TECHNIQUES FOR SAMPLING AND INFERENCE

MPA lecture series, July 25, 2024
Julia Stadler

1. EXPLICIT LIKELIHOOD INFERENCE

We here consider a (cosmological) data analysis with the goal to derive parameter constraints from noisy, incomplete information within a Bayesian framework. In particular, we suppose that we have the following components available in our analysis

- a model $\mathbf{m}(\boldsymbol{\theta})$, that predicts observables from a set of free parameters $\boldsymbol{\theta}$,
- some observations $\mathbf{d}$, whose noise properties or likelihood $\mathcal{P}(\mathbf{d}|\mathbf{m})$ we understand,

All information on the cosmological parameters $\boldsymbol{\theta}$ under the model $\mathbf{m}$ is encoded in the posterior distribution

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))\,\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\mathbf{d})}\,, \tag{1}$$

where the evidence $\mathcal{P}(\mathbf{d}) = \int d\boldsymbol{\theta}\ \mathcal{P}(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))\,\mathcal{P}(\boldsymbol{\theta})$ ensures proper normalization, $\mathcal{P}(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))$ is the likelihood and $\mathcal{P}(\boldsymbol{\theta})$ the prior. We here specifically focus on continuous parameters $\boldsymbol{\theta}$, so the posterior probability density function (PDF) is a function over multiple dimensions. It is usually compressed into some low-order moments such as the parameter mean and covariance

$$\mathbb{E}[\boldsymbol{\theta}] = \int d\boldsymbol{\theta}\ \theta\,\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})\ , \tag{2}$$

$$\mathbb{V}[\theta]_{ij} = \int d\boldsymbol{\theta}\ \theta_i\theta_j\,\mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) - \mathbb{E}[\theta_i]\,\mathbb{E}[\theta_j]\ . \tag{3}$$

More generically, we are interested in the expectation of some function $\phi(\boldsymbol{\theta})$ under this distribution

$$\Phi = \mathbb{E}[\phi(\boldsymbol{\theta})] \int d\boldsymbol{\theta}\ \phi(\boldsymbol{\theta})\,\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})\ . \tag{4}$$

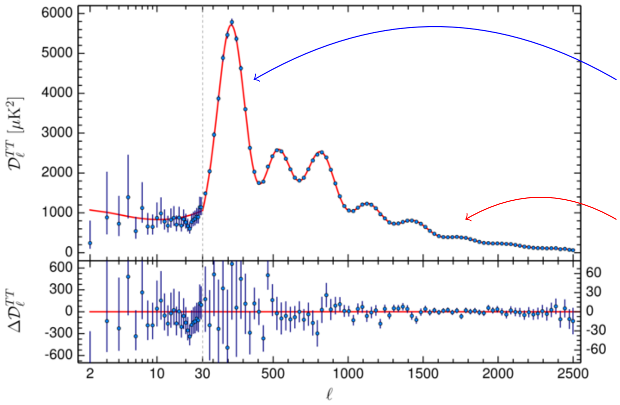
While these Bayesian parameter estimates are straightforward in principle, they can be very hard to implement in practice. For an example, consider the example in figure 1, estimating the ΛCDM parameters from the Planck CMB spectrum. Indeed, for most interesting cases, there is no closed analytical formulation for $\mathcal{P}(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))$ that would permit to compute aboves expectation values. Even though we are able to evaluate the likelihood for specific parameter queries, the dimensionality of the parameter space d typically prohibits solving eq. (4) by numerical integration; as rough limit numerical integration becomes infeasible for $d \geq 4$.

2. ESTIMATING EXPECTATION VALUES FROM SAMPLES

If we are able to draw random samples from the posterior, we can use these to estimate the desired expectation values in eq.(4). This section discusses how the estimates are constructed and their statistical properties, while section 3 goes into detail how the samples are obtained. For lighter notation, the explicit model dependence is dropped in the following, and the posterior is simply denote as $\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})$.

We start from an ensemble of N independent random samples $\left\{\boldsymbol{\theta}^{(n)}\right\}_{n=1}^{N}$, drawn from $\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})$. That is, the probability of having a sample with value θ is proportional to $\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})$. From the samples, we can construct the Monte Carlo estimator for Φ

$$\hat{\Phi} = \frac{1}{N}\sum_n \phi\left(\boldsymbol{\theta}^{(n)}\right)\ . \tag{5}$$



Data **d**: Planck measurement of the temperature auto-correlation. The likelihood $\mathcal{P}(\mathbf{d}|\mathbf{m})$ can be evaluated by a C/Fortran code.

Model **m**: CMB spectrum for ΛCDM by some Boltzmann code, e.g. CAMB, CLASS ...
→ model parameters $\boldsymbol{\theta} = \{\omega_b, \omega_c, \theta_* \tau, \ln A_s, n_s\}$

Figure 1. One example for a Bayesian parameter inference problem: measuring ΛCDM parameters from the CMB temperature auto-correlation. One likelihood evaluation takes $\mathcal{O}(\mathrm{s})$. In addition to the cosmological parameters, there are ~ 16 nuisance parameter, so the space to be sampled has dimensionality $\mathcal{O}(20)$. Derived parameters, e.g. $\mathbb{E}[H_0]$ can be obtained from eq. (4).

If the mean and the variance of ϕ exists, this estimator behaves as

$$\mathbb{E}\left[\hat{\Phi}\right] = \Phi\,, \quad \text{and} \quad \mathbb{V}\left[\hat{\Phi}\right] = \mathbb{V}[\phi]/N\,. \tag{6}$$

The *Central Limit Theorem* states that in the limit of large N the distribution of $\hat{\Phi}$ tends to a Normal distribution, which is completely determined by the mean and variance.

- The accuracy of the Monte Carlo estimate (eq. 5) only depends on the variance of ϕ and on the number of samples, not on the dimensionality of the parameter vector.
- In a typical application, we might draw 100 independent sample, to estimate the mean of a parameter to 10% of its variance.
- The usefulness of the Monte Carlo estimator hinges on the existence of μ and σ for the posterior distribution. Since we typically have no closed-form expression for the posterior, the existence of mean and variance can not straightforwardly demonstrated. In section 2.1, we give an example for sampling from a distribution with ill-defined mean and variance.

Mean and Variance of the estimator $\hat{\Phi}$ We abbreviate $\phi\left(\theta^{(n)}\right) = \phi^{(n)}$. If $\boldsymbol{\theta}^{(n)}$ are independent samples, $\phi^{(n)}$ are independent as well.

$$\mathbb{E}\left[\hat{\Phi}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\phi^{(n)}\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\phi] = \mathbb{E}[\phi] \tag{7}$$

$$\begin{aligned}\mathbb{V}\left[\hat{\Phi}\right] &= \mathbb{E}\left[\left(\left[\frac{1}{N}\sum_{n=1}^{N}\phi^{(n)}\right] - \mathbb{E}[\phi]\right)^2\right] = \frac{1}{N^2}\,\mathbb{E}\left[\left(\sum_{n=1}^{N}\left[\phi^{(n)} - \mathbb{E}[\phi]\right]^2\right)\right] \\ &= \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\left[\left(\phi^{(n)} - \mathbb{E}[\phi]\right)^2\right] = \mathbb{V}[\phi]/N \end{aligned} \tag{8}$$

The central limit theorem We show that the variable

$$Z_N = \frac{\sqrt{N}}{\sigma}\left(\hat{\Phi} - \Phi\right) \tag{9}$$

follows a standard Normal distribution in the limit of large N under the assumption that the moment generation function (MGF) M_{Z_N} exists. This actually is a stronger assumption than the existence of

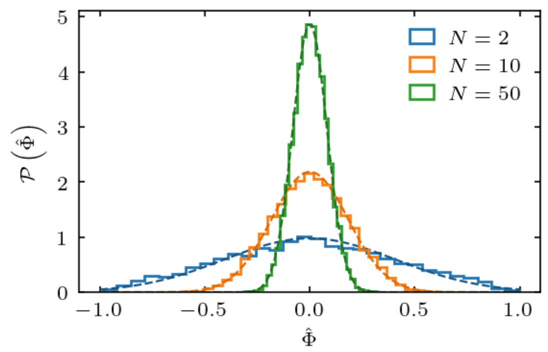


Figure 2. Monte Carlo estimation of the sample mean from a uniform distribution. Solid lines show the histogram of $\hat{\Phi}$ over 10,000 tries for different number of samples N per try. Dashed lines indicate the corresponding Gaussian distribution that $\mathcal{P}\left(\hat{\Phi}\right)$ tends to for a large number of samples, see eq. (16).

mean $\Phi = \mathbb{E}\left[\phi\right]$ and variance $\sigma^2 = \mathbb{V}\left[\phi\right]$. The MGF can be written as,

$$\begin{aligned} M_{Z_N} &= \mathbb{E}\left[\exp\left\{\sqrt{N}\left(\hat{\Phi}-\Phi\right)t/\sigma\right\}\right] = \mathbb{E}\left[\exp\left\{\sum_{n=1}^{N}\frac{\left(\phi^{(n)}-\Phi\right)t}{\sigma\sqrt{N}}\right\}\right] \\ &= \mathbb{E}\left[\prod_{n=1}^{N}\exp\left\{\frac{\left(\phi^{(n)}-\Phi\right)t}{\sigma\sqrt{N}}\right\}\right] = \prod_{n=1}^{N}\mathbb{E}\left[\exp\left\{\frac{\left(\phi^{(n)}-\Phi\right)t}{\sigma\sqrt{N}}\right\}\right] = \left[M_{\phi^{(n)}-\Phi}\left(\frac{t}{\sigma\sqrt{N}}\right)\right]^N , \end{aligned} \tag{10}$$

where we used the independence of samples and the fact that all samples are drawn from the same distribution. The MGF for a single sample can be expanded as

$$M_{\phi^{(n)}-\Phi}\left(t\right) = 1 + \frac{\sigma t^2}{2} + \frac{\nu_3 t^3}{3!} + \dots , \tag{11}$$

where ν_3 is the third moment of ϕ. With that,

$$M_{Z_N} = \left[1 + \frac{1}{N}\left(\frac{t^2}{2} + \frac{\nu_3 t^3}{3!\sigma^3\sqrt{N^3}} + \dots\right)\right]^N , \tag{12}$$

and higher-order moments are suppressed by successively higher powers of N. With the asymptotic behavior

$$\begin{aligned} &\lim_{N\to\infty} \frac{\nu_3 t^3}{3!\sigma\sqrt{N^3}} = 0 , \\ &\lim_{N\to\infty} \left(1 + \frac{a + b(N)}{n}\right)^n = e^a \quad \text{if} \quad \lim_{N\to\infty} b(N) = 0 , \end{aligned} \tag{13}$$

It becomes clear that

$$\lim_{n\to\infty} M_{Z_N}\left(t\right) = e^{t^2/2} , \tag{14}$$

which is the MGF of a Gaussian. The *Uniqueness Theorem* states that if two random variables have the same MGF $M(t)$ for all values of t where the MGF is defined, they have the same PDF. Besides known cases, however, there is no general method to recover the PDF from the MGF.

