On Monte Carlo Methods for Intractable Latent Variable Models

Sebastian Schmon
Magdalen College
University of Oxford

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY

Viva: 18TH of May 2020

Abstract

This thesis provides novel methodological and theoretical contributions to the area of Monte Carlo methods for intractable Bayesian models. Such intractability can come in various forms.

The first project of this thesis considers Markov chain Monte Carlo methods for Bayesian models with intractable likelihood functions. In such cases, the posterior distribution, which is proportional to the product of likelihood and prior, is intractable as well. However, inference based on Markov chain Monte Carlo algorithms is still possible for such models. If the likelihood is replaced by a non-negative unbiased estimate, the resulting algorithm will still target the correct invariant distribution. Such algorithms often result in a trade-off; if the number of samples producing the estimator is increased the algorithm mixes faster, but at the cost of a higher computational effort. In the paper 'Large Sample Asymptotics of the Pseudo-Marginal Algorithm' we analyse the trade-off between the number of samples and speed of mixing under assumptions on the target distribution and the noise of the logarithm of the likelihood estimate. We assume that the Bayesian posterior obeys the regularity conditions which are commonly referred to as Bernstein–von Mises theorems, i.e. the posterior will be similar to a concentrating normal distribution as the number of data increases. Considering the estimator of the log-likelihood we assume that a logarithmic central limit theorem holds as we increase the number of samples with the number of observed data points. Under these conditions, we derive an asymptotic expression of the pseudo-marginal algorithm, which we subsequently exploit as a guide for optimizing such algorithms. We derive novel dimension dependent tuning guidelines for the noise in the likelihood estimate as well as the scaling of the proposal distribution.

Similar intractability problems can occur in the case of sequential Monte Carlo algorithms. Such algorithms require a resampling step where random variables are drawn from a discrete distribution with probabilities proportional to some importance weights. However, these importance weights can be intractable for complicated models or importance distributions. In the second paper, called 'Bernoulli Race Particle Filters', we present a novel resampling algorithm that enables exact sampling of these intractable discrete distributions when only a non-negative unbiased estimator is available. We analyse the resulting algorithm theoretically and investigate its performance on various examples.

Contents

Preface

This thesis follows an *integrated* format. As such its main contribution consists of two research articles which form Chapters 4 and 7, respectively. The remaining chapters aim to embed these two papers in the broader research area of Bayesian statistics and related computational methods, and can be skipped by a reader familiar with the matter. The background material begins with a short introduction to the relevant Bayesian statistical theory in Chapter 1, followed by an introduction to elements of general state space Markov chain theory in Chapter 2. Subsequently, Chapter 3 considers the optimal tuning of Metropolis-Hastings algorithms where we review current state-of-the-art arguments based on scaling limits as the dimension of the parameter space increases. In addition, we provide a simple alternative rule based on a large sample argument.

The paper titled 'Large Sample Asymptotics of the Pseudo-Marginal Algorithm', presented in Chapter 4 considers an original analysis of the pseudo-marginal algorithm leading to a better understanding of previous contributions and novel tuning guidelines. Chapter 5 concludes the discussion of the pseudo-marginal algorithm and outlines potential future research areas.

After a brief introduction to sequential Monte Carlo algorithm in Chapter 6, the final part of thesis consists of the article 'Bernoulli Race Particle Filters' presented in Chapter 7. Chapter 8 concludes this thesis.

Acknowledgement

I am deeply indebted to my supervisors Arnaud Doucet and George Deligiannidis for their continuous guidance and support throughout the course of my studies without which this thesis would not have been possible. Both have been incredible mentors and certainly have helped me become a better researcher. In addition, I want to thank my collaborator Mike for valuable insights and helpful discussions.

During my time in Oxford I had the great pleasure to meet and exchange ideas with many faculty members and fellow students and I want to thank everyone of them. Working with so many brilliant and talented people taught me a lot and has helped me greatly to improve my understanding of statistics and machine learning.

I am also grateful to the Department of Statistics, Magdalen College and the EPSRC for their financial support through their scholarship programmes and travel funding.

Lastly, but most importantly, I want to thank my family, my mother Cleta and my brother Fabian, and my girlfriend Olga for their unending support and encouragement.

Sebastian Schmon, Oxford
April 2020

Part I

Introduction

1

Introduction to Bayesian Statistical Inference

1.1 Statistical Modelling

STATISTICAL procedures provide the connection between mathematical models and real world observations. As such they enable scientists (e.g. physicists, chemists, biologists etc.) to reconcile theory with observed measurements and data. The fundamental principle underlying this method is the rigorous development of probabilistic models and estimation procedures. Formally, we proceed by assuming that the data is a realisation of a random variable taking values in a measure space $(\mathsf{Y}, \mathcal{Y})$ distributed according to an underlying probability measure $\mathbb{P}^Y$. The aim of a statistical procedure is to find or approximate this distribution from the observed data. In order to do so, the statistician will prescribe a *model*, a set of probability measures $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is some arbitrary index set and $\mathcal{M} \subseteq \mathcal{P}$, where $\mathcal{P}$ denotes the set of all probability measures, sometimes also called the *full model*. Usually, $\mathcal{M}$ will be a true subset of $\mathcal{P}$ which will induce the problem that $\mathcal{M}$ might not contain the true distribution $\mathbb{P}^Y$. In this case the model is said to be *misspecified*. If indeed $\mathbb{P}^Y \in \mathcal{M}$ the model is called *well-specified*. For now, we will assume that the model is well-specified, i.e. that there exists a parameter value $\bar{\theta} \in \Theta$ such that $\mathbb{P}_{\bar{\theta}} = \mathbb{P}^Y$. Misspecified models are usually more difficult to analyse. However, considering that misspecification is a realistic assumption for many practical applications this case deserves special attention and we will consider this case briefly in Section 1.6.

When $\Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$, the model is called *parametric* and in this thesis we will confine attention to this case. In the following we will be exclusively interested in the dominated case, i.e. there exists a measure μ such that for every θ we can write

$$\mathbb{P}_\theta(\mathrm{d}y) = p(y \mid \theta)\mu(\mathrm{d}y).$$

The parametrized density $p(y \mid \theta)$ interpreted as a map $\theta \mapsto p(y \mid \theta)$ is referred to as the *likelihood function*. Relevant for this thesis will be the case where μ is the Lebesgue measure[1], which we denote $\mathrm{d}y$ (with some abuse of notation). In the next sections we will consider what statements can be made about the unknown distribution $\mathbb{P}_{\bar{\theta}}$.

1.2 Frequentist Inference

What can be said about $\mathbb{P}_{\bar{\theta}} = \mathbb{P}^Y$ on the basis of some observation $Y = y$? A common approach is to approximate $\mathbb{P}_{\bar{\theta}}$ by optimizing a cost function over all models $\mathcal{M}$. Formally, we try to find a measure $\mathbb{P}^*$

$$\mathcal{D}(\mathbb{P}_{\bar{\theta}}, \mathbb{P}^*) = \inf_{\mathbb{P} \in \mathcal{M}} \mathcal{D}(\mathbb{P}_{\bar{\theta}}, \mathbb{P}) = \inf_{\theta \in \Theta} \mathcal{D}(\mathbb{P}_{\bar{\theta}}, \mathbb{P}_\theta), \tag{1.1}$$

where $\mathcal{D}$ induces some notion of distance on $\mathcal{P}$ but not necessarily a metric. The popular Kullback–Leibler (KL) divergence, for instance, is not a metric, but it does generate a topology. If $\mathcal{D}$ is taken to be the KL-divergence the objective function becomes

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\bar{\theta}}, \mathbb{P}_\theta) &= \int_{\mathsf{Y}} \log \frac{p(y \mid \bar{\theta})}{p(y \mid \theta)} p(y \mid \bar{\theta}) \mathrm{d}y \\ &= \int_{\mathsf{Y}} \log p(y \mid \bar{\theta}) p(y \mid \bar{\theta}) \mathrm{d}y - \int \log p(y \mid \theta) p(y \mid \bar{\theta}) \mathrm{d}y. \end{aligned} \tag{1.2}$$

The first part of the objective function (1.2) can be disregarded, since it is independent of θ. The second integral, the expectation $E\{\log p(Y \mid \theta)\}$ taken under $\mathbb{P}_{\bar{\theta}}$, is generally unknown and thus (1.2) is intractable. However, under the frequentist assumption $\mathbb{P}_{\bar{\theta}}$ can be approximated by its empirical measure

$$\widehat{\mathbb{P}}_T^Y(\mathrm{d}y) = \frac{1}{T} \sum_{t=1}^{T} \delta_{Y_t}(\mathrm{d}y),$$

where δ_y denotes the Dirac measure at point y and $Y_1, \ldots, Y_T$ are independent replications drawn from $\mathbb{P}^Y$. The frequentist assumption in this context means that we suppose that the process generating Y could be repeated indefinitely yielding independent replications $Y_1, Y_2, \ldots$. Then, as the number of data, T, increases it was already shown by Borel (1909) that

$$\mathbb{P}_{\bar{\theta}}(A) = \lim_{T \to \infty} \widehat{\mathbb{P}}_T^Y(A) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \delta_{Y_t}(A)$$

[1] The Lebesgue measure is defined by $\mu((a, b]) = b - a$ for intervals $(a, b]$, $a < b$.

almost surely for every measurable set A by a strong law of large numbers. The optimization problem (1.2) then becomes

$$\inf_{\theta\in\Theta} \mathcal{D}_{\mathrm{KL}}(\widehat{\mathbb{P}}_T^Y, \mathbb{P}_\theta) = \text{constant} - \sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \log p(Y_t \mid \theta),$$

which leads to maximum likelihood estimation. The minimizing value $\hat{\theta}$ is called the *maximum likelihood* (ML) estimator. The property of approximation of probability distributions through long run frequencies is what lends its name to frequentist statistics.

1.3 Bayesian Inference

In contrast to frequentist statistics, Bayesian approaches assume a broader definition of probability. In addition to the data, we assume that the parameter θ is random as well. The parameter distribution, however, can not directly be grounded in long run frequencies since the parameter will only be indirectly observed through the observations, Y. Hence, the distribution on the parameter space Θ is defined by the statisticians degrees of believe *prior* to observing any data. This motivates naming the distribution $p(\theta)$ on Θ the *prior distribution*.

The assignment of probability measures to degrees of believes is contentious. However, the Dutch Book argument (Ramsey 1926), for instance, connects Kolmogorov's axioms of probability with rational beliefs. In brief, the Dutch Book argument states that a betting agent, say, must assign odds to events that are consistent with the axioms of probability (with finite additivity) so no set of bets with guaranteed win could be made against him[2]. Other motivations for assigning distributions on the space Θ are possible, see e.g. Jaynes (2003) who uses an approach based on logic.

Incorporating the prior distribution $p(\theta)$ on $\{\Theta, \mathcal{B}(\Theta)\}$ we can define a probability measure on the joint space $\{\mathsf{Y}\times\Theta, \mathcal{Y}\otimes\mathcal{B}(\Theta)\}$ by

$$p(y, \theta) = p(y \mid \theta)p(\theta),$$

where $p(y \mid \theta)$ is again the likelihood function. As the prior distributions $p(\theta)$ encodes the statisticians beliefs before observing data, a natural description of our knowledge after observing data is the conditional distribution

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

[2] A set of bets with guaranteed profits independent of the outcome of the gamble is called a *Dutch Book*, suggesting the name of the argument.

where

$$p(y) = \int_\Theta p(y \mid \theta) p(\theta) \mathrm{d}\theta \tag{1.3}$$

is the *marginal* or *prior predictive distribution*. In Bayesian statistics $p(\theta \mid y)$ is called the *posterior distribution* and is the main inferential object of interest. For example, prediction can be performed by virtue of the *posterior predictive distribution*

$$p(y' \mid y) = \int_\Theta p(y' \mid \theta) p(\theta \mid y) \mathrm{d}\theta.$$

However, in many cases interest lies in the original model described through the likelihood function $p(y \mid \theta)$. This model could, for instance, describe an interpretable physical process. In such cases we are not interested in integrating out the parameters as in the posterior predictive distribution. Instead, we prefer to obtain point estimators as in the frequentist setting. Common choices are the *posterior mean*

$$\hat{\theta}_{\text{post}} = \int_\Theta \theta p(\theta \mid y) \mathrm{d}\theta \tag{1.4}$$

or the *maximum a posteriori* (MAP) estimator

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \Theta} p(\theta \mid y).$$

But how do such point estimators fare against classical frequentist estimators such as the maximum likelihood estimator introduced in the previous section? To answer this question it is useful to revert back to a frequentist setting, i.e. assuming that the data follows a true distribution $\mathbb{P}^Y$ and then consider the behaviour of Bayesian posterior distribution and point estimators under classical statistical asymptotics.

1.4 Frequentist Properties of Bayesian Methods

We want to compare the Bayesian point estimators arising from the posterior distribution with classical frequentist point estimators. However, assuming the data has been generated by a true distribution, $\mathbb{P}^Y$, leads to a predicament in the Bayesian setting as we have already defined a probability distribution over the data, namely the prior predictive distribution (1.3). These distributions are in general not equal. If the data consists of independent identical observations ($i.i.d$), then $p(y)$ as a mixture of $i.i.d$. data will be only exchangeable (see Kleijn 2020, Remark 2.1). Mathematically this is important since the posterior distribution $p(\theta \mid y)$ is

only well-defined[3] almost surely with respect to the prior predictive distribution. To alleviate this issue, we will introduce the following domination assumption

$$\int_A p(y)\mathrm{d}y = 0 \Rightarrow \mathbb{P}^Y(A) = 0$$

for all $A \in \mathcal{Y}$.

In the following we will investigate the behaviour of the posterior distribution as the number of data increases. We will therefore make the dependence of the data on the number of observations more explicit and denote $y_{1:T} = (y_1, \dots, y_T)$ where $T \in \mathbb{N}$ is the number of observations. One of the most important theorems describing the posterior as $T \to \infty$ is the so-called *Bernstein–von Mises Theorem*.

1.5 Asymptotic Normality and the Bernstein–von Mises Theorem

The Bernstein–von Mises Theorem describes the behaviour of posterior distribution as the number of data, T, tends to infinity. Under regularity conditions, we have

$$\int_\Theta |p(\theta \mid y_{1:T}) - \varphi(\theta, \hat{\theta}_T, \Sigma/T)|\mathrm{d}\theta \overset{\mathbb{P}^Y}{\to} 0, \tag{1.5}$$

where $\hat{\theta}_T \to \bar{\theta}$ in $\mathbb{P}^Y$-probability, $\bar{\theta}$ is a fixed parameter, Σ a positive-definite covariance matrix and $\varphi(\,\cdot\,; \mu, \Lambda)$ denotes a normal density with mean μ and covariance matrix Λ. We have left the precise form of $(\hat{\theta}_T)_{T\geqslant 1}$ and Σ unspecified in order to keep the statement more general. In particular, (1.5) encompasses the well-specified and the misspecified case.

Before we describe the precise conditions on the statistical model we will inspect the above statement a bit further. There are three statements that are commonly derived from results like (1.5), which we want to discuss now informally:

i) **Posterior normality.** As the number of data increases, the posterior approaches a Gaussian distribution, centred around $\hat{\theta}_T$ and with variance Σ/T. If these quantities can be easily computed and the data is sufficiently large, the respective Gaussian distribution can serve as a good approximation of the posterior and can be used for inference.

ii) **Posterior concentration.** As mentioned in the previous item, the variance of the approximating Gaussian distribution concentrates at rate $1/T$ meaning that the distribution concentrates to a point mass at $\bar{\theta}$ as $T \to \infty$. Indeed, the posterior will accumulate most of its probability mass near this point when the number of data increases.

[3] With well-defined we mean that there exists a regular version of the conditional expectation for $p(y)$-almost every y.

iii) **Asymptotic independence of the prior.** Note that the limiting Gaussian distribution is not affected by the prior at all. Hence, the Bernstein–von Mises Theorem shows that the prior influence 'washes out' as more data is accumulated for all priors that fulfil the assumptions of the theorem. We will see that the requirement is very mild, only assuming that the prior is continuous and positive around $\bar{\theta}$. Since the prior effect vanishes as $T \to \infty$, the MAP estimator will be close to the ML estimator, motivating the notion that Bayesian MAP estimation is asymptotically equivalent to a maximum likelihood approach.

The Bernstein–von Mises Theorem requires the statistical model to fulfil some regularity conditions. These conditions can again be summarized in three basic ideas. Firstly, the log-likelihood needs to be close to a Gaussian with constant Hessian. Formally, this means that the log-likelihood process allows for an quadratic expansion about $\bar{\theta}$

$$\log \prod_{i=1}^{T} \frac{p(y \mid \bar{\theta} + h_T/\sqrt{T})}{p(y \mid \bar{\theta})} = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} h^\top \ell'_{\bar{\theta}}(y_i) - \frac{1}{2} h^\top I_{\bar{\theta}} h + o_{\mathbb{P}_{\bar{\theta}}}(1)$$

for any sequence $h_T \to h$ as $T \to \infty$ and where we introduce the shorthand $\ell_\theta(y) = \log p(y \mid \theta)$ and $\ell'_\theta(y)$ for its partial derivative with respect to θ. If such an expansion is valid, the model is called *locally asymptotically normal* at $\bar{\theta}$ or short LAN.

1.1 Remark. *The Bernstein-von Mises Theorem and the (asymptotic) normality of maximum likelihood estimators are closely related. In particular, the expansion of the log-likelihood is a fundamental ingredient of asymptotic theory for maximum likelihood estimation as well as posterior normality. We illustrate this by considering the maximum likelihood estimator in the context of locally asymptotically normal models. We present here a nice short argument by Geyer et al. (2013). Assume the log-likelihood is exactly quadratic, i.e.*

$$\log p(y \mid \theta) = U + \theta^\top X - \frac{1}{2} \theta^\top I_{\bar{\theta}} \theta$$

with U a random scalar, X is multivariate standard normal and $I_{\bar{\theta}}$ the Fisher information matrix at $\bar{\theta}$. Then the maximum likelihood estimator is

$$\hat{\theta} = I_{\bar{\theta}}^{-1} X.$$

Hence, for a model that is approximately quadratic in the sense of the preceding display, we would expect that the maximum likelihood estimator is approximately normal, see Geyer et al. (2013) and Van der Vaart (2000) for details.

In order to understand how local asymptotic normality leads to asymptotic posterior normality we investigate the following short heuristic argument (see also Kleijn 2020). Consider the *local parameter* $h = \sqrt{T}(\theta - \bar{\theta})$. Then the posterior is given by

$$\begin{aligned} p(\theta \mid y_{1:T}) &= \frac{\prod_{t=1}^{T} p(y_i \mid \theta + h/\sqrt{T}) p(\bar{\theta} + h/\sqrt{T})}{\int \prod_{t=1}^{T} p(y_i \mid \theta + h'/\sqrt{T}) p(\bar{\theta} + h'/\sqrt{T}) \mathrm{d}h'} \\ &\approx \frac{\exp\left(h^\top I_{\bar{\theta}} \Delta_{T,\bar{\theta}} - \frac{1}{2} h^\top I_{\bar{\theta}} h\right)}{\int \exp\left(h'^\top I_{\bar{\theta}} \Delta_{T,\bar{\theta}} - \frac{1}{2} h'^\top I_{\bar{\theta}} h'\right) \mathrm{d}h'} \\ &= \frac{\exp\left\{-\frac{1}{2}\left(h - \Delta_{T,\bar{\theta}}\right)^\top I_{\bar{\theta}} \left(h - \Delta_{T,\bar{\theta}}\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(h' - \Delta_{T,\bar{\theta}}\right)^\top I_{\bar{\theta}} \left(h' - \Delta_{T,\bar{\theta}}\right)\right\} \mathrm{d}h'}, \end{aligned}$$

where

$$\Delta_{T,\bar{\theta}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} I_{\bar{\theta}}^{-1} \ell'_{\bar{\theta}}(Y_t).$$

converges to Gaussian $X \sim \mathcal{N}(0, I_{\bar{\theta}}^{-1})$. Heuristically, $\Delta_{T,\bar{\theta}}$ can be thought of as the scaled and centred maximum likelihood estimator, see Example 1.1. Indeed, in regular models this will be true up to a function that vanishes in probability (see Van der Vaart (2000) Theorem 5.39). Thus

$$\frac{\exp\left\{-\frac{1}{2}\left(h - \Delta_{T,\bar{\theta}}\right)^\top I_{\bar{\theta}} \left(h - \Delta_{T,\bar{\theta}}\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(h' - \Delta_{T,\bar{\theta}}\right)^\top I_{\bar{\theta}} \left(h' - \Delta_{T,\bar{\theta}}\right)\right\} \mathrm{d}h'} \to \frac{\exp\left\{-\frac{1}{2}(h - X)^\top I_{\bar{\theta}} (h - X)\right\}}{\int \exp\left\{-\frac{1}{2}(h' - X)^\top I_{\bar{\theta}} (h' - X)\right\} \mathrm{d}h'}$$

in distribution, where the right-hand side coincides with a Gaussian density $\varphi(h; X, I_{\bar{\theta}}^{-1})$. Turning the above argument into something reminiscent of (1.5) requires a bit more work and further assumptions. One additional assumption, however, can be seen from the above display. For the prior influence to vanish we require $p(\theta)$ to be positive and continuous at $\bar{\theta}$. A prior distribution that fulfils such a condition is called *thick* at $\bar{\theta}$.

We now have established that local asymptotic normality is a key ingredient for asymptotic posterior normality. But when can we expect a model to be locally asymptotically normal? It turns out, that in the case of $i.i.d.$ observations the necessary conditions are relatively weak. The right condition to ensure the LAN condition is *differentiability in quadratic mean* (DQM), see e.g. Van der Vaart (2000).

1.2 Definition. Let $p(y \mid \theta)$ be a likelihood of an i.i.d. model. Then p is called *differentiable in quadratic mean* at θ if there exists $\varsigma_\theta \in L^2(\mathbb{P}_\theta)$, such that

$$\int \left(\sqrt{p(y \mid \theta + h)} - \sqrt{p(y \mid \theta)} - \frac{1}{2} h^\mathsf{T} \varsigma_\theta(y) \sqrt{p(y \mid \theta)} \right)^2 \mathrm{d}y \Big/ \|h\|^2 \to 0,$$

as $h \to 0$.

The function $\varsigma_\theta(y)$ is called the *score* and will often coincide with the classical definition, i.e. $\varsigma_\theta(y) = \ell'_\theta(y)$. If $h \mapsto \sqrt{p(y \mid \theta + h)}$ is differentiable at $h = 0$ then

$$\frac{\partial}{\partial h} \sqrt{p(y \mid \theta + h)} = \frac{1}{2} \varsigma_\theta(y) \sqrt{p(y \mid \theta)}$$

and hence

$$\varsigma_\theta(y) = \frac{p'(y \mid \theta)}{p(y \mid \theta)} = \frac{\partial}{\partial \theta} \log p(y \mid \theta).$$

The assumption of continuous differentiability of the root density together with a condition on the Fisher information matrix is indeed sufficient for differentiability in quadratic mean.

1.3 Lemma. *Let $\theta \mapsto \sqrt{p(y \mid \theta)}$ be continuously differentiable for every y and assume that the elements of the matrix*

$$I_\theta = E_\theta \left[\ell'(\theta) \ell'(\theta)^\mathsf{T} \right]$$

are well defined and continuous in θ. Then the model $p(y \mid \theta)$ is differentiable in quadratic mean and $\varsigma_\theta(y) = \ell'(\theta)$.

Proof. This is Lemma 7.6 in Van der Vaart (2000). □

We can now state the result linking differentiability in quadratic mean to local asymptotic normality.

1.4 Theorem. *Let $\Theta \subset \mathbb{R}^d$ be open and assume that the model $\{p(y \mid \theta); \theta \in \Theta\}$ is differentiable in quadratic mean at θ. Then the score function has expectation 0, i.e.*

$$E_\theta \left[\varsigma_\theta(Y) \right] = 0$$

and the Fisher information matrix I_θ exists. In addition, the model is locally asymptotically normal, i.e. for every sequence $h_T \to h$ as $T \to \infty$

$$\log \prod_{i=1}^{T} \frac{p(y \mid \theta + h_T/\sqrt{T})}{f(y \mid \theta)} = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} h^\mathsf{T} \ell'_\theta(y_i) - \frac{1}{2} h^\mathsf{T} I_\theta h + o_{\mathbb{P}_\theta}(1)$$

Proof. This is Theorem 7.2 in Van der Vaart (2000). □

The above condition of is remarkably weak; we do not even require differentiability of the density. This is in contrast to the standard Cramer conditions (Ferguson 2017, Theorem 18 or Van der Vaart 2000, Section 5.6).

1.5 Theorem (Bernstein–von Mises)**.** *Let* $(\mathbb{P}_\theta : \theta \in \Theta)$ *be differentiable in quadratic mean at* $\bar{\theta}$ *with non-singular Fisher information matrix* $I_{\bar{\theta}}$*. Suppose further that the prior density is continuous and positive in a neighbourhood of* $\bar{\theta}$*. Further assume that for any* $\varepsilon > 0$ *there exists a sequence of tests* $\phi_T : \mathcal{Y}^T \to [0,1]$ *such that we have the uniform testing condition*

$$\int \phi_T(y_{1:T})\mathbb{P}_{\bar{\theta}}(\mathrm{d}y_{1:T}) \to 0, \qquad \sup_{\|\theta-\bar{\theta}\|\geqslant\varepsilon} \int \left(1-\phi_T(y_{1:T})\right)\mathbb{P}_\theta(\mathrm{d}y_{1:T}) \to 0 \tag{1.6}$$

as $T \to \infty$*. Then*

$$\int \left| p(\theta \mid y_{1:T}) - \varphi\left(\theta; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}/\sqrt{T}\right)\right| \mathrm{d}\theta \to 0$$

in $\mathbb{P}^Y$*-probability where*

$$\Delta_{T,\bar{\theta}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} I_{\bar{\theta}}^{-1} \ell'_{\bar{\theta}}(Y_t).$$

Proof. This is Van der Vaart (2000, Theorem 10.1). The proof is very similar to the one shown later in Section 4.11.2. □

The Bernstein–von Mises Theorem requires one further assumption that we have not investigated in more detail, namely the testing condition (1.6) which requires that we are able to separate the true model from complements of balls centred around the true parameter. For compact $\Theta \subset \mathbb{R}^d$ such tests exist under mere regularity conditions. It is sufficient, for example, to require $\theta \mapsto F_\theta$ to be continuous and identifiable, see e.g. the discussion in Van der Vaart (2000, pp. 144-145).

For non-compact parameter spaces establishing such testing conditions can be quite difficult. However, modifying the model assumptions one can circumvent the requirement for the tests to be uniformly consistent on the whole of (non-compact) Θ. For example, in Section 4.11.2 we will show that a different version of the Bernstein–von Mises Theorem holds when the prior has exponential tails. Such statements are unusual in the literature since they partially defeat the purpose of the theorem (item *iii*), asymptotic independence of the prior) in our initial outline. However, for the purposes of this thesis we will be mostly concerned with points *i)* and *ii)*.

1.6 Model Misspecification

Our previous sections all assume that the true data generating process can be identified as a member of the statistical models under considerations, i.e.

$$\mathbb{P}^Y \in \mathcal{M} = \left\{\mathbb{P}_\theta : \theta \in \Theta\right\}.$$

In other words, there exists $\bar{\theta} \in \Theta$ such that $\mathbb{P}_{\bar{\theta}} = \mathbb{P}^Y$.

In practical application, where the data may arise from natural experiments such an assumption is unrealistic. Well-specified models are usually easier to analyse and therefore serve as good first starting point for a theoretical investigation. However, in general it is to expect that 'all models are wrong' (Box 1976) and hence well-specification is unrealistic. Analysing models in the misspecified case, however, offers an understanding on their robustness.

Our main interest in this section is to establish that the Gaussian posterior approximation established through the Bernstein–von Mises Theorem still holds in the misspecified setting under mild assumptions. For a more detailed review of Bayesian models in the misspecified case, see Kleijn (2020) and Kleijn and Van der Vaart (2012, Chapter 5).

Of course, the value $\bar{\theta}$ around which the posterior concentrates cannot be the the 'true' parameter because such an element does not exist. However, in many cases the sequence $\hat{\theta}_T \to \bar{\theta}$ can be taken as the maximum likelihood estimators for the respective data and we have already seen in Section 1.2 that maximum likelihood theory aims to minimize the Kullback–Leibler divergence. This suggest to take $\bar{\theta}$ as the model closest to the true distribution in Kullback–Leibler divergence, i.e.

$$\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\bar{\theta}}, \mathbb{P}^Y) = \inf_{\theta\in\Theta} \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_\theta, \mathbb{P}^Y).$$

1.6 Theorem (Misspecified Bernstein–von Mises Theorem)**.** *Let $\bar{\theta}$ be as above and assume that the prior is positive and continuous at $\bar{\theta}$ and let $\theta \mapsto \log p(y \mid \theta)$ be differentiable at $\bar{\theta}$ with $\mathbb{P}^Y$-probability one. Assume there exists an open neighbourhood U around $\bar{\theta}$ and a square integrable function $m_{\bar{\theta}}$ with*

$$\left|\log \frac{p(y \mid \theta')}{p(y \mid \theta)}\right| \leqslant m_{\bar{\theta}}(y)\|\theta' - \theta\|, \quad \theta, \theta' \in U, \tag{1.7}$$

$\mathbb{P}^Y$-almost surely. Further, assume the second-order expansion

$$-\int \log \frac{p(y \mid \theta)}{p(y \mid \bar{\theta})} \mathbb{P}^Y(\mathrm{d}y) = \frac{1}{2}(\theta - \bar{\theta})^\intercal V_{\bar{\theta}}(\theta - \bar{\theta}) + o(\|\theta - \bar{\theta}\|^2), \quad (\theta \to \bar{\theta}), \tag{1.8}$$

where $V_{\bar{\theta}}$ is a positive-definite matrix with dimensions $d \times d$. In addition, assume that

$$\int \mathbb{1}\left\{\|\theta - \bar{\theta}\| \geqslant M_T/\sqrt{T}\right\} \pi_T(\mathrm{d}\theta) \to 0, \tag{1.9}$$

in $\mathbb{P}^Y$-probability for all $M_T \to \infty$. Then

$$\int |p(\theta \mid y_{1:T}) - \varphi\left(\theta; \Delta_{T,\bar{\theta}}, V_{\bar{\theta}}^{-1}/\sqrt{T}\right)| \mathrm{d}\theta \to 0$$

in $\mathbb{P}^Y$-probability where

$$\Delta_{T,\bar{\theta}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{\bar{\theta}}^{-1} \ell'_{\bar{\theta}}(Y_t).$$

Proof. This follows from Lemma 2.1 and Theorem 2.1 in Kleijn and Van der Vaart (2012). □

Let us relate this result to the standard version of the Bernstein–von Mises Theorem 1.5. It is no longer enough that the model is mean square differentiable to establish the LAN expansion. Now we need differentiability in conjunction with (1.7) and (1.8) in order to ensure a modified (stochastic) LAN expansion Kleijn and Van der Vaart (see 2012, Lemma 2.1). The posterior concentration, in turn, is related to the uniform testing condition in Theorem 1.5. Indeed, under additional regularity conditions Kleijn and Van der Vaart (2012, Theorem 3.1) show that the testing condition (1.6) implies (1.9).

References

Borel, M Émile (1909). 'Les probabilités dénombrables et leurs applications arithmétiques'. *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 27.1, pp. 247–271 (cit. on p. 6).

Box, George EP (1976). 'Science and statistics'. *Journal of the American Statistical Association* 71.356, pp. 791–799 (cit. on p. 14).

