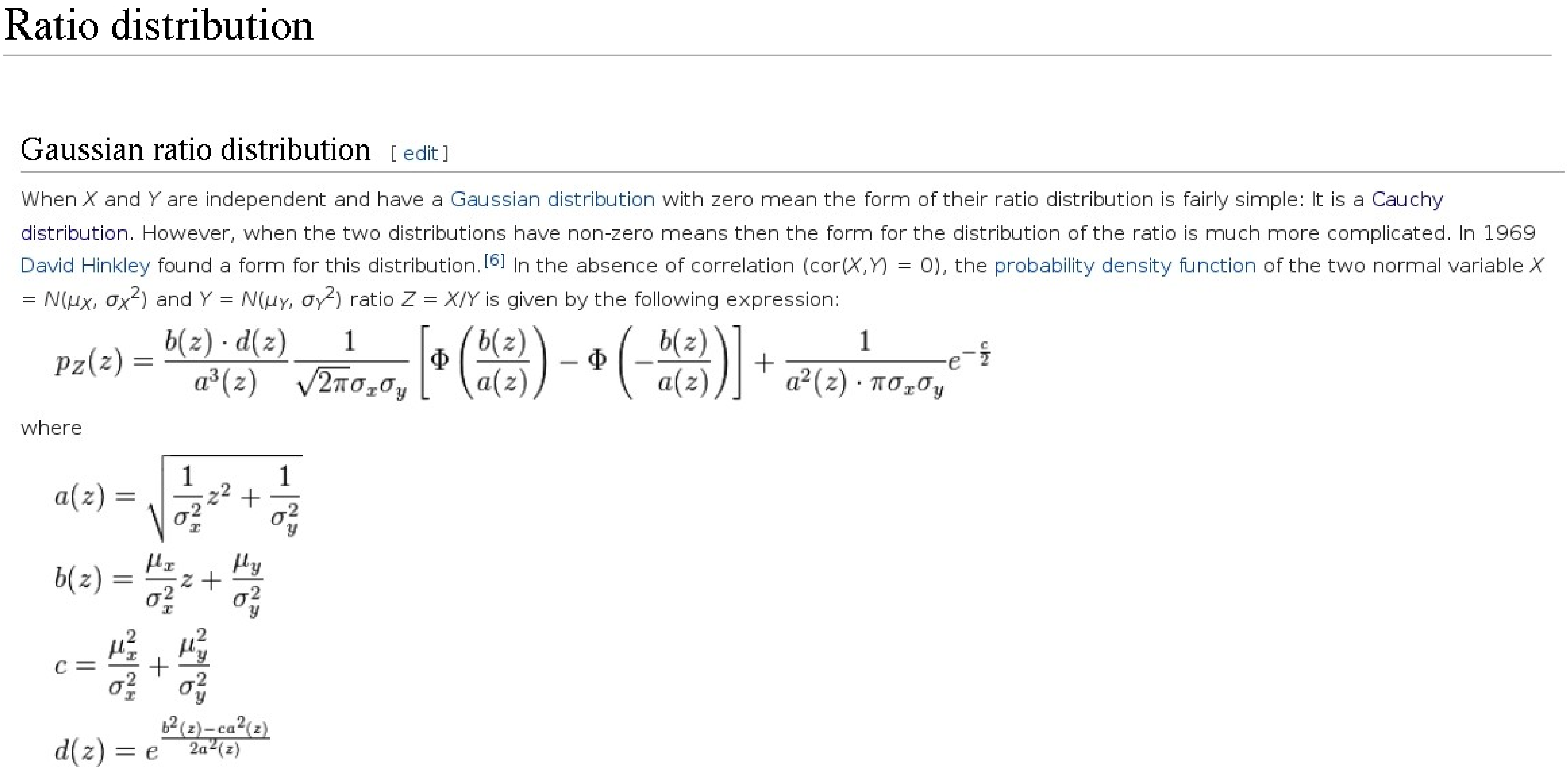


A probability distribution function has units or dimensions. Don't ignore them. For example, if you have a continuous parameter a, and a pdf $p(a)$ for a, it must obey the normalization condition

$$1 = \int p(a)\,\mathrm{d}a \quad , \tag{1}$$

where the limits of the integral should be thought of as going over the entire domain of a. This (along with, perhaps, $p(a) \geq 0$ everywhere) is almost the *definition* of a pdf, from my (pragmatic, informal) point of view. This normalization condition shows that $p(a)$ has units of a^{-1}. Nothing else would integrate properly to a dimensionless result. Even if a is a multi-dimensional vector or list or tensor or field or even point in function space, the pdf must have units of a^{-1}.

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)



Ratio distribution

Gaussian ratio distribution [edit]

When *X* and *Y* are independent and have a Gaussian distribution with zero mean the form of their ratio distribution is fairly simple: It is a Cauchy distribution. However, when the two distributions have non-zero means then the form for the distribution of the ratio is much more complicated. In 1969 David Hinkley found a form for this distribution.[6] In the absence of correlation (cor(*X*,*Y*) = 0), the probability density function of the two normal variable $X = N(\mu_X, {\sigma_X}^2)$ and $Y = N(\mu_Y, {\sigma_Y}^2)$ ratio $Z = X/Y$ is given by the following expression:

$$p_Z(z) = \frac{b(z) \cdot d(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x \sigma_y} \left[\Phi\left(\frac{b(z)}{a(z)}\right) - \Phi\left(-\frac{b(z)}{a(z)}\right) \right] + \frac{1}{a^2(z) \cdot \pi \sigma_x \sigma_y} e^{-\frac{c}{2}}$$

where

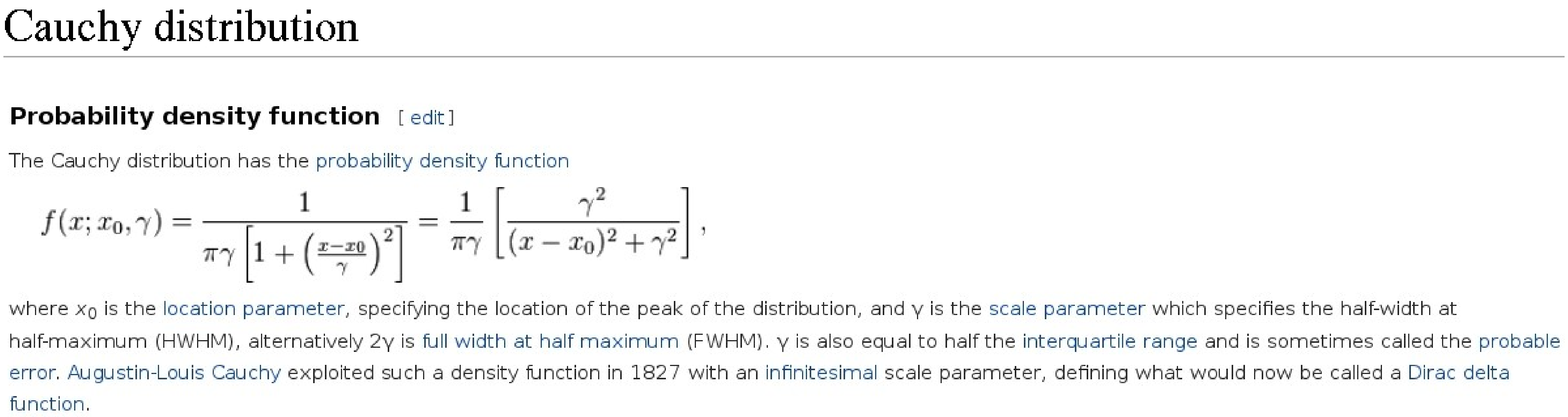
$$a(z) = \sqrt{\frac{1}{\sigma_x^2} z^2 + \frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2} z + \frac{\mu_y}{\sigma_y^2}$$

$$c = \frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}$$

$$d(z) = e^{\frac{b^2(z) - ca^2(z)}{2a^2(z)}}$$

But, there are other possibilities for the random variables in the numerator and denominator...



Cauchy distribution

Probability density function [edit]

The Cauchy distribution has the probability density function

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x0}{\gamma}\right)^2\right]} = \frac{1}{\pi\gamma}\left[\frac{\gamma^2}{(x-x_0)^2+\gamma^2}\right],$$

where x_0 is the location parameter, specifying the location of the peak of the distribution, and γ is the scale parameter which specifies the half-width at half-maximum (HWHM), alternatively 2γ is full width at half maximum (FWHM). γ is also equal to half the interquartile range and is sometimes called the probable error. Augustin-Louis Cauchy exploited such a density function in 1827 with an infinitesimal scale parameter, defining what would now be called a Dirac delta function.

Chi-squared distribution

Definition [edit]

If Z_1, ..., Z_k are independent, standard normal random variables, then the sum of their squares,

$$Q \ = \sum_{i=1}^{k} Z_i^2,$$

Let $(X_1, X_2, \ldots, X_i, \ldots, X_k)$ be k independent, normally distributed random variables with means μ_i and unit variances. Then the random variable

$$\sum_{i=1}^{k} X_i^2$$

is distributed according to the noncentral chi-squared distribution. It has two parameters: k which specifies the number of degrees of freedom (i.e. the number of X_i), and λ which is related to the mean of the random variables X_i by:

$$\lambda = \sum_{i=1}^{k} \mu_i^2.$$

λ is sometimes called the noncentrality parameter. Note that some references define λ in other ways, such as half of the above sum, or its square root.

Student's *t*-distribution

If we take a sample of n observations from a normal distribution, then the *t*-distribution with $\nu = n - 1$ degrees of freedom can be defined as the distribution of the location of the true mean, relative to the sample mean and divided by the sample standard deviation, after multiplying by the normalizing term $\sqrt{n}$. In this way, the *t*-distribution can be used to estimate how likely it is that the true mean lies in any given range.

F-distribution

A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates:[7]

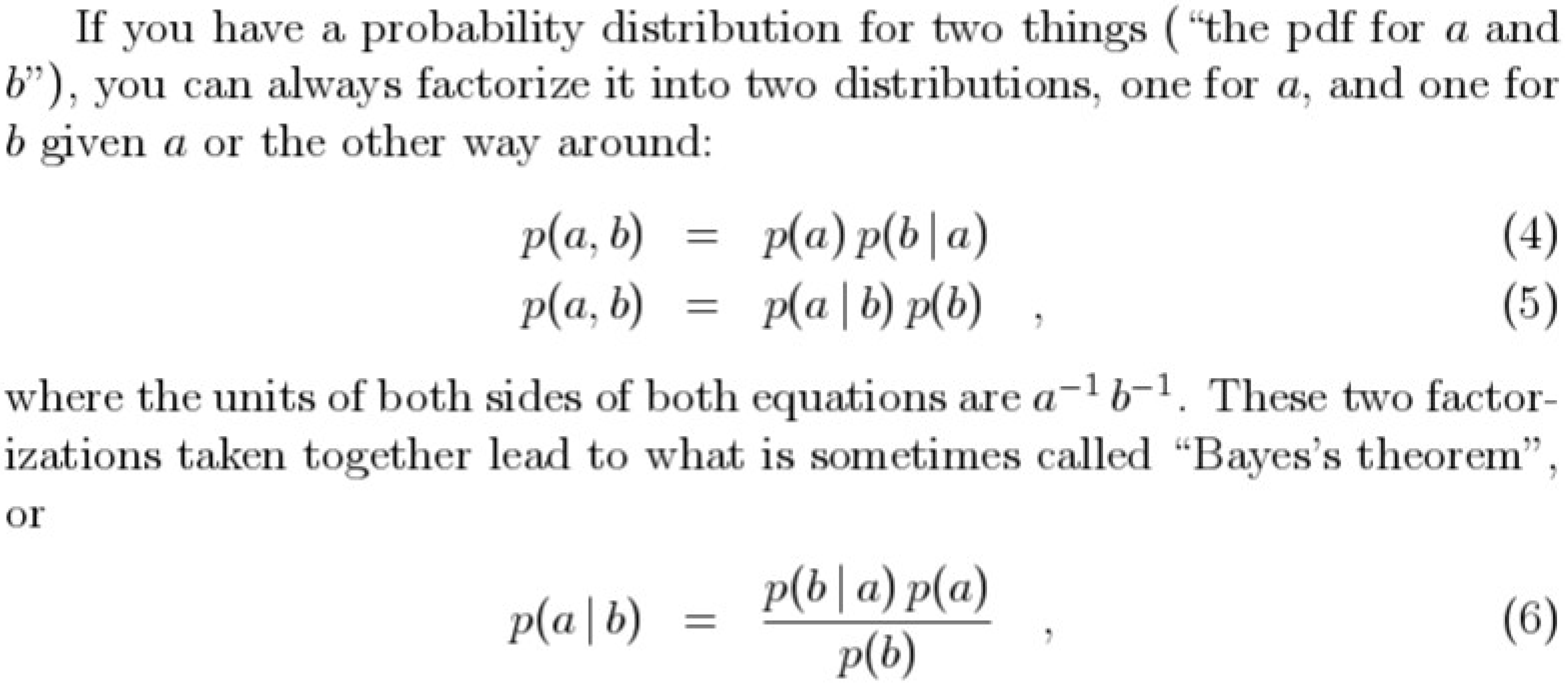
$$X = \frac{U_1/d_1}{U_2/d_2}$$

where

- U_1 and U_2 have chi-squared distributions with d_1 and d_2 degrees of freedom respectively, and
- U_1 and U_2 are independent.

In instances where the F-distribution is used, for example in the analysis of variance, independence of U_1 and U_2 might be demonstrated by applying Cochran's theorem.

Conditional probabilities



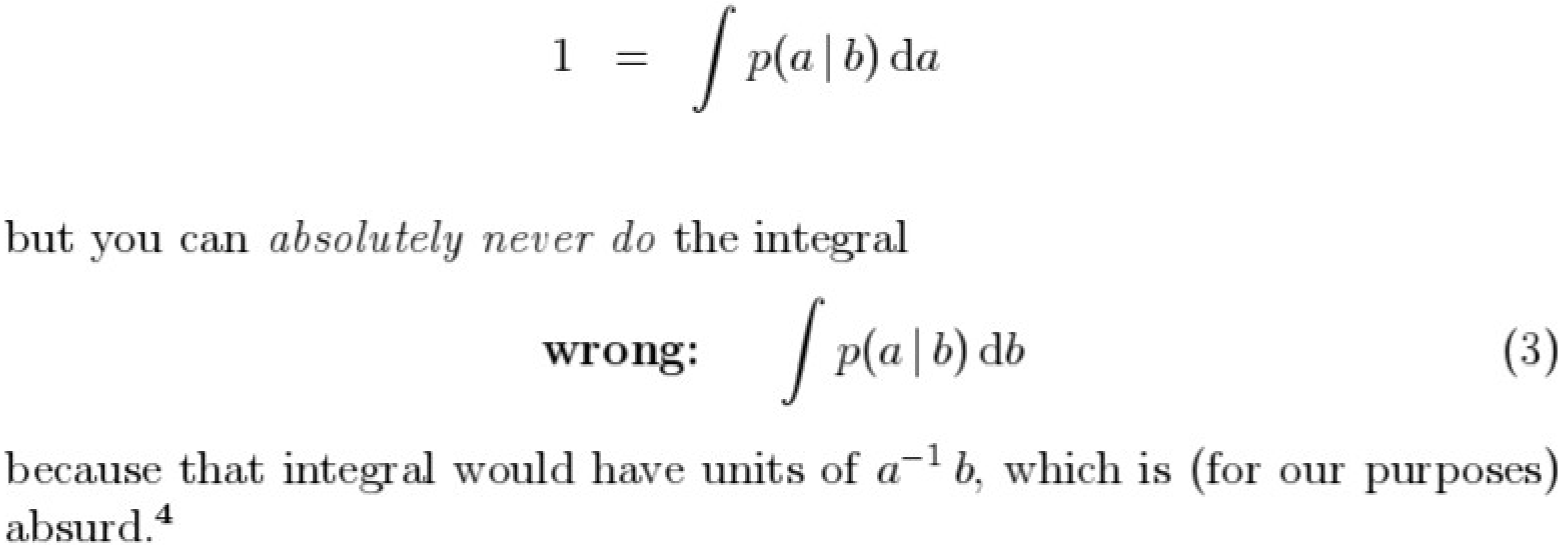
If you have a probability distribution for two things ("the pdf for a and b"), you can always factorize it into two distributions, one for a, and one for b given a or the other way around:

$$\begin{aligned} p(a,b) &= p(a)\,p(b\,|\,a) \qquad (4)\\ p(a,b) &= p(a\,|\,b)\,p(b) \quad , \qquad (5) \end{aligned}$$

where the units of both sides of both equations are $a^{-1}\,b^{-1}$. These two factorizations taken together lead to what is sometimes called "Bayes's theorem", or

$$p(a\,|\,b) = \frac{p(b\,|\,a)\,p(a)}{p(b)} \quad , \qquad (6)$$

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)