Coordinate transformations We might be interested in a re-parametrization the parameter vector $\boldsymbol{\rho} = \boldsymbol{\rho}\left(\theta\right)$. From the Monte Carlo estimate (eq. 5) it is clear that we can compute expectation values for $\boldsymbol{\rho}$ directly from samples $\{\boldsymbol{\theta}\}$. Following the transformation of probability density functions, this implies the following priors on $\boldsymbol{\rho}$

$$\mathcal{P}\left(\boldsymbol{\rho}\right) = \mathcal{P}\left(\boldsymbol{\theta}\right)\left|\frac{d\boldsymbol{\rho}\left(\boldsymbol{\theta}\right)}{d\boldsymbol{\theta}}\right|^{-1} . \tag{15}$$

2.1 The Monte Carlo estimate in action – two examples

Uniform distribution As first example, we consider samples drawn from the uniform distribution, $x \hookleftarrow \mathcal{U}\left(-1, 1\right)$. We know that the target distribution has the mean $\mathbb{E}\left[x\right] = 0$ and variance $\mathbb{V}\left[x\right] = 1/3$.

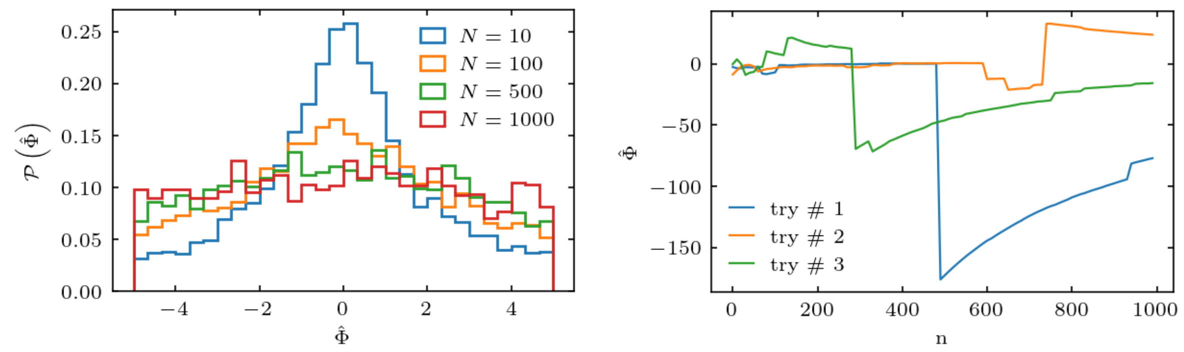


Figure 3. Monte Carlo estimate of the sample mean with samples drawn from a Student's t distribution that has $\nu = 0.7$. That is, mean and variance of the underlying distribution are not well defined. The histograms on the left show the distribution of $\hat{\Phi}$ estimated from 10,000 tries with different numbers of samples N. On the right, we show the evolution of $\hat{\Phi}$ with increasing sample sizes for three tries.

Hence, estimating the distribution mean from N samples we expect that

$$\text{for } \hat{\Phi} = \frac{1}{N} \sum_{n=1}^{N} x_n \,, \quad \mathcal{P}\left(\hat{\Phi}\right) \to \mathcal{N}\left(0, \left(3\sqrt{N}\right)^{-1}\right) . \tag{16}$$

We can test this expectation by repeating the estimate multiple times, each time drawing a new set of samples, and computing the histogram of the distribution of $\hat{\Phi}$. Indeed, these histograms match the expectation well, as figure 2 illustrates.

Student's t distribution As second example, we consider the Student's t distribution

$$\mathcal{P}\left(x|\nu\right) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma\left(\nu/2\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} . \tag{17}$$

For $\nu < 2$, the variance of the distribution is ill-defined and for $\nu < 1$ mean and variance are ill defined. The estimation of the distribution mean from samples is summarized in figure 3 for such a case. As the number of samples increases, the distribution of $\hat{\Phi}$ does not narrow down (left) as we would expect from eq. (6) and observed in case of the Uniform distribution. Rather, adding more samples can lead to sudden and dramatic jumps in $\hat{\Phi}$ (right).

2.2 Why not do something simpler

For a multivariate Normal distribution in d dimensions

$$\mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \Sigma\right) = \frac{1}{\sqrt{\left(2\pi\right)^d \, |\Sigma|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)\right] , \tag{18}$$

we know that the mean and the mode are identical and that the second derivative around this maximum yields the covariance

$$\max_{\boldsymbol{\theta}} \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \Sigma\right) = \boldsymbol{\mu} \tag{19}$$

$$\frac{\partial^2}{\theta_i \, \theta_j} \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \Sigma\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\mu}} = \Sigma^{-1}_{ij} \,. \tag{20}$$

This can motivate a *Maximum a Posteriori (MAP)* estimate, where $\mathbb{E}\left[\boldsymbol{\theta}\right]$ is approximated by the maximum of $\mathcal{P}\left(\boldsymbol{\theta}|\mathbf{d}\right)$ and its variance by the Hessian matrix ath the maximum. However, as long as we don't know the full posterior it is hard to say how good an approximation this is.

- For asymmetric distributions, the mean, the mode and the median are not the same. By optimization we can only obtain the mode.

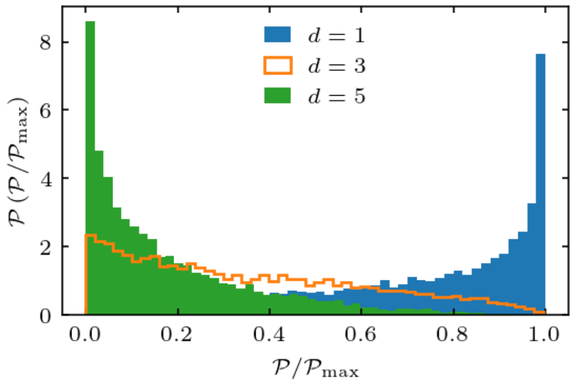


Figure 4. Distribution of the probability for samples from a multivariate Gaussian in d dimensions. In low dimensions, most samples have a high probability, close to the maximum. As the dimensionality increases, samples from lower probability regions become more frequent. **Credit:** Figure adapted from Will Handley.

- For multi-modal distributions, the minimizer might get stuck in the wrong (local) maximum (although multi modal distributions also are challenging to sample).
- For distributions that are heavy-tailed, the error bar will be under-estimated.
- In high-dimensional parameter spaces, the *typical set* is not at the peak (see figure 4 for an illustration), so we might miss a significant portion of the posterior mass.

Typical Set is the region of parameter space V, where most of the posterior mass $\sim \int_V \mathcal{P}\,(\boldsymbol{\theta}|\mathbf{d})$ is located. This can be though of as a balance between probability density and volume or alternatively between likelihood and prior. If we draw samples from the unnormalized posterior, these effectively allow us to explore and characterize the typical set.

Credible Intervals A more robust way of reporting parameter uncertainties for non-Gaussian distributions is the 100 $(1-\alpha)\,\%$ credible interval $[\theta_-, \theta_+]$, defined as

$$\mathcal{P}\,(\theta_- \leq \theta \leq \theta_+|\mathbf{d}) = 1-\alpha \quad \text{or, alternatively} \quad \mathcal{P}\int_{\theta_-}^{\theta_+} \mathcal{P}\,(\theta|\mathbf{d}) = 1-\alpha\,. \tag{21}$$

- This definition makes no statement about the center of the interval, but the most useful choices are *centered credible intervals* around the median (or the mean).
- Typical values for the level are 68%, 95% or 99%, inspired by the probability contained within the 1,2,3-σ contours of a Gaussian.
- The 95% credible interval relies on only 2.5% of the posterior draws and hence require a higher number of samples to be computationally stable. The stan software, for example, reports 90% credible intervals by default for this reason.
- The literature differentiates between frequentist *confidence intervals* and Bayesian *credible intervals.*

2.3 Estimating parameters & estimating the evidence

In many analysis scenarios, we can compute the likelihood at some reasonable computational cost and we also assume some prior distribution. This enables us to compute the posterior up to some normalization constant

$$\mathcal{P}^*\,(\boldsymbol{\theta}|\mathbf{d}) = Z \times \mathcal{P}\,(\boldsymbol{\theta}|\mathbf{d}) = \mathcal{P}\,(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))\,\mathcal{P}\,(\boldsymbol{\theta})\;, \tag{22}$$

where the evidence (sometimes also called the marginal posterior) is

$$Z = \mathcal{P}\,(\mathbf{d}) = \int d\boldsymbol{\theta}\;\mathcal{P}\,(\mathbf{d}|\mathbf{m}(\boldsymbol{\theta}))\,\mathcal{P}\,(\boldsymbol{\theta})\;. \tag{23}$$

The evidence is independent of the parameters θ. Thus, if we manage to draw samples whose frequency is proportional to $\mathcal{P}^*\,(\boldsymbol{\theta})$ they will automatically follow $\mathcal{P}\,(\boldsymbol{\theta})$ and we can obtain the Monte Carlo estimate in eq. (5) equally from these samples.