Ferguson, Thomas S (2017). *A course in large sample theory*. Routledge (cit. on p. 13).

Geyer, Charles J et al. (2013). 'Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity'. *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*. Institute of Mathematical Statistics, pp. 1–24 (cit. on p. 10).

Jaynes, Edwin T (2003). *Probability theory: The logic of science*. Cambridge university press (cit. on p. 7).

Kleijn, Bas J. K. (2020). *The frequentist theory of Bayesian statistics*. Springer, New York (cit. on pp. 8, 11, 14).

Kleijn, Bas J K and Aad W Van der Vaart (2012). 'The Bernstein-Von-Mises theorem under misspecification.' *Electron. J. Statist.* 6, pp. 354–381 (cit. on pp. 14–15).

Ramsey, F.P. (1926). 'Truth and Probability'. *Ramsey, 1931, The Foundations of Mathematics and other Logical Essays, Ch. VII*. Ed. by R.B. Braithwaite. London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company., pp. 156–198 (cit. on p. 7).

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Cambridge University Press (cit. on pp. 10–13).

2

Markov Chain Monte Carlo Simulation

2.1 *Monte Carlo Simulation for Bayesian Posteriors*

As outlined in the previous section, Bayesian posterior distributions are often not analytically tractable. In particular, the normalizing constant

$$p(y) = \int p(y \mid \theta) p(\theta) \mathrm{d}\theta$$

in the Bayesian update

$$p(\theta \mid y) = \frac{p(y \mid \theta) p(\theta)}{p(y)}$$

is usually not available analytically and thus the posterior distribution cannot be computed exactly. However, underlying the prevalence of Monte Carlo algorithms in Bayesian statistics is the idea that properties of any distribution can be learned by drawing samples from said distribution. Hence, the aim is to construct algorithms to sample from the posterior which we denote $\pi(\theta) = p(\theta \mid y)$ for brevity, suppressing for now the dependency on the data. Thus, the difficulty of performing Bayesian inference has shifted from calculating $\pi(A)$ for measurable sets A to being able to draw samples from this distribution. We will see in the following that sampling (at least approximately) can be achieved with fairly simple algorithms.

2.2 *Markov chain Monte Carlo*

Among the class of Monte Carlo algorithm, Markov chain Monte Carlo (MCMC) methods have become particularly popular in Bayesian inference as they do not require access to the intractable normalizing constant $p(y)$. The aim of Markov chain Monte Carlo algorithms is to

simulate a Markov chain $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$ such that for n large enough

$$\mathbb{P}(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_0 = \theta_0) \approx \pi(A) \tag{2.1}$$

for all (measurable) sets A. Here, $\boldsymbol{\theta}_0 = \theta_0$ denotes an arbitrary starting point of the Markov chain. In general, Markov chains can be started from arbitrary distributions and the previous display then just corresponds to a Dirac mass at θ_0. Of course, if the Markov chain is already started from the target distribution π a desirable property would be that this distribution is retained after application of one time step, formally,

$$\int_\Theta P(\theta, A)\pi(\theta)\mathrm{d}\theta = \pi(A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d) \tag{2.2}$$

where

$$P(\theta, A) = \mathbb{P}(\boldsymbol{\theta}_1 \in A \mid \boldsymbol{\theta}_0 = \theta)$$

is the one-step *transition kernel*. The distribution π that fulfils (2.2) for a transition kernel P is called the Markov chain's *stationary* or *invariant* distribution. This condition is not always easy to verify. A stronger condition that implies stationarity of P with respect to π is reversibility

$$\int_A P(\theta, B)\pi(\mathrm{d}\theta) = \int_B P(\theta, A)\pi(\mathrm{d}\theta)$$

for all measurable sets A, B.

2.1 Proposition. *If a Markov transition kernel P is reversible with respect to π then π is an invariant distribution.*

Proof. Straightforward calculation yields

$$\begin{aligned}\int_\Theta P(\theta, A)\pi(\mathrm{d}\theta) &= \int_A P(\theta, \Theta)\pi(\mathrm{d}\theta) \\ &= \int_A \pi(\mathrm{d}\theta) = \pi(A).\end{aligned}$$

□

We want to make the statement (2.1) more precise and therefore we need to introduce some notation. We begin with the n-step transition probability which can be defined recursively by

$$P^n(\theta, A) = \int P(\theta', A)P^{n-1}(\theta, \mathrm{d}\theta') = \int P^{n-1}(\theta', A)P(\theta, \mathrm{d}\theta').$$

To measure the distance between the Markov chain after n steps and the target distribution we

introduce the *total variation distance* between probability measures ν, μ defined by

$$\|\nu - \mu\|_{\text{tv}} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\nu(A) - \mu(A)|.$$

2.2 Remark. *If the two probability measures are dominated by a common measure, ρ, say, the total variation distance has a simple expression as half the L^1-distance between the respective densities (see e.g. Roberts and Rosenthal 2004, Proposition 3)*

$$\|\nu - \mu\|_{\text{tv}} = \frac{1}{2} \int_{\mathbb{R}} \left| \frac{\mathrm{d}\nu}{\mathrm{d}\rho}(y) - \frac{\mathrm{d}\mu}{\mathrm{d}\rho}(y) \right| \rho(\mathrm{d}y),$$

where we write $\frac{\mathrm{d}\nu}{\mathrm{d}y}$ for the Radon–Nikodym derivative of ν with respect to $\mathrm{d}y$.

We will say that a Markov chain is *ergodic* if the marginal distribution of the current state approaches the invariant distribution in the limit, i.e.

$$\|P^n(\theta, \cdot) - \pi(\cdot)\|_{\text{tv}} \to 0 \tag{2.3}$$

as $n \to \infty$ for π-almost every starting point θ. Convergence in total variation as defined in (2.3) has the useful implication that $P^n(\theta, A) \to \pi(A)$ for any measurable set. Other notions of convergence placing emphasis on other aspects of distributions are also possible. Convergence analyses in Wasserstein distance, for example, have become more and more popular (see e.g. Rudolf and Schweizer 2018). Here we will only consider convergence in total variation. Ergodicity can be established under fairly weak conditions as we will demonstrate now.

2.3 Theorem. *Consider a Markov chain with state space $\mathbb{R}^d$. Assume the Markov chain is ϕ-irreducible and aperiodic. Then the Markov chain is ergodic in the sense of* (2.3).

Proof. This is Theorem 4 in Roberts and Rosenthal (2004). □

ϕ-irreducibility and aperiodicity are mild conditions. A chain is called ϕ-irreducible if there exists a non-trivial σ-finite measure on Θ such that for all measurable set A with $\phi(A) > 0$ and all $\theta \in \Theta$ there exists an integer $n(\theta, A)$ with $P^n(\theta, A) > 0$ (e.g. Roberts and Rosenthal 2004). Hence, this condition ensures that all sets with positive measure are eventually reached. In continuous state spaces such a measure could be the Lebesgue measure. Aperiodicity ensures that there do not exist disjoint subsets the Markov chain cycles through periodically. Formally, we require that there does not exist disjoint subset $\Theta_1, \Theta_2, \ldots, \Theta_d, d \geqslant 2$, such that $P(\theta, \Theta_{i+1}) = 1$ for all $\theta \in \Theta_i$ and $P(\theta, \Theta_1) = 1$ for $\theta \in \Theta_d$.

Very often we are interested in the estimation of expectations of functions $h : \Theta \to \mathbb{R}$ with respect to the target distribution π. A common Markov chain Monte Carlo approximation consists of the ergodic average

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(\theta_i).$$

We shall note that Theorem 2.3 ensures the convergence of such averages as $n \to \infty$,

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(\theta_i) \to \pi(h)$$

almost surely (Roberts and Rosenthal 2004).

While it might appear cumbersome to have to construct a Markov chain that converges to a particular target distribution, it turns out that simple 'accept-reject'-type recipes provide powerful and easy-to-implement algorithms. In the next section we will introduce one such recipe, called the Metropolis–Hastings algorithms.

2.3 *Metropolis–Hastings Algorithm*

Originally introduced by Metropolis et al. (1953) and extended by Hastings (1970), the Metropolis–Hastings algorithm is one of the most popular MCMC algorithms, both for its simplicity and efficiency. The algorithm proceeds as follows: Given a current state of the chain, θ_k for some $k \geqslant 1$, a new state is proposed according to some distribution $\theta' \sim q(\theta_k, \cdot)$. Then with probability $\alpha(\theta_k, \theta')$, given by

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')} \right\}$$

set $\theta_{k+1} = \theta'$, otherwise set $\theta_{k+1} = \theta_k$. This induces a transition kernel

$$P_{\mathrm{MH}}(\theta, \mathrm{d}\theta') = \alpha(\theta, \theta') q(\theta, \theta') \mathrm{d}\theta' + \rho(\theta) \delta_\theta(\mathrm{d}\theta')$$

where

$$\rho(\theta) = 1 - \int_{\mathbb{R}^d} \alpha(\theta, \theta') q(\theta, \theta') \mathrm{d}\theta'.$$

The above transition kernel is reversible, i.e. for $\theta \neq \theta'$ we have

$$\begin{aligned} \pi(\mathrm{d}\theta) P_{\mathrm{MH}}(\theta, \mathrm{d}\theta') &= \pi(\theta) q(\theta, \theta') \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')} \right\} \mathrm{d}\theta \mathrm{d}\theta' \\ &= \min \left\{ \pi(\theta) q(\theta, \theta'), \pi(\theta') q(\theta', \theta) \right\} \mathrm{d}\theta \mathrm{d}\theta' \end{aligned}$$

$$= \pi(\mathrm{d}\theta')P_{\mathrm{MH}}(\theta', \mathrm{d}\theta)$$

and therefore P_{MH} has π as its invariant distribution.

It can subsequently be established that the Metropolis–Hastings algorithm using a proposal q that is continuous and positive in both components and targeting the Lebesgue density π is π-irreducible and aperiodic (see the running example in Roberts and Rosenthal 2004). To fulfil these assumption, the proposal could be taken as a multivariate normal distribution, i.e.

$$q(\theta, \theta') = \varphi(\theta'; \theta, l \cdot \Sigma),$$

where Σ is some covariance matrix, l is a tuning parameter and $\varphi(\,\cdot\,; \mu, \Lambda)$ denotes a normal density with mean μ and covariance matrix Λ. The Metropolis–Hastings algorithm with this proposal—called the *random walk Metropolis–Hastings*—is therefore ergodic.

2.4 Remark (Choice of the proposal)**.** *The performance of any implemented Metropolis–Hastings algorithm crucially depends on the proposal distribution $q(\theta, \theta')$. For this thesis we will only consider random walk proposal, i.e. $q(\theta, \theta') = \nu(\theta' - \theta)$.*

2.5 Remark (Extensions and alternatives)**.** *Since the original Metropolis algorithm, many extensions and alternatives have been introduced in the literature. Amongst them state-of-the-art algorithms such as the Gibbs sampler (Geman and Geman 1984), Hamiltonian-Monte Carlo (Duane et al. 1987), the bouncy particle sampler (Bouchard-Côté, Vollmer and Doucet 2018). However, a treatment of these algorithms is beyond the scope of this thesis.*

2.6 Remark (Sampling from Bayesian posteriors)**.** *When implementing the Metropolis–Hastings algorithms to sample from Bayesian posterior distribution we are interested in $\pi(\theta) \propto p(y \mid \theta)p(\theta)$. Since the parameter does not appear in the normalizing constant, $p(y)$, the evaluation of $p(y \mid \theta)p(\theta)$ is sufficient to successfully sample from the posterior.*

2.4 Ergodicity and Spectral Gaps

In Theorem 2.3 we have already introduced conditions to ensure that a Markov chain $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$ is ergodic, i.e. it converges to its stationary distribution in total variation distance. However, this statement is of limited practical use, since it does not provide any guaranteed *rate of convergence* and thus convergence to the stationary distribution could be arbitrarily slow. For this reason, we introduce a stronger notion of convergence that does provide such a rate and we will describe conditions under which such convergence holds. The following qualitative

concept of convergence is commonly used in practice.

2.7 Definition (Geometric ergodicity)**.** A Markov chain is called *geometrically ergodic* if there exists $\gamma < 1$ and a function $C(\theta) < \infty$ such that

$$\|P^n(\theta, \cdot) - \pi\|_{\text{tv}} \leqslant C(\theta)\gamma^n, \qquad n = 1, 2, \ldots$$

for π-almost every θ.

In contrast to showing ergodicity, establishing geometric ergodicity for Metropolis–Hastings algorithms can be quite challenging. The result is usually proven by establishing a drift condition towards a small set (see e.g. Jones and Hobert 2001). For random walk Metropolis–Hastings algorithms necessary and sufficient conditions on the target and proposal are given by Jarner and Hansen (2000) and involve the curvature of the target distribution.

We want to introduce an alternative condition for reversible Markov chains arising from the spectral theory of bounded self-adjoint operators. Let P denote a reversible Markov transition kernel. We want to explore the properties of P as an operator acting on the Lebesgue space $L^2(\pi)$ defined by

$$Pf(\theta) = \int_\Theta f(\theta')P(\theta, \mathrm{d}\theta').$$

By definition, reversibility of P with respect to π can be written as

$$\int_\Theta \mathbb{1}_A(\theta) \int_\Theta \mathbb{1}_B(\theta')P(\theta, \mathrm{d}\theta')\pi(\mathrm{d}\theta) = \int_\Theta \mathbb{1}_B(\theta) \int_\Theta \mathbb{1}_A(\theta')P(\theta, \mathrm{d}\theta')\pi(\mathrm{d}\theta)$$

for all measurable sets A, B. Rewriting the above display we get

$$\langle P\mathbb{1}_B, \mathbb{1}_A\rangle_\pi = \langle \mathbb{1}_B, P\mathbb{1}_A\rangle_\pi\,, \tag{2.4}$$

where

$$\langle f, g\rangle_\pi = \int_\Theta f(\theta)g(\theta)\pi(d\theta)$$

is the usual inner product in $L^2(\pi)$. A standard argument using monotone convergence can generalize equality (2.4) to simple functions, non-negative measurable functions and ultimately to all functions in $L^2(\pi)$, yielding

$$\langle Pf, g\rangle_\pi = \langle f, Pg\rangle_\pi \qquad \forall f, g \in L^2(\pi). \tag{2.5}$$

Hence, the operator P acting on function in $L^2(\pi)$ is self-adjoint. Denoting the norm $\|f\|_\pi =$

$\sqrt{\langle f, f\rangle_\pi}$ we can find the bound

$$\begin{aligned}\|Pf\|_\pi^2 &= \int_\Theta \left| \int_\Theta f(\theta')P(\theta, \mathrm{d}\theta') \right|^2 \pi(\mathrm{d}\theta) \\ &\leqslant \int_\Theta \left|f(\theta')\right|^2 P(\theta, \mathrm{d}\theta')\pi(\mathrm{d}\theta) \\ &= \int_\Theta \left|f(\theta')\right|^2 \pi(\mathrm{d}\theta') \\ &= \|f\|_\pi^2.\end{aligned}$$

This ensures that the operator norm $\|P\|_{\mathrm{op}} = \sup_{f \in L^2(\pi)} \|Pf\|_\pi / \|f\|_\pi$ is bounded by one. That is, we know that P is a bounded operator and we can leverage spectral theory for bounded self-adjoint operators. The spectrum of an operator P is defined by the set

$$\sigma(P) = \{\lambda : \ I - \lambda P \text{ is not invertible}\}.$$

Using the $L^2(\pi)$ contraction property above and the fact that $\sigma(P)$ is contained in the closure of the numerical range (Gustafson and Rao 1997, Theorem 1.2-1)

$$r(P) = \left\{\langle Pf, f\rangle : \ f \in L^2(\pi), \|f\|_\pi = 1\right\}$$

it follows that $\sigma(P) \subset [-1, 1]$. Indeed, we know that $P1 = 1$ and thus every Markov kernel has eigenvalue $\lambda = 1$. What can be said about the rest of the spectrum? We restrict attention to the orthogonal complement to the constant functions in $L^2(\pi)$, i.e.

$$L_0^2(\pi) = \{f \in L^2(\pi) : \pi(f) = 0\}.$$

Indeed, for all $f \in L_0^2(\pi)$ we can compute $\langle f, 1\rangle_\pi = E_\pi\{f(X)\} = 0$ and thus $L^2(\pi) = L_0^2(\pi)^\perp \oplus L_0^2(\pi)$. This last equality means that we can write any $f \in L^2(\pi)$ as $f = \langle f, 1\rangle_\pi 1 + f_0$ with $f_0 \in L_0^2(\pi)$. When considering the n-step transition operator P^n we have

$$P^n f = E_\pi\{f(X)\} + P^n f_0. \tag{2.6}$$

Thus, in order to ensure $P^n f \to E_\pi\{f(X)\}$ we need to argue that $P^n f_0 \to 0$. For clarity we will use the notation $P|_{L_0^2(\pi)}$ for P acting on $L_0^2(\pi)$. Then for $f_0 \in L_0^2(\pi)$

$$\left\langle P|_{L_0^2(\pi)}^n f_0, P|_{L_0^2(\pi)}^n f_0 \right\rangle_\pi \leqslant \|P|_{L_0^2(\pi)}^{2n}\|_{\mathrm{op}} \|f_0\|_\pi$$

and thus we obtain our desired result if we can show $\|P|^{2n}_{L^2_0(\pi)}\|_{\text{op}} \to 0$. Here, $\|P|_{L^2_0(\pi)}\|_{\text{op}} = \lambda = \max\left\{|\lambda_{\min}|, |\lambda_{\max}|\right\}$, where $\lambda_{\min}, \lambda_{\max}$ are the smallest and largest eigenvalues of P acting on $L^2_0(\pi)$, respectively[1]. Since the operator norm is submultiplicative it holds $\|P^{2n}\|_{\text{op}} \leqslant \lambda^{2n}$ and thus $Pf_0 \to 0$ for $f_0 \in L^2_0(\pi)$ if the spectral radius, λ, is strictly smaller than one. Heuristically, this means that there is a gap between the largest eigenvalue and the remaining spectrum and thus an operator with the above property is said to possess a *spectral gap*. Spectral gaps turn out the be very useful when analysing the behaviour of reversible Markov chains. Mathematically, we can use the definition of the numerical range of $P|_{L^2_0(\pi)}$

$$r(P|_{L^2_0(\pi)}) = \left\{ \langle P|_{L^2_0(\pi)} f, f\rangle : \ f \in L^2_0(\pi), \|f\|_\pi = 1 \right\}$$

to write the right spectral gap as the distance $b_{\text{right}}(P) = 1 - \lambda_{\max}$ and the left spectral gap as $b_{\text{left}}(P) = 1 + \lambda_{\min}$. To find these quantities we can write

$$\begin{aligned} b_{\text{right}}(P) &= 1 - \lambda_{\max} \\ &= 1 - \sup_{f \in L^2_0(\pi), \|f\|_\pi = 1} \langle Pf, f\rangle \\ &= \inf_{f \in L^2_0(\pi), \|f\|_\pi = 1} \langle f, f\rangle - \langle f, Pf\rangle \\ &= \inf_{f \in L^2_0(\pi), \|f\|_\pi = 1} \langle f, (I - P)f\rangle. \end{aligned}$$

The expression inside the infimum motivates the definition of the Dirichlet form of a Markov chain with transition kernel P and stationary distribution π as

$$\mathcal{E}_P(f) = \langle f, (I - P)f\rangle_\pi = \frac{1}{2} \int \left\{f(\theta) - f(\theta')\right\}^2 \pi(\mathrm{d}\theta) P(\theta, \mathrm{d}\theta').$$

Using the definition of the Dirichlet form, the right spectral gap of a Markov chain is then

$$b_{\text{right}} = \inf_{f : \mu(f) = 0, \|f\|_\mu = 1} \mathcal{E}_P(f).$$

In analogy to the above one can similarly calculate the left spectral gap as

$$b_{\text{left}} = \inf_{f : \mu(f) = 0, \|f\|_\mu = 1} \left\{2 + \mathcal{E}_P(f)\right\}.$$

[1] For self-adjoint operators, operator norm, spectral radius and the radius of the numerical range are identical (Gustafson and Rao 1997, Theorems 1.2-3 and 1.2-4). Hence, we change our point of view whenever convenient for the argument.

The following theorem connects the spectral gap of a Markov chain with geometric ergodicity.

2.8 Theorem. *Let P denote a Markov transition kernel of a reversible Markov chain (i.e. P is self-adjoint in $L^2(\pi)$). Then the chain is geometrically ergodic if and only if P possesses an $L^2(\pi)$ spectral gap.*

Proof. This follows from Theorem 2.1 in Roberts and Rosenthal (1997). □

2.5 Markov Chain Central Limit Theorems

In practice, the objects of interest in Bayesian inference are integrals with respect to the posterior distribution such as the posterior mean $\hat{\theta}_{\text{post}}$ defined in (1.4). For $h : \Theta \to \mathbb{R}$ denote the expectation with respect to the posterior as

$$\pi(h) = \int_{\Theta} h(\theta)\pi(\mathrm{d}\theta).$$

This quantity can be approximated by the ergodic average of the Markov chain

$$\hat{\pi}(h) = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}_i)$$

where n denotes the number of Markov chain iterations, i.e. the length of the simulated Markov chain. In order to quantify the uncertainty of the estimate $\hat{\pi}(h)$, e.g. establishing (approximate) confidence intervals, it is helpful to derive central limit theorems for such Markov chain estimates. We say that a Markov chain satisfied a $\sqrt{n}$-central limit theorem if

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(h(\boldsymbol{\theta}_i) - \pi(h)\big) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(h, P))$$

where

$$\mathrm{Var}(h, P) = \lim_{n\to\infty} E\left[\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(h(\boldsymbol{\theta}) - \pi(h))\right\}^2\right].$$

When $\mathrm{Var}(h, P) < \infty$ we can write

$$\mathrm{Var}(h, P) = \mathrm{IAT}(h, P)\mathrm{Var}_{\pi}(h)$$

where $\text{IAT}(h, P)$ is the integrated autocorrelation time defined by

$$\text{IAT}(h, P) = 1 + 2 \sum_{i=1}^{\infty} \frac{\text{Cov}\left\{h(\theta_0), h(\theta_i)\right\}}{\text{Var}_\pi(\theta_0)}.$$

Hence, the integrated autocorrelation time can be seen as a measure of inefficiency due to the correlation in the Markov chain sampler. This inefficiency is sometimes measured by the number of independent samples one would need to obtain the same variance. To see this, denote n^* the number of independent samples from π, then equating

$$\frac{\text{Var}_\pi(h)}{n^*} = \frac{\text{Var}_\pi(h)\text{IAT}(h, P)}{n}$$

we can write $n^* = n/\text{IAT}(h, P)$. For this reason n^* is referred to as the *effective sample size*. The following theorem is due to Kipnis and Varadhan (1986).

2.9 Theorem. *Let $(\theta_n)_{n\in\mathbb{N}}$ be reversible, ϕ-irreducible and aperiodic (compare this to Theorem 2.3). Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{h(\theta_i) - \pi(h)\right\} \rightsquigarrow \mathcal{N}\left\{0, \text{Var}(h, P)\right\}$$

whenever

$$\text{IAT}(h, P) < \infty.$$

Hence, a central limit theorem can be established for some function h whenever we can establish that the integrated autocorrelation time $\text{IAT}(h, P) < \infty$. As a result of the above theorem Kipnis and Varadhan (1986) also provide an alternative formula of the asymptotic variance and hence for the integrated autocorrelation time.

2.10 Corollary. *Consider an aperiodic and ergodic Markov chain evolving according to some π-reversible Markov kernel P. Let $h \in L^2(\pi)$ with non-zero variance. Then there exists a spectral measure $e(h, P)$ on $[-1, 1]$ such that*

$$\text{Var}_\pi(h) = \int_{-1}^{1} e(h, P)(\text{d}\lambda), \qquad \text{Cov}_\pi\left\{h(X_0), h(X_k)\right\} = \int_{-1}^{1} \lambda^k e(h, P)(\text{d}\lambda)$$

and

$$\text{Var}(h, P) = \int_{-1}^{1} \frac{1+\lambda}{1-\lambda} e(h, P)(\text{d}\lambda).$$

Proposition 2.10 relates the asymptotic variance and therefore the integrated autocorrelation

time to the spectrum of the self-adjoint operator P. Using this connection, the following result provides a sufficient condition for $\mathrm{IAT}(h, P)$ to be finite.

2.11 Corollary. *Assume that a reversible Markov transition kernel has a (right) spectral gap,* $b_{\mathrm{right}} > 0$. *Then for* $h \in L^2(\pi)$ *we have* $\mathrm{IAT}(h, P) < \infty$.

Proof. Denote by $b = b_{\mathrm{right}} > 0$ the spectral gap of the operator P. Then

$$\mathrm{Var}(h, P) = \int_{-1}^{1-b} \frac{1+\lambda}{1-\lambda} e(h, P)(\mathrm{d}\lambda) \leqslant \frac{2-b}{b} \mathrm{Var}_{\pi}(h). \qquad \square$$

2.6 Pseudo-Marginal Metropolis–Hastings Algorithms

Even though the Metropolis–Hastings algorithm only requires access to an unnormalized version of the target density, in many situations even this quantity is not tractable. A common example are latent variable models consisting of latent data X and observed data Y given by densities

$$Y \mid (X = x) \sim f(\cdot \mid x, \theta) \quad \text{and} \quad X \sim g(\cdot \mid \theta).$$

The (marginal) likelihood function

$$p(y \mid \theta) = \int f(y \mid x, \theta) g(x \mid \theta) \mathrm{d}x$$

is intractable in general but it can be estimated unbiasedly. Sample, for example, $X \sim g(\cdot \mid \theta)$ and set $\hat{p}(y \mid \theta) = f(y \mid X, \theta)$. This is an unbiased estimator since

$$E\{\hat{p}(y \mid \theta)\} = \int f(y \mid x, \theta) g(x \mid \theta) \mathrm{d}x.$$

Of course, instead of sampling from $g(\cdot \mid \theta)$ any importance sampling estimator can be employed. The naïve approach is to apply the Metropolis–Hastings algorithm to sample from the Bayesian posterior $\pi(\theta) \propto p(y \mid \theta) p(\theta)$ by replacing the true likelihood $p(y \mid \theta)$ with the unbiased estimate. Remarkably, if correctly implemented, such a replacement does preserve the correct target distribution. The resulting algorithm is called the *pseudo-marginal Metropolis–Hastings algorithm* (Andrieu, Doucet and Holenstein 2010; Andrieu and Roberts 2009; Beaumont 2003; Lin, Liu and Sloan 2000).

To outline this algorithm and to understand how it preserves the correct target distribution by using an unbiased estimator we will abstract away from the setting above. Denote $\pi(\theta)$ the correct but intractable target distribution. Further, let us denote $\hat{\pi}(\theta; u)$ an non-negative

unbiased estimator of $\pi(\theta)$ where $u \sim m_\theta$ are the random variables used to generate this estimator. The pseudo-marginal algorithm proceeds as follows: At current state (θ_k, u_k), $k \geqslant 1$ with estimator $\hat{\pi}(\theta_k; u_k)$ propose $\theta' \sim q(\theta_k, \cdot)$. Subsequently, sample $u' \sim m_{\theta'}$ and compute $\hat{\pi}(\theta', u')$. With probability

$$\alpha(\theta_k, u_k; \theta', u') = \min\left\{1, \frac{\hat{\pi}(\theta'; u')}{\hat{\pi}(\theta_k; u_k)} \frac{q(\theta', \theta_k)}{q(\theta_k, \theta')}\right\}$$

set $(\theta_{k+1}, u_{k+1}) = (\theta', u')$ otherwise $(\theta_{k+1}, u_{k+1}) = (\theta_k, u_k)$.

To see that this algorithm indeed has the correct invariant distribution it is helpful to consider the algorithm as a Markov chain on the joint space $\Theta \times \mathsf{U}$ instead of a noisy algorithm on Θ. Defining the extended target

$$\pi(\theta, u) = \hat{\pi}(\theta; u) m_\theta(u)$$

we note that

$$\alpha(\theta_k, u_k; \theta', u') = \min\left\{1, \frac{\hat{\pi}(\theta'; u')}{\hat{\pi}(\theta_k; u_k)} \frac{q(\theta', \theta_k)}{q(\theta_k, \theta')}\right\} = \min\left\{1, \frac{\pi(\theta', u')}{\pi(\theta_k, u_k)} \frac{q(\theta', \theta_k) m_{\theta_k}(u_k)}{q(\theta_k, \theta') m_{\theta'}(u')}\right\}.$$

Inspecting the above display it is clear why the pseudo-marginal method works: it corresponds to a Metropolis–Hastings algorithm on the extended space targeting $\pi(\theta, u)$. The joint density by definition has the correct marginal distribution as unbiasedness implies

$$\int \pi(\theta, u) \mathrm{d}u = \int \hat{\pi}(\theta; u) m_\theta(u) \mathrm{d}u = \pi(\theta).$$

To analyse this algorithm it is helpful to change the parametrization of the target distribution. Write

$$Z(\theta) = \log \hat{\pi}(\theta; u) - \log \pi(\theta), \tag{2.7}$$

for the error in the logarithm of the estimator and we will write $g(\mathrm{d}z \mid \theta)$ for its induced distribution. Then the invariant distribution of the pseudo-marginal methods can be written as a density

$$\pi(\theta, z) = \pi(\theta) e^z g(z \mid \theta)$$

on $\Theta \times \mathbb{R}$. The respective Metropolis–Hastings algorithm has transition kernel

$$P_{\mathrm{PMH}}\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q(\theta, \mathrm{d}\theta') g(\mathrm{d}z' \mid \theta') \alpha\left(\theta, z; \theta', z'\right) + \rho(\theta, z) \delta_{(\theta, z)}(\mathrm{d}\theta', \mathrm{d}z'),$$

with

$$\alpha\big(\theta, z; \theta', z'\big) = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{q(\theta', \theta)}{q(\theta, \theta')}e^{z'-z}\right\}.$$

2.12 Remark. *An alternative way to represent the pseudo-marginal algorithm using a one-dimensional auxiliary state is to write*

$$\pi(\theta, w) = \pi(\theta) w Q_\theta(w).$$

with

$$W(\theta) = \frac{\hat{\pi}(\theta; u)}{\pi(\theta)}, \qquad W(\theta) \sim Q_\theta(\cdot).$$

This is done, for example, by Andrieu and Vihola (2015) where the authors consider convergence rates of the pseudo-marginal algorithm. We use the parametrization through $Z(\theta)$ since in many settings $Z(\theta)$ obeys a central limit theorem, see e.g. Section 4.10 and Bérard, Del Moral and Doucet (2014).

We can now make a statement about the ergodicity of the pseudo-marginal algorithm. First, we consider an extension of Theorem 2.3 which is inspired by Andrieu and Roberts (2009). Since the marginal algorithm is defined on a different space than the pseudo-marginal we define an embedded kernel

$$\bar{P}_{\mathrm{MH}}(\theta, z; \mathrm{d}\theta', z') = \alpha(\theta, \theta') q(\theta, \theta') g(\mathrm{d}z' \mid \theta) + \rho(\theta, z)\delta_{\theta,z}(\mathrm{d}\theta', \mathrm{d}z').$$

The corresponding algorithm is the same as the marginal Metropolis–Hastings algorithm which additionally samples $Z' \sim g(\cdot \mid \theta')$ at each iteration.

2.13 Theorem. *Let P_{MH} denote the transition kernel of an aperiodic ϕ-irreducible Markov chain targeting the intractable marginal distribution $\pi(\theta)$. Then the pseudo-marginal kernel P_{PMH} targeting $\pi(\theta, u)$ is ergodic in the sense of* (2.3).