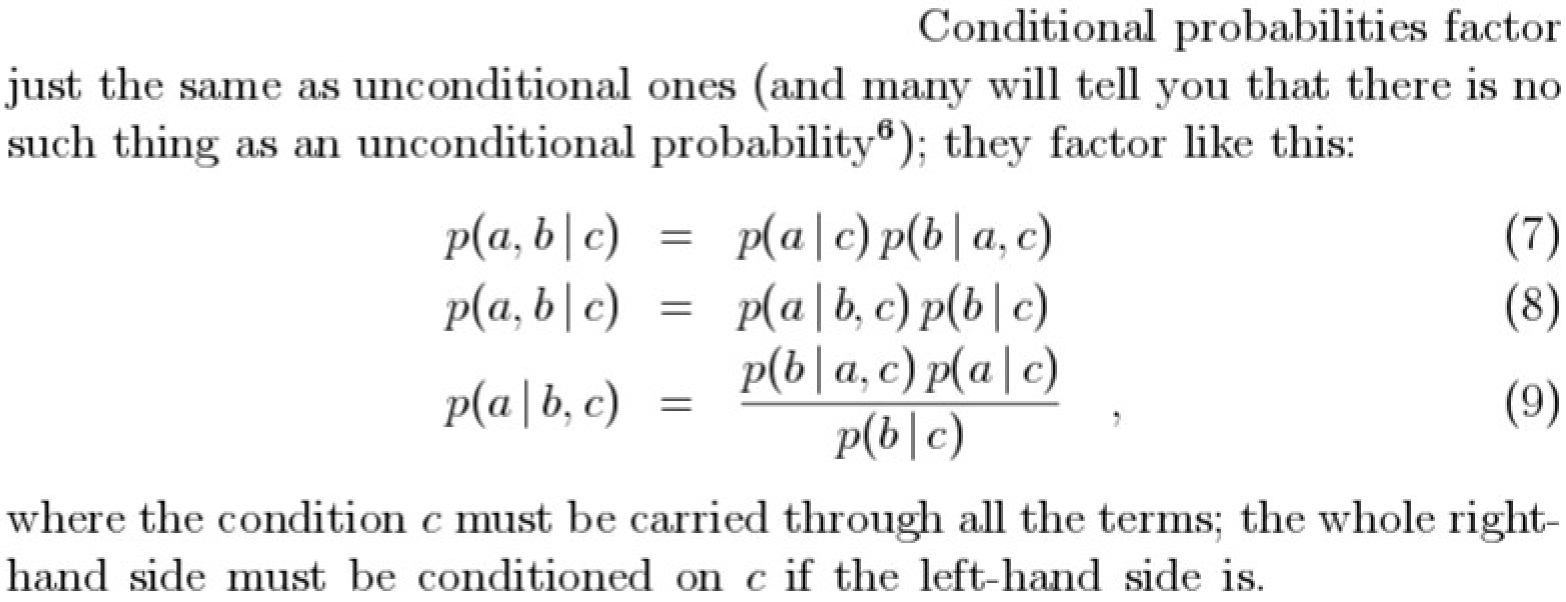
$$1 = \int p(a\,|\,b)\,\mathrm{d}a$$

but you can *absolutely never do* the integral

$$\textbf{wrong:} \quad \int p(a\,|\,b)\,\mathrm{d}b \qquad (3)$$

because that integral would have units of $a^{-1}\,b$, which is (for our purposes) absurd.[4]

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)



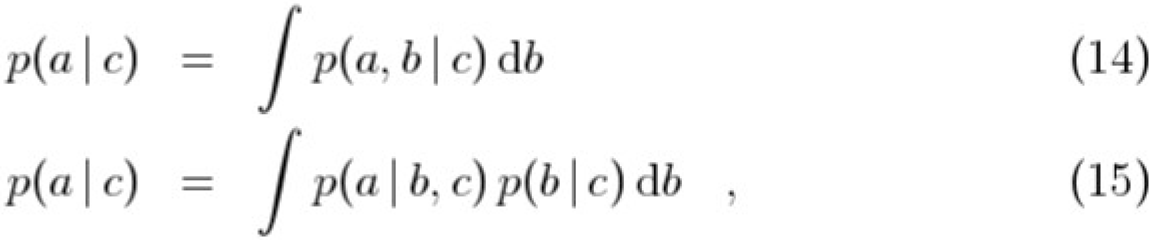
Conditional probabilities factor just the same as unconditional ones (and many will tell you that there is no such thing as an unconditional probability[6]); they factor like this:

$$
\begin{aligned}
p(a, b \,|\, c) &= p(a \,|\, c)\, p(b \,|\, a, c) && (7)\\
p(a, b \,|\, c) &= p(a \,|\, b, c)\, p(b \,|\, c) && (8)\\
p(a \,|\, b, c) &= \frac{p(b \,|\, a, c)\, p(a \,|\, c)}{p(b \,|\, c)} \quad , && (9)
\end{aligned}
$$

where the condition c must be carried through all the terms; the whole right-hand side must be conditioned on c if the left-hand side is.

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)

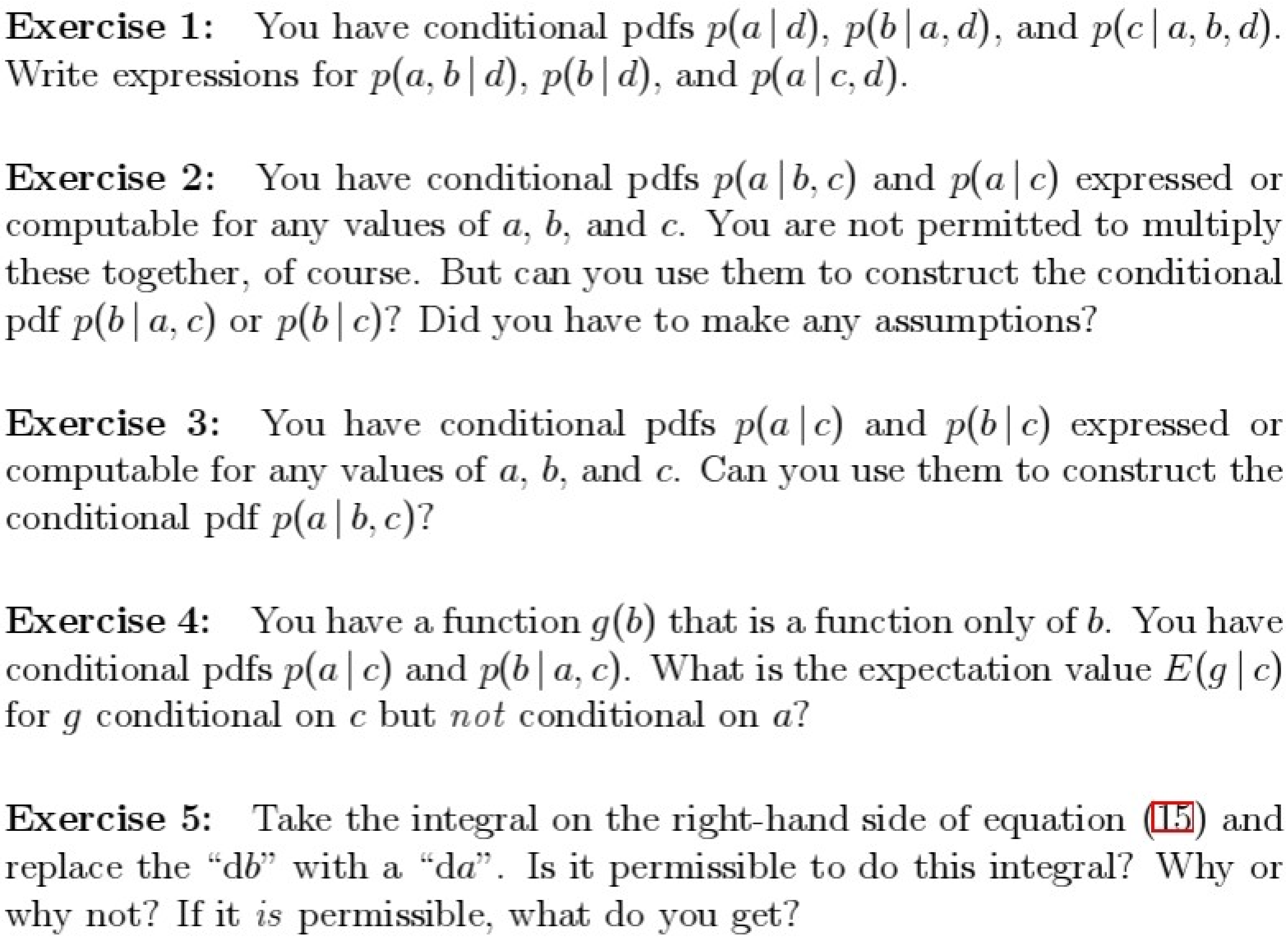
Marginalization



$$p(a\,|\,c) \;=\; \int p(a,b\,|\,c)\,\mathrm{d}b \qquad (14)$$

$$p(a\,|\,c) \;=\; \int p(a\,|\,b,c)\,p(b\,|\,c)\,\mathrm{d}b \quad , \qquad (15)$$

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)



Exercise 1: You have conditional pdfs $p(a\,|\,d)$, $p(b\,|\,a,d)$, and $p(c\,|\,a,b,d)$. Write expressions for $p(a,b\,|\,d)$, $p(b\,|\,d)$, and $p(a\,|\,c,d)$.

Exercise 2: You have conditional pdfs $p(a\,|\,b,c)$ and $p(a\,|\,c)$ expressed or computable for any values of a, b, and c. You are not permitted to multiply these together, of course. But can you use them to construct the conditional pdf $p(b\,|\,a,c)$ or $p(b\,|\,c)$? Did you have to make any assumptions?

Exercise 3: You have conditional pdfs $p(a\,|\,c)$ and $p(b\,|\,c)$ expressed or computable for any values of a, b, and c. Can you use them to construct the conditional pdf $p(a\,|\,b,c)$?

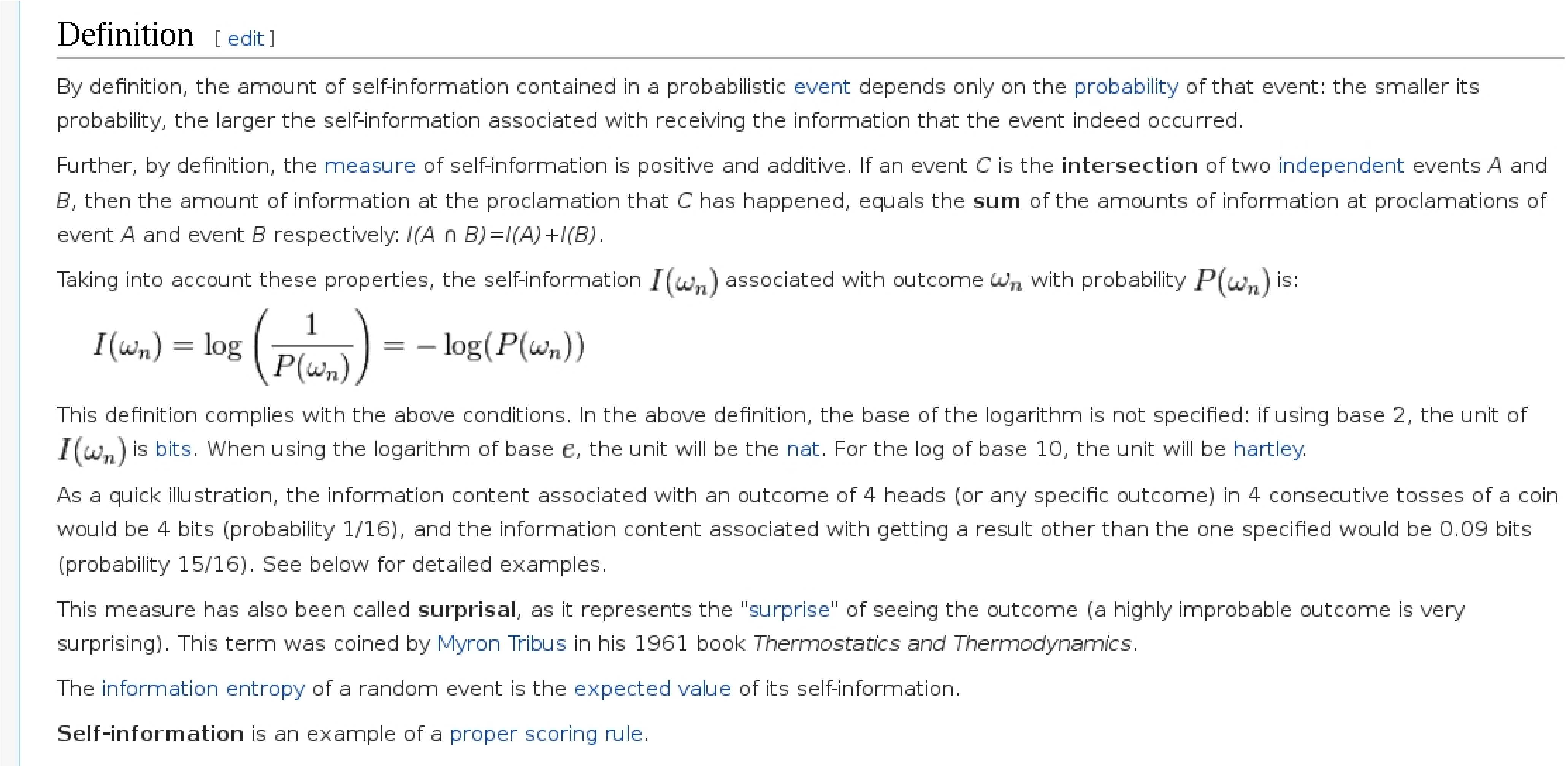
Exercise 4: You have a function $g(b)$ that is a function only of b. You have conditional pdfs $p(a\,|\,c)$ and $p(b\,|\,a,c)$. What is the expectation value $E(g\,|\,c)$ for g conditional on c but *not* conditional on a?

Exercise 5: Take the integral on the right-hand side of equation (15) and replace the "db" with a "da". Is it permissible to do this integral? Why or why not? If it *is* permissible, what do you get?

Data analysis recipes: fitting a model to data (Hogg, Bovy & Lang, 2010)

Concepts from Information Theory

Self-information/Information content



Definition [edit]

By definition, the amount of self-information contained in a probabilistic event depends only on the probability of that event: the smaller its probability, the larger the self-information associated with receiving the information that the event indeed occurred.

Further, by definition, the measure of self-information is positive and additive. If an event *C* is the **intersection** of two independent events *A* and *B*, then the amount of information at the proclamation that *C* has happened, equals the **sum** of the amounts of information at proclamations of event *A* and event *B* respectively: $I(A \cap B) = I(A) + I(B)$.

Taking into account these properties, the self-information $I(\omega_n)$ associated with outcome ω_n with probability $P(\omega_n)$ is:

$$I(\omega_n) = \log\left(\frac{1}{P(\omega_n)}\right) = -\log(P(\omega_n))$$

This definition complies with the above conditions. In the above definition, the base of the logarithm is not specified: if using base 2, the unit of $I(\omega_n)$ is bits. When using the logarithm of base e, the unit will be the nat. For the log of base 10, the unit will be hartley.

As a quick illustration, the information content associated with an outcome of 4 heads (or any specific outcome) in 4 consecutive tosses of a coin would be 4 bits (probability 1/16), and the information content associated with getting a result other than the one specified would be 0.09 bits (probability 15/16). See below for detailed examples.

This measure has also been called **surprisal**, as it represents the "surprise" of seeing the outcome (a highly improbable outcome is very surprising). This term was coined by Myron Tribus in his 1961 book *Thermostatics and Thermodynamics*.

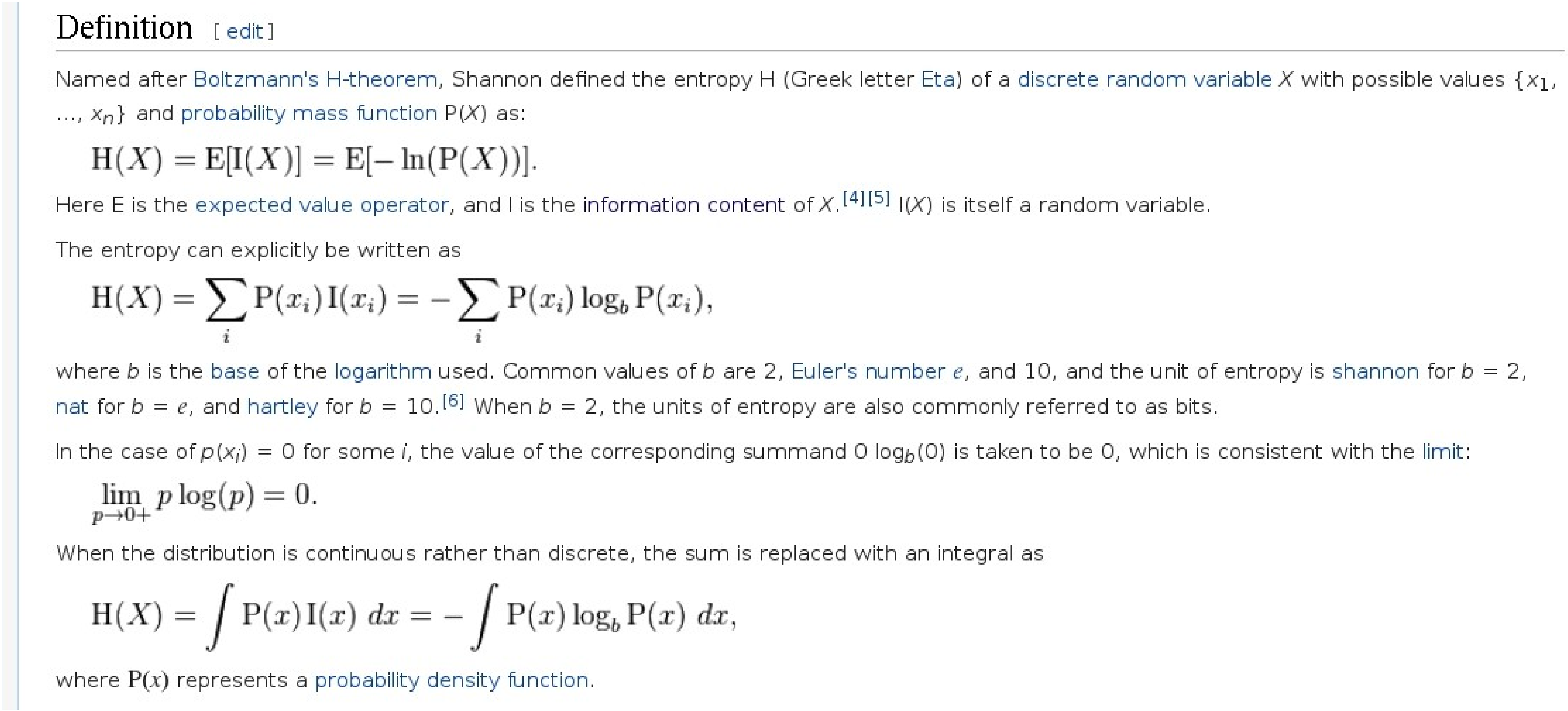
The information entropy of a random event is the expected value of its self-information.

