Safe models for risky decisions

Steingröver, H.M.

Publication date
2017
Document Version
Other version
License
Other

Chapter 8

A Tutorial on Bridge Sampling

This chapter is currently in preparation as:
Quentin F. Gronau, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers, Helen Steingroever (in preparation).
A tutorial on bridge sampling.

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 sampling methods. Here we provide a tutorial on bridge sampling (Meng & Wong, 1996), a reliable and relatively straightforward sampling method that allows researchers to obtain 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.

Bayesian statistics has become increasingly popular in mathematical psychology (Andrews & Baguley, 2013; Bayarri, Benjamin, Berger, & Sellke, 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen, Levy, Millsap, & Fox, 2015; Wetzels et al., 2016). The Bayesian approach is conceptually simple, theoretically coherent, and easily applied to relatively complex problems. These problems involve, for instance, hierarchical modeling (Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Matzke & Wagenmakers, 2009; Rouder & Lu, 2005; Rouder et al., 2005, 2007) or the comparison of non-nested models (Lee, 2008; Lee & Wagenmakers, 2005; Pitt, Myung, & Zhang, 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 & Meng, 1998).

First, in parameter estimation, we consider a single model and aim to quantify the uncertainty for a parameter of interest θ after having observed the data y. This is realized by means of a posterior distribution; here, the marginal likelihood of the data $p(y)$ ensures that the posterior distribution is a proper probability density function (PDF) in the sense that it integrates to 1. This is evident from Bayes theorem:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta')\, p(\theta')\, \mathrm{d}\theta'} = \frac{\overbrace{p(y \mid \theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(y)}_{\text{marginal likelihood}}}. \tag{8.1}$$

Equation 8.1 shows that the marginal likelihood is independent of the parameter θ, and is just a single number. This illustrates why in parameter estimation the marginal likelihood is referred to as a normalizing constant.

Second, in model comparison, we consider m ($m \in \mathbb{N}$) competing models, and are interested in the evidence that the data y provide for a particular model $\mathcal{M}_i$ ($i \in \{1, 2, \ldots, m\}$) given the prior probabilities of all models under consideration (see three special issues on this topic in the *Journal of Mathematical Psychology*: Mulder & Wagenmakers, 2016; I. J. Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). This evidence is quantified by the so-called posterior model probability $p(\mathcal{M}_i \mid y)$ of model $\mathcal{M}_i$ given the data y (Berger & Molina, 2005):

$$p(\mathcal{M}_i \mid y) = \frac{p(y \mid \mathcal{M}_i)\, p(\mathcal{M}_i)}{\sum_{j=1}^{m} p(y \mid \mathcal{M}_j)\, p(\mathcal{M}_j)}, \tag{8.2}$$

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 evidence or likelihood of the model (Kass & Raftery, 1995; Ntzoufras, 2009).

If only two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are considered, Equation 8.2 can be used to quantify the relative posterior model plausibility of model $\mathcal{M}_1$ compared to model $\mathcal{M}_2$. This relative plausibility is given by the ratio of the posterior probabilities of both models, and is referred to as the posterior model odds:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid y)}{p(\mathcal{M}_2 \mid y)}}_{\substack{\text{posterior}\\ \text{odds}}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\substack{\text{prior}\\ \text{odds}}} \times \underbrace{\frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_2)}}_{\substack{\text{Bayes}\\ \text{factor}}}. \tag{8.3}$$

Equation 8.3 illustrates that the posterior model odds are the product of two factors: The first factor is the ratio of the prior probabilities of both models—the prior model odds. The second factor is the ratio of the marginal likelihoods of both models—the so-called Bayes factor (Etz & Wagenmakers, submitted; Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 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 & 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 also plays an important role in Bayesian model averaging (BMA; Hoeting, Madigan, Raftery, & Volinsky, 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, BMA aims to quantify the evidence for a specific component (e.g., a specific learning rule) that is assumed by several models under consideration. Specifically, BMA quantifies the model-averaged evidence in favor of that specific component by computing the so-called posterior inclusion probability. The posterior inclusion probability is given by the sum of the posterior model probabilities of all models that contain the specific component, and as such depends on the marginal likelihood of the models.

A problem that arises in all three areas –parameter estimation, model comparison and BMA– is that the marginal likelihood involves an integral (see Equation 8.1) that is tractable only in certain restricted examples. This is a pressing problem 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 Bayesian 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 & Lu, 2005; Rouder et al., 2005, 2008; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, et al., 2010). For instance, consider a four-parameter Bayesian hierarchical model with four group-level distributions each characterized by two parameters and a group size of 30 participants; this then results in 30×4 individual-level parameters and 2×4 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, the marginal likelihood can be approximated using several Monte Carlo sampling methods. In this tutorial we focus on four such methods: the bridge sampling estimator and its three commonly used special cases—the naive Monte Carlo estimator, the importance sampling estimator, and the generalized harmonic mean estimator (for alternative methods see Gamerman & Lopes, 2006, Chapter 7; and for alternative approximation methods relevant to model comparison and BMA see Carlin & Chib, 1995; Green, 1995).[1] As we will illustrate throughout this tutorial, the bridge sampler is accurate, efficient, and relatively straightforward to implement (e.g., DiCiccio, Kass, Raftery, & Wasserman, 1997; Frühwirth-Schnatter, 2004; Meng & 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 generalized 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 & 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 article, we use the software package R to implement the bridge sampling

[1] Appendix F gives a derivation showing that the first three estimators are indeed special cases of the bridge sampler.

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`.

8.1 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). The marginal likelihood is the probability of the observed data y given a specific model of interest $\mathcal{M}$, and is defined as the integral of the likelihood over the prior:

$$\underbrace{p(y \mid \mathcal{M})}_{\substack{\text{marginal} \\ \text{likelihood}}} = \int \underbrace{p(y \mid \theta, \mathcal{M})}_{\text{likelihood}} \underbrace{p(\theta \mid \mathcal{M})}_{\text{prior}} \, \mathrm{d}\theta, \tag{8.4}$$

with θ a vector containing the model parameters. Equation 8.4 illustrates that the marginal likelihood can be interpreted as a weighted average of the likelihood of the data given a specific value for θ where the weight is the a priori plausibility of that specific value. Equation 8.4 can therefore be written as an expected value:

$$p(y \mid \mathcal{M}) = \mathbb{E}_{\text{prior}}(p(y \mid \theta, \mathcal{M})),$$

where the expectation is taken with respect to the prior distribution. This idea is central to the four sampling methods that we discuss in this tutorial.

Introduction of the Running Example: The Beta-Binomial Model

Our running example focuses on estimating the marginal likelihood for a binomial model assuming a uniform prior on the rate parameter θ (i.e., the beta-binomial model). Consider a single participant who answered $k = 2$ out of $n = 10$ true/false questions correctly. Assume that the number of correct answers follows a binomial distribution, that is, $k \sim \text{Binomial}(n, \theta)$ with $\theta \in (0, 1)$, where θ represents the latent probability for answering any one question correctly. The probability mass function (PMF) of the binomial distribution is given by:

$$\text{Binomial}(k \mid n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \tag{8.5}$$

where $n \in \mathbb{Z}_{\geq 0}$, and $k \leq n$. The PMF of the binomial distribution serves as the likelihood function in our running example.

In the Bayesian framework, we also have to specify the prior distribution of the model parameters; the prior distribution expresses our knowledge about the parameters before the data have been observed. In our running example, we assume that all values of θ are equally likely a priori. This prior belief is captured by a uniform distribution across the range of θ, that is, $\theta \sim \text{Uniform}(0, 1)$ which can equivalently be written in terms of a beta distribution $\theta \sim \text{Beta}(1, 1)$. This prior distribution is represented by the dotted line in Figure 8.1. It is evident that the density of the prior distribution equals 1 for all values of θ. One advantage of expressing the

prior distribution by a beta distribution is that its two parameters (i.e., in its general form the shape parameters α and β) can be thought of as counts of "prior successes" and "prior failures", respectively. In its general form, the PDF of a Beta(α, β) distribution $(\alpha, \beta > 0)$ is given by:

$$\text{Beta}(x \mid \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta)$ is the beta function that is defined as: $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

Analytical derivation of the marginal likelihood

As we will see in this section, for the beta-binomial model the marginal likelihood is analytic. Assuming a general k and n, we obtain the marginal likelihood as:

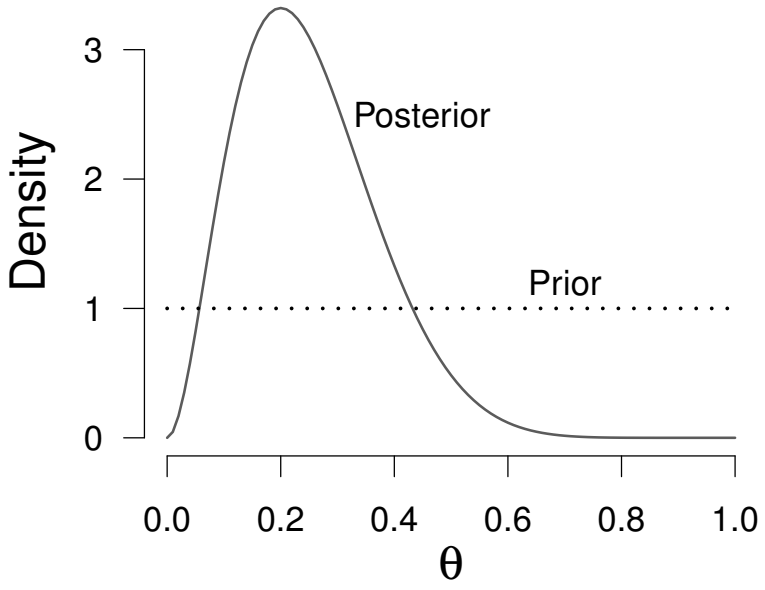

Figure 8.1: Prior and posterior distribution for the rate parameter θ from the beta-binomial model. The Beta$(1, 1)$ prior on the rate parameter θ is represented by the dotted line; the Beta$(3, 9)$ posterior distribution is represented by the solid line and was obtained after having observed 2 correct responses out of 10 trials.