Proof. This is Theorem 1 in Andrieu and Roberts (2009). □

2.14 Proposition. *Assume there exists a constant c such that*

$$\int_0^c g(z \mid \theta)\mathrm{d}z = 1 \quad \pi\text{-almost surely}.$$

Then we have the following relations between the Dirichlet forms of the kernels P_{PMH} *and* $\bar{P}_{\text{MH}}$

$$\mathcal{E}_{P_{\text{PMH}}}(f) \geqslant \frac{\mathcal{E}_{\bar{P}_{\text{MH}}}(f)}{c}$$

for any function $f : \Theta \times \mathbb{R} \to \mathbb{R}$ *with*

$$\pi(f^2) = \int |f(\theta, z)|^2 \pi(\theta, z) \mathrm{d}\theta \mathrm{d}z < \infty$$

and as a consequence $b_{\text{right}}(\bar{P}_{\text{PMH}}) \geqslant c^{-1} b_{\text{right}}(\bar{P}_{\text{MH}})$.

Proof. This is Proposition 10 in Andrieu and Vihola (2015). □

The above proposition implies that the pseudo-marginal chain inherits the (right) spectral gap of the marginal chain if the noise of the estimator is bounded. For positive operators (i.e. $\langle f, Pf \rangle \geqslant 0$ for all f with $\pi(f) < \infty$) Theorem 2.8 implies that the inheritance of the right spectral gap is equivalent to the inheritance of geometric ergodicity. This is true, for example, for the (Gaussian) random walk Metropolis–Hastings algorithm as defined in Section 2.3 (Andrieu and Vihola 2015, Proposition 16).

The boundedness of the weight distribution is crucial for geometric ergodicity of the pseudo-marginal algorithm as the following proposition shows.

2.15 Proposition. *Let* P_{PMH} *have a non-zero spectral gap. Then there exists a function* $c : \Theta \mapsto [1, \infty)$ *with*

$$\int_0^{c(\theta)} g(z \mid \theta) \mathrm{d}z = 1$$

for π*-almost all* θ.

Proof. This is Proposition 13 in Andrieu and Vihola (2015). □

References

Andrieu, Christophe, Arnaud Doucet and Roman Holenstein (2010). 'Particle Markov chain Monte Carlo methods (with Discussion)'. *J. R. Statist. Soc.* B 72.3, pp. 269–342 (cit. on p. 27).

Andrieu, Christophe and Gareth Owen Roberts (2009). 'The pseudo-marginal approach for efficient Monte Carlo computations'. *Ann. Statist.* 37, pp. 697–725 (cit. on pp. 27, 29).

Andrieu, Christophe and Matti Vihola (2015). 'Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms'. *Ann. Appl. Probab.* 25.2, pp. 1030–1077 (cit. on pp. 29–30).

Beaumont, Mark A (2003). 'Estimation of population growth or decline in genetically monitored populations'. *Genetics* 164.3, pp. 1139–1160 (cit. on p. 27).

Bérard, Jean, Pierre Del Moral and Arnaud Doucet (2014). 'A lognormal central limit theorem for particle approximations of normalizing constants'. *Electron. J. Probab.* 19.94, pp. 1–28 (cit. on p. 29).

Bouchard-Côté, Alexandre, Sebastian J Vollmer and Arnaud Doucet (2018). 'The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method'. *Journal of the American Statistical Association* 113.522, pp. 855–867 (cit. on p. 21).

Duane, Simon, Anthony D Kennedy, Brian J Pendleton and Duncan Roweth (1987). 'Hybrid monte carlo'. *Physics letters B* 195.2, pp. 216–222 (cit. on p. 21).

Geman, Stuart and Donald Geman (1984). 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images'. *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741 (cit. on p. 21).

Gustafson, Karl E and Duggirala KM Rao (1997). *Numerical Range - The Field of Values of Linear Operators and Matrices*. Springer, pp. 1–26 (cit. on pp. 23–24).

Hastings, W Keith (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. *Biometrika* 57.1, pp. 97–109 (cit. on p. 20).

Jarner, Søren Fiig and Ernst Hansen (2000). 'Geometric ergodicity of Metropolis algorithms'. *Stochastic processes and their applications* 85.2, pp. 341–361 (cit. on p. 22).

Jones, Galin L and James P Hobert (2001). 'Honest exploration of intractable probability distributions via Markov chain Monte Carlo'. *Statistical Science*, pp. 312–334 (cit. on p. 22).

Kipnis, Claude and SR Srinivasa Varadhan (1986). 'Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions'. *Communications in Mathematical Physics* 104.1, pp. 1–19 (cit. on p. 26).

Lin, L, KF Liu and J Sloan (2000). 'A noisy Monte Carlo algorithm'. *Phys. Rev.* D 61.7, p. 074505 (cit. on p. 27).

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller and Edward Teller (1953). 'Equation of state calculations by fast computing machines'. *The journal of chemical physics* 21.6, pp. 1087–1092 (cit. on p. 20).

Roberts, Gareth O and Jeffrey S Rosenthal (2004). 'General state space Markov chains and MCMC algorithms'. *Probability surveys* 1, pp. 20–71 (cit. on pp. 19–21).

Roberts, Gareth and Jeffrey Rosenthal (1997). 'Geometric ergodicity and hybrid Markov chains'. *Electronic Communications in Probability* 2, pp. 13–25 (cit. on p. 25).

Rudolf, Daniel and Nikolaus Schweizer (2018). 'Perturbation theory for Markov chains via Wasserstein distance'. *Bernoulli* 24.4A, pp. 2610–2639 (cit. on p. 19).

Part II

Large Sample Asymptotics for Tuning of Metropolis–Hastings Algorithms

3

Weak Convergence Analysis of Metropolis–Hastings Algorithms

3.1 Scaling Gaussian Random Walk Metropolis-Hastings Algorithms

THE efficiency of Metropolis–Hastings algorithms heavily relies on the choice of the transition distribution $q(\theta, \cdot)$. The Gaussian random walk proposal introduced in Section 2.3 which will be relevant for this thesis is

$$q(\theta, \theta') = \varphi(\theta'; \theta, l^2 \cdot \Sigma), \tag{3.1}$$

where Σ is the posterior covariance matrix and l is a tuning parameter.

3.1 Remark. *Many other (and often better) choices for the proposal are possible, but they usually require additional information about the target, for example, the gradient at any given point. While gradients are easily computed for many target densities, the intractable models we are interested in do usually not fall into that class (see Section 2.6).*

In this section we want to investigate how to optimally select the parameter l in order to maximize the efficiency of the Markov chain Monte Carlo algorithm. To illustrate the problem, consider the one-dimensional version of the above Markov chain, targeting a univariate Gaussian distribution. As a proposal we also choose a normal distribution $\varphi(\theta'; \theta, l)$ having set $\Sigma = 1$. But how should l be chosen? Recall, we are often interested in ergodic averages

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(\theta_i).$$

For a chain started at stationarity this estimator is unbiased with variance

$$\operatorname{Var}\left\{\hat{\pi}_n(h)\right\} = \frac{1}{n}\operatorname{Var}_\pi\{h\} + \frac{1}{n^2}\sum_{i\neq j}^{n}\operatorname{Cov}\left\{h(\theta_i), h(\theta_j)\right\}. \tag{3.2}$$

Hence, the performance of the estimator heavily depends on how fast the covariance decays as $|i-j| \to \infty$. In case of the simple Gaussian random walk example, we can identify two sources of large covariance or autocorrelation:

i) **Stepsize too small.** If l is chosen very small then the acceptance rate will be close to 1, but the chain will only move very slowly. Hence, we have high autocorrelation in subsequent states.

ii) **Stepsize too large.** If l is chosen very large the chain will often propose new values deep in the tails of the distribution. Such moves are rejected with high probability and the Markov chain then is stuck.

The task is consequently to choose l so as to balance both the above considerations and to find an optimal trade-off. In the general (multi-dimensional) setting of (3.1) the tuning parameter l can not be expected to be the same regardless of the dimension d of $\Theta = \mathbb{R}^d$. Therefore we will consider the tuning parameter $l = \ell/\sqrt{d}$, where d is the dimension of the parameter space Θ.

In the following we will consider how the tuning parameter ℓ should be chosen as to maximize efficiency of the algorithm under various assumptions as well as why the factor $\sqrt{d}$ is precisely the correct scaling as the dimension increases. First, however, we need to introduce a notion of efficiency. As mentioned above, when started at stationarity, estimators using ergodic averages are unbiased. In Section 2.5 we have introduced central limit theorems for such ergodic averages and introduced descriptions of the asymptotic variance as well as conditions for its finiteness. In addition, we outlined how this quantity corresponds to the effective sample size, i.e. the equivalent number of independent samples the Markov chain has drawn. Thus, for any particular test function of interest h, a good measure of the efficiency of a Markov chain Monte Carlo algorithm with transition kernel P is $\mathrm{IAT}(h, P)$.

The most influential heuristic to maximize the efficiency of a random walk Metropolis–Hastings algorithm is derived from the classical diffusion limit (as $d \to \infty$) argument, first introduced by Roberts, Gelman and Gilks (1997), which we introduce in the next section. Afterwards, we will develop an alternative approach, considering the case where the number of data T tends to infinity.

3.2 Classical Diffusion Limit

An important step in the analysis of the performance and optimization of Metropolis–Hastings algorithms was the introduction of the diffusion limits by Roberts, Gelman and Gilks (1997). In this work the authors analyse the random walk Metropolis–Hastings algorithm, interpreted as a continuous time Markov process with exponential waiting times, by reducing the time steps as the parameter dimension, d, increases. As a result the authors obtain a weak convergence result of the continuous time process in the Skorokhod topology (Ethier and Kurtz 2005) showing that as $d \to \infty$, the first component of the process converges to a Langevin diffusion which subsequently can be optimized. In this section we want to make this result more precise and introduce the assumption required in more detail.

***Assumption* 3.1** (Product density)**.** *The target distribution π factorizes into a product of d components*

$$\pi(\theta) = \prod_{i=1}^{d} f(\theta_i). \tag{3.3}$$

This is a very unrealistic assumption in practice and will never be satisfied. Further, if it is satisfied every single of the d coordinates can be simulated separately. However, assumptions similar in spirit to (3.3) are crucial for the subsequent argument. In addition, we need f to be smooth as follows.

***Assumption* 3.2** (Smoothness)**.** *The density f is positive with continuous first and second derivatives. In addition, $\partial_x \log f(x)$ is Lipschitz continuous and*

$$\int \left(\frac{f'(\theta)}{f(\theta)} \right)^4 f(\theta) \mathrm{d}\theta < \infty,$$

and

$$\int \left(\frac{f''(\theta)}{f(\theta)} \right)^8 f(\theta) \mathrm{d}\theta < \infty.$$

***Assumption* 3.3** (Proposal)**.** *The proposal is the Gaussian random walk defined in* (3.1) *with $l = \ell / \sqrt{d}$ and $\Sigma = I$, the d-dimensional identity matrix.*

In order to compare the discrete time Markov chain $(\theta_n)_{n \in \mathbb{N}}$ to the limiting diffusion, we can interpret it as a continuous time *càdlàg* process[1] by defining a Poisson process $N(t)$ with

[1] Càdlàg processes are right continuous with left limits and allow for jumps. In our case the jumps correspond to acceptance events of the Metropolis–Hastings chain.

rate d and taking $\boldsymbol{\theta}_{N(t)}$. Ultimately, we will be interested in the first component of the process $U_t^d = \boldsymbol{\theta}_{N(t),1}^d$, where the superscript d emphasizes the dependence on the dimension.

3.2 Theorem (Diffusion Limit of the Random Walk Metropolis–Hastings Algorithm)**.** *Let Assumptions 3.1, 3.2 and 3.3 hold. Consider the process $U_t^d = \boldsymbol{\theta}_{N(t),1}^d$ started from stationarity, where $N(t)$ is a Poisson process with rate d. Then $U^d \rightsquigarrow U$ (in the Skorokhod topology), where U is described by the Langevin stochastic differential equation (SDE)*

$$\mathrm{d}U_t = \sqrt{h(\ell)}\mathrm{d}B_t + h(\ell)\frac{1}{2}\nabla \log \pi(U_t)\mathrm{d}t$$

with

$$h(\ell) = \ell^2 \cdot 2\Phi\left(-\frac{\ell\sqrt{I}}{2}\right), \qquad I = \int \left\{\nabla_\theta \log f(\theta)\right\}^2 f(\theta)\mathrm{d}\theta.$$

Proof. This is Theorem 1.1 in Roberts, Gelman and Gilks (1997). □

The quantity $h(\ell)$ is called the *speed measure* since we can write $U_t = V_{h(\ell)t}$ where V_t is a diffusion with unit speed

$$\mathrm{d}V_t = \mathrm{d}B_t + \frac{1}{2}\nabla \log \pi(V_t)\mathrm{d}t,$$

where we denote B_t a Brownian motion. Consequently, this result provides a precise optimization criterion as increasing the speed measure improves the mixing of the diffusion. Similarly, arguing according to (3.2), Roberts and Rosenthal (2001) also note that for small $\varepsilon > 0$ maximizing $h(\ell)$ minimizes $\mathrm{Cov}\left\{h(\theta_0), h(\theta_\varepsilon)\right\}$.

Numerically, we find that $h(\ell)$ is optimized at $\ell = \hat{\ell}_\infty \approx 2.38$. As a corollary of the above result we also know that the acceptance rate

$$\mathrm{pr}_{\mathrm{acc},d}(\ell) \to \mathrm{pr}_{\mathrm{acc}}(\ell) = 2\Phi\left(-\ell\sqrt{I}/2\right) \tag{3.4}$$

as $d \to \infty$ and we can calculate $\mathrm{pr}_{\mathrm{acc}}(\hat{\ell}_\infty) \approx 23.38\%$.

3.3 Large Sample Asymptotics of the Random Walk Metropolis–Hastings Algorithm

As already mentioned above, the assumptions required to establish Theorem 3.2 are very unrealistic in practice. Especially condition (3.3) is almost never fulfilled. Here we want to introduce a different asymptotic analysis. Instead of considering the case of the parameter dimension, d, increasing we consider the case where the amount of data, T, increases thereby

leveraging results from Bayesian asymptotic theory. In particular, the Bernstein–von Mises Theorem (Theorem 1.5) provides a precise structure of the posterior distribution when the number of data tends to infinity. In contrast to the conditions required for the diffusion limit (Theorem 3.2), the Bernstein–von Mises can be observed to hold in many practical cases.

Following this motivation, we will from now on explicitly consider Bayesian posterior targets

$$\pi_T(\theta) = p(\theta \mid y_{1:T}) \propto p(y_{1:T} \mid \theta)p(\theta),$$

where $y_{1:T} = (y_1, \dots, y_T)$ is a vector of T data points. A Bernstein–von Mises-type result will be our first assumption.

***Assumption* 3.4** (Posterior concentration)**.** *The posterior distribution converges towards a normal distribution, i.e. as $T \to \infty$*

$$\int \left| \pi_T(\theta) - \varphi(\theta; \hat{\theta}_T, \Sigma/T) \right| \mathrm{d}\theta \to 0, \quad \hat{\theta}_T \to \bar{\theta},$$

both in $\mathbb{P}^Y$-probability and Σ is a positive definite covariance matrix.

We leave the precise structure of $\hat{\theta}_T$ and Σ unspecified here, because it is not necessary for the remainder of this analysis. For conditions under which Assumption 3.4 holds, see the summary in Section 1.5 or Van der Vaart (2000). For simplicity, we will only consider the case $d = 1$, i.e. $\theta \in \mathbb{R}$, but the argument holds similarly for higher dimensions.

The classical version of the Metropolis-Hastings algorithm proceeds with proposing random walk moves according to some density ν. If the posterior concentrates at rate $1/\sqrt{T}$ we need to adjust the scale of the random walk. This is summarized in the next assumption.

***Assumption* 3.5.** *The proposal is a scaled random walk*

$$\theta' = \theta + \frac{\xi}{\sqrt{T}}, \quad \xi \sim \nu(\cdot),$$

where ν is a continuous density.

This yields the transition density

$$q_T(\theta, \theta') = \sqrt{T}\nu\left\{\sqrt{T}\left(\theta' - \theta\right)\right\}.$$

Rescaling to the local parameter $\tilde{\theta} = \sqrt{T}\left(\theta - \hat{\theta}_T\right)$ yields

$$\tilde{q}_T(\tilde{\theta}, \tilde{\theta}') = \nu(\tilde{\theta}' - \tilde{\theta}) = \tilde{q}(\tilde{\theta}, \tilde{\theta}'),$$

independent of the number of data T. Under the rescaled parameter $\tilde{\theta}$ the posterior distribution is $\tilde{\pi}_T\left(\tilde{\theta}\right) = \pi_T\left(\hat{\theta}_T + \theta/\sqrt{T}\right)/\sqrt{T}$ and

$$\int \left|\tilde{\pi}_T(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma)\right| \mathrm{d}\tilde{\theta} \to 0$$

in $\mathbb{P}^Y$-probability. We consider now the Metropolis–Hastings algorithm to construct a Markov chain with stationary distribution $\tilde{\pi}_T(\mathrm{d}\theta)$ and transition kernel

$$\tilde{P}_T(\tilde{\theta}, d\tilde{\theta}') = \min\left\{1, \frac{\tilde{\pi}_T(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{\pi}_T(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')}\right\} \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') + \rho(\tilde{\theta})\delta_\theta(\mathrm{d}\theta'),$$

where again we make the dependence on the data, T, explicit.

3.3 Theorem. *Under Assumptions 3.4 and 3.5 the Markov chain $\chi_T = \left(\tilde{\theta}_{T,n}\right)_{n\in\mathbb{N}}$ with rescaled parameter $\tilde{\theta} = \sqrt{T}\left(\theta - \hat{\theta}_T\right)$ and started from stationarity converges to a Markov chain with stationary distribution $\varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma)$ and transition kernel*

$$\tilde{P}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') = \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}')}{\varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta})}\right\} \tilde{q}(\tilde{\theta}', \tilde{\theta})\mathrm{d}\theta' + \rho(\tilde{\theta})\delta_\theta(\mathrm{d}\theta') \tag{3.5}$$

where $\rho(\tilde{\theta})$ is the corresponding rejection probability.

Let μ_T denote the distribution of χ_T defined on the space on $\bigotimes_{i=1}^{\infty}\mathbb{R}$. In order to prove convergence of μ_T it is sufficient to ensure convergence of the distribution's first k coordinates (that is, the projections on the space $\bigotimes_{i=1}^{k}\mathbb{R}$) for every $k \in \mathbb{N}$, see Section 4.20. This, in turn, can be shown by considering products, i.e.

$$\int \prod_{i=1}^{k} f_i(\theta_i)\mu_T(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_k) \to \int \prod_{i=1}^{k} f_i(\theta_i)\mu(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_k)$$

as $T \to \infty$ for all $k \in \mathbb{N}$, $f_i \in C_b(\mathbb{R})$, see e.g. Ethier and Kurtz (2005, Chapter 3, Proposition 4.6). For the Markov chain in Theorem 3.3 we can show convergence for all k by a simple induction argument. First, we need to show that the stationary initial distribution converges,

$$E\left\{f(\tilde{\boldsymbol{\theta}}_{T,1})\right\} \to E\left\{f(\tilde{\boldsymbol{\theta}}_1)\right\} \tag{3.6}$$

for all $f \in C_b(\mathbb{R})$. In the induction step, we assume that we have established convergence of the first k coordinates, i.e. for a collection of functions $f_i \in C_b(\mathbb{R}), i = 1, \ldots, k$

$$E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k})\right\} \to E\left\{f_1(\tilde{\theta}_1) \cdot \ldots \cdot f_k(\tilde{\theta}_k)\right\}.$$

Without loss of generality take $\|f_i\|_\infty \leqslant 1, i = 1, \ldots, k$

$$\begin{aligned}
&\left|E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) f_{k+1}(\tilde{\theta}_{T,k+1})\right\} - E\left\{f_1(\tilde{\theta}_1) \cdot \ldots \cdot f_k(\tilde{\theta}_k) f_{k+1}(\tilde{\theta}_{k+1})\right\}\right| \\
&= \left|E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P}_T f_{k+1}(\tilde{\theta}_{T,k})\right\} - E\left\{f_1(\tilde{\theta}_1) \cdot \ldots \cdot f_k(\tilde{\theta}_k) \tilde{P} f_{k+1}(\tilde{\theta}_k)\right\}\right| \\
&\leqslant \left|E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P}_T f_{k+1}(\tilde{\theta}_{T,k})\right\} - E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P} f_{k+1}(\tilde{\theta}_{T,k})\right\}\right| \\
&\quad + \left|E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P} f_{k+1}(\tilde{\theta}_{T,k})\right\} - E\left\{f_1(\tilde{\theta}_1) \cdot \ldots \cdot f_k(\tilde{\theta}_k) \tilde{P} f_{k+1}(\tilde{\theta}_k)\right\}\right|.
\end{aligned}$$

For the second part we can use the induction assumption by noting that $\tilde{P} f \in C_b(\mathbb{R})$ whenever $f \in C_b(\mathbb{R})$. For the first term we have

$$\begin{aligned}
&\left|E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P}_T f_{k+1}(\tilde{\theta}_{T,k})\right\} - E\left\{f_1(\tilde{\theta}_{T,1}) \cdot \ldots \cdot f_k(\tilde{\theta}_{T,k}) \tilde{P} f_{k+1}(\tilde{\theta}_{T,k})\right\}\right| \\
&\leqslant E\left\{\left|\tilde{P}_T f_{k+1}(\tilde{\theta}_{T,k}) - \tilde{P} f_{k+1}(\tilde{\theta}_{T,k})\right|\right\}.
\end{aligned}$$

As a consequence, in order to show that the Markov chain χ_T converges in distribution to the limiting Markov chain defined in Theorem 3.3, we need to prove

$$\tilde{\pi}_T(\mathrm{d}\theta) \rightsquigarrow \varphi(\mathrm{d}\theta; 0, \Sigma), \tag{3.7}$$

where we use the symbol $\rightsquigarrow$ to denote weak convergence and

$$\int \left|\tilde{P}_T f(\tilde{\theta}) - \tilde{P} f(\tilde{\theta})\right| \tilde{\pi}_T(\mathrm{d}\tilde{\theta}) \to 0 \tag{3.8}$$

as $T \to \infty$.

Let us note in passing that the above argument is not entirely complete. This is because we have ignored the fact that the posterior distribution is a random measure (with respect to the data distribution) and its convergence to a Gaussian occurs only in $\mathbb{P}^Y$-probability. However, this will not lead to any issues as shown by Schmon et al. (2021) and detailed in Section 4.8. Equipped with these results we can now proceed with the proof of Theorem 3.3.

Proof of Theorem 3.3. We proceed by showing that (3.7) and (3.8) hold in $\mathbb{P}^Y$-probability.

The stationary distributions converge in probability by Assumption 3.4. Write

$$\Pi_T f(\tilde{\theta}) = \int f(\tilde{\theta}') \min\left\{1, \frac{\tilde{\pi}_T(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{\pi}_T(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')}\right\} \tilde{q}(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}'$$

and Π analogously for the limiting kernel. Then, for $\tilde{\boldsymbol{\theta}}_0^T \sim \tilde{\pi}_T(\,\cdot\,)$,

$$\begin{aligned}
&E\left\{\left|\tilde{P}_T f(\tilde{\boldsymbol{\theta}}_0^T) - \tilde{P} f(\tilde{\boldsymbol{\theta}}_0^T)\right|\right\} \\
&= E\left[\left|\Pi_T f(\tilde{\boldsymbol{\theta}}_0^T) + f(\tilde{\boldsymbol{\theta}}_0^T)\left\{1 - \Pi_T 1(\tilde{\boldsymbol{\theta}}_0^T)\right\} - \Pi f(\tilde{\boldsymbol{\theta}}_0^T) - f(\tilde{\boldsymbol{\theta}}_0^T)\left\{1 - \Pi 1(\tilde{\boldsymbol{\theta}}_0^T)\right\}\right|\right] \\
&\leqslant E\left\{\left|\Pi_T f(\tilde{\boldsymbol{\theta}}_0^T) - \Pi f(\tilde{\boldsymbol{\theta}}_0^T)\right|\right\} + E\left\{\left|\Pi_T 1(\tilde{\boldsymbol{\theta}}_0^T) - \Pi 1(\tilde{\boldsymbol{\theta}}_0^T)\right|\right\}.
\end{aligned}$$

Since $1 \in C(\mathbb{R}^d)$ it is thus sufficient to show that for any choice of $f \in C(\mathbb{R}^d)$ we have

$$E\left\{\left|\Pi_T f(\tilde{\theta}) - \Pi f(\tilde{\theta})\right|\right\} \overset{\mathbb{P}^Y}{\to} 0.$$

Writing $\varphi(\tilde{\theta}) = \varphi(\tilde{\theta}; 0, \Sigma)$ we have for $f \in C_b(\mathbb{R})$ with $\|f\|_\infty \leqslant 1$

$$\begin{aligned}
&\int \left|\Pi_T f(\tilde{\theta}) - \Pi f(\tilde{\theta})\right| \pi_T(\mathrm{d}\tilde{\theta}) \\
&= \int \left| \int f(\tilde{\theta}) \min\left\{1, \frac{\tilde{\pi}_T(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})}{\tilde{\pi}_T(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')}\right\} q(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}' \right. \\
&\qquad \left. - \int f(\tilde{\theta}) \min\left\{1, \frac{\varphi(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})}{\varphi(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')}\right\} q(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}' \right| \tilde{\pi}_T(\mathrm{d}\tilde{\theta}) \\
&\leqslant \int \left| \int f(\tilde{\theta}) \min\left\{q(\tilde{\theta}, \tilde{\theta}')\tilde{\pi}_T(\tilde{\theta}), \tilde{\pi}_T(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})\right\} \mathrm{d}\tilde{\theta}' \right. \\
&\qquad \left. - \int f(\tilde{\theta}) \min\left\{\varphi(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})\right\} \mathrm{d}\tilde{\theta}' \right| \mathrm{d}\tilde{\theta} \\
&\quad + \int \left| \int f(\tilde{\theta}) \min\left\{1, \frac{\varphi(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})}{\varphi(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')}\right\} \varphi(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}' \right. \\
&\qquad \left. - \int f(\tilde{\theta}) \min\left\{1, \frac{\varphi(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})}{\varphi(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')}\right\} \tilde{\pi}_T(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}' \right| \mathrm{d}\tilde{\theta} \\
&\leqslant \iint \left| \min\left\{\tilde{\pi}_T(\tilde{\theta})q(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T(\tilde{\theta}')q(\tilde{\theta}', \tilde{\theta})\right\}\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - \min\left\{\varphi(\tilde{\theta})q(\tilde{\theta},\tilde{\theta}'), \varphi(\tilde{\theta}')q(\tilde{\theta}',\tilde{\theta})\right\}\Big| \mathrm{d}\tilde{\theta}\mathrm{d}\tilde{\theta}' \\
&+ \iint \Bigg| f(\tilde{\theta}) \min\left\{1, \frac{\varphi(\tilde{\theta}')q(\tilde{\theta}',\tilde{\theta})}{\varphi(\tilde{\theta})q(\tilde{\theta},\tilde{\theta}')}\right\} \varphi(\tilde{\theta})q(\tilde{\theta},\tilde{\theta}') \\
&\quad - \int f(\tilde{\theta}) \min\left\{1, \frac{\varphi(\tilde{\theta}')q(\tilde{\theta}',\tilde{\theta})}{\varphi(\tilde{\theta})q(\tilde{\theta},\tilde{\theta}')}\right\} \tilde{\pi}_T(\tilde{\theta})q(\tilde{\theta},\tilde{\theta}')\Bigg| \mathrm{d}\tilde{\theta}\mathrm{d}\tilde{\theta}' \\
&\leqslant \iint \left|\tilde{\pi}_T(\tilde{\theta}) - \varphi(\tilde{\theta})\right| q(\tilde{\theta},\tilde{\theta}')\mathrm{d}\tilde{\theta}\mathrm{d}\tilde{\theta}' \\
&\quad + \iint \left|\tilde{\pi}_T(\tilde{\theta}') - \varphi(\tilde{\theta}')\right| q(\tilde{\theta}',\tilde{\theta})\mathrm{d}\tilde{\theta}\mathrm{d}\tilde{\theta}' \\
&\quad + \iint \left|\tilde{\pi}_T(\tilde{\theta}) - \varphi(\tilde{\theta})\right| \mathrm{d}\theta q(\tilde{\theta},\tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&= 3 \iint \left|\tilde{\pi}_T(\tilde{\theta}) - \varphi(\tilde{\theta})\right| \mathrm{d}\tilde{\theta},
\end{aligned}
$$

which vanishes in probability. □

3.4 Convergence of Spectral Gap and Integrated Autocorrelation Time

Large sample asymptotics aim to approximate Markov chain Monte Carlo algorithms for Bayesian posteriors with a limiting kernel that is obtained when the number of data is increased. As such, we would like to make statements about a Markov chain targeting $\tilde{\pi}_T$ by examining the limiting kernel targeting $\tilde{\pi}(\mathrm{d}\tilde{\theta}) = \tilde{\varphi}(\mathrm{d}\tilde{\theta}; 0, \Sigma)$. For example, we might want to optimize the integrated autocorrelation time of the limiting kernel as a proxy for finding the optimal Metropolis–Hastings kernel for finite T. However, our weak convergence result does not provide precise conditions for $\mathrm{IAT}(h, \tilde{P}_T) \to \mathrm{IAT}(h, \tilde{P})$ as $T \to 0$. Similarly, we would like to make statements about the qualitative convergence behaviour of Markov chains by considering the limiting kernel. If the limiting algorithm has a spectral gap and $\|\tilde{P}_T|_{L^2_0(\tilde{\pi}_T)}\|_{\mathrm{op}} \to \|\tilde{P}|_{L^2_0(\tilde{\pi})}\|_{\mathrm{op}}$ in $\mathbb{P}^Y$-probability we know that for T large enough $\tilde{P}_T$ also must have a spectral gap. We can use this argument to provide sufficient conditions for the integrated autocorrelation time to converge.

3.4 Proposition. *Consider the setting of Theorem 3.3. Let* $\|\tilde{P}_T|_{L^2_0(\tilde{\pi}_T)}\|_{\mathrm{op}} \to \|\tilde{P}|_{L^2_0(\tilde{\pi})}\|_{\mathrm{op}}$ *in* $\mathbb{P}^Y$*-probability. Then for* $h \in L^2(\tilde{\pi})$

$$
\mathrm{cov}\left\{h(X_{0,T}), h(X_{k,T})\right\} \xrightarrow{\mathbb{P}^Y} \mathrm{cov}\left\{h(X_0), h(X_k)\right\}
$$

for all $k \in \mathbb{N}$ *and*

$$\mathrm{IAT}(h, \tilde{P}_T) \to \mathrm{IAT}(h, \tilde{P})$$

in $\mathbb{P}^Y$*-probability.*

Proof. Since the spectral radius converges (because the operator norm does) there exists and interval $[-b, b] \subset [-1, 1)$ with $0 < b < 1$ such that the support of the spectral measure $e(h, \tilde{P}_T)$ is contained in $[-b, b]$ for T large enough. Together with Theorem 3.3 we can conclude that for all k and for T large

$$\int_{-b}^{b} \lambda^k e(h, \tilde{P}_T)(\mathrm{d}\lambda) = \mathrm{cov}\left\{h(X_{0,T}), h(X_{k,T})\right\} \overset{\mathbb{P}^Y}{\to} \mathrm{cov}\left\{h(X_0), h(X_k)\right\} = \int_{-b}^{b} \lambda^k e(h, \tilde{P})(\mathrm{d}\lambda). \tag{3.9}$$

Since polynomials are dense in $C[-b, b]$ by Stone-Weierstrass we have $e(h, \tilde{P}_T) \to e(h, \tilde{P})$ weakly. The function $f(\lambda) = (1 + \lambda)/(1 - \lambda)$ is continuous on $[-b, b]$ and we have

$$\mathrm{IAT}(h, \tilde{P}_T) = \int_{-b}^{b} \frac{1+\lambda}{1-\lambda} e(h, \tilde{P}_T)(\mathrm{d}\lambda) \to \int_{-b}^{b} \frac{1+\lambda}{1-\lambda} e(h, \tilde{P})(\mathrm{d}\lambda) = \mathrm{IAT}(h, \tilde{P}). \qquad \square$$

The assumptions of Proposition 3.4 are quite strong. To use this result in order to show convergence of the integrated autocorrelation time we require convergence of the transition operators in the norm topology induced by the operator norm

$$\| P \mid_{L_0^2(\pi)} \|_{\mathrm{op}} = \sup_{f \in L_0^2(\pi)} \{\langle Pf, f \rangle : \|f\|_\pi = 1\}$$

and therefore uniform convergence for a large class of functions.