Self-information is an example of a proper scoring rule.

Not to be mistaken with Fisher's information!

(more on this later)

Entropy



Definition

[edit]

Named after Boltzmann's H-theorem, Shannon defined the entropy H (Greek letter Eta) of a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ and probability mass function P(X) as:

$$\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\ln(\mathrm{P}(X))].$$

Here E is the expected value operator, and I is the information content of X.[4][5] I(X) is itself a random variable.

The entropy can explicitly be written as

$$\mathrm{H}(X) = \sum_i \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_i \mathrm{P}(x_i) \log_b \mathrm{P}(x_i),$$

where b is the base of the logarithm used. Common values of b are 2, Euler's number e, and 10, and the unit of entropy is shannon for $b = 2$, nat for $b = e$, and hartley for $b = 10$.[6] When $b = 2$, the units of entropy are also commonly referred to as bits.

In the case of $p(x_i) = 0$ for some i, the value of the corresponding summand $0 \log_b(0)$ is taken to be 0, which is consistent with the limit:

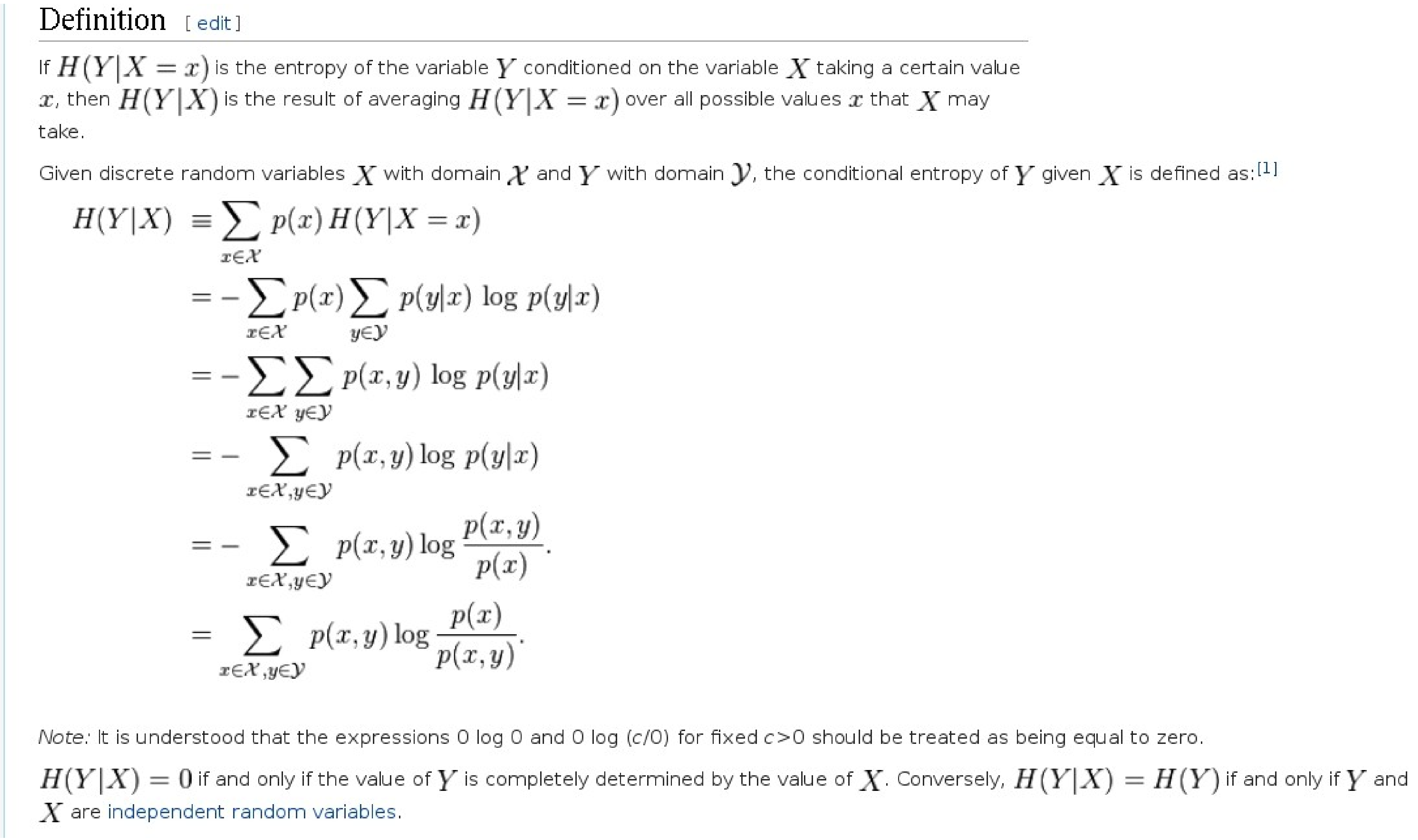
$$\lim_{p \to 0+} p \log(p) = 0.$$

When the distribution is continuous rather than discrete, the sum is replaced with an integral as

$$\mathrm{H}(X) = \int \mathrm{P}(x)\,\mathrm{I}(x)\;dx = -\int \mathrm{P}(x) \log_b \mathrm{P}(x)\;dx,$$

where $\mathrm{P}(x)$ represents a probability density function.

Conditional Entropy



Definition [edit]

If $H(Y|X=x)$ is the entropy of the variable Y conditioned on the variable X taking a certain value x, then $H(Y|X)$ is the result of averaging $H(Y|X=x)$ over all possible values x that X may take.

Given discrete random variables X with domain $\mathcal{X}$ and Y with domain $\mathcal{Y}$, the conditional entropy of Y given X is defined as:[1]

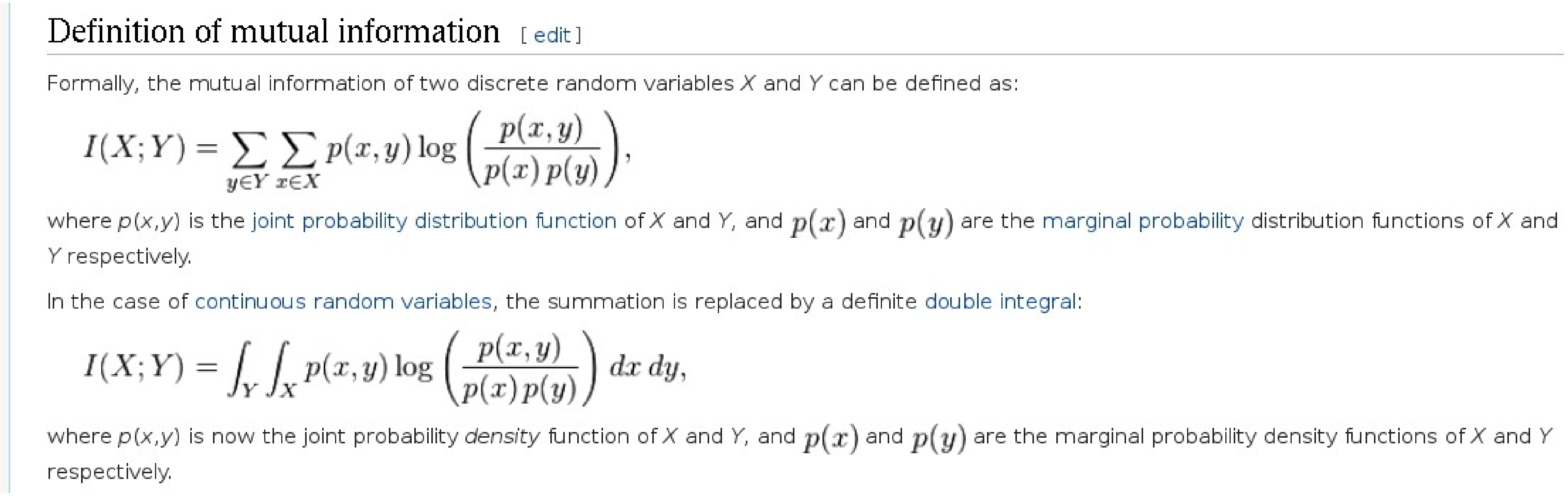
$$\begin{aligned} H(Y|X) &\equiv \sum_{x\in\mathcal{X}} p(x)\, H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\, \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\, \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)}. \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)}. \end{aligned}$$

Note: It is understood that the expressions 0 log 0 and 0 log (*c*/0) for fixed *c*>0 should be treated as being equal to zero.

$H(Y|X)=0$ if and only if the value of Y is completely determined by the value of X. Conversely, $H(Y|X)=H(Y)$ if and only if Y and X are independent random variables.

Mutual Information

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$



Definition of mutual information [edit]

Formally, the mutual information of two discrete random variables *X* and *Y* can be defined as:

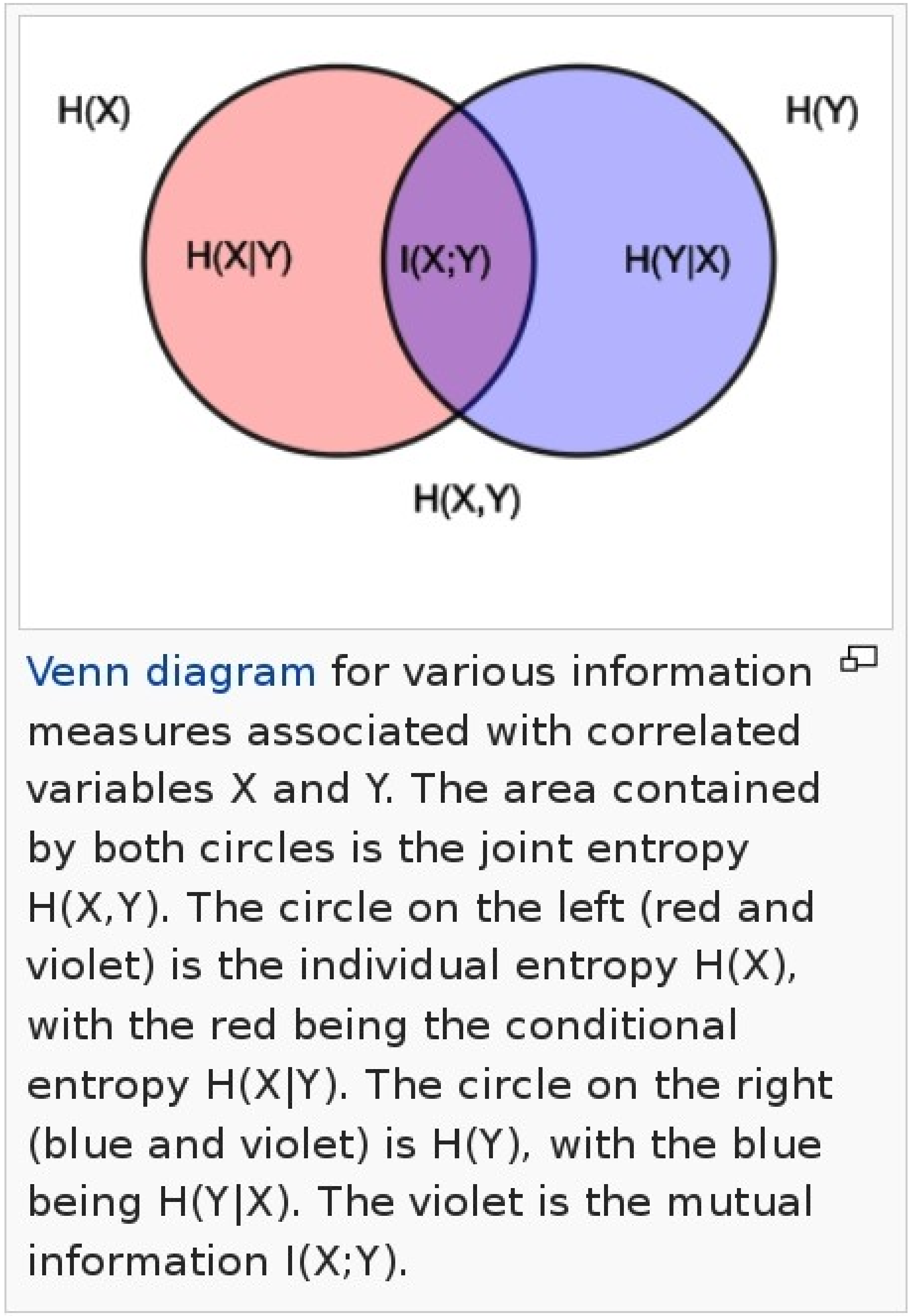
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right),$$

where *p*(*x*,*y*) is the joint probability distribution function of *X* and *Y*, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of *X* and *Y* respectively.

In the case of continuous random variables, the summation is replaced by a definite double integral:

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) dx\, dy,$$

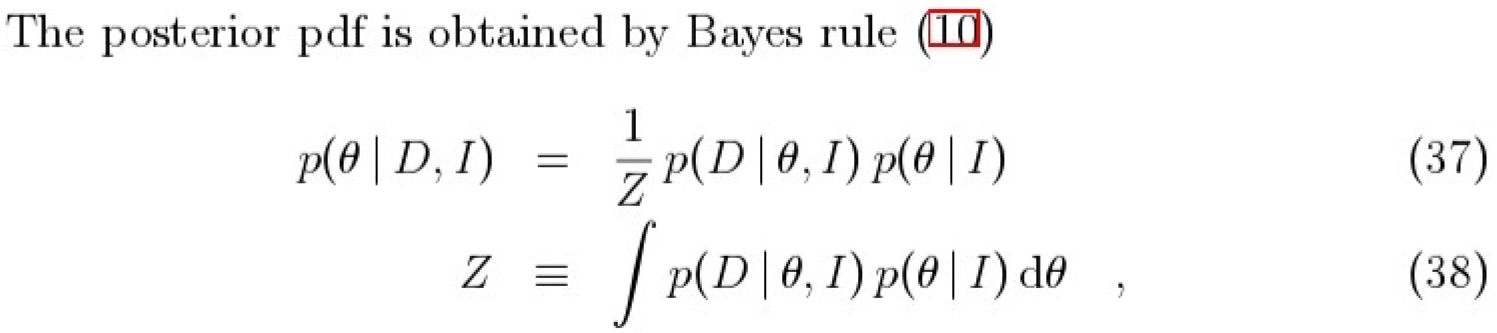
where *p*(*x*,*y*) is now the joint probability *density* function of *X* and *Y*, and $p(x)$ and $p(y)$ are the marginal probability density functions of *X* and *Y* respectively.



Venn diagram for various information measures associated with correlated variables X and Y. The area contained by both circles is the joint entropy H(X,Y). The circle on the left (red and violet) is the individual entropy H(X), with the red being the conditional entropy H(X|Y). The circle on the right (blue and violet) is H(Y), with the blue being H(Y|X). The violet is the mutual information I(X;Y).

Session 2: Fitting a model to data.

