9.1 Importance sampling to estimate tail probabilities (based on Robert and Casella, 2010, Exercise 3.5)

We would like to use importance sampling to compute the probability that a standard Gaussian random variable x takes on a value larger than 5, i.e

$$\mathbb{P}(x > 5) = \int_5^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{9.1}$$

We know that the probability equals

$$\begin{aligned}
\mathbb{P}(x > 5) &= 1 - \int_{-\infty}^5 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx && (9.2)\\
&= 1 - \Phi(5) && (9.3)\\
&\approx 2.87 \cdot 10^{-7} && (9.4)
\end{aligned}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable.

(a) With the indicator function $\mathbb{1}_{x>5}(x)$, which equals one if x is larger than 5 and zero otherwise, we can write $\mathbb{P}(x > 5)$ in form of the expectation

$$\mathbb{P}(x > 5) = \mathbb{E}[\mathbb{1}_{x>5}(x)], \tag{9.5}$$

where the expectation is taken with respect to the density $\mathcal{N}(x; 0, 1)$ of a standard normal random variable,

$$\mathcal{N}(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{9.6}$$

This suggests that we can approximate $\mathbb{P}(x > 5)$ by a Monte Carlo average

$$\mathbb{P}(x > 5) \approx \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x>5}(x_i), \qquad x_i \sim \mathcal{N}(x; 0, 1). \tag{9.7}$$

Explain why this approach does not work well.

(b) Another approach is to use importance sampling with an importance distribution $q(x)$ that is zero for $x < 5$. We can then write $\mathbb{P}(x > 5)$ as

$$\begin{aligned}
\mathbb{P}(x > 5) &= \int_5^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx && (9.8)\\
&= \int_5^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{q(x)}{q(x)} dx && (9.9)\\
&= \mathbb{E}_{q(x)}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{q(x)}\right] && (9.10)
\end{aligned}$$

and estimate $\mathbb{P}(x > 5)$ as a sample average.

We here use an exponential distribution shifted by 5 to the right. It has pdf

$$q(x) = \begin{cases} \exp(-(x-5)) & \text{if } x \geq 5 \\ 0 & \text{otherwise} \end{cases} \tag{9.11}$$

For background on the exponential distribution, see e.g. https://en.wikipedia.org/wiki/Exponential_distribution.

Provide a formula that approximates $\mathbb{P}(x > 5)$ as a sample average over n samples $x_i \sim q(x)$.

9.2 Monte Carlo integration and importance sampling

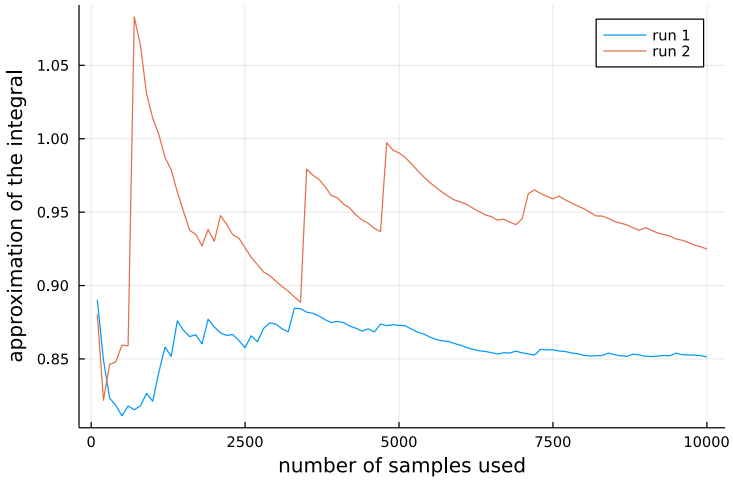
A standard Cauchy distribution has the density function (pdf)

$$p(x) = \frac{1}{\pi} \frac{1}{1+x^2} \tag{9.12}$$

with $x \in \mathbb{R}$. A friend would like to verify that $\int p(x)dx = 1$ but doesn't quite know how to solve the integral analytically. They thus use importance sampling and approximate the integral as

$$\int p(x)dx \approx \frac{1}{n} \sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \qquad x_i \sim q \tag{9.13}$$

where q is the density of the auxiliary/importance distribution. Your friend chooses a standard normal density for q and produces the following figure:



The figure shows two independent runs. In each run, your friend computes the approximation with different sample sizes by subsequently including more and more x_i in the approximation, so that, for example, the approximation with $n = 2000$ shares the first 1000 samples with the approximation that uses $n = 1000$.