3.5 Comparison of Tuning Guidelines

In this section, we want to compare the guidelines that result from the diffusion limit with the asymptotic kernel of the big data limit. We already established that under the conditions of Theorem 3.2 the (asymptotically) optimal scaling factor for the random walk Metropolis–Hastings algorithm is given by $\ell = \hat{\ell}_\infty = 2.38$ with associated optimal acceptance rate $\mathrm{pr}_{\mathrm{acc}}(\hat{\ell}_\infty) = 23.38\%$. These results are derived from optimizing the limiting diffusion for any test function.

We contrast this result with optimizing the limiting kernel (3.5) arising in the large data setting (see Theorem 3.3). Since the parameter dimension, d, is fixed in Theorem 3.3 the

resulting optimal value will vary as d changes. In the limiting kernel we set

$$q(\theta, \theta') = \varphi\left(\theta'; \theta, \frac{\ell}{\sqrt{d}} I\right)$$

and the target will be a centred Gaussian with covariance I. We run the resulting algorithm for every value on a grid $\mathcal{L} = \{2, 2.1, \ldots, 3\}$ for selected values of d. On this grid we find ℓ such that the integrated autocorrelation time of the identity function is minimized, where the integrated autocorrelation time is estimated using the overlapping batch means estimator. For $d \geqslant 2$ we estimate $\mathrm{IAT}(\ell)$ as the average integrated autocorrelation time over all one-dimensional projections. Table 3.1 shows the result of the grid search for the optimal parameter.

Dimension d	$\hat{\ell}_{\mathrm{opt}}$	$\mathrm{IAT}(\hat{\ell}_{\mathrm{opt}})$	$\mathrm{pr}_{\mathrm{acc}}(\hat{\ell}_{\mathrm{opt}})$
$d = 1$	2.46 (0.25)	4.23	43.42%
$d = 2$	2.44 (0.11)	7.53	34.70%
$d = 3$	2.53 (0.16)	10.40	29.65%
$d = 5$	2.52 (0.15)	15.99	26.31%
$d = 10$	2.42 (0.15)	31.61	25.43%
$d = 15$	2.45 (0.07)	46.15	24.00%
$d = 20$	2.39 (0.10)	60.44	24.55%
$d = 30$	2.40 (0.08)	88.41	23.97%
$d = 50$	2.38 (0.10)	140.70	23.89%

Table 3.1: Optimal values for scaling ℓ and associated value of computing time and average acceptance probability: mean and standard deviation of the minimizers over 10 runs.

It can be seen that the optimal value, ℓ, minimizing the integrated autocorrelation time is fairly constant and approaches $\hat{\ell}_{\infty} = 2.38$ as $d \to \infty$. However, while we see that the respective acceptance probability approaches the asymptotically optimal 23.38% it can also be observed that in low dimensions a deviation from this guideline can be beneficial, even if all assumptions of Theorem 3.2 are fulfilled (which is clearly the case for Gaussian distributions with identity covariance matrix).

3.6 Tuning Pseudo-Marginal Methods

In this section we will introduce one of the main topics of this thesis: the tuning of the pseudo-marginal algorithm. In contrast to the simple random walk Metropolis–Hastings algorithm analysed in the previous section, the pseudo-marginal method introduced in Section 2.6 has the additional complication that the likelihood evaluation in each step is noisy, i.e. it has a

random component. Recall the pseudo-marginal transition kernel

$$P_{\text{PMH}}\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q(\theta, \mathrm{d}\theta')g(\mathrm{d}z' \mid \theta')\alpha\left(\theta, z; \theta', z'\right) + \rho(\theta, z)\delta_{(\theta,z)}(\mathrm{d}\theta', \mathrm{d}z'),$$

targeting the distribution $\pi(\theta)e^z g(z \mid \theta)$ on the extended space $\Theta \times \mathbb{R}$ and

$$\alpha\left(\theta, z; \theta', z'\right) = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{q(\theta', \theta)}{q(\theta, \theta')}e^{z'-z}\right\}.$$

We can rewrite $\hat{\pi}(\theta; u) = \pi(\theta)e^z$ (see Section 2.6), that is, we can denote the estimator as the true density with a multiplicative noise component. This enables a comparison of the pseudo-marginal chain with the idealized marginal Metropolis–Hastings algorithm. Intuition yields that reducing the variance of the estimator produces an algorithm that performs similar to the standard Metropolis–Hastings algorithm. However, lower variance often comes at higher computational cost.

Analysing this trade-off is not straightforward for general target distributions and general classes of estimators. Hence, previous works introduce further assumptions—either on the target or on the noise—that simplify the analysis. The first contribution in this line of work is that of Pitt et al. (2012), where the authors assume that $q(\theta, \theta') = \pi(\theta')$. Even though this is a very unrealistic assumption, it simplifies the analysis considerably. In particular, the acceptance probability becomes

$$\alpha\left(\theta, z; \theta', z'\right) = \min\left\{1, e^{z'-z}\right\}.$$

In order to make this quantity tractable we need to introduce assumptions on the distribution of z either under the proposal or under stationarity (that is, the distribution of z conditional on being accepted). Pitt et al. (2012) introduce a Gaussian assumption which can be motivated by a central limit theorem. In many cases asymptotically $g(z \mid \theta) = \varphi\left\{z; -\sigma^2(\theta)/2; \sigma^2(\theta)\right\}$ and $\pi(z \mid \theta) = e^z g(z \mid \theta) = \varphi\left\{z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\}$ see e.g. Example 3.5.

In order to balance the computational cost with the algorithmic efficiency Pitt et al. (2012) introduce the *computing time*

$$\text{CT}(h, P) = \frac{\text{IAT}(h, P)}{\sigma^2}$$

for some target variance σ^2 which is motivated by the fact that in the above setting the number of particles $N \propto 1/\sigma^2$. In the simple setting above $\text{IAT}(h, P) = \text{IAT}(\sigma)$ independent of the function h. Subsequently, it can be shown that $\text{CT}(\sigma) = \text{CT}(h, P)$ can be minimized at $\sigma = 0.92$ (Pitt et al. 2012, Lemma 5), independent of h.

3.5 Example. *To understand how a central limit theorem for the noise could be derived consider an i.i.d. latent variable model with (single) observation densities*

$$Y_t \mid (X_t = x) \sim g(\cdot \mid x, \theta) \quad X_t \sim f(\cdot \mid \theta), \quad t = 1, \dots, T$$

where T is the number observations. An unbiased estimator for the intractable likelihood can be constructed by

$$\hat{p}(y_{1:T} \mid \theta) = \prod_{t=1}^{T} \hat{p}(y_t \mid U_{t,i}, \theta), \quad \hat{p}(y_t \mid U_{t,i}, \theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{g(y_t \mid U_{t,i}, \theta) f(U_{t,i} \mid \theta)}{h(U_{t,i} \mid y, \theta)}.$$

Then, under regularity conditions, we can use an expansion of $\log(1 + x)$ *about* 1 *to compute*

$$\begin{aligned} Z_T(\theta) &= \sum_{t=1}^{T} \log \frac{p(Y_t \mid U_{t,1:N}, \theta)}{p(y_t \mid \theta)} \\ &= \sum_{t=1}^{T} \log \left\{ 1 + \frac{p(Y_t \mid U_{t,1:N}, \theta) - p(Y_t \mid \theta)}{p(Y_t \mid \theta)} \right\} \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \epsilon_T(Y_t; \theta) - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_T^2(Y_t; \theta) + o_{\mathbb{P}^Y}(1), \end{aligned}$$

where

$$\epsilon_T(Y_t; \theta) = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} \left\{ \overline{w}(Y_t, U_{i,t}, \theta) - 1 \right\}, \quad \overline{w}(Y_t, U_{t,i}, \theta) = \frac{g(y_t \mid U_{t,i}, \theta) f(U_{t,i} \mid \theta)}{h(U_{t,i} \mid y, \theta)} \Big/ p(y_t \mid \theta).$$

Under regularity conditions we expect

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \epsilon_T(Y_t; \theta) \rightsquigarrow \mathcal{N}\left\{0, \sigma^2(\theta)\right\} \quad \text{and} \quad -\frac{1}{2T} \sum_{t=1}^{T} \epsilon_T^2(Y_t; \theta) \to -\frac{\sigma^2(\theta)}{2}.$$

where $E\left\{\epsilon_T^2(Y_t; \theta)\right\} = \sigma^2(\theta)$. *Thus*

$$Z_T(\theta) \rightsquigarrow \mathcal{N}\left\{-\frac{\sigma^2(\theta)}{2}, \sigma^2(\theta)\right\}.$$

This argument is not very precise and we will provide a rigorous proof in Section 4.10.

The above arguments cover the setting where $q(\theta, \theta') = \pi(\theta')$ which is not interesting in practice. Indeed, we could just sample from π directly, so no Markov chain approximation is

required. In two follow-up works, both Doucet et al. (2015) and Sherlock et al. (2015) aim to provide guidelines for more general settings.

Doucet et al. (2015) introduce an auxiliary kernel

$$P^*(\theta, z; \mathrm{d}\theta', \mathrm{d}z') = \alpha_\theta(\theta, \theta')\alpha_z(z, z')q(\theta, \theta')g(z) + \left\{1 - \varrho_\theta(\theta)\varrho_z(z)\right\} \delta_{\theta,z}(\mathrm{d}\theta', \mathrm{d}z'),$$

where

$$\rho_\theta(\theta) = \int a_\theta(\theta, \theta')q(\theta, \theta')\mathrm{d}\theta', \qquad \rho_z(z) = \int a_z(z, z')g(z)\mathrm{d}z,$$
$$a_\theta(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta')}\frac{q(\theta, \theta')}{q(\theta', \theta)}\right\}, \qquad a_z(z, z') = \min\left\{1, e^{z'-z}\right\}.$$

Here the noise distribution $g(z) = g(z \mid \theta)$ is assumed independent of the current parameter. Using the inequality $\min(1, a)\min(1, b) \leqslant \min(1, ab)$ for non-negative a, b we know that the kernels P_{PMH} and P^* are Peskun-Tierney ordered (Peskun 1973; Tierney 1998) and hence $\mathrm{Var}(h, P_{\mathrm{PMH}}) \leqslant \mathrm{Var}(h, P^*)$ for every $h \in L_0^2(\Theta \times \mathbb{R}, \pi)$. Based on this the authors propose to minimize bounds on the computing time relative to the exact Metropolis–Hastings algorithm using the exact target density. We already know from Pitt et al. (2012) under very efficient proposals (e.g. $q(\theta, \theta') = \pi(\theta')$) the optimal noise is given by $\sigma = 0.92$. On the other hand, Doucet et al. (2015) show that for very inefficient proposals the relative inefficiency is optimized at $\sigma = 1.68$. Since the efficiency of the exact marginal algorithm is unknown, the authors propose to choose $\sigma = 1.2$ as a robust default choice.

In a different approach Sherlock et al. (2015) pursue a diffusion limit in the line of Theorem 3.2. As in the classical paper by Roberts, Gelman and Gilks (1997) the authors require that the target factorizes as a product of *i.i.d* densities, see Assumption 3.1. In addition—and like all other contributions—Sherlock et al. (2015) require the noise distribution to be independent of the current state of the chain while also employing a Gaussian assumption on this distribution. As $d \to \infty$ the first component of the pseudo-marginal random walk Metropolis–Hastings algorithm converges to a Langevin diffusion

$$\mathrm{d}U_t = \sqrt{h(\ell)}\mathrm{d}B_t + h(\ell)\frac{1}{2}\nabla \log \pi(U_t)\mathrm{d}t$$

with

$$h(\ell) = \ell^2 \cdot 2\Phi\left(-\frac{\sqrt{2\sigma^2 + \ell^2}}{2}\right) \Big/ I, \qquad I = \int \left\{\nabla_\theta \log f(\theta)\right\}^2 f(\theta)\mathrm{d}\theta.$$

As before we can take into account the inefficiency of having to estimate π by considering

to maximize $\sigma^2 h(\ell)$. The optimal values are $\hat{\ell}_\infty = 2.562, \hat{\sigma}_\infty = 1.81$ and $\text{pr}_{\text{acc}}(\hat{\ell}_\infty, \hat{\sigma}_\infty) = 7.001\%$.

All previous contributions aiming to optimize the efficiency of the pseudo-marginal algorithm rely on the unrealistic assumption that the noise is independent of the current state of the chain. In Pitt et al. (2012) and Doucet et al. (2015) this is motivated by the fact that the posterior concentrates around some fixed parameter, $\bar{\theta}$, as the number of data T increases due to the Bernstein–von Mises Theorem (e.g. Theorem 1.5). Hence, the asymptotic noise distribution $g(z \mid \theta) = \varphi\left\{z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} \approx \varphi\left\{z; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta})\right\}$ for $\theta \sim \pi$. At the same time the Bernstein–von Mises Theorem is a much more realistic assumption on the target than the classical factorizing $i.i.d.$ distribution required for the diffusion limit.

Based on these arguments the question arises how the pseudo–marginal Metropolis–Hastings algorithm behaves when the posterior concentrates according to the Bernstein–von Mises Theorem and at the same time the noise distribution $g(z \mid \theta)$ becomes $\varphi\left\{z; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\}$. This is the subject of the article presented in following chapter, which is accepted for publication in *Biometrika* (Schmon et al. 2021).

References

Doucet, Arnaud, Michael K Pitt, George Deligiannidis and Robert Kohn (2015). 'Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator'. *Biometrika* 102.2, pp. 295–313 (cit. on pp. 48–49).

Ethier, Stewart N and Thomas G Kurtz (2005). *Markov Processes: Characterization and Convergence*. Vol. 282. John Wiley & Sons (cit. on pp. 37, 40).

Peskun, Peter H (1973). 'Optimum monte-carlo sampling using markov chains'. *Biometrika* 60.3, pp. 607–612 (cit. on p. 48).

Pitt, Michael K, Ralph dos Santos Silva, Paolo Giordani and Robert Kohn (2012). 'On some properties of Markov chain Monte Carlo simulation methods based on the particle filter'. *J. Econometrics* 171.2, pp. 134–151 (cit. on pp. 46, 48–49).

Roberts, Gareth O and Jeffrey S Rosenthal (2001). 'Optimal scaling for various Metropolis-Hastings algorithms'. *Statistical science* 16.4, pp. 351–367 (cit. on p. 38).

Roberts, G.O., A. Gelman and W.R. Gilks (1997). 'Weak convergence and optimal scaling of random walk Metropolis algorithms.' *Ann. Appl. Probab.* 7, pp. 110–120 (cit. on pp. 36–38, 48).

Schmon, Sebastian M, George Deligiannidis, Arnaud Doucet and Michael K Pitt (2021). 'Large Sample Asymptotics of the Pseudo-Marginal Algorithm'. *Biometrika* 108.1, pp. 37–51 (cit. on pp. 41, 49).

Sherlock, Chris, Alexandre H Thiery, Gareth O Roberts, Jeffrey S Rosenthal et al. (2015). 'On the efficiency of pseudo-marginal random walk Metropolis algorithms'. *Ann. Statist.* 43.1, pp. 238–275 (cit. on p. 48).

Tierney, Luke (1998). 'A note on Metropolis-Hastings kernels for general state spaces'. *Ann. Appl. Probab.* pp. 1–9 (cit. on p. 48).

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Cambridge University Press (cit. on p. 39).

4

Large Sample Asymptotics of the Pseudo-Marginal Method

Accepted for publication in *Biometrika*, 2020

(with George Deligiannidis, Arnaud Doucet and Michael Pitt)

4.1 Introduction

THE pseudo-marginal algorithm is a variant of the popular Metropolis–Hastings algorithm where an unnormalized version of the target density is replaced by a non-negative unbiased estimate. The algorithm first appeared in the physics literature (Lin, Liu and Sloan 2000) and has become popular in Bayesian statistics as many intractable likelihood functions can be estimated unbiasedly using importance sampling or particle filters (Andrieu, Doucet and Holenstein 2010; Andrieu and Roberts 2009; Beaumont 2003).

Replacing the true likelihood in the Metropolis-Hastings algorithm with an estimate results in a trade-off: the asymptotic variance of an ergodic average of a pseudo-marginal chain typically decreases as the number of Monte Carlo samples, N, used to obtain the likelihood estimator increases, as established by Andrieu and Vihola (2016) for importance sampling estimators; however, this comes at the cost of a higher computational burden. An important task in practice is thus to choose N such that the computational resources required to obtain a given asymptotic variance are minimized. This problem has already been investigated by Pitt et al. (2012), Doucet et al. (2015) and Sherlock, Thiery, Roberts et al. (2015) where guidelines have been obtained under various assumptions either on the proposal (Doucet et al. 2015; Pitt et al. 2012) or on the proposal and target distribution (Sherlock, Thiery, Roberts et al. 2015).

Additionally, all these contributions make the assumption that the noise in the log-likelihood estimator, that is the difference between this estimator and the true log-likelihood, is Gaussian

with variance inversely proportional to N, its mean and variance being independent of the parameter value at which it is evaluated. A similar assumption has also been used by Nemeth, Sherlock and Fearnhead (2016) for the analysis of a related algorithm. This assumption can cast doubts over the practical relevance of the guidelines provided in these contributions. The normal noise assumption was motivated by Pitt et al. (2012), Doucet et al. (2015) and Sherlock, Thiery, Roberts et al. (2015) by the fact that the error in the log-likelihood estimator for state-space models computed using a particle filter is asymptotically normal of variance proportional to γ as $T \to \infty$ with $N = T/\gamma$ (Bérard, Del Moral and Doucet 2014). In addition, the constant variance assumption over the parameter space was motivated in Pitt et al. (2012) and Doucet et al. (2015) by the fact that the posterior typically concentrates as T increases. However, no formal argument justifying why the pseudo-marginal chain would behave as a Markov chain for which these assumptions hold has been provided.

Here, we carry out a novel weak convergence analysis of the pseudo-marginal algorithm in a Bayesian setting which not only justifies these assumptions but also allows us to obtain more precise guidelines on how to optimally tune this algorithm as a function of the parameter dimension d. Weak convergence techniques have become very popular in the Markov chain Monte Carlo literature since their introduction in the seminal paper of Roberts, Gelman and Gilks (1997). With the recent exception of Deligiannidis, Doucet and Pitt (2018), all these analyses have been performed in the asymptotic regime where the parameter dimension $d \to \infty$. Results of this type typically require making strong structural assumptions on the target distribution such as having d independent and identically distributed components as in Sherlock, Thiery, Roberts et al. (2015). We analyse here the pseudo-marginal scheme in the large-sample asymptotic regime where the number of data points T goes to infinity while d is fixed. Under weak regularity conditions, we show that a space-rescaled version of the pseudo-marginal chain converges to a pseudo-marginal chain targeting a normal distribution for which the noise in the log-likelihood estimator is indeed also normal of constant mean and variance. We provide numerical results to optimally scale normal random walk proposals and the noise variance to optimize the performance of this limiting Markov chain as a function of d. These guidelines complement and validate the results obtained in Doucet et al. (2015) and Sherlock, Thiery, Roberts et al. (2015). All proofs can be found in the supplementary material.

4.2 The Pseudo-Marginal Algorithm

4.2.1 Background

Consider a Bayesian model on the Borel space $\{\Theta, \mathcal{B}(\Theta)\}$ where $\Theta \subseteq \mathbb{R}^d$. The parameter $\theta \in \Theta$ follows a prior distribution $p(\mathrm{d}\theta)$ while $\theta \mapsto p(y \mid \theta)$ denotes the likelihood function, where $y = (y_1, \ldots, y_T)$ denotes the vector of observations. When the likelihood arises from a complex latent variable model an analytic expression of $p(y \mid \theta)$ might not be available. Hence, the standard Metropolis–Hastings algorithm cannot be used to sample the posterior distribution $p(\mathrm{d}\theta \mid y) \propto p(\mathrm{d}\theta)\, p(y \mid \theta)$ as the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ appearing in the Metropolis–Hastings acceptance probability, when at parameter θ and proposing θ', cannot be computed. Assume we have access to an unbiased positive estimator $\hat{p}(y \mid \theta, U)$ of the intractable likelihood $p(y \mid \theta)$, where $U \sim m_\theta$ represents the auxiliary variables on $\{\mathcal{U}, \mathcal{B}(\mathcal{U})\}$ used to compute this estimator. We introduce the following probability measure on $\{\Theta \times \mathcal{U}, \mathcal{B}(\Theta) \times \mathcal{B}(\mathcal{U})\}$

$$\pi(\mathrm{d}\theta, \mathrm{d}u) = p(\mathrm{d}\theta \mid y) \frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m_\theta(\mathrm{d}u),$$

which satisfies $\pi(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y)$. The pseudo-marginal algorithm is a Metropolis–Hastings scheme targeting $\pi(\mathrm{d}\theta, \mathrm{d}u)$, hence marginally $p(\mathrm{d}\theta \mid y)$, by using the proposal distribution $Q\left(\theta, u; \mathrm{d}\theta', \mathrm{d}u'\right) = q(\theta, \mathrm{d}\theta') m_{\theta'}\left(\mathrm{d}u'\right)$. This yields the acceptance probability

$$\alpha(\theta, u; \theta', u') = \min\left\{1, r(\theta, \theta') \frac{\hat{p}(y \mid \theta', u')/p(y \mid \theta')}{\hat{p}(y \mid \theta, u)/p(y \mid \theta)}\right\}, \text{ where } r(\theta, \theta') = \frac{\pi(\mathrm{d}\theta')}{\pi(\mathrm{d}\theta)} \frac{q(\theta', \mathrm{d}\theta)}{q(\theta, \mathrm{d}\theta')}.$$

As in previous contributions (Andrieu and Roberts 2009; Andrieu and Vihola 2015; Doucet et al. 2015; Pitt et al. 2012; Sherlock, Thiery, Roberts et al. 2015), we analyse the pseudo-marginal algorithm using additive noise in the log-likelihood estimator, writing $Z(\theta) = \log \hat{p}(y \mid \theta, U) - \log p(y \mid \theta)$. This parameterization allows us to write the target distribution as a measure on $\{\Theta \times \mathbb{R}, \mathcal{B}(\Theta) \times \mathcal{B}(\mathbb{R})\}$ with

$$\pi(\mathrm{d}\theta, \mathrm{d}z) = p(\mathrm{d}\theta \mid y) \exp(z)\, g(\mathrm{d}z \mid \theta),$$

where $Z(\theta) \sim g(\cdot \mid \theta)$ when $U \sim m_\theta$ and the pseudo-marginal kernel is

$$P\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q(\theta, \mathrm{d}\theta') g(\mathrm{d}z' \mid \theta') \alpha\left(\theta, z; \theta', z'\right) + \rho(\theta, z) \delta_{(\theta, z)}(\mathrm{d}\theta', \mathrm{d}z'),$$

with acceptance probability

$$\alpha\left(\theta, z; \theta', z'\right) = \min\left\{1, r(\theta, \theta')\exp\left(z' - z\right)\right\},$$

and corresponding rejection probability $\rho(\theta, z)$.

4.2.2 *Literature review*

We review here recent research motivating this work. To this end, we need to introduce a few additional notations. Let μ be a probability measure on $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$ and $\Pi : \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ a Markov transition kernel. For any measurable function f and measurable set A, we write $\mu(f) = \int f(x)\mu(\mathrm{d}x)$, $\mu(A) = \mu\left\{\mathbb{I}_A(\cdot)\right\}$ and $\Pi f(x) = \int \Pi(x, \mathrm{d}y)\, f(y)$. We consider the Hilbert space $L^2(\mu)$ with inner product $\langle f, g\rangle_\mu = \int f(x)g(x)\mu(\mathrm{d}x)$. For a function $f \in L^2(\mu)$, the asymptotic variance of averages of a stationary Markov chain $\left(X_k\right)_{k\geqslant 1}$ of μ-invariant transition kernel Π is defined as

$$\mathrm{var}(f, \Pi) = \lim_{M\to\infty} \frac{1}{M} E\left\{\sum_{k=1}^{M} f(X_k) - \mu(f)\right\}^2,$$

and $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f)\,\mathrm{IAT}(f, \Pi)$ when the integrated autocorrelation time given by

$$\mathrm{IAT}(f, \Pi) = 1 + 2\sum_{k=1}^{\infty} \frac{\mathrm{cov}\left\{f(X_0), f(X_k)\right\}}{\mathrm{var}\left\{f(X_0)\right\}}$$

is finite. We denote by $\varphi(x; m, \Lambda)$ the normal density of argument x, mean m and covariance Λ.

In order to obtain guidelines to balance computational cost and accuracy of the likelihood estimator Pitt et al. (2012), Doucet et al. (2015) and Sherlock, Thiery, Roberts et al. (2015) make the simplifying assumption that $g(\mathrm{d}z \mid \theta) = \varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)$, that $\sigma^2 \propto 1/N$, and focus on functions $f \in L^2(\pi)$ such that $f(\theta, z) = f\left(\theta, z'\right)$ for any z, z'. Under these assumptions, it was first proposed by Pitt et al. (2012) to minimize

$$\mathrm{CT}(f, P_\sigma) = \frac{\mathrm{IAT}(f, P_\sigma)}{\sigma^2}, \tag{4.1}$$

with respect to σ where

$$P_\sigma\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q(\theta, \mathrm{d}\theta')\varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)\alpha\left(\theta, z; \theta', z'\right) + \rho_\sigma(\theta, z)\delta_{(\theta,z)}(\mathrm{d}\theta', \mathrm{d}z'), \tag{4.2}$$

$\rho_\sigma(\theta, z)$ being the corresponding rejection probability. The criterion (4.1) arises from the fact

that the computational time required to evaluate the likelihood is typically proportional to N. Under the additional assumption that $q(\theta, \mathrm{d}\theta') = \pi(\mathrm{d}\theta')$, the minimizer of $\mathrm{CT}(f, P_\sigma)$ is $\sigma = 0{\cdot}92$ (Pitt et al. 2012). For general proposal distributions Doucet et al. (2015) minimize upper bounds on $\mathrm{CT}(f, P_\sigma)$. This results in guidelines stating that one should indeed select σ around $1{\cdot}0$ when the Metropolis–Hastings algorithm using the exact likelihood would provide an estimator having a small integrated autocorrelation time and around $1{\cdot}7$ when this autocorrelation time is very large (Doucet et al. 2015). In practical scenarios, the integrated autocorrelation time of the Metropolis–Hastings algorithm using the exact likelihood is unknown and the results in Doucet et al. (2015) suggest to select σ around $1{\cdot}2$ as a robust default choice. A slightly different approach is taken by Sherlock, Thiery, Roberts et al. (2015). In addition to similar noise assumptions, it is assumed that the posterior factorizes into d independent and identically distributed components and that one uses an isotropic normal random walk proposal of jump size proportional to ℓ. In this context, one maximizes with respect to (σ, ℓ) the expected squared jump distance associated to the pseudo-marginal sequence of the first parameter component $(\vartheta_{1,k})_{k\geqslant 0}$ divided by the noise variance as $d \to \infty$. In this asymptotic regime, a time-rescaled version of $(\vartheta_{1,k})_{k\geqslant 0}$ converges weakly to a diffusion process and the adequately rescaled expected squared jumping distance converges to the squared diffusion coefficient of this process. Maximizing this squared jump distance is asymptotically equivalent to minimizing $\mathrm{CT}(f, P_\sigma)$ irrespective of f (see Roberts and Rosenthal 2014) and its maximizing arguments are $\sigma = 1{\cdot}8$ and $\ell = 2{\cdot}56$ (Sherlock, Thiery, Roberts et al. 2015, Corollary 1). In practice, the standard deviation of the log-likelihood estimator varies over the parameter space and one selects N such that this standard deviation is approximately equal to the desired σ for a parameter value around the mode of the posterior obtained through a preliminary run.

The strong assumptions made in those contributions can bring into question the merits of the guidelines provided within these papers. Our novel weak convergence analysis of the pseudo-marginal algorithm justifies this assumption in the large sample regime, as $T \to \infty$. This convergence occurs under fairly weak regularity assumptions on the posterior distribution. The resulting limiting algorithms can be optimized to provide guidelines for random walk proposals without relying on any upper bound as in Doucet et al. (2015).

4.3 Large Sample Asymptotics of the Pseudo-Marginal Algorithm

4.3.1 Notation and assumptions

Our analysis of the pseudo-marginal algorithm relies on the assumption that the posterior concentrates (Assumption 4.1) which is most commonly formulated using convergence in probability with respect to the data distribution, denoted $\mathbb{P}^Y$. For our result to hold under this weak assumption we take into account the randomness induced by the data, resulting in a random Markov chain and requiring us to deal with weak convergence of random probability measures. To make this more precise we introduce the following notation.

The observations $(Y_t)_{t\geqslant 1}$ are regarded as random variables defined on a probability space $\left\{\mathsf{Y}^{\mathbb{N}}, \mathcal{B}(\mathsf{Y})^{\mathbb{N}}, \mathbb{P}^Y\right\}$, where $\mathcal{B}(\mathsf{Y})^{\mathbb{N}}$ denotes the Borel σ-algebra and we write $\Omega = \mathsf{Y}^{\mathbb{N}}$ for brevity. For $T \geqslant 1$ we can define the random variables $Y_{1:T} = \left(Y_1, \ldots, Y_T\right)$ as the coordinate projections to Y^T. Then, for $\omega = \left(y_t\right)_{t\geqslant 1} \in \Omega$, $\pi_T^\omega(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y_{1:T})$ denotes a regular version of the target posterior distribution and, for any $\theta \in \Theta$, $g_T^\omega(\mathrm{d}z \mid \theta)$ the conditional distribution of the error in the log-likelihood estimator given observations $y_{1:T}$. The measures π_T^ω and g_T^ω can be interpreted as random measures. Relevant results for random measures are briefly discussed in Section 4.4 and in more detail in the supplementary material. In the following we will use a superscript ω to highlight that a certain quantity depends on the data. All probability densities considered hereafter are with respect to the Lebesgue measure and we use the same symbols for distributions and densities, for example $\mu(\mathrm{d}\theta) = \mu(\theta)\,\mathrm{d}\theta$.