Bayes' Theorem

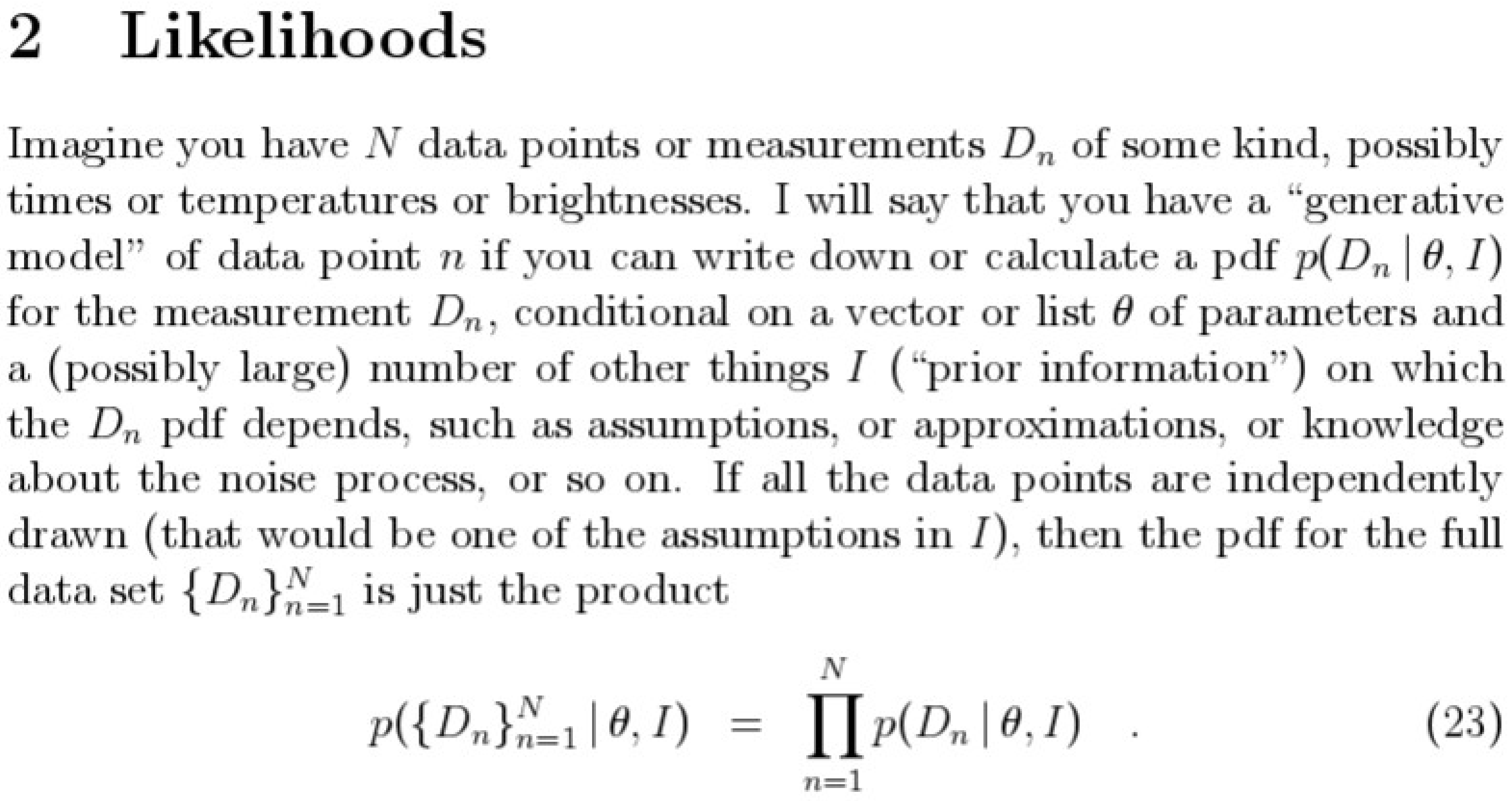


The posterior pdf is obtained by Bayes rule (10)

$$p(\theta \,|\, D, I) \;=\; \frac{1}{Z}\, p(D \,|\, \theta, I)\, p(\theta \,|\, I) \qquad (37)$$

$$Z \;\equiv\; \int p(D \,|\, \theta, I)\, p(\theta \,|\, I)\, \mathrm{d}\theta \quad , \qquad (38)$$

Data Analysis Recipes: Probability Calculus for inference (Hogg, 2012)

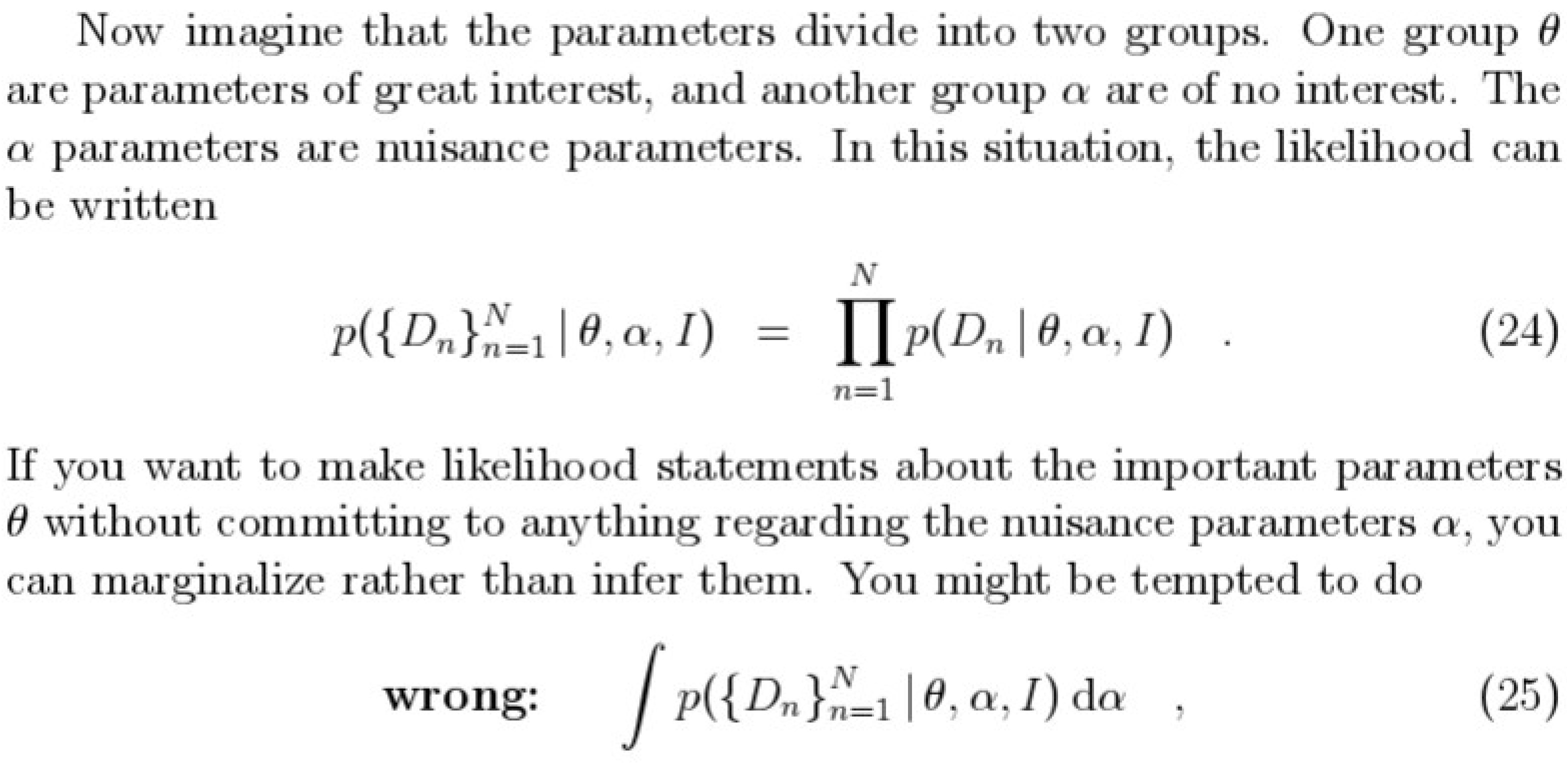


2 Likelihoods

Imagine you have N data points or measurements D_n of some kind, possibly times or temperatures or brightnesses. I will say that you have a "generative model" of data point n if you can write down or calculate a pdf $p(D_n\,|\,\theta, I)$ for the measurement D_n, conditional on a vector or list θ of parameters and a (possibly large) number of other things I ("prior information") on which the D_n pdf depends, such as assumptions, or approximations, or knowledge about the noise process, or so on. If all the data points are independently drawn (that would be one of the assumptions in I), then the pdf for the full data set $\{D_n\}_{n=1}^N$ is just the product

$$p(\{D_n\}_{n=1}^N\,|\,\theta, I) \;=\; \prod_{n=1}^N p(D_n\,|\,\theta, I) \quad . \tag{23}$$

Data Analysis Recipes: Probability Calculus for inference (Hogg, 2012)



Now imagine that the parameters divide into two groups. One group θ are parameters of great interest, and another group α are of no interest. The α parameters are nuisance parameters. In this situation, the likelihood can be written

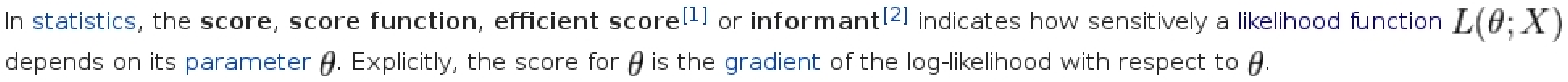
$$p(\{D_n\}_{n=1}^{N}\,|\,\theta,\alpha,I) \;=\; \prod_{n=1}^{N} p(D_n\,|\,\theta,\alpha,I) \quad . \tag{24}$$

If you want to make likelihood statements about the important parameters θ without committing to anything regarding the nuisance parameters α, you can marginalize rather than infer them. You might be tempted to do

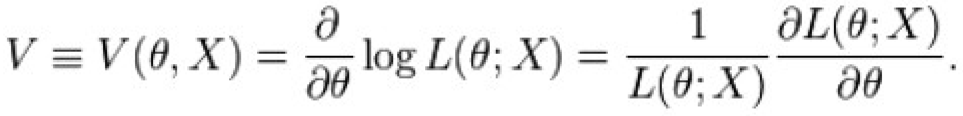
$$\textbf{wrong:} \qquad \int p(\{D_n\}_{n=1}^{N}\,|\,\theta,\alpha,I)\,\mathrm{d}\alpha \quad , \tag{25}$$

Data Analysis Recipes: Probability Calculus for inference (Hogg, 2012)

The score

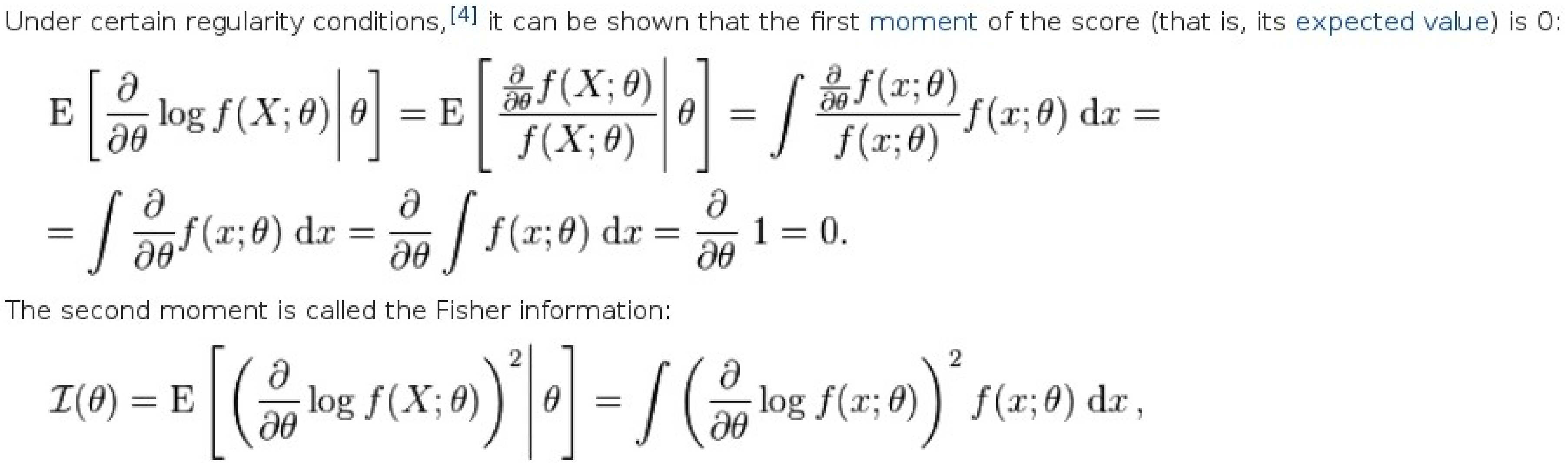


In statistics, the **score**, **score function**, **efficient score**[1] or **informant**[2] indicates how sensitively a likelihood function $L(\theta;X)$ depends on its parameter θ. Explicitly, the score for θ is the gradient of the log-likelihood with respect to θ.



$$V \equiv V(\theta, X) = \frac{\partial}{\partial\theta} \log L(\theta;X) = \frac{1}{L(\theta;X)} \frac{\partial L(\theta;X)}{\partial\theta}.$$

The Fisher Information



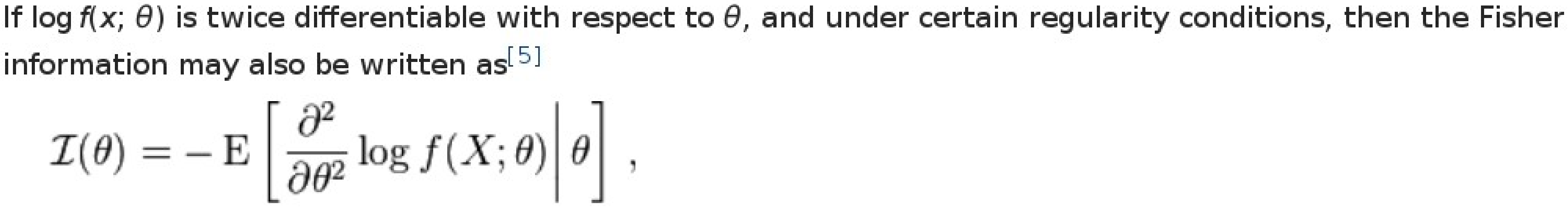
Under certain regularity conditions,[4] it can be shown that the first moment of the score (that is, its expected value) is 0:

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] = \mathrm{E}\left[\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\middle|\theta\right] = \int \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta)\,\mathrm{d}x =$$

$$= \int \frac{\partial}{\partial\theta}f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\,1 = 0.$$

The second moment is called the Fisher information:

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\middle|\theta\right] = \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,\mathrm{d}x\,,$$



If log $f(x;\,\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as[5]

$$\mathcal{I}(\theta) = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\middle|\theta\right]\,,$$

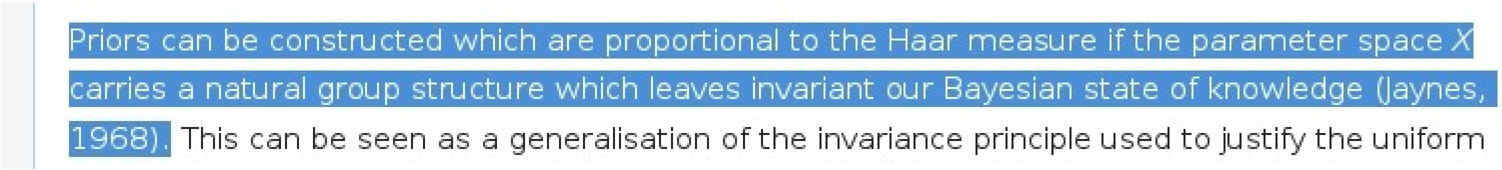
The Crame-Rao bound

$$\mathrm{Var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}\left(\theta\right)}.$$

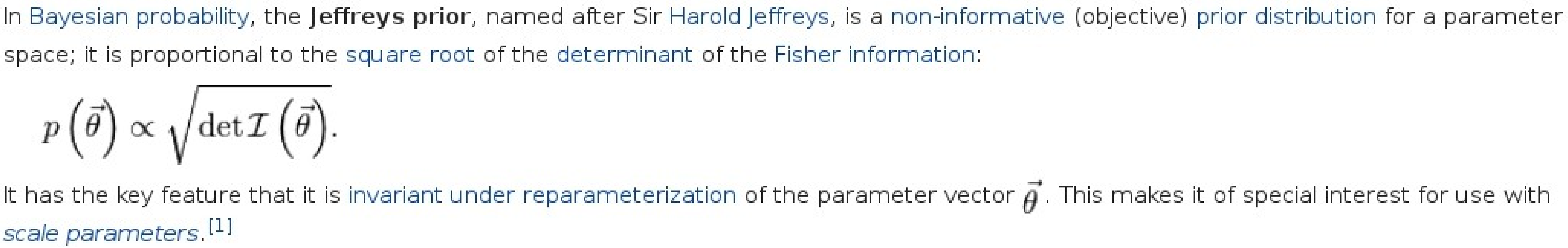
Priors

- Informative priors
- ~~Uninformative~~ Objective priors
  1. Principle of indifference
  2. Principle of invariance under symmetries
  3. Principle of Maximum Entropy
  4. Reference priors

Principle of invariance



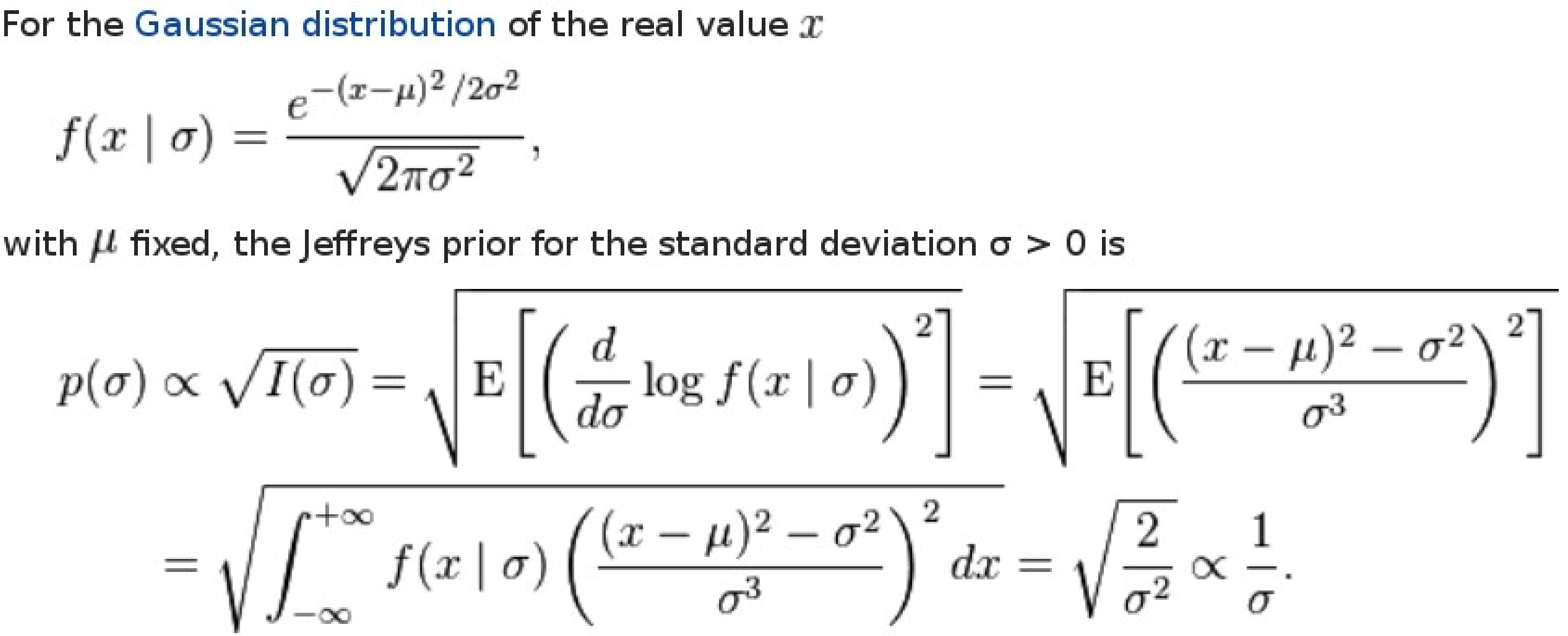
> Priors can be constructed which are proportional to the Haar measure if the parameter space *X* carries a natural group structure which leaves invariant our Bayesian state of knowledge (Jaynes, 1968). This can be seen as a generalisation of the invariance principle used to justify the uniform



