Bayes factors for research workers

Ly, A.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Abstract

The marginal likelihood plays an important role in many areas of Bayesian statistics such as parameter estimation, model comparison, and model averaging. In most applications, however, the marginal likelihood is not analytically tractable and must be approximated using numerical methods. Here we provide a tutorial on bridge sampling (Bennett, 1976; Meng and Wong, 1996), a reliable and relatively straightforward sampling method that allows researchers to estimate the marginal likelihood for models of varying complexity. First, we introduce bridge sampling and three related sampling methods using the beta-binomial model as a running example. We then apply bridge sampling to estimate the marginal likelihood for the Expectancy Valence (EV) model, a popular model for reinforcement learning. Our results indicate that bridge sampling provides accurate estimates for both a single participant and a hierarchical version of the EV model. We conclude that bridge sampling is an attractive method for mathematical psychologists who typically aim to approximate the marginal likelihood for a limited set of possibly high-dimensional models.

Keywords: Bayes factor, hierarchical model, marginal likelihood, normalising constant, predictive accuracy, reinforcement learning.

12.1 Introduction

Bayesian statistics has become increasingly popular in mathematical psychology (Andrews and Baguley, 2013; Bayarri et al., 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen et al., 2015; Wetzels et al., 2016). The Bayesian approach is conceptually...
simple, theoretically coherent, and easily applied to relatively complex problems. These problems include hierarchical modelling (Matzke et al., 2015a; Matzke and Wagenmakers, 2009; Rouder and Lu, 2005; Rouder et al., 2005, 2007) or the comparison of non-nested models (Lee, 2008; Pitt et al., 2002; Shiffrin et al., 2008). Three major applications of Bayesian statistics concern parameter estimation, model comparison, and Bayesian model averaging. In all three areas, the marginal likelihood—that is, the probability of the observed data given the model of interest—plays a central role (see also Gelman and Meng, 1998).

First, in parameter estimation, we consider a single model and aim to quantify the uncertainty for a parameter of interest \( \theta \) after having observed the data \( d \). This is realised by means of a posterior distribution that can be obtained using Bayes theorem, as

\[
\pi(\theta | d) = \frac{f(d | \theta) \pi(\theta)}{\int f(d | \theta) \pi(\theta) d\theta} = \frac{p(d | \theta) \pi(\theta)}{p(d)}.
\] (12.1.1)

Here, the marginal likelihood of the data \( p(d) \) ensures that the posterior distribution \( \pi(\theta | d) \) is a proper probability density function (pdf) in the sense that it integrates to one. This illustrates why in parameter estimation the marginal likelihood is referred to as a normalising constant.

Second, in model comparison, we consider \( m \in \mathbb{N} \) number of competing models, and are interested in the relative plausibility of a particular model \( M_i \), where \( i = 1, 2, \ldots, m \), given the prior model probability and the evidence from the data \( d \) (see three special issues on this topic in the Journal of Mathematical Psychology: Mulder and Wagenmakers, 2016; Myung et al., 2000b; Wagenmakers and Waldorp, 2006a). This relative plausibility is quantified by the so-called posterior model probability \( P(M_i | d) \) of model \( M_i \) given the data \( d \) (Berger and Molina, 2005)

\[
P(M_i | d) = \frac{p(d | M_i) P(M_i)}{\sum_{j=1}^{m} p(d | M_j) P(M_j)},
\] (12.1.2)

where \( p(d | M_j) \) denotes the marginal likelihood of model \( M_j \), and where the denominator is the sum of the marginal likelihood times the prior model probability of all \( m \) models. In model comparison, the marginal likelihood for a specific model is also referred to as the model evidence (Didelot et al., 2011), the integrated likelihood (Kass and Raftery, 1995), and the predictive likelihood of the model (Gamerman and Lopes, 2006, Chapter 7). As a function of the data it is also known as the predictive probability of the data (Kass and Raftery, 1995), or the prior predictive density (Ntzoufras, 2009). Note that computationally the marginal likelihood of Eq. (12.1.2) is the same as the marginal likelihood of Eq. (12.1.1). However, for the latter equation we dropped the model index because in parameter estimation we consider only one model.

If only two models \( M_1 \) and \( M_2 \) are considered, Eq. (12.1.2) can be used to quantify the relative posterior model plausibility of model \( M_1 \) compared to model \( M_2 \). This relative plausibility is given by the ratio of the posterior probabilities of both models, and is referred to as the posterior model odds.
12.1. Introduction

\[
\begin{align*}
\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_2 | d)} &= \frac{P(d | \mathcal{M}_1)}{P(d | \mathcal{M}_2)} \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}.
\end{align*}
\]

Eq. (12.1.3) illustrates that the posterior model odds are the product of two factors: The right most factor is the ratio of the prior probabilities of the models also known as the prior model odds. The factor in the middle is the ratio of the marginal likelihoods—the so-called Bayes factor (Etz and Wagenmakers, 2017; Jeffreys, 1961; Ly et al., 2016a, 2016b; Robert, 2016). The Bayes factor plays an important role in model comparison and is referred to as the “standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis and Raftery, 1997, p. 648) and “the primary tool used in Bayesian inference for hypothesis testing and model selection” (Berger, 2006, p. 378).

Third, the marginal likelihood plays an important role in Bayesian model averaging (BMA; Hoeting et al., 1999) where aspects of parameter estimation and model comparison are combined. As in model comparison, BMA considers several models; however, it does not aim to identify a single best model. Instead, it fully acknowledges model uncertainty. Model averaged parameter inference can be obtained by combining, across all models, the posterior distribution of the parameter of interest weighted by each model’s posterior model probability, and as such depends on the marginal likelihood of the models. This procedure assumes that the parameter of interest has identical interpretation across the different models. Model averaged predictions can be obtained in a similar manner.

A problem that arises in all three areas—parameter estimation, model comparison, and BMA—is that an analytical expression of the marginal likelihood can be obtained only for certain restricted examples. This is a pressing problem in Bayesian modelling, and in particular in mathematical psychology where models can be non-linear and equipped with a large number of parameters, especially when the models are implemented in a hierarchical framework. Such a framework incorporates both commonalities and differences between participants of one group by assuming that the model parameters of each participant are drawn from a group-level distribution (for advantages of the Bayesian hierarchical framework see Ahn et al., 2011; Navarro et al., 2006; Rouder and Lu, 2005; Rouder et al., 2005, 2008; Scheibehenne and Pachur, 2015; Shiffrin et al., 2008; Wetzels et al., 2010b). For instance, consider a four-parameter Bayesian hierarchical model with four group-level distributions each characterised by two parameters and a group size of 30 participants; this then results in $30 \times 4$ individual-level parameters and $2 \times 4$ group-level parameters for a total of 128 parameters. In sum, even simple models quickly become complex once hierarchical aspects are introduced and this frustrates the derivation of the marginal likelihood.

To overcome this problem, several Monte Carlo sampling methods have been proposed to approximate the marginal likelihood. In this tutorial we focus on four such methods: the bridge sampling estimator (Bennett, 1976; Chapter 5 of Chen et al., 2012; Meng and Wong, 1996), and its three commonly used special cases—the naive Monte Carlo estimator, the importance sampling estimator, and
the generalised harmonic mean estimator (for alternative methods see Gamerman and Lopes, 2006, Chapter 7; and for alternative approximation methods relevant to model comparison and BMA, see Carlin and Chib (1995) and Green (1995). As we will illustrate throughout this tutorial, the bridge sampler is accurate, efficient, and relatively straightforward to implement (e.g., DiCiccio et al., 1997; Frühwirth-Schnatter, 2004; Meng and Wong, 1996).

The goal of this tutorial is to bring the bridge sampling estimator to the attention of mathematical psychologists. We aim to explain this estimator and facilitate its application by suggesting a step-by-step implementation scheme. To this end, we first show how bridge sampling and the three special cases can be used to approximate the marginal likelihood in a simple beta-binomial model. We begin with the naive Monte Carlo estimator and progressively work our way up –via the importance sampling estimator and the generalised harmonic mean estimator– to the most general case considered: the bridge sampling estimator. This order was chosen such that key concepts are introduced gradually and estimators are of increasing complexity and sophistication. The first three estimators are included in this tutorial with the sole purpose of facilitating the reader’s understanding of bridge sampling. In the second part of this tutorial, we outline how the bridge sampling estimator can be used to derive the marginal likelihood for the Expectancy Valence (EV; Busemeyer and Stout, 2002) model, a popular, yet relatively complex reinforcement-learning model for the Iowa gambling task (Bechara et al., 1994). We apply bridge sampling to both an individual-level and a hierarchical implementation of the EV model.

Throughout the chapter, we use the software package R to implement the bridge sampling estimator for the various models. The interested reader is invited to reproduce our results by downloading the code and all relevant materials from our Open Science Framework folder at osf.io/f9cq4.

12.2 Four sampling methods to approximate the marginal likelihood

In this section we outline four standard methods to approximate the marginal likelihood. For more detailed explanations and derivations, we recommend Ntzoufras (2009, Chapter 11) and Gamerman and Lopes (2006, Chapter 7); a comparative review of the different sampling methods is presented in DiCiccio et al. (1997).

For concreteness let \( Y \) represent the number of correct responses given by a participant in \( n \) test items. We assume that \( Y \) follows a binomial distribution given by

\[
f(d | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y},
\]

where \( d \) refers to the number of successes \( y \) and \( n \) the number of trials, thus, \( y = 0, 1, \ldots, n \) and \( n \in \mathbb{N} \). Data are assumed to be known. For instance, suppose

\(^1\)The appendix gives a derivation showing that the first three estimators are indeed special cases of the bridge sampler.
our participant answered \( y = 2 \) items correctly in \( n = 10 \) trials and plugging these observations into Eq. (12.2.1) yields a function of \( \theta \), i.e.,

\[
f(d \mid \theta) = \binom{10}{2} \theta^2 (1 - \theta)^8.
\]

(12.2.2)

Such a function is in general known as a likelihood function. The parameter \( \theta \in (0, 1) \) can be thought of as the participant’s latent ability, which is unknown. To learn \( \theta \) we assign it a so-called prior distribution \( \pi(\theta) \). The prior can be thought of as our knowledge about the participant’s ability before we observe the data. For the running example it is computationally convenient to choose a so-called beta distribution for \( \theta \) with \( \alpha, \beta > 0 \), that is,

\[
\pi(\theta) = \text{Beta}(\theta; \alpha, \beta) = \frac{1}{\mathcal{B}(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1},
\]

(12.2.3)

where \( \mathcal{B}(\alpha, \beta) \) is the beta function defined as \( \mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \), and where \( \Gamma(n) = (n-1)! \) whenever \( n \in \mathbb{N} \). To ease the exposition, we set \( \alpha = \beta = 1 \). This choice corresponds to the uniform prior on \( \theta \), which is depicted as the dotted line in Fig. 12.1. The uniform prior on \( \theta \) is interpreted as each value of \( \theta \) being equally probable.\(^2\)

Using Bayes rule we can update our prior knowledge about the participant’s latent ability \( \theta \) into a posterior as

\[
\pi(\theta \mid d) = \frac{f(d \mid \theta)\pi(\theta)}{p(d)},
\]

(12.2.4)

where the marginal likelihood \( p(d) \) is defined as

\[
\int_{\text{likelihood prior}} f(d \mid \theta) \pi(\theta) d\theta.
\]

(12.2.5)

The marginal likelihood makes the posterior for \( \theta \) a proper probability function so that it integrates to one, which is why \( p(d) \) is also referred to as the normalising constant of the posterior. In general, we cannot perform this integral analytically and have to resort to numerical methods such as the bridge sampler, instead.

The running example, however, is chosen in such a way that both the posterior and the target of estimation \( p(d) = \int f(d \mid \theta)\pi(\theta)d\theta \) can be calculated explicitly. Filling in the binomial likelihood and the beta prior, we see that for the running example the posterior is proportional to

\[
\pi(\theta \mid d) \propto \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1}.
\]

(12.2.6)

Note that this expression is of the same form as the beta distribution given in Eq. (12.2.3). Consequently, the posterior for \( \theta \) is also a beta distribution, namely

\[
\pi(\theta \mid d) = \text{Beta}(\theta \mid y + \alpha, n - y + \beta) = \frac{\binom{n}{y} \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1}}{p(d)},
\]

(12.2.7)

\(^2\)But see Ly et al. (2017c).
Figure 12.1: Prior and posterior distribution for the rate parameter $\theta$ from the beta-binomial model. The Beta($\theta$; 1, 1) prior on the parameter $\theta$ is represented by the dotted line; the Beta($\theta$; 3, 9) posterior distribution is represented by the solid line and was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials.

The denominator is the marginal likelihood and given by

$$p(d) = \int_0^1 f(d | \theta) \pi(\theta) \, d\theta = \int_0^1 \binom{n}{y} \theta^{y+\alpha-1}(1 - \theta)^{n-y+\beta-1} \, d\theta$$  \hspace{1cm} (12.2.8)

$$= \binom{n}{y} B(y + \alpha, n - y + \beta).$$  \hspace{1cm} (12.2.9)

For $\alpha = \beta = 1$ and the observations $y = 2$ out of $n = 10$, we get

$$p(d) = \binom{10}{2} B(3, 9) = \frac{1}{11} \approx 0.0909,$$  \hspace{1cm} (12.2.10)

the target of the four estimation methods.