$$p(k \mid n) \overset{Eq.\ 8.4}{=} \int_0^1 p(k \mid n, \theta)\, p(\theta)\, \mathrm{d}\theta \overset{Eq.\ 8.5}{=} \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k}\, 1\, \mathrm{d}\theta$$
$$= \binom{n}{k} B(k+1, n-k+1) = \frac{1}{n+1}\,,$$

where we suppress the "model" in the conditioning part of the probability statements because we focus on a single model in this running example. Using $k = 2$ and $n = 10$ of our example, we obtain:

$$p(k = 2 \mid n = 10) = \frac{1}{11} \approx 0.0909.$$

The marginal likelihood can now be used to derive the posterior distribution for θ after having observed the data. Using Bayes theorem, we obtain:

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

which we recognize as the PDF of the Beta$(k+1, n-k+1)$ distribution. Thus, if we assume a uniform prior on θ and observe $k = 2$ correct responses out of $n = 10$ trials, we obtain a Beta$(3, 9)$ distribution as posterior distribution. This distribution is represented by the solid line in Figure 8.1. In general, if $k \mid n, \theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(1, 1)$, then $\theta \mid n, k \sim \text{Beta}(k+1, n-k+1)$.

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 & Handscomb, 1964; Raftery & Banfield, 1991). This method uses the standard definition of the marginal likelihood (Equation 8.4), and relies on the central idea that the marginal likelihood can be written as an expected value with respect to the prior distribution:

$$\underbrace{p(y)}_{\substack{\text{marginal}\\ \text{likelihood}}} = \int \underbrace{p(y \mid \theta)}_{\text{likelihood}}\, \underbrace{p(\theta)}_{\text{prior}}\, d\theta = \mathbb{E}_{\text{prior}}(p(y \mid \theta)).$$

The expected value of the likelihood with respect to the prior can be approximated by evaluating the likelihood in N samples from the prior distribution for θ and averaging the resulting values. This yields the naive Monte Carlo estimator $\hat{p}_1(y)$:

$$\hat{p}_1(y) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} p(y \mid \tilde{\theta}_i)}_{\text{average likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim p(\theta)}_{\substack{\text{samples from the}\\ \text{prior distribution}}}\,. \tag{8.6}$$

Running example

To obtain the naive Monte Carlo estimate of the marginal likelihood in our running example, we need N samples from the Beta$(1, 1)$ prior distribution for θ. For illustrative purposes, we limit the number of samples to 12. We obtain the following samples:

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

where we use the tilde symbol to emphasize that we refer to a sampled value. All sampled values are represented by the gray dots in Figure 8.2.

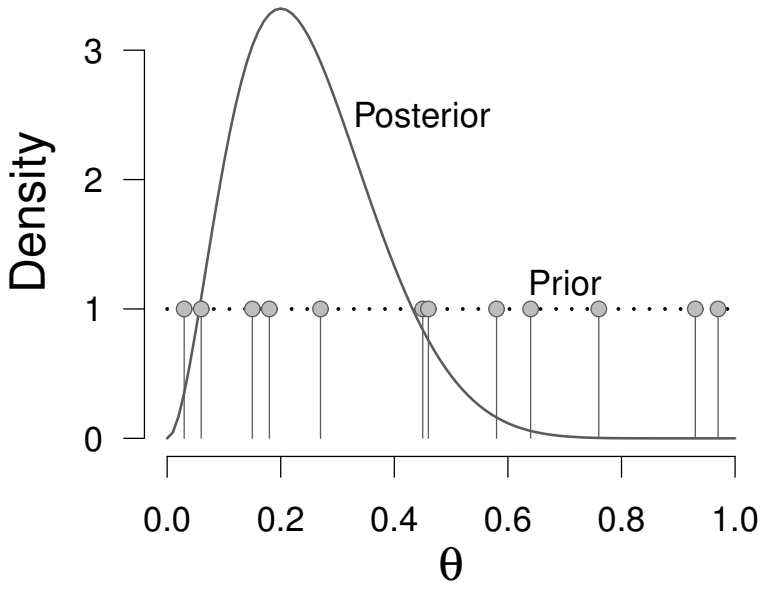

Figure 8.2: Illustration of the naive Monte Carlo estimator for the beta-binomial example. The dotted line represents the prior distribution and the solid line represents the posterior distribution that was obtained after having observed 2 correct responses out of 10 trials. The gray dots represent the 12 samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\}$ randomly drawn from the Beta$(1, 1)$ prior distribution.

Following Equation 8.6, the next step is to calculate the likelihood (Equation 8.5) for each $\tilde{\theta}_i$, and then to average all obtained likelihood values. This yields the naive Monte Carlo estimate of the marginal likelihood:

$$\begin{aligned}\hat{p}_1(y) &= \frac{1}{12}\sum_{i=1}^{12} p(k = 2 \mid n = 10, \tilde{\theta}_i) = \frac{1}{12}\sum_{i=1}^{12} \binom{n}{k} (\tilde{\theta}_i)^k (1 - \tilde{\theta}_i)^{n-k} \\ &= \frac{1}{12}\binom{10}{2}\left(0.58^2(1 - 0.58)^8 + \ldots + 0.15^2(1 - 0.15)^8\right) \\ &= 0.0945,\end{aligned}$$

where we use $\hat{p}_1(y)$ to refer to $\hat{p}_1(k = 2 \mid n = 10)$—a notation that we adopt throughout this article.

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. However, the estimator is unstable if the posterior distribution is peaked relative to the prior (Gamerman & Lopes, 2006; Ntzoufras, 2009). In such a situation, most of the sampled values for θ 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 dominate estimates of the marginal likelihood. Consequently, the variance of the estimator is increased (Newton & Raftery, 1994; Pajor, 2016).[2]

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 Equation 8.4 is large. This is realized 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 θ that result in high likelihood values are sampled most frequently, whereas values for θ with low likelihood values are sampled only rarely.

To derive the importance sampling estimator, Equation 8.4 is used as starting point which is then extended by the importance density $g_{IS}(\theta)$:

$$\begin{aligned}p(y) &= \int p(y \mid \theta)\, p(\theta)\, \mathrm{d}\theta = \int p(y \mid \theta)\, p(\theta)\, \frac{g_{IS}(\theta)}{g_{IS}(\theta)}\, \mathrm{d}\theta = \int \frac{p(y \mid \theta)\, p(\theta)}{g_{IS}(\theta)}\, g_{IS}(\theta)\, \mathrm{d}\theta \\ &= \mathbb{E}_{g_{IS}(\theta)}\left(\frac{p(y \mid \theta)\, p(\theta)}{g_{IS}(\theta)}\right).\end{aligned}$$

This yields the importance sampling estimator $\hat{p}_2(y)$:

$$\hat{p}_2(y) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \frac{p(y \mid \tilde{\theta}_i)\, p(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}}_{\text{average adjusted likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim g_{IS}(\theta).}_{\substack{\text{samples from the} \\ \text{importance density}}} \tag{8.7}$$

A suitable importance density 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 have little impact on the estimator (Neal, 2001).

[2]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.

Running example

To obtain the importance sampling estimate of the marginal likelihood in our running example, we first need to choose an importance density $g_{IS}(\theta)$. An importance density that fulfills the four above mentioned desiderata is a mixture between a beta density that provides the best fit to the posterior distribution and a uniform density across the range of θ (Vandekerckhove et al., 2015). The relative impact of the uniform density is quantified by a mixture weight γ that ranges between 0 and 1. The larger γ, the higher the influence of the uniform density resulting in a less peaked distribution with thick tails. If $\gamma = 1$, the beta mixture density simplifies to the uniform distribution on $[0, 1]$;[3] and if $\gamma = 0$, the beta mixture density simplifies to the beta density that provides the best fit to the posterior distribution.

In our specific example, we already know that the Beta$(3, 9)$ density is the beta density that provides the best fit to the posterior distribution because this is the analytic expression of the posterior distribution. However, to demonstrate the general case, we show how we can find the beta distribution with the best fit to the posterior distribution using the method of moments. This method requires draws from the Beta$(3, 9)$ posterior distribution. We obtain:

$$\{\theta^*_1, \theta^*_2, \ldots, \theta^*_{12}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\},$$

which yields a mean of $\bar{\theta}^* = 0.232$ and a variance of $s^2_{\theta*} = 0.014$. Note that here we use θ^*_i to refer to the i^{th} sample from the posterior distribution to distinguish it from the previously used $\tilde{\theta}_i$—the i^{th} sample from a distribution other than the posterior distribution, such as a prior distribution or an importance density.

Knowing that, if $X \sim \text{Beta}(\alpha, \beta)$, then $\mathbb{E}(X) = \alpha/(\alpha + \beta)$ and $V(X) = \alpha\beta/\left[(\alpha + \beta)^2(\alpha + \beta + 1)\right]$, we obtain the following method of moment estimates for α and β:

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

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

Using a mixture weight on the uniform component of $\gamma = 0.30$ –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: $\gamma \times \text{Beta}(\theta \mid 1, 1) + (1 - \gamma) \times \text{Beta}(\theta \mid \hat{\alpha}, \hat{\beta}) = .3 + .7\ \text{Beta}(\theta \mid 2.673, 8.865)$. This importance density is represented by the dashed line in Figure 8.3. 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—a crucial criterion of the importance density in importance sampling.

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

Drawing $N = 12$ samples for θ from our beta mixture importance density results in:

[3] In our running example, the importance sampling estimator then reduces to the naive Monte Carlo estimator.