In Bayesian probability, the **Jeffreys prior**, named after Sir Harold Jeffreys, is a non-informative (objective) prior distribution for a parameter space; it is proportional to the square root of the determinant of the Fisher information:

$$p\left(\vec{\theta}\right) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}.$$

It has the key feature that it is invariant under reparameterization of the parameter vector $\vec{\theta}$. This makes it of special interest for use with *scale parameters*.[1]



For the Gaussian distribution of the real value x

$$f(x \mid \sigma) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}},$$

with μ fixed, the Jeffreys prior for the standard deviation σ > 0 is

$$p(\sigma) \propto \sqrt{I(\sigma)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\sigma}\log f(x \mid \sigma)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2\right]}$$

$$= \sqrt{\int_{-\infty}^{+\infty} f(x \mid \sigma)\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2 dx} = \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma}.$$

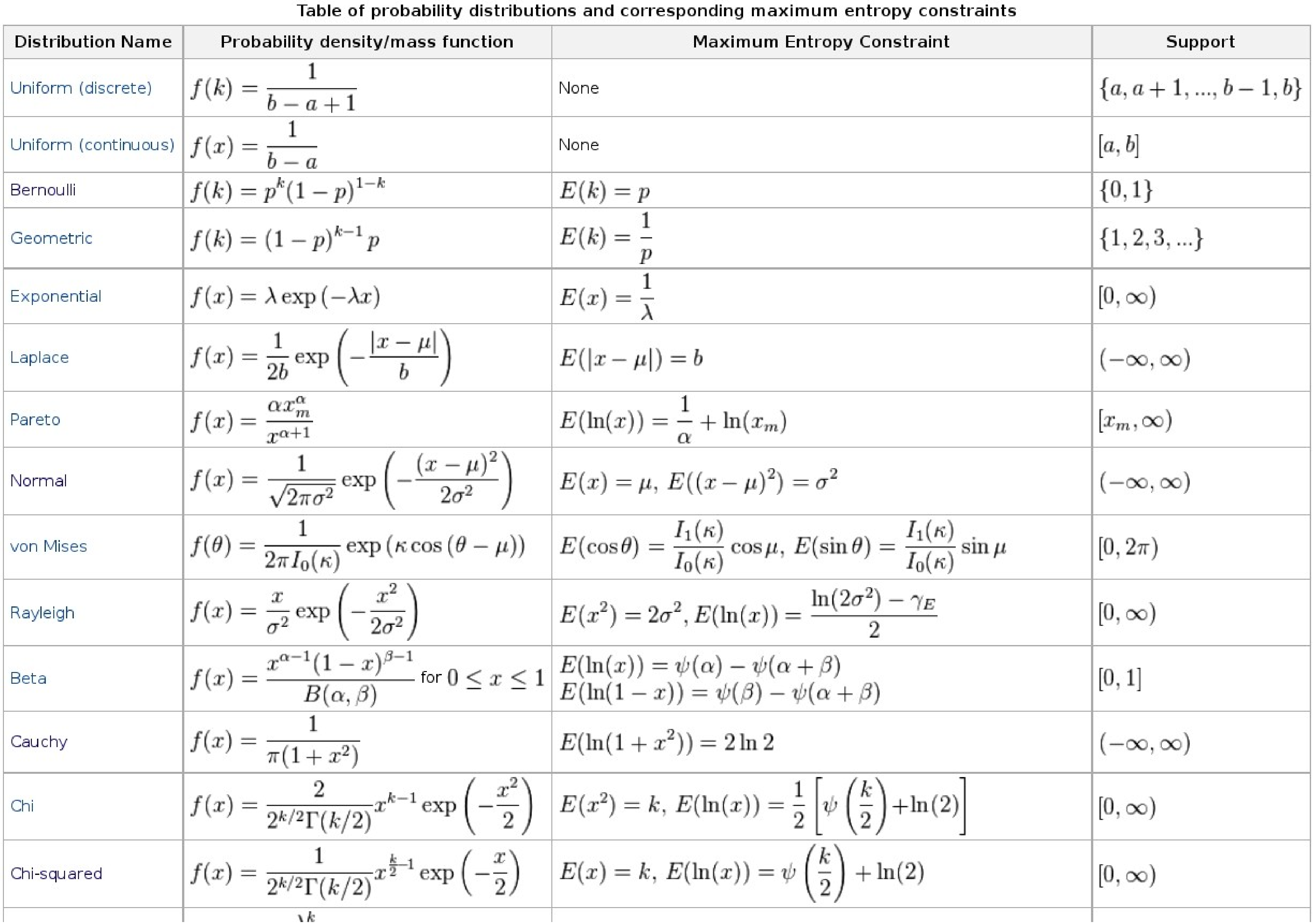
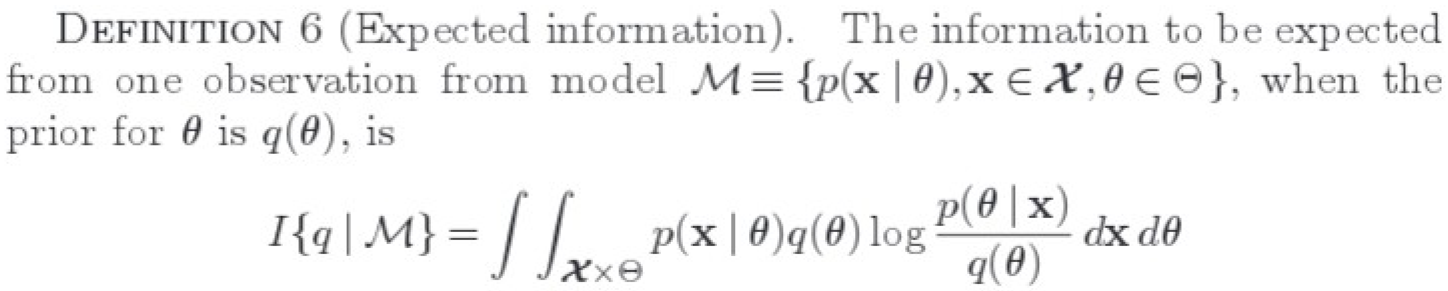


Table of probability distributions and corresponding maximum entropy constraints

Distribution Name	Probability density/mass function	Maximum Entropy Constraint	Support
Uniform (discrete)	$f(k) = \frac{1}{b-a+1}$	None	$\{a, a+1, ..., b-1, b\}$
Uniform (continuous)	$f(x) = \frac{1}{b-a}$	None	$[a, b]$
Bernoulli	$f(k) = p^k(1-p)^{1-k}$	$E(k) = p$	$\{0, 1\}$
Geometric	$f(k) = (1-p)^{k-1}\, p$	$E(k) = \frac{1}{p}$	$\{1, 2, 3, ...\}$
Exponential	$f(x) = \lambda \exp\left(-\lambda x\right)$	$E(x) = \frac{1}{\lambda}$	$[0, \infty)$
Laplace	$f(x) = \frac{1}{2b} \exp\left(-\frac{\|x-\mu\|}{b}\right)$	$E(\|x-\mu\|) = b$	$(-\infty, \infty)$
Pareto	$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$E(\ln(x)) = \frac{1}{\alpha} + \ln(x_m)$	$[x_m, \infty)$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$E(x) = \mu,\ E((x-\mu)^2) = \sigma^2$	$(-\infty, \infty)$
von Mises	$f(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \cos\left(\theta - \mu\right)\right)$	$E(\cos\theta) = \frac{I_1(\kappa)}{I_0(\kappa)} \cos\mu,\ E(\sin\theta) = \frac{I_1(\kappa)}{I_0(\kappa)} \sin\mu$	$[0, 2\pi)$
Rayleigh	$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$	$E(x^2) = 2\sigma^2, E(\ln(x)) = \frac{\ln(2\sigma^2) - \gamma_E}{2}$	$[0, \infty)$
Beta	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ for $0 \leq x \leq 1$	$E(\ln(x)) = \psi(\alpha) - \psi(\alpha+\beta)$ $E(\ln(1-x)) = \psi(\beta) - \psi(\alpha+\beta)$	$[0, 1]$
Cauchy	$f(x) = \frac{1}{\pi(1+x^2)}$	$E(\ln(1+x^2)) = 2\ln 2$	$(-\infty, \infty)$
Chi	$f(x) = \frac{2}{2^{k/2}\Gamma(k/2)} x^{k-1} \exp\left(-\frac{x^2}{2}\right)$	$E(x^2) = k,\ E(\ln(x)) = \frac{1}{2}\left[\psi\left(\frac{k}{2}\right) + \ln(2)\right]$	$[0, \infty)$
Chi-squared	$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right)$	$E(x) = k,\ E(\ln(x)) = \psi\left(\frac{k}{2}\right) + \ln(2)$	$[0, \infty)$
	λ^k		

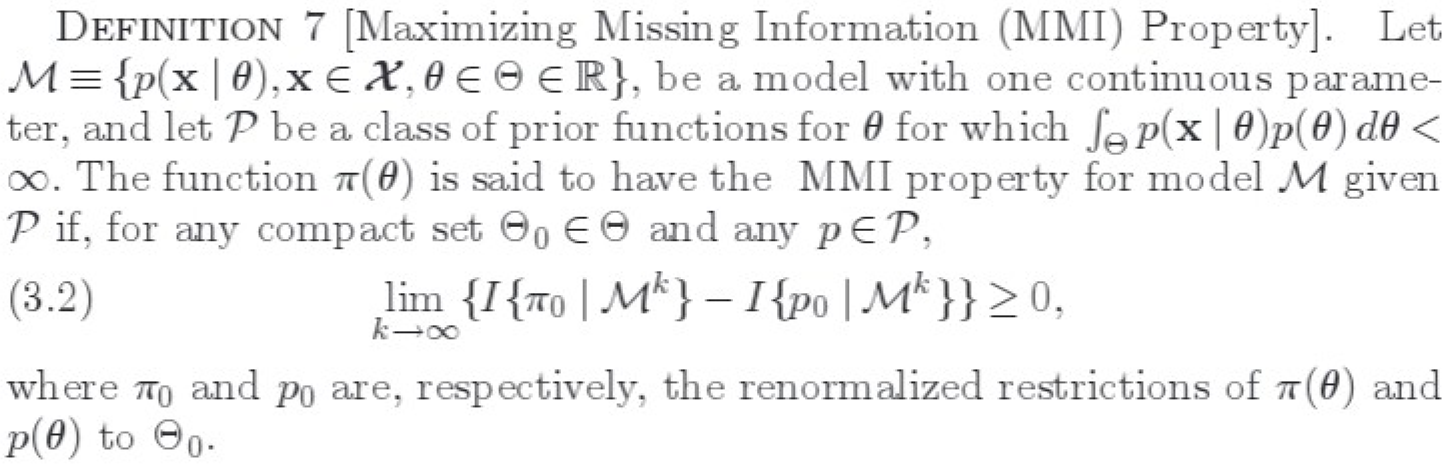
The formal definition of Reference priors

(Berger, Bernardo & Sun, 2009)



DEFINITION 6 (Expected information). The information to be expected from one observation from model $\mathcal{M} \equiv \{p(\mathbf{x} \mid \theta), \mathbf{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, when the prior for θ is $q(\theta)$, is

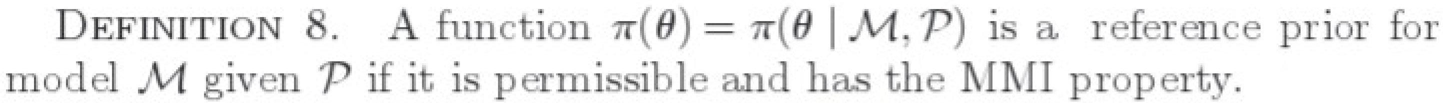
$$I\{q \mid \mathcal{M}\} = \int\int_{\boldsymbol{\mathcal{X}} \times \Theta} p(\mathbf{x} \mid \theta) q(\theta) \log \frac{p(\theta \mid \mathbf{x})}{q(\theta)} \, d\mathbf{x} \, d\theta$$



DEFINITION 7 [Maximizing Missing Information (MMI) Property]. Let $\mathcal{M} \equiv \{p(\mathbf{x} \mid \theta), \mathbf{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta \in \mathbb{R}\}$, be a model with one continuous parameter, and let $\mathcal{P}$ be a class of prior functions for θ for which $\int_\Theta p(\mathbf{x} \mid \theta) p(\theta) \, d\theta < \infty$. The function $\pi(\theta)$ is said to have the MMI property for model $\mathcal{M}$ given $\mathcal{P}$ if, for any compact set $\Theta_0 \in \Theta$ and any $p \in \mathcal{P}$,

$$\lim_{k \to \infty} \{I\{\pi_0 \mid \mathcal{M}^k\} - I\{p_0 \mid \mathcal{M}^k\}\} \geq 0, \tag{3.2}$$

where π_0 and p_0 are, respectively, the renormalized restrictions of $\pi(\theta)$ and $p(\theta)$ to Θ_0.

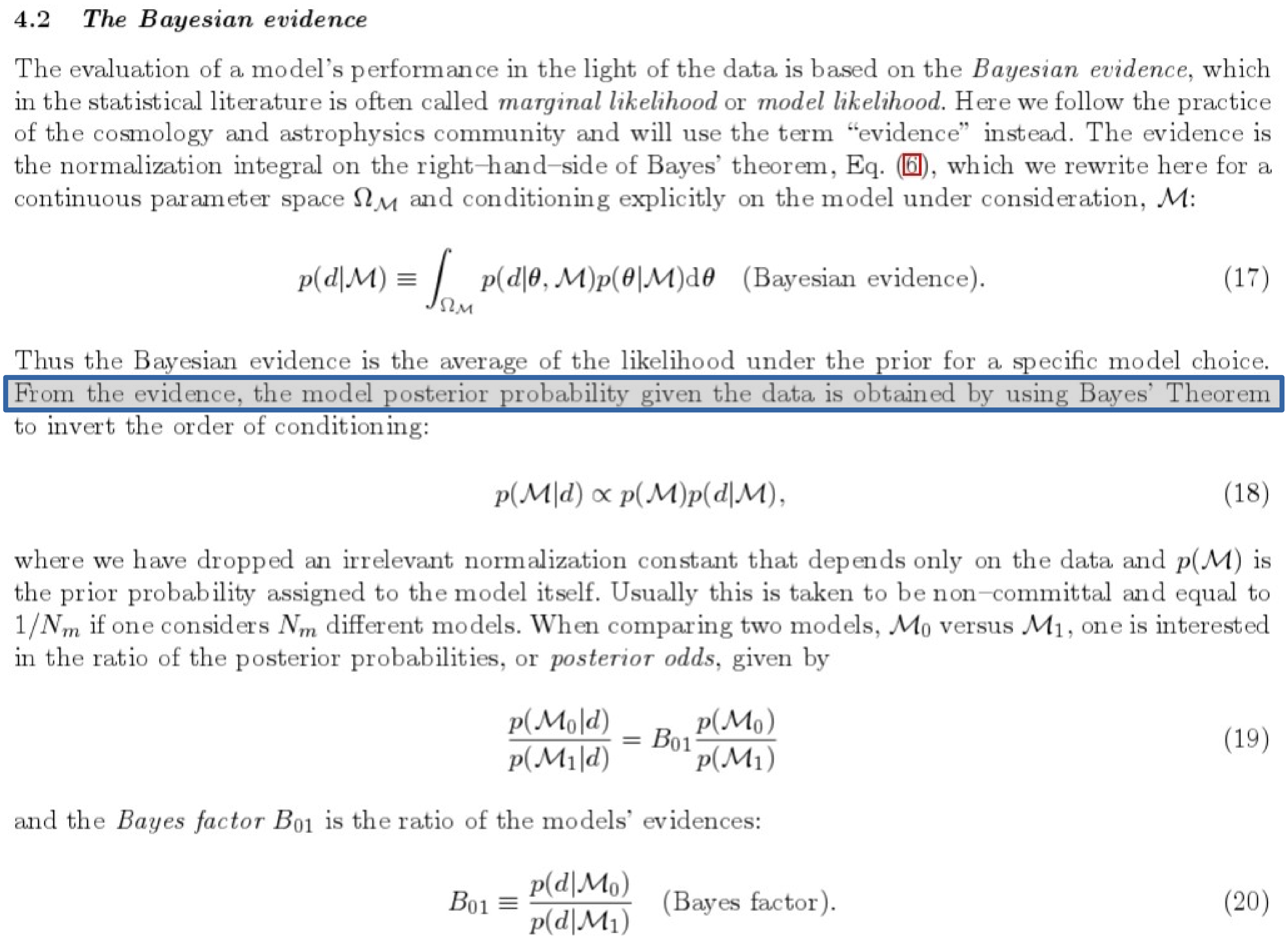


DEFINITION 8. A function $\pi(\theta) = \pi(\theta \mid \mathcal{M}, \mathcal{P})$ is a reference prior for model $\mathcal{M}$ given $\mathcal{P}$ if it is permissible and has the MMI property.