Your friend is puzzled that the two runs give rather different results (which are not equal to one), and also that within each run, the estimate very much depends on the sample size. Explain these findings.

9.10 Mixing and convergence of Metropolis-Hasting MCMC

Under weak conditions, an MCMC algorithm is an asymptotically exact inference algorithm, meaning that if it is run forever, it will generate samples that correspond to the desired probability distribution. In this case, the chain is said to converge.

In practice, we want to run the algorithm long enough to be able to approximate the posterior adequately. How long is long enough for the chain to converge varies drastically depending on the algorithm, the hyperparameters (e.g. the variance vari), and the target posterior distribution. It is impossible to determine exactly whether the chain has run long enough, but there exist various diagnostics that can help us determine if we can "trust" the sample-based approximation to the posterior.

A very quick and common way of assessing convergence of the Markov chain is to visually inspect the *trace plots* for each parameter. A trace plot shows how the drawn samples evolve

through time, i.e. they are a time-series of the samples generated by the Markov chain. Figure 9.6 shows examples of trace plots obtained by running the Metropolis Hastings algorithm for different values of the hyperparameters vari and param_init. Ideally, the time series covers the whole domain of the target distribution and it is hard to "see" any structure in it so that predicting values of future samples from the current one is difficult. If so, the samples are likely independent from each other and the chain is said to be well "mixed".

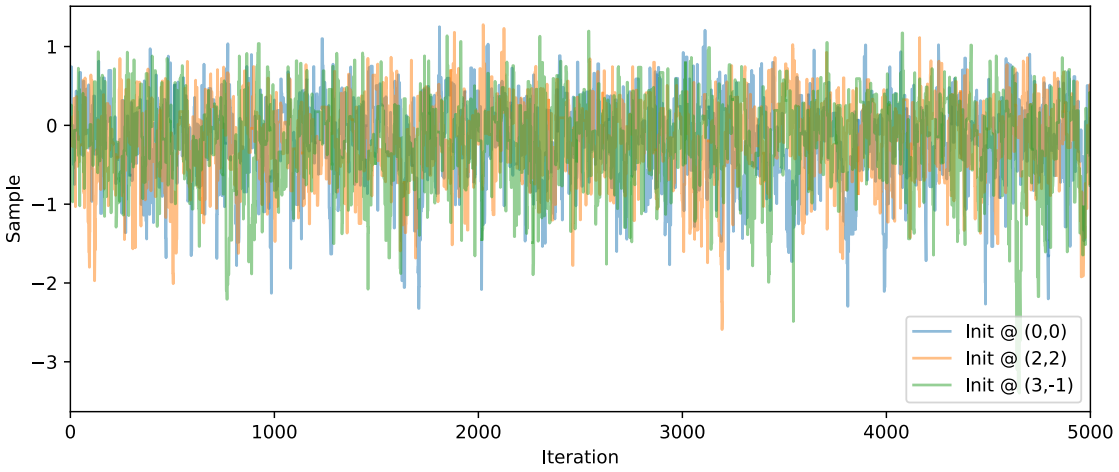
(a) Consider the trace plots in Figure 9.6: Is the variance vari used in Figure 9.6b larger or smaller than the value of vari used in Figure 9.6a? Is vari used in Figure 9.6c larger or smaller than the value used in Figure 9.6a?

In both cases, explain the behaviour of the trace plots in terms of the workings of the Metropolis Hastings algorithm and the effect of the variance vari.