### 12.2.1 Method 1: The naive Monte Carlo estimator of the marginal likelihood

The simplest method to approximate the marginal likelihood is provided by the naive Monte Carlo estimator (Hammersley and Handscomb, 1964; Raftery and
Banfield, 1991). This method uses the standard definition of the marginal likelihood, Eq. (12.2.5), and relies on viewing integrals as sums. The integral implies that the marginal likelihood \( p(d) \) is a weighted average of the likelihood where the weights correspond to the prior distribution for the parameters. In other words, the marginal likelihood is the expected value of the likelihood with respect to the prior, that is,

\[
p(d) = \mathbb{E}_{\pi}[f(d | \theta)]. \tag{12.2.11}
\]

To estimate this population mean, we use a sample mean by sampling, say, \( K \) samples from the prior and subsequently average the values of the integrand, the likelihood, at these \( K \) samples. This yields the naive Monte Carlo estimator \( \hat{p}_1(d) \)

\[
\hat{p}_1(d) = \frac{1}{K} \sum_{i=1}^{K} f(d | \tilde{\theta}_i), \quad \tilde{\theta}_i \sim \pi(\theta) \tag{12.2.12}
\]

for the target \( p(d) \).

### 12.2.1.1 Running example

To obtain the naive Monte Carlo estimate of the marginal likelihood in our running example, we need \( K \) samples from the Beta(\( \theta; 1, 1 \)) prior distribution for \( \theta \). For illustrative purposes, we limit the number of samples to \( K = 12 \) whereas in practice one should take \( K \) to be very large. To do so in R, we use the command

\[
\text{priorSamples <- rbeta(n=12, shape=1, shape=1)}
\]

and we obtained the following samples

\[
\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\} = \{0.58, 0.76, 0.03, 0.93, 0.27, 0.97, 0.45, 0.46, 0.18, 0.64, 0.06, 0.15\}.
\]

We use the tilde to make explicit that these values for \( \theta \) are sampled. All sampled values are represented by the grey dots in Fig. 12.2. Following Eq. (12.2.12), the next step is to evaluate the likelihood, Eq. (12.2.1), at each sampled value \( \tilde{\theta}_i \), weight these values by \( 1/K \), and sum them to obtain the average likelihood \( \hat{p}_1(d) \), thus,

\[
\hat{p}_1(d) = \frac{1}{12} \sum_{i=1}^{12} f(d | \tilde{\theta}_i) = \frac{1}{12} \sum_{i=1}^{12} \binom{n}{y} \tilde{\theta}_i^n (1 - \tilde{\theta}_i)^{n-y}, \tag{12.2.13}
\]

\[
= \frac{1}{12} \binom{10}{2} (0.58^2 (1 - 0.58)^8 + \ldots + 0.15^2 (1 - 0.15)^8), \tag{12.2.14}
\]

\[
= 0.0945, \tag{12.2.15}
\]

where in the second line we filled in \( y = 2 \) and \( n = 10 \). To evaluate the likelihood for the first posterior sample \( \tilde{\theta}_1 = 0.58 \) we use the command \( \text{dbinom(x=2, size=10, prob=0.58)} \), while the estimate \( \hat{p}_1(d) \) is obtained from the command \( 1/12* \text{sum(dbinom(x=2, size=10, prob=priorSamples))} \) in R.
12. A Tutorial on Bridge Sampling

12.2.2 Method 2: The importance sampling estimator of the marginal likelihood

The naive Monte Carlo estimator introduced in the last section performs well if the prior and posterior distribution have a similar shape and strong overlap. However, the estimator is unstable if the posterior distribution is peaked relative to the prior (Gamerman and Lopes, 2006; Ntzoufras, 2009). In such a situation, most of the sampled values for $\theta$, say, 98 out of $K = 100$, result in likelihood values close to zero and contribute only minimally to the estimate. This means that those few samples that result in high likelihood values, say, 2 out of $K = 100$, dominate the estimate of the marginal likelihood, which in effect results in high variance of the estimator (Newton and Raftery, 1994; Pajor, 2016).\footnote{The interested reader is referred to Pajor (2016) for a recent improvement on the calculation of the naive Monte Carlo estimator. The proposed improvement involves trimming the prior distribution in such a way that regions with low likelihood values are eliminated, thereby increasing the accuracy and efficiency of the estimator.} The importance sampling estimator, on the other hand, overcomes this shortcoming by boosting sampled values in regions of the parameter space where the integrand of Eq. (12.2.5) is large. This

---

Figure 12.2: Illustration of the naive Monte Carlo estimator for the beta-binomial example. The dotted line represents the prior distribution and the solid line represents the posterior distribution that was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials. The grey dots represent the $K = 12$ samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\}$ randomly drawn from the Beta($\theta$; 1, 1) prior.
is realised by using samples from a so-called importance density \( g_{IS}(\theta) \) instead of the prior distribution. The advantage of sampling from an importance density is that values for \( \theta \) that result in high likelihood values are sampled most frequently, whereas values for \( \theta \) with low likelihood values are sampled only rarely.

To derive the importance sampling estimator, the definition of the marginal likelihood, Eq. (12.2.5), is once again used as the starting point. The trick is to multiply and divide by the importance density \( g_{IS}(\theta) \) as follows

\[
\begin{align*}
p(d) &= \int f(d | \theta) \pi(\theta) d\theta = \int f(d | \theta) \pi(\theta) \frac{g_{IS}(\theta)}{g_{IS}(\theta)} d\theta = \int \frac{f(d | \theta) \pi(\theta)}{g_{IS}(\theta)} g_{IS}(\theta) d\theta, \\
&= \mathbb{E}_{g_{IS}(\theta)} \left[ \frac{f(d | \theta) \pi(\theta)}{g_{IS}(\theta)} \right].
\end{align*}
\] (12.2.16)

In other words, the marginal likelihood is the expected value of the ratio \( \frac{f(d | \theta) \pi(\theta)}{g_{IS}(\theta)} \) with respect to the importance density \( g_{IS}(\theta) \). To estimate the population mean \( \mathbb{E}_{g_{IS}(\theta)} \left[ \frac{f(d | \theta) \pi(\theta)}{g_{IS}(\theta)} \right] \), we use a sample mean by sampling \( K \) samples from \( g_{IS}(\theta) \) and subsequently average the values of the integrand at these \( K \) samples. This yields the importance estimator \( \hat{p}_2(d) \)

\[
\hat{p}_2(d) = \frac{1}{K} \sum_{i=1}^{K} \frac{f(d | \tilde{\theta}_i) \pi(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)} , \quad \tilde{\theta}_i \sim g_{IS}(\theta).
\] (12.2.17)

Choosing a suitable importance density is crucial and should (1) be easy to evaluate; (2) have the same domain as the posterior distribution; (3) closely resemble the posterior distribution, and (4) have fatter tails than the posterior distribution (Neal, 2001; Vandekerckhove et al., 2015). The latter criterion ensures that values in the tails of the distribution cannot misleadingly dominate the estimate (Neal, 2001).\(^4\)

### 12.2.2.1 Running example

To obtain the importance sampling estimate of the marginal likelihood for our running example, we first need to choose an importance density \( g_{IS}(\theta) \). An importance density that fulfils the four above mentioned desiderata is a mixture between (i) a first rough approximation of the posterior based on the posterior samples, and (ii) a uniform density across the range of \( \theta \) (Vandekerckhove et al., 2015).

\(^4\)To illustrate the need for an importance density with fatter tails than the posterior distribution, imagine you sample from the tail region of an importance density with thinner tails. In this case, the numerator in Eq. (12.2.17) would be substantially larger than the denominator resulting in a very large ratio. Since this specific ratio is only one component of the sum displayed in Eq. (12.2.17), this component would dominate the importance sampling estimate. Hence, thinner tails of the importance density run the risk of producing unstable estimates across repeated computations. In fact, the estimator may have infinite variance (e.g., Ionides, 2008; Owen and Zhou, 2000).
The rough approximation of the posterior leads to an importance density that closely resemble the posterior distribution, while the uniform distribution secures that the importance density has the same range as posterior. By mixing these two distributions we get an importance sampler that has thick enough tails. For the rough approximation we use a beta distribution, because we can sample from it easily.

The relative impact of the uniform density is quantified by a mixture weight \( \gamma \) that ranges between 0 and 1. The larger \( \gamma \), the higher the influence of the uniform density resulting in a less peaked distribution with thick tails. If \( \gamma = 1 \), the importance density simplifies to the uniform distribution on \([0, 1]\), and if \( \gamma = 0 \) the importance density simplifies to the rough first approximation to the posterior distribution.

First, we have to sample \( K = 12 \) samples from the posterior distribution. Posterior samples can be obtained without knowing the normalising constant \( p(d) \) using so-called Markov chain Monte Carlo (MCMC) methods that are made accessible through software packages such as WinBUGS, JAGS and Stan. These MCMC methods exploit the fact that the posterior is known up to a constant whenever the observations are given, the likelihood is chosen, and a prior is specified, see for instance Ntzoufras (2009) and Robert (2015) for an introduction.

For the running example, however, we do not need WinBUGS, JAGS or Stan, as we can generate posterior samples directly in R. Recall that for the data at hand, the posterior distribution is proportional to the beta distribution Beta(\( \alpha \); 3, 9) and to obtain, say, \( K = 12 \) posterior samples we use the R command \texttt{rbeta(n=12, shape1=3, shape1=9)}, which for us resulted in

\[
\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{12}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.
\]

We use \( \hat{\theta}_i \) to refer to the \( i \)th sample from the posterior distribution to distinguish it from the previously used \( \theta_i \)—the \( i \)th sample from a distribution other than the posterior distribution, such as a prior distribution or an importance density.

Second, as a first approximation to the posterior we use a beta distribution fitted to these posterior samples using the methods of moments. Recall that a beta distributed random variable \( X \sim \text{Beta}(\alpha, \beta) \) has a mean of \( \mathbb{E}(X) = \alpha / (\alpha + \beta) \) and a variance of \( V(X) = \alpha \beta / [(\alpha + \beta)^2(\alpha + \beta + 1)] \). Filling in the mean \( \hat{\theta} = 0.232 \) and variance of \( s^2_{\hat{\theta}} = 0.014 \) of our posterior sample \( \{\hat{\theta}_1, \ldots, \hat{\theta}_{12}\} \) and solving for \( \alpha \) and \( \beta \), we retrieve the parameters

\[
\hat{\alpha} = \hat{\theta} \left( \frac{\hat{\theta}(1 - \hat{\theta})}{s^2_{\hat{\theta}}} - 1 \right) = 0.232 \left( \frac{0.232(1 - 0.232)}{0.0142} - 1 \right) = 2.673,
\]

\[
\hat{\beta} = (1 - \hat{\theta}) \left( \frac{\hat{\theta}(1 - \hat{\theta})}{s^2_{\hat{\theta}}} - 1 \right) = (1 - 0.232) \left( \frac{0.232(1 - 0.232)}{0.0142} - 1 \right) = 8.865.
\]

\[\text{In our running example, the importance sampling estimator then reduces to the naive Monte Carlo estimator.}\]
As such, we use the beta distribution $B(\theta; 2.673, 8.865)$ as the first component in the importance sampler.

Third, we choose a mixture weight. With a mixture weight of $\gamma = 0.30$ on the uniform component—a choice that was made to ensure that, visually, the tails of the importance density are clearly thicker than the tails of the posterior distribution—we obtain the following importance density

$$\gamma B(\theta; 1, 1) + (1 - \gamma) B(\theta; \hat{\alpha}, \hat{\beta}) = .3 + .7 \, B(\theta; 2.673, 8.865).$$  \hfill (12.2.18)

Note that the importance distribution is a mixture of the prior and the fitted beta distribution, both from which we can easily draw samples from. This importance density is represented by the dashed line in Fig. 12.3. The figure also shows the posterior distribution (solid line). As is evident from the figure, the beta mixture importance density resembles the posterior distribution, but has fatter tails.

In general, it is advised to choose the mixture weight on the uniform component $\gamma$ small enough to make the estimator efficient, yet large enough to produce fat tails to stabilise the estimator. A suitable mixture weight can be realised by gradually decreasing the mixture weight and investigating whether stability is still guaranteed (i.e., robustness analysis).

Fourth, to draw one sample from the importance density we first draw a dummy variable $Z$ that takes on the value 1 with 30% chance and 0 otherwise. If $Z = 1$ we draw from the uniform distribution, otherwise we draw from the fitted beta distribution. For $K = 12$ the R code simplifies to

$$K <- 12$$
$$\text{numFittedBeta} <- \text{rbinom}(n=1, \text{size}=K, \text{prob}=0.3)$$
$$\text{postSamples} <- \text{c}(\text{rbeta}(n=\text{numFittedBeta}, \text{shape1}=2.673, \text{shape2}=8.865), \text{rbeta}(n=K-\text{numFittedBeta}, \text{shape1}=1, \text{shape2}=1))$$

which for us resulted in

$$\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\} = \{0.11, 0.07, 0.33, 0.25, 0.41, 0.39, 0.25, 0.13, 0.64, 0.26, 0.74, 0.92\}.$$

These samples are represented by the grey dots in Fig. 12.3. The final step is to compute the average adjusted likelihood, i.e., $f(d \mid \theta) \pi(\theta) / g_{IS}(\theta)$, at the $K = 12$ samples. This yields the following importance sampling estimate for the marginal likelihood

$$\hat{p}_2(d) = \frac{1}{12} \sum_{i=1}^{12} \frac{f(d \mid \tilde{\theta}_i) \pi(\tilde{\theta}_i)}{.3 + .7 \, B(\tilde{\theta}_i; 2.673, 8.865)}$$

$$= \frac{1}{12} \left( \frac{10}{2} 0.11^2 (1 - 0.11)^8 \times 1 + \cdots + \frac{10}{2} 0.92^2 (1 - 0.92)^8 \times 1 \right)$$

$$= \frac{1}{12} \left( \frac{10}{2} \right) (0.0021 + \cdots + 4.7 \times 10^{-9})$$

$$= 0.0829. \hfill (12.2.19)$$

where the $.3$ in the numerator is multiplied with the density of the prior, which is one for every $\theta$. To evaluate the density of the beta density at the first value
Figure 12.3: Illustration of the importance sampling estimator for the beta-binomial model. The dashed line represents our beta mixture importance density and the solid grey line represents the posterior distribution that was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials. The grey dots represent the $K = 12$ samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\}$ randomly drawn from our beta mixture importance density.

of the importance density, say, $\tilde{\theta}_i = 0.11$ we use the command \texttt{dbeta(x=0.11, shape1=2.673, shape2=8.865)} in R.