In this context, the target distribution of the pseudo-marginal algorithm is

$$\pi_T^\omega(\mathrm{d}\theta, \mathrm{d}z) = \pi_T^\omega(\mathrm{d}\theta)\exp(z)\, g_T^\omega(\mathrm{d}z \mid \theta),$$

and its transition kernel is

$$P_T^\omega\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q_T(\theta, \mathrm{d}\theta') g_T^\omega(\mathrm{d}z' \mid \theta') \alpha_T^\omega\left(\theta, z; \theta', z'\right) + \rho_T^\omega(\theta, z)\delta_{(\theta,z)}\left(\mathrm{d}\theta', \mathrm{d}z'\right),$$

where

$$\alpha_T^\omega\left(\theta, z; \theta', z'\right) = \min\left\{1, \frac{\pi_T^\omega(\mathrm{d}\theta')}{\pi_T^\omega(\mathrm{d}\theta)} \frac{q_T(\theta', \mathrm{d}\theta)}{q_T(\theta, \mathrm{d}\theta')} \exp\left(z' - z\right)\right\},$$

$\rho_T^\omega(\theta, z)$ is the corresponding rejection probability.

Our first assumption is that the posterior distributions concentrate towards a Gaussian at rate $1/\sqrt{T}$. We denote by $\mathcal{Y}_T$ the σ-algebra spanned by $Y_{1:T}$.

***Assumption* 4.1.** *The posterior distributions* $\{\pi_T^\omega(\mathrm{d}\theta)\}_{T\geqslant 1}$ *admit Lebesgue densities and there exists a* $d \times d$ *positive definite matrix* Σ*, a parameter value* $\bar{\theta} \in \Theta$ *and a sequence* $(\hat{\theta}_T^\omega)_{T\geqslant 1}$ *of* $\mathcal{Y}_T$*-adapted random variables such that as* $T \to \infty$

$$\int \left|\pi_T^\omega(\theta) - \varphi\left(\theta; \hat{\theta}_T^\omega, \Sigma/T\right)\right| \mathrm{d}\theta \to 0, \qquad \hat{\theta}_T^\omega \to \bar{\theta}, \tag{4.3}$$

both limits being in $\mathbb{P}^Y$*-probability.*

Assumption 4.1 is satisfied if a Bernstein-von Mises theorem holds; see Van der Vaart (2000, Theorem 10.1) and Kleijn and Van der Vaart (2012). Our second assumption is that we use random walk proposal distributions with appropriately scaled increments.

***Assumption* 4.2.** *The proposal distributions* $\{q_T(\theta, \mathrm{d}\theta')\}_{T\geqslant 1}$ *admit densities of the form*

$$q_T(\theta, \theta') = \sqrt{T}\nu\left\{\sqrt{T}(\theta' - \theta)\right\},$$

where ν *is a continuous density on* $\mathbb{R}^d$*.*

Finally, we assume that the error in the log-likelihood estimator satisfies a central limit theorem conditional upon $\mathcal{Y}_T$ and that this convergence holds uniformly in a neighbourhood of $\bar{\theta}$.

***Assumption* 4.3.** *There exists an* ε-ball $B(\bar{\theta})$ *around* $\bar{\theta}$ *such that the distributions of the error in the log-likelihood estimator* $\left\{g_T^\omega(\mathrm{d}z \mid \theta)\right\}_{T\geqslant 1}$ *satisfy as* $T \to \infty$

$$\sup_{\theta\in B(\bar{\theta})} d_{\mathrm{BL}}\left[g_T^\omega(\cdot \mid \theta), \varphi\left\{\cdot; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\}\right] \to 0, \quad \textit{in } \mathbb{P}^Y\textit{-probability}, \tag{4.4}$$

where $d_{\mathrm{BL}}(\cdot,\cdot)$ *denotes the bounded Lipschitz metric and the function* $\sigma : \Theta \to [0,\infty)$ *is continuous at* $\bar{\theta}$ *with* $0 < \sigma(\bar{\theta}) < \infty$*. An analogous result holds for* $\bar{g}_T^\omega(\mathrm{d}z \mid \theta) = \exp(z) g_T^\omega(\mathrm{d}z \mid \theta)$*, the distribution of this error at equilibrium, that is as* $T \to \infty$

$$\sup_{\theta\in B(\bar{\theta})} d_{\mathrm{BL}}\left[\bar{g}_T^\omega(\cdot \mid \theta), \varphi\left\{\cdot; \sigma^2(\theta)/2, \sigma^2(\theta)\right\}\right] \to 0 \quad \textit{in } \mathbb{P}^Y\textit{-probability}. \tag{4.5}$$

We will refer to convergence in probability with respect to the bounded Lipschitz metric as weak convergence in probability. In Section 4.5, we provide sufficient conditions under which Assumption 3 is satisfied for random effect models where the likelihood estimator is a product of T independent importance sampling estimators. This differs from scenarios where the likelihood estimator is given by one single importance sampling estimator studied in

Sherlock, Thiery and Lee (2017). Empirical evidence in Pitt et al. 2012 and Doucet et al. 2015 also suggests that Assumption 3 might hold for a large class of state-space models when the likelihood is estimated using particle filters. Under strong assumptions, a standard central limit theorem has been established in Bérard, Del Moral and Doucet 2014 for $g_T^\omega(\cdot \mid \theta)$. However, it would be technically very challenging to provide weak sufficient conditions under which Assumption 3 holds in this context.

4.3.2 *Weak convergence in the large sample regime*

Denote by $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ the stationary Markov chain defined by the pseudo-marginal kernel, $(\vartheta_{T,0}^\omega, Z_{T,0}^\omega) \sim \pi_T^\omega$ and $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega) \sim P_T^\omega(\vartheta_{T,k-1}^\omega, Z_{T,k-1}^\omega; \cdot)$ for $k \geqslant 1$. Let $\chi_T^\omega = (\tilde{\vartheta}_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ where $\tilde{\vartheta}_{T,k}^\omega = \sqrt{T}(\vartheta_{T,k}^\omega - \hat{\theta}_T^\omega)$ is the Markov chain arising from rescaling the parameter component of the pseudo-marginal chain. Its transition kernel is thus

$$\tilde{P}_T^\omega(\tilde{\theta}, z; \mathrm{d}\tilde{\theta}', \mathrm{d}z') = \tilde{q}_T(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{g}_T^\omega(\mathrm{d}z'|\tilde{\theta}')\tilde{\alpha}_T^\omega\left(\tilde{\theta}, z; \tilde{\theta}', z'\right) + \tilde{\rho}_T^\omega(\tilde{\theta}, z)\delta_{(\tilde{\theta},z)}(\mathrm{d}\tilde{\theta}', \mathrm{d}z'), \quad (4.6)$$

where

$$\tilde{\alpha}_T^\omega(\tilde{\theta}, z; \tilde{\theta}', z') = \min\left\{1, \frac{\tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}')}{\tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta})}\frac{\tilde{q}_T(\tilde{\theta}', \mathrm{d}\tilde{\theta})}{\tilde{q}_T(\tilde{\theta}, \mathrm{d}\tilde{\theta}')}\exp\left(z' - z\right)\right\},$$

$\tilde{\rho}_T^\omega(\theta, z)$ is the corresponding rejection probability, $\tilde{\pi}_T^\omega(\tilde{\theta}) = \pi_T^\omega(\hat{\theta}_T^\omega + \tilde{\theta}/\sqrt{T})/\sqrt{T}$, $\tilde{q}_T(\tilde{\theta}, \tilde{\theta}') = q_T(\hat{\theta}_T^\omega + \tilde{\theta}/\sqrt{T}, \hat{\theta}_T^\omega + \tilde{\theta}'/\sqrt{T})/\sqrt{T}$ and $\tilde{g}_T^\omega(z \mid \tilde{\theta}) = g_T^\omega(z \mid \hat{\theta}_T^\omega + \tilde{\theta}/\sqrt{T})$. Under Assumption 4.2, we have $\tilde{q}_T(\tilde{\theta}, \tilde{\theta}') = \nu(\tilde{\theta}' - \tilde{\theta}) = \tilde{q}(\tilde{\theta}, \tilde{\theta}')$. We now state the main result of this paper.

4.1 Theorem. *Under Assumptions 4.1, 4.2 and 4.3, the sequence of stationary Markov chains $(\chi_T^\omega)_{T\geqslant 1}$ converges weakly in $\mathbb{P}^Y$-probability as $T \to \infty$ to the law of a stationary Markov chain of initial distribution*

$$\tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma)\varphi\left(\mathrm{d}z; \sigma^2/2, \sigma^2\right) \quad (4.7)$$

and transition kernel

$$\tilde{P}(\tilde{\theta}, z; \mathrm{d}\tilde{\theta}', \mathrm{d}z') = \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\varphi\left(\mathrm{d}z'; -\sigma^2/2, \sigma^2\right)\tilde{\alpha}(\tilde{\theta}, z; \tilde{\theta}', z') + \tilde{\rho}(\tilde{\theta}, z)\delta_{(\tilde{\theta},z)}(\mathrm{d}\tilde{\theta}', \mathrm{d}z') \quad (4.8)$$

where $\sigma = \sigma(\bar{\theta})$,

$$\tilde{\alpha}(\tilde{\theta}, z; \tilde{\theta}', z') = \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)}\frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')}\exp\left(z' - z\right)\right\},$$

and $\tilde{\rho}(\theta, z)$ is the corresponding rejection probability.

Under this asymptotic regime, the limiting transition kernel $\tilde{P}$ in (4.8) is also a pseudo-marginal kernel where the noise distribution is $\varphi\left(\mathrm{d}z; -\sigma^2/2, \sigma^2\right)$ as assumed in previous analyses (Doucet et al. 2015; Pitt et al. 2012; Sherlock, Thiery, Roberts et al. 2015). As Theorem 4.1 is a weak convergence result, it does not imply that the integrated autocorrelation time of the pseudo-marginal kernel $\tilde{P}_T^{\omega}$ converges to the one of $\tilde{P}$. However, for large T, this suggests that some characteristics of $\tilde{P}_T^{\omega}$ can indeed be captured by those of the kernel (4.2) which can be obtained from $\tilde{P}$ by using the change of variables $\theta = \hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T}$ and substituting the true target for its normal approximation $\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)$, hence removing a level of approximation.

4.4 Outline of the Proof of the Main Result

4.4.1 Random Markov chains

The proof of Theorem 4.1 follows from a slightly more general result on weak convergence of random Markov chains on Polish spaces given in Theorem 4.5 below. We introduce here some notation and recall some definitions concerning random probability measures needed to define random Markov chains; see the supplementary material or Crauel (2003) for more details.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and S a Polish space endowed with its Borel σ-algebra $\mathcal{B}(S)$. We equip the product space $\Omega \times S$ with the product σ-algebra $\mathcal{F} \otimes \mathcal{B}(S)$. We denote by $\mathcal{P}(S)$ the space of Borel probability measures which is itself endowed with the Borel σ-algebra $\mathcal{B}\{\mathcal{P}(S)\}$ generated by the weak topology. Finally, $C_b(S)$, respectively $\mathrm{BL}(S)$, denote the sets of continuous bounded functions, respectively the set of bounded Lipschitz functions.

4.2 Definition. A *random probability measure* is a map $\mu : \Omega \times \mathcal{B}(S) \to [0, 1]$, $(\omega, B) \mapsto \mu(\omega, B) = \mu^{\omega}(B)$, such that for every $B \in \mathcal{B}(S)$ the map $\omega \mapsto \mu(\omega, B)$ is measurable while $\mu^{\omega} \in \mathcal{P}(S)$ $\mathbb{P}-$almost surely.

For all bounded and measurable functions $g : \Omega \times S \to \mathbb{R}$, $\omega \mapsto \int_S g(\omega, x)\mu^{\omega}(\mathrm{d}x)$ is measurable (Crauel 2003, Proposition 3.3) and thus the map $\omega \mapsto \mu^{\omega}(f)$ is a random variable for bounded measurable functions $f : S \to \mathbb{R}$. Consequently, $\mu^{\omega} : \Omega \to \mathcal{P}(S)$ is a Borel measurable map. Conversely, it can be shown that any random element of $\{\mathcal{P}(S), \mathcal{B}(\mathcal{P}(S))\}$ fulfils the conditions set out in Definition 4.2; see Crauel (2003, Remark 3.20 (i)) or Kallenberg (2006, Lemma 1.37).

4.3 Definition. A *random Markov kernel* is a map $K : \Omega \times S \times \mathcal{B}(S) \to [0,1]$, $(\omega, x, B) \mapsto K(\omega, x, B) = K^\omega(x, B)$, such that

(i) $(\omega, x) \mapsto K^\omega(x, B)$ is $\mathcal{F} \otimes \mathcal{B}(S)$-measurable for every $B \in \mathcal{B}(S)$,

(ii) $K^\omega(x, \cdot) \in \mathcal{P}(S)$ $\mathbb{P}-$almost surely for every $x \in S$.

4.4 Lemma. Given a random probability measure μ^ω and random Markov kernel K^ω, there exists an almost surely unique random probability measure $\mu^{\mathbb{N},\omega}$ on $S^{\mathbb{N}}$ such that

$$\mu^{\mathbb{N},\omega}(A_1 \times \ldots \times A_k \times E_{k+1}) = \int_{A_1} \mu^\omega(\mathrm{d}x_1) \int_{A_2} K^\omega(x_1, \mathrm{d}x_2) \ldots \int_{A_k} K^\omega(x_{k-1}, \mathrm{d}x_k)$$

for any $A_i \in \mathcal{B}(S)$ $(i = 1, \ldots, k)$, $k \in \mathbb{N}$ *and* $E_{k+1} = \times_{i=k+1}^{\infty} S$.

4.4.2 *Convergence of random Markov chains*

For a sequence of random probability measures $(\mu_n^\omega)_{n\geqslant 1}$, respectively a sequence of random Markov kernels $(K_n^\omega)_{n\geqslant 1}$, converging in a suitable sense towards a probability measure μ, respectively a Markov kernel K, we show here that the distributions of the associated Markov chains $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ defined in Lemma 4.4 converge weakly in probability to the distribution $\mu^{\mathbb{N}}$ of the homogeneous Markov chain of initial distribution μ and Markov kernel K.

4.5 Theorem. *If the following assumptions hold,*

(T.1) the random probability measures $\left(\mu_n^\omega\right)_{n\geqslant 1}$ *converge weakly in probability to a probability measure* μ *as* $n \to \infty$*,*

(T.2) the random Markov transition kernels $\left(K_n^\omega\right)_{n\geqslant 1}$ *satisfy*

$$\int \left| K_n^\omega f(x) - K f(x) \right| \mu_n^\omega(\mathrm{d}x) \to 0$$

in probability as $n \to \infty$ *for all* $f \in \mathrm{BL}(S)$ *where* K *is a Markov transition kernel ,*

(T.3) the transition kernel K *is such that* $x \mapsto K f(x)$ is continuous for any $f \in C_b(S)$,

then, as $n \to \infty$*, the measures* $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ *on* $S^{\mathbb{N}}$ *converge weakly in probability to the measure* $\mu^{\mathbb{N}}$ *induced by the Markov chain with initial distribution* μ *and transition kernel* K*.*

4.4.3 Application to the pseudo-marginal algorithm

Theorem 4.1 follows from Theorem 4.5 by showing that, under Assumptions 4.1, 4.2 and 4.3, all conditions set out in Theorem 4.5 are fulfilled. Firstly, as we increase the number of data points, the stationary distribution of the Markov chain will converge weakly to the limiting stationary distribution of Theorem 4.5.

4.6 Proposition. *Under Assumptions 4.1 and 4.3, we have*

$$\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) \to \tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z),$$

weakly in $\mathbb{P}^Y$*-probability as* $T \to \infty$ *where* $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})\exp(z)\,\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})$. $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})$ *and* $\tilde{g}_T^{\omega}(\mathrm{d}z)$ *are defined in Section 4.3.2 and* $\tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z)$ *in equation* (4.7).

This follows as the marginal $\pi_T^{\omega}(\mathrm{d}\theta)$ concentrates around the limiting parameter value $\bar{\theta}$ while the noise uniformly converges towards a normal distribution in a neighbourhood around $\bar{\theta}$. The next proposition ensures the stability of the transition and can be proven using similar arguments.

4.7 Proposition. *Under Assumptions 4.1, 4.2 and 4.3, as* $T \to \infty$ *we have for any* $f \in \mathrm{BL}(\mathbb{R}^{d+1})$

$$\int |\tilde{P}_T^{\omega} f(\theta, z) - \tilde{P} f(\theta, z)| \tilde{\pi}_T^{\omega}(\mathrm{d}\theta, \mathrm{d}z) \to 0, \quad \text{in } \mathbb{P}^Y\text{-probability},$$

where the transition kernels $\tilde{P}_T^{\omega}$ *and* $\tilde{P}$ *are defined in equations* (4.6) *and* (4.8).

A further requirement to ensure the stability of the transition is that the application of the transition operator conserves continuity.

4.8 Proposition. *Under Assumption 4.2, the map* $(\theta, z) \mapsto \tilde{P} f(\theta, z)$ *is continuous for every* $f \in C_b(\mathbb{R}^{d+1})$.

Theorem 4.1 now follows from a direct application of Theorem 4.5 as the assumptions *(T.1)*, *(T.2)* and *(T.3)* hold by Proposition 4.6, 4.7 and 4.8, respectively.

4.5 Random effects models

4.5.1 Statistical model and likelihood estimator

We provide here sufficient conditions under which weak convergence of the pseudo-marginal algorithm is verified for an important class of latent variable models. Consider the model

$$X_t \sim f(\cdot \mid \theta), \qquad Y_t \mid X_t \sim g(\cdot \mid X_t, \theta), \tag{4.9}$$

where $(X_t)_{t\geqslant 1}$ are independent $\mathbb{R}^k$-valued latent variables, $f(x \mid \theta)$ is a density with respect to Lebesgue measure and $\left(Y_t\right)_{t\geqslant 1}$ are Y-valued observations distributed according to a conditional density $g(y \mid x, \theta)$ with respect to a dominating measure, Y being a topological space. For observations $Y_{1:T} = y_{1:T}$ the likelihood is

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid \theta) = \prod_{t=1}^{T} \int g(y_t \mid x_t, \theta) f(x_t \mid \theta) \mathrm{d}x_t.$$

In many scenarios, this likelihood is not available analytically. In order to perform Bayesian inference about the parameter θ, we can thus use the pseudo-marginal algorithm as it is possible to obtain an unbiased non-negative estimator of $p(y_{1:T} \mid \theta)$ using importance sampling. Indeed, we can consider $\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \hat{p}(y_t \mid \theta, U_t)$ where $U = \left(U_1, ..., U_T\right)$, $U_t = \left(U_{t,1}, ..., U_{t,N}\right)$, $U_{t,i}$ is $\mathbb{R}^k$-valued, N denotes the number of Monte Carlo samples and $\hat{p}(y_t \mid \theta, U_t)$ is an importance sampling estimator of $p(y_t \mid \theta)$ is

$$\hat{p}(y_t \mid \theta, U_t) = \frac{1}{N} \sum_{i=1}^{N} w(y_t, U_{t,i}, \theta), \qquad w(y_t, U_{t,i}, \theta) = \frac{g(y_t \mid U_{t,i}, \theta) f(U_{t,i} \mid \theta)}{h(U_{t,i} \mid y_t, \theta)},$$

where $U_{t,i} \sim h(\cdot \mid y_t, \theta)$, $h(\cdot \mid y_t, \theta)$ being a probability density on $\mathbb{R}^k$ with respect to Lebesgue measure. In this case, the joint density $m_{T,\theta}(u)$ of all the auxiliary variates used to obtain the likelihood estimator is given by the product over $t = 1, ..., T$ and $i = 1, ..., N$ of $h(u_{t,i} \mid y_t, \theta)$. We will assume subsequently that the true observations are independent and identically distributed samples taken from a probability measure μ so the joint data distribution is the product measure $\mathbb{P}^Y(\mathrm{d}\omega) = \prod_{t=1}^{\infty} \mu(\mathrm{d}y_t)$.

4.5.2 Verifying the assumptions

The Bernstein–von Mises theorem holds under weak regularity assumptions; see Van der Vaart (2000, Theorem 10.1) and the supplementary material (Section S3·2) for the case of generalized linear mixed models presented in Section 4.5.3. This ensures Assumption 4.1 is

satisfied while Assumption 4.2 is easy to satisfy, selecting for example a multivariate normal proposal of covariance scaling as $1/T$. Assumption 4.3 is more complicated as it requires to establish uniform conditional central limit theorems for $\hat{p}(Y_{1:T} \mid \theta, U)$ in scenarios where $U \sim m_{T,\theta}$ arise from the proposal, so $Z \sim g_T^\omega(\cdot \mid \theta)$, or at stationarity where $U \sim \pi_T^\omega(\cdot \mid \theta)$ with

$$\pi_T^\omega(u \mid \theta) = \frac{\hat{p}(y_{1:T} \mid \theta, u)}{p(y_{1:T} \mid \theta)} m_{T,\theta}(u),$$

implying that $Z \sim \bar{g}_T^\omega(\cdot \mid \theta)$. We denote

$$\sigma^2(y, \theta) = \operatorname{var}\left\{\overline{w}(y, U_{1,1}, \theta)\right\}, \quad \sigma^2(\theta) = E\left\{\sigma^2(Y_1, \theta)\right\},$$

with $U_{1,1} \sim h(\cdot \mid y, \theta)$, $Y_1 \sim \mu$ and

$$\overline{w}(Y_t, U_{t,i}, \theta) = \frac{w(Y_t, U_{t,i}, \theta)}{p(Y_t \mid \theta)}. \tag{4.10}$$

However, under the following assumption, we show here that Assumption 4.3 holds.

***Assumption* 4.4.** *There exists a closed ε-ball $B(\bar{\theta})$ around $\bar{\theta}$ and a function g such that the normalized weight $\overline{w}(y, U_{1,1}, \theta)$ defined in (4.10) satisfies for some $\Delta > 0$*

$$\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U_{1,1}, \theta)^{2+\Delta}\right\} \leqslant g(y),$$

where $U_{1,1} \sim h(\cdot \mid y, \theta)$ and $\mu(g) < \infty$. Additionally, $\theta \mapsto \sigma^2(y, \theta)$ is continuous in θ on $B(\bar{\theta})$ for all $y \in \mathsf{Y}$.

4.9 Theorem. *Under Assumption 4.4, Assumption 4.3 is satisfied.*

Theorem 4.9 strengthens earlier results of Deligiannidis, Doucet and Pitt (2018, Theorem 1) which obtain standard central limit theorems for the error in the log-likelihood estimator.

4.5.3 Generalized linear mixed models

A common example of random effects models is the class of generalized linear mixed models (McCulloch and Neuhaus 2005), where the observation density is a member of the exponential family and the latent variable follows a centred Gaussian distribution. The densities with

respect to some dominating measure can be written as

$$g(y \mid x, \theta) = \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) T(y_j) - A\{\eta_j(x)\}\right], \quad f(x \mid \theta) = \varphi(x; 0, \tau^2), \tag{4.11}$$

where $\eta_j(x) = c_j^{\mathsf{T}}\beta + x$, c is a vector of covariates with corresponding parameter vector β, $A(\eta)$ denotes the log-partition function and $m(y)$ is a base measure. In section S3·2 of the supplementary material, we show that for many such models the assumptions of Theorem 4.1 can be verified when using appropriately chosen importance sampling proposals. In particular, we show that Assumption 4 holds and consequently Assumption 3 holds by Theorem 4.9.

4.6 *Efficient Implementation of the Pseudo-Marginal Random Walk Algorithm*

4.6.1 *Optimal tuning*

We optimize the performance of the limiting pseudo-marginal chain identified in Theorem 4.1 as a proxy for the optimization of the original pseudo-marginal chain. We assume that the limiting covariance matrix Σ in (4.3) is the identity matrix I_d with d denoting the parameter dimension. For general covariance matrices, we can use a Cholesky decomposition and a change of variables as in (Nemeth, Sherlock and Fearnhead 2016; Sherlock, Thiery, Roberts et al. 2015). We denote by $\tilde{P}_{\ell,\sigma}$ the transition kernel (4.8) using the proposal density

$$q(\theta, \theta') = \varphi\left(\theta'; \theta, \ell^2 I_d / d\right).$$

As in Pitt et al. (2012) and Doucet et al. (2015), we propose to minimize $\mathrm{CT}(f, \tilde{P}_{\ell,\sigma})$, as defined in (4.1), with respect to the noise standard deviation σ but, contrary to these contributions, also with respect to the scale parameter ℓ. We restrict attention here to the case where $f(\theta, z) = \theta_1$, the first component of θ, and write $\mathrm{CT}(f, \tilde{P}_{\ell,\sigma}) = \mathrm{CT}(\ell, \sigma)$ in this case. As this criterion is not available in closed-form, we simulate the limiting Markov chain initialized in its stationary regime with different noise levels σ and scales ℓ on a fine grid to obtain empirical estimates of $\mathrm{CT}(\ell, \sigma)$ computed using the overlapping batch mean estimator. This simulation is straightforward as the target and noise distributions in this limiting case are both Gaussian. We then find the approximate minimizer $(\hat{\ell}_{\mathrm{opt}}, \hat{\sigma}_{\mathrm{opt}})$ of $\mathrm{CT}(\ell, \sigma)$ over this grid. This set-up is applied for parameter dimension d ranging from 1 to 50. The results are summarized in Table 4.1.

Table 4.1 also lists the computing time at these values and the average acceptance probability

Dimension d	$\hat{\ell}_{\text{opt}}$	$\hat{\sigma}_{\text{opt}}$	$\text{CT}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$	$\text{pr}_{\text{acc}}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$
$d = 1$	2·05 (0·25)	1·16 (0·07)	8·47	25·73%
$d = 2$	1·97 (0·14)	1·21 (0·06)	12·71	22·92%
$d = 3$	2·11 (0·07)	1·24 (0·05)	16·79	19·97%
$d = 5$	2·17 (0·12)	1·30 (0·05)	23·18	17·35%
$d = 10$	2·20 (0·08)	1·44 (0·05)	37·93	14·27%
$d = 15$	2·33 (0·08)	1·50 (0·00)	53·43	12·07%
$d = 20$	2·34 (0·10)	1·54 (0·05)	65·62	11·44%
$d = 30$	2·36 (0·11)	1·61 (0·03)	90·46	10·41%
$d = 50$	2·41 (0·10)	1·74 (0·05)	136·38	8·66%

Table 4.1: Optimal values for scaling ℓ and noise σ and associated value of computing time and average acceptance probability: mean and standard deviation of the minimizers over 10 runs.

of the proposal under $\tilde{P}_{\ell,\sigma}$ at stationarity by using 5 million iterates of the chain. The obtained results are consistent with those in Doucet et al. (2015) and Sherlock, Thiery, Roberts et al. (2015). For low dimensions, $1 \leqslant d \leqslant 5$, the ideal Metropolis–Hastings algorithm mixes well and $\hat{\sigma}_{\text{opt}}$ is around 1·1-1·3 as suggested by Doucet et al. (2015) and it increases slowly as d increases to the values $(\ell_\infty, \sigma_\infty) = (2{\cdot}56, 1{\cdot}81)$ obtained by the diffusion limit (Sherlock, Thiery, Roberts et al. 2015). For example, for $d = 50$, we obtain $(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}}) = (2{\cdot}41, 1{\cdot}74)$ and the resulting optimal computing time $\text{CT}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$ is close to $\text{CT}(\ell_\infty, \sigma_\infty)$. For lower dimensions, however, the performance in terms of computing time can be increased by reducing σ and ℓ in comparison to σ_∞ and ℓ_∞; see Table 4.2. We also observed empirically that the cost function $\ell \mapsto \text{CT}(\ell, \sigma)$ is fairly flat as noticed in the limiting case by Sherlock, Thiery, Roberts et al. (2015).

4.6.2 Implementation

We now show how to exploit the results of the last section in practice to design an efficient implementation of the pseudo-marginal algorithm. Using a preliminary run, we compute estimates $\hat{\theta}$, $\hat{\Sigma}$ of the posterior mean and posterior covariance matrix. For the parameter dimension d, we choose ℓ according to Table 4.1 and use a Gaussian random walk proposal with covariance matrix $\hat{\ell}^2_{\text{opt}}\hat{\Sigma}/d$. Finally we select the number of Monte Carlo samples N such that the sample standard deviation of the log-likelihood estimate at $\hat{\theta}$ matches the optimal value $\hat{\sigma}_{\text{opt}}$ listed in Table 4.1. This approach is similar to the one followed in Sherlock, Thiery, Roberts et al. (2015) except for the dimension dependence of the recommended parameters $(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$.

Dimension d	CT($\ell_\infty, \hat{\sigma}_{\text{opt}}$)	CT($\ell_\infty, \sigma = 1.2$)	CT($\ell_\infty, \sigma_\infty$)
$d = 1$	9·04 (0·25)	9·05 (0·21)	17·10 (1·34)
$d = 2$	13·48 (0·32)	13·37 (0·28)	22·45 (0·81)
$d = 3$	17·63 (0·28)	17·43 (0·26)	26·71 (0·64)
$d = 5$	24·38 (0·44)	24·72 (0·31)	34·14 (0·88)
$d = 10$	40·17 (0·71)	41·60 (0·24)	47·08 (1·03)
$d = 15$	53·69 (0·72)	58·01 (0·50)	59·08 (0·79)
$d = 20$	67·15 (0·53)	74·34 (0·36)	71·41 (1·48)
$d = 30$	91·36 (0·95)	106·08 (0·34)	93·73 (1·08)
$d = 50$	136·49 (1·18)	167·83 (0·75)	135·92 (1·27)

Table 4.2: Comparison of the computing time for different noise levels. $\hat{\sigma}_{\text{opt}}$ denotes the minimizer of the estimated integrated autocorrelation time, as shown in Table 4.1.