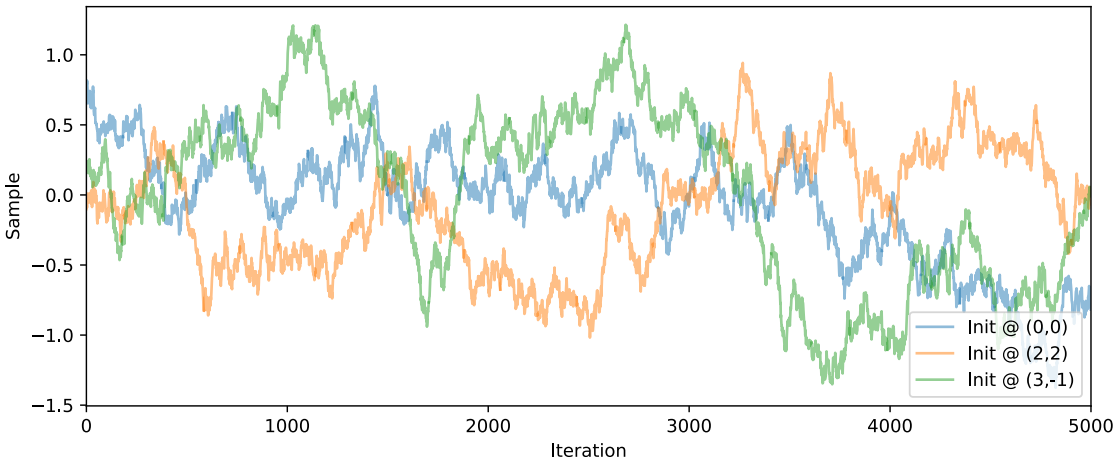
(b) In Metropolis-Hastings, and MCMC in general, any sample depends on the previously generated sample, and hence the algorithm generates samples that are generally statistically dependent. The *effective sample size* of a sequence of dependent samples is the number of independent samples that are, in some sense, equivalent to our number of dependent samples. A definition of the effective sample size (ESS) is

$$\text{ESS} = \frac{S}{1 + 2\sum_{k=1}^{\infty} \rho(k)} \tag{9.34}$$

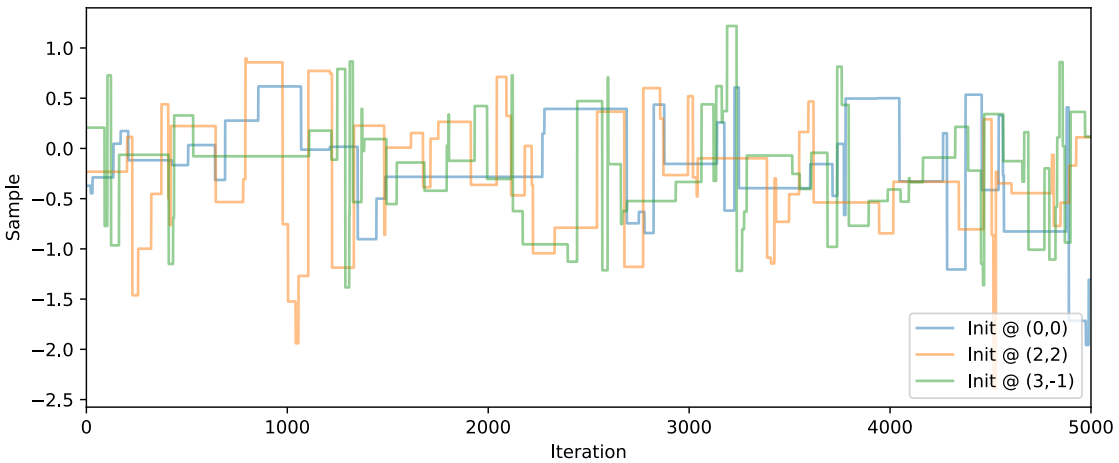
where S is the number of dependent samples drawn and $\rho(k)$ the correlation coefficient between two samples in the Markov chain that are k time points apart. We can



(a) variance vari: 1



(b) Alternative value of vari



(c) Alternative value of vari

Figure 9.6: For Question 9.10(a): Trace plots of the parameter β from Question 9.9 drawn using Metropolis-Hastings with different variances of the proposal distribution.

see that if the samples are strongly correlated, $\sum_{k=1}^{\infty} \rho(k)$ is large and the effective sample size is small. On the other hand, if $\rho(k) = 0$ for all k, the effective sample size is S.