12.2.3 Method 3: The generalised harmonic mean estimator of the marginal likelihood

Just as the importance sampling estimator, the generalised harmonic mean estimator focuses on regions of the parameter space where the integrand of Eq. (12.2.5) is large by using an importance density $g_{IS}(\theta)$ (Gelfand and Dey, 1994). However, in contrast to the importance sampling estimator, the generalised harmonic mean estimator requires an importance density with thinner tails for an analogous reason as in importance sampling.

To derive the generalised harmonic mean estimator, also known as reciprocal importance sampling estimator (Frühwirth-Schnatter, 2004), we integrate

\footnote{Note that the generalised harmonic mean estimator is a more stable version of the harmonic mean estimator (Newton and Raftery, 1994). A problem of the harmonic mean estimator is that it is dominated by the samples that have small likelihood values.}
12.2. Four sampling methods to approximate the marginal likelihood

\[ 1/p(d) = \frac{\pi(\theta \mid d)}{f(d \mid \theta) \pi(\theta)} \]

with respect to a proposal density \( g_{IS}(\theta) \) that integrates to one, that is,

\[
\frac{1}{p(d)} = \int \frac{1}{p(d)} g_{IS}(\theta) \, d\theta = \int \frac{\pi(\theta \mid d)}{f(d \mid \theta) \pi(\theta)} \, g_{IS}(\theta) \, d\theta = \int \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \, \pi(\theta \mid d) \, d\theta
\]

\[
= \mathbb{E}_{\text{post}} \left[ \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \right],
\]

which in turn leads to

\[
p(d) = \left( \mathbb{E}_{\text{post}} \left[ \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \right] \right)^{-1}. \tag{12.2.21}
\]

In other words, the reciprocal of the marginal likelihood is the expected value of the ratio \( \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \) with respect to the posterior. To estimate the population mean \( \mathbb{E}_{\text{post}} \left[ \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \right] \), we use a sample mean by sampling \( K \) samples from the posterior and subsequently average the values of the integrand at these \( K \) samples. This yields the generalised harmonic mean estimator \( \hat{p}_3(d) \) (Gelfand and Dey, 1994), where

\[
\hat{p}_3(d) = \left( \frac{1}{K} \sum_{j=1}^{K} \frac{g_{IS}(\theta_j)}{f(d \mid \theta_j) \pi(\theta_j)} \right)^{-1}, \quad \hat{\theta}_j \sim \pi(\theta \mid d) \tag{12.2.22}
\]

Note that the generalised harmonic mean estimator –in contrast to the importance sampling estimator– evaluates samples from the posterior distribution. Consequently, the sum in Eq. (12.2.22) will contain relatively few terms with \( \theta_j \) coming from the tail of the posterior. To avoid having the estimator \( \hat{p}_3(d) \) miss out on the contribution of the integrand for \( \theta \) from the tail, we require that for these values of \( \theta \) that the ratio \( \frac{g_{IS}(\theta)}{f(d \mid \theta) \pi(\theta)} \) itself is already small. This condition implies that \( g_{IS}(\theta) < f(d \mid \theta) \pi(\theta) \propto \pi(\theta \mid d) \) for \( \theta \) in the tail of the posterior.

Thus, an importance density for the generalised harmonic mean estimator should (1) have thinner tails than the posterior distribution (Newton and Raftery, 1994; DiCiccio et al., 1997), (2) be easy to evaluate, (3) have the same domain as the posterior distribution, and (4) closely resemble the posterior distribution.

12.2.3.1 Running example

To obtain a generalised harmonic mean estimate of \( p(d) \), we need to choose a suitable importance density. The trick is to transform the parameters onto the real line and use a normal distribution fitted to the posterior samples as the importance density. First, we draw \( K = 12 \) samples from the posterior distribution. Reusing the samples from the last section, we obtain

\[
\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{12}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.
\]
Second, to fit a normal distribution to the posterior samples, we probit transform the posterior samples $\xi_j = \Phi^{-1}(\theta_j)$ that range over the entire real line. For the first sample, we use the command `qnorm(0.22)` in R. Applying this to our particular posterior samples $\hat{\theta}_1, \ldots, \hat{\theta}_{12}$ yields

$$\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{12}\} = \{-0.77, -0.99, -1.34, -0.39, -1.55, -0.61, -0.64, -0.23, -0.84, -0.18, -0.81, -1.17\}.$$ 

These probit-transformed samples are represented by the grey dots in Fig. 12.4.

Third, we search for the normal distribution that provides the best fit to the probit-transformed posterior samples $\hat{\xi}_j$. Using the method of moments, we obtain $\hat{\mu} = -0.793$ and $\hat{\sigma} = 0.423$. Note that the choice of a normal importance density justifies step 2; the probit transformation (or an equivalent transformation) was required to match the range of the posterior distribution to the one of the normal distribution.

Finally, as importance density we choose a normal distribution with mean $\mu = -0.793$ and standard deviation $\sigma = 0.423/1.5$. The additional division by 1.5 ensures that the importance density has thinner tails than the probit-transformed posterior distribution (for a discussion of alternative importance densities see DiCiccio et al., 1997). We decided to divide $\hat{\sigma}$ by 1.5 for illustrative purposes only. Our importance density is displayed in Fig. 12.4 (dashed line) together with the probit-transformed posterior distribution (solid line).

A generalised harmonic mean estimate can now be obtained using either the original posterior samples $\hat{\theta}_j$ or the probit-transformed samples $\hat{\xi}_j$. Here we choose for the latter option (see also Overstall and Forster, 2010). Incorporating our specific importance density and a correction for using the probit-transformation, Eq. (12.2.22) becomes

$$\hat{p}_3(d) = \left( \frac{1}{K} \sum_{j=1}^{K} \frac{1}{\hat{\sigma}} \phi\left( \frac{\hat{\xi}_j - \hat{\mu}}{\hat{\sigma}} \right) \right)^{-1} \left[ \frac{1}{\hat{\sigma}} \phi\left( \frac{\hat{\xi}_j - \hat{\mu}}{\hat{\sigma}} \right) \right] \left( \frac{f(d | \Phi(\hat{\xi}_j)) \phi(\hat{\xi}_j)}{\hat{\sigma} \phi(\hat{\xi}_j)} \right) \cdot \hat{\xi}_j = \Phi^{-1}(\hat{\theta}_j) \text{ and } \hat{\theta}_j \sim \pi(\theta | d).$$

Note that $\Phi(\hat{\xi}_j) = \hat{\theta}_j$, thus, to evaluate the likelihood, we can simply use the untransformed sample $\hat{\theta}_j$, but to evaluate the importance density and the prior, we have to use the probit-transformed samples $\hat{\xi}_j$ instead. For the particular samples $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{12}$ obtained above, we can now use the generalised harmonic

---

7 Other transformation are conceivable (e.g., logit transformation).
8 A detailed explanation is provided in the appendix. Note that using the original posterior samples $\hat{\theta}_j$ would involve transforming the importance density (e.g., the normal density on $\xi$) to the (0, 1) interval.
12.2. Four sampling methods to approximate the marginal likelihood

Figure 12.4: Illustration of the generalised harmonic mean estimator for the beta-binomial model. The solid line represents the probit-transformed Beta(θ; 3, 9) posterior distribution that was obtained after having observed \( y = 2 \) correct responses out of \( n = 10 \) trials, and the dashed line represents the importance density \( N(\xi; \mu = -0.793, \sigma = 0.423/1.5) \). The grey dots represent the \( K = 12 \) probit-transformed posterior samples.

mean estimator to calculate the following estimate for \( p(d) \)

\[
\hat{p}_3(d) = \left( \frac{1}{12} \sum_{j=1}^{12} \frac{1}{1.5} \frac{0.423^{1/1.5}}{f(d \mid \Phi(\xi_j)) \phi(\xi_j)} \right)^{-1} \\
= \left( \frac{1}{12} \left( \frac{1}{0.423^{1/1.5}} \frac{0.423^{1/1.5}}{0.423^{1/1.5}} + \ldots + \frac{1}{0.423^{1/1.5}} \frac{0.423^{1/1.5}}{0.423^{1/1.5}} \right) \right)^{-1} \\
= \left( \frac{1}{12} \frac{1}{10 \choose 2} (716.81 + \ldots + 556.38) \right)^{-1} \\
= 0.092.
\]

For the first posterior sample \( \xi_1 = -0.77 \), thus, \( \hat{\theta}_1 = 0.22 \) in the original parameterisation, we evaluate the numerator using the R command \( \text{dnorm}(x=-0.77, \text{mean}=0.793, \text{sd}=0.423/1.5) \), the prior using \( \text{dnorm}(x=-0.77) \), and the likelihood using \( \text{dbinom}(x=2, \text{size}=10, \text{prob}=0.22) \).
12.2.4 Method 4: The bridge sampling estimator of the marginal likelihood

As became evident in the last two sections, both the importance sampling estimator and the generalised harmonic mean estimator for $p(d)$ impose strong constraints on the tail behaviour of the importance density relative to the posterior distribution. These conditions make the choice for a suitable importance density complicated, especially when $\theta$ is high dimensional. The bridge sampler, on the other hand, alleviates such requirements (e.g., Frühwirth-Schnatter, 2004).

The bridge sampler was originally developed to directly estimate the Bayes factor, that is, the ratio of the marginal likelihoods of two models $M_1$ and $M_2$ (e.g., Jeffreys, 1961; Kass and Raftery, 1995). However, in this tutorial, we use a version of bridge sampling that allows us to approximate the marginal likelihood of a single model (for an earlier application see for example Overstall and Forster, 2010). This version is based on the following identity

$$1 = \frac{\int f(d \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int f(d \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}, \quad (12.2.25)$$

where $g(\theta)$ is known as the proposal distribution and $h(\theta)$ the so-called bridge function. Multiplying both sides of Eq. (12.2.25) by the marginal likelihood $p(d)$ leads to

$$p(d) = \frac{\int f(d \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int \frac{f(d \mid \theta)\pi(\theta)}{p(d)}h(\theta)g(\theta)d\theta} = \frac{\int f(d \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\frac{\pi(\theta \mid d)}{p(d)}\int h(\theta)g(\theta)d\theta}$$

$$= \frac{E_{g(\theta)}[f(d \mid \theta)\pi(\theta)h(\theta)]}{E_{\text{post}}[h(\theta)g(\theta)]}. \quad (12.2.26)$$

Hence, the marginal likelihood is the expected value $E_{g(\theta)}[f(d \mid \theta)\pi(\theta)h(\theta)]$ with respect to the proposal density $g(\theta)$ divided by the expected value $E_{\text{post}}[h(\theta)g(\theta)]$ with respect to the posterior distribution. To estimate the population mean of the numerator, we use a sample mean by generating $K_2$ samples from the proposal distribution and average the integrand $f(d \mid \theta)\pi(\theta)h(\theta)$ at these samples. Analogously, to estimate the population mean of the denominator, we use a sample mean by generating $K_1$ samples from the posterior distribution and average the integrand $h(\theta)g(\theta)$ at these samples. This yields the bridge sampling estimator $\hat{p}(d)$

$$\hat{p}(d) = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} f(d \mid \tilde{\theta}_i)\pi(\tilde{\theta}_i)h(\tilde{\theta}_i)}{\frac{1}{K_1} \sum_{j=1}^{K_1} h(\hat{\theta}_j)g(\hat{\theta}_j)}, \quad \tilde{\theta}_i \sim g(\theta), \quad \hat{\theta}_j \sim \pi(\theta \mid d). \quad (12.2.27)$$
Conceptually, the proposal distribution is similar to an importance density, it should resemble the posterior distribution, and should have sufficient overlap with the posterior distribution. In fact, we follow Overstall and Forster (2010) and use a normal distribution fitted to probit-transformed samples as the proposal distribution as in the case for the generalised harmonic mean estimator. In our experience, this choice for the proposal distribution works well for a wide range of scenarios. However, this proposal distribution might produce unstable estimates in case of high-dimensional posterior distributions that clearly do not follow a multivariate normal distribution. In such cases, it might be advisable to consider more sophisticated versions of bridge sampling (e.g., Frühwirth-Schnatter, 2004; Meng and Schilling, 2002; Wang and Meng, 2016).

12.2.4.1 Choosing the optimal bridge function

In this tutorial we use the bridge function from Meng and Wong (1996) defined as

\[ h(\theta) = C \frac{1}{q_1 f(\theta | \hat{\theta}) \pi(\hat{\theta}) + q_2 p(d) g(\theta)}, \]  

(12.2.28)

where \( q_1 = \frac{K_1}{K_2 + K_1}, \) \( q_2 = \frac{K_2}{K_2 + K_1}, \) and \( C \) a constant; its particular value is not required because \( h(\theta) \) appears in both the numerator and the denominator of Eq. (12.2.27) and therefore cancels. This particular bridge function is referred to as the “optimal bridge function” because Meng and Wong (1996, p. 837) proved that it minimises the relative mean-squared error, Eq. (12.2.34).