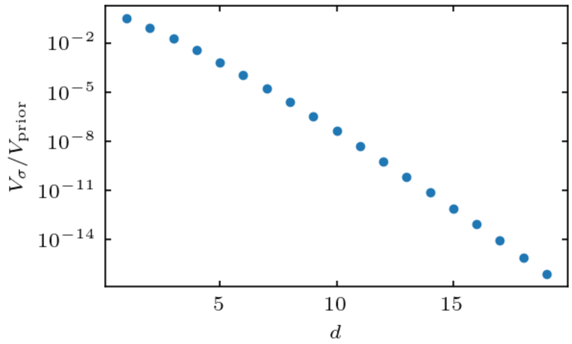


Figure 5. Illustration how the typical set gets increasingly concentrated in higher dimensions d. Here, the likelihood is a multivariate Normal distribution and the prior a uniform distribution, $\mathcal{U}[-3,3]$, i.e. the prior covers the 3σ interval of the likelihood in each dimension. The volume of the standard deviation ellipse is $V_d = \pi^{d/2}/\Gamma(n/2+1)$, and its share of the prior volume decreases dramatically with dimensionality.

Still, the evidence is an interesting quantity for model selection. In principle, we could try to estimate it directly from Monte Carlo samples by a method called *Harmonic mean estimator* [1]

$$\mathbb{E}\left[\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\right]_{\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})} = \int d\boldsymbol{\theta}\, \frac{\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})}{\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})} = \int d\boldsymbol{\theta}\, \frac{\mathcal{P}(\boldsymbol{\theta})}{Z} = \frac{1}{Z} \tag{24}$$

$$\hat{\rho} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{P}^{-1}\left(\mathbf{d}|\boldsymbol{\theta}^{(n)}\right) \quad \text{with} \quad \boldsymbol{\theta}^{(n)} \hookleftarrow \mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) \tag{25}$$

The mean of $\hat{\rho}$ converges to Z^{-1}, however, its variance is ill behaved and can be infinite in many cases. The intuitive reason for this poor behavior is that posterior samples usually are dominated by the likelihood, which is narrower than the prior in most cases. The evidence, however, should depend sensitively on the prior; with an improper prior it reduces to zero. The harmonic mean estimator cannot reproduce this behavior [2]. More elaborate methods are required for reliable evidence estimation, such as *nested sampling* [3].

3. DRAWING SAMPLES

For simplicity of notation, I will write a generic PDF as $\mathcal{P}(\mathbf{x})$ and an un-normalized distribution as $\mathcal{P}^*(\mathbf{x})$ from here on, and the samples are $\left\{\mathbf{x}^{(n)}\right\}_{n=1}^{N}$.

Analytic distributions There is a simple trick to draw samples from an analytic PDF if its *cumulative distribution function (CDF)* and its inverse are known or can be approximated,

$$F_{\mathcal{P}}(y) = \int_{-\infty}^{y} dt\ \mathcal{P}(t)\ . \tag{26}$$

For simplicity, we here consider the one-dimensional case. By drawing samples $y^{(n)} \hookleftarrow \mathcal{U}[0,1]$ and evaluating $x^{(n)} = F_{\mathcal{P}}^{-1}(y)$ one can obtain samples from the original distribution, i.e. $x^{(n)} \hookleftarrow \mathcal{P}(x)$. However, in most practical cases this is not very useful because we do not deal with an analytic PDF.

Uniform Sampling We could sample uniformly from the prior and use these samples to estimate the normalization constant and the desired parameter means as

$$\hat{Z}_N = \sum_{n=1}^{N} \mathcal{P}^*\left(\mathbf{x}^{(n)}\right), \quad \hat{\Phi}_{\text{US}} = \sum_{n=1}^{N} \phi\left(\mathbf{x}^{(n)}\right) \frac{\mathcal{P}^*\left(\mathbf{x}^{(n)}\right)}{\hat{Z}_N} \tag{27}$$

These estimators will be dominated by samples from the typical set. However, as the dimensionality increases, the typical set becomes more concentrated and sampling uniformly from the prior it is less and less likely that any samples will ly in the typical set (see figure 5 for an illustration). In this case, uniform sampling gives a very poor estimate.

Importance sampling One attempt to improve over uniform sampling is drawing samples from an auxiliary distribution $\mathcal{Q}(\mathbf{x})$ which is non-zero for all $\mathbf{x}$ where $\mathcal{P}(\mathbf{x}) \neq \mathbf{x}$. Each sample the is assigned the weight $w_n = \mathcal{P}^*\left(\mathbf{x}^{(n)}\right)/\mathcal{Q}\left(\mathbf{x}^{(n)}\right)$, and the weight is used to adjust the "importance" of each sample in the estimator

$$\hat{\Phi}_{\mathrm{IS}} = \sum_{n=1}^{N} w_n\, \phi\left(\mathbf{x}^{(n)}\right) / \sum_{n=1}^{N} w_n\,. \tag{28}$$

However, the problem here is similar to uniform sampling, and importance sampling will not give a good estimate unless $\mathcal{Q}$ is very close to $\mathcal{P}$.

3.1 Markov Chains

Markov chains allow to explore the typical set more effectively. A Markov chain is a stochastic process, defined as a sequence of random variables $\{x_t\}_{t=1}^{T}$ (I have switched the index from n to t here to indicate that these variables are no longer independent). The distribution of states in the chain is denoted as $\mathcal{P}_{\mathrm{MC}}^{(t)}(\mathbf{x})$ and the Markov chain can be specified by

- the initial probability distribution $\mathcal{P}_{\mathrm{MC}}^{(0)}(\mathbf{x})$
- the transition probability between states $\mathcal{T}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right)$.

The transition probability $\mathcal{T}$ formulated above already implements the *Markov property*, i.e. that each state in the chain only depends on its immediate predecessor.

By constructing a Markov chain whose distribution $\mathcal{P}_{\mathrm{MC}}^{(t)}(\mathbf{x})$ tends to the target distribution $\mathcal{P}(\mathbf{x})$ in the limit of large t, it is possible to effectively sample from the target distribution. The methods discussed below equally apply to normalized and unnormalized target distribution, for simplicity the target is denoted as $\mathcal{P}(\mathbf{x})$.

Subsequent samples in the Markov chain will be correlated, but the correlation decays with increasing distance between samples and we can obtain a set of *effectively independent samples* by thinning out the chain. Actually, dependent samples will not lead to a bias in the estimator $\hat{\Phi}$, so we can combine all samples from a chain. When we assess the accuracy of $\hat{\Phi}$, we however have to take account of the number of *independent samples*, see section 4.3.

Two criteria need to be fulfilled so the Markov chain converges to the desired target distribution $\mathcal{P}(\mathbf{x})$.

- *Invariance* of the target distribution $\mathcal{P}(\mathbf{x})$ under the chain. A distribution $\pi(\mathbf{x})$ is an invariant distribution of the transition probability if

$$\pi(\mathbf{x}') = \int d\mathbf{x}\; \mathcal{T}(\mathbf{x}'|\mathbf{x})\; \pi(\mathbf{x}) \tag{29}$$

  An invariant distribution is an eigenvector of the transition probability with eigenvalue 1.

- *Ergodicity* of the chain

$$\mathcal{P}_{\mathrm{MC}}^{(t)} \to \pi(\mathbf{x}) \quad \text{as} \quad t \to \infty \quad \text{for any} \quad \mathcal{P}_{\mathrm{MC}}^{(0)}(\mathbf{x}) \tag{30}$$

  A possible reason for the chain not to be ergodic is if the state space contains two or more subsets that can never be reached from another. The transition probability of such a chain has more than one eigenvector with eigenvalue 1.

Detailed balance Markov chains which satisfy the detailed balance property,

$$\mathcal{T}(\mathbf{x}|\mathbf{x}')\,\mathcal{P}(\mathbf{x}') = \mathcal{T}(\mathbf{x}'|\mathbf{x})\,\mathcal{P}(\mathbf{x}) \quad \text{for all} \quad \mathbf{x}, \mathbf{x}'\,, \tag{31}$$

are called *reversible* Markov chains. Detailed balance implies invariance of the transition under the target distribution, but is actually a more strict criterion,

$$\int d\mathbf{x}\; \mathcal{T}(\mathbf{x}'|\mathbf{x})\; \mathcal{P}(\mathbf{x}) = \int d\mathbf{x}\; \mathcal{T}(\mathbf{x}|\mathbf{x}')\; \mathcal{P}(\mathbf{x}') = \mathcal{P}(\mathbf{x}') \int d\mathbf{x}\; \mathcal{T}(\mathbf{x}|\mathbf{x}') = \mathcal{P}(\mathbf{x}')\,. \tag{32}$$

3.2 Gibbs Sampling

Lets assume that the full distribution $\mathcal{P}(\mathbf{x})$ is too complicated to sample from, but the conditional distributions $\mathcal{P}\left(x_i | \{x_j\}_{j\neq i}\right)$ are traceable In this case, we can draw samples from the full distribution by iteratively sampling from the conditional distributions

$$\begin{aligned} x_1^{(t+1)} &\hookleftarrow \mathcal{P}\left(x_1 | x_2^{(t)}, \dots x_d^{(t)}\right) \\ x_2^{(t+1)} &\hookleftarrow \mathcal{P}\left(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots x_d^{(t)}\right) \\ x_d^{(t+1)} &\hookleftarrow \mathcal{P}\left(x_d | x_1^{(t+1)}, \dots x_{d-1}^{(t+1)}\right) \end{aligned} \tag{33}$$

We show that the target distribution is invariant under this transition for the two-dimensional case $\mathbf{x} = (x_1, x_2)$, but the calculation easily extends to higher dimensions. The transition kernel between two states $\mathbf{x}$ and $\mathbf{x}'$ is

$$\mathcal{T}(\mathbf{x}' | \mathbf{x}) = \mathcal{P}(x_2' | x_1') \, \mathcal{P}(x_1' | x_2) \tag{34}$$

and integration over the initial state $\mathbf{x}$ yields

$$\begin{aligned} \int d\mathbf{x} \; \mathcal{T}(\mathbf{x}' | \mathbf{x}) \, \mathcal{P}(\mathbf{x}) &= \mathcal{P}(x_2' | x_1') \int d\mathbf{x} \; \mathcal{P}(x_1' | x_2) \, \mathcal{P}(x_1, x_2) \\ &= \mathcal{P}(x_2' | x_1') \, \mathcal{P}(x_1') = \mathcal{P}(\mathbf{x}') \; . \end{aligned} \tag{35}$$

The advantage of Gibbs sampling is that there are no tunable parameters, but still we need to be able to sample from the one-dimensional marginal distributions.