Catalog of reference priors

Session 3: Model Selection

Model selection: the bayesian approach



4.2 *The Bayesian evidence*

The evaluation of a model's performance in the light of the data is based on the *Bayesian evidence*, which in the statistical literature is often called *marginal likelihood* or *model likelihood*. Here we follow the practice of the cosmology and astrophysics community and will use the term "evidence" instead. The evidence is the normalization integral on the right–hand–side of Bayes' theorem, Eq. (6), which we rewrite here for a continuous parameter space $\Omega_{\mathcal{M}}$ and conditioning explicitly on the model under consideration, $\mathcal{M}$:

$$p(d|\mathcal{M}) \equiv \int_{\Omega_{\mathcal{M}}} p(d|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \mathrm{d}\theta \quad \text{(Bayesian evidence)}. \tag{17}$$

Thus the Bayesian evidence is the average of the likelihood under the prior for a specific model choice. From the evidence, the model posterior probability given the data is obtained by using Bayes' Theorem to invert the order of conditioning:

$$p(\mathcal{M}|d) \propto p(\mathcal{M}) p(d|\mathcal{M}), \tag{18}$$

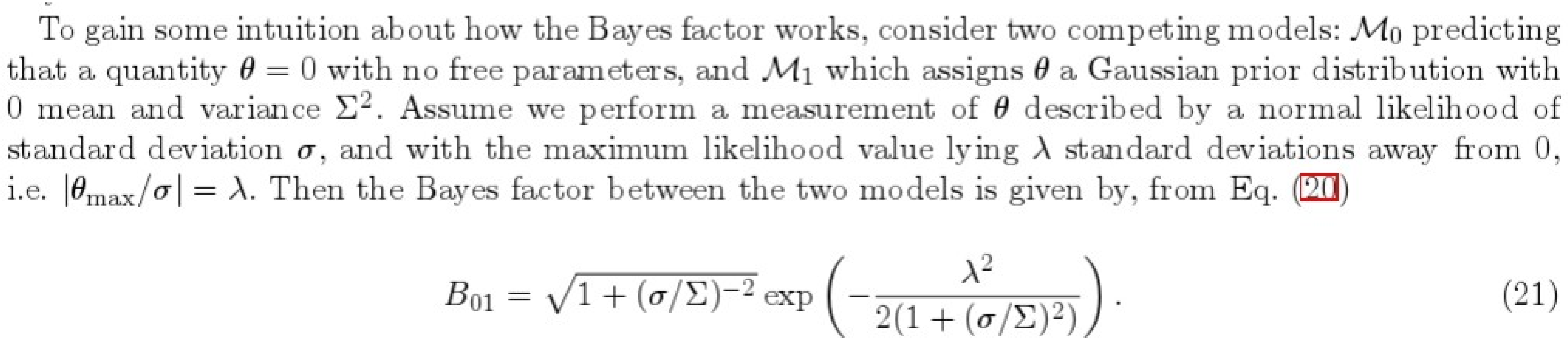
where we have dropped an irrelevant normalization constant that depends only on the data and $p(\mathcal{M})$ is the prior probability assigned to the model itself. Usually this is taken to be non–committal and equal to $1/N_m$ if one considers N_m different models. When comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{p(\mathcal{M}_0|d)}{p(\mathcal{M}_1|d)} = B_{01} \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)} \tag{19}$$

and the *Bayes factor* B_{01} is the ratio of the models' evidences:

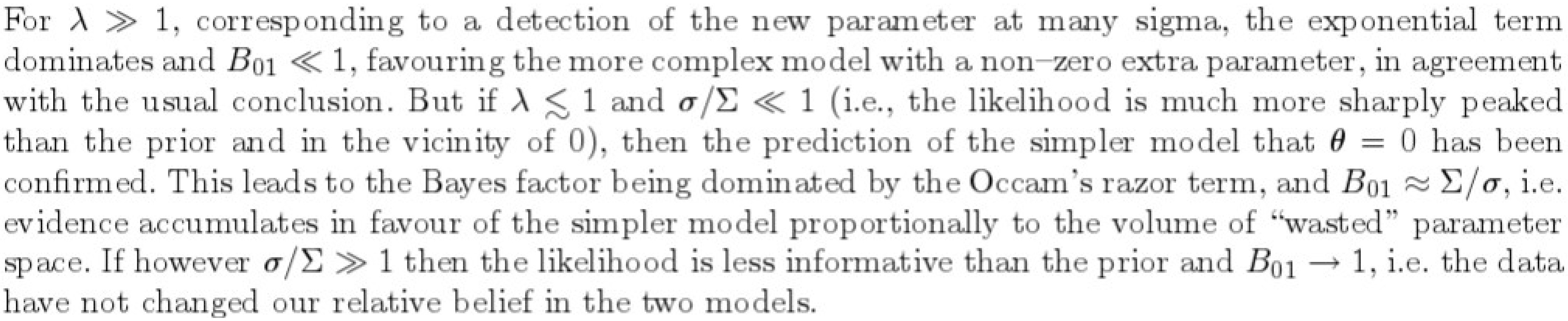
$$B_{01} \equiv \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)} \quad \text{(Bayes factor)}. \tag{20}$$

Bayes in the sky, R. Trotta (2008)



To gain some intuition about how the Bayes factor works, consider two competing models: $\mathcal{M}_0$ predicting that a quantity $\theta = 0$ with no free parameters, and $\mathcal{M}_1$ which assigns θ a Gaussian prior distribution with 0 mean and variance Σ^2. Assume we perform a measurement of θ described by a normal likelihood of standard deviation σ, and with the maximum likelihood value lying λ standard deviations away from 0, i.e. $|\theta_{\max}/\sigma| = \lambda$. Then the Bayes factor between the two models is given by, from Eq. (20)

$$B_{01} = \sqrt{1 + (\sigma/\Sigma)^{-2}} \exp\left(-\frac{\lambda^2}{2(1 + (\sigma/\Sigma)^2)} \right). \tag{21}$$



For $\lambda \gg 1$, corresponding to a detection of the new parameter at many sigma, the exponential term dominates and $B_{01} \ll 1$, favouring the more complex model with a non–zero extra parameter, in agreement with the usual conclusion. But if $\lambda \lesssim 1$ and $\sigma/\Sigma \ll 1$ (i.e., the likelihood is much more sharply peaked than the prior and in the vicinity of 0), then the prediction of the simpler model that $\theta = 0$ has been confirmed. This leads to the Bayes factor being dominated by the Occam's razor term, and $B_{01} \approx \Sigma/\sigma$, i.e. evidence accumulates in favour of the simpler model proportionally to the volume of "wasted" parameter space. If however $\sigma/\Sigma \gg 1$ then the likelihood is less informative than the prior and $B_{01} \to 1$, i.e. the data have not changed our relative belief in the two models.

Bayes in the sky, R. Trotta (2008)

(iv) An instructive approximation to the Bayesian evidence can be obtained when the likelihood function is unimodal and approximately Gaussian in the parameters. Expanding the likelihood around its peak to second order one obtains the Laplace approximation

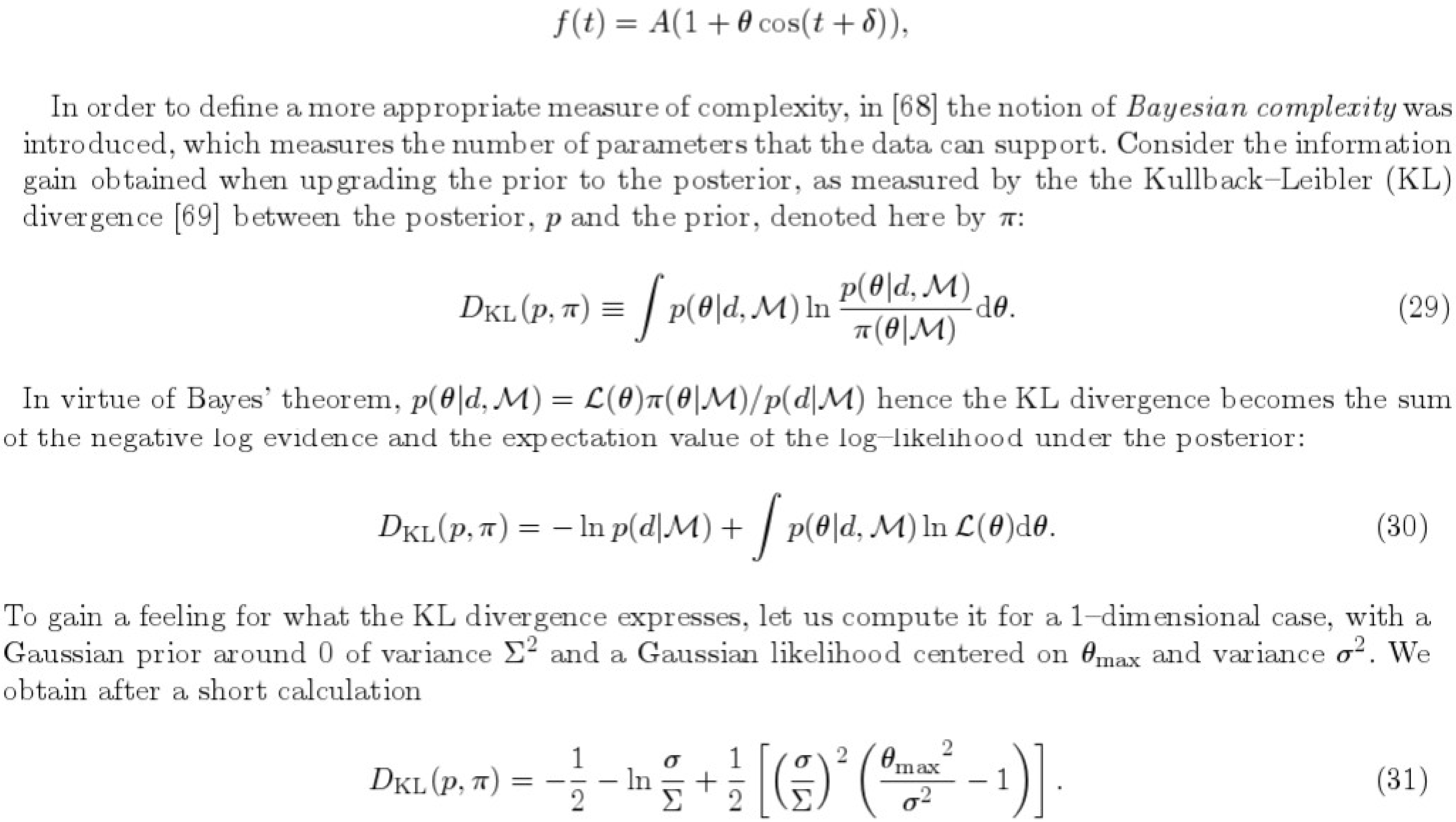
$$p(d|\boldsymbol{\theta}, \mathcal{M}) \approx \mathcal{L}_{\max} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})^t L(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})\right], \tag{24}$$

where $\boldsymbol{\theta}_{\max}$ is the maximum–likelihood point, $\mathcal{L}_{\max}$ the maximum likelihood value and L the likelihood Fisher matrix (which is the inverse of the covariance matrix for the parameters). Assuming as a prior a multinormal Gaussian distribution with zero mean and Fisher information matrix P one obtains for the evidence, Eq. (17)

$$p(d|\mathcal{M}) = \mathcal{L}_{\max} \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_{\max}{}^t L \boldsymbol{\theta}_{\max} - \overline{\boldsymbol{\theta}}^t F \overline{\boldsymbol{\theta}})\right], \tag{25}$$

From Eq. (25) we can deduce a few qualitatively relevant properties of the evidence. First, the quality of fit of the model is expressed by $\mathcal{L}_{\max}$, the best–fit likelihood. Thus a model which fits the data better will be favoured by this term. The term involving the determinants of P and F is a volume factor, encoding the Occam's razor effect. As $|P| \leq |F|$, it penalizes models with a large volume of wasted parameter space, i.e. those for which the parameter space volume $|F|^{-1/2}$ which survives after arrival of the data is much smaller than the initially available parameter space under the model prior, $|P|^{-1/2}$. Finally, the exponential term suppresses the likelihood of models for which the parameters values which maximise the likelihood, $\boldsymbol{\theta}_{\max}$, differ appreciably from the expectation value under the posterior, $\overline{\boldsymbol{\theta}}$. Therefore when we consider a model with an increased number of parameters we see that *its evidence will be larger only if the quality–of–fit increases enough to offset the penalizing effect of the Occam's factor* (see also the discussion in [57]).

On the other hand, it is important to notice that the Bayesian evidence does *not* penalizes models with parameters that are unconstrained by the data. It is easy to see that unmeasured parameters (i.e., parameters whose posterior is equal to the prior) do not contribute to the evidence integral, and hence model comparison does not act against them, awaiting better data.



$$f(t) = A(1 + \theta\cos(t + \delta)),$$

In order to define a more appropriate measure of complexity, in [68] the notion of *Bayesian complexity* was introduced, which measures the number of parameters that the data can support. Consider the information gain obtained when upgrading the prior to the posterior, as measured by the the Kullback–Leibler (KL) divergence [69] between the posterior, p and the prior, denoted here by π:

$$D_{\rm KL}(p, \pi) \equiv \int p(\theta|d, \mathcal{M}) \ln \frac{p(\theta|d, \mathcal{M})}{\pi(\theta|\mathcal{M})} {\rm d}\theta. \tag{29}$$

In virtue of Bayes' theorem, $p(\theta|d, \mathcal{M}) = \mathcal{L}(\theta)\pi(\theta|\mathcal{M})/p(d|\mathcal{M})$ hence the KL divergence becomes the sum of the negative log evidence and the expectation value of the log–likelihood under the posterior:

$$D_{\rm KL}(p, \pi) = -\ln p(d|\mathcal{M}) + \int p(\theta|d, \mathcal{M}) \ln \mathcal{L}(\theta){\rm d}\theta. \tag{30}$$

To gain a feeling for what the KL divergence expresses, let us compute it for a 1–dimensional case, with a Gaussian prior around 0 of variance Σ^2 and a Gaussian likelihood centered on $\theta_{\rm max}$ and variance σ^2. We obtain after a short calculation

$$D_{\rm KL}(p, \pi) = -\frac{1}{2} - \ln\frac{\sigma}{\Sigma} + \frac{1}{2}\left[\left(\frac{\sigma}{\Sigma}\right)^2\left(\frac{{\theta_{\rm max}}^2}{\sigma^2} - 1\right)\right]. \tag{31}$$

Bayes in the sky, R. Trotta (2008)

In order to define a more appropriate measure of complexity, in [68] the notion of *Bayesian complexity* was introduced, which measures the number of parameters that the data can support. Consider the information gain obtained when upgrading the prior to the posterior, as measured by the the Kullback–Leibler (KL) divergence [69] between the posterior, p and the prior, denoted here by π:

$$D_{\mathrm{KL}}(p,\pi) \equiv \int p(\boldsymbol{\theta}|d,\mathcal{M}) \ln \frac{p(\boldsymbol{\theta}|d,\mathcal{M})}{\pi(\boldsymbol{\theta}|\mathcal{M})} \mathrm{d}\boldsymbol{\theta}. \tag{29}$$

$$D_{\mathrm{KL}}(p,\pi) = -\ln p(d|\mathcal{M}) + \int p(\boldsymbol{\theta}|d,\mathcal{M}) \ln \mathcal{L}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Let us now define an effective χ^2 through the likelihood as $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\chi^2/2)$. Then Eq. (30) gives

$$D_{\mathrm{KL}}(p, \pi) = -\frac{1}{2}\overline{\chi^2(\boldsymbol{\theta})} + \ln p(d|\mathcal{M}), \tag{32}$$

where the bar indicates a mean taken over the posterior distribution. The posterior average of the effective chi–square is a quantity can be easily obtained by Markov chain Monte Carlo techniques (see section 3.2). We then subtract from the "expected surprise" the estimated surprise in the data after we have actually fitted the model parameters, denoted by

$$\widehat{D_{\mathrm{KL}}} \equiv -\frac{1}{2}\chi^2(\hat{\boldsymbol{\theta}}) + \ln p(d|\mathcal{M}), \tag{33}$$

$$\mathcal{C}_b \equiv -2\left(D_{\mathrm{KL}}(p, \pi) - \widehat{D_{\mathrm{KL}}}\right) = \overline{\chi^2(\boldsymbol{\theta})} - \chi^2(\hat{\boldsymbol{\theta}}) \quad \text{(Bayesian complexity)},$$

Model Selection: the Information approach

- **Akaike Information Criterion (AIC):** Introduced by Akaike [74], the AIC is an essentially frequentist criterion that sets the penalty term equal to twice the number of free parameters in the model, k:

$$\mathrm{AIC} \equiv -2 \ln \mathcal{L}_{\mathrm{max}} + 2k \tag{36}$$

where $\mathcal{L}_{\mathrm{max}} \equiv p(d|\boldsymbol{\theta}_{\mathrm{max}}, \mathcal{M})$ is the maximum likelihood value. The derivation of the AIC follows from an approximate minimization of the KL divergence between the true model distribution and the distribution being fitted to the data.

- **Bayesian Information Criterion (BIC):** Sometimes called "Schwarz Information Criterion" (from the name of its proposer [75]), the BIC follows from a Gaussian approximation to the Bayesian evidence in the limit of large sample size:

  $$\mathrm{BIC} \equiv -2 \ln \mathcal{L}_{\mathrm{max}} + k \ln N \tag{37}$$

  where k is the number of fitted parameters as before and N is the number of data points. The best model is again the one that minimizes the BIC.