Eq. (12.2.28) shows that the optimal bridge function depends on the marginal likelihood \( p(d) \) which is the very entity we want to estimate. We can resolve this issue by applying an iterative scheme that updates an initial guess of the marginal likelihood until the estimate of the marginal likelihood has converged according to a predefined tolerance level. To do so, we insert the expression for the optimal bridge function, Eq. (12.2.28), into Eq. (12.2.27) as was discussed in Meng and Wong (1996). The formula to approximate the marginal likelihood on iteration \( t + 1 \) is then specified as

\[ \hat{p}(d)^{(t+1)} = \frac{1}{K_2} \sum_{i=1}^{K_2} f(d | \hat{\theta}_i) \pi(\hat{\theta}_i) + q_2 \hat{p}(d)^{(t)} g(\hat{\theta}_i), \]  

(12.2.29)

\[ \frac{1}{K_1} \sum_{j=1}^{K_1} g(\hat{\theta}_j), \]  

(12.2.30)

where \( \hat{p}(d)^{(t)} \) denotes the estimate of the marginal likelihood on iteration \( t \) of the iterative scheme. Note that Eq. (12.2.29) illustrates why bridge sampling is robust to the tail behaviour of the proposal distribution relative to the posterior distribution; the difference to the importance sampling and generalised harmonic
mean estimator is that, in the case of the bridge sampling estimator, samples from the tail region cannot inflate individual summation terms and thus dominate the estimate. This is because both sums displayed in Eq. (12.2.29) involve a ratio that has a sum in the denominator. Nevertheless it should be noted that the posterior distribution and the proposal distribution need to have sufficient overlap. In the extreme scenario of no overlap the bridge sampling estimator is not defined because both sums of Eq. (12.2.29) would be zero.

To simplify matters, we multiply the numerator of the right side of Eq. (12.2.29) by \(1/g(\tilde{\theta}_i)\), the denominator by \(1/g(\tilde{\theta}_j)\), and define \(l_{1,j} := \frac{f(d|\tilde{\theta}_j)\pi(\tilde{\theta}_j)}{g(\theta_j)}\) with samples \(\tilde{\theta}_j\) from the posterior as in the generalised harmonic mean estimator, and \(l_{2,i} := \frac{f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}\) with samples \(\tilde{\theta}_i\) from the proposal distribution as in importance sampling. Once we calculated the values \(l_{1,j}\) and \(l_{2,i}\), for \(j = 1, 2, \ldots, K_1\) and \(i = 1, 2, \ldots, K_2\) respectively, we obtain the formula for the iterative scheme of the bridge sampling estimator \(\hat{p}_4(d)^{(t+1)}\) at iteration \(t + 1\) (Meng and Wong, 1996, p. 837), that is,

\[
\hat{p}_4(d)^{(t+1)} = \frac{1}{K_2} \sum_{i=1}^{K_2} \frac{f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i)}{g(\theta_i)} \frac{1}{1/g(\tilde{\theta}_i)} + \frac{1}{K_1} \sum_{j=1}^{K_1} \frac{g(\theta_j)}{q_1 f(d|\theta_j)\pi(\theta_j) + q_2 \hat{p}_4(d)^{(t)} g(\theta_j)} \frac{1}{1/g(\theta_j)}
\]

Eq. (12.2.31) suggests that, in order to obtain a bridge sampling estimate of the marginal likelihood, a number of requirements need to be fulfilled. First, we need \(K_2\) samples from the proposal distribution \(g(\theta)\) and \(K_1\) samples from the posterior distribution \(\pi(\theta|d)\). Second, for all \(K_2\) samples from the proposal distribution, we evaluate \(l_{2,i}\). This involves obtaining the value of the unnormalised posterior (i.e., the product of the likelihood times the prior) and of the proposal distribution for all samples. Third, we evaluate \(l_{1,j}\) for all \(K_1\) samples from the posterior distribution. This is analogous to evaluating \(l_{2,i}\). Fourth, we need to choose the number of samples \(K_1\) and \(K_2\) for the constants \(q_1\) and \(q_2\). Fifth, we need an initial guess of the marginal likelihood \(\hat{p}_4(d)\). Since some of these five requirements can be obtained easier than others, we will point out possible challenges.

A first challenge is that using a suitable proposal distribution may involve transforming the posterior samples. Consequently, we have to determine how the transformation affects the definition of the bridge sampling estimator for the marginal likelihood, Eq. (12.2.31).

A second challenge is how to use the \(K_1\) samples from the posterior distribution.
12.2. Four sampling methods to approximate the marginal likelihood

One option is to use all $K_1$ samples for both fitting the proposal distribution and for computing the numerator of the bridge estimator. However, Overstall and Forster (2010) showed that such a procedure may result in an underestimation of the marginal likelihood. To obtain more reliable estimates they propose to divide the posterior samples into two parts; the first part is used to obtain the best-fitting proposal distribution in the numerator of $\hat{p}(d)$, and the second part is used to compute the bridge sampling estimate, thus, the denominator of $\hat{p}(d)$. Throughout this tutorial, we use two equally large parts. In the remainder we therefore state that we draw $2K_1$ samples from the posterior distribution. Out of these $2K_1$ posterior samples, we use samples with even index numbers for the first part; posterior samples with odd index numbers constitute the second part.

To summarise, the discussion of the requirements and challenges encountered in bridge sampling illustrated that the bridge sampling estimator imposes less strict requirements on the proposal distribution than the importance sampling and generalised harmonic mean estimator and allows for an almost automatic application due to the default choice of the bridge function, see also the R package bridgesampling by the first author.

12.2.4.2 Running example

![Diagram of bridge sampling steps]

To obtain a bridge sampling estimate of the marginal likelihood in the beta-binomial example, we follow the eight steps illustrated in Fig. 12.5:

1. We draw $2K_1 = 24$ samples from the Beta($\theta$; 3, 9) posterior distribution for $\theta$. 

Figure 12.5: Schematic illustration of the steps involved in constructing a bridge sampling estimator for the marginal likelihood.
Using the R command `rbeta(n=24, shape1=3, shape2=9)`, we obtained the following sample of 24 values
\[ \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{24}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, \]
\[ 0.21, 0.12, 0.15, 0.21, 0.24, 0.18, 0.12, 0.22, 0.15, 0.22, \]
\[ 0.23, 0.26, 0.29, 0.28\}. \]

Note that the first 12 samples equal the ones used in the last section.

2. We choose a proposal distribution.
Here we opt for an approach that can be easily generalised to models with multiple parameters and select a normal distribution as the proposal distribution \( g(\theta) \).

3. We transform the first batch of \( K_1 \) posterior samples.
Since we use a normal proposal distribution, we have to transform the posterior samples from the rate scale to the real line so that the range of the posterior distribution matches the range of the proposal distribution. This can be achieved by probit-transforming the posterior samples, that is, \( \xi_j = \Phi^{-1}(\hat{\theta}_j) \) with \( j \in \{2, 4, \ldots, 24\} \). Using the R command `qnorm(p=firstBatch)` we obtained
\[ \{\hat{\xi}_2, \hat{\xi}_4, \ldots, \hat{\xi}_{24}\} = \{-0.99, -0.39, -0.61, -0.23, -0.18, -1.17, -0.81, \]
\[ -0.92, -0.77, -0.77, -0.64, -0.58\}. \]

4. We fit the proposal distribution to the first batch of \( K_1 \) probit-transformed posterior samples.
For the proposal distribution we use a normal distribution fitted with the method of moments. After probit-transforming the first batch of \( K_1 \) and applying the R commands `mean` and `sd`, we retrieve \( \hat{\mu} = -0.672 \) and \( \hat{\sigma} = 0.298 \). We therefore use \( g(\xi) = \frac{1}{0.298} \phi\left(\frac{\xi + 0.672}{0.298}\right) \) as the proposal density.

5. We draw \( K_2 \) samples from the proposal distribution.
In R we use the command `rnorm(n=12, mean=-0.672, sigma=0.298)` to sample from the fitted normal proposal and obtained
\[ \{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{12}\} = \{-0.90, -0.55, -1.16, -0.53, -0.45, -0.60, -0.63, -0.48, \]
\[ -0.69, -1.20, -0.65, -0.79\}. \]

6. We calculate \( l_{2,i} \) for all \( K_2 \) samples from the proposal distribution.
For this step we evaluate the likelihood and prior at the samples \( \hat{\xi}_i \) for all \( i = 1, 2, \ldots, K_2 \). Recall that the prior was specified in the original parameterisation \( \theta \), while the samples from the proposal \( \xi \) range over the real line. The uniform prior in terms of \( \theta \) implies a standard normal prior in terms of \( \xi \) due to the change-of-variables rule, see Appendix 12.C.

---

9There exist several candidates for the proposal distribution. Alternative proposal distributions are, for example, the importance density that we used for the importance sampling estimator or for the generalised harmonic mean estimator.
As the likelihood function is also specified in terms of the original parameters \(\theta\), we first transform the samples \(\hat{\xi}\) from the proposal distribution that range over the real line to the original parameterisation resulting in \(\hat{\theta}_i = \Phi(\hat{\xi})\).

Thus, to calculate \(l_{2,i}\) for all \(i = 1, 2, \ldots, K_2\) we evaluate the likelihood with the sample in the original parameterisation \(\hat{\theta}_i\), while we use the samples \(\hat{\xi}_i\) to evaluate the normal proposal density and the prior. This can be done in R with the functions `dbinom` and `dnorm` respectively.

7. We transform the second batch of the \(K_1\) posterior samples.
   As in step 2, we use the probit transformation and obtained
   
   \[
   \{\hat{\xi}_1, \hat{\xi}_3, \ldots, \hat{\xi}_{23}\} = \{-0.77, -1.34, -1.55, -0.64, -0.84, -0.81, -1.04, -0.71, -1.17, -1.04, -0.74, -0.55\}.
   \]

8. We calculate \(l_{1,j}\) for the second batch of \(K_1\) probit-transformed samples from the posterior distribution.
   This is analogous to step 6. We plug in the probit-transformed samples \(\hat{\xi}_j\) into the prior and the proposal density in terms of \(\xi\), while we use the untransformed samples \(\hat{\theta}_j\) for the likelihood.

9. We run the iterative scheme, Eq. (12.2.31), until our predefined tolerance criterion is reached.
   With \(l_{1,j}\) and \(l_{2,i}\) at hand we can now run the iterative scheme. As tolerance criterion we choose \(|\hat{p}_4(d)^{(t+1)} - \hat{p}_4(d)^{(t)}|/\hat{p}_4(d)^{(t+1)} \leq 10^{-10}\). This requires an initial guess for the marginal likelihood \(\hat{p}_4(d)^{(0)}\) which we set to 0.10

The simplicity of the beta-binomial model allows us to calculate a bridge sampling estimate for the \(p(d)\) by hand. To determine \(\hat{p}_4(d)^{(t+1)}\) according to Eq. (12.2.31), we need to calculate the constants \(q_1\) and \(q_2\). Since \(K_1 = K_2 = 12\), we obtain \(q_1 = q_2 = 0.5\). In addition, we need to calculate \(l_{2,i}\) for \(i = 1, 2, \ldots, 12\) for all samples from the fitted normal proposal distribution, and \(l_{1,j}\) for \(j = 1, 3, \ldots, 23\) for the second batch of the probit-transformed samples from the posterior distribution. Here we show how to calculate \(l_{2,1}\) and \(l_{1,1}\) using the first sample from the proposal distribution \(\hat{\xi}_1 = -0.9\), thus, \(\hat{\theta}_1 = 0.18\), and the posterior distribution, thus, \(\hat{\xi}_1 = -0.77\), respectively

\[
l_{2,1} = \frac{f(d | \hat{\theta}_1) \phi(\hat{\xi}_1)}{g(\hat{\xi}_1)} = \left(\frac{10}{2}\right) 0.18^2 (1 - 0.18)^8 \cdot 0.27 \cdot \frac{1}{0.298} \cdot \phi\left(\frac{-0.90 + 0.072}{0.298}\right) = 0.080,
\]

\[
l_{1,1} = \frac{f(d | \hat{\theta}_1) \phi(\hat{\xi}_1)}{g(\hat{\xi}_1)} = \left(\frac{10}{2}\right) 0.22^2 (1 - 0.22)^8 \cdot 0.30 \cdot \frac{1}{0.298} \cdot \phi\left(\frac{-0.77 + 0.672}{0.298}\right) = 0.070.
\]

\(^{10}\) A better initial guess can be obtained from the generalised harmonic mean estimator explained in the previous section. In our experience, however, the exact choice of the initial value does not seem to influence the convergence of the bridge sampler much.
12. A Tutorial on Bridge Sampling

For $\hat{p}_4(d)^{(t+1)}$, we then get

$$
\hat{p}_4(d)^{(t+1)} = \frac{1}{K_2} \sum_{i=1}^{K_2} \frac{l_{2,i}}{q_1l_{2,i} + q_2\hat{p}_4(d)^{(t)}} ,
$$

(12.2.32)

$$
\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1l_{1,2j-1} + q_2\hat{p}_4(d)^{(t)}}
$$

$$
= \frac{1}{12} \left( \frac{0.080}{0.5-0.080+0.5\hat{p}_4(d)^{(t)}} + \cdots + \frac{0.071}{0.5-0.085+0.5\hat{p}_4(d)^{(t)}} \right)
$$

(12.2.33)

Using $\hat{p}_4(d)^{(0)} = 0$, we obtain as updated estimate of the marginal likelihood $\hat{p}_4(d)^{(1)} = 0.091$. This iterative procedure has to be repeated until our predefined tolerance criterion is reached. For our running example, this criterion is reached after six iterations. We now obtain $\hat{p}_4(d)^{(6)} = 0.0894$ as a bridge sampling estimate of the marginal likelihood $p(d)$.

12.2.5 Interim summary

So far we used the beta-binomial model to illustrate the computation of four different estimators of the marginal likelihood. These four estimators were discussed in order of increasing sophistication, such that the first three estimators provided the proper context for understanding the fourth—the bridge sampler. This estimator is the focus in the remainder of this tutorial. The goal of the next sections is to demonstrate that bridge sampling is particularly suitable to estimate the marginal likelihood of popular models in mathematical psychology. Importantly, bridge sampling may be used to obtain accurate estimates of the marginal likelihood of hierarchical models (for a detailed comparison of bridge sampling versus its special cases see Frühwirth-Schnatter, 2004; Sinharay and Stern, 2005).

12.2.6 Assessing the accuracy of the bridge sampling estimator

In this section we show how to quantify the accuracy of the bridge estimator. A straightforward approach would be to apply the bridge estimator multiple times and investigate the variability of the marginal likelihood estimate. In practice, however, this solution is often impractical due to the substantial computational burden of obtaining the posterior samples and evaluating the relevant quantities in the bridge sampling procedure.