3.3 Slice Sampling

Slice sampling [4] allows to draw samples from a one-dimensional probability distribution, and by a Gibbs sampling approach it can be extended to multi-dimensional cases. It builds on the idea that sampling from $\mathcal{P}_{\mathrm{1D}}(x)$ corresponds to uniformly sampling from the curve under the PDF. To that effect, we introduce a joint distribution

$$\mathcal{P}_{\mathrm{2D}}(x, y) = \begin{cases} 1 & \text{if} \quad 0 < y < \mathcal{P}_{\mathrm{1D}}(x) \\ 0 & \text{else} \end{cases} \tag{36}$$

Marginalizing over the 2D distribution yields the original, up to a constant, target as expected

$$\int dy \; \mathcal{P}_{\mathrm{2D}}(x, y) = \int_0^{\mathcal{P}_{\mathrm{1D}}(x)} dy = \mathcal{P}_{\mathrm{1D}}(x) \; . \tag{37}$$

Sampling a new point (x',y') from $\mathcal{P}_{\mathrm{2D}}$ proceeds in two steps (see figure 6 for an illustration):

(1) Draw y' uniformly from $[0, \mathcal{P}_{\mathrm{1D}}(x)]$.

(2) Draw x' uniformly from the slice $S = \{x : y' < \mathcal{P}(x)\}$

- Randomly position an interval $[x_{\mathrm{l}}, x_{\mathrm{r}}]$ of size w around x. From (1) it is clear that x and thus at least parts of the interval are within the slice.

$$\begin{aligned} r &\hookleftarrow \mathcal{U}[0,1] \\ x_{\mathrm{l}} &= x - rw \\ x_{\mathrm{r}} &= x + (1-r)w \end{aligned}$$

- Enlarge the interval to both sides until there ends fall outside the slice.

$$\begin{aligned} \text{while} \quad \mathcal{P}(x_{\mathrm{l}}) > y' \quad x_{\mathrm{l}} \leftarrow x_{\mathrm{l}} - w \\ \text{while} \quad \mathcal{P}(x_{\mathrm{r}}) > y' \quad x_{\mathrm{r}} \leftarrow x_{\mathrm{r}} + w \end{aligned}$$

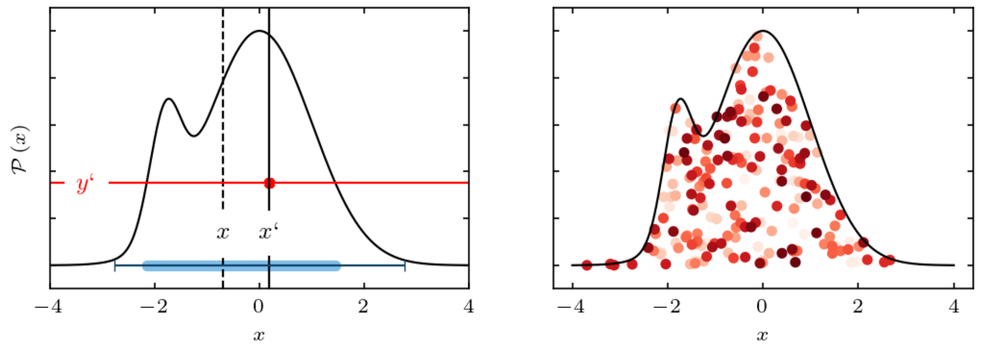


Figure 6. Illustration of slice sampling in one-dimension. The left panel shows all steps of a single parameter update: the dashed vertical line indicates the original position, the red horizontal line is the new value drawn for y', the thick blue region illustrates the slice from which x' is drawn and the interval the initial search region, the new value x' is indicated by a black vertical line. The right panel shows the x, y coordinates of 200 samples for a step size $w = 0.3$ where darker color indicates higher sample indices.

- Randomly draw x' from the interval and accept the sample if it is inside the slice. Otherwise, use the sample to shrink the interval.

$$\text{if} \quad (x' > x) \;\; x_\mathrm{r} \leftarrow x' \quad \text{else} \quad x_\mathrm{l} \leftarrow x' \,.$$

There are alternative methods for constructing the slice and adjusting the its size by rejected samples, the ones described above are called *stepping out* procedure and *shrinkage* procedure.

Convergence to the target density Showing that the target probability is invariant under the transition is a bit tedious. Step (1) leaves the distribution $\mathcal{P}_{2\mathrm{D}}$ invariant as we have shown for Gibbs sampling above. It remains to show that the update of x in step (2) leaves the joint distribution invariant. Various random choices r are made in step (2) and one can show that this step satisfies

$$\mathcal{P}\left(x', r|x\right) = \mathcal{P}\left(x, \pi(r)|x'\right) \tag{38}$$

where $\pi(r)$ is a one-to-one mapping with unit Jacobian. Marginalizing both sides over r then shows detailed balance in step (2).

Efficiency The only tunable parameter in slice sampling is the step size w. Luckily, through the stepping out and shrinking procedure, the efficiency of the algorithm does not depend on w very crucially. For optimal performance, in the sense of minimal likelihood evaluations per accepted sample, w should be chosen of similar order as the parameter variance. In case of doubt, a larger step size is less disadvantageous than a too low one. This is because every enlargement step increases the interval by $2w$, while the shrinkage step typically decreases the interval by a factor 0.6. So if the interval width is off by a factor f, enlargement scales linearly in f but shrinkage only logarithmically.

3.4 Metropolis-Hastings

In contrast to slice sampling, the Metropolis-Hastings algorithm [5, 6] updates all components of a multi-dimensional parameter vector simultaneously. It consists of the following steps

(1) Given the current state $\mathbf{x}^{(t)}$, draw a new proposed state from the proposal distribution $\mathcal{Q}\left(\mathbf{x}'|\mathbf{x}^{(t)}\right)$. Note that by the Markov property, the proposal distribution can only depend on the current state, but there are no further restrictions. Many implementations use simple proposal densities such as a Gaussian distribution centered at the current state.

(2) Accept the proposal with probability

$$\alpha\left(\mathbf{x}'|\mathbf{x}^{(t)}\right) = \min\left[1\,,\; \frac{\mathcal{P}\left(\mathbf{x}'\right)}{\mathcal{P}\left(\mathbf{x}^{(t)}\right)} \frac{\mathcal{Q}\left(\mathbf{x}^{(t)}|\mathbf{x}'\right)}{\mathcal{Q}\left(\mathbf{x}'|\mathbf{x}^{(t)}\right)}\right]\,, \tag{39}$$

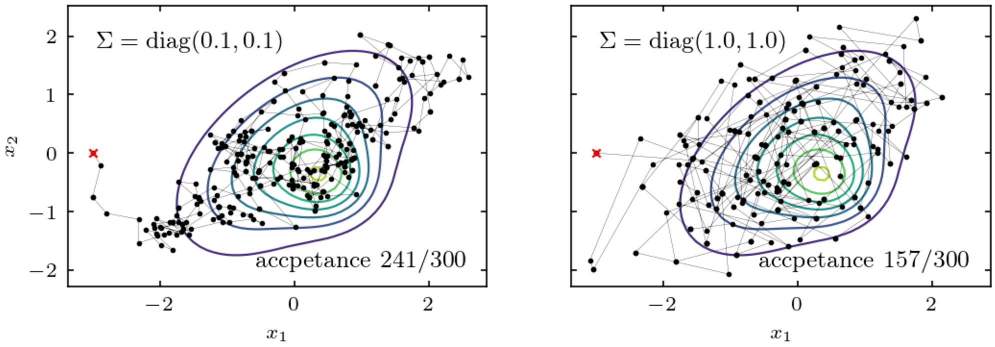


Figure 7. Sampling of a two-dimensional target density with the Metropolis-Hastings algorithm using a Gaussian proposal density. The proposal density is narrower on the left, resulting in a high number of accepted samples, which however, a strongly correlated. The wider proposal density on the right has the opposite effect. A red cross marks the initial point of the chains.

that is

$$\begin{aligned} &\text{draw } u \hookleftarrow \mathcal{U}\left(0,1\right) \ , \\ \text{if } \alpha \geq u \quad &\mathbf{x}^{(t+1)} = \mathbf{x}' \, , \\ \text{else} \quad &\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} \, . \end{aligned}$$

The posterior normalization cancels in α, so this algorithm is explicitly independent of Z.

Convergence to the target density The transition probability of the Metropolis-Hastings algorithm is

$$\mathcal{T}_{\mathrm{MH}}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) = \alpha\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) \mathcal{Q}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) \, . \tag{40}$$

Obviously, if $\alpha\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) < 1$, then $\alpha\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right) = 1$ and vice versa. Therefore, to show detailed balance, we can assume $\alpha\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) < 1$ without loss of generality. In this case, we have

$$\begin{aligned} \mathcal{T}_{\mathrm{MH}}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) \mathcal{P}\left(\mathbf{x}^{(t)}\right) &= \alpha\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) \mathcal{Q}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) \mathcal{P}\left(\mathbf{x}^{(t+1)}\right) \\ &= \mathcal{Q}\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right) \mathcal{P}\left(\mathbf{x}^{(t)}\right) = \mathcal{T}_{\mathrm{MH}}\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right) \mathcal{P}\left(\mathbf{x}^{(t+1)}\right) \, , \end{aligned} \tag{41}$$

and the last equality follows from $\alpha\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right) = 1$.