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

These samples are represented by the gray dots in Figure 8.3.

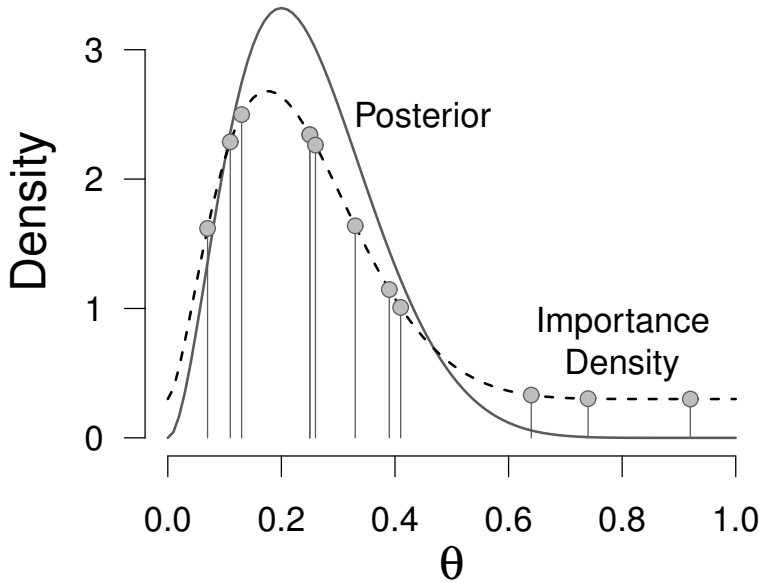

Figure 8.3: Illustration of the importance sampling estimator for the beta-binomial model. The dashed line represents our beta mixture importance density and the solid gray line represents the posterior distribution that was obtained after having observed 2 correct responses out of 10 trials. The gray dots represent the 12 samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{12}\}$ randomly drawn from our beta mixture importance density.

The final step is to compute the average adjusted likelihood for the 12 samples using Equation 8.7. This yields the importance sampling estimate of the marginal likelihood as:

$$\hat{p}_2(y) = \frac{1}{12}\sum_{i=1}^{12} \frac{p(k=2 \mid n=10, \tilde{\theta}_i)\; p(\tilde{\theta}_i)}{.3 + .7\; \text{Beta}(\tilde{\theta}_i \mid 2.673, 8.865)}$$

$$= \frac{1}{12}\left(\frac{\binom{10}{2} 0.11^2(1-0.11)^8 \times 1}{.3 + .7\; \text{Beta}(0.11 \mid 2.673, 8.865)} + \ldots + \frac{\binom{10}{2} 0.92^2(1-0.92)^8 \times 1}{.3 + .7\; \text{Beta}(0.92 \mid 2.673, 8.865)} \right)$$
$$= \frac{1}{12}\binom{10}{2}(0.0021 + \ldots + 4.7 \times 10^{-9})$$
$$= 0.0829.$$

Method 3: The Generalized Harmonic Mean Estimator of the Marginal Likelihood

Just as the importance sampling estimator, the generalized harmonic mean estimator focuses on regions of the parameter space where the integrand of Equation 8.4 is large by using an importance density $g_{IS}(\theta)$ (Gelfand & Dey, 1994).[4] However, in contrast to the importance sampling estimator, the generalized harmonic mean estimator requires an importance density with thinner tails, and it only requires samples from one density (i.e., the posterior density).

To derive the generalized harmonic mean estimator, also known as reciprocal importance sampling estimator (Frühwirth-Schnatter, 2004), we use the following identity:

$$\frac{1}{p(y)} = \int \frac{1}{p(y)}\, g_{IS}(\theta)\; \mathrm{d}\theta = \int \frac{p(\theta \mid y)}{p(y \mid \theta)p(\theta)}\, g_{IS}(\theta)\; \mathrm{d}\theta = \int \frac{g_{IS}(\theta)}{p(y \mid \theta)p(\theta)}\, p(\theta \mid y)\; \mathrm{d}\theta$$
$$= \mathbb{E}_{\text{post}}\left(\frac{g_{IS}(\theta)}{p(y \mid \theta)\; p(\theta)} \right).$$

Rewriting results in:

$$p(y) = \left(\mathbb{E}_{\text{post}}\left(\frac{g_{IS}(\theta)}{p(y \mid \theta)p(\theta)} \right) \right)^{-1},$$

which is used to define the generalized harmonic mean estimator $\hat{p}_3(y)$ (Gelfand & Dey, 1994) as follows:

$$\hat{p}_3(y) = \left(\frac{1}{N}\sum_{j=1}^{N} \frac{\overbrace{g_{IS}(\theta_j^*)}^{\text{importance density}}}{\underbrace{p(y \mid \theta_j^*)}_{\text{likelihood}}\; \underbrace{p(\theta_j^*)}_{\text{prior}}} \right)^{-1} \quad , \quad \underbrace{\theta_j^* \sim p(\theta \mid y)\; .}_{\substack{\text{samples from the} \\ \text{posterior distribution}}} \tag{8.8}$$

[4] Note that the generalized harmonic mean estimator is a more stable version of the harmonic mean estimator (Newton & Raftery, 1994). A problem of the harmonic mean estimator is that it is dominated by samples that have small likelihood values.

Note that the generalized harmonic mean estimator –in contrast to the importance sampling estimator– evaluates samples from the posterior distribution.

As for the importance sampling estimator, the importance density for the generalized harmonic mean estimator should (1) be easy to evaluate; (2) have the same domain as the posterior distribution; (3) closely resemble the posterior distribution; but (4) should have thinner tails than the posterior distribution—a crucial difference to the importance density of the importance sampling estimator (Newton & Raftery, 1994; DiCiccio et al., 1997).

Running example

To obtain the generalized harmonic mean estimate of the marginal likelihood in our running example, we need to choose a suitable importance density. In our running example, an importance density that fulfills the four above mentioned desiderata can be obtained by following four steps: First, we draw $N = 12$ samples from the posterior distribution. Reusing the samples from the last section, we obtain:

$$\{\theta_1^*, \theta_2^*, \ldots, \theta_{12}^*\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\},$$

Second, we probit-tranform all posterior samples (i.e., $\xi_j^* = \Phi^{-1}(\theta_j^*),\ j = \{1, 2, \ldots, 12\}$). The result of this transformation is that the samples range across the entire real line instead of the $(0, 1)$ interval only. We obtain:

$$\begin{aligned}\{\xi_1^*, \xi_2^*, \ldots, \xi_{12}^*\} = \{&-0.77, -0.99, -1.34, -0.39, -1.55, -0.61, -0.64, -0.23, -0.84, -0.18,\\ &-0.81, -1.17\}.\end{aligned}$$

These probit-transformed samples are represented by the gray dots in Figure 8.4.

Third, we search for the normal distribution that provides the best fit to the probit-transformed posterior samples ξ_j^*. Using the method of moments, we obtain as estimates $\hat{\mu} = -0.793$ and $\hat{\sigma} = 0.423$. Note that the choice of a normal importance density justifies step 2; the probit 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 choice of a smaller standard deviation ensures thinner tails of the importance density than of 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 Figure 8.4 (dashed line) together with the probit-transformed posterior distribution (solid line).

The generalized harmonic mean estimate can now be obtained using either the original posterior samples θ_j^* or the probit-transformed samples ξ_j^*. Here we use the latter ones (see also Overstall & Forster, 2010). Incorporating our specific importance density and a correction for having used the probit-transformation, Equation 8.8 becomes:[5]

[5] A detailed explanation is provided in Appendix F. Note that using the original posterior samples θ_j^* would involve transforming the importance density that ranges across the real line to the $(0, 1)$ interval.

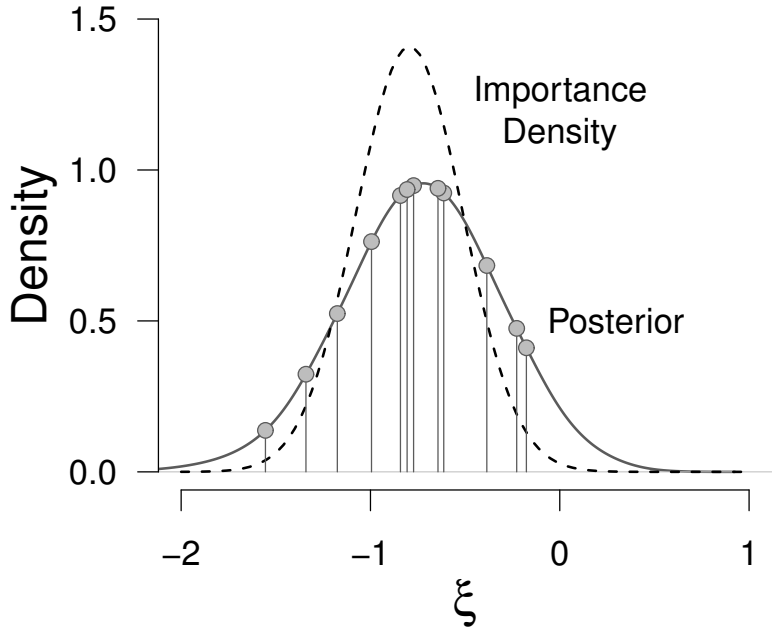

Figure 8.4: Illustration of the generalized harmonic mean estimator for the beta-binomial model. The solid line represents the probit-transformed Beta$(3, 9)$ posterior distribution that was obtained after having observed 2 correct responses out of 10 trials, and the dashed line represents the importance density $\mathrm{N}\left(\xi;\ \mu = -0.793, \sigma = 0.423/1.5\right)$. The gray dots represent the 12 probit-transformed samples $\{\xi_1^*, \xi_2^*, \ldots, \xi_{12}^*\}$ randomly drawn from the Beta$(3, 9)$ posterior distribution.