ESS, as defined above, is the number of independent samples which are needed to obtain a sample average that has the same variance as the sample average computed from correlated samples.

To illustrate how correlation between samples is related to a reduction of sample size, consider two pairs of samples (θ_1, θ_2) and (ω_1, ω_2). All variables have variance σ^2 and the same mean μ, but ω_1 and ω_1 are uncorrelated while the covariance matrix for θ_1, θ_2 is $\mathbf{C}$,

$$\mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{9.35}$$

with $\rho > 0$. The variance of the average $\bar{\omega} = 0.5(\omega_1 + \omega_2)$ is

$$\mathbb{V}(\bar{\omega}) = \frac{\sigma^2}{2}, \tag{9.36}$$

where the 2 in the denominator is the sample size.

Derive an equation for the variance of $\bar{\theta} = 0.5(\theta_1 + \theta_2)$ and compute the reduction α of the sample size when working with the correlated (θ_1, θ_2). In other words, derive an equation of α in

$$\mathbb{V}\left(\bar{\theta}\right) = \frac{\sigma^2}{2/\alpha}. \tag{9.37}$$

What is the effective sample size $2/\alpha$ as $\rho \to 1$?

10.1 Mean field variational inference I

Let $\mathcal{L}_{\mathbf{x}}(q)$ be the evidence lower bound for the marginal $p(\mathbf{x})$ of a joint pdf/pmf $p(\mathbf{x}, \mathbf{y})$,

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]. \tag{10.1}$$

Mean field variational inference assumes that the variational distribution $q(\mathbf{y}|\mathbf{x})$ fully factorises, i.e.

$$q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{d} q_i(y_i|\mathbf{x}), \tag{10.2}$$

when $\mathbf{y}$ is d-dimensional. An approach to learning the q_i for each dimension is to update one at a time while keeping the others fixed. We here derive the corresponding update equations.

(a) Show that the evidence lower bound $\mathcal{L}_{\mathbf{x}}(q)$ can be written as

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q_1(y_1|\mathbf{x})}\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right] - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})}\left[\log q_i(y_i|\mathbf{x})\right] \tag{10.3}$$

where $q(\mathbf{y}_{\setminus 1}|\mathbf{x}) = \prod_{i=2}^{d} q_i(y_i|\mathbf{x})$ is the variational distribution without $q_1(y_1|\mathbf{x})$.

(b) Assume that we would like to update $q_1(y_1|\mathbf{x})$ and that the variational marginals of the other dimensions are kept fixed. Show that

$$\underset{q_1(y_1|\mathbf{x})}{\operatorname{argmax}} \mathcal{L}_{\mathbf{x}}(q) = \underset{q_1(y_1|\mathbf{x})}{\operatorname{argmin}} \mathrm{KL}(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) \tag{10.4}$$

with

$$\log \bar{p}(y_1|\mathbf{x}) = \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x},\mathbf{y})\right] + \text{const}, \tag{10.5}$$

where const refers to terms not depending on y_1. That is,

$$\bar{p}(y_1|\mathbf{x}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x},\mathbf{y})\right]\right], \tag{10.6}$$

where Z is the normalising constant. Note that variables $y_2, \ldots, y_d$ are marginalised out due to the expectation with respect to $q(\mathbf{y}_{\setminus 1}|\mathbf{x})$.

(c) Conclude that given $q_i(y_i|\mathbf{x})$, $i = 2, \ldots, d$, the optimal $q_1(y_1|\mathbf{x})$ equals $\bar{p}(y_1|\mathbf{x})$.