Efficiency The efficiency at which the Metropolis-Hastings algorithm explores the parameter space depends on the choice of the proposal density. If the typical step size is large (compared to the parameter variance), a single update is likely to propose a sample outside the typical set and the acceptance rate will be low. In contrast, if the step size is too small, subsequent samples are highly correlated and it takes a long chain to explore the typical set effectively. Figure 7 illustrates the exploration of a two-dimensional posterior for different proposal densities. Strong parameter degeneracies confine the typical set to a thin sheet and reduce the acceptance rate additionally. If known beforehand, one can try to avoid such situations by a reparametrization.

To optimize the proposal density, it can be useful to run a brief initial chain from which a rough estimate for the posterior covariance is obtained. These samples are then discarded. It is not valid to adapt the proposal density based on previous samples; this would validate detailed balance.

3.5 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo [7] makes use of gradient information to reduce the random walk behavior of the Metropolis-Hastings algorithm and to propose samples that can be accepted with a high probability.

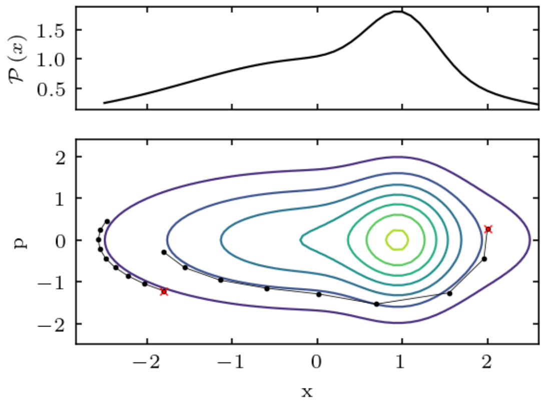


Figure 8. Examples for two HMC updates (**bottom**) sampling a one-dimensional target distribution (**top**). The mass matrix is $\mathbb{1}$ and each update evolves for $N_\epsilon = 8$ leapfrog steps. Red crosses indicate the starting point. One update uses a larger step size $\epsilon = 0.5$, the other a smaller one of $\epsilon = 0.2$. The chain with small ϵ starts at the endpoint of the larger-ϵ one, due to the random momentum draw at the beginning of each iteration the evolve along different energy levels.

At the same time, it allows for longer step sizes, reducing the correlation between samples. It is of particular importance to the sampling of high-dimensional parameter spaces, where the acceptance rate and hence the efficiency of other methods reduces drastically.

In Hamiltonian Monte Carlo, the parameter space is augmented by auxiliary momentum variables $\mathbf{p}$ to define the Hamiltonian

$$H(\mathbf{x}, \mathbf{p}) = E(\mathbf{x}) + K(\mathbf{p}) = -\ln \mathcal{P}(\mathbf{x}) + \frac{1}{2}\mathbf{p}^T M^{-1} \mathbf{p} \,. \tag{42}$$

In analogy to classical mechanics, the negative log-posterior is interpreted as potential energy. The kinetic term can in principle be chosen freely but most implementations assume a quadratic form with mass matrix M. This Hamiltonian defines the joint probability distribution of position $\mathbf{x}$ and momentum $\mathbf{p}$ variables,

$$\mathcal{P}_{\mathrm{H}}(\mathbf{x}, \mathbf{p}) = \frac{1}{Z_{\mathrm{H}}} e^{-H(\mathbf{x},\mathbf{p})} = \frac{1}{Z_{\mathrm{H}}} \mathcal{P}(\mathbf{x})\, e^{-\frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}} \,, \tag{43}$$

where we see that the quadratic form for the kinetic energy implies a Gaussian distribution of the momenta. Because the probability distribution of positions and momenta are independent, sampling from the joint distribution and marginalization over $\mathbf{p}$, yields samples from the desired target $\mathcal{P}(\mathbf{x})$. Each update proceeds in the following steps:

(1) At the current position $\mathbf{x}^{(t)}$, draw new random momenta from $e^{-K(\mathbf{p})}$.

(2) Evolve to a new state $(\mathbf{x}', \mathbf{p}')$ by solving the Hamiltonian equations

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i} \,, \quad \text{and} \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i} \,. \tag{44}$$

and negating the momentum at the end. In practice, the solution is found numerically for some finite time step ϵ, evolving the system over N_ϵ steps. Symplectic integrators, like the Hamiltonian equations, are time reversible and preserve volume. Their solution cannot drift too far away from the true trajectory, and it rather oscillates around it for long integration times.

(3) Accept the new state with probability

$$\alpha\left(\mathbf{x}', \mathbf{p}' | \mathbf{x}^{(t)}, \mathbf{p}^{(t)}\right) = \min\left[1\,,\; e^{H\left(\mathbf{x}^{(t)},\mathbf{p}^{(t)}\right) - H\left(\mathbf{x}',\mathbf{p}'\right)}\right] \,, \tag{45}$$

which is the Metropolis-Hastings acceptance criterion for the joint probability distribution. Because the Hamiltonian is a constant of motion, the proposals can always be accepted in case of an ideal time integration and with a high probability if a suitable step size ϵ is chosen.

The update in the position-momentum plane for a one-dimensional target distribution is illustrated in figure 8 for two different step sizes ϵ.

The leapfrog integrator Most HMC implementations use a first-order leapfrog integrator. For simplicity, I here assume a diagonal mass matrix, where $K = \sum_i p_i^2/m_i$. The update rule is

$$\begin{aligned} p_i\left(t+\epsilon/2\right) &= p_i\left(t\right) - \frac{\epsilon}{2}\frac{\partial E}{\partial x_i}\left(\mathbf{x}\left(t\right)\right) \\ x_i\left(t+\epsilon\right) &= x_i\left(t\right) + \epsilon\,\frac{p_i\left(t+\epsilon/2\right)}{m_i} \\ p_i\left(t+\epsilon\right) &= p_i\left(t+\epsilon/2\right) - \frac{\epsilon}{2}\frac{\partial E}{\partial x_i}\left(\mathbf{x}\left(t+\epsilon\right)\right)\ . \end{aligned} \tag{46}$$

Convergence to the target density For a given $\mathbf{x}$ and $\mathbf{p}$ the mapping to a new state is deterministic and one-to-one. For the moment, lets assume for simplicity that there is only one momentum $\tilde{\mathbf{p}}^{(t)}$ which transforms $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$. The transition probability then is $\mathcal{T}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) = \mathcal{P}\left(\tilde{\mathbf{p}}^{(t)}\right)$, and we denote the final momentum, at the end of the trajectory, as $\tilde{\mathbf{p}}^{(t+1)}$. The mapping is reversible by negating the final momentum and $\mathcal{T}\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right) = \mathcal{P}\left(\tilde{-\mathbf{p}}^{(t+1)}\right)$. With these properties, we can show detailed balance,

$$\begin{aligned} \mathcal{T}\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right)\mathcal{P}\left(\mathbf{x}^{(t)}\right) &= \mathcal{P}\left(\mathbf{x}^{(t)}\right)\mathcal{P}\left(\tilde{\mathbf{p}}^{(t)}\right) \\ &= \frac{1}{Z_\mathrm{H}}\exp\left[-H\left(\mathbf{x}^{(t)},\tilde{\mathbf{p}}^{(t)}\right)\right] \\ &= \frac{1}{Z_\mathrm{H}}\exp\left[-H\left(\mathbf{x}^{(t+1)},\tilde{\mathbf{p}}^{(t+1)}\right)\right] \\ &= \mathcal{P}\left(\mathbf{x}^{(t+1)}\right)\mathcal{P}\left(\tilde{-\mathbf{p}}^{(t+1)}\right) \\ &= \mathcal{T}\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}\right)\mathcal{P}\left(\mathbf{x}^{(t+1)}\right)\ , \end{aligned} \tag{47}$$

where we used that the mapping conserves the Hamiltonian and that the Gaussian momentum distribution is symmetric under $\mathbf{p} \to -\mathbf{p}$. Because the mapping is one-to-one, this argumentation generalizes to the case where more than one momenta map $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$.

Efficiency The HMC algorithm contains several tunable parameters:

- **Mass Matrix** To understand the role of the mass matrix, note that the dynamics are invariant under the transformation
$$\mathbf{x}' = A\mathbf{x} \quad \text{and} \quad \mathbf{p}' = A\mathbf{p}\,, \tag{48}$$
for an orthogonal matrix A. If the distribution of $\mathbf{q}$ is close to a Gaussian with covariance Σ and $\Sigma = LL^T$ is the Cholesky decomposition, we could define $\mathbf{x}' = L^{-1}\mathbf{x}$ and use $M = \mathbb{1}$. In this case, the momentum variables are independent and the position variables are almost independent and they all have a variance close to one. HMC should perform well and deliver nearly independent samples for a small number of leapfrog steps. The same can be achieved by keeping the original position variables $\mathbf{x}$ but using $M = \Sigma$. In high dimensions, often only the use of a diagonal mass matrix is feasible. Still, an adequate choice of the diagonal elements can improve the performance considerably.

- **Leapfrog step size** A too large step size ϵ can lead to instable integration and dramatically reduce the acceptance rate. Too small steps waste computational resources by unnecessary posterior evaluations. As rough criterion, the stability is determined by the width of the distribution in the most constrained direction.

- **Number of leapfrog steps** A suitable trajectory length ϵN_ϵ is crucial to suppress the random walk behavior and explore the state space systematically. Longer trajectories also are effective in exploring less constrained directions. However, too long trajectories become inefficient in traversing the typical set multiple times.