4.7 Simulation study: Random Effects Model

We now illustrate how the guidelines derived from the limiting pseudo-marginal chain compare to a practical implementation of the pseudo-marginal algorithm. We consider a logistic mixed effects model applied to a real data set. Mixed models are popular in econometrics, survey analysis and medical statistics amongst others and are often used to describe heterogeneity between groups. Here we consider a subset of a cohort study of Indonesian preschool children. This dataset was previously analysed using Bayesian mixed models by Zeger and Karim (1991). It contains 1200 observations of 275 children. We model the probability of a respiratory infection based on the following covariates: age, sex, height, an indicator for presence of vitamin deficiency, an indicator for subnormal height and two seasonal components. Including the intercept we have 8 covariates. Cluster effects due to repeated measurements of the same children are modelled with individual random intercepts. In this case the linear predictor of a regression model based on covariates $c_{t,j}$ ($t = 1, \dots, T, j = 1, \dots J$) reads $\eta_{t,j} = c_{t,j}^{\mathsf{T}}\beta + X_t$ where $X_t \sim \mathcal{N}(0, \tau)$ denotes the random intercept for children $t = 1, \dots, T$ and β the regression parameters. For every child, we have an observation vector $y_t = \left(y_{t,1} \dots, y_{t,J}\right) \in \{0, 1\}^J$. The unknown parameter is $\theta = (\beta, \tau) \in \mathbb{R}^d$ where $d = 9$. The observations are assumed conditionally independent given the random effects and are modelled through

$$g(y_t \mid x_t, \theta) = \prod_{j=1}^{J} \frac{\exp(y_{t,j}\eta_{t,j})}{1 + \exp(\eta_{t,j})}, \quad f(x_t \mid \theta) = \varphi(x_t; 0, \tau), \quad t = 1, \dots, T.$$

Inference in mixed effects models often aims at finding the population effects and thus one is interested in integrating out the random effects. Since the marginal likelihood contains intractable integrals, this model lends itself to the pseudo-marginal approach. We obtain an unbiased estimator of the marginal likelihood by estimating the integrals using an importance

Particles N	$\hat{\sigma}$	$\widehat{\text{IAT}}$	$\hat{\text{pr}}_{\text{acc}}$	$\widehat{\text{IAT}}\left(\tilde{P}_{\ell=2{\cdot}2,\sigma=\hat{\sigma}}\right)$	$\hat{\text{pr}}_{\text{acc}}\left(\tilde{P}_{\ell=2{\cdot}2,\sigma=\hat{\sigma}}\right)$
12	2·00	140·22	8·93%	162·57	7·67%
15	1·76	112·06	10·70%	121·70	9·93%
18	1·63	98·69	12·30%	94·14	11·73%
21	1·46	72·42	13·93%	72·31	14·00%
24	1·34	66·29	15·10%	64·45	15·55%
27	1·29	61·95	16·08%	58·08	16·39%
30	1·22	58·70	16·85%	54·12	17·52%
33	1·16	52·39	17·77%	50·26	18·16%

Table 4.3: For N particles: standard deviation $\hat{\sigma}$ of the log-likelihood estimator at the mean, average integrated autocorrelation time $\widehat{\text{IAT}}$ and average acceptance probability $\hat{\text{pr}}_{\text{acc}}$ for pseudo-marginal kernel and limiting kernel $\tilde{P}_{\ell,\hat{\sigma}}$ for $\ell = 2{\cdot}2$.

sampling estimator

$$h(u \mid y_t, \theta) = \varphi(u; \widehat{x}_t, \tau_q^2), \quad \widehat{x}_t = \arg\max_{x_t} g(y_t \mid x_t, \theta) f(x_t \mid \theta)$$

with proposal variance $\tau_q > 0$. We provide more details to importance sampling for mixed effects models in Section S3·3 where we also show that Assumption 4 is satisfied in the present example. For the covariate parameters we assume a diffuse Gaussian prior and the variance of the random effects are assigned an inverse gamma prior. We run a pseudo-marginal algorithm with a Gaussian random walk proposal for 500000 iterations. The covariance of the proposal is set equal to the posterior covariance of the parameters estimated in a preliminary run and scaled by $\ell^2/d = (2{\cdot}2)^2/9$. We compare the average integrated autocorrelation time and the acceptance rate with that of the limiting chain using the same $\ell = 2{\cdot}2$ and $\sigma = \hat{\sigma}$, the average being defined as $\widehat{\text{IAT}}(P_T^\omega) = \sum_{i=1}^d \text{IAT}(f_i, P_T^\omega)$ for $f_i(\theta, z) = \theta_i$ the i^{th} parameter component. Here, $\hat{\sigma}$ is the standard deviation of the log-likelihood estimator obtained using 10000 samples of the marginal likelihood evaluated at an estimate $\hat{\theta}$ of the posterior mean. The results are summarized in Table 4.3. For a given number of particles N, we report the associated estimate of the noise in the log-likelihood estimator, the average integrated autocorrelation time averaged and the average acceptance rate.

The average integrated autocorrelation time and the acceptance rate are very close to those of the limiting algorithm. This is visualized in Figure 4.1 where we plot the same quantities against the number of particles N. The computing time of the pseudo-marginal algorithm targeting the posterior, $\widehat{\text{CT}}(P_T^\omega) = \widehat{\text{IAT}}(P_T^\omega)/\hat{\sigma}^2$, and the computing time of the limiting algorithm, $\widehat{\text{CT}}(\tilde{P}_{\ell,\sigma})$, are both optimized for $\hat{\sigma} = 1{\cdot}46$, as expected from Table 4.1. In this example, the limiting kernel captures very well the behaviour of the pseudo-marginal algorithm for large data sets and Table 4.1 thus provides useful guidelines on how to tune this

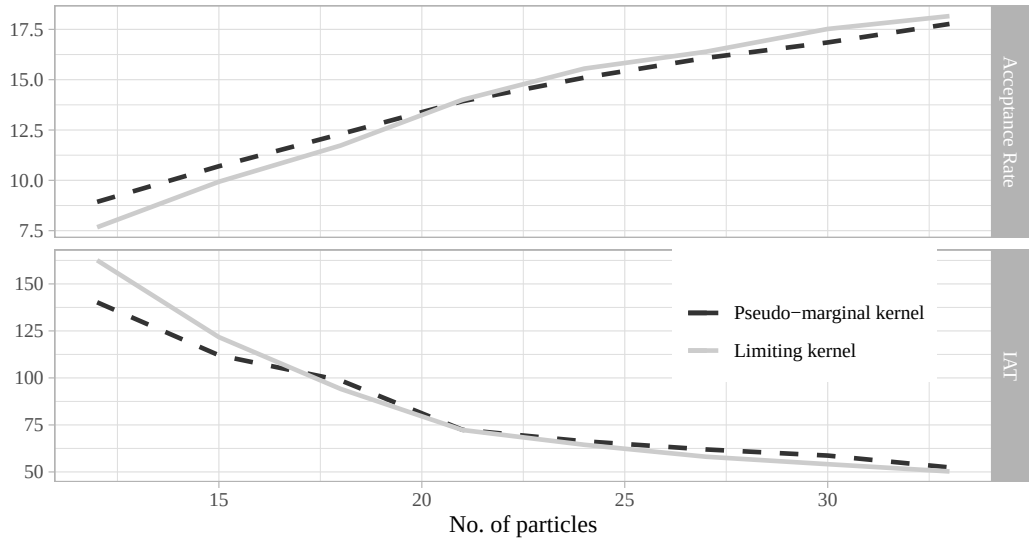


Figure 4.1: Average integrated autocorrelation time (top) and average acceptance rate (bottom) for the pseudo-marginal algorithm as a function of N and the limiting transition kernel $\tilde{P}_{\ell,\hat{\sigma}}$ for $\ell = 2{\cdot}2$.

scheme.

References

Andrieu, Christophe, Arnaud Doucet and Roman Holenstein (2010). 'Particle Markov chain Monte Carlo methods (with Discussion)'. *J. R. Statist. Soc.* B 72.3, pp. 269–342 (cit. on p. 51).

Andrieu, Christophe and Gareth Owen Roberts (2009). 'The pseudo-marginal approach for efficient Monte Carlo computations'. *Ann. Statist.* 37, pp. 697–725 (cit. on pp. 51, 53).

Andrieu, Christophe and Matti Vihola (2015). 'Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms'. *Ann. Appl. Probab.* 25.2, pp. 1030–1077 (cit. on p. 53).

– (2016). 'Establishing some order amongst exact approximations of MCMCs'. *Ann. Appl. Probab.* 26.5, pp. 2661–2696 (cit. on p. 51).

Beaumont, Mark A (2003). 'Estimation of population growth or decline in genetically monitored populations'. *Genetics* 164.3, pp. 1139–1160 (cit. on p. 51).

Bérard, Jean, Pierre Del Moral and Arnaud Doucet (2014). 'A lognormal central limit theorem for particle approximations of normalizing constants'. *Electron. J. Probab.* 19.94, pp. 1–28 (cit. on pp. 52, 58).

Crauel, Hans (2003). *Random Probability Measures on Polish Spaces*. CRC Press (cit. on p. 59).

Deligiannidis, George, Arnaud Doucet and Michael K Pitt (2018). 'The correlated pseudo-marginal method'. *J. R. Statist. Soc.* B 80.5, pp. 839–870 (cit. on pp. 52, 63).

Doucet, Arnaud, Michael K Pitt, George Deligiannidis and Robert Kohn (2015). 'Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator'. *Biometrika* 102.2, pp. 295–313 (cit. on pp. 51–55, 58–59, 64–65).

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer-Verlag: New York (cit. on p. 59).

Kleijn, Bas J K and Aad W Van der Vaart (2012). 'The Bernstein-Von-Mises theorem under misspecification.' *Electron. J. Statist.* 6, pp. 354–381 (cit. on p. 57).

Lin, L, KF Liu and J Sloan (2000). 'A noisy Monte Carlo algorithm'. *Phys. Rev.* D 61.7, p. 074505 (cit. on p. 51).

McCulloch, Charles E and John M Neuhaus (2005). 'Generalized linear mixed models'. *Encyclopedia of Biostatistics* 4 (cit. on p. 63).

Nemeth, Christopher, Chris Sherlock and Paul Fearnhead (2016). 'Particle Metropolis-adjusted Langevin algorithms'. *Biometrika* 103.3, pp. 701–717 (cit. on pp. 52, 64).

Pitt, Michael K, Ralph dos Santos Silva, Paolo Giordani and Robert Kohn (2012). 'On some properties of Markov chain Monte Carlo simulation methods based on the particle filter'. *J. Econometrics* 171.2, pp. 134–151 (cit. on pp. 51–55, 58–59, 64).

Roberts, Gareth O and Jeffrey S Rosenthal (2014). 'Minimising MCMC variance via diffusion limits, with an application to simulated tempering.' *Ann. Appl. Probab.* 24, pp. 131–149 (cit. on p. 55).

Roberts, G.O., A. Gelman and W.R. Gilks (1997). 'Weak convergence and optimal scaling of random walk Metropolis algorithms.' *Ann. Appl. Probab.* 7, pp. 110–120 (cit. on p. 52).

Sherlock, Chris, Alexandre H Thiery and Anthony Lee (2017). 'Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators'. *Biometrika* 104.3, pp. 727–734 (cit. on p. 58).

Sherlock, Chris, Alexandre H Thiery, Gareth O Roberts, Jeffrey S Rosenthal et al. (2015). 'On the efficiency of pseudo-marginal random walk Metropolis algorithms'. *Ann. Statist.* 43.1, pp. 238–275 (cit. on pp. 51–55, 59, 64–65).

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Cambridge University Press (cit. on pp. 57, 62).

Zeger, S. L. and M. R. Karim (1991). 'Generalized linear models with random effects; a Gibbs sampling approach'. *J. Am. Statist. Ass.* 86.413, pp. 79–86 (cit. on p. 66).

Supplementary Material

This supplementary material contains the proofs to all theorems and propositions, some background material and additional simulation studies. Section 4.8 includes a brief survey of weak convergence results for random probability measures on Polish spaces which play an important role in this chapter. We have not been able to find some of the precise statements we require in the literature so we present their proofs here without any claim of originality. Sections 4.9 and 4.10 provide the proofs for sections 4.4 and 4.5, respectively. Finally, section 4.12 includes some additional numerical examples: a toy example and a Lotka-Volterra model where the likelihood is estimated using a particle filter as opposed to importance sampling.

4.8 Random Measures and Weak Convergence on Polish Spaces

4.8.1 Weak Convergence

Let S be a Polish space, endowed with the Borel σ-algebra $\mathcal{B}(S)$. We denote d the metric inducing the topology on S and $\mathcal{P}(S)$ the space of Borel probability measures on S. In the following, we will only consider (random) probability measures in $\mathcal{P}(S)$ unless stated otherwise.

4.10 Definition (*Weak convergence*)**.** A sequence of probability measures $(\mu_n)_{n\geqslant 1}$ converges

weakly to a probability measure μ, denoted $\mu_n \rightsquigarrow \mu$, if for all $f \in C_b(S)$

$$\mu_n(f) \to \mu(f) \quad \text{as } n \to \infty, \tag{4.12}$$

where $C_b(S)$ is the set of bounded continuous real-valued functions of domain S.

The set of test functions generating this topology can be restricted to bounded continuous functions $f : S \to [0,1]$ or bounded Lipschitz functions, see for example Crauel (2003, Lemma A.1 and Theorem A.2). The topology of weak convergence can be metrized using the bounded Lipschitz metric which is given for $\mu, \nu \in \mathcal{P}(S)$ by

$$d_{\mathrm{BL}}(\mu,\nu) = \sup\left\{|\mu(f) - \nu(f)| \,;\, f \in \mathrm{BL}(S), \|f\|_{\mathrm{BL}} \leqslant 1\right\}, \tag{4.13}$$

see for example Dudley (2002, Proposition 11.3.2). Here, the set $\mathrm{BL}(S)$ denotes the set of bounded Lipschitz functions and we follow Pollard (2002) by defining the norm

$$\|f\|_{\mathrm{BL}} = \max\left\{\|f\|_{\mathrm{L}}, 2\|f\|_{\infty}\right\}, \tag{4.14}$$

where

$$\|f\|_{\mathrm{L}} = \sup_{x,y:\, x\neq y} \frac{|f(x) - f(y)|}{d(x,y)} \quad \text{and} \quad \|f\|_{\infty} = \sup_x |f(x)|. \tag{4.15}$$

This definition gives us the inequality

$$|f(x) - f(y)| \leqslant \|f\|_{\mathrm{BL}} \left[\min\{1, d(x,y)\}\right] \tag{4.16}$$

for every x, y.

4.8.2 Weak Convergence of Random Measures

We recall here some facts about random probability measures. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. We equip the product space $\Omega \times S$ with the product σ-algebra, $\mathcal{F} \otimes \mathcal{B}(S)$.

4.11 Definition (*Random probability measure*)**.** A *random probability measure* is a map $\mu : \Omega \times \mathcal{B}(S) \to [0,1]$ such that for every $B \in \mathcal{B}(S)$ the map $\omega \mapsto \mu(\omega, B) = \mu^{\omega}(B)$ is measurable while $\mu(\omega, \cdot) \in \mathcal{P}(S)$ for almost every $\omega \in \Omega$.

For all bounded and measurable functions $g : \Omega \times S \to \mathbb{R}$, the assignment $\omega \mapsto \int_S g(\omega, x)\mu^{\omega}(\mathrm{d}x)$ is measurable (see, for example, Crauel 2003, Proposition 3.3) and thus, for random meas-

ures, the map $\omega \mapsto \mu^\omega(f)$ is a random variable. As a consequence we have that $\mu^\omega : \Omega \to \mathcal{P}(S)$ is a Borel measurable map. Conversely, it can be shown that any random element of $[\mathcal{P}(S), \mathcal{B}\{\mathcal{P}(S)\}]$ fulfils the condition set out in Definition 4.2, see (Crauel 2003, Remark 3.20 (i)) or (Kallenberg 2006, Lemma 1.37) for details.

4.12 Definition (*Weak convergence of random measures*)**.** A sequence of random probability measures $(\mu_n^\omega)_{n\geqslant 1}$ converges weakly almost surely to a probability measure μ, denoted $\mu_n^\omega \rightsquigarrow_{a.s.} \mu$, if

$$\mathbb{P}\left(\omega \in \Omega : \ \mu_n^\omega \rightsquigarrow \mu\right) = 1. \tag{4.17}$$

Further, we say that $(\mu_n^\omega)_{n\geqslant 1}$ converges weakly in probability, denoted $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$, if every subsequence contains a further subsequence which converges weakly almost surely.

One can easily verify that the above definition of almost sure weak convergence, respectively weak convergence in probability, is equivalent to $\rho(\mu_n^\omega, \mu) \to 0$ almost surely, respectively in probability, for some metric ρ on $\mathcal{P}(S)$ metrizing weak convergence, e.g., the bounded Lipschitz metric (4.13), see for example Theorem 4.16.

4.13 Remark (*Measurability of probability metric*)**.** As already mentioned above, for any random measure the map $\omega \mapsto \mu^\omega$ is measurable with respect to the Borel σ-algebra $\mathcal{B}\{\mathcal{P}(S)\}$. Moreover, any metric ρ inducing the weak topology on $\mathcal{P}(S)$ is trivially continuous in its first argument and hence the map $\mu^\omega \mapsto \rho(\mu^\omega, \nu)$ for some fixed measure ν is measurable with respect to the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. This implies (Borel) measurability of the map $\omega \mapsto \rho(\mu^\omega, \nu)$ for a non-random measure ν.

In light of the definition of weak convergence (4.12) it is natural to ask whether almost sure weak convergence holds if

$$\mu_n^\omega(f) \xrightarrow{\text{a.s.}} \mu(f) \qquad \text{for all} \qquad f \in C_b(S), \tag{4.18}$$

and similarly whether weak convergence in probability holds if

$$\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f) \qquad \text{for all} \qquad f \in C_b(S). \tag{4.19}$$

In many practical applications, it appears easier to check (4.18) rather than (4.17), similarly checking (4.19) appears easier than having to check that every subsequence of $(\mu_n^\omega)_{n\geqslant 1}$ contains a subsequence which converges weakly almost surely. Relating those statements is inconvenienced by the fact that weak convergence is usually checked using an uncountable

convergence determining class of functions, e.g., the space of bounded continuous functions. However, we show here that these equivalences hold true for Polish spaces; see Theorem 4.16 below.

Almost sure weak convergence can be shown using the existence of a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S) \subset C_b(S)$. Considering subsequences and using a diagonal argument we can show the equivalence of the statement also holds if almost sure convergence is replaced by convergence in probability. For the purposes of this paper we confine our attention to weak convergence in probability. To prove the statements above we first need an auxiliary result, which also appeared in Sweeting (1989, Lemma 4).

4.14 Proposition. *Suppose A is a countable set and consider random variables $X_n(a) : \Omega \to \mathbb{R}$ indexed by $a \in A$ and $n \in \mathbb{N}$. Moreover, assume that for every $a \in A$ the sequence $\{X_n(a)\}_{n\geqslant 1}$ converges to $X(a)$ in probability, i.e.,*

$$X_n(a) \xrightarrow{\mathbb{P}} X(a) \quad \forall a \in A.$$

Then there exists a subsequence $N' \subset \mathbb{N}$ such that along N'

$$\mathbb{P}\left\{\omega : X_n(a) \to X(a) \quad \forall a \in A\right\} = 1.$$

Proof. Choose $a_1 \in A$. Since we have $X_n(a_1) \xrightarrow{\mathbb{P}} X(a_1)$ we can extract a subsequence $n_{1,1}, n_{1,2}, \ldots$ such that

$$\left\{X_{n_{1,1}}(a_1), X_{n_{1,2}}(a_1), X_{n_{1,3}}(a_1), \ldots\right\}$$

converges almost surely. Pick now $a_2 \in A$, we can now extract a further subsequence

$$\left\{X_{n_{2,1}}(a_2), X_{n_{2,2}}(a_2), X_{n_{2,3}}(a_2), \ldots\right\}$$

along which we have almost sure convergence. We can iterate this procedure to get another subsequence

$$\left\{X_{n_{3,1}}(a_3), X_{n_{3,2}}(a_3), X_{n_{3,3}}(a_3), \ldots\right\}.$$

Along the subsequence $N' = \left(n_{1,1}, n_{2,2}, n_{3,3}, \ldots\right)$, we have almost sure convergence of $X_{n'}(a) \to X(a)$ for all $a \in A$. □

The existence of a countable convergence determining class for Polish spaces is guaranteed by the following Proposition. The proof is adapted from Berti, Pratelli and Rigo (2006, Theorem 2.2).

4.15 Proposition. *Consider $\mathcal{P}(S)$ equipped with the Borel σ-algebra generated by the topology of weak convergence. There exists a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S)$.*

Proof. Take a countable set $\{s_1, s_2, \dots\}$ dense in S and let $H = [0,1]^{\mathbb{N}}$ be the Hilbert cube. For $x \in S$, define the map $h : S \to H$ by

$$h(x) = \left\{ d(x, s_1) \wedge 1, d(x, s_2) \wedge 1, \dots \right\}.$$

We can equip H with the topology of coordinate wise convergence. Writing $u = (u_1, u_2, \dots)$ and $v = (v_1, v_2, \dots)$ for elements $u, v \in H$, this topology is induced by the metric

$$\alpha(u, v) = \sum_{i=1}^{\infty} \frac{|u_i - v_i|}{2^i}.$$

The Hilbert cube H is compact by Tychonoff's Theorem (see for example Dudley 2002, Theorem 2.2.8.), h is a homeomorphism from S to $h(S)$ (Borkar 1991, Theorem A.1.1.) and its closure $\overline{h(S)} \subset H$ is compact. For $\mu \in \mathcal{P}(S)$ denote $\nu = \mu \circ h^{-1}$ the image measure on $h(S)$.

Note that any Lipschitz continuous function on $h(S)$ can be extended to $\overline{h(S)}$ without increasing its norm (Dudley 2002, Proposition 11.2.3.). By the Arzelà–Ascoli theorem, the sets $B_n = [f \in \mathrm{BL}\{\overline{h(S)}\} : \|f\|_{\mathrm{BL}} \leqslant n]$ are compact and thus separable under the $\|\cdot\|_\infty$-norm. Therefore $\mathrm{BL}\{\overline{h(S)}\} = \bigcup_{n=1}^{\infty} B_n$ is separable under the $\|\cdot\|_\infty$-norm and so is $\mathrm{BL}\{h(S)\}$. Hence, we can pick a countable set $\mathcal{D}$ which is dense in $\mathrm{BL}\{h(S)\}$. Defining $\mathcal{C} = \{g \circ h : g \in \mathcal{D}\}$ we have $\mathcal{C} \subset \mathrm{BL}(S)$ since for all $x, y \in S$ and $i \in \mathbb{N}$

$$|d(x, s_i) \wedge 1 - d(y, s_i) \wedge 1| \leqslant d(x, y)$$

and thus

$$|g \circ h(x) - g \circ h(y)| \leqslant L_g \alpha\{h(x), h(y)\} = L_g \sum_{i=1}^{\infty} \frac{|d(x, s_i) \wedge 1 - d(y, s_i) \wedge 1|}{2^i} \leqslant L_g d(x, y),$$

where L_g denotes the Lipschitz constant of the function g.

Now assume that $\mu_n(f) \to \mu(f)$ for all $f \in \mathcal{C}$. Then by a change of variable

$$\int_S f \, d\mu_n = \int_S g \circ h \, d\mu_n = \int_{h(S)} g \, d\nu_n \to \int_{h(S)} g \, d\nu$$

for all $g \in \mathcal{D}$. Since $\mathcal{D}$ is dense in $\mathrm{BL}\{h(S)\}$ with respect to the $\|\cdot\|_\infty$-norm we have convergence for all bounded Lipschitz functions and thus $\nu_n \rightsquigarrow \nu$. By continuity of h^{-1} we also have convergence $\mu_n \rightsquigarrow \mu$. □

Equipped with these results we can now prove some equivalences which facilitate the verification of weak convergence of random probability measures in the sense introduced above. We will prove the following statements only for convergence in probability. The modifications for almost sure convergence are obvious.

4.16 Theorem. *Let $\left(\mu_n^\omega\right)_{n\geqslant 1}$ be a sequence of random probability measures and μ a probability measure. Then the following statements are equivalent*

(i) $d_{\mathrm{BL}}(\mu_n^\omega, \mu) \xrightarrow{\mathbb{P}} 0$,

(ii) $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$

(iii) $\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f) \qquad$ *for all* $\qquad f \in C_b(S)$

(iv) $\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f) \qquad$ *for all* $\qquad f \in \mathrm{BL}(S)$.

The same results hold if convergence in probability is replaced by almost sure convergence throughout.

Proof. The equivalence $(i) \Leftrightarrow (ii)$ is immediate since d_{BL} metrizes weak convergence. The implications $(ii) \Rightarrow (iii) \Rightarrow (iv)$ are trivial. To show $(iv) \Rightarrow (ii)$, note that by Proposition 4.15 there exists a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S)$. By virtue of Proposition 4.14 there exists a subsequence $(n_1, n_2, \ldots)$ such that for all $g \in \mathcal{C}$

$$\mu_{n_k}^\omega(g) \xrightarrow{\text{a.s.}} \mu(g) \qquad \text{as } k \to \infty.$$

Now, given $(n_k)_{k\in\mathbb{N}}$ define

$$A(g) = \left\{ \omega \in \Omega : \mu_{n_k}^\omega(g) \longrightarrow \mu(g) \quad \text{as} \quad k \to \infty \right\}.$$

We have $\mathbb{P}\{A(g)\} = 1$ for all $g \in \mathcal{C}$ and for $\bigcap_{g\in\mathcal{C}} A(g) = A \in \mathcal{B}(S)$ we find $\mathbb{P}(A) = 1$. Since we can apply this reasoning to any subsequence we always find a further subsequence such that $(\mu_{n_{k_j}}^\omega)$ converges almost surely. See also Sweeting (1989, Theorem 9) and Berti, Pratelli and Rigo (2006, Theorem 2.2). □

4.17 Remark. *If the random measure is induced by a regular conditional distribution, i.e., let* $\left(\mu_n^\omega\right)_{n\geqslant 1}$ *denote a sequence of transition kernels such that*

$$\mu_n^\omega(\cdot) = \mathbb{P}(X_n \in \cdot \mid \mathcal{F}_n)(\omega) \qquad \mathbb{P}-a.s.$$

for some filtration $\left(\mathcal{F}_n\right)_{n\geqslant 1}$*, we have*

$$\int f(x)\mu_n^\omega(\mathrm{d}x) = E\left\{f(X_n) \mid \mathcal{F}_n\right\}(\omega) \qquad \mathbb{P}-a.s.$$

and thus equivalently to $\mu_n \rightsquigarrow_{\mathbb{P}} \mu$ *then we can write*

$$E\left\{f(X_n) \mid \mathcal{F}_n\right\} \xrightarrow{\mathbb{P}} E\left\{f(X)\right\}, \tag{4.20}$$

where $X \sim \mu$*. For brevity we will also use the notation* $X_n \mid \mathcal{F}_n \rightsquigarrow_{\mathbb{P}} \mu$ *instead of (4.20).*

4.8.3 Product Spaces

We address here the setting where the spaces are of the form $S^k = S \times S \times \cdots \times S$ or $S^{\mathbb{N}} = S \times S \times \ldots$. We will equip these product spaces with the product topology and the respective Borel σ-algebra. The following lemma is helpful to characterize weak convergence in probability in this context.

4.18 Lemma. *For fixed k, let* $(\mu_n^\omega)_{n\geqslant 1}$ *denote random measures on* S^k *and* μ *a non-random measure on* S^k*. Then the following are equivalent*

(i)

$$\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu,$$

(ii)

$$\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f)$$

for all $f \in C_b(S^k)$.

(iii)

$$\int_{S^k} \prod_{i=1}^k f_i(x_i)\mu_n^\omega(\mathrm{d}x_1 \ldots \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^k f_i(x_i)\mu(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$

for all $f_1, \ldots f_k \in C_b(S)$.

(iv)

$$\int_{S^k} \prod_{i=1}^{k} f_i(x_i) \mu_n^\omega(\mathrm{d}x_1 \dots \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^{k} f_i(x_i) \mu(\mathrm{d}x_1 \dots \mathrm{d}x_k)$$

for all $f_1, \dots f_p \in \mathrm{BL}(S)$.

Proof. The implications $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv)$ are trivial. Thus, we only need to show $(iv) \Rightarrow (i)$. We now by Proposition 4.15 that there exists a countable convergence determining class $\mathcal{C} \subset \mathrm{BL}(S)$, so we can assume $f_1, f_2, \dots \in \mathcal{C}$. Without loss of generality we can assume $\|f_i\|_\infty \leqslant 1$ for all i and $1 \in \mathcal{C}$. Then we have that for every $i \in \{1, \dots, k\}$ the marginal of the ith coordinate, denoted $\mu_{n,i}^\omega$, converges to μ_i weakly in probability, i.e. for all i and all $f_i \in \mathcal{C}$ we have

$$\int_S f_i(x) \mu_{n,i}^\omega(\mathrm{d}x_i) \xrightarrow{\mathbb{P}} \int_S f_i(x) \mu_i(\mathrm{d}x_i).$$

Now by Proposition 4.14 for every $i \in \{1, \dots, k\}$ every subsequence $N \subset \mathbb{N}$ contains a further subsequence $N' \subset N$ such that we have convergence almost sure convergence for all $g \in \mathcal{C}$, i.e. denoting

$$A_i := \left\{ \omega \in \Omega : \int_S g(x_i) \mu_{n',i}^\omega(\mathrm{d}x_i) \longrightarrow \int_S g(x_i) \mu_i(\mathrm{d}x_i) \quad \text{for all } g \in \mathcal{C} \right\}$$

we have $P(A_i) = 1$. We can extract a further subsequence $N'' \subset N'$ such that along N'' we have convergence almost surely for all i and all g and thus for $\omega \in A := \cap_{i=1}^k A_i$ the sequence $\left\{ \mu_n^\omega; n \in N'' \right\}$ is tight, since $\left\{ \mu_{n,i}^\omega; n \in N'' \right\}$ is tight for every i (see Ethier and Kurtz 2005, Chapter 3 Proposition 2.4.). We can conclude that for every such ω every subsequence of $(\mu_n^\omega)_{n \geqslant 1}$ has a further subsequence that converges. It remains to show that the functions of the form $\prod_{i=1}^k f_i$ are measure determining. However, by Ethier and Kurtz (2005, Chapter 2 Proposition 4.6.) if $\mathcal{C}$ is measure determining on S then so is the product for S^k. □

If $S = \mathbb{R}^k$ for some $k \in \mathbb{N}$ we can check weak convergence in probability by considering moment generating functions. The following result is shown by Sweeting (1989, Corollary 3); see also Castillo and Rousseau (2015, Lemma 1).

4.19 Proposition. *Let $\left(\mu_n^\omega\right)_{n \geqslant 1}$ be a sequence of random probability measures and assume there exists $u_0 > 0$ such that for all $n \in \mathbb{N}$ the moment generating functions*

$$m_n(u, \omega) = \int \exp\left(u^\mathsf{T} x\right) \mu_n^\omega(\mathrm{d}x)$$

exist for $|u| < u_0$ *then* $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ *if and only if for every* $u \in \mathbb{R}^k$

$$m_n(u, \cdot) \xrightarrow{\mathbb{P}} m(u, \cdot) = \int \exp\left(u^\intercal x\right) \mu^\omega(\mathrm{d}x).$$

Proof. This can be seen by considering the class of functions of the form $f_u(x) = \exp(u^\intercal x)$ for $u \in \mathbb{Q}, \quad |u| < u_0$ and showing that they form a countable convergence determining class, see Sweeting (1989, Corollary 3). Consider the case $k = 1$ and a sequence of measures $(\mu_n)_{n\geqslant 1}$ and μ such that

$$m_n(u) = \int e^{ux} \mu_n(\mathrm{d}x) \to m(u) = \int e^{ux} \mu(\mathrm{d}x).$$

Denote a compact set $K = [-c, c]$. Then by the Markov inequality

$$\mu_n\left(K^\complement\right) = \int_{|x|\geqslant c} \mu_n(dx) \leqslant \frac{m_n(u_0)}{e^{u_0 c}}$$

and $m_n(u_0) \to m(u_0)$. Hence, $\mu_n(K^\complement)$ is bounded and we can find c such that $\sup_n \mu_n(K^\complement) < \epsilon$ and $(\mu_n)_{n\geqslant 1}$ is tight. By continuity the f_u are measure determining so we can conclude that the limit is unique. For $k > 1$ we can use the same argument to show that the marginals are tight, see the proof of Lemma 4.18. □

Lemma 4.18 can be readily extended to countably infinite product spaces by considering convergence of the finite dimensional distribution. Let us therefore denote $\mu \circ \pi_k^{-1} : S^{\mathbb{N}} \to S^k; k \in \mathbb{N}$ the canonical projections. For non-random measures, it is well-known that convergence of the projections already implies convergence on the whole of $S^{\mathbb{N}}$ (Billingsley 1999, Example 2.6). Since there are countably many such projections, we can apply the reasoning of Proposition 4.14 to conclude that for checking $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ on $S^{\mathbb{N}}$ we just need to show

$$\int_{S^k} \prod_{i=1}^k f_i(x_i) \mu_n^\omega(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^k f_i(x_i) \mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k)$$

for all $f_1, \ldots f_k \in \mathrm{BL}(S)$ and $k \in \mathbb{N}$. The following Lemma is essentially a version of Ethier and Kurtz (2005, Chapter 3 Proposition 4.6 b) extended to random measures.

4.20 Lemma. *Let* $\left(\mu_n^\omega\right)_{n\geqslant 1}$ *be a sequence of random probability measures and* μ *a non-random probability measure on* $S^{\mathbb{N}}$*. Then* $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ *is equivalent to*

$$\int_{S^k} \prod_{i=1}^k f_i(x_i) \mu_n^\omega(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^k f_i(x_i) \mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k)$$

for all $f_1, \dots f_k \in \mathrm{BL}(S)$ *and* $k \in \mathbb{N}$.