Frühwirth-Schnatter (2004) proposed an alternative approach that approximates the estimator’s expected relative mean-squared error

$$
RE^2 = \frac{\mathbb{E}[(\hat{p}_4(d) - p(d))^2]}{p(d)^2}
$$

(12.2.34)

The derivation of this approximate relative mean-squared error by Frühwirth-Schnatter takes into account that the samples from the proposal distribution

192


12.3. Case study: Bridge sampling for reinforcement learning models

$g(\theta)$ are independent, whereas the MCMC samples from the posterior distribution $\pi(\theta | d)$ may be autocorrelated. The approximate relative mean-squared error is given by

$$
RE^2 = \frac{1}{K_2} \frac{V_{g(\theta)}[r_1(\theta)]}{\mathbb{E}^2_{g(\theta)}[r_1(\theta)]} + \frac{\rho_{r_2}(0)}{K_1} \frac{V_{\text{post}}[r_2(\theta)]}{\mathbb{E}^2_{\text{post}}[r_2(\theta)]},
$$

(12.2.35)

where $r_1(\theta) = \frac{\pi(\theta | d)}{q_1 \pi(\theta | d) + q_2 g(\theta)}$, $r_2(\theta) = \frac{g(\theta)}{q_1 \pi(\theta | d) + q_2 g(\theta)}$, and where $V_{g(\theta)}[r_1(\theta)] = \int (r_1(\theta) - \mathbb{E}[r_1(\theta)])^2 g(\theta) d\theta$ denotes the variance of $r_1(\theta)$ with respect to the proposal distribution $g(\theta)$ (the variance $V_{\text{post}}[r_2(\theta)]$ is defined analogously), and $\rho_{r_2}(0)$ corresponds to the normalised spectral density of the autocorrelated process $r_2(\theta)$ at the frequency 0.

In practice, we approximate the unknown variances and expected values by the corresponding sample variances and means. Hence, for evaluating the variance and expected value with respect to $g(\theta)$, we use the $K_2$ samples for $\hat{\theta}_i$ from the proposal distribution. To evaluate the variance and expected value with respect to the posterior distribution, we use the second batch of $K_1$ samples $\hat{\theta}_j$ from the posterior distribution which we also use in the iterative scheme for computing the marginal likelihood. Because the posterior samples are obtained via an MCMC procedure and are, hence, autocorrelated, the second term in Eq. (12.2.35) is adjusted by the normalised spectral density (for details see Frühwirth-Schnatter, 2004). The spectral density at frequency zero can be estimated by first fitting an autoregressive model using the \texttt{spectrum0.ar()} function as implemented in the \texttt{coda} R package (Plummer et al., 2006). To evaluate the normalised posterior density which appears in the numerator of $r_1(\theta)$ and the denominator of both $r_1(\theta)$ and $r_2(\theta)$, we use the bridge sampling estimate as normalising constant.

Note that, under the assumption that the bridge sampling procedure $\hat{p}_4(d)$ is an unbiased estimator of the marginal likelihood $p(d)$, the square root of the expected relative mean-squared error, Eq. (12.2.34), can be interpreted as the coefficient of variation (i.e., the ratio of the standard deviation and the mean; Brown, 1998). In the remainder of this chapter, we report the coefficient of variation to quantify the accuracy of the bridge sampling estimator.

12.3 Case study: Bridge sampling for reinforcement learning models

In this section, we illustrate the computation of the marginal likelihood using bridge sampling in the context of a published data set (Busemeyer and Stout, 2002) featuring the Expectancy Valence (EV) model—a popular reinforcement learning model for the Iowa gambling task (IGT; Bechara et al., 1994). We first introduce the task and the model, and then use bridge sampling to estimate the marginal likelihood of the EV model implemented in both an individual-level and a hierarchical Bayesian framework. For the individual-level framework, we compare estimates obtained from bridge sampling to importance sampling estimates published in Steingroever et al. (2016b). For the hierarchical framework, we compare our results to estimates from the Savage-Dickey density ratio test (Dickey, 1971; Dickey and Lientz, 1970; Wagenmakers et al., 2010; Wetzels et al., 2010a).

193
12.3.1 The Iowa gambling task

In this section we describe the IGT (see also Steingroever et al., 2013a, 2013b, 2013c, 2014, 2016a, 2016b). Originally, Bechara et al. (1994) developed the IGT to distinguish decision-making strategies of patients with lesions to the ventromedial prefrontal cortex from the ones of healthy controls (see also Bechara et al., 1998, 1999, 2000). During the last decades, the scope of application of the IGT has increased tremendously covering clinical populations with, for example, pathological gambling tendencies (Cavedini et al., 2002b), obsessive-compulsive disorder (Cavedini et al., 2002a), psychopathic tendencies (Blair et al., 2001), and schizophrenia (Bark et al., 2005; Martino et al., 2007).

The IGT is a card game that requires participants to choose, over several rounds, cards from four different decks in order to maximise their long-term net outcome Bechara et al. (1994, 1997). The four decks differ in their payoffs, and two of them result in negative long-term outcomes (i.e., the bad decks), whereas the remaining two decks result in positive long-term outcomes (i.e., the good decks). After each choice, participants receive feedback on the rewards and losses (if any) associated with that card, as well as their running tally of net outcomes over all trials so far. Unbeknownst to the participants, the task (typically) contains \( N = 100 \) trials.

Table 12.1: Summary of the payoff scheme of the traditional IGT as developed by Bechara et al. (1994)

<table>
<thead>
<tr>
<th>Deck A</th>
<th>Deck B</th>
<th>Deck C</th>
<th>Deck D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad deck with frequent losses</td>
<td>Bad deck with infrequent losses</td>
<td>Good deck with frequent losses</td>
<td>Good deck with infrequent losses</td>
</tr>
<tr>
<td>Reward/trial</td>
<td>100</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Number of losses/10 cards</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Loss/10 cards</td>
<td>-1250</td>
<td>-1250</td>
<td>-250</td>
</tr>
<tr>
<td>Net outcome/10 cards</td>
<td>-250</td>
<td>-250</td>
<td>250</td>
</tr>
</tbody>
</table>

A question is whether and to what extent participants eventually learn to prefer the good decks that allow them to maximise their long-term net outcome. The good decks are typically labelled as decks C and D, whereas the bad decks are labelled as decks A and B. Table 12.1 presents a summary of the traditional payoff scheme as developed by Bechara et al. (1994). This table illustrates that decks A and B yield high constant rewards, but even higher unpredictable losses, thus, a negative long-term net outcome. Decks C and D, on the other hand, yield low constant rewards, but even lower unpredictable losses: hence, the long-term net outcome is positive. In addition to the different payoff magnitudes, the decks also differ in the frequency of losses: decks A and C yield frequent losses, while decks B and D yield infrequent losses.
12.3. Case study: Bridge sampling for reinforcement learning models

Table 12.2: Example data of chosen decks $y_n$ and experienced payoffs $x_n$ for $n = 10$ trials.

<table>
<thead>
<tr>
<th>Trial $n$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_n$</td>
<td>D</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>D</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$x_n$</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>-1250</td>
</tr>
</tbody>
</table>

12.3.2 The Expectancy Valence model

In this section, we describe the EV model (see also Steingroever et al., 2013b, 2014, 2016a, 2016b). Originally proposed by Busemeyer and Stout (2002), the EV model is arguably the most popular model for the IGT (for references see Steingroever et al., 2013b, and for alternative IGT models see Ahn et al., 2008; Dai et al., 2015; Steingroever et al., 2014; Worthy et al., 2013; Worthy and Maddox, 2014).

The model specifies how previous experienced payoffs affect the participant’s next choice in the IGT through the interaction of three model parameters $\theta = (w, a, c)$ that represent distinct psychological processes. In essence, the participant’s next choice at trial $n + 1$ depends on her expected utility for each deck based on the previous $n$ trials and her willingness to exploit this knowledge. It is assumed that the deck with the highest expected utility at the trial $n$ has the largest probability to be chosen in the next trial. The participant’s expected utilities evolve over time and depend on previously choices $y^n = (y_1, y_2, \ldots, y_n)$ through the subsequently observed payoffs $x^n = (x_1, x_2, \ldots, x_n)$, such as the ones depicted in Table 12.2. The model assumes that the participant summarises the experienced payoffs for each deck $y \in \mathcal{Y} = \{A, B, C, D\}$ with a weighted mean of the experienced wins $W_n(y)$ and losses $L_n(y)$ to obtain the utility $v_n(y \mid w)$ where

$$v_n(y \mid w) = (1 - w)W_n(y) + wL_n(y). \quad (12.3.1)$$

Hence, at any trial $n$ there are four utilities $v_n(y \mid w)$. The weight that the participant assigns to losses relative to rewards is referred to as the attention weight parameter $w$. A small value of $w$, that is, $w < 0.5$, is characteristic for decision makers who put more weight on immediate rewards and can thus be described as reward-seeking, whereas a large value of $w$, that is, $w > 0.5$, is characteristic for decision makers who put more weight on the immediate losses and can, thus, be described as loss averse (Ahn et al., 2008; Busemeyer and Stout, 2002).

The actually observed utility $v_n(y \mid w)$ corresponding to the chosen deck $y$ at trial $n$ after observing the payoff $x_n$ might be higher or lower than what the decision maker expected about deck $y$. We write $Ev_{n-1}(y \mid w)$ for the expected utility of deck $y$ extracted from information up to and including trial $n-1$. That is, before the participant has made her choice $Y_n = y$ and before she is presented with the payoff $x_n$ at trial $n$. If the observed utility $v_n(y \mid w)$ is higher (lower) than what was expected $Ev_{n-1}(y \mid w)$, then the expected utility for deck $y$ is shifted upwards (downwards) for the next trial $n + 1$. This learning process is described by the delta learning rule, also known as the Rescorla-Wagner rule (Rescorla and Wagner,
and formalised as

\[ Ev_n(y \mid w, a) = Ev_{n-1}(y \mid w) + a [v_n(y \mid w) - Ev_{n-1}(y \mid w)], \quad \text{for } y \in \mathcal{Y}, \quad (12.3.2) \]

where the parameter \( a \) quantifies the memory for rewards and losses. A value of \( a \) close to zero indicates slow forgetting and weak recency effects, whereas a value of \( a \) close to one indicates rapid forgetting and strong recency effects.

We set \( Ev_0(y \mid w) = 0 \) for every deck \( y \) to convey that the participant has no prior knowledge about the payoffs of the decks. Furthermore, we assume that at trial \( n \) only the expected utility of the chosen deck is updated and that the expected utility of the decks that are not chosen stay untouched. Consequently, Eq. (12.3.2) implies that the expected utility of each deck remains zero until the first time the deck is chosen. For instance, for the example data in Table 12.2 the expected utility of deck A is zero, until \( n = 4 \). The change of expected utility for deck A then plays a role relative to the expected utilities of decks B, C and D in the next trials. This value of the updated expected utility for deck A remains the same from trial \( n + 1 = 5 \) up to trial 8, but before she is presented with the payoff at trial 8.

We assume that the probability with which the participant chooses deck \( y \) at trial \( n + 1 \) is given by the following softmax choice rule\(^{11}\)

\[ Pr(Y_{n+1} = y \mid x^n, \theta) = \frac{e^{u(n) \cdot Ev_n(y \mid w, a)}}{\sum_{y \in \mathcal{Y}} e^{u(n) \cdot Ev_n(y \mid w, a)}}, \quad \text{for } y \in \mathcal{Y}. \quad (12.3.3) \]

The function \( u \) measures how sensitive the participant is to the expected utilities collected up to trial \( n \) for the decision at trial \( n + 1 \). Values of \( u(n) \) close to zero indicate random choice behaviour (i.e., strong exploration), whereas large values of \( u(n) \) indicate choice behaviour that is strongly determined by the expected utilities (i.e., strong exploitation). We parameterise the between trial-dependent sensitivity function \( u(n) \) with the following function

\[ u(n) = (n/10)^c, \quad \text{for } n = 1, 2, \ldots, N, \quad (12.3.4) \]

where \( c \in [-5, 5] \). For positive \( c \), successive choices become less random and more determined by the expected deck utilities; if \( c \) is negative, successive choices become more random and less determined by the expected deck utilities, a pattern that is clearly non-optimal. We restricted the consistency parameter \( c \) of the EV model to the range \([-2, 2]\) instead of the proposed range \([-5, 5]\) (Busemeyer and Stout, 2002). This modification improved the estimation of the EV model and prevented the choice rule from producing numbers that exceed machine precision (see also Steingroever et al., 2014).

In sum, to specify how past experience is processed for the choice in the next trial the EV model uses three parameters \( \theta = (w, a, c) \): (1) the attention weight

\(^{11}\)This rule is also known as the ratio-of-strength choice rule (Luce, 1959). Furthermore, we wrote \( Y_{n+1} \) for the random choice the participant will make before seeing the payoff \( x_{n+1} \). After we observed \( x_{n+1} \), we write \( y_{n+1} \) for the deck that is chosen to convey that it is not random anymore.
parameter \( w \in [0, 1] \) quantifies the weight of losses over rewards at each trial \( n \), (2) the updating parameter \( a \in [0, 1] \) determines how the the observed utility \( v_n(y \mid w) \) of the choosing deck \( y \) affects the expected utility for the next trial, and (3) the response consistency parameter \( c \in [-2, 2] \) determines the balance between exploitation and exploration.