$$\hat{p}_3(y) = \left(\frac{1}{N} \sum_{j=1}^{N} \frac{\overbrace{\frac{1}{\hat{\sigma}} \phi \left(\frac{\xi_j^* - \hat{\mu}}{\hat{\sigma}} \right)}^{\text{importance density}}}{\underbrace{p\left(y \mid \Phi\left(\xi_j^*\right)\right)}_{\text{likelihood}} \underbrace{p\left(\Phi\left(\xi_j^*\right)\right) \phi\left(\xi_j^*\right)}_{\text{prior}}} \right)^{-1}, \quad \underbrace{\xi_j^* = \Phi^{-1}(\theta_j^*) \text{ and } \theta_j^* \sim p(\theta \mid y)\,.}_{\substack{\text{probit-transformed samples} \\ \text{from the posterior distribution}}} \tag{8.9}$$

Equation 8.9 correctly suggests that the uniform prior on θ translates to a standard normal prior ϕ on ξ because $p(\Phi(\xi)) = 1\ \forall \xi$. For our beta-binomial model, we now obtain the generalized harmonic mean estimate of the marginal likelihood as:

$$
\begin{aligned}
\hat{p}_3(y) &= \left(\frac{1}{12} \sum_{j=1}^{12} \frac{\frac{1}{0.423/1.5}\, \phi\left(\frac{\xi_j^* + 0.793}{0.423/1.5}\right)}{p(k = 2 \mid n = 10, \Phi(\xi_j^*))\; \phi(\xi_j^*)} \right)^{-1} \\
&= \left(\frac{1}{12} \left(\frac{\frac{1}{0.423/1.5}\, \phi\left(\frac{-0.77 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.22^2 (1 - 0.22)^8\; \phi(-0.77)} + \ldots + \frac{\frac{1}{0.423/1.5}\, \phi\left(\frac{-1.17 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.12^2 (1 - 0.12)^8\; \phi(-1.17)} \right) \right)^{-1} \\
&= \left(\frac{1}{12}\, \frac{1}{\binom{10}{2}}\; (716.81 + \ldots + 556.38) \right)^{-1} \\
&= 0.092.
\end{aligned}
$$

Method 4: The Bridge Sampling Estimator of the Marginal Likelihood

Both the importance sampling estimator and the generalized harmonic mean estimator are sensitive to the tail behavior of the importance density relative to the posterior distribution. The bridge sampler, on the other hand, is robust to the tail behavior of the posterior and therefore produces more stable estimates (DiCiccio et al., 1997; Frühwirth-Schnatter, 2004; Gelman & Meng, 1998; Meng & Wong, 1996). Another distinctive feature of the bridge sampling estimator is that it relies on draws from two distributions to approximate the marginal likelihood: the posterior distribution $p(\theta|y)$ and a proposal distribution $g(\theta)$.

The proposal distribution $g(\theta)$ is conceptually similar to an importance density and should resemble the posterior distribution. The distance between the proposal distribution and the posterior distribution is reduced by a so-called bridge function $h(\theta)$. Due to this function, the bridge sampler is relatively robust to the choice of the proposal distribution

Originally, bridge sampling was developed to directly estimate the Bayes factor, that is, the ratio of the marginal likelihoods of two models $\mathcal{M}_1$ and $\mathcal{M}_2$ (e.g., Jeffreys, 1961; Kass & Raftery, 1995):

$$
\underbrace{\frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_2)}}_{\substack{\text{Bayes} \\ \text{factor}}} = \underbrace{\frac{\int p(y \mid \theta, \mathcal{M}_1)\; p(\theta \mid \mathcal{M}_1)\; \mathrm{d}\theta}{\int p(y \mid \theta, \mathcal{M}_2)\; p(\theta \mid \mathcal{M}_2)\; \mathrm{d}\theta}}_{\substack{\text{ratio of} \\ \text{marginal likelihoods}}},
$$

which is equivalent to:

$$
\frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_2)} = \frac{\int p(y \mid \theta, \mathcal{M}_1)\; p(\theta \mid \mathcal{M}_1)\; \underbrace{h(\theta)}_{\substack{\text{bridge} \\ \text{function}}}\; \underbrace{g(\theta)}_{\substack{\text{proposal} \\ \text{distribution}}}\; \mathrm{d}\theta}{\int p(y \mid \theta, \mathcal{M}_2)\; p(\theta \mid \mathcal{M}_2)\; h(\theta)\; g(\theta)\; \mathrm{d}\theta}. \tag{8.10}
$$

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 & Forster, 2010). In the remainder, we therefore suppress the "model" in the conditioning part of the probability statements. Equation 8.10 then simplifies to:

$$1 = \frac{\int p(y \mid \theta)\, p(\theta)\, h(\theta)\, g(\theta)\, \mathrm{d}\theta}{\int p(y \mid \theta)\, p(\theta)\, h(\theta)\, g(\theta)\, \mathrm{d}\theta}\,. \tag{8.11}$$

Multiplying the left side of Equation 8.11 with the marginal likelihood $p(y)$, and dividing the right side by its reciprocal results in:

$$p(y) = \frac{\int p(y \mid \theta)\, p(\theta)\, h(\theta)\, g(\theta)\, \mathrm{d}\theta}{\int \frac{p(y \mid \theta)\, p(\theta)}{p(y)}\, h(\theta)\, g(\theta)\, \mathrm{d}\theta} = \frac{\int p(y \mid \theta)\, p(\theta)\, h(\theta) \quad \overbrace{g(\theta)}^{\substack{\text{proposal}\\ \text{distribution}}} \quad \mathrm{d}\theta}{\int h(\theta) g(\theta) \quad \underbrace{p(\theta \mid y)}_{\substack{\text{posterior}\\ \text{distribution}}} \quad \mathrm{d}\theta}$$

$$= \frac{\mathbb{E}_{g(\theta)}(p(y \mid \theta)\, p(\theta)\, h(\theta))}{\mathbb{E}_{\text{post}}(h(\theta)\, g(\theta))}\,.$$

The marginal likelihood can now be approximated using:

$$\hat{p}(y) = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} p(y \mid \tilde{\theta}_i)\, p(\tilde{\theta}_i)\, h(\tilde{\theta}_i)}{\frac{1}{N_1}\sum_{j=1}^{N_1} h(\theta_j^*)\, g(\theta_j^*)}, \qquad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\substack{\text{samples from the}\\ \text{proposal distribution}}} \quad , \quad \underbrace{\theta_j^* \sim p(\theta \mid y)\,.}_{\substack{\text{samples from the}\\ \text{posterior distribution}}} \tag{8.12}$$

Equation 8.12 shows that, in order to compute $\hat{p}(y)$, we need N_2 samples from the proposal distribution $g(\theta)$; for each sample $\tilde{\theta}_i$, $i = \{1, 2, \ldots, N_2\}$, we need to compute $p(y \mid \tilde{\theta}_i)\, p(\tilde{\theta}_i)\, h(\tilde{\theta}_i)$. In addition, we need N_1 samples from the posterior distribution $p(\theta \mid y)$; for each sample θ_j^*, $j = \{1, 2, \ldots, N_1\}$, we need to compute $h(\theta_j^*)\, g(\theta_j^*)$. With these ingredients we can obtain an estimate of the marginal likelihood according to Equation 8.12. However, before we can apply Equation 8.12 to our running example, we have to discuss how we can obtain a suitable bridge function.

Choosing the optimal bridge function

The purpose of the bridge function $h(\theta)$ is to reduce the distance between the proposal distribution $g(\theta)$ and the posterior distribution $p(\theta|y)$ by linking them together. Hence, the optimal bridge function should represent this link by supporting both distributions. Meng and Wong (1996, p. 837) showed that the Monte Carlo error is minimized if the bridge function is defined as:

$$h(\theta) = C \cdot \frac{1}{s_1 p(y \mid \theta) p(\theta) + s_2 p(y) g(\theta)}\,, \tag{8.13}$$

where $s_1 = \frac{N_1}{N_2+N_1}$, $s_2 = \frac{N_2}{N_2+N_1}$, and C a constant. Equation 8.13 shows that the optimal choice of $h(\theta)$ depends on the marginal likelihood $p(y)$ which is the very entity we want to approximate. 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 (Equation 8.13)

in Equation 8.12 (Meng & Wong, 1996). The formula to approximate the marginal likelihood on iteration $t+1$ is then specified as follows:

$$\hat{p}(y)^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{p(y \mid \tilde{\theta}_i)p(\tilde{\theta}_i)}{s_1 p(y \mid \tilde{\theta}_i)p(\tilde{\theta}_i) + s_2 \hat{p}(y)^{(t)} g(\tilde{\theta}_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(y \mid \theta_j^*)p(\theta_j^*) + s_2 \hat{p}(y)^{(t)} g(\theta_j^*)}}, \tag{8.14}$$

$$\underbrace{\tilde{\theta}_i \sim g(\theta)}_{\substack{\text{samples from the} \\ \text{proposal distribution}}} \quad , \quad \underbrace{\theta_j^* \sim p(\theta \mid y)}_{\substack{\text{samples from the} \\ \text{posterior distribution}}} \quad ,$$

where $\hat{p}(y)^{(t)}$ denotes the estimate of the marginal likelihood on iteration t of the iterative scheme. Extending the numerator of the right side of the Equation 8.14 with $\frac{1/g(\tilde{\theta}_i)}{1/g(\tilde{\theta}_i)}$, and the denominator with $\frac{1/g(\theta_j^*)}{1/g(\theta_j^*)}$, and subsequently defining $l_{1,j} := \frac{p(y \mid \theta_j^*)p(\theta_j^*)}{g(\theta_j^*)}$ and $l_{2,i} := \frac{p(y \mid \tilde{\theta}_i)p(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}$, we obtain the formula for the iterative scheme of the bridge sampling estimator $\hat{p}_4(y)^{(t+1)}$ at iteration $t+1$ (Meng & Wong, 1996, p. 837).