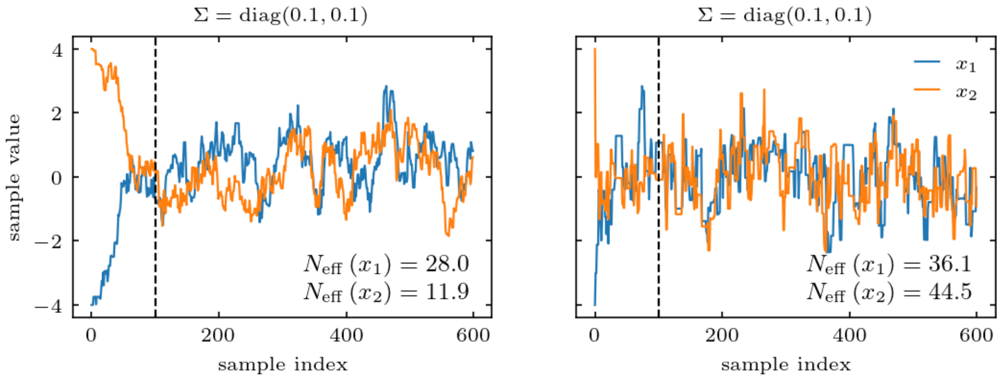


Figure 9. Illustration of *burn-in* and correlations between chains for the two Metropolis-Hastings examples from figure 7. To illustrate the burn-in phase, we now start the chains further away from the maximum likelihood point. The chain with a narrower proposal density takes shorter steps and more time to reach the typical set. Due to the shorter step size, samples are also more strongly correlated resulting in a lower number of effective samples, after we have removed the first 100 for burn-in. Note that the situation at some point would reverse as we continued to increase the width of the proposal distribution; the low acceptance rate then causes a drop in the efficiency at which the typical set is explored.

4. CONVERGENCE

After running a Markov chain, we need to answer the following questions in order to build reliable Monte Carlo estimators for the parameters of interest.

- Has the chain converged to the stationary distribution? Are the samples representative of the target distribution and not biased by the starting point of the chain?
- Have we collected enough *independent* samples for reliable parameter estimates?

Too short chains lead to unreliable and inaccurate results; too long chains wast computational resources.

4.1 Burn-in

After a chain starts from a random initial position, it requires some steps to drift towards regions of high probability. Only once a chain has reached the typical set, the samples will be representative of the underlying target distribution. This initial period is called *burn-in*, and its length depends target distribution and on the settings of the MCMC algorithm, such as proposal distribution or step size.

The burn in for the two previous Metropolis-Hastings examples (figure 7) is illustrated in figure 9. These type of *trace plots* are an important too to assess the convergence of a chain. Unfortunately, it is quite difficult to estimate the length of the burn in period automatically. There is no standard procedure and many MCMC packages require the user to specify the burn-in length by hand. Below, we discuss two attempts to nevertheless give a burn-in estimate based on the chain samples.

Burn-in estimation in Monte Python Monte Python [8] is a python package for cosmological parameter inference that provides a convenient interface with the Boltzmann solver CLASS. It automatically discards all samples before the chain has reached an effective χ^2 less that $\chi^2_{\text{min}} + 6$ for the first time.

The Geweke Z-Score test compares the parameter estimates from different segments of the chain.

$$\begin{aligned}\hat{\Phi}_a &= \frac{1}{aT}\sum_{t=1}^{aT}\phi^{(t)} \\ \hat{\Phi}_b &= \frac{1}{bT}\sum_{t=(1-b)T}^{T}\phi^{(t)}\end{aligned} \tag{49}$$

For the comparison, an estimate of the parameter variances $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$ are also required. These have to be computed from thinned chains, because subsequent samples within a chain are correlated (see below, here the problem starts going in circles a bit). The *Z-score statistics* is

$$Z_{\mathrm{G}} = \frac{\hat{\Phi}_a + \hat{\Phi}_b}{\sqrt{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}} \, . \tag{50}$$

If both segments represent the same target distribution, Z_{G} should follow a standard normal distribution. A possible application of this test is to place the interval b at the end of the chain and move the a interval forward until Z_{G} doesn't exceed approximately ±3.

4.2 Gelman-Rubin Test

The Gelman-Rubin test [9] judges the convergence of an MC estimate based on the comparison between a number N_{C} of parallel chains with different starting points. It is based on the comparison of the *within-chain variance* σ_{W}^2 and the *between-chain variance* σ_{B}^2.

For each of the N_{C} chains with T samples, one can compute the sample mean $\hat{\Phi}_i$ according to eq. (5) and the mean of means

$$\bar{\hat{\Phi}} = \frac{1}{N_{\mathrm{C}}}\sum_{i=1}^{N_{\mathrm{C}}}\hat{\Phi}_i \, , \tag{51}$$

The within-chain variance and the between-chain variance are defined respectively as

$$\begin{aligned}\hat{\sigma}_{\mathrm{W}}^2 &= \frac{1}{N_{\mathrm{C}}}\sum_{i=1}^{N_{\mathrm{C}}}\left[\frac{1}{T-1}\sum_{t=1}^{T}\left(\phi_i^{(t)} - \hat{\Phi}_i\right)^2\right] . \\ \hat{\sigma}_{\mathrm{B}}^2 &= \frac{T}{N_{\mathrm{C}}-1}\sum_{i=1}^{N_{\mathrm{C}}}\left(\hat{\Phi}_i - \bar{\hat{\Phi}}\right)^2 .\end{aligned} \tag{52}$$

With these definitions and under the assumption that all chains sample the same stationary distribution, the following unbiased estimator can be constructed for $\mathbb{V}[\phi]$

$$\hat{\sigma}_\phi^2 = \left(\frac{T-1}{T}\right)\hat{\sigma}_{\mathrm{W}}^2 + \frac{1}{T}\hat{\sigma}_{\mathrm{B}}^2 \, . \tag{53}$$

Before the stationary distribution is reached, $\hat{\sigma}_\phi^2$ overestimates the variance due to different starting points. If the posterior is multi modal and individual chains explore different local maxima, the between-chain variance is much larger than the within-chain variance. Convergence is monitored using the Gelman-Rubin test statistic

$$\hat{R} = \sqrt{\frac{\hat{\sigma}_\phi^2}{\hat{\sigma}_{\mathrm{W}}^2}} \, , \tag{54}$$

and it is common in MCMC analyses to quote a maximum limit for $1 - \hat{R}$ in addition to the parameter estimates. Figure 10 and table 1 give an example for poorly converged chains due to a multi modal posterior that are correctly recognized by the Gelman-Rubin test.

4.3 Effective Sample Size

The samples within a MCMC chain typically are correlated and this increases the uncertainty of parameter estimates. One can account for this by estimating the effective sample size N_{eff}, and replacing the

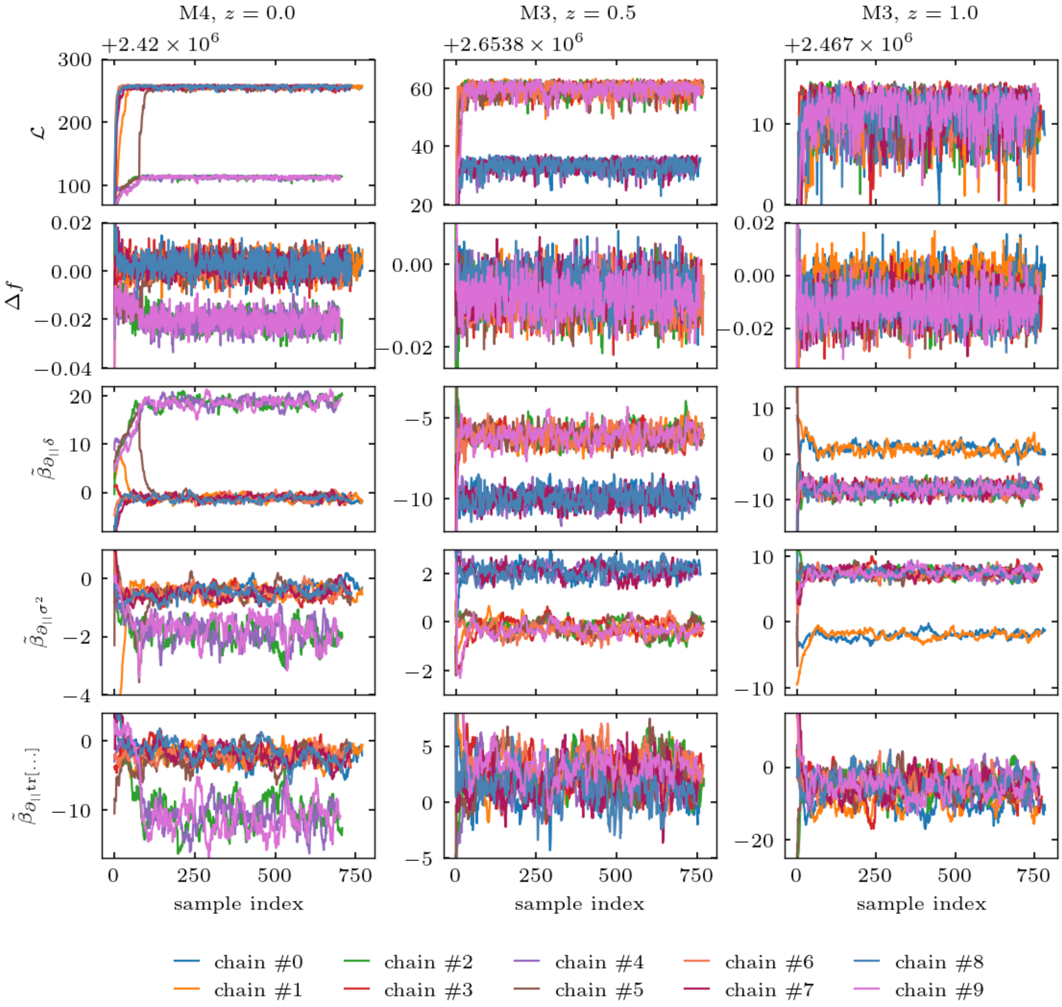


Figure 10. Convergence tests for an analysis that uses the slice sampler. The top row shows the likelihood of each sample, while the three bottom rows show the sample values for different parameters of interest. The three columns consider three different data sets. In all cases, there are indications for a multi modal posterior distribution and chains typically get stuck in a local maximum. In the first two cases, there is a clear hierarchy between the two maxima, with one of them exhibiting significantly higher likelihood values. Traditional Markov Chains become very inefficient in such multi modal settings, and non of these examples can be considered converged.
Credit: figure taken from [10].