Proof. Suppose for any k that the above convergence holds for all test functions $f_1, \dots f_k \in \mathrm{BL}(S)$. We have shown in Lemma 4.18 that this is equivalent of convergence of the canonical projections $\mu_n^\omega \circ \pi_k^{-1}$ on S^k (in probability) for any given k. Hence, using Proposition 4.14 for every subsequence $N \subset \mathbb{N}$ there is a subsequence $N' \subset N$ such that along N'

$$\mathbb{P}\left(\omega \in \Omega : \mu_n^\omega \circ \pi_k^{-1} \rightsquigarrow \mu \circ \pi_k^{-1} \text{ as } n \to \infty \quad \text{for all } k \in \mathbb{N}\right) = 1.$$

An application of Ethier and Kurtz (2005, Chapter 3 Proposition 4.6 b) concludes the proof. □

4.9 Proofs of Section 4.4

4.9.1 Proofs for Section 4.4.1

4.21 Lemma. Given a random probability measure μ^ω and random Markov kernel K^ω, there exists an almost surely unique random probability measure $\mu^{\mathbb{N},\omega}$ on $S^{\mathbb{N}}$ such that

$$\mu^{\mathbb{N},\omega}(A_1 \times \ldots \times A_k \times E_{k+1}) = \int_{A_1} \mu^\omega(\mathrm{d}x_1) \int_{A_2} K^\omega(x_1, \mathrm{d}x_2) \ldots \int_{A_k} K^\omega(x_{k-1}, \mathrm{d}x_k)$$

for any $A_i \in \mathcal{B}(S)$ $(i = 1, \dots, k)$, $k \in \mathbb{N}$ *and* $E_{k+1} = \bigtimes_{i=k+1}^{\infty} S$.

Proof of Lemma 4.4. For $\mathbb{P}-$almost all ω, the existence and uniqueness of the distribution $\mu^{\mathbb{N},\omega}$ on $\{S^{\mathbb{N}}, \mathcal{B}(S)^{\mathbb{N}}\}$ can be obtained using the Ionescu-Tulcea extension theorem; see, e.g., Kallenberg (2006, Theorem 6.17) or Klenke (2013, Theorem 14.32). Measurability follows analogously by noting that $\omega \mapsto \mu^{\mathbb{N}}(\omega, A)$ is measurable for any $A \in \mathcal{E} = \{A_1 \times \ldots \times A_k \times E_{k+1}; A_i \in \mathcal{B}(S), i = 1, \dots, k, k \in \mathbb{N}\}$ and that $\mathcal{E}$ forms a $\pi-$system that generates $\mathcal{B}(S)^{\mathbb{N}}$. By Crauel (2003, Remark 3.2) this is enough to obtain measurability for every $A \in \mathcal{B}(S)^{\mathbb{N}}$. □

4.22 Theorem. *If the following assumptions hold,*

(T.1) the random probability measures $\left(\mu_n^\omega\right)_{n \geqslant 1}$ *converge weakly in probability to a probability measure* μ *as* $n \to \infty$*,*

(T.2) the random Markov transition kernels $\left(K_n^\omega\right)_{n \geqslant 1}$ *satisfy*

$$\int \left| K_n^\omega f(x) - K f(x) \right| \mu_n^\omega(\mathrm{d}x) \to 0$$

in probability as $n \to \infty$ *for all* $f \in \mathrm{BL}(S)$ *where* K *is a Markov transition kernel ,*

(T.3) the transition kernel K is such that $x \mapsto Kf(x)$ is continuous for any $f \in C_b(S)$,

then, as $n \to \infty$, the measures $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ on $S^{\mathbb{N}}$ converge weakly in probability to the measure $\mu^{\mathbb{N}}$ induced by the Markov chain with initial distribution μ and transition kernel K.

Proof of Theorem 4.5. By Section 4.8.2 Lemma 4.20, we need to show that for any $k \geqslant 0$ and any $f_0, \ldots, f_k \in \mathrm{BL}(S)$

$$E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) \right\} \xrightarrow{\mathbb{P}} E \left\{ f_0(X_0) \cdots f_k(X_k) \right\} \tag{4.21}$$

where E^ω, resp. E, denotes the expectation w.r.t. the law of $\mathbf{X}_n^\omega$, respectively w.r.t. the law of $\mathbf{X}$. We prove this by induction. For $k = 0$, this follows directly from $(T.1)$. Now assume that (4.21) is true for $k \geqslant 0$, i.e.

$$\left| E^\omega \left\{ f_0(X_{n,0}^\omega) f_1(X_{n,1}^\omega) \cdots f_k(X_{n,k}^\omega) \right\} - E \left\{ f_0(X_0) f_1(X_1) \cdots f_k(X_k) \right\} \right| \xrightarrow{\mathbb{P}} 0.$$

By Lemma 4.18 this is equivalent to weak convergence in probability of the vector of the first k states, i.e., for all $f \in C_b(S^k)$

$$E^\omega \left\{ f(X_0^n, \ldots, X_k^n) \right\} \xrightarrow{\mathbb{P}} E \left\{ f(X_0, \ldots, X_k) \right\}. \tag{4.22}$$

For $k+1$, we have

$$\begin{aligned}
&\left| E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) f_{k+1}(X_{n,k+1}^\omega) \right\} - E \left\{ f_0(X_0) \cdots f(X_k) f_{k+1}(X_{k+1}) \right\} \right| \\
&= \left| E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) K_n^\omega f_{k+1}(X_{n,k}^\omega) \right\} - E \left\{ f_0(X_0) \cdots f(X_k) K f_{k+1}(X_k) \right\} \right| \\
&\leqslant \left| E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) K_n^\omega f_{k+1}(X_{n,k}^\omega) - f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) K f_{k+1}(X_{n,k}^\omega) \right\} \right| \\
&\qquad + \left| E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) K f_{k+1}(X_{n,k}^\omega) \right\} - E \left\{ f_0(X_0) \cdots f_k(X_k) K f_{k+1}(X_k) \right\} \right| \\
&\leqslant E^\omega \left\{ \left| K_n^\omega f_{k+1}(X_{n,k}^\omega) - K f_{k+1}(X_{n,k}^\omega) \right| \right\} \qquad (4.23) \\
&\qquad + \left| E^\omega \left\{ f_0(X_{n,0}^\omega) \cdots f_k(X_{n,k}^\omega) K f_{k+1}(X_{n,k}^\omega) \right\} - E \left\{ f_0(X_0) \cdots f_k(X_k) K f_{k+1}(X_k) \right\} \right|. \qquad (4.24)
\end{aligned}$$

The term (4.23) converges due to (4.5). For the term (4.24), the function Kf_{k+1} is bounded and it is assumed continuous so the function $f_0 \cdots f_k K f_{k+1} \in C_b(S^k)$. Hence this term vanishes by (4.22). □

4.9.2 Some Auxiliary Results

4.23 Lemma. *Under Assumption 4.1, we have*

$$\varphi(\mathrm{d}\theta; \hat{\theta}_T^\omega, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$$

and

$$\pi_T^\omega(\mathrm{d}\theta) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta).$$

Proof. Using the moment generating function of the normal distribution, we have as $T \to \infty$

$$\int e^{u^\mathsf{T}\theta} \varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)\mathrm{d}\theta = \exp\left(u^\mathsf{T}\hat{\theta}_T^\omega + u^\mathsf{T}\Sigma u/2T\right) \xrightarrow{\mathbb{P}^Y} \exp(u^\mathsf{T}\bar{\theta}) = \int \exp(u^\mathsf{T}\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta),$$

where $\delta_{\bar{\theta}}$ denotes the Dirac measure at $\bar{\theta}$ and thus $\varphi(\mathrm{d}\theta; \hat{\theta}_T^\omega, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$ by Proposition 4.19. This implies that for $f \in C_b(\mathbb{R}^d)$

$$\begin{aligned}
&\left|\int f(\theta)\pi_T^\omega(\theta)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right| \\
&\leqslant \left|\int f(\theta)\pi_T^\omega(\theta)\mathrm{d}\theta - \int f(\theta)\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)\mathrm{d}\theta\right| + \left|\int f(\theta)\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right| \\
&\leqslant \|f\|_\infty \int \left|\pi_T^\omega(\theta) - \varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)\right| \mathrm{d}\theta + \left|\int f(\theta)\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right|,
\end{aligned}$$

where the first term on the r.h.s. converges to zero in probability under Assumption 4.1 while the second term converges to zero as $\varphi(\mathrm{d}\theta; \hat{\theta}_T^\omega, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$. Hence, it follows that $\pi_T^\omega(\mathrm{d}\theta) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$. □

To analyse the asymptotic properties of the pseudo-marginal algorithm, we rescale the parameter component. A simple change of variables and the fact that convergence in total variation in probability implies weak convergence in probability shows that the following result holds.

4.24 Lemma. *Under Assumption 4.1, we have*

$$\int \left|\tilde{\pi}_T^\omega(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma)\right| \mathrm{d}\theta \xrightarrow{\mathbb{P}^Y} 0, \quad \text{as} \quad T \to \infty,$$

and thus $\tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) \rightsquigarrow_{\mathbb{P}^Y} \varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma)$.

4.25 Lemma (Convergence of marginal distributions)**.** *Under Assumptions 4.1 and 4.2, the*

marginal distribution of the proposal at stationarity

$$\pi_T^\omega q_T(\mathrm{d}\vartheta) = \int \pi_T^\omega(\mathrm{d}\theta) q_T(\theta, \mathrm{d}\vartheta)$$

satisfies

$$\pi_T^\omega q_T(\mathrm{d}\vartheta) \overset{\mathbb{P}^Y}{\rightsquigarrow} \delta_{\bar{\theta}}(\mathrm{d}\vartheta).$$

Proof. Let $f \in \mathrm{BL}(\mathbb{R})$, then we have

$$\begin{aligned}
\left| \int f(\vartheta) \pi_T^\omega q_T(\mathrm{d}\vartheta) - f(\bar{\theta}) \right| &= \left| \int f(\theta + \xi/\sqrt{T}) \int \pi_T^\omega(\mathrm{d}\theta) \nu(\mathrm{d}\xi) - f(\bar{\theta}) \right| \\
&\leqslant \left| \iint \left(f(\theta + \xi/\sqrt{T}) - f(\theta) \right) \nu(\mathrm{d}\xi) \pi_T^\omega(\mathrm{d}\theta) \right| + \left| \iint f(\theta) \pi_T^\omega(\mathrm{d}\theta) \nu(\mathrm{d}\xi) - f(\bar{\theta}) \right| \\
&\leqslant \iint \left| f(\theta + \xi/\sqrt{T}) - f(\theta) \right| \nu(\mathrm{d}\xi) \pi_T^\omega(\mathrm{d}\theta) + \left| \int f(\theta) \pi_T^\omega(\mathrm{d}\theta) - f(\bar{\theta}) \right|.
\end{aligned}$$

The second term on the r.h.s. vanishes due to Lemma 4.23. For the first term we use the fact that f is bounded Lipschitz, hence

$$\begin{aligned}
\iint \left| f(\theta + \xi/\sqrt{T}) - f(\theta) \right| \nu(\mathrm{d}\xi) \pi_T^\omega(\mathrm{d}\theta) &\leqslant \|f\|_{\mathrm{BL}} \iint \min\left\{ 1, \frac{\|\xi\|}{\sqrt{T}} \right\} \nu(\mathrm{d}\xi) \pi_T^\omega(\mathrm{d}\theta) \\
&= \|f\|_{\mathrm{BL}} \iint \min\left\{ 1, \frac{\|\xi\|}{\sqrt{T}} \right\} \nu(\mathrm{d}\xi) \to 0.
\end{aligned}$$

□

The proof of the following Lemmas are straightforward and thus omitted.

4.26 Lemma. *The map $x \mapsto \min(1, ae^x)$ with $a > 0$ is $1-$Lipschitz, i.e., for all $x, y \in \mathbb{R}$*

$$|\min(1, ae^x) - \min(1, ae^y)| \leqslant |x - y|.$$

4.27 Lemma. *Under Assumption 4.3*

(i) the function

$$\theta \mapsto d_{\mathrm{BL}}\left[\varphi\left\{ \cdot ; \sigma^2(\theta)/2, \sigma^2(\theta) \right\}, \varphi\left\{ \cdot ; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta}) \right\} \mid \mathcal{Y}_T \right]$$

is bounded for all θ and continuous at $\bar{\theta}$;

(ii) for all $f \in \mathrm{BL}(\mathbb{R})$ *the functions*

$$\theta \mapsto \left| \int f(z)\varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} - \int f(z)\varphi\left\{\mathrm{d}z; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta})\right\} \right|$$

are bounded for all θ *and continuous at* $\bar{\theta}$.

4.9.3 Proof of Theorem 4.1

In order to prove Theorem 4.1, we need to prove Propositions 4.6, 4.7 and 4.8 of Section 4·3.

4.28 Proposition. *Under Assumptions 4.1 and 4.3, we have*

$$\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) \to \tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z),$$

weakly in $\mathbb{P}^Y$*-probability as* $T \to \infty$ *where* $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})\exp(z)\,\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})$.

Proof of Proposition 4.6. As established in Lemma 4.18, it is enough to check convergence for products of bounded Lipschitz functions. Now, without loss of generality, assume that $\|f_1\|_\infty, \|f_2\|_\infty \leqslant 1/2$. Then we have

$$\begin{aligned}
&\left| \iint f_1(\tilde{\theta}) f_2(z) \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) - \iint f_1(\tilde{\theta}) f_2(z) \varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma) \varphi\left\{\mathrm{d}z; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta})\right\} \right| \\
&\leqslant \iint e^z \tilde{g}_T^{\omega}(z \mid \tilde{\theta}) \mathrm{d}z \left| \tilde{\pi}_T^{\omega}(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma) \right| \mathrm{d}\tilde{\theta} \\
&\quad + \int \varphi(\tilde{\theta}; 0, \Sigma) \left| \int f_2(z) e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta})\right\} \right| \mathrm{d}\tilde{\theta} \\
&\leqslant \int \left| \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma) \right| \mathrm{d}\tilde{\theta} \qquad (4.25) \\
&\quad + \int \varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T) \left| \int f_2(z) e^z g_T^{\omega}(\mathrm{d}z \mid \theta) - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)2, \sigma^2(\theta)\right\} \right| \mathrm{d}\theta \qquad (4.26) \\
&\quad + \int \varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T) \left| \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\bar{\theta})/2, \sigma^2(\bar{\theta})\right\} \right| \mathrm{d}\theta \\
&\qquad\qquad (4.27)
\end{aligned}$$

The term (4.25) converges to zero in $\mathbb{P}^Y$-probability by Lemma 4.24. For (4.26), write $B(\bar{\theta}) \subset \Theta$ for the ε-ball on which the uniform CLT in Assumption 4.3 holds, that is

$$\sup_{\theta \in B(\bar{\theta})} h_T(\theta) = \sup_{\theta \in B(\bar{\theta})} \left| \int f_2(z) e^z g_T^{\omega}(\mathrm{d}z \mid \theta) \mathrm{d}z - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} \mathrm{d}z \right| \xrightarrow{\mathbb{P}^Y} 0.$$

We can bound (4.26) as follows

$$\begin{aligned}
&\int_{B(\bar\theta)} \varphi(\theta; \hat\theta_T^\omega, \Sigma/T) \left| \int f_2(z) e^z g_T^\omega(\mathrm{d}z \mid \theta) - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} \right| \mathrm{d}\theta \\
&+ \int_{B(\bar\theta)^\complement} \varphi(\theta; \hat\theta_T^\omega, \Sigma/T) \left| \int f_2(z) e^z g_T^\omega(\mathrm{d}z \mid \theta) - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} \right| \mathrm{d}\theta \\
&\leqslant \sup_{\theta \in B(\bar\theta)} h_T(\theta) + \int_{B(\bar\theta)^\complement} \varphi(\theta; \hat\theta_T^\omega, \Sigma/T) \mathrm{d}\theta,
\end{aligned}$$

since $\|f_2\|_\infty \leqslant 1/2$. We have already mentioned that the first term vanishes in probability whereas for the second term we have

$$\int_{B(\bar\theta)^\complement} \varphi(\theta; \hat\theta_T^\omega, \Sigma/T) \mathrm{d}\theta \xrightarrow{\mathbb{P}^Y} \delta_{\bar\theta}\left\{B(\bar\theta)^\complement\right\} = 0,$$

by Lemma 4.23. Thus (4.26) vanishes in $\mathbb{P}^Y$-probability. Finally we consider (4.27). By Lemma 4.27

$$h(\theta) = \left| \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} - \int f_2(z) \varphi\left\{\mathrm{d}z; \sigma^2(\bar\theta)/2, \sigma^2(\bar\theta)\right\} \right|$$

is bounded and continuous at $\bar\theta$. Since $\varphi(\mathrm{d}\theta; \hat\theta_T^\omega, \Sigma/T)$ converges weakly in probability to a point mass in $\bar\theta$ (by Lemma 4.23) we can conclude that

$$\int f(\theta) \varphi(\mathrm{d}\theta; \hat\theta_T^\omega, \Sigma/T) \xrightarrow{\mathbb{P}^Y} \int f(\theta) \delta_{\bar\theta}(\mathrm{d}\theta)$$

for every bounded function f which is continuous at $\overline{\theta}$. In particular,

$$\int h(\theta) \varphi(\mathrm{d}\theta; \hat\theta_T^\omega, \Sigma/T) \xrightarrow{\mathbb{P}^Y} 0.$$

□

4.29 Proposition. *Under Assumptions 4.1, 4.2 and 4.3, as $T \to \infty$ we have for any $f \in \mathrm{BL}(\mathbb{R}^{d+1})$*

$$\int |\tilde P_T^\omega f(\theta, z) - \tilde P f(\theta, z)| \tilde\pi_T^\omega(\mathrm{d}\theta, \mathrm{d}z) \to 0, \quad \textit{in } \mathbb{P}^Y\textit{-probability}.$$

Proof of Proposition 4.7. Let $f \in \mathrm{BL}(\mathbb{R}^{d+1})$. Denote

$$\Pi_T^\omega f(\tilde\theta, z) = \iint f(\tilde\theta', z') \tilde\alpha_T^\omega\left\{(\tilde\theta, z), (\tilde\theta', z')\right\} \tilde q(\tilde\theta, \mathrm{d}\tilde\theta') \tilde g_T^\omega(\mathrm{d}z' \mid \tilde\theta')$$

and

$$\Pi f(\tilde{\theta}, z) = \iint f(\tilde{\theta}', z')\tilde{\alpha}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\}\tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')g(\mathrm{d}z' \mid \overline{\theta}),$$

where $g(\cdot \mid \vartheta) = \varphi\{\cdot; -\sigma^2(\vartheta)/2, \sigma^2(\vartheta)\}$. Then we have

$$\tilde{P}_T^{\omega} f(\tilde{\theta}, z) = \Pi_T^{\omega} f(\tilde{\theta}, z) + f(\tilde{\theta}, z)\left\{1 - \Pi_T^{\omega} 1(\tilde{\theta}, z)\right\}$$

and

$$\tilde{P} f(\tilde{\theta}, z) = \Pi f(\tilde{\theta}, z) + f(\tilde{\theta}, z)\left\{1 - \Pi 1(\tilde{\theta}, z)\right\}. \tag{4.28}$$

Because

$$\begin{aligned}
&E^{\omega}\left\{\left|\tilde{P}_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) - \tilde{P} f(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\} \\
&= E^{\omega}\Bigg[\Big|\Pi_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) + f(\tilde{\vartheta}_0^T, Z_0^T)\left\{1 - \Pi_T^{\omega} 1(\tilde{\vartheta}_0^T, Z_0^T)\right\} \\
&\qquad - \Pi f(\tilde{\vartheta}_0^T, Z_0^T) - f(\tilde{\vartheta}_0^T, Z_0^T)\left\{1 - \Pi 1(\tilde{\vartheta}_0^T, Z_0^T)\right\}\Big|\Bigg] \\
&\leqslant E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) - \Pi f(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\} + E^{\omega}\left\{\left|\Pi_T^{\omega} 1(\tilde{\vartheta}_0^T, Z_0^T) - \Pi 1(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\}
\end{aligned}$$

and $1 \in \mathrm{BL}(\mathbb{R}^{d+1})$ it is sufficient to show that for any choice of $f \in \mathrm{BL}(\mathbb{R}^{d+1})$ we have $E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\theta}, z) - \Pi f(\tilde{\theta}, z)\right|\right\} \overset{\mathbb{P}^Y}{\to} 0$. Thus

$$\begin{aligned}
&E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\theta}, z) - \Pi f(\tilde{\theta}, z)\right|\right\} \\
&= \iint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z)\Bigg|\iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}_T^{\omega}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z')g(\mathrm{d}z' \mid \overline{\theta})\Bigg| \\
&= \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\Bigg|\iint \min\left\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z')g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}'\Bigg|\mathrm{d}\tilde{\theta} \\
&\leqslant \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\Bigg|\iint \min\left\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'
\end{aligned}$$

$$- \iint \min\left\{ \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta}$$

$$+ \iint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \min\left\{ \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'$$

$$- \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta}. \tag{4.29}$$

By taking $\varphi(\tilde{\theta}; 0, \Sigma)$ out in last two lines of (4.29), this can be rewritten as

$$\iint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \min\left\{ \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'$$

$$- \iint \min\left\{ \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta} \tag{4.30}$$

$$+ \iint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'$$

$$- \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta}. \tag{4.31}$$

For (4.30), we use the inequality $|\min(a, b) - \min(c, d)| \leqslant |a - c| + |b - d|$:

$$\iint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \min\left\{ \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')$$

$$- \min\left\{ \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta}) e^{z'-z} \right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta}$$

$$\leqslant \|f\|_{\infty} \iiiint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma)\right| \mathrm{d}\tilde{\theta}$$

$$+ \|f\|_{\infty} \iiiint \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) e^{z'} \tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta}) \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}') - \varphi(\tilde{\theta}'; 0, \Sigma)\right| \mathrm{d}\tilde{\theta}'\mathrm{d}\tilde{\theta}$$

$$= 2\|f\|_{\infty} \int \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma)\right| \mathrm{d}\tilde{\theta} \xrightarrow{\mathbb{P}^Y} 0,$$

by Lemma 4.24. For the part (4.31) note that

$$\iint e^z \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')$$

$$- \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}\left\{(\tilde{\theta}, z), (\tilde{\theta}', z')\right\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta}) \Bigg| \mathrm{d}\tilde{\theta}$$

$$
\begin{aligned}
&\leqslant \iint e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \varphi(\tilde{\theta}; 0, \Sigma) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \iint \tilde{\pi}_T^\omega(\tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \Bigg| \mathrm{d}\tilde{\theta} \qquad (4.32) \\
&\qquad + \iint \tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \\
&\qquad - \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') g(\mathrm{d}z' \mid \overline{\theta}) \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \Bigg| . \qquad (4.33)
\end{aligned}
$$

For the first part (4.32) we have

$$
\begin{aligned}
&\iint e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \varphi(\tilde{\theta}; 0, \Sigma) \tilde{q}(\tilde{\theta}, \tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \tilde{\pi}_T^\omega(\tilde{\theta}) \tilde{q}(\tilde{\theta}, \tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \mathrm{d}\tilde{\theta}' \Bigg| \mathrm{d}\tilde{\theta} \\
&\leqslant \|f\|_\infty \iiiint e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \big|\varphi(\tilde{\theta}; 0, \Sigma) - \tilde{\pi}_T^\omega(\tilde{\theta})\big| \, \mathrm{d}\tilde{\theta} \\
&= \|f\|_\infty \int \big|\varphi(\tilde{\theta}; 0, \Sigma) - \tilde{\pi}_T^\omega(\tilde{\theta})\big| \, \mathrm{d}\tilde{\theta} \xrightarrow{\mathbb{P}^Y} 0,
\end{aligned}
$$

again by Lemma 4.24. The second part (4.33)

$$
\begin{aligned}
&\iint \tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \Bigg| \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \\
&\qquad - \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') g(\mathrm{d}z' \mid \overline{\theta}) \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \Bigg| \\
&\leqslant \iiint \tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \Bigg| \int \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \\
&\qquad - \int \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} g(\mathrm{d}z' \mid \hat{\theta}_T^\omega + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \Bigg| \qquad (4.34) \\
&\qquad + \iiint \tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \Bigg| \int g(\mathrm{d}z' \mid \hat{\theta}_T^\omega + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z') \\
&\qquad - \int \tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} g(\mathrm{d}z' \mid \bar{\theta}) f(\tilde{\theta}', z') \Bigg| \qquad (4.35)
\end{aligned}
$$

We first consider (4.34) using $\theta = \hat{\theta}_T^\omega + \tilde{\theta}/\sqrt{T}$, and similarly for θ',

$$
\begin{aligned}
&\iiint \tilde{\pi}_T^\omega(\mathrm{d}\tilde{\theta}) e^z \tilde{g}_T^\omega(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \Bigg| \int \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} e^{z'-z}\right\} \tilde{g}_T^\omega(\mathrm{d}z' \mid \tilde{\theta}') f(\tilde{\theta}', z') \\
&\qquad - \int \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} e^{z'-z}\right\} g(\mathrm{d}z' \mid \hat{\theta}_T^\omega + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \Bigg|
\end{aligned}
$$

$$
\begin{aligned}
&= \iiint \pi_T^\omega(\mathrm{d}\theta) e^z g_T^\omega(\mathrm{d}z \mid \theta) q_T(\theta, \mathrm{d}\theta') \\
&\quad \times \Bigg| \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^\omega, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'-z}\right\} g_T^\omega(\mathrm{d}z' \mid \theta') f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), z'\right\} \\
&\qquad - \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^\omega, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'-z}\right\} g(\mathrm{d}z' \mid \theta') f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), z'\right\} \Bigg|
\end{aligned}
$$

In the rest of the proof, without loss of generality, we will consider f such that $\|f\|_{\mathrm{L}} \leqslant 1$

$$
\begin{aligned}
&\left| f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), x\right\} - f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), y\right\} \right| \\
&\leqslant d\left[\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), x\right\}, \left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), y\right\}\right] = |x - y|
\end{aligned}
$$

and thus $x \mapsto f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), x\right\}$ is Lipschitz with coefficient 1 uniformly in T. Moreover, due to Lemma 4.26, the map

$$
z' \mapsto \min\left\{1, e^{-z} \frac{\varphi(\theta'; \hat{\theta}_T^\omega, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'}\right\}
$$

is Lipschitz with Lipschitz constant 1 uniformly for all θ, θ', z and T. Thus, using the triangle inequality, we can write

$$
\begin{aligned}
&\iiint \pi_T^\omega(\mathrm{d}\theta) e^z g_T^\omega(\mathrm{d}z \mid \theta) q_T(\theta, \mathrm{d}\theta') \Bigg| \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^\omega, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)} e^{z'-z}\right\} f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), z'\right\} g_T^\omega(z' \mid \theta') \\
&\qquad - \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^\omega, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^\omega, \Sigma/T)} e^{z'-z}\right\} f\left\{\sqrt{T}(\theta' - \hat{\theta}_T^\omega), z'\right\} g(z' \mid \theta') \Big| \mathrm{d}z' \\
&\leqslant 2 \iint \pi_T^\omega(\mathrm{d}\theta) q_T(\theta, \mathrm{d}\theta') \cdot \sup_{f \in \mathrm{BL}(\mathbb{R}),\, \|f\|_{\mathrm{BL}} \leqslant 1} \left| \int f(z') g_T^\omega(\mathrm{d}z' \mid \theta') - \int f(z') g(\mathrm{d}z' \mid \theta') \right| \mathrm{d}\theta \\
&= 2 \iint \pi_T^\omega(\mathrm{d}\theta) q_T(\theta, \mathrm{d}\theta') d_{\mathrm{BL}}\left\{g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right\} \\
&= 2 \int_{B(\bar{\theta})} \pi_T^\omega q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\left\{g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right\} + 2 \int_{B(\bar{\theta})^{\complement}} \pi_T^\omega q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\left(g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right),
\end{aligned}
$$

where $B(\bar{\theta})$ is given in Assumption 4.3. Since the bounded Lipschitz norm metrizes weak convergence (for non-random probability measures) we know that for $\theta' \in B(\bar{\theta})$

$$
d_{\mathrm{BL}}\left(g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right) = \sup_{f \in \mathrm{BL}(\mathbb{R}),\, \|f\|_{\mathrm{BL}} \leqslant 1} \left| \int f(z') g_T^\omega(\mathrm{d}z' \mid \theta') - \int f(z') g(\mathrm{d}z' \mid \theta') \right|
$$

vanishes in $\mathbb{P}^Y$-probability by Assumption 4.3. From Lemma 4.25 we know that the marginal distribution of the proposal at stationarity $\pi_T^\omega q_T(\mathrm{d}\theta') = \int \pi_T^\omega(\mathrm{d}\theta) q(\theta, \mathrm{d}\theta')$ concentrates around the true parameter value. Since the bounded Lipschitz metric cannot exceed 1 we have

$$\int \pi_T^\omega q_T(\mathrm{d}\theta') \mathbb{1}_{B(\bar\theta)^\complement}(\theta') d_{\mathrm{BL}}\left(g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right) \leqslant \pi_T^\omega q_T\left\{B(\bar\theta)^\complement\right\} \xrightarrow{\mathbb{P}^Y} \delta_{\bar\theta}\left\{B(\bar\theta)^\complement\right\} = 0.$$

In addition from Assumption 4.3

$$\left|\int_{B(\bar\theta)} \pi_T^\omega q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\left\{g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right\}\right| \leqslant \sup_{\theta\in B(\bar\theta)} \left|d_{\mathrm{BL}}\left\{g_T^\omega(\cdot|\theta), g(\cdot|\theta)\right\}\right| \xrightarrow{\mathbb{P}^Y} 0.$$

Finally, using a similar argument for (4.35) we have

$$\begin{aligned}
&\iiint \tilde\pi_T^\omega(\mathrm{d}\tilde\theta) e^z \tilde g_T^\omega(\mathrm{d}z \mid \tilde\theta) \tilde q(\tilde\theta, \mathrm{d}\tilde\theta') \left| \int g(z' \mid \hat\theta_T^\omega + \tilde\theta'/\sqrt{T}) f(\tilde\theta', z') \tilde\alpha\left\{(\tilde\theta, z), (\tilde\theta', z')\right\} f(\tilde\theta', z') \right. \\
&\qquad \left. - \tilde q(\tilde\theta, \mathrm{d}\tilde\theta') \tilde\alpha\left\{(\tilde\theta, z), (\tilde\theta', z')\right\} g(z' \mid \bar\theta) f(\tilde\theta', z') \right| \mathrm{d}z' \\
&\leqslant 2 \iint \pi_T^\omega(\theta) q_T(\theta, \theta') \mathrm{d}\theta d_{\mathrm{BL}}\left(g(\cdot|\theta'), g(\cdot|\bar\theta)\right) \mathrm{d}\theta'. \qquad (4.36)
\end{aligned}$$

By Lemma 4.27 the bounded Lipschitz metric, $d_{\mathrm{BL}}\left\{g(\cdot|\theta'), g(\cdot|\bar\theta)\right\}$, is bounded and continuous at $\bar\theta$. Thus (4.36) converges to zero by Lemma 4.25. □

4.30 Proposition. *Under Assumption 4.2, the map $(\theta, z) \mapsto \tilde P f(\theta, z)$ is continuous for every $f \in C_b(\mathbb{R}^{d+1})$.*

Proof of Proposition 4.8. Without loss of generality let $\|f\|_\infty \leqslant 1$, consider $(\theta^*, z^*) \in \Theta \times \mathbb{R}$ and denote $(\theta_n, z_n)_{n\in\mathbb{N}}$ a sequence converging to (θ^*, z^*) as $n \to \infty$. Using the decomposition (4.28) we have

$$\begin{aligned}
&\left|\tilde P f(\theta_n, z_n) - \tilde P f(\theta^*, z^*)\right| \\
&= \left|\Pi f(\theta_n, z_n) + f(\theta_n, z_n)\left\{1 - \Pi 1(\theta_n, z_n)\right\} - \Pi f(\theta^*, z^*) - f(\theta^*, z^*)\left\{1 - \Pi 1(\theta^*, z^*)\right\}\right| \\
&\leqslant \left|\Pi f(\theta_n, z_n) - \Pi f(\theta^*, z^*)\right| + \left|f(\theta_n, z_n) - f(\theta^*, z^*)\right| + \left|\Pi 1(\theta_n, z_n) - \Pi 1(\theta^*, z^*)\right|
\end{aligned}$$

By continuity of f we have $f(\theta_n, z_n) \to f(\theta^*, z^*)$ as $n \to \infty$. Since $1 \in C_b(\mathbb{R}^{d+1})$ it remains to show that Πf is continuous for every $f \in C_b(\mathbb{R}^{d+1})$. Now

$$\begin{aligned}
&\left|\Pi f(\theta_n, z_n) - \Pi f(\theta^*, z^*)\right| \\
&= \left| \int f(\theta', z') \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z' - z_n}\right\} \nu(\theta' - \theta_n) g(\mathrm{d}z' \mid \overline{\theta}) \mathrm{d}\theta' \right. \qquad (4.37)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \int f(\theta', z') \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z'-z^*}\right\} \nu(\theta' - \theta^*) g(\mathrm{d}z' \mid \overline{\theta}) \mathrm{d}\theta' \Big| \\
&\leqslant \int \left|\nu(\theta' - \theta_n) - \nu(\theta' - \theta^*)\right| \mathrm{d}\theta' \qquad (4.38) \\
&\quad + \int \Bigg| \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z'-z_n}\right\} \\
&\quad - \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z'-z^*}\right\} \Bigg| \nu(\theta' - \theta^*) g(\mathrm{d}z' \mid \overline{\theta}) \mathrm{d}\theta'. \qquad (4.39)
\end{aligned}
$$

For (4.38), Assumption 4.2 implies $\nu(\theta' - \theta_n) \to \nu(\theta' - \theta^*)$ as $n \to \infty$ and hence Scheffé's lemma yields

$$
\int \left|\nu(\theta' - \theta_n) - \nu(\theta' - \theta^*)\right| \mathrm{d}\theta' \to 0.
$$

For (4.39), the map

$$
(\theta, z) \mapsto \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta; 0, \Sigma)} \frac{\nu(\theta - \theta')}{\nu(\theta' - \theta)} e^{z'-z}\right\}
$$

is continuous for all θ', z' since it is just a composition of continuous functions. Hence,

$$
\left|\min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z'-z_n}\right\} - \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z'-z^*}\right\}\right| \to 0
$$

for every (θ', z') and an application of dominated convergence shows that (4.39) goes to zero. □

4.10 Proofs of Section 4.5

4.10.1 Central Limit Theorem for Likelihood Estimators

We detail here the proof of Theorem 4.9. For clarity we explicitly state the probability space supporting all random variables that are used to prove our limit theorem. For integers N, T, k we introduce the space $E_T = \Theta \times \mathbb{R}^{NTk}$ where $\Theta \subset \mathbb{R}^d$ is the parameter space equipped with the Borel σ-algebra and probability measure $\mathbb{P}_T(\mathrm{d}\theta, \mathrm{d}u) = \pi_T^{\omega}(\mathrm{d}\theta) m_{T,\theta}(\mathrm{d}u)$. Finally, we will work with the Borel probability measure $\mathbb{P}$ on E where $E = \mathsf{Y}^{\mathbb{N}} \times \prod_{T=1}^{\infty} E_T$, $\mathbb{P} = \mathbb{P}^Y \otimes \bigotimes_{T=1}^{\infty} \mathbb{P}_T$.