### 12.3.3 Data

We applied bridge sampling to a data set published by Busemeyer and Stout (2002). The data set consists of \( S = 30 \) healthy participants each contributing \( N = 100 \) IGT card selections (see Busemeyer and Stout for more details on the data sets).\(^{12}\)

### 12.3.4 Application of bridge sampling to an individual-level implementation of the EV model

In this section we describe how we use bridge sampling to estimate the marginal likelihood of an individual-level implementation of the EV model. This implementation estimates model parameters for each participant separately. We therefore have as many data sets \( d_1, d_2, \ldots, d_S \) as there are participants. Accordingly, we also obtain a marginal likelihood of the EV model for every participant. The likelihood of the \( s \)th participant follows from plugging in the sequence of observed choices \( y_s^N \) and payoffs \( x_s^N \) into Eq. (12.3.3) gradually resulting in

\[
f(d_s \mid \theta_s) = Pr(Y_{s,1} = y_{s,1} \mid x_0^0, \theta_s) \times Pr(Y_{s,2} = y_{s,2} \mid x_1^1, \theta_s) \times \cdots \times Pr(Y_{s,N} = y_{s,N} \mid x_{N-1}^{N-1}, \theta_s),
\]

where \( x_0^0 = 0 \) resulting in \( Ev_0(y) = 0 \) for every deck \( y = A, B, C, D, \) and \( \theta_s = (w_s, a_s, c_s) \) as before. For each individual \( s \) we use the uniform priors \( w_s \sim U[0, 1], \) \( a_s \sim U[0, 1], \) \( c_s \sim U[-2, 2]. \) As a result, we have

\[
p(d_s \mid \text{Ind}_s) = \int \int \int f(d_s \mid w_s, a_s, c_s) \frac{1}{4} dw_s da_s dc_s,
\]

for every participant \( s \), see Steingroever et al. (2016b) for more details on this prior choice and model implementations.

#### 12.3.4.1 Schematic execution of the bridge sampler

To obtain a bridge sampling estimate of the marginal likelihood for the \( s \)th participant where \( s = 1, 2, \ldots, 30 \) we follow the steps outlined in Fig. 12.5. We proceed as follows:

1. **For each parameter, we draw** \( 2K_s,1 \) **samples from the posterior distribution.**
   
   We use the posterior samples from Steingroever et al. (2016b) who fit an individual-level implementation of the EV model separately to the data of 30 participants due to incomplete choice data.

\(^{12}\)Note that we excluded three participants due to incomplete choice data.
each participant in Busemeyer and Stout (2002). For each participant we have at least 5,000 posterior samples; whenever this number of samples was insufficient to ensure convergence of the Hamiltonian Monte Carlo (HMC) chains, Steingroever et al. (2016b) repeated the fitting routine with 5,000 additional samples. Steingroever et al. (2016b) confirmed convergence of the HMC chains by reporting that all $\hat{R}$ statistics were below 1.05. The posterior samples were split into two batches each consisting of $K_{s,1}$ number of samples.

2. We choose a proposal distribution.
We generalise our approach from the running example and use a multivariate normal distribution as a proposal distribution.

3. We transform the first batch of $K_{s,1}$ posterior samples.
Since we use a multivariate normal distribution as a proposal distribution, we transform all posterior samples to the real line using the probit function, that is, we obtain $\xi_{s,j} = (\hat{\omega}_{s,j}, \hat{\alpha}_{s,j}, \hat{\gamma}_{s,j}) \in \mathbb{R}^3$, where $\hat{\omega}_{s,j} = \Phi^{-1} (\tilde{w}_{s,j})$, $\hat{\alpha}_{s,j} = \Phi^{-1} (\tilde{a}_{s,j})$, $\hat{\gamma}_{s,j} = \Phi^{-1} ((\tilde{c}_{s,j} + 2) / 4)$ for $j = 2, 4, \ldots, 2K_1$.

4. We fit the proposal distribution to the first batch of $K_1$ probit-transformed posterior samples.
We use method-of-moment estimates and use the mean vector and the covariance matrix obtained from the first batch of $K_{s,1}$ probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. We draw $K_{s,2}$ samples from the proposal distribution.
We use R to randomly draw $K_2$ samples from the proposal distribution obtained in step 4.

6. We calculate $l_{s,2,i}$ for all $K_{s,2}$ samples from the proposal distribution.
For this step we evaluate the likelihood and prior at the samples $\xi_{s,i}$ for all $i = 1, 2, \ldots, K_{s,2}$. Recall that the prior was specified in the original parameterisation, while the samples range over the real line. The uniform priors in terms of $\theta$ change into standard normal priors in terms of $\xi$ due to the change-of-variables rule as before, see Appendix 12.C and Appendix 12.D for a more detailed explanation.

As the likelihood function is specified in terms of the parameters $\theta_s$, we first transform the proposal samples that range over the real line to the original parameterisation resulting in $\tilde{\theta}_{s,i} = (\tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i})$, where $\tilde{w}_{s,i} = \Phi (\tilde{\omega}_{s,i})$, $\tilde{a}_{s,i} = \Phi (\tilde{\alpha}_{s,i})$ and $\tilde{c}_{s,i} = 4\Phi (\tilde{\gamma}_{s,i}) - 2$.

Thus, to calculate $l_{s,2,i}$ for all $i = 1, 2, \ldots, K_{s,2}$ we evaluate the likelihood with the sample in the original parameterisation $\tilde{\theta}_{s,i}$, while we use the samples $\xi_{s,i}$ to evaluate the multivariate normal proposal density and the prior, that is,

$$
\frac{\pi(\xi_{s,i})}{f(d_s \mid \tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i}) \phi(\tilde{\omega}_{s,i}) \phi(\tilde{\alpha}_{s,i}) \phi(\tilde{\gamma}_{s,i})} g(\tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i}) \quad (12.3.8)
$$
where \( \phi \) denotes the standard normal density.

7. **We transform the second batch of** \( K_{s,1} \) **posterior samples.**
   This is analogous to step 2.

8. **We calculate** \( l_{s,1,j} \) **for the second batch of** \( K_{s,1} \) **probit-transformed samples from the posterior distribution.**
   This is analogous to step 6. We plug in the probit-transformed samples \( \tilde{\xi}_{s,j} \) into the prior and the proposal density in terms of \( \xi_s \), while we use the untransformed samples \( \tilde{\theta}_{s,j} \) for the likelihood.

9. **We run the iterative scheme, Eq. (12.2.31), until our predefined tolerance criterion is reached.**
   With \( l_{s,1,j} \) and \( l_{s,2,i} \) at hand we can now run the iterative scheme. We use the same tolerance criterion and initialisation \( \tilde{p}_4(d)^0 = 0 \) as in running example. Once convergence is reached, we obtain an estimate of the marginal likelihood for each participant, and derive the coefficient of variation for each participant using Eq. (12.2.35). The largest coefficient of variation is 1.94% suggesting that the bridge sampler has low variance.\(^{13}\)

### 12.3.4.2 Assessing the accuracy of our implementation

To assess the accuracy of our implementation, we compared the marginal likelihood estimates obtained with our bridge sampler to the estimates obtained with importance sampling (Steingroever et al., 2016b). Fig. 12.6 shows the logarithm of the marginal likelihoods \( p(d_1 \mid \text{Ind}_1), p(d_2 \mid \text{Ind}_2), \ldots, p(d_S \mid \text{Ind}_S) \) for the \( S = 30 \) participants of Busemeyer and Stout (2002) obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016b; y-axis). The two sets of estimates correspond almost perfectly. These results indicate a successful implementation of the bridge sampler. Thus, this section emphasises that both the importance sampler and bridge sampler can be used to estimate the marginal likelihood for the data of individual participants. However, when we want to estimate the marginal likelihood of a Bayesian hierarchical model, it may be difficult to find a suitable importance density. The bridge sampler, on the other hand, can be applied more easily and more efficiently.

### 12.3.5 Application of bridge sampling to a hierarchical Implementation of the EV model

In this section we illustrate how bridge sampling is used to estimate the marginal likelihood of a hierarchical EV model. This hierarchical implementation assumes that the parameters \( w_s, a_s, \) and \( c_s \) from each participant are drawn from three separate group-level distributions. This model specification incorporates both the differences and the similarities between participants. We illustrate this application with the data from Busemeyer and Stout (2002) as we have done before, but we now also assume that these participants belong to one group.

\(^{13}\)Note that this measure relates to the marginal likelihoods, not to the log marginal likelihoods.
This one group assumption is formalised using a hierarchical model, where the \( s \)th participant’s parameters \( \theta_s \) are drawn from a group distribution \( f(\theta_s | \eta) \). Given the group-level distribution \( f(\theta_s | \eta) \) with group-level parameters \( \eta \), the data of the individuals are conditionally independent which implies that the likelihood is given by

\[
f(d_{\text{all}} | \theta_1, \ldots, \theta_S, \eta) = \prod_{s=1}^{S} f(d_s | \theta_s) f(\theta_s | \eta) d\theta_s, \tag{12.3.9}
\]

where \( f(d_s | \theta_s) \) is the likelihood of each individual as specified in Eq. (12.3.5). We focus on the group-level parameters \( \eta \) for inference about the group. For a posterior on \( \eta \), we have to choose a prior \( \pi(\eta) \), and the resulting posterior has as normalising constant

\[
p(d_{\text{all}} | \text{Hier}) = \int f(d_{\text{all}} | \theta_1, \ldots, \theta_S, \eta) \pi(\eta) d\eta. \tag{12.3.10}
\]

We use bridge sampling to estimate \( p(d_{\text{all}} | \text{Hier}) \).

### 12.3.5.1 Schematic execution of the bridge sampler

To compute the marginal likelihood, we again follow the steps outlined in Fig. 12.5, with a few minor modifications.
1. **For each parameter, that is, all individual-level and group-level parameters, we draw** $2K_1 = 60,000$ **samples from the posterior distribution.**

To obtain the posterior samples, we fit a hierarchical Bayesian implementation of the EV model to the Busemeyer and Stout (2002) data set using the JAGS software package (Plummer, 2003). Each participant’s parameters are assumed to be drawn from a group-level distribution. As group-level distribution we use a normal distribution characterised by the group-level means and standard deviation parameters expressed in terms of the probit-transformed parameters $\xi_s = (\omega_s, \alpha_s, \gamma_2)$. More specifically, we assume that the $s$th participant’s probit-transformed parameters are drawn from the following distribution

$$f(\xi_s | \eta) = N(\omega_s | \mu_\omega, \sigma^2_\omega) N(\alpha_s | \mu_\alpha, \sigma^2_\alpha) N(\gamma_s | \mu_\gamma, \sigma^2_\gamma), \quad (12.3.11)$$

where $\eta = (\eta_\mu, \eta_\sigma)$ with $\eta_\mu = (\mu_\omega, \mu_\alpha, \mu_\gamma)$ and $\eta_\sigma = (\sigma_\omega, \sigma_\alpha, \sigma_\gamma)$. For inference about these group-level parameters we use standard normal priors on the group means and uniform priors between 0 and 1.5 on the group-level standard deviations, that is, $\pi(\eta) = \pi(\eta_\mu) \pi(\eta_\sigma)$, where

$$\pi(\eta_\mu) = N(\mu_\omega; 0, 1) N(\mu_\alpha; 0, 1) N(\mu_\gamma; 0, 1), \quad (12.3.12)$$

$$\pi(\eta_\sigma) = U(\sigma_\omega; 0, 1.5) U(\sigma_\alpha; 0, 1.5) U(\sigma_\gamma; 0, 1.5). \quad (12.3.13)$$

For a detailed explanation of the hierarchical implementation of the EV model, see Wetzels et al. (2010b).

To reach convergence and reduce autocorrelation, we collect two MCMC chains, each with 120,000 samples from the posterior distributions after having excluded the first 30,000 samples as burn-in. Out of these 120,000 samples per chain, we retained every 4th value yielding 30,000 samples per chain. This setting resulted in all $\hat{R}$ statistics below 1.05 suggesting that all chains have successfully converged from their starting values to their stationary distributions. The resulting posterior samples of the six group-level parameters $\eta$ together with the three individual-level parameters $\theta_s$ for every participant in the group of $S = 30$ individuals are used in the bridge sampler to estimate the marginal likelihood $p(d_{all} | \text{Hier})$.

2. **We choose a proposal distribution.**

We use a multivariate normal distribution as a proposal distribution.

3. **We transform the first batch of $K_1$ posterior samples.**

As before, we ensure that the range of the posterior distribution matches the range of the proposal distribution by using the probit transformation as described above. Hence, for the even samples of the posteriors $j = 2, 4, \ldots, 2K_1$ and each $s = 1, \ldots, 30$ we write $\tilde{\xi}_{s,j} = (\tilde{\omega}_{s,j}, \tilde{\alpha}_{s,j}, \tilde{\gamma}_{s,j})$ for the probit-transformed individual parameters, and $\tilde{\xi}_{\sigma,j} = (\tilde{\omega}_{\sigma,j}, \tilde{\alpha}_{\sigma,j}, \tilde{\gamma}_{\sigma,j})$ for the the probit-transformed group-level standard deviations, where $\tilde{\omega}_{j} =$

---

14 We used a model file that is an adapted version of the model file used by Ahn et al. (2011).

15 As before we define $\omega_s = \Phi^{-1}(w_s)$, $\alpha_s = \Phi^{-1}(a_s)$, $\gamma_s = \Phi^{-1}((c_s + 2)/4))$. 

201
12. A Tutorial on Bridge Sampling

\[ \Phi^{-1}\left(\left(\hat{\sigma}_{\omega,j}\right)/1.5\right), \hat{\tau}_{\alpha,j} = \Phi^{-1}\left(\left(\hat{\sigma}_{\alpha,j}\right)/1.5\right), \hat{\tau}_{\gamma,j} = \Phi^{-1}\left(\left(\hat{\sigma}_{\gamma,j}\right)/1.5\right). \]

The group-level mean parameters do not have to be transformed because they already range across the entire real line.

4. We fit the proposal distribution to the first batch of the \( K_1 \) probit-transformed posterior samples.

We use method-of-moment estimates for the mean vector and the covariance matrix obtained from the first batch of \( K_1 \) probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. We draw \( K_2 \) samples from the proposal distribution.