			Gelman-Rubin criterion				
data set and analysis		N	f	b_δ	$\tilde{\beta}_{\partial_{||}\delta}$	$\tilde{\beta}_{\partial_{||}(\sigma^2)}$	$\tilde{\beta}_{\partial_{||}\mathrm{tr}[\ldots]}$
halos, sim. 1, M4, z=0, Λ=0.16 h/Mpc	*	6.5×10^3	3.0	1.4	12.8	2.7	2.9
halos, sim. 1, M3, z=0.5, Λ=0.16 h/Mpc	*	7.6×10^3	1.04	1.04	4.2	4.5	1.9
halos, sim. 1, M3, $z = 1$, Λ=0.16 h/Mpc	*	7.8×10^3	1.2	1.2	4.2	7.4	1.3

Table 1: Gelman-Rubin criterion $\hat{R}$ for the chains shown in figure 10. In all three cases, the chains are poorly converged and correspondingly $\hat{R}$ deviates significantly from unity.

number of independent samples N in eq. (6) by the effective sample size N_{eff}. To see how this works, we start from the (normalized) auto-correlation of the chain

$$\rho\left(\Delta t\right) = \sigma^{-2}\left[\mathbb{E}\left[\phi^{(t)}\phi^{(t+\Delta t)}\right] - \mu^2\right] , \tag{55}$$

where μ and σ are the mean and variance of ϕ, respectively. Typically, the auto-correlation decays exponentially for large Δt. It can be estimated from the samples in a chain as

$$\hat{\rho}\left(\Delta t\right) = \frac{1}{T - \Delta t}\sum_{t=1}^{T-\Delta t}\left(\phi^{(t)} - \mu\right)\left(\phi^{(t+\Delta t)} - \mu\right) , \tag{56}$$

this convolution is usually implemented more efficiently by Fast Fourier Transforms than computing the sum directly.

If we estimate ϕ by the Monte Carlo estimator in eq. (6) from correlated samples, its variance is

$$\begin{aligned} \mathbb{V}\left[\hat{\Phi}\right] &= \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\phi^{(t)}\right)^2\right] = \frac{1}{T^2}\sum_{t,t'=1}^{T}\mathbb{E}\left[\left(\phi^{(t)} - \mu\right)\left(\phi^{(t')} - \mu\right)\right] \\ &= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{\Delta t=t-T}^{T-t}\mathbb{E}\left[\left(\phi^{(t)} - \mu\right)\left(\phi^{(t+\Delta t)} - \mu\right)\right] \simeq \frac{\sigma^2}{T}\sum_{\Delta t=-T}^{T}\rho\left(\Delta t\right) = \frac{\sigma^2}{N_{\text{eff}}} . \end{aligned} \tag{57}$$

In the last equality, we defined the effective sample size $N_{\text{eff}} = T/\hat{\rho}_M$, where the correlation length of the chain is

$$\hat{\rho}_M = \sum_{\Delta t=-M}^{M}\rho\left(\Delta t\right) = 1 + 2\sum_{\Delta t=1}^{M}\rho\left(\Delta t\right) . \tag{58}$$

At long lags, $\hat{\rho}\left(\Delta t\right)$ becomes increasingly noisy, and in practice it is therefore advisable truncate the sum at some $M < T$.

5. PRIORS & MARGINALIZATION EFFECTS

Bayes theorem (eq. 1) provides a well-defined way to extract knowledge from noisy, incomplete observations and to quantify our uncertainty on that knowledge. However, even if our Markov Chains are well converged, marginalization over high-dimensional parameter spaces, prior choices and the parametrization of the parameter space might have unintuitive consequences one should be aware of when interpreting the results. In this context I give some examples from the literature which are by no means exhaustive.

5.1 Prior volume effects

Marginalization translates the distribution of samples in a high-dimensional parameter space into one- and two-dimensional confidence regions, which can be visualized and interpreted more easily. However, the marginalized posterior

$$\mathcal{P}\left(\theta_i|\mathbf{d}\right) = \int\left(\prod_{j\neq i}d\theta_j\right)\mathcal{P}\left(\boldsymbol{\theta}|\mathbf{d}\right) \tag{59}$$

shows the distribution of a parameter according to its integrated probability weight. As such, it can hide points of parameter space that well explain the data but cover only a small volume fraction of the high-dimensional parameter space.

The following example from [11] illustrates this issue. Imagine a two-dimensional parameter space $\boldsymbol{\theta} = (\alpha, \beta)$ that is constrained by experiment 1, and a second experiment that measures only α

$$\begin{aligned} \mathcal{P}_{\text{exp1}}\left(\alpha, \beta\right) &= \frac{5}{21\pi}\left[e^{-\frac{1}{4}\left(\alpha^2+\beta^2\right)} + e^{-\frac{1}{4}(\alpha-3.5)^2-100\beta^2}\right] \\ \mathcal{P}_{\text{exp2}}\left(\alpha\right) &= \frac{1}{\sqrt{\pi/2}}e^{-2(\alpha-5.0)^2} \end{aligned}$$

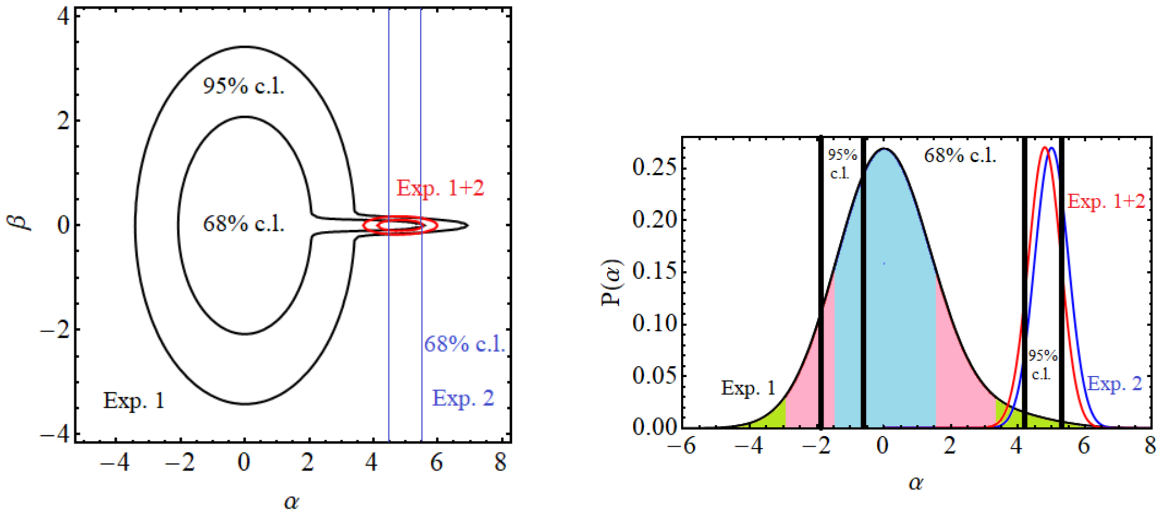


Figure 11. Two-dimensional posterior distribution (**left**) to illustrate prior volume effects. The two experiments are compatible at 1σ and experiment 1 agrees with one of the posterior maxima of experiment 2. Nevertheless, there is an apparent 3σ discrepancy after marginalizing the posteriors over β (**right**), which is driven by the large volume of $\mathcal{P}_{\text{exp1}}(\alpha, \beta)$ that lies left of the best fit value for experiment 2.
Credit: figures taken from [11]

The two dimensional posterior distributions are perfectly consistent as figure 11 illustrates. Nevertheless, after marginalization over β, experiment 2 is almost 3σ discrepant with experiment 1. This discrepancy is driven by the large posterior volume of experiment 1 at low values of α.

Prior volume effects are not per se an error of the analysis, but reflect the Bayesian interpretation of the inference problem. However, they might point to situations where our conclusions are driven by the prior or by the choice of parametrization. As the example above illustrates, the marginalization might hide regions of parameter space which only have a small volume but provide a good explanation to the data. One possibility to diagnose prior volume effects is by comparing the marginalized Bayesian contours to a frequentist profile-likelihood analysis

$$\mathcal{P}_{\text{prof.}}(\theta_i) = \max_{\{\theta_j\}_{j \neq i}} \mathcal{P}(\mathbf{d}|\boldsymbol{\theta}) \, . \tag{60}$$

The latter is insensitive to the volume of the parameter space and will identify those parameters which best describe the data. Computation of the profile likelihood is often expensive, but a rough estimate can also be derived from existing Monte Carlo chains [11]. Prior volume effects might be particularly harmful in scenarios with a large number of nuisance parameters which correlate with the parameters of interest.

In cosmological research, prior volume effects often occur when two parameters are measured simultaneously, one of which becomes unbound as the other approaches a certain value (e.g. zero). The larger posterior volume towards zero then drives very tight constraints on the latter parameter.

- Early dark energy (EDE) was introduced to resolve the Hubble tension, by postulating a scalar field that behaves as Dark Energy at early redshifts and decays to a dark-matter like behavior at the critical redshift z_{c}. The model contains at least two parameters, z_{c} and the maximum fraction that EDE contributes to the energy density f_{EDE} at z_{c}. As f_{EDE} approaches zero, z_{c} becomes unbound leading to tight upper bounds on f_{EDE}. A profile likelihood analysis, instead, reveals a preference for $f_{\text{EDE}} \neq 0$ [12].
- The primordial spectrum of tensor perturbations can be parametrized as

$$P_{\text{t}}(k) = r A_{\text{s}} \left(\frac{k}{k_*} \right)^{n_{\text{t}}} , \tag{61}$$

where r is the tensor-to-scalar ratio, A_{s} the amplitude of the scalar perturbations, n_{t} the spectral tilt and the pivot scale k_* is fixed by the analysis. Currently, there exists only upper bounds on r, and a profile likelihood analysis demonstrated consistency with the Bayesian result at fixed n_{t} [13]. Simultaneous inference of n_{t}, on the other hand appears to be more prior-dependent [14].