- **Deviance Information Criterion (DIC):** Introduced by [68], the DIC can be written as

$$\mathrm{DIC} \equiv -2\widehat{D_{\mathrm{KL}}} + 2\mathcal{C}_b. \tag{38}$$

  In this form, the DIC is reminiscent of the AIC, with the $\ln \mathcal{L}_{\max}$ term replaced by the estimated KL divergence and the number of free parameters by the effective number of parameters, $\mathcal{C}_b$, from Eq. (34). Indeed, in the limit of well–constrained parameters, the AIC is recovered from (38), but the DIC has the advantage of accounting for unconstrained directions in parameters space.

Session 4: Hands-on session

Let me quote directly from Coryn's paper:

- I have (he has) shown that estimating the distance given a parallax is not trivial.
- The naive approach of reporting $1/p \pm \sigma_p / p^2$ fails for non-positive parallaxes, is extremely noisy for fractional parallax errors larger than about 20%, and gives an incorrect (symmetric) error estimate.
- Probability-based inference for the general case is unavoidable.
- Adopting an “uninformative” improper uniform prior over all positive r does not solve any of these problems, and is in fact both informative and implausible (viz. a volume density of stars varying as $1/r^2$).
- The problems can be avoided by using a properly normalized prior which necessarily decreases after some distance.

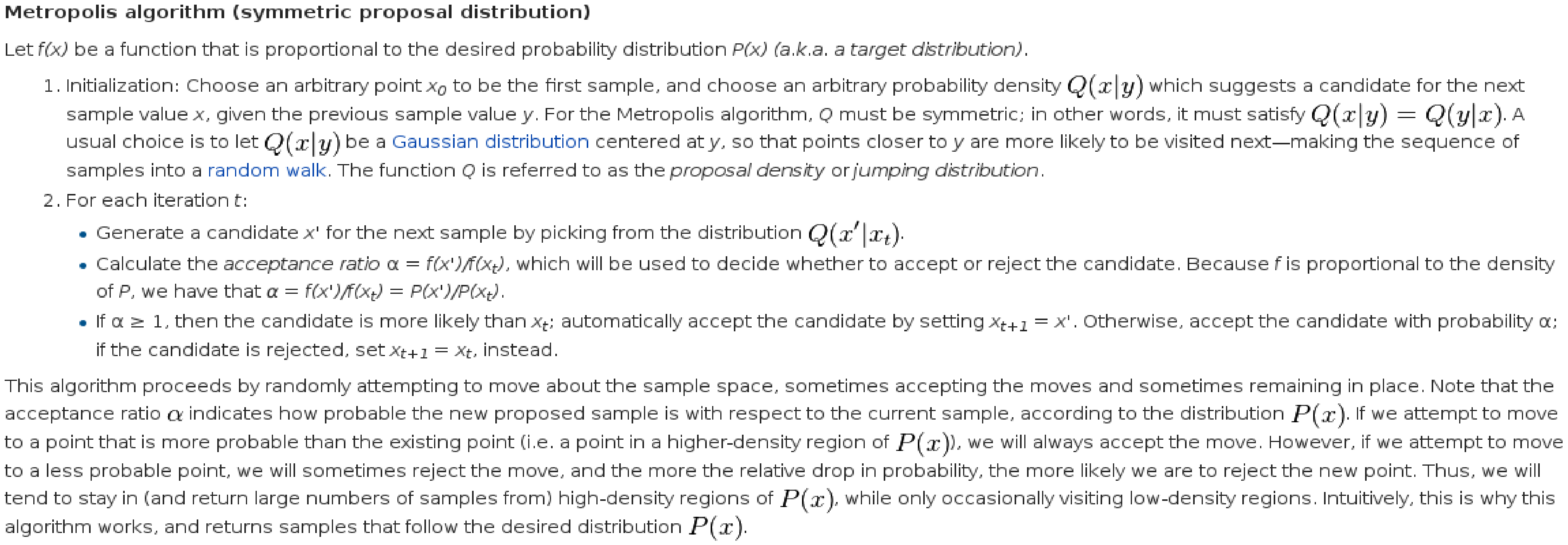
- The use of a prior with a sharp cut-off is not recommended, because it introduces significant biases for low accuracy measurements at all distances (not just those near the cut-off).
- Instead, a prior which converges asymptotically to zero as distance goes to infinity should be used.

- The mode of the corresponding posterior is an unbiased estimator (for the case that the population observed conforms to the prior) and it provides meaningful estimates for arbitrarily large parallax errors as well as non-positive parallaxes. Bias corrections are not necessary. The variance of this estimator behaves as well as one could expect given the irreducible measurement errors.
- The median and mean perform less well than the mode.

- Distance estimates must be accompanied by uncertainty estimates. For fractional parallax errors larger than about 20%, the posterior over distance is significantly asymmetric, so we should always report confidence intervals using quantiles (e.g. 5% and 95%) rather than standard deviations.

Session 5: MCMC techniques and hierarchical Bayes

Metropolis-Hastings algorithm, the wikipedia



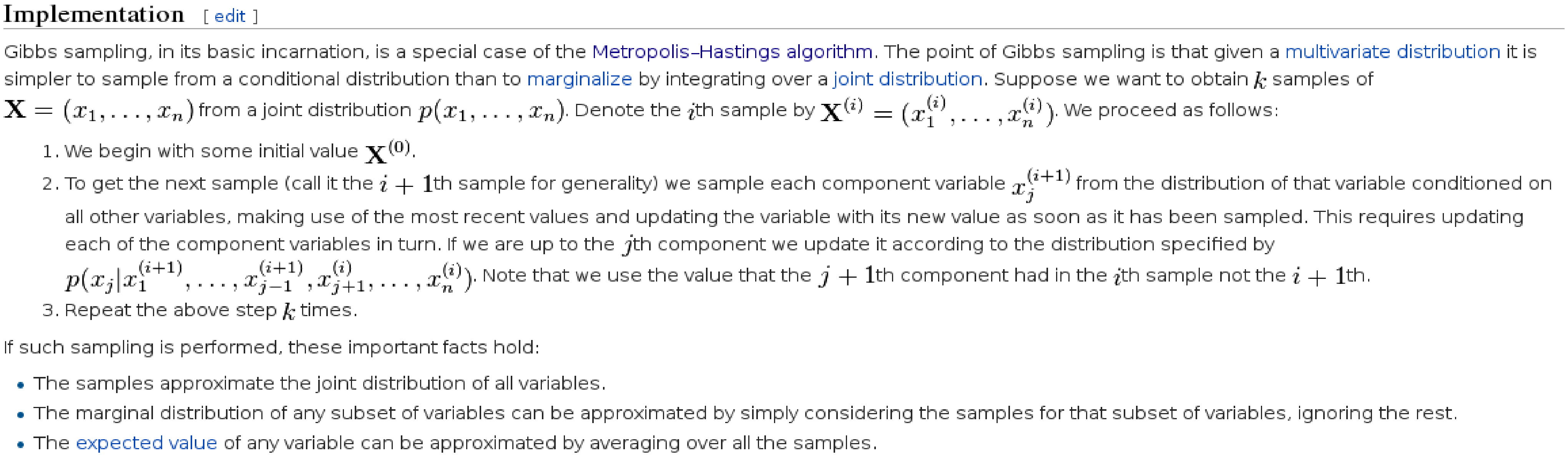
Metropolis algorithm (symmetric proposal distribution)

Let *f(x)* be a function that is proportional to the desired probability distribution *P(x) (a.k.a. a target distribution)*.

1. Initialization: Choose an arbitrary point x_0 to be the first sample, and choose an arbitrary probability density $Q(x|y)$ which suggests a candidate for the next sample value *x*, given the previous sample value *y*. For the Metropolis algorithm, *Q* must be symmetric; in other words, it must satisfy $Q(x|y) = Q(y|x)$. A usual choice is to let $Q(x|y)$ be a Gaussian distribution centered at *y*, so that points closer to *y* are more likely to be visited next—making the sequence of samples into a random walk. The function *Q* is referred to as the *proposal density* or *jumping distribution*.
2. For each iteration *t*:
   - Generate a candidate *x*' for the next sample by picking from the distribution $Q(x'|x_t)$.
   - Calculate the *acceptance ratio* $\alpha = f(x')/f(x_t)$, which will be used to decide whether to accept or reject the candidate. Because *f* is proportional to the density of *P*, we have that $\alpha = f(x')/f(x_t) = P(x')/P(x_t)$.
   - If $\alpha \geq 1$, then the candidate is more likely than x_t; automatically accept the candidate by setting $x_{t+1} = x'$. Otherwise, accept the candidate with probability α; if the candidate is rejected, set $x_{t+1} = x_t$, instead.

This algorithm proceeds by randomly attempting to move about the sample space, sometimes accepting the moves and sometimes remaining in place. Note that the acceptance ratio α indicates how probable the new proposed sample is with respect to the current sample, according to the distribution $P(x)$. If we attempt to move to a point that is more probable than the existing point (i.e. a point in a higher-density region of $P(x)$), we will always accept the move. However, if we attempt to move to a less probable point, we will sometimes reject the move, and the more the relative drop in probability, the more likely we are to reject the new point. Thus, we will tend to stay in (and return large numbers of samples from) high-density regions of $P(x)$, while only occasionally visiting low-density regions. Intuitively, this is why this algorithm works, and returns samples that follow the desired distribution $P(x)$.

Gibbs sampling, the wikipedia



Implementation [edit]

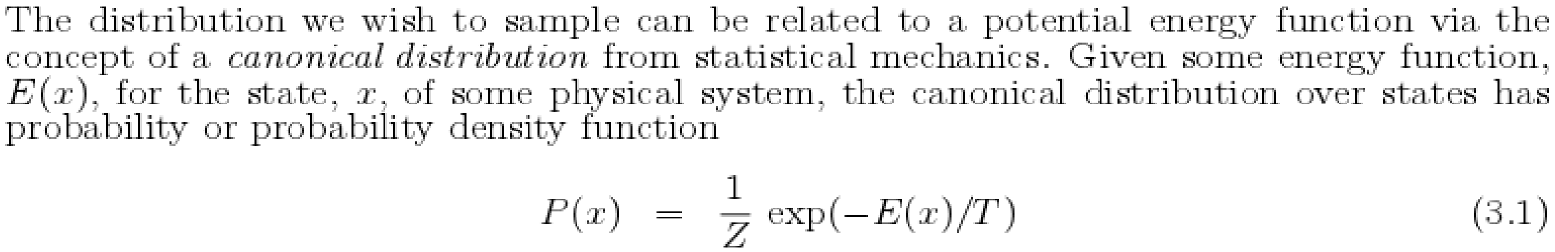
Gibbs sampling, in its basic incarnation, is a special case of the Metropolis–Hastings algorithm. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Suppose we want to obtain k samples of $\mathbf{X} = (x_1, \ldots, x_n)$ from a joint distribution $p(x_1, \ldots, x_n)$. Denote the ith sample by $\mathbf{X}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$. We proceed as follows:

1. We begin with some initial value $\mathbf{X}^{(0)}$.
2. To get the next sample (call it the $i+1$th sample for generality) we sample each component variable $x_j^{(i+1)}$ from the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled. This requires updating each of the component variables in turn. If we are up to the jth component we update it according to the distribution specified by $p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$. Note that we use the value that the $j+1$th component had in the ith sample not the $i+1$th.
3. Repeat the above step k times.

If such sampling is performed, these important facts hold:

- The samples approximate the joint distribution of all variables.
- The marginal distribution of any subset of variables can be approximated by simply considering the samples for that subset of variables, ignoring the rest.
- The expected value of any variable can be approximated by averaging over all the samples.

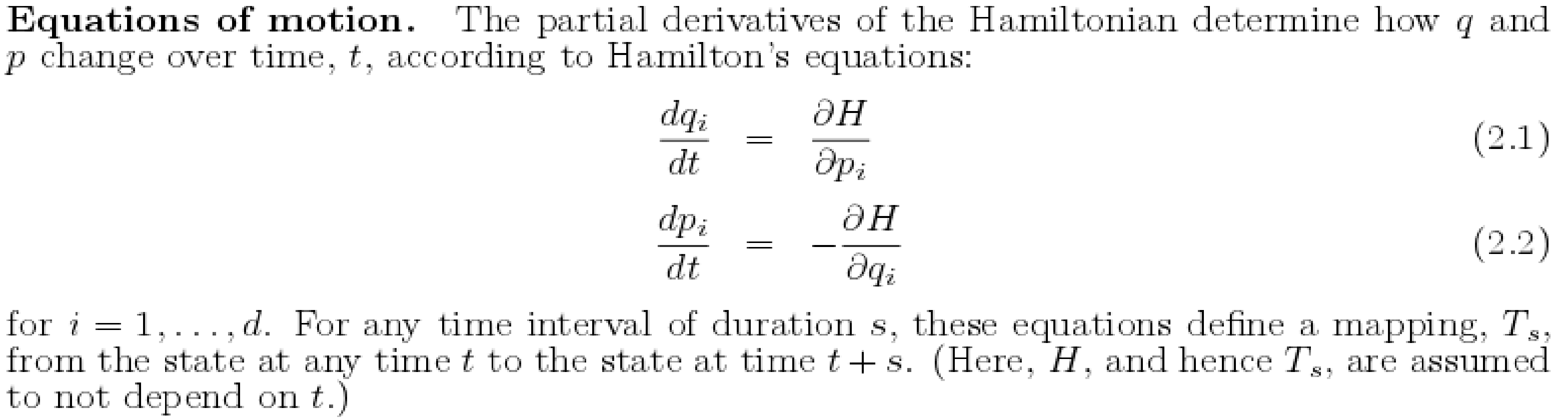
Hamiltonian Monte Carlo



The distribution we wish to sample can be related to a potential energy function via the concept of a *canonical distribution* from statistical mechanics. Given some energy function, $E(x)$, for the state, x, of some physical system, the canonical distribution over states has probability or probability density function

$$P(x) \;=\; \frac{1}{Z}\exp(-E(x)/T) \tag{3.1}$$

R.M. Neal, MCMC using Hamiltonian dynamics



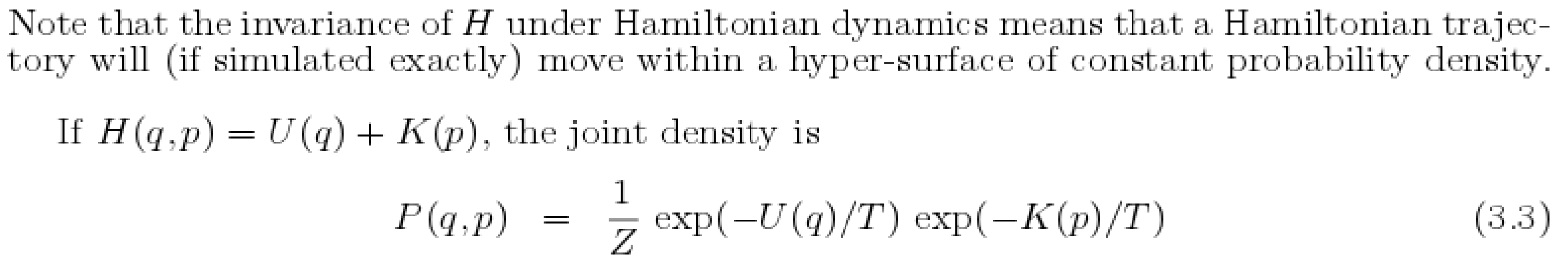
Equations of motion. The partial derivatives of the Hamiltonian determine how q and p change over time, t, according to Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{2.1}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{2.2}$$

for $i = 1, \ldots, d$. For any time interval of duration s, these equations define a mapping, T_s, from the state at any time t to the state at time $t + s$. (Here, H, and hence T_s, are assumed to not depend on t.)

R.M. Neal, MCMC using Hamiltonian dynamics

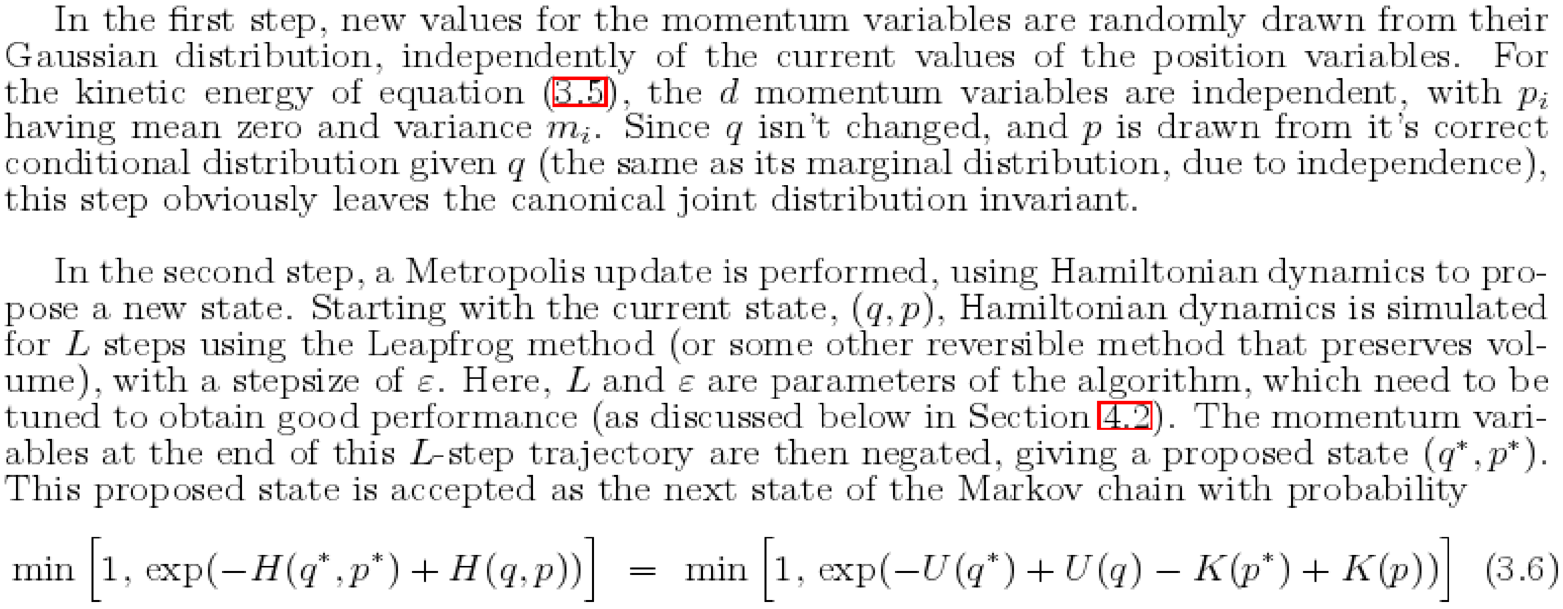


Note that the invariance of H under Hamiltonian dynamics means that a Hamiltonian trajectory will (if simulated exactly) move within a hyper-surface of constant probability density.