We use R to randomly draw \( K_2 \) samples from the proposal distribution fitted in step 4. For \( i = 1, 2, \ldots, K_2 \) we obtain group-level parameters \( \tilde{\eta}_i = (\tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i}) \), and \( S = 30 \) number of individual-levels parameter each consisting of \( \xi_{s,i} = (\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i}) \).

6. We calculate \( l_{2,i} \) for all \( K_2 \) samples from the proposal distribution.

For this step we evaluate the prior, the group-level distribution at the samples \( \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i} \) and, subsequently, all individual-level likelihood functions at the samples \( \xi_{1,i}, \xi_{2,i}, \ldots, \xi_{S,i} \) for \( S = 30 \). As the prior on the group-level means are already specified as normal distributions, we simply need to evaluate the normal density at the samples \( \tilde{\eta}_{\mu,i} \). On the other hand, the prior on the group-level standard deviations were specified in the original parameterisation, while the samples \( \tilde{\eta}_{\sigma,i} \) range over the real line. The uniform priors in terms of the \( \sigma s \) change into standard normal priors in terms of \( \zeta \) due to the change-of-variables rule as before, see Appendix 12.C and Appendix 12.E for a more detailed explanation.

As the likelihood \( f(d_{all} \mid \theta_1, \ldots, \theta_S, \eta) \) and the group-level distribution are in terms of the group-level \( \sigma s \) and the original parameterisation \( \theta_s \), we transform the proposal samples that range over the real line to the original parameterisation resulting in \( \tilde{\eta}_{\sigma,i} = (\tilde{\sigma}_{\omega,i}, \tilde{\sigma}_{\alpha,i}, \tilde{\sigma}_{\gamma,i}) \) for the group-level standard deviations, and \( \tilde{\theta}_{s,i} = (\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i}) \), where \( \tilde{\omega}_{s,i} = \Phi(\tilde{\omega}_{s,i}), \tilde{\alpha}_{s,i} = \Phi(\tilde{\alpha}_{s,i}) \), and \( \tilde{\gamma}_{s,i} = 4\Phi(\tilde{\gamma}_{s,i}) - 2 \) for every individual \( s = 1, 2, \ldots, 30 \).

Thus, to calculate \( l_{2,i} \) for all \( i = 1, 2, \ldots, K_2 \) we evaluate the individual-level likelihood functions with the sample in the original parameterisation \( \tilde{\theta}_{s,i} \) and the argument of the group-level density with \( \tilde{\xi}_{s,i} \) for \( s = 1, 2, \ldots, 30 \) and group-level density parameters \( \tilde{\eta}_{\mu,i} \), the original parameterisation \( \tilde{\eta}_{\sigma,i} \) and the prior at \( \tilde{\eta}_{\mu,i} \) and \( \tilde{\eta}_{\sigma,i} \), that is,

\[
\prod_{s=1}^{S} f(d_s \mid \tilde{\theta}_{s,i}) f(\tilde{\xi}_{s,i} \mid \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i}) \phi(\tilde{\omega}_{s,i}) \phi(\tilde{\alpha}_{s,i}) \phi(\tilde{\gamma}_{s,i})
\]

\[
\pi(\tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i})
\]

\[
g(\xi_{1,i}, \ldots, \xi_{S,i}, \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i})
\]

\[ (12.3.14) \]

where the function \( g \) refers to the multivariate normal distribution obtained in step 4.
7. We follow steps 7 – 9, as outlined for the bridge sampler of the individual-level implementation of the EV model.

This procedure yields a logarithm of the marginal likelihood of $-3801.877$ with a coefficient of variation of 10.53%.

12.3.5.2 Assessing the accuracy of our implementation

To investigate the accuracy of our implementation, we compare Bayes factors obtained with bridge sampling to Bayes factors obtained from the Savage-Dickey density ratio test (Dickey and Lientz, 1970; Dickey, 1971; for a tutorial, see Wagenmakers et al., 2010).

Recall that a Bayes factor is a ratio of two marginal likelihood functions. For nested model comparisons with, say, a restricted model $M_r$ within the full model $M_f$, the Savage-Dickey density ratio implies that the Bayes factor can be computed as a ratio of the prior divided by the posterior of the full model at the restriction $\theta = \theta_0$, that is,

$$BF_{fr}(d) = \frac{p(d \mid M_f)}{p(d \mid M_r)} = \frac{\pi(\theta = \theta_0 \mid M_f)}{\pi(\theta = \theta_0 \mid d, M_f)}.$$  \hfill (12.3.15)

As the full model $M_f$ we take the EV model in which all group-level parameters are free to vary. In what follows we also consider three restricted models each with one of the three group-level mean parameters, that is, $\mu_\omega$, $\mu_\alpha$, and $\mu_\gamma$, fixed at a predefined value. These predefined values are choosing such that the Savage-Dickey density ratio for the data at hand is one. To do so, we fit the full EV model to the Busemeyer and Stout (2002) data set (i.e., step 1 of Section 12.3.5) and then apply a nonparametric logspline density estimator (Stone et al., 1997) to the posterior samples. Fig. 12.7 shows the posterior for $\mu_\gamma$ as the full curve, while the prior is shown as the dotted curve. Furthermore, Fig. 12.7 also shows that the prior and posterior for $\mu_\alpha$ evaluated at $\mu_\alpha,0 = -0.604$ have the same value, i.e., the grey dot. Hence, the Bayes factor computed using the Savage-Dickey method is one when we compare the full model in which all parameters are free to vary against the model $M_{r2}$ with $\mu_\alpha$ fixed at $\mu_\alpha,0$. To compute the Bayes factor between the full model and the artificially constructed model $M_{r2}$ with $\mu_\alpha,0 = -0.604$ using bridge sampling, we have to estimate the marginal likelihood of both $M_f$ and $M_{r2}$. The logarithm of the marginal likelihood $M_f$ was already computed in Section 12.3.5 and presented in the top row of Table 12.3. To estimate the marginal likelihood of $M_{r2}$ we first need posterior samples of this restricted model for which we use JAGS as before. This time however, we use the likelihood Eq. (12.3.9) with $\mu_\alpha$ fixed at $\mu_\alpha,0 = -0.604$ in the group distribution Eq. (12.3.11). As $\mu_\alpha$ is known and fixed within $M_{r2}$, it is not random anymore, and therefore does not need a prior. As such, for the restricted model $M_{r2}$ we use the priors

$$\pi(\eta_\mu) = N(\mu_\omega ; 0, 1)N(\mu_\gamma ; 0, 1),$$ \hfill (12.3.16)

$$\pi(\eta_\sigma) = U(\sigma_\omega ; 0, 1.5)U(\sigma_\alpha ; 0, 1.5)U(\sigma_\gamma ; 0, 1.5).$$ \hfill (12.3.17)

\[\text{Under certain regularity conditions that are met in our example, see Marin and Robert (2010); Verdinelli and Wasserman (1995); Wetzels et al. (2010a) for more details.}\]
Figure 12.7: Prior and posterior distributions of the group-level mean $\mu_\alpha$ in the Busemeyer and Stout (2002) data set. The figure shows the posterior distribution (solid line) and the prior distribution (dotted line). The grey dot indicates the intersection of the prior and the posterior distributions, for which the Savage-Dickey density ratio equals 1.

Table 12.3: Bayes factors comparing the full EV model to the restricted EV models, log marginal likelihoods, and coefficient of variation (with respect to the marginal likelihood) expressed as a percentage.

<table>
<thead>
<tr>
<th>Model</th>
<th>BF$<em>{fr}(d</em>{all})$</th>
<th>log marginal likelihood</th>
<th>CV[?]</th>
</tr>
</thead>
<tbody>
<tr>
<td>full model</td>
<td>1.000</td>
<td>$-3801.877$</td>
<td>10.53</td>
</tr>
<tr>
<td>restricted at $\mu_\omega = -0.334$</td>
<td>0.729</td>
<td>$-3801.561$</td>
<td>14.21</td>
</tr>
<tr>
<td>restricted at $\mu_\alpha = -0.604$</td>
<td>0.826</td>
<td>$-3801.686$</td>
<td>9.99</td>
</tr>
<tr>
<td>restricted at $\mu_\gamma = 0.92$</td>
<td>0.710</td>
<td>$-3801.535$</td>
<td>13.15</td>
</tr>
</tbody>
</table>

instead, which is simply the prior of the full model with the prior for the fixed parameter, in this case $\mu_\alpha$ removed. With the posterior samples for the parameters of $\mathcal{M}_{r_2}$ at hand we proceed the estimation procedure as described in Section 12.3.5 from step 2 onwards. This lead to an estimate of the logarithm of the marginal likelihood $p(d_{all} \mid \mathcal{M}_{r_2})$ of $-3801.686$ with a coefficient of variation of 9.99%. Dividing the estimate of the marginal likelihood $p(d_{all} \mid \mathcal{M}_f)$ by the estimate of the marginal likelihood of $p(d_{all} \mid \mathcal{M}_{r_2})$ yields a Bayes factor BF$_{fr_2}(d_{all}) = 0.826$, see the third row in Table 12.3. This table also shows two other restricted models:
\( \mathcal{M}_{r_1} \) with \( \mu_\omega \) fixed at \( \mu_{\omega,0} = -0.334 \), and \( \mathcal{M}_{r_3} \) with \( \mu_\gamma \) fixed at \( \mu_{\gamma,0} = 0.92 \). As before these restrictions were chosen such that the Savage-Dickey density ratio equals one. The corresponding Bayes factors were derived by dividing the estimated marginal likelihood of the full model and the restricted model using the bridge sampler. It is evident that Bayes factors derived from bridge sampling closely approximate the Savage-Dickey density ratio of one. These results suggest a successful implementation of the bridge sampler. This is also reflected by the close match between the log marginal likelihoods of the four models presented in the third column of Table 12.3.

Finally, we confirm that the bridge sampler has low variance; the coefficient of variation with respect to the marginal likelihood of the full model and the three restricted models ranges between 9.99 and 14.21\%.

12.4 Discussion

In this tutorial, we explained how bridge sampling can be used to estimate the marginal likelihood of popular models in mathematical psychology. As a running example, we used the beta-binomial model to illustrate step-by-step the bridge sampling estimator. To facilitate the understanding of the bridge sampler, we first discussed three of its special cases—the naive Monte Carlo estimator, the importance sampling estimator, and the generalised harmonic mean estimator. Consequently, we introduced key concepts that became gradually more complicated and sophisticated. In the second part of this tutorial, we showed how bridge sampling can be used to estimate the marginal likelihood of both an individual-level and a hierarchical implementation of the EV model (Busemeyer and Stout, 2002) for the Iowa gambling task (Bechara et al., 1994). The running example and the application of bridge sampling to the EV model demonstrated the positive aspects of the bridge sampling estimator, that is, its accuracy, reliability, practicality, and ease-of-implementation (DiCiccio et al., 1997; Frühwirth-Schnatter, 2004; Meng and Wong, 1996).

The bridge sampling estimator is superior to the naive Monte Carlo estimator, the importance sampling estimator, and the generalised harmonic mean estimator for several reasons. First, Meng and Wong (1996) showed that, among the four estimators discussed in this chapter, the bridge sampler minimises the mean-squared error because it uses the optimal bridge function. Second, in bridge sampling, choosing a suitable proposal distribution is much easier than choosing a suitable importance density for the importance sampling estimator or the generalised harmonic mean estimator because bridge sampling is more robust to the tail behaviour of the proposal distribution relative to the posterior distribution. This advantage facilitates the application of the bridge sampler to higher dimensional and complex models. This characteristic of the bridge sampler combined with the popularity of higher dimensional and complex models in mathematical psychology suggests that bridge sampling can advance model comparison exercises in many areas of mathematical psychology (e.g., reinforcement-learning models, response time models, multinomial processing tree models, etc.). Third, bridge sampling is relatively straightforward to implement. In particular, our step-by-step procedure
can be easily applied to other models with only minor changes of the code (i.e., the unnormalised posterior and potentially the proposal function have to be adapted).

Despite the numerous advantages of the bridge sampler, the take-home message of this tutorial is not that the bridge sampler should be used blindly. There exist a large variety of methods to approximate the marginal likelihood that differ in their efficiency. The most appropriate method optimises the trade-off between accuracy and implementation effort. This trade-off depends on a number of aspects such as the complexity of the model, the number of models under consideration, the statistical experience of the researcher, and the time available. This suggests that the choice of the method should be reconsidered each time a marginal likelihood needs to be obtained. Obviously, when the marginal likelihood can be determined analytically, bridge sampling is not needed at all. If the goal is to compare (at least) two nested models, the Savage-Dickey density ratio test (Dickey and Lientz, 1970; Dickey, 1971) might be a better alternative. Note however that the Savage-Dickey density ratio is not free of caveats—the full curve depicted in Fig. 12.7 is just an estimate of the posterior and also subject to estimation error.\footnote{In fact, uncertainty quantification of frequentist nonparametric methods has not yet become satisfactory.} If only an individual-level implementation of a model is used, importance sampling may be easier to implement and may require less computational effort. If the goal is to obtain the marginal likelihood of a large number of relatively simple models, the product space or reversible jump method might be more appropriate (Carlin and Chib, 1995; Green, 1995). If a researcher with a limited programming background and/or little time resources wants to conduct a model comparison exercise, rough approximations of the Bayes factor, such as the Bayesian information criterion, might be more suitable (Schwarz, 1978). On the other hand, a researcher with an extensive background in programming and mathematical statistics might consider using path sampling—a generalisation of bridge sampling (Gelman and Meng, 1998).

To conclude, in this tutorial we showed that bridge sampling offers a reliable and easy-to-implement approach to estimate a model’s marginal likelihood, see also bridgesampling R package of the first author. Bridge sampling can be profitably applied to a wide range of problems in mathematical psychology involving parameter estimation, model comparison, and Bayesian model averaging.
12.A The bridge sampling estimator as a general case of methods 1 – 3

Table 12.4 shows how the naive Monte Carlo, the importance sampling, and the generalised harmonic mean estimators are special cases of the bridge sampling estimator under specific choices of the bridge function $h(\theta)$ and the proposal distribution $g(\theta)$.