5.2 One-sided priors

In case of parameters where only an upper limit exists, priors can become important and drive the strength of parameter constraints. For illustration, we can consider the case of interacting dark matter, which is captured by a single parameter

$$u_X = \frac{\sigma_{\mathrm{DM}-X}}{\sigma_{\mathrm{Th.}}} \left(\frac{m_{\mathrm{DM}}}{100\,\mathrm{GeV}}\right) , \tag{62}$$

and in principle can span many orders of magnitude. It is therefore often customary to assume a logarithmic prior. This choice, however, comes with a lower bound that needs to be specified explicitly. For interacting dark matter constraints from the CMB, [15] showed that the upper bound becomes tighter as the lower limit decreases and is less conservative for a log-prior than if a linear or a Jeffreys prior were assumed. The reason is that the log-prior puts more volume towards small values of u_X, where the data is completely unconstraining, and this effect is the stronger the smaller the lower limit.

As second example, the recent DESI BAO analysis combined with CMB data [16] yields a tight constraint for the sum of neutrino masses $\sum m_\nu$

$$\begin{aligned}
\sum m_\nu &< 0.072\,\mathrm{eV at}\ 95\%\ \mathrm{CL} \quad \text{for} \quad & \sum m_\nu &> 0.000\,\mathrm{eV}\,, \\
\sum m_\nu &< 0.113\,\mathrm{eV at}\ 95\%\ \mathrm{CL} \quad \text{for} \quad & \sum m_\nu &> 0.059\,\mathrm{eV}\,, \\
\sum m_\nu &< 0.145\,\mathrm{eV at}\ 95\%\ \mathrm{CL} \quad \text{for} \quad & \sum m_\nu &> 0.100\,\mathrm{eV}\,.
\end{aligned}$$

Neutrino oscillations imply a lower bound on $\sum m_\nu$ which depends on the mass ordering; it is $\sum m_\nu > 0.059\,\mathrm{eV}$ for normal hierarchy and $\sum m_\nu > 0.1\,\mathrm{eV}$ for inverted hierarchy. The collaboration cautions that, despite the impressive limit in the first line, the preference for a normal neutrino mass hierarchy actually is only at the 2σ level [16].

5.3 Reparametrization

One possibility to deal with degeneracies and volume effects is to attempt a re-parametrization which avoids one parameter becoming unconstrained in the limit of the other going to zero. Such a re-parametrization will inevitably imply a change of prior according tp eq. (15). One the one hand, this effect is desired to reduce prior volume effects, but it might have unintended consequences.

Coming back to the example of the primordial tensor spectrum, [14] compared the parametrization in terms of r and n_{t} to an approach where r is measured at two different scales r_1 and r_2, such that

$$\begin{aligned}
n_{\mathrm{t}} &= \frac{\log r_2/r_1}{\log k_2/k_1} + n_{\mathrm{s}} - 1 \\
r_{k_*} &= r_1\ (k_*/k_1)^{n_{\mathrm{t}}-n_{\mathrm{s}}-1}\ ,
\end{aligned}$$

where n_{s} is the spectral index of scalar perturbations. They found that the latter approach actually implies rather informative priors in r and n_{t}, leading to $r = 0$ being disfavored at 95 % CL from the combination of CMB data with gravitational wave interferometers. Figure 12 illustrates how this happens by translating the uniform prior on r_1 and r_2 back into the r-n_{t} plane.

6. REPORTING THE RESULTS

Analysis method The description of the analysis method should allow for *reproducibility* and *usability* of the results. Other researchers should be able to understand the results, including the assumptions they depend on, and use them in their own work. If a similar analysis where to reach different conclusions, one should be able to understand the differences between the works and what might cause the discrepancy. To this end, the following information on the analysis method should be provided:

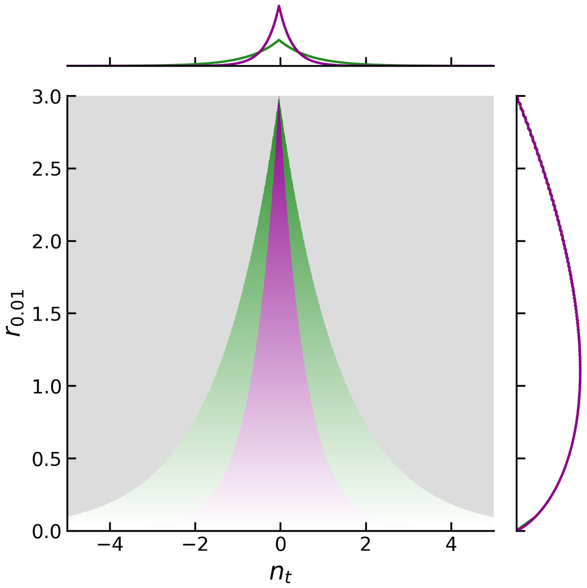


Figure 12. Impact of reparametrization on constraining the amplitude r and tilt n_{t} of the primordial power spectrum of tensor perturbations. Uniform prior contours $r_1, r_2 \hookleftarrow \mathcal{U}(0,3)$ are transformed into the r-n_{t} plane. The green contour is for $(k_1, k_2) = (0.005, 0.02)\,\mathrm{Mpc}^{-1}$ the purple for $(k_1, k_2) = (0.005, 0.05)\,\mathrm{Mpc}^{-1}$, the color gradient indicates the likelihood value at a location, and gray regions are excluded by the prior. Histograms to the top and the right depict the 1 dimensional marginalized priors. The two-scale prior implicitly puts tight constraints on n_{t} and there is little prior mass close to $r = 0$.
Credit: figure taken from [14].

- Assumed priors, also for nuisance parameters.
- The algorithm used and its implementation (i.e. the software package); if tunable parameters differ from defaults, they should also be mentioned.
- The data and likelihood used. In most cases, an MCMC analysis includes a forward model which connects model parameters to observables. This should also be specified.
- How where the chains initialized.
- Number of chains, steps per chain, points discarded for burn-in and possibly thinning factor.

To avoid cluttering the main points and the conclusions, some of this information might be moved to the appendix or summarized in tables.

Convergence The goal of reporting convergence is to demonstrate that the samples sufficiently capture the true shape of the posterior. This should include at least the following tests:

- The autocorrelation length and the effective sample size of the individual parameters. If there is a large number of parameters, the information can also be summarized e.g. shortest and longest correlation length and correspondingly highest and lowest effective sample size.
- The acceptance fraction. Very low or very high acceptance suggests that the sampler is not well-matched to the target distribution and are cause for concern.
- A quantitative convergence criterion such as the Gelman-Rubin criterion (eq. 54). It is very common to quote a maximum value of $\hat{R} - 1$ for all parameter estimates. As a side note, the Gelman-Rubin criterion is not suitable for the *emcee* sampler, which imposes correlations between chains.

Results The results reported should provide an accurate and fair characterization of the high-dimensional posterior distribution implied by the samples. Reporting the results might include the following considerations:

- Corner plots of two-dimensional marginalized parameter constraints. If feasible, it is good practice to include a corner plot with all nuisance parameters in the appendix from which the reader can judge possible parameter degeneracies and prior effects.
- One-dimensional parameter constraints in the form of *median value* and *68% credible interval* for parameters with a unimodal, approximately Gaussian distribution. For one-sided constrained parameters, it is customary to quote a 95% upper limit.

- For unimodal parameters, it can be interesting to check for fat-tailed distribution which contain more "outliers" than what would be expected for a Gaussian. Also, multi-modal and highly correlated parameters should be emphasized. If parameters are correlated but approximately Gaussian distributed (i.e. their 2D contour is a tilted ellipse), one can consider reporting their posterior covariance.
- The shrinkage of the posterior with respect to the prior should be investigated. If the constraint or parts of it are driven by the prior, this should be stated explicitly.
- One can consider publishing the full chains in machine-readable format, including parameter, likelihood and prior values of every samples. This will allow subsequent studies based on the detailed characteristics of the posterior.

7. LITERATURE

- David J.C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.
- Michael Betancourt, *Markov Chain Monte Carlo in Practice*, 05/20.
- Slides by Will Handley.
- G.P. Beaumont, *Probability and Random Variables*, Horwood Publishing, 2005
- Massimiliano Bonamente, *Statistics and Analysis of Scientific Data, Springer Science+Business, 2017*
- Alan D. Sokal, *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*, Lectures at the Cargese Summer School on "Functional Integration: Basics and Applications", 1996.
- Peter K. G. Williams, *Recommendations for Reporting MCMC Analyses in Academic Literature*, published on GitHub

[1] M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994.

[2] R. Neal. Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Newton MA, Raftery AE. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:41–42, 1994.

[3] J. Skilling. Nested Sampling. In R. Fischer, R. Preuss, and U. V. Toussaint, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *American Institute of Physics Conference Series*, pages 395–405. AIP, November 2004.

[4] R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705 – 767, 2003.

[5] W. K. Hastings. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57(1):97–109, April 1970.

[6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[7] R. Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

[8] T. Brinckmann and J. Lesgourgues. MontePython 3: boosted MCMC sampler and other features. *Phys. Dark Univ.*, 24:100260, 2019.

[9] A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7:457–472, January 1992.

[10] J. Stadler, F. Schmidt, and M. Reinecke. Cosmology inference at the field level from biased tracers in redshift-space. *JCAP*, 10:069, 2023.

[11] A. Gómez-Valent. Fast test to assess the impact of marginalization in Monte Carlo analyses and its application to cosmology. *Phys. Rev. D*, 106(6):063506, 2022.

[12] L. Herold, E. G. M. Ferreira, and E. Komatsu. New Constraint on Early Dark Energy from Planck and BOSS Data Using the Profile Likelihood. *Astrophys. J. Lett.*, 929(1):L16, 2022.

[13] P. Campeti and E. Komatsu. New Constraint on the Tensor-to-scalar Ratio from the Planck and BICEP/Keck Array Data Using the Profile Likelihood. *Astrophys. J.*, 941(2):110, 2022.

[14] G. Galloni, N. Bartolo, S. Matarrese, M. Migliaccio, A. Ricciardone, and N. Vittorio. Updated constraints on amplitude and tilt of the tensor primordial spectrum. *JCAP*, 04:062, 2023.

[15] J. A. D. Diacoumis and Y. Y. Y. Wong. On the prior dependence of cosmological constraints on some dark matter interactions. *JCAP*, 05:025, 2019.

[16] A. G. Adame et al. DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations. 4 2024.