$$\begin{aligned}
\hat{p}_4(y)^{(t+1)} &= \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{p(y \mid \tilde{\theta}_i)p(\tilde{\theta}_i)}{s_1 p(y \mid \tilde{\theta}_i)p(\tilde{\theta}_i) + s_2 \hat{p}_4(y)^{(t)} g(\tilde{\theta}_i)} \frac{1/g(\tilde{\theta}_i)}{1/g(\tilde{\theta}_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(y \mid \theta_j^*)p(\theta_j^*) + s_2 \hat{p}_4(y)^{(t)} g(\theta_j^*)} \frac{1/g(\theta_j^*)}{1/g(\theta_j^*)}} \\
&= \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_4(y)^{(t)}}}, \quad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\substack{\text{samples from the} \\ \text{proposal distribution}}} \quad , \quad \underbrace{\theta_j^* \sim p(\theta \mid y)}_{\substack{\text{samples from the} \\ \text{posterior distribution}}} \quad .
\end{aligned} \tag{8.15}$$

Equation 8.15 suggests that, in order to obtain the bridge sampling estimate of the marginal likelihood, a number of requirements need to be fulfilled. First, we need N_2 samples from the proposal distribution $g(\theta)$ and N_1 samples from the posterior distribution $p(\theta|y)$. Second, for all N_2 samples from the proposal distribution, we have to evaluate $l_{2,i}$. This involves obtaining the value of the unnormalized posterior (i.e., the product of the likelihood times the prior) and of the proposal distribution for all samples. Third, we have to evaluate $l_{1,j}$ for all N_1 samples from the posterior distribution. This is analogous to evaluating $l_{2,i}$. Fourth, we have to determine the constants s_1 and s_2 that only depend on N_1 and N_2. Fifth, we need an initial guess of the marginal likelihood $\hat{p}_4(y)$. Since some of these five requirements can be obtained easier than others, we will point to possible challenges.

A first challenge is that using a suitable proposal distribution may involve transforming the posterior samples. Consequently, as in the running example of the generalized harmonic mean estimator, we have to determine how the transformation affects the definition of the bridge sampling estimator for the marginal likelihood (Equation 8.15).

A second challenge is how to use the N_1 posterior samples for fitting the proposal distribution and for evaluating $l_{1,j}$ most efficiently. One option is to use all N_1 samples for both fitting the proposal distribution and for evaluating $l_{1,j}$. However, Overstall and Forster (2010) showed that it is more efficient to divide the posterior samples in two parts; the first part is used to obtain the best-fitting proposal distribution, and the second part is used for the function evaluations. Throughout this tutorial, we use two equally large parts. In the remainder we therefore state that we draw $2N_1$ samples from the posterior distribution. In addition, we obtain the first part by selecting only those posterior samples with even index numbers; posterior samples with odd index numbers constitute the second part (for an alternative way of splitting the posterior samples see Overstall & Forster, 2010).

Running example

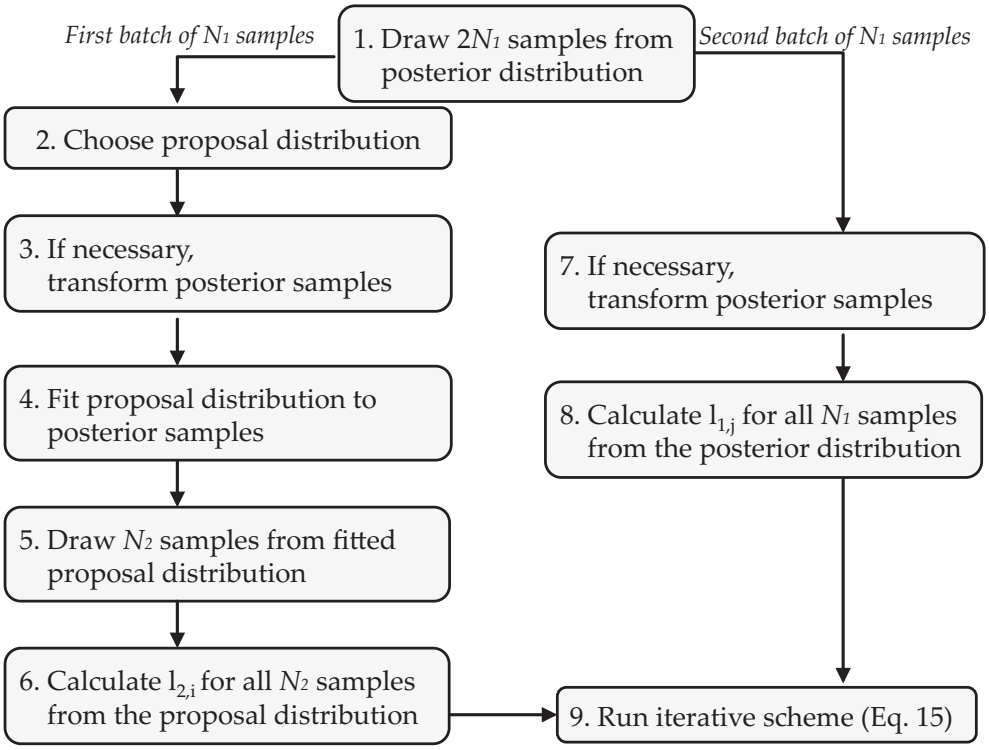

Figure 8.5: Schematic illustration of the steps involved in obtaining the bridge sampling estimate of the marginal likelihood.

To obtain the bridge sampling estimate of the marginal likelihood in the beta-binomial example, we follow the eight steps illustrated in Figure 8.5:

1. *We draw* $2N_1 = 24$ *samples from the Beta*$(3, 9)$ *posterior distribution for* θ.
   We obtain the following sample of 24 values:

$$\begin{aligned}\{\theta_1^*, \theta_2^*, \ldots, \theta_{24}^*\} = &\{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12,\\ &0.15, 0.21, 0.24, 0.18, 0.12, 0.22, 0.15, 0.22, 0.23, 0.26, 0.29, 0.28\},\end{aligned}$$

   Note that the first 12 samples equal the ones used in the last section, whereas the last 12 samples were obtained from drawing again 12 values from the Beta$(3, 9)$ posterior distribution for θ.

2. *We choose a proposal distribution.*
   Here we opt for an approach that can be easily generalized to models with multiple parameters and select a normal distribution as the proposal distribution $g(\theta)$.[6]

3. *We transform the first batch of N_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, $\xi_j^* = \Phi^{-1}(\theta_j^*)$ with $j \in \{2, 4, \ldots, 24\}$. We obtain:

$$\begin{aligned}\{\xi_2^*, \xi_4^*, \ldots, \xi_{24}^*\} =& \{-0.99, -0.39, -0.61, -0.23, -0.18, -1.17, -0.81, -0.92, -0.77, -0.77, \\ & -0.64, -0.58\}.\end{aligned}$$

4. *We fit the proposal distribution to the first batch of N_1 probit-transformed posterior samples.*
   We use the method of moment estimates $\hat{\mu} = -0.672$ and $\hat{\sigma} = 0.298$ from the first batch of N_1 probit-transformed posterior samples to obtain our proposal distribution $g(\xi \mid \mu = -0.672, \sigma = 0.298) = \frac{1}{0.298}\,\phi\left(\frac{\xi + 0.672}{0.298}\right)$.

5. *We draw N_2 samples from the proposal distribution.*
   We obtain:

$$\begin{aligned}\{\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_{12}\} =& \{-0.90, -0.55, -1.16, -0.53, -0.45, -0.60, -0.63, -0.48, -0.69, \\ & -1.20 - 0.65 - 0.79\}.\end{aligned}$$

6. *We calculate $l_{2,i}$ for all N_2 samples from the proposal distribution.*
   This step involves assessing the value of the unnormalized posterior and the proposal distribution for all N_2 samples from the proposal distribution. As in the running example for the generalized harmonic mean estimator, we obtain the unnormalized posterior as: $p\left(y \mid \Phi\left(\tilde{\xi}_i\right)\right) p\left(\Phi\left(\tilde{\xi}_i\right)\right) \phi\left(\tilde{\xi}_i\right)$, where the extra term $\phi\left(\tilde{\xi}_i\right)$ comes from using the change-of-variable method (see running example for the generalized harmonic mean estimator and Appendix F for details). Thus, as in the case of the generalized harmonic mean estimator, the uniform prior on θ translates to a standard normal prior on ξ. The values of the proposal distribution can easily be obtained (for example using the R software).

7. *We transform the second batch of the N_1 posterior samples.*
   As in step 2, we use the probit transformation and obtain:

$$\begin{aligned}\{\xi_1^*, \xi_3^*, \ldots, \xi_{23}^*\} =& \{-0.77, -1.34, -1.55, -0.64, -0.84, -0.81, -1.04, -0.71, -1.17, -1.04, \\ & -0.74, -0.55\}.\end{aligned}$$

8. *We calculate $l_{1,j}$ for the second batch of N_1 probit-transformed samples from the posterior distribution.*
   This is analogous to step 6.

[6] There exist several candidates for the proposal distribution. Alternative proposal distributions are, for example, the importance density that we used for the importance sampling estimator (i.e., a beta mixture importance density) or for the generalized harmonic mean estimator (i.e., a normal distribution with best fit to the posterior distribution, but with a smaller standard deviation), or the analytically derived Beta$(3, 9)$ posterior distribution.