If $H(q,p) = U(q) + K(p)$, the joint density is

$$P(q,p) \;=\; \frac{1}{Z}\exp(-U(q)/T)\,\exp(-K(p)/T) \tag{3.3}$$

R.M. Neal, MCMC using Hamiltonian dynamics



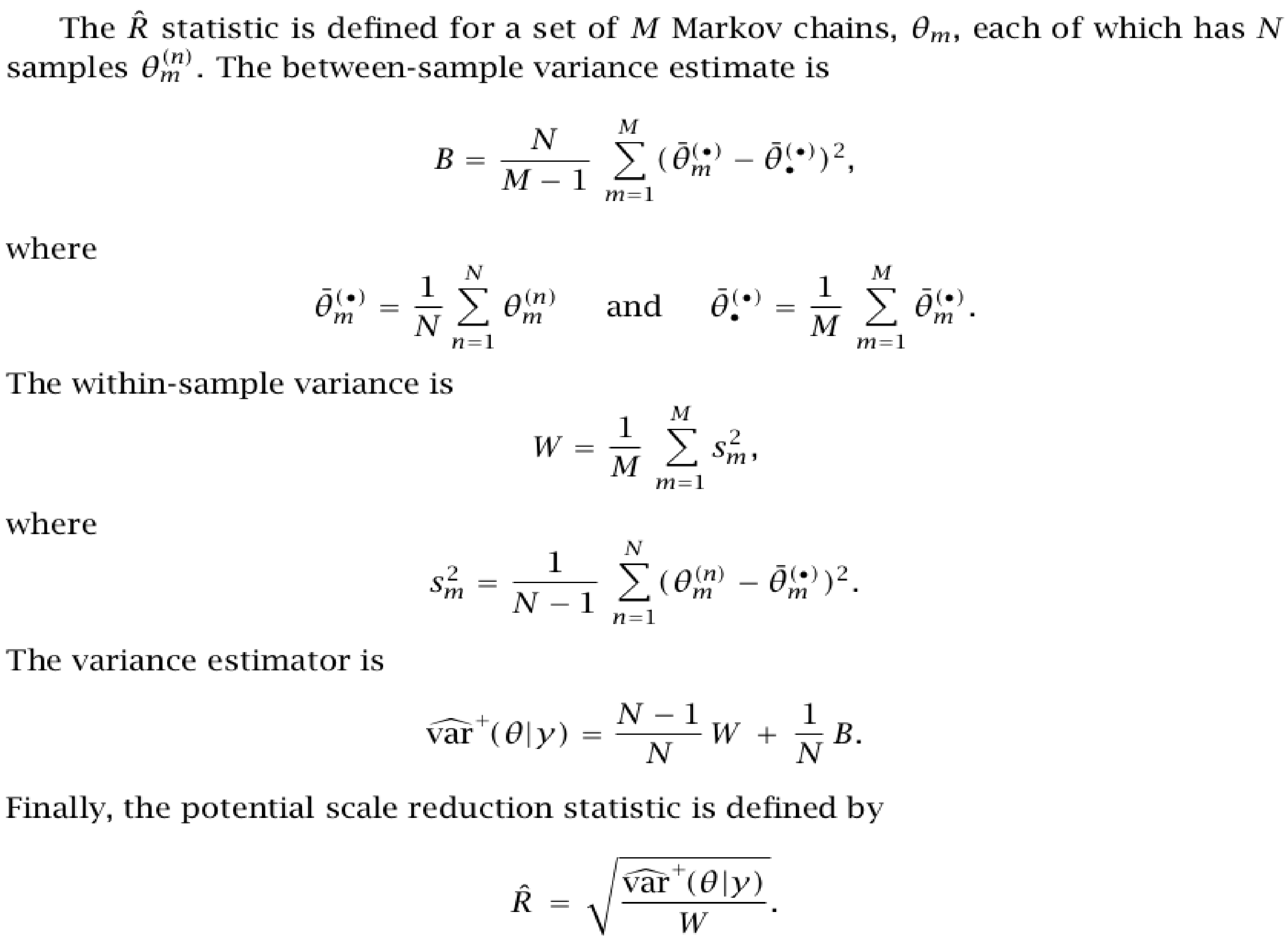
In the first step, new values for the momentum variables are randomly drawn from their Gaussian distribution, independently of the current values of the position variables. For the kinetic energy of equation (3.5), the d momentum variables are independent, with p_i having mean zero and variance m_i. Since q isn't changed, and p is drawn from it's correct conditional distribution given q (the same as its marginal distribution, due to independence), this step obviously leaves the canonical joint distribution invariant.

In the second step, a Metropolis update is performed, using Hamiltonian dynamics to propose a new state. Starting with the current state, (q,p), Hamiltonian dynamics is simulated for L steps using the Leapfrog method (or some other reversible method that preserves volume), with a stepsize of ε. Here, L and ε are parameters of the algorithm, which need to be tuned to obtain good performance (as discussed below in Section 4.2). The momentum variables at the end of this L-step trajectory are then negated, giving a proposed state (q^*,p^*). This proposed state is accepted as the next state of the Markov chain with probability

$$\min\Big[1,\, \exp(-H(q^*,p^*)+H(q,p))\Big] \;=\; \min\Big[1,\, \exp(-U(q^*)+U(q)-K(p^*)+K(p))\Big] \quad (3.6)$$

R.M. Neal, MCMC using Hamiltonian dynamics

Checking for convergence: the R-hat statistic.



The $\hat{R}$ statistic is defined for a set of M Markov chains, θ_m, each of which has N samples $\theta_m^{(n)}$. The between-sample variance estimate is

$$B = \frac{N}{M-1} \sum_{m=1}^{M} (\bar{\theta}_m^{(\bullet)} - \bar{\theta}_{\bullet}^{(\bullet)})^2,$$

where

$$\bar{\theta}_m^{(\bullet)} = \frac{1}{N} \sum_{n=1}^{N} \theta_m^{(n)} \quad \text{and} \quad \bar{\theta}_{\bullet}^{(\bullet)} = \frac{1}{M} \sum_{m=1}^{M} \bar{\theta}_m^{(\bullet)}.$$

The within-sample variance is

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2,$$

where

$$s_m^2 = \frac{1}{N-1} \sum_{n=1}^{N} (\theta_m^{(n)} - \bar{\theta}_m^{(\bullet)})^2.$$

The variance estimator is

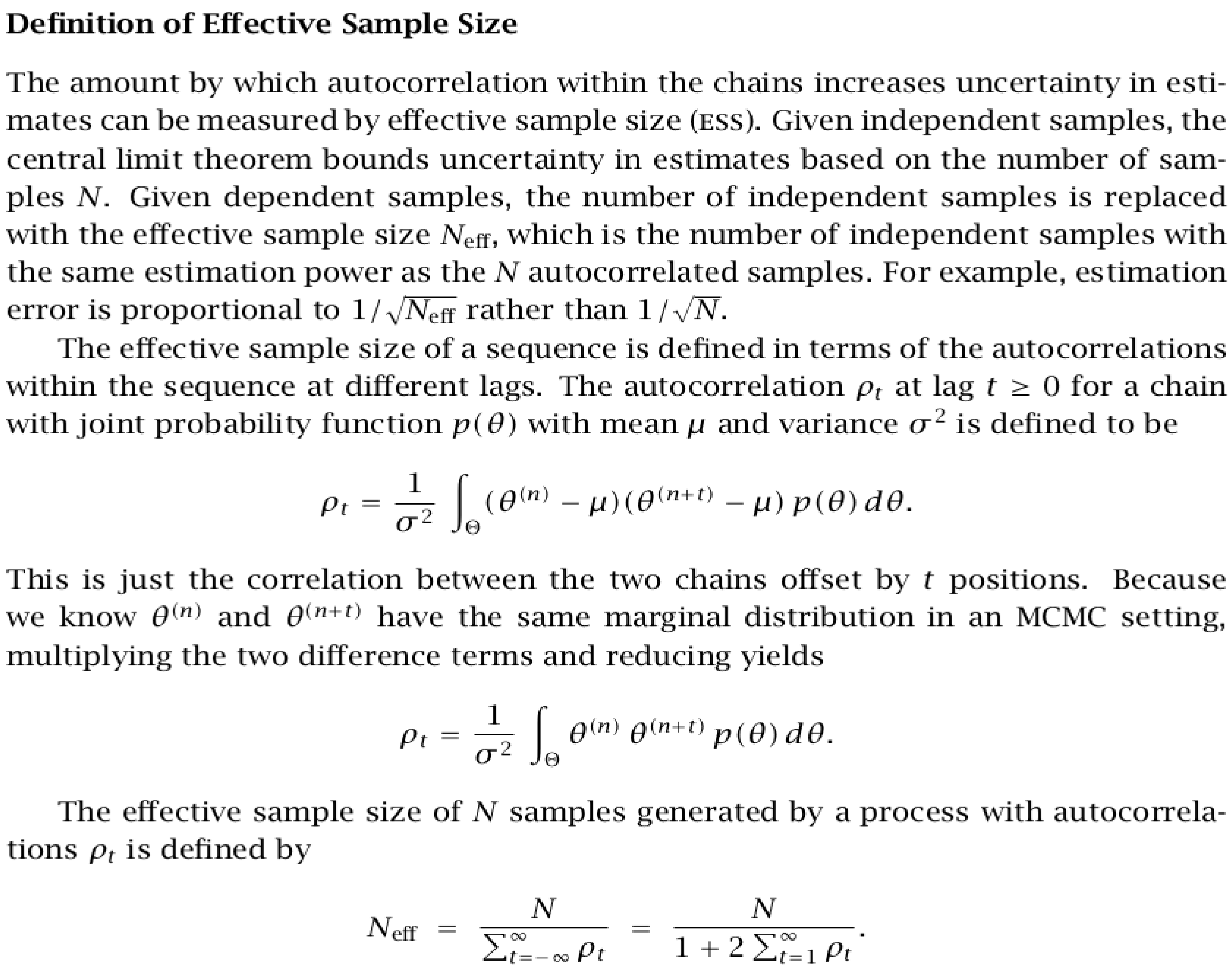
$$\widehat{\text{var}}^{+}(\theta|y) = \frac{N-1}{N} W + \frac{1}{N} B.$$

Finally, the potential scale reduction statistic is defined by

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\theta|y)}{W}}.$$

The stan reference manual and user's guide.

Autocorrelation of samples and effective sample size



Definition of Effective Sample Size

The amount by which autocorrelation within the chains increases uncertainty in estimates can be measured by effective sample size (ESS). Given independent samples, the central limit theorem bounds uncertainty in estimates based on the number of samples N. Given dependent samples, the number of independent samples is replaced with the effective sample size N_{eff}, which is the number of independent samples with the same estimation power as the N autocorrelated samples. For example, estimation error is proportional to $1/\sqrt{N_{\text{eff}}}$ rather than $1/\sqrt{N}$.

The effective sample size of a sequence is defined in terms of the autocorrelations within the sequence at different lags. The autocorrelation ρ_t at lag $t \geq 0$ for a chain with joint probability function $p(\theta)$ with mean μ and variance σ^2 is defined to be

$$\rho_t = \frac{1}{\sigma^2} \int_{\Theta} (\theta^{(n)} - \mu)(\theta^{(n+t)} - \mu)\, p(\theta)\, d\theta.$$

This is just the correlation between the two chains offset by t positions. Because we know $\theta^{(n)}$ and $\theta^{(n+t)}$ have the same marginal distribution in an MCMC setting, multiplying the two difference terms and reducing yields

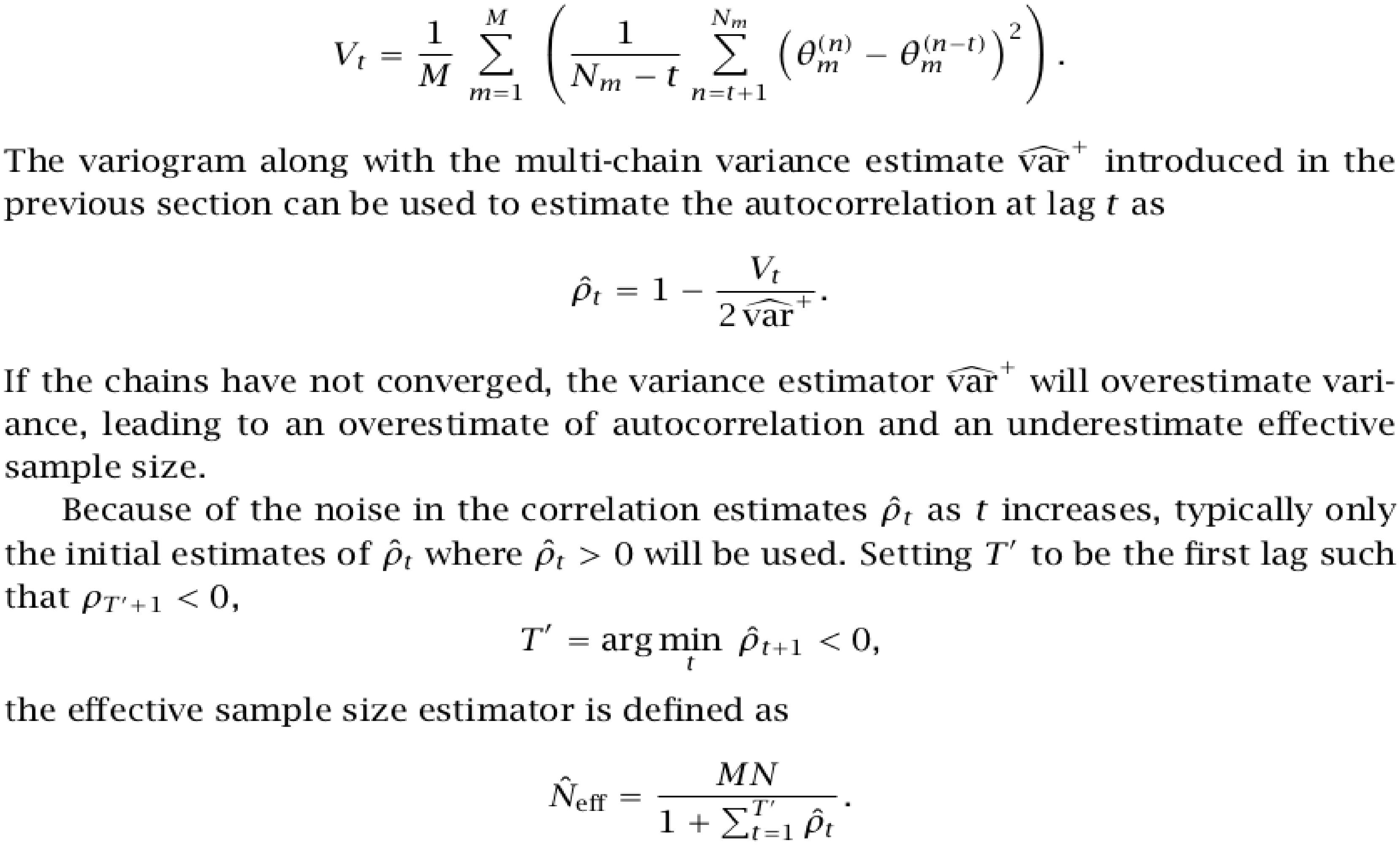
$$\rho_t = \frac{1}{\sigma^2} \int_{\Theta} \theta^{(n)}\, \theta^{(n+t)}\, p(\theta)\, d\theta.$$

The effective sample size of N samples generated by a process with autocorrelations ρ_t is defined by

$$N_{\text{eff}} = \frac{N}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}.$$

The stan reference manual and user's guide.

Autocorrelation and effective sample sizes: the practical approach



$$V_t = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{N_m - t} \sum_{n=t+1}^{N_m} \left(\theta_m^{(n)} - \theta_m^{(n-t)} \right)^2 \right).$$

The variogram along with the multi-chain variance estimate $\widehat{\text{var}}^+$ introduced in the previous section can be used to estimate the autocorrelation at lag t as

$$\hat{\rho}_t = 1 - \frac{V_t}{2\,\widehat{\text{var}}^+}.$$

If the chains have not converged, the variance estimator $\widehat{\text{var}}^+$ will overestimate variance, leading to an overestimate of autocorrelation and an underestimate effective sample size.

Because of the noise in the correlation estimates $\hat{\rho}_t$ as t increases, typically only the initial estimates of $\hat{\rho}_t$ where $\hat{\rho}_t > 0$ will be used. Setting T' to be the first lag such that $\rho_{T'+1} < 0$,

$$T' = \arg\min_t \; \hat{\rho}_{t+1} < 0,$$

the effective sample size estimator is defined as

$$\hat{N}_{\text{eff}} = \frac{MN}{1 + \sum_{t=1}^{T'} \hat{\rho}_t}.$$

The stan reference manual and user's guide.