Table 12.4: Summary of the bridge sampling estimators for the marginal likelihood and its special cases: The naive Monte Carlo, importance sampling, and generalised harmonic mean estimator

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimator</th>
<th>Samples</th>
<th>Bridge function $h(\theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridge sampling</td>
<td>$\frac{1}{K} \sum_{i=1}^{K} f(d</td>
<td>\tilde{\theta}_i) \pi(\tilde{\theta}_i) h(\tilde{\theta}_i)$</td>
<td>$\tilde{\theta}_i \sim g(\theta)$</td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{K} \sum_{j=1}^{K} h(\hat{\theta}_j) g(\hat{\theta}_j)$</td>
<td>$\hat{\theta}_j \sim \pi(\theta</td>
<td>d)$</td>
</tr>
<tr>
<td>Naive Monte Carlo</td>
<td>$\frac{1}{K} \sum_{i=1}^{K} f(d</td>
<td>\tilde{\theta}_i)$</td>
<td>$\tilde{\theta}_i \sim \pi(\theta)$</td>
</tr>
<tr>
<td>Importance sampling</td>
<td>$\frac{1}{K} \sum_{i=1}^{K} \frac{f(d</td>
<td>\tilde{\theta}_i) \pi(\tilde{\theta}<em>i)}{g</em>{IS}(\tilde{\theta}_i)}$</td>
<td>$\tilde{\theta}<em>i \sim g</em>{IS}(\theta)$</td>
</tr>
<tr>
<td>Generalised harmonic mean</td>
<td>$\left( \frac{1}{K} \sum_{i=1}^{K} \frac{g_{IS}(\tilde{\theta}_i)}{p(d</td>
<td>\tilde{\theta}_i) \pi(\tilde{\theta}_i)} \right)^{-1}$</td>
<td>$\tilde{\theta}_i \sim \pi(\theta</td>
</tr>
</tbody>
</table>

In the table above $\pi(\theta)$ denotes the prior distribution, $g_{IS}(\theta)$ the importance density, $\pi(\theta | d)$ the posterior distribution, $g(\theta)$ the proposal distribution, $h(\theta)$ the bridge function, and $C$ a constant. The last column shows the bridge function needed to obtain the special cases.

12.B Bridge sampling implementation: Avoiding numerical issues

In order to avoid numerical issues, we rewrite Eq. (12.2.31) as

\[\text{Note that bridge sampling is also a general case of the Chib and Jeliazkov (2001) method of estimating the marginal likelihood using the Metropolis-Hastings acceptance probability (Meng and Schilling, 2002; Mira and Nicholls, 2004).} \]
the generalised harmonic mean estimator, Eq. (12.2.22), to yield Eq. (12.2.23). Rectifies our expression of
In this section we describe how the probit transformation \( \hat{r} \) yields
add the marginal likelihood if we multiply \( \hat{r} \)

where \( l^* \) is a constant which we can choose in a way that keeps the terms in the
sums manageable. We used \( l^* = \text{median}(\log(l_{1,j})) \). To further simplify matters,
we defined \( \hat{r}(t) = \hat{p}_4(y^{(t)}) \exp(-l^*) \) and multiply the above expressions by \( \exp(-l^*) \)
on both sides resulting in

\[
\hat{r}(t+1) = \frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp \left( \log(l_{2,i}) - l^* \right)}{q_1 \exp \left( \log(l_{2,i}) - l^* \right) + q_2 \hat{p}_4(y^{(t)}) \exp(-l^*)} \]

(12.B.6)

Hence, we can run the iterative scheme with respect to \( \hat{r} \) which is more convenient
because it keeps the terms in the sums manageable. We obtain an estimate of
the marginal likelihood if we multiply \( \hat{r} \) by \( \exp(l^*) \). Equivalently, we obtain an
estimate of the logarithm of the marginal likelihood if we take the logarithm of \( \hat{r} \)
and add \( l^* \).

12.C Correcting for the probit transformation

In this section we describe how the probit transformation affects our expression of
the generalised harmonic mean estimator, Eq. (12.2.22), to yield Eq. (12.2.23). Recall that we derived the generalised harmonic mean estimator using the following
$$\frac{1}{p(d)} = \int \frac{g_{IS}(\theta)}{f(d|\theta)\pi(\theta|d)} \pi(\theta|d)d\theta. \quad (12.C.1)$$

For practical reasons, in the running example, we used a fitted normal distribution in terms of the probit-transformed parameters $\xi$ as importance density, that is, $g_{IS}(\xi) = \frac{1}{\sigma} \phi\left(\frac{\xi - \mu}{\sigma}\right)$. Note that this importance density is a function of $\xi$, whereas the general importance density $g_{IS}$ in Eq. (12.C.1) is specified in terms of $\theta$.

We now have two choices, we either (i) express the likelihood, prior and posterior in terms of $\xi$, or (ii) we express the importance sampler in terms of $\theta$ instead. For (i) we recall that $\theta = \Phi(\xi)$ and that the derivative of $\theta$ with respect to $\xi$ is then $\frac{d\theta}{d\xi} = \phi(\xi)$. As such, we fill in $\theta = \Phi(\xi)$ in the likelihood, prior, posterior and $\int d\theta = \int \phi(\xi)d\xi$ resulting in

$$p(d) = \left[ \mathbb{E}_{post} \left[ \begin{array}{c} \text{importance density} \\
\frac{1}{\sigma} \phi\left(\frac{\xi - \mu}{\sigma}\right) \\
\frac{f(d|\Phi(\xi))\pi(\Phi(\xi))\phi(\xi)}{\text{likelihood} \quad \text{prior}} \end{array} \right] \right]^{-1}, \quad (12.C.2)$$

where the expectation is with respect to the posterior in terms of $\xi$. Note that $\Phi(\xi) = \theta$, thus, we plugged in the untransformed $\theta$ in the likelihood, the prior and multiply it with the standard normal density $\phi$ to compensate for the probit-transform. In the online-provided code, we use this approach (see also Overstall and Forster, 2010). Note that in our running example with a uniform prior we have $\pi(\theta) = 1$ for every $\theta \in \Theta$ and therefore $\pi(\Phi(\xi)) = 1$ for every $\xi \in \mathbb{R}$.

Alternatively, for (ii) we recall that the integral of importance density is given by $\int g_{IS}(\xi)d\xi$ and that $\xi = \Phi^{-1}(\theta)$. The derivative of $\xi$ with respect to $\theta$ is then $\frac{d\xi}{d\theta} = \frac{1}{\phi(\theta)}$ due to the inverse function theorem. As such, we get the equivalent expression

$$p(d) = \left[ \mathbb{E}_{post} \left[ \begin{array}{c} \text{importance density} \\
\frac{1}{\sigma} \phi\left(\frac{\Phi^{-1}(\theta) - \mu}{\sigma}\right) \\
\frac{1}{\phi(\Phi^{-1}(\theta))} \frac{f(d|\theta)\pi(\theta)}{\text{likelihood} \quad \text{prior}} \end{array} \right] \right]^{-1}, \quad (12.C.3)$$

where the expectation is with respect to the posterior in terms of $\theta$.

These two population means can be estimated by their respective sample
means, that is,

\[ \hat{\rho}_3(d) = \left( \frac{1}{K} \sum_{j=1}^{K} \frac{1}{\sigma} \left( \frac{\hat{\xi}_j - \bar{\mu}}{\sigma} \right) \right)^{-1}, \quad \hat{\xi}_j = \Phi^{-1}(\theta^*) \]  
[\Theta \sim \pi(\theta | d), \]  
(probit-transformed samples from the posterior distribution)

\[ \hat{\varphi}_j = \varphi^{-1}(\hat{\theta}_j - \bar{\mu}) \]  
[\varphi^{-1}(\hat{\theta}_j) \]  
(samples from the posterior distribution)

which is a result of interpreting integration as glorified summation.

### 12.D Details on the application of bridge sampling to the individual-level EV model

In this section, we provide more details on how we obtained the unnormalised posterior distribution for a specific participant \( s \), for \( s = 1, 2, \ldots, 30 \), with choices \( y^N_s = (y_{s,1}, y_{s,2}, \ldots, y_{s,N}) \) and corresponding payoffs \( x^N_s = (x_{s,1}, x_{s,2}, \ldots, x_{s,N}) \).

As explained in Appendix 12.B, we run the iterative scheme with respect to \( \hat{\rho} \) to avoid numerical issues. Consequently, we have to compute \( \log(l_{1,j}) \) and \( \log(l_{2,i}) \). We do so by transforming the priors specified with the original parameterisation to the real line using the probit transform. For the parameters \( w_s \sim U[0, 1] \) and \( \alpha_s \sim U[0, 1] \) we get standard normal priors on \( w_s \) and \( \alpha_s \) as was elaborated on in Appendix 12.C. For the parameter \( c \), the uniform prior \( U[-2, 2] \) implies that \( \int_{-2}^{2} \pi(c) dc = 0.25 \). To apply the change-of-variable rule we recall that \( c = 4\varphi(\gamma) + 2 \) and subsequently take the derivative of \( c \) with respect to \( \gamma \), which results in \( \frac{dc}{d\gamma} = 4\phi(\gamma) \). Hence, the uniform prior on \( \gamma \) is also just the normal density.

As such, to calculate \( \log(l_{2,i}) \) with \( \hat{\xi}_{s,i} \) for the \( i \)th sample from the proposal distribution we get

\[ \log(l_{2,i}) = \log \left( \frac{f(d_s | \hat{\theta}_{s,i})\pi(\hat{\theta}_{s,i})\phi(\hat{\xi}_{s,i})}{g(\hat{\xi}_{s,i})} \right), \]  
(12.D.1)

where \( \hat{\theta}_{s,i} \) refers to the sampled \( \hat{\xi}_{s,i} \) transformed to the original parameterisation. Taking the logarithm simplifies matters as multiplication then becomes a summation. That is,

\[ \log(l_{2,i}) = \sum_{n=1}^{N} \log Pr(y_{s,n} | x^{n-1}, \hat{\theta}_{s,i}) + \log \phi(\hat{\omega}_{s,i}) + \log \phi(\hat{\alpha}_{s,i}) + \log \phi(\hat{\gamma}_{s,i}) \]

\[ - \log g(\hat{\xi}_{s,i}), \]  
(12.D.2)
as a result of taking independent uniform priors on the parameters and because $\log 1 = 0$.

### 12.E Details on the application of bridge sampling to the hierarchical EV model

Analogous to the last section, we explain here how we obtained the logarithm of the unnormalised posterior for the hierarchical implementation of the EV model. As in Appendix 12.D, we run the iterative scheme with respect to $\hat{r}$ and, therefore, compute $\log(l_{1,j})$ and $\log(l_{2,i})$ in terms of the probit-transformed parameters. The priors on the group-level means are just standard normal, while the prior on the group-level standard deviations were given in terms of the $\sigma$s. These prior in terms of the probit-transformed $\tau$s are also standard normal, which can be derived analogously to how we showed that the uniform prior of $c$ on $[-2, 2]$ results in a standard normal density on $\gamma$, see Appendix 12.D. As the hierarchical model also incorporates a group-level distribution the logarithm of $l_{2,i}$ is now given by

$$
\log(l_{2,i}) = \sum_{s=1}^{S} \left[ \log f(d_s | \hat{\theta}_{s,i}) + \log f(\hat{\xi}_{s,i} | \hat{\eta}_{\mu,i}, \hat{\eta}_{\sigma,i}) \right],
$$

(12.E.1)

$$
+ \log \phi(\bar{\mu}_{\omega,i}) + \log \phi(\bar{\mu}_{\alpha,i}) + \log \phi(\bar{\mu}_{\gamma,i}),
$$

(12.E.2)

$$
+ \log \phi(\bar{\tau}_{\omega,i}) + \log \phi(\bar{\tau}_{\alpha,i}) + \log \phi(\bar{\tau}_{\gamma,i}) - \log g(\bar{\xi}_{1,i}, \ldots, \bar{\xi}_{S,i}, \tilde{\eta}_i),
$$

(12.E.3)

where $\log f(d_s | \hat{\theta}_{s,i}) = \sum_{n=1}^{N} \log Pr(y_{s,n} | x^{n-1}, \hat{\theta}_{s,i})$ and the logarithm of the group-level distribution is

$$
\log f(\hat{\xi}_{s,i} | \hat{\eta}_{\mu,i}, \hat{\eta}_{\sigma,i}) = \log \frac{1}{\bar{\sigma}_{\omega,i}} \phi \left( \frac{\bar{\omega}_{s,i} - \bar{\mu}_i}{\bar{\sigma}_{\omega,i}} \right) + \log \frac{1}{\bar{\sigma}_{\alpha,i}} \phi \left( \frac{\bar{\alpha}_{s,i} - \bar{\mu}_i}{\bar{\sigma}_{\alpha,i}} \right)
$$

(12.E.4)

$$
+ \log \frac{1}{\bar{\sigma}_{\gamma,i}} \phi \left( \frac{\bar{\gamma}_{s,i} - \bar{\mu}_i}{\bar{\sigma}_{\gamma,i}} \right)
$$

(12.E.5)

Note that the hierarchical implementation implies that each draw from the proposal $g$, say, the $i$th, consists of six group-level draws $\hat{\eta}_{\mu,i}$ and $\hat{\xi}_{\sigma,i}$, and $S = 30$ individual-level $\hat{\xi}_{s,i}$, each consisting of three parameters $\hat{\xi}_{s,i} = (\bar{\omega}_{s,i}, \bar{\alpha}_{s,i}, \bar{\gamma}_{s,i})$. As such, each draw of the proposal is a vector of length 96. To evaluate $\log l_{2,i}$ we transform these samples to the parameters in which the individual-level likelihood and the group-level distribution are specified.