9. *We run the iterative scheme (Equation 8.15) until our predefined tolerance criterion is reached.*
   As tolerance criterion we choose $|\hat{p}_4(y)^{(t+1)} - \hat{p}_4(y)^{(t)}| \,/\, \hat{p}_4(y)^{(t+1)} \leq 10^{-10}$. This requires an initial guess for the marginal likelihood $\hat{p}_4(y)^{(0)}$ which we set to 0.[7]

The simplicity of the beta-binomial model allows us to calculate the bridge sampling estimate by hand. To determine $\hat{p}_4(y)^{(t+1)}$ according to Equation 8.15, we need to calculate the constants s_1 and s_2. Since $N_1 = N_2$, we obtain: $s_1 = s_2 = N_2/(N_2 + N_1) = 12/(12 + 12) = 0.5$. In addition, we need to calculate $l_{2,i}$ ($i \in \{1, 2, \ldots, 12\}$) for all samples from the proposal distribution, and $l_{1,j}$ ($j \in \{1, 3, \ldots, 23\}$) for the second batch of the probit-transformed samples from the posterior distribution. Here we show how to calculate $l_{2,1}$ and $l_{1,1}$ for the first sample from the proposal distribution and the posterior distribution, respectively:

$$l_{2,1} = \frac{p(k \mid n, \Phi(\tilde{\xi}_1))p(\Phi(\tilde{\xi}_1))\phi(\tilde{\xi}_1)}{g(\tilde{\xi}_1)} = \left(\frac{\binom{10}{2} 0.18^2(1-0.18)^8 \cdot 1 \cdot 0.27}{\frac{1}{0.298}\,\phi\left(\frac{-0.90+0.672}{0.298}\right)} \right) = 0.080.$$

$$l_{1,1} = \frac{p(k \mid n, \Phi(\xi_1^*))p(\Phi(\xi_1^*))\phi(\xi_1^*)}{g(\xi_1^*)} = \frac{\binom{10}{2} 0.22^2(1-0.22)^8 \cdot 1 \cdot 0.30}{\frac{1}{0.298}\,\phi\left(\frac{-0.77+0.672}{0.298}\right)} = 0.070, \text{ and}$$

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

$$\hat{p}_4(y)^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}(y)^{(t)}}}{\frac{1}{N_1}\sum_{j=1}^{2N_1} \frac{1}{s_1 l_{1,2j-1} + s_2 \hat{p}(y)^{(t)}}}$$

$$= \frac{\frac{1}{12}\left(\frac{0.080}{0.5 \cdot 0.080 + 0.5 \cdot \hat{p}_4(y)^{(t)}} + \ldots + \frac{0.071}{0.5 \cdot 0.085 + 0.5 \cdot \hat{p}_4(y)^{(t)}} \right)}{\frac{1}{12}\left(\frac{1}{0.5 \cdot 0.070 + 0.5 \cdot \hat{p}_4(y)^{(t)}} + \ldots + \frac{1}{0.5 \cdot 0.068 + 0.5 \cdot \hat{p}_4(y)^{(t)}} \right)}.$$

Using $t = 0$, we obtain as updated estimate of the marginal likelihood $\hat{p}_4(y)^{(1)} = 0.091$. This iterative procedure has to be repeated until our predefined tolerance criterion is reached. For our running example, this criterion is reached after $t = 6$ iterations. We now obtain the bridge sampling estimate of the marginal likelihood as $\hat{p}_4(y)^{(7)} = 0.0894$.

[7] A better initial guess can be obtained from the generalized 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.

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, most general estimator—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 & Stern, 2005).

Assessing the Accuracy of the Bridge Sampling Estimate

In this section we show how to quantify the accuracy of the bridge sampling estimate. A straightforward approach would be to apply the bridge sampling procedure 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:

$$RE^2 = \frac{\mathbb{E}\left[\left(\hat{p}_4(y) - p(y)\right)^2\right]}{p(y)^2}. \tag{8.16}$$

The derivation of this approximate relative mean-squared error by Frühwirth-Schnatter takes into account that the samples from the proposal distribution $g(\theta)$ are independent, whereas the MCMC samples from the posterior distribution $p(\theta|y)$ may be autocorrelated. The approximate relative mean-squared error is given by:

$$\widehat{RE}^2 = \frac{1}{N_2}\frac{V_{g(\theta)}(f_1(\theta))}{\mathbb{E}^2_{g(\theta)}(f_1(\theta))} + \frac{\rho_{f_2}(0)}{N_1}\frac{V_{\text{post}}(f_2(\theta))}{\mathbb{E}^2_{\text{post}}(f_2(\theta))}, \tag{8.17}$$

where $f_1(\theta) = \frac{p(\theta|y)}{s_1 p(\theta|y) + s_2 g(\theta)}$, $f_2(\theta) = \frac{g(\theta)}{s_1 p(\theta|y) + s_2 g(\theta)}$, $V_{g(\theta)}(f_1(\theta)) = \int \left(f_1(\theta) - \mathbb{E}\left[f_1(\theta)\right]\right)^2 g(\theta)\, \mathrm{d}\theta$ denotes the variance of $f_1(\theta)$ with respect to the proposal distribution $g(\theta)$ (the variance $V_{\text{post}(\theta)}(f_2(\theta))$ is defined analogously), and $\rho_{f_2}(0)$ corresponds to the normalized spectral density of the autocorrelated process $f_2(\theta)$ 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 N_2 samples for $\tilde{\theta}_i$ from the proposal distribution. To evaluate the variance and expected value with respect to the posterior distribution, we use the second batch of N_1 samples θ^*_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 Equation 8.17 is adjusted by the normalized spectral density (for details see Frühwirth-Schnatter, 2004).[8] To evaluate the normalized posterior density which appears in the numerator of $f_1(\theta)$ and the denominator of both $f_1(\theta)$ and $f_2(\theta)$, we use the bridge sampling estimate as normalizing constant.

[8] We estimate the spectral density at frequency zero by fitting an autoregressive model using the `spectrum0.ar()` function as implemented in the `coda` R package (Plummer, Best, Cowles, & Vines, 2006).

Note that, under the assumption that the bridge sampling estimator $\hat{p}_4(y)$ is an unbiased estimator of the marginal likelihood $p(y)$, the square root of the expected relative mean-squared error (Equation 8.16) can be interpreted as the coefficient of variation (i.e., the ratio of the standard deviation and the mean; C. E. Brown, 1998). In the remainder of this article, we report the coefficient of variation to quantify the accuracy of the bridge sampling estimate.

8.2 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 & Stout, 2002) featuring the Expectancy Valence (EV) model—a popular reinforcement learning (RL) 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. (2016). For the hierarchical framework, we compare our results to estimates from the Savage-Dickey density ratio test (Dickey, 1971; Dickey & Lientz, 1970; Wagenmakers et al., 2010; Wetzels, Grasman, & Wagenmakers, 2010).

The Iowa Gambling Task

In this section we describe the IGT (see also Steingroever et al., submitted; Steingroever, Wetzels, Horstmann, et al., 2013; Steingroever, Wetzels, & Wagenmakers, 2013a, 2013b; Steingroever et al., 2014, 2016). 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 (Cavedini, Riboldi, Keller, et al., 2002), obsessive-compulsive disorder (Cavedini, Riboldi, D'Annucci, et al., 2002), 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 maximize 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 100 trials.

A crucial aspect of the IGT is whether and to what extent participants eventually learn to prefer the good decks because only choosing from the good decks maximizes their long-term net outcome. The good decks are typically labeled as decks C and D, whereas the bad decks are labeled as decks A and B. Table 8.1 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: hence, the long-term net outcome is negative. 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

Table 8.1: *Summary of the Payoff Scheme of the Traditional IGT as Developed by Bechara et al. (1994)*

	Deck A	Deck B	Deck C	Deck D
	Bad deck with frequent losses	Bad deck with infrequent losses	Good deck with frequent losses	Good deck with infrequent losses
Reward/trial	100	100	50	50
Number of losses/10 cards	5	1	5	1
Loss/10 cards	−1250	−1250	−250	−250
Net outcome/10 cards	−250	−250	250	250

differ in the frequency of losses: decks A and C yield frequent losses, while decks B and D yield infrequent losses.

The Expectancy Valence Model

In this section, we describe the EV model (see also Steingroever, Wetzels, & Wagenmakers, 2013a; Steingroever et al., 2014, 2016, submitted). Originally proposed by Busemeyer and Stout (2002), the EV model is arguably the most popular model for the IGT (for references see Steingroever, Wetzels, & Wagenmakers, 2013a, and for alternative IGT models see Ahn et al., 2008; Dai, Kerestes, Upton, Busemeyer, & Stout, 2015; Steingroever et al., 2014; Worthy, Pang, & Byrne, 2013; Worthy & Maddox, 2014). The model formalizes participants' performance on the IGT through the interaction of three model parameters that represent distinct psychological processes. The first model assumption is that after choosing a card from deck k, $k \in \{1, 2, 3, 4\}$, on trial t, participants compute a weighted mean of the experienced reward *W(t)* and loss *L(t)* to obtain the utility of deck k on trial t, $u_k(t)$:

$$u_k(t) = (1 - w)W(t) + wL(t).$$

The weight that participants assign to losses relative to rewards is the attention weight parameter w. A small value of w, that is, $w < .5$, is characteristic for decision makers who put more weight on the immediate rewards and can thus be described as reward-seeking, whereas a large value of w, that is, $w > .5$, 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 & Stout, 2002).

The EV model further assumes that decision makers use the utility of deck k on trial t, $u_k(t)$, to update only the expected utility of deck k, $Ev_k(t)$; the expected utilities of the unchosen decks are left unchanged. This updating process is described by the Delta learning rule, also known as the Rescorla-Wagner rule (Rescorla & Wagner, 1972):

$$Ev_k(t) = Ev_k(t - 1) + a(u_k(t) - Ev_k(t - 1)).$$

If the experienced utility $u_k(t)$ is higher than expected, the expected utility of deck k is adjusted upward. If the experienced utility $u_k(t)$ is lower than expected, the expected utility of deck k is adjusted downward. This updating process is influenced by the second model parameter—the updating parameter a. This parameter 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. For all models, we initialized the expectancies

of all decks to zero, $Ev_k(0) = 0$ ($k \in \{1, 2, 3, 4\}$). This setting reflects an absence of prior knowledge about the payoffs of the decks.