This then leads to an iterative updating scheme where we cycle through the different dimensions, each time updating the corresponding marginal variational distribution according to:

$$q_i(y_i|\mathbf{x}) = \bar{p}(y_i|\mathbf{x}), \qquad \bar{p}(y_i|\mathbf{x}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus i}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right]\right] \tag{10.7}$$

where $q(\mathbf{y}_{\setminus i}|\mathbf{x}) = \prod_{j \neq i} q(y_j|\mathbf{x})$ is the product of all marginals without marginal $q_i(y_i|\mathbf{x})$.

10.2 Mean field variational inference II

Assume random variables y_1, y_2, x are generated according to the following process

$$y_1 \sim \mathcal{N}(y_1; 0, 1) \qquad y_2 \sim \mathcal{N}(y_2; 0, 1) \tag{10.8}$$

$$n \sim \mathcal{N}(n; 0, 1) \qquad x = y_1 + y_2 + n \tag{10.9}$$

where y_1, y_2, n are statistically independent.

(a) y_1, y_2, x are jointly Gaussian. Determine their mean and their covariance matrix.

(b) The conditional $p(y_1, y_2|x)$ is Gaussian with mean $\mathbf{m}$ and covariance $\mathbf{C}$,

$$\mathbf{m} = \frac{x}{3}\begin{pmatrix}1\\1\end{pmatrix} \qquad \mathbf{C} = \frac{1}{3}\begin{pmatrix}2 & -1\\-1 & 2\end{pmatrix} \tag{10.10}$$

Since x is the sum of three random variables that have the same distribution, it makes intuitive sense that the mean assigns $1/3$ of the observed value of x to y_1 and y_2. Moreover, y_1 and y_2 are negatively corrected since an increase in y_1 must be compensated with a decrease in y_2.

Let us now approximate the posterior $p(y_1, y_2|x)$ with mean field variational inference. Determine the optimal variational distribution using the method and results from Exercise 10.1. You may use that

$$p(y_1, y_2, x) = \mathcal{N}\left((y_1, y_2, x); \mathbf{0}, \mathbf{\Sigma}\right) \quad \mathbf{\Sigma} = \begin{pmatrix}1 & 0 & 1\\0 & 1 & 1\\0 & 1 & 3\end{pmatrix} \quad \mathbf{\Sigma}^{-1} = \begin{pmatrix}2 & 1 & -1\\1 & 2 & -1\\-1 & -1 & 1\end{pmatrix} \tag{10.11}$$

10.3 Variational posterior approximation

We have seen that maximising the evidence lower bound (ELBO) with respect to the variational distribution q minimises the Kullback-Leibler divergence to the true posterior p. We here assume that q and p are probability density functions so that the Kullback-Leibler divergence between them is defined as

$$\mathrm{KL}(q||p) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_q \left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right]. \tag{10.12}$$

(a) You can here assume that $\mathbf{x}$ is one-dimensional so that p and q are univariate densities. Consider the case where p is a bimodal density but the variational densities q are unimodal. Sketch a figure that shows p and a variational distribution q that has been learned by minimising $\mathrm{KL}(q||p)$. Explain qualitatively why the sketched q minimises $\mathrm{KL}(q||p)$.

(b) Assume that the true posterior $p(\mathbf{x}) = p(x_1, x_2)$ factorises into two Gaussians of mean zero and variances σ_1^2 and σ_2^2,

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{x_1^2}{2\sigma_1^2}\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{x_2^2}{2\sigma_2^2}\right]. \tag{10.13}$$

Assume further that the variational density $q(x_1, x_2; \lambda^2)$ is parametrised as

$$q(x_1, x_2; \lambda^2) = \frac{1}{2\pi\lambda^2} \exp\left[-\frac{x_1^2 + x_2^2}{2\lambda^2}\right] \tag{10.14}$$

where λ^2 is the variational parameter that is learned by minimising $\mathrm{KL}(q||p)$. If σ_2^2 is much larger than σ_1^2, do you expect λ^2 to be closer to σ_2^2 or to σ_1^2? Provide an explanation.