In the next step, the model assumes that the expected utilities of each deck guide participants' choices on the next trial $t + 1$. This assumption is formalized by the softmax choice rule, also known as the ratio-of-strength choice rule (Luce, 1959):

$$P[S_k(t+1)] = \frac{e^{\theta(t) \cdot Ev_k(t)}}{\sum_{j=1}^{4} e^{\theta(t) \cdot Ev_j(t)}}.$$

The EV model uses this rule to compute the probability of choosing each deck on each trial. This rule contains a sensitivity parameter θ that indexes the extent to which trial-by-trial choices match the expected deck utilities. Values of θ close to zero indicate random choice behavior (i.e., strong exploration), whereas large values of θ indicate choice behavior that is strongly determined by the expected utilities (i.e., strong exploitation).
The EV model uses a trial-dependent sensitivity parameter $\theta(t)$, which also depends on the final model parameter, response consistency $c \in [-5, 5]$:

$$\theta(t) = (t/10)^c.$$

If c is positive, 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 of the EV model to the range $[-2, 2]$ instead of the proposed range $[-5, 5]$ (Busemeyer & 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, the EV model has three parameters: (1) the attention weight parameter $w \in [0, 1]$, which quantifies the weight of losses over rewards; (2) the updating parameter $a \in [0, 1]$, which determines the memory for past expectancies; and (3) the response consistency parameter $c \in [-2, 2]$, which determines the balance between exploitation and exploration.

Data

We applied bridge sampling to a data set published by Busemeyer and Stout (2002). The data set consists of 30 healthy participants each contributing $T = 100$ IGT card selections (see Busemeyer and Stout for more details on the data sets).[9]

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. Accordingly, we also obtain a marginal likelihood of the EV model for every participant.

[9] Note that we excluded three participants due to incomplete choice data.

Schematic execution of the bridge sampler

To obtain the bridge sampling estimate of the marginal likelihood for each participant, we follow the steps outlined in Figure 8.5.

For each participant s, $s \in \{1, 2, \ldots, 30\}$, we proceed as follows:

1. *For each parameter, we draw $2N_1$ samples from the posterior distribution.*
   Since Steingroever et al. (2016) already fit an individual-level implementation of the EV model separately to the data of each participant in Busemeyer and Stout (2002), we reuse their posterior samples (see Steingroever et al., 2016, for details on the implementation). Therefore, $2N_1$ matches the number of samples obtained from Steingroever et al. (2016) which was at least $5,000$; however, whenever this number of samples was insufficient to ensure convergence of the Hamiltonian Monte Carlo (HMC) chains, Steingroever et al. (2016) repeated the fitting routine with $5,000$ additional samples. Steingroever et al. (2016) confirmed convergence of the HMC chains by reporting that all $\hat{R}$ statistics were below 1.05.

2. *We choose a proposal distribution.*
   We use a multivariate normal distribution as a proposal distribution because it is easy to fit, easy to evaluate, and easy to sample from.

3. *We transform the first batch of N_1 posterior samples.*
   Since we use a multivariate normal distribution as a proposal distribution, we have to transform all posterior samples to the real line using the probit transformation, that is, $\omega^*_{s,j} = \Phi^{-1}(w^*_{s,j})$, $\alpha^*_{s,j} = \Phi^{-1}(a^*_{s,j})$, $\gamma^*_{s,j} = \Phi^{-1}\left((c^*_{s,j} + 2)\,/\,4\right)$, $j = \{2, 4, \ldots, 2N_1\}$.

4. *We fit the proposal distribution to the first batch of N_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 N_1 probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. *We draw N_2 samples from the proposal distribution.*
   We use the R software to randomly draw N_2 samples from the proposal distribution obtained in step 4. We obtain $(\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$ with $i \in \{1, 2, \ldots, N_2\}$.

6. *We calculate $l_{2,i}$ for all N_2 samples from the proposal distribution.*
   This step involves assessing the value of the unnormalized posterior and the proposal distribution for all N_2 samples from the proposal distribution. Before we can assess the value of the unnormalized posterior (i.e., the product of the likelihood and the prior), we have to derive how our transformation in step 3 affects the unnormalized posterior.
   First, we derive how our transformation affects the likelihood. Following the change-of-variable method, we have to obtain the value of the likelihood in $(\Phi(\tilde{\omega}_{s,i}), \Phi(\tilde{\alpha}_{s,i}), \Phi(\tilde{\gamma}_{s,i}))$. Before formalizing the likelihood of the observed choices of participant s, we define the following variables: We define $Ch_s(t)$ as a vector containing the sequence of choices made by participant s up to and including trial t, and $X_s(t)$ as a vector containing the corresponding sequence of net outcomes. We now obtain the following expression for the likelihood of the observed choices of participant s:

$$p(Ch_s(T) \mid \Phi(\tilde{\omega}_{s,i}), \Phi(\tilde{\alpha}_{s,i}), \Phi(\tilde{\gamma}_{s,i}), X_s(T)) = \prod_{t=1}^{T} \prod_{k=1}^{4} Pr(S_k(t+1) \mid Ch_s(t), X_s(t)) \cdot \delta_k(t+1). \tag{8.18}$$

Here T is the total number of trials, $Pr(S_k(t+1) \mid Ch_s(t), X_s(t))$ is the probability of choosing deck k on trial $t+1$ given all previous choices and net outcomes, and $\delta_k(t+1)$ is a dummy variable which is 1 if deck k is chosen on trial $t+1$ and 0 otherwise.
Second, we have to derive how our transformation affects the priors on each EV model parameter to yield priors on the probit-transformed model parameters. Since Steingroever et al. (2016) used independent uniform priors on $[0, 1]$ –a prior setting that requires transforming the c parameter to the $[0, 1]$ interval prior to model fitting– we obtain standard normal priors on the probit-transformed model parameters (see beta-binomial example and Appendix F for an explanation).
Finally, we note that the value of the proposal distribution can be obtained using the R software.

7. *We transform the second batch of N_1 posterior samples.*
   This is analogous to step 2.

8. *We calculate $l_{1,j}$ for the second batch of N_1 probit-transformed samples from the posterior distribution.*
   This is analogous to step 6.

9. *We run the iterative scheme (Equation 8.15) until our predefined tolerance criterion is reached.*
   We use a tolerance criterion and initialization analogous to the running example. Once convergence is reached, we receive an estimate of the marginal likelihood for each participant. We use these estimates to derive the coefficient of variation for each participant. The largest coefficient of variation is 1.94% suggesting that the bridge sampler has low variance.[10]

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., 2016). Figure 8.6 shows the log marginal likelihoods for the 30 participants of Busemeyer and Stout (2002) obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016; y-axis). The two sets of estimates correspond almost perfectly. These results indicate a successful implementation of the bridge sampler. Thus, this section emphasizes 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.

[10] Note that this measure relates to the marginal likelihoods, not to the log marginal likelihoods.

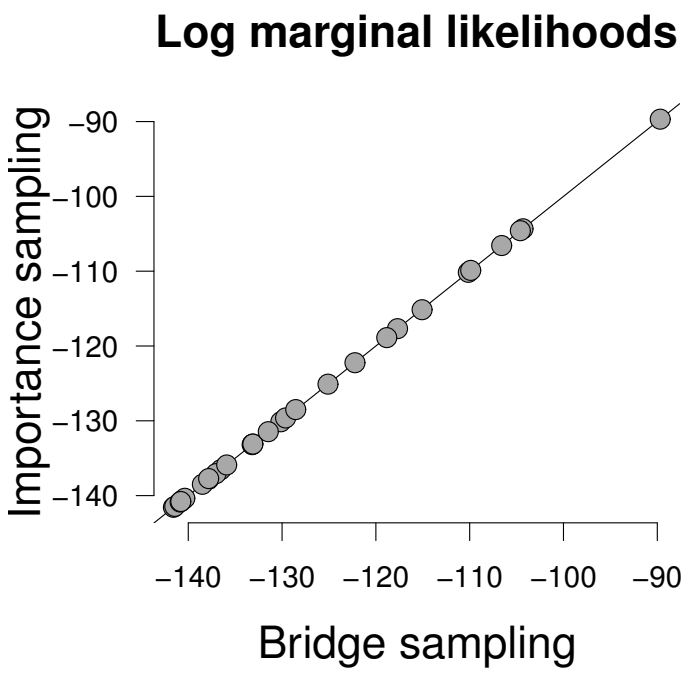

Figure 8.6: Comparison of the log marginal likelihoods obtained with bridge sampling (x-axis) and importance sampling reported by (2016) (2016; y-axis). The main diagonal indicates perfect correspondence between the two methods.

Application of Bridge Sampling to a Hierarchical Implementation of the EV Model

Using the Busemeyer and Stout (2002) data set in this section again, we illustrate how we used bridge sampling to estimate the marginal likelihood of a hierarchical EV model. In the hierarchical case of the EV model, we account for the hierarchical structure of the data and hence incorporate both the differences and the similarities between participants. Accordingly, we assume that the parameters w, a and c from each participant are drawn from a group-level distribution.

Schematic execution of the bridge sampler

To compute the marginal likelihood, we again follow the steps outlined in Figure 8.5, with a few minor modifications.

1. *For each parameter, that is, all individual-level and group-level parameters, we draw* $2N_1 = 60,000$ *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 software JAGS (Plummer, 2003).[11] We assume that, for each participant s, $s \in \{1, 2, \ldots, 30\}$, each probit-transformed individual-level parameter (i.e., $\omega_s = \Phi^{-1}(w_s)$, $\alpha_s = \Phi^{-1}(a_s)$, $\gamma_s = \Phi^{-1}\left((c_s + 2)/4\right)$) is drawn from a group-level normal distribution characterized by a group-level mean and

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

standard deviation parameter. For all group-level mean parameters $\mu_\omega, \mu_\alpha, \mu_\gamma$ we assume a standard normal distribution, and for all group-level standard deviation parameters $\sigma_\omega, \sigma_\alpha, \sigma_\gamma$ a uniform distribution ranging from 0 to 1.5. For a detailed explanation of the hierarchical implementation of the EV model, see Wetzels, Vandekerckhove, et al. (2010).
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.

2. *We choose a proposal distribution.*
We use a multivariate normal distribution as a proposal distribution.

3. *We transform the first batch of N_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, that is, $\omega^*_{s,j} = \Phi^{-1}(w^*_{s,j})$, $\alpha^*_{s,j} = \Phi^{-1}(a^*_{s,j})$, $\gamma^*_{s,j} = \Phi^{-1}\left((c^*_{s,j} + 2) \,/\, 4\right)$, $\tau^*_{\omega,j} = \Phi^{-1}((\sigma^*_{\omega,j}) \,/\, 1.5)$, $\tau^*_{\alpha,j} = \Phi^{-1}((\sigma^*_{\alpha,j}) \,/\, 1.5)$, and $\tau^*_{\gamma,j} = \Phi^{-1}((\sigma^*_{\gamma,j}) \,/\, 1.5)$, $j = \{2, 4, \ldots, 2N_1\}$. 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 N_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 N_1 probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. *We draw N_2 samples from the proposal distribution.*
We use the R software to randomly draw N_2 samples from the proposal distribution obtained in step 4. We obtain $(\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$ and $(\tilde{\mu}_{\omega,i}, \tilde{\tau}_{\omega,i}, \tilde{\mu}_{\alpha,i}, \tilde{\tau}_{\alpha,i}, \tilde{\mu}_{\gamma,i}, \tilde{\tau}_{\gamma,i})$ with $i \in \{1, 2, \ldots, N_2\}$ and $s \in \{1, 2, \ldots, 30\}$.

6. *We calculate $l_{2,i}$ for all N_2 samples from the proposal distribution.*
This step involves assessing the value of the unnormalized posterior and the proposal distribution for all N_2 samples from the proposal distribution. The unnormalized posterior is defined as:
$\left(\prod_{s=1}^{30} p(Ch_s(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{s,j}}), X_s(T))\, p(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{s,i}} \mid \tilde{\boldsymbol{\zeta}}_i)\right)\; p(\tilde{\boldsymbol{\zeta}}_i)$, where $Ch_s(T)$ refers to all choices of subject s, $X_s(T)$ to the corresponding net outcomes, $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{s,i}} = (\tilde{\omega}_{s,j}, \tilde{\alpha}_{s,j}, \gamma_{s,j})$ to the i^{th} sample from the proposal distribution for the individual-level parameters of subject s, and $\tilde{\boldsymbol{\zeta}}_i$ to the i^{th} sample from the proposal distribution for all group-level parameters (e.g., $\tilde{\boldsymbol{\zeta}}_i = (\tilde{\mu}_{\omega,i}, \tilde{\tau}_{\omega,i}, \tilde{\mu}_{\alpha,i}, \tilde{\tau}_{\alpha,i}, \tilde{\mu}_{\gamma,j}, \tilde{\tau}_{\gamma,j})$).
The likelihood function for a given participant is the same as in the individual case. However, for each participant we now have to add besides the prior on the individual-level parameters also the prior on the group-level parameters. The product of the likelihood and the priors gives the unnormalized posterior density (see Appendix F for details).
Finally, we note that the value of the proposal distribution can be obtained using the R software.

7. We follow steps 7 – 9, as outlined for the bridge sampler of the individual-level implementation of the EV model.

Table 8.2: 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 an Percentage

Model	Bayes factor	log marginal likelihood	$CV[\%]$
full model	–	−3801.877	10.53
restricted at $\mu_\omega = -.334$	0.729	−3801.561	14.21
restricted at $\mu_\alpha = -0.604$	0.826	−3801.686	9.99
restricted at $\mu_\gamma = 0.92$	0.710	−3801.535	13.15

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 & Lientz, 1970; Dickey, 1971; for a tutorial, see Wagenmakers et al., 2010). The Savage-Dickey density ratio is a simple method for computing Bayes factors for nested models. We artificially create three nested models by taking the full EV model $\mathcal{M}_f$ in which all parameters are free to vary, and then restrict one of the three group-level mean parameters, that is, μ_ω, μ_α, or μ_γ , to a predefined value. For these values we choose the intersection point of the prior and posterior distribution of each group-level mean parameter. To obtain these intersection points, we fit the full EV model and then use a nonparametric logspline density estimator (Stone, Hansen, Kooperberg, Truong, et al., 1997). The obtained values are presented in Table 8.2. Since we compare the full model to each restricted model, we obtain three Bayes factors.

According to the Savage-Dickey density ratio test, the Bayes factor for the full model versus a specific restricted model $\mathcal{M}_r$ can be obtained by dividing the height of the prior density at the predefined parameter value θ_0 by the height of the posterior at the same location:

$$\text{BF}_{\mathcal{M}_f,\mathcal{M}_r} = \frac{p(y \mid \mathcal{M}_f)}{p(y \mid \mathcal{M}_r)} = \frac{p(\theta = \theta_0 \mid \mathcal{M}_f)}{p(\theta = \theta_0 \mid y, \mathcal{M}_f)}, \tag{8.19}$$

Since we choose θ_0 to be the intersection point of the prior and posterior distribution, $\text{BF}_{\mathcal{M}_f,\mathcal{M}_r}$ equals 1. This Savage-Dickey Bayes factor of 1 indicates that the marginal likelihood under the full model equals the marginal likelihood under the restricted model. Figure 8.7 illustrates the Savage-Dickey Bayes factor comparing the full model to the model assuming μ_α fixed to -0.604.

The computation of the three bridge sampling Bayes factors, on the other hand, works as follows: First, we follow the steps outlined above to obtain the bridge sampling estimate of the full EV model. Second, we obtain the bridge sampling estimate of the marginal likelihood for the three restricted models. This requires adapting the steps outlined above to each of the three restricted models. Lastly, we use Equation 8.19 to obtain the three Bayes factors.

The Bayes factors derived from bridge sampling are reported in Table 8.2. It is evident that Bayes factors derived from bridge sampling closely approximate the Savage-Dickey Bayes factors of 1. 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 8.2.

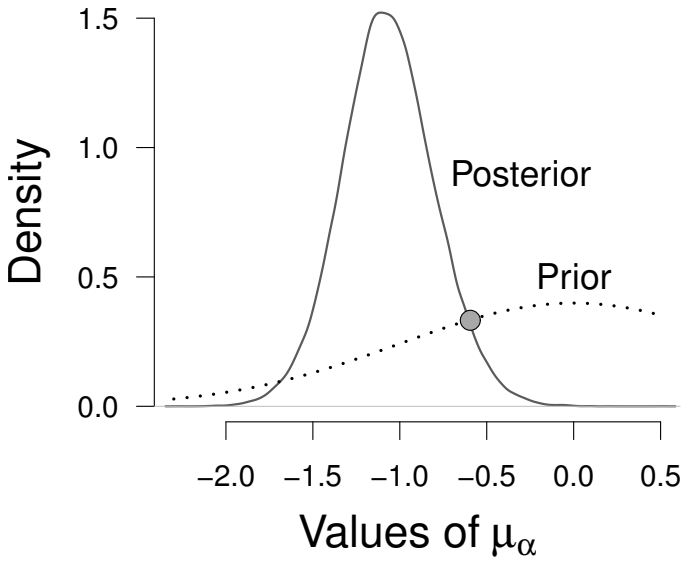

Figure 8.7: Prior and posterior distributions of the group-level mean μ_α in the Busemeyer and Stout (2002) data set. The figure shows the posterior distribution (solid line) and the prior distribution (dotted line). The gray dot indicates the intersection of the prior and the posterior distributions, for which the Savage-Dickey Bayes factor equals 1.

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%.

8.3 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 generalized 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 Expectancy Valence (EV; Busemeyer & Stout, 2002) model—a popular reinforcement-learning model for the Iowa gambling task (IGT; 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 & Wong, 1996).

The bridge sampling estimator is superior to the naive Monte Carlo estimator, the importance

sampling estimator, and the generalized harmonic mean estimator for several reasons. First, Meng and Wong (1996) showed that, among the four estimators discussed in this article, the bridge sampler minimizes the Monte Carlo errors 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 generalized harmonic mean estimator. In particular, the specific shape of the proposal distribution in bridge sampling is not as relevant because the bridge function corrects for the deviances between the posterior and proposal distribution. Third, due to its robustness to the tail behavior of the proposal distribution relative to the posterior distribution, the bridge sampler can be applied 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.). Fourth, 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 unnormalized posterior and 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 optimizes 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 has to 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 & Lientz, 1970; Dickey, 1971) might be a better alternative. 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 & 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 (Gelman & Meng, 1998). This method generalizes bridge sampling by using an infinite number of bridges.

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. Bridge sampling can be profitably applied to a wide range of problems in mathematical psychology involving parameter estimation, model comparison, and Bayesian model averaging.

Acknowledgements

We thank Busemeyer and Stout (2002) for providing the data used in this article.
This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to UB (406-12-125) and to HS (404-10-086), a European Research Council (ERC) grant to EJW (283876), and a Veni grant (451-15-010) from the NWO to DM.