We are interested in the asymptotic distribution of the relative error of the log-likelihood

$$
Z_T(\theta) = \log \widehat{p}(Y_{1:T} \mid \theta, U) - \log p(Y_{1:T} \mid \theta),
$$

where $U \sim m_{T,\theta}(\cdot)$ or $U \sim \pi_T^{\omega}(\cdot \mid \theta)$. Indeed, we have $\mathcal{L}\text{aw}\left\{Z_T(\theta)\right\} = g_T^{\omega}(\cdot \mid \theta)$ when $U \sim m_{T,\theta}(\cdot)$ and $\mathcal{L}\text{aw}\left\{Z_T(\theta)\right\} = \bar{g}_T^{\omega}(\cdot \mid \theta)$ when $U \sim \pi_T^{\omega}(\cdot \mid \theta)$. Weak convergence results for $Z_T(\theta)$ have been established in Deligiannidis, Doucet and Pitt (2018, Theorem 1) using a Taylor expansion. However, the CLTs introduced therein do not provide a bound on the Lipschitz metric d_{BL} and are not uniform in the parameter θ as required in Assumption 4.3. In order to obtain a uniform bound for all functions in $\mathrm{BL}(\mathbb{R})$ with $\|f\|_{\mathrm{BL}} \leqslant 1$ and all parameter values for some neighbourhood $B(\bar{\theta})$ we need to introduce further assumptions. We follow the approach in Deligiannidis, Doucet and Pitt (2018) and write

$$\begin{aligned} Z_T(\theta) &= \sum_{t=1}^{T} \log\left\{1 + \frac{\widehat{p}(Y_t \mid \theta, U_t) - p(Y_t \mid \theta)}{p(Y_t \mid \theta)}\right\} \\ &= \sum_{t=1}^{T} \log\left\{1 + \frac{\epsilon_N(Y_t, \theta)}{\sqrt{N}}\right\} \end{aligned}$$

where

$$\epsilon_N(Y_t, \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{\overline{w}(Y_t, U_{t,i}, \theta) - 1\right\},$$

$\overline{w}(Y_t, U_{t,i}, \theta)$ being a normalized importance weight defined in (4.10). Recall that

$$\sigma^2(y, \theta) = E\left\{\epsilon_T(y, \theta)^2\right\} = \mathrm{var}\left\{\overline{w}(y, U_{1,1}, \theta)\right\}, \quad \sigma^2(\theta) = E\left\{\sigma^2(Y_1, \theta)\right\}.$$

Here the number of particles, N, is scaled proportionally to the number of observations, that is $N = \lceil \gamma T \rceil$ for some $\gamma > 0$. In the following we will take $\gamma = 1$ (that is $N = T$) for simplicity and without loss of generality. In order to show convergence of the bounded Lipschitz metric uniformly in θ, we will exploit the relation

$$\log(1 + x) = x - \frac{x^2}{2} + \int_0^x \frac{u^2}{1 + u} \mathrm{d}u,$$

where for $x < 0$ we use the convention

$$\int_0^x \frac{u^2}{1 + u} \mathrm{d}u = -\int_x^0 \frac{u^2}{1 + u} \mathrm{d}u.$$

We thus obtain

$$Z_T(\theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \epsilon_T(Y_t, \theta) - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_T(Y_t, \theta)^2 + \sum_{t=1}^{T} R_T(Y_t, \theta), \tag{4.40}$$

with

$$R_T(y,\theta) = \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u}\mathrm{d}u. \tag{4.41}$$

We recall the following assumptions regarding the normalized weights.

***Assumption* 4.4.** *There exists a closed ε-ball $B(\bar{\theta})$ around $\bar{\theta}$ and a function g such that the normalized weight $\overline{w}(y, U_{1,1}, \theta)$ defined in (4.10) satisfies for some $\Delta > 0$*

$$\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U_{1,1}, \theta)^{2+\Delta}\right\} \leqslant g(y),$$

where $U_{1,1} \sim h(\cdot \mid y, \theta)$ and $\mu(g) < \infty$. Additionally, $\theta \mapsto \sigma^2(y,\theta)$ is continuous in θ on $B(\bar{\theta})$ for all $y \in \mathsf{Y}$.

We can relate expectations of powers of $\epsilon_T(y,\theta)$ to that of $\overline{w}(y, U_{1,1}, \theta)$ in the following way.

4.31 Lemma. *For any $k \geqslant 2$ and any $T \geqslant 1$*

$$E\left\{\left|\epsilon_T(y,\theta)\right|^k\right\} \leqslant c(k)\left[E\left\{\overline{w}(y, U_{1,1}, \theta))^k\right\} + 1\right]$$

where $c(k)$ is a constant only depending on k.

Proof. This is Lemma 2 in Deligiannidis, Doucet and Pitt (2018). We repeat it here for convenience. It holds

$$\begin{aligned}
E\left\{\left|\epsilon_T(y,\theta)\right|^k\right\} &= E\left[\left|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\left\{\overline{w}(y, U_{1,i}, \theta) - 1\right\}\right|^k\right] \\
&\leqslant c_1(k) E\left[\left|\frac{1}{T}\sum_{i=1}^{T}\left\{\overline{w}(y, U_{1,i}, \theta) - 1\right\}^2\right|^{k/2}\right] \\
&\leqslant c_1(k)\frac{1}{T}\sum_{i=1}^{T} E\left\{\left|\overline{w}(y, U_{1,i}, \theta) - 1\right|^k\right\} \\
&\leqslant c_1(k)c_2(k)\left[E\left\{\overline{w}(y, U_{1,1}, \theta))^k\right\} + 1\right]
\end{aligned}$$

for some constants $c_1(k), c_2(k)$ by application of the Marcinkiewicz–Zygmund, Jensen and c_r-inequalities. □

As a result we have thus

$$\sup_{\theta\in B(\bar{\theta})} E\left\{\left|\epsilon_T(y,\theta)\right|^k\right\} \leqslant c(k) \sup_{\theta\in B(\bar{\theta})} \left[E\left\{\overline{w}(y,U_{1,1},\theta))^k\right\}+1\right] \tag{4.42}$$

and the left-hand-side is finite whenever the right-hand-side is finite.

4.10.2 Moment Conditions for Weak Convergence

Denote $\mathcal{Y}_T$ the $\sigma-$algebra spanned by the data $Y_{1:T} = (Y_1, \dots, Y_T)$ observed up to T.

4.32 Theorem (Moment conditions for UCLT)**.** *Under Assumption 4.4 we have the following uniform central limit theorems*

a)

$$\sup_{\theta\in B(\bar{\theta})} d_{\mathrm{BL}}\left[g_T^{\omega}(\cdot \mid \theta), \varphi\left\{\cdot; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\} \mid \mathcal{Y}_T\right] \xrightarrow{\mathbb{P}} 0,$$

and

b)

$$\sup_{\theta\in B(\bar{\theta})} d_{\mathrm{BL}}\left[\bar{g}_T^{\omega}(\cdot \mid \theta), \varphi\left\{\cdot; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} \mid \mathcal{Y}_T\right] \xrightarrow{\mathbb{P}} 0.$$

We will need the following auxiliary results.

4.33 Lemma. *Let $S_T(\theta) = \sum_{i=1}^T \xi_i(\theta)$ denote the sum of zero mean independent random variables $\xi_1(\theta), \dots, \xi_T(\theta)$ such that $\mathrm{var}(S_T) = 1$. Then for any Lipschitz function f with Lipschitz constant L and $Z \sim \mathcal{N}(0,1)$*

$$\left|E\left[f\left\{S_T(\theta)\right\} - f\left(Z\right)\right]\right| \leqslant L\left(4E\left[\sum_{i=1}^T \xi_i^2(\theta) 1_{\{|\xi_i(\theta)|>1\}}\right] + 3E\left[\sum_{i=1}^T |\xi_i(\theta)|^3 1_{\{|\xi_i(\theta)|\leqslant 1\}}\right]\right).$$

Proof. This is Theorem 3.2 in Chen, Goldstein and Shao (2010). □

The above result reduces the problem of showing weak convergence uniformly over some neighbourhood $B(\bar{\theta})$ to uniform laws of large numbers for conditional higher order moments. Conditions to ensure uniformity in the convergence of averages are widely established. We will use the following result given in (Jennrich 1969, Theorem 2).

4.34 Lemma. *Let $A \subset \mathbb{R}^d$ be compact and let $f : \mathbb{R}^k \times A \to \mathbb{R}$ be continuous in θ for each $y \in \mathbb{R}^k$ and measurable in y for each $\theta \in A$. Further assume that there exists an integrable*

function g, such that $|f(y,\theta)| \leqslant g(y)$ for all y and θ. For independent random variables $Y_i \sim \mu$ $(i = 1, \ldots, T)$ then $\mathbb{P}^Y$-almost surely

$$\sup_{\theta \in A} \left| \frac{1}{T} \sum_{t=1}^{T} f\left(Y_t, \theta\right) - E\left\{ f(Y_1, \theta) \right\} \right| \to 0,$$

as $T \to \infty$.

Before we proceed with the proof of Theorem 4.32, we note that Lemma 4.33 is not formulated in terms of conditional laws. However, considering conditionally (upon $\mathcal{Y}_T$) centred and independent random variables $\xi_{T,1}, \ldots, \xi_{T,T}$ such that $\sum_{i=1}^{T} \operatorname{var}\left\{ \xi_i(\theta) | Y_{1:T} \right\} = 1$, we can apply the above lemma for every realization $Y_{1:T} = y_{1:T}$. Denote P_T^y a regular conditional distribution associated with the law of $S_T = \xi_{T,1} + \ldots + \xi_{T,T}$ given $Y_{1:T} = y_{1:T}$. By applying Lemma 4.33, we get

$$\begin{aligned} &d_{\mathrm{BL}}\left\{ P_T^y, \varphi(\cdot\,; 0, 1) \mid Y_{1:T} = y_{1:T} \right\} \\ &\leqslant 4E\left[\sum_{i=1}^{T} \xi_i^2(\theta) 1_{\{|\xi_i(\theta)|>1\}} \mid Y_{1:T} = y_{1:T} \right] + 3E\left[\sum_{i=1}^{T} |\xi_i(\theta)|^3 1_{\{|\xi_i(\theta)|\leqslant 1\}} \mid Y_{1:T} = y_{1:T} \right]. \end{aligned} \tag{4.43}$$

Thus, if the terms on the r.h.s. go to zero in $\mathbb{P}^Y$-probability then $d_{\mathrm{BL}}\left\{ P_T^Y, \varphi(\cdot\,; 0, 1) \right\} \xrightarrow{\mathbb{P}^Y} 0$. With this reasoning we can apply Lemma 4.33 to prove Theorem 4.32.

Proof of Theorem 4.32, part a). Define

$$\xi_{T,t}(\theta) = \frac{\epsilon_T(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}, \quad S_T(\theta) = \sum_{t=1}^{T} \xi_{T,t}(\theta),$$

where

$$\sigma_T^2(Y_{1:T}, \theta) = \frac{1}{T} \sum_{t=1}^{T} \operatorname{var}\left\{ \epsilon_{T,t}(\theta) \mid \mathcal{Y}_T \right\}. \tag{4.44}$$

Thus

$$\operatorname{var}\left\{ S_T(\theta) \mid \mathcal{Y}_T \right\} = \sum_{t=1}^{T} \operatorname{var}\left\{ \xi_{T,t}(\theta) \right\} = 1.$$

In the following we will use the shorthand $\sigma_T(Y_{1:T}, \theta) = \left\{ \sigma_T^2(Y_{1:T}, \theta) \right\}^{1/2}$ and $\sigma_T^r(Y_{1:T}, \theta) = \left\{ \sigma_T^2(Y_{1:T}, \theta) \right\}^{r/2}$ for any real value r.

Then $S_T(\theta)$ fulfils the conditions of Lemma 4.33 conditionally on $\mathcal{Y}_T$. The random variable $Z_T(\theta)$ defined in (4.40) can be rewritten as

$$Z_T(\theta) = S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T}R_T(Y_t,\theta).$$

We have for $Z \sim \mathcal{N}(0,1)$

$$\begin{aligned}
&\sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\left[\mathcal{L}aw\left\{Z_T(\theta)\right\}, \varphi\left\{\cdot; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\} \mid \mathcal{Y}_T\right] \\
&= \sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\left[\mathcal{L}aw\left\{Z_T(\theta)\right\}, \mathcal{L}aw\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] \\
&= \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T}R_T(Y_t,\theta)\right\} \mid \mathcal{Y}_T\right]\right. \\
&\qquad \left. - E\left[f\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\}\right]\right| \\
&\leqslant \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T}R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} + \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right. \qquad (4.45) \\
&\qquad \left. - E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T}R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right| \\
&\quad + \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T}R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right. \\
&\qquad \left. - E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right| \qquad (4.46) \\
&\quad + \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] - E\left[f\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\}\right]\right| \qquad (4.47)
\end{aligned}$$

Now we have for (4.45)

$$(4.45) \leqslant \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T}R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} + \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right.$$

$$
\begin{aligned}
&-E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)+\sum_{t=1}^{T}R_T(Y_t,\theta)-\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]\Bigg| \\
&\leqslant \sup_{\theta\in B(\bar{\theta})} E\left[\min\left\{1,\left|\frac{\sigma^2(\theta)}{2}-\frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2\right|\right\}\mid \mathcal{Y}_T\right]
\end{aligned}
$$

where we use that f is bounded and Lipschitz. We can bound this term by

$$
\begin{aligned}
&\sup_{\theta\in B(\bar{\theta})} E\left(\min\left\{1,\left|\frac{\sigma^2(\theta)}{2}-\frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2\right|\right\}\Bigg|\mathcal{Y}_T\right) \\
&\leqslant \sup_{\theta\in B(\bar{\theta})} E\left(\min\left\{1,\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right|\right\}\Bigg|\mathcal{Y}_T\right) \\
&\qquad + \sup_{\theta\in B(\bar{\theta})} E\left(\min\left\{1,\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\epsilon_T(Y_t,\theta)^2-\sigma^2(Y_t,\theta)\right\}\right|\right\}\Bigg|\mathcal{Y}_T\right).
\end{aligned} \tag{4.48}
$$

For any $0<\delta<1$, we can bound the first term on the r.h.s. of (4.48) by

$$
\begin{aligned}
&\sup_{\theta\in B(\bar{\theta})} E\left[\min\left\{1,\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\epsilon_T(Y_t,\theta)^2-\sigma^2(Y_t,\theta)\right\}\right|\right\}\Bigg|\mathcal{Y}_T\right] \\
&\leqslant \sup_{\theta\in B(\bar{\theta})}\left(E\left[\min\left\{1,\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\epsilon_T(Y_t,\theta)^2-\sigma^2(Y_t,\theta)\right\}\right|^{1+\delta}\right\}\Bigg|\mathcal{Y}_T\right]\right)^{\frac{1}{1+\delta}} \\
&\leqslant \left[\frac{C}{2^{1+\delta}T^{1+\delta}}\sum_{t=1}^{T}\sup_{\theta\in B(\bar{\theta})} E\left\{\left|\epsilon_T(Y_t,\theta)^2-\sigma^2(Y_t,\theta)\right|^{1+\delta}\Bigg|\mathcal{Y}_T\right\}\right]^{\frac{1}{1+\delta}} \\
&\leqslant C\left[\frac{C'}{2^{1+\delta}T^{1+\delta}}\sum_{t=1}^{T}\left\{1+g(Y_t)\right\}\right]^{\frac{1}{1+\delta}}\to 0
\end{aligned}
$$

in $\mathbb{P}^Y$-probability by the law of large numbers using, in turn, Jensen's inequality, von Bahr–Esseen inequality (Bahr and Esseen 1965) as $E\left\{\epsilon_T(Y_t,\theta)^2\mid\mathcal{Y}_T\right\}=\sigma^2(Y_t,\theta)$, c_r-inequality, (4.42) and Assumption 4.4 for $\Delta=2\delta$, noting that

$$
\begin{aligned}
\sigma^2(Y_t,\theta)=E\left\{\epsilon_T(Y_t,\theta)^2\mid\mathcal{Y}_T\right\}&\leqslant E\left\{|\epsilon_T(Y_t,\theta)|^{2+\Delta}\right\}^{2/(2+\Delta)} \\
&\leqslant C\cdot\left\{g(Y_t)+1\right\}^{2/(2+\Delta)},
\end{aligned}
$$

where the last inequality is due to (4.42). The second term on the right-hand side of (4.48)

can be bounded

$$\sup_{\theta\in B(\bar{\theta})} E\left(\min\left\{1, \left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right|\right\} \mid \mathcal{Y}_T\right)$$
$$\leqslant E\left(\min\left\{1, \sup_{\theta\in B(\bar{\theta})}\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right|\right\} \mid \mathcal{Y}_T\right).$$

Noting that $\sigma^2(y,\theta)$ is continuous in θ for all y by Assumption 4.4 and $\sigma^2(y,\theta) \leqslant C \cdot \{1+g(y)\}^{2/(2+\Delta)}$ we can apply Lemma 4.34 to get

$$\sup_{\theta\in B(\bar{\theta})}\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right| \xrightarrow{\mathbb{P}^Y} 0$$

and we can use dominated convergence to conclude that

$$E\left(E\left[\min\left\{1, \sup_{\theta\in B(\bar{\theta})}\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right|\right\} \mid \mathcal{Y}_T\right]\right) \to 0$$

and thus

$$E\left[\min\left\{1, \sup_{\theta\in B(\bar{\theta})}\left|\frac{1}{2T}\sum_{t=1}^{T}\left\{\sigma^2(Y_t,\theta)-\sigma^2(\theta)\right\}\right|\right\} \mid \mathcal{Y}_T\right] \xrightarrow{\mathbb{P}^Y} 0.$$

The quantity (4.46) can be upper bounded by

$$(4.46) \leqslant E\left[\min\left\{1, \left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right] \leqslant \sum_{t=1}^{T} E\left[\min\left\{1, \left|R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right]. \tag{4.49}$$

We will split the expectation into two terms

$$E\left[\min\left\{1, \left|R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right]$$
$$= E\left[\min\left\{1, \left|R_T(Y_t,\theta)\right|\right\} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\leqslant 1\right\}} \mid \mathcal{Y}_T\right] + E\left[\min\left\{1, \left|R_T(Y_t,\theta)\right|\right\} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|> 1\right\}} \mid \mathcal{Y}_T\right]. \tag{4.50}$$

Recall

$$R_T(y,\theta) = \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u}\mathrm{d}u.$$

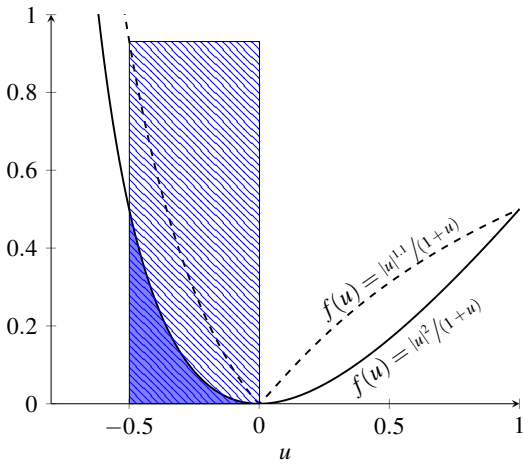


Figure 4.2: For $x \in (-1, 1]$, $x = -0.5$ on the graph, the remainder of our expansion is estimated by the integral under the solid curve (blue shaded area). We bound this integral first by dashed line and then we approximate the integral by the box containing the area (lines).

We investigate the integral

$$\Psi(x) = \int_0^x \frac{u^2}{1+u} \mathrm{d}u \tag{4.51}$$

in more detail (see also Figure 4.2), where in the case $x < 0$, we interpret the above as an integral over the interval $[x, 0]$. Without loss of generality, we can always select $0 < \Delta < 1$ in Assumption 4.4. On the interval $(-1, 1]$ we can bound the function

$$\frac{u^2}{1+u} \leqslant \frac{|u|^{1+\Delta}}{1+u},$$

as $0 < \Delta < 1$ where we show $\Delta = 0.1$ as an example in Figure 4.2. Subsequently, we bound for $x \in (-1, 1]$

$$\left| \int_0^x \frac{u^2}{1+u} \mathrm{d}u \right| \leqslant \left| \int_0^x \frac{|u|^{1+\Delta}}{1+u} \mathrm{d}u \right| \leqslant \left| x \cdot \frac{|x|^{1+\Delta}}{1+x} \right|,$$

i.e. the box containing the area under the curve. This is visualized in Figure 4.2. The integral (shaded blue) is bounded by the striped box. Hence, on the set $|\epsilon_T(y, \theta)/\sqrt{T}| \leqslant 1$

$$\left| \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u} \mathrm{d}u \right| \leqslant \left| \frac{|\epsilon_T(y, \theta)|^{2+\Delta}}{T^{1+\Delta/2}} \frac{1}{1+\epsilon_T(y, \theta)/\sqrt{T}} \right|.$$

For any non-negative random variable X and event A, we have the identity

$$E\left\{\min(1, X) 1_A\right\} \leqslant E\left\{X 1_{X \leqslant 1} 1_A\right\} + \mathbb{P}(X > 1),$$

so we can bound the first term on the right-hand side of (4.50) for every $t = 1, \dots, T$

$$
\begin{aligned}
& E\left[\min\left\{1, \left|R_T(Y_t,\theta)\right|\right\} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\leqslant 1\right\}} \mid \mathcal{Y}_T\right] \\
& \leqslant E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right|\leqslant 1\right\}} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\leqslant 1\right\}} \middle| \mathcal{Y}_T\right] \\
& \quad + \mathbb{P}\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| > 1 \middle| \mathcal{Y}_T\right\}.
\end{aligned}
$$

By inspection of the function, similarly to before,

$$
u \mapsto \frac{|u|^{2+\Delta}}{1+u}
$$

one can easily verify that there exist $0 < \delta_1 < 1$ and $\delta_2 > 0$ such that

$$
\frac{|u|^{2+\Delta}}{1+u} \leqslant 1 \Leftrightarrow -\delta_1 \leqslant u \leqslant \delta_2.
$$

Thus we have

$$
\begin{aligned}
& E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\left\{\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\leqslant 1\right\}} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\leqslant 1\right\}} \middle| \mathcal{Y}_T\right] \\
& \leqslant E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\{-\delta_1\leqslant \epsilon_T(Y_t,\theta)/\sqrt{T}\leqslant \delta_2\}} \middle| \mathcal{Y}_T\right] \\
& \leqslant \frac{1}{(1-\delta_1)T^{1+\Delta/2}} E\left[\left|\epsilon_T(Y_t,\theta)\right|^{2+\Delta} \middle| \mathcal{Y}_T\right], \qquad (4.52)
\end{aligned}
$$

while

$$
\begin{aligned}
& \mathbb{P}\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| > 1 \middle| \mathcal{Y}_T\right\} \\
& \leqslant \mathbb{P}\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{T^{1/2}}\right| > \min\{\delta_1,\delta_2\} \middle| \mathcal{Y}_T\right\} \\
& \leqslant \frac{1}{\min\{\delta_1,\delta_2\}^{2+\Delta}T^{1+\Delta/2}} E\left[\left|\epsilon_T(Y_t,\theta)\right|^{2+\Delta} \middle| \mathcal{Y}_T\right]. \qquad (4.53)
\end{aligned}
$$

The second term on the right-hand side of (4.50) is bounded by

$$E\left[\min\left\{1,\left|R_T(Y_t,\theta)\right|\right\}1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|>1\right\}}\mid \mathcal{Y}_T\right]$$
$$\leqslant E\left[\left|R_T(Y_t,\theta)\right|1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|>1\right\}}\mid \mathcal{Y}_T\right]. \tag{4.54}$$

As $\epsilon_T(Y_t,\theta)/\sqrt{T}\geqslant -1$, (4.54) is null for $\epsilon_T(Y_t,\theta)<-1$ so writing $X^+=\max\{0,X\}$ this can be rewritten as

$$E\left\{\int_0^{\epsilon_T(Y_t,\theta)/\sqrt{T}}\frac{u^2}{1+u}\mathrm{d}u1_{\{\epsilon_T(Y_t,\theta)/\sqrt{T}\geqslant 1\}}\Big|\mathcal{Y}_T\right\}$$
$$\leqslant E\left[\int_0^{(\epsilon_T(Y_t,\theta)/\sqrt{T})^+}\frac{u^2}{1+u}\mathrm{d}u\Big|\mathcal{Y}_T\right]$$
$$=\int_0^\infty\frac{u^2}{1+u}\mathbb{P}\left\{\epsilon_T(Y_t,\theta)^+>\sqrt{T}u\Big|\mathcal{Y}_T\right\}\mathrm{d}u,$$

where we have used that for the function (4.51) is increasing and differentiable on its domain so

$$E\left\{\Psi\left(|X|\right)\right\}=\Psi(0)+\int_0^\infty\Psi'(u)P(|X|>u)\mathrm{d}u.$$

For $\Delta\in(0,1)$, we bound the remainder using

$$=\int_0^\infty\frac{u^2}{1+u}\mathbb{P}\left\{\epsilon_T(Y_t,\theta)^+>\sqrt{T}u\Big|\mathcal{Y}_T\right\}\mathrm{d}u$$
$$\leqslant\int_0^\infty\frac{u^2}{1+u}\frac{E\left\{\left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta}\Big|\mathcal{Y}_T\right\}}{T^{(2+\Delta)/2}u^{2+\Delta}}\mathrm{d}u$$
$$=\int_0^\infty\frac{1}{(1+u)\,u^\Delta}\mathrm{d}u\frac{1}{T^{1+\Delta/2}}E\left\{\left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta}\Big|\mathcal{Y}_T\right\}$$
$$=C(\Delta)\frac{E\left\{\left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta}\Big|\mathcal{Y}_T\right\}}{T^{1+\Delta/2}} \tag{4.55}$$

noting that

$$\int_0^\infty\frac{1}{(1+u)\,u^\Delta}du=C(\Delta)<\infty$$

for $\Delta\in(0,1)$. Hence we can bound (4.50) by the sum of (4.52), (4.53) and (4.55) so, by

using (4.49), we obtain a bound for (4.46)

$$
\begin{aligned}
(4.46) \leqslant & \frac{1}{(1-\delta_1)T^{(1+\Delta)/2}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ \left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta} \middle| \mathcal{Y}_T \right\} \\
& + \frac{1}{\min\{\delta_1,\delta_2\}^{2+\Delta} T^{(1+\Delta)/2}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ \left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta} \middle| \mathcal{Y}_T \right\} \\
& + C(\Delta) \frac{1}{T^{1+\Delta/2}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ \left|\varepsilon_T(Y_t,\theta)\right|^{2+\Delta} \middle| \mathcal{Y}_T \right\} \to 0
\end{aligned}
$$

which all converge in $\mathbb{P}^Y$-probability by (4.42), Assumption 4.4 and the law of large numbers.

We are now going to bound (4.47). We will use the fact that any constant c and any two random variables X_1, X_2 we have for $c > 0$

$$
\begin{aligned}
\sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant 1}} \left| E\left[f(cX_1) - f(cX_2)\right] \right| &\leqslant \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant c}} \left| E\left[f(X_1) - f(X_2)\right] \right| \\
&= \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant c}} \left| E\left[c \left\{ \frac{f(X_1)}{c} - \frac{f(X_2)}{c} \right\} \right] \right| \\
&\leqslant c \cdot \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \leqslant 1}} \left| E\left[\left\{ f(X_1) - f(X_2) \right\} \right] \right|.
\end{aligned}
$$

Note that we only require $\|f\|_{\mathrm{L}} \leqslant 1$ ($\|f\|_{\mathrm{L}}$ denoting the Lipschitz constant) in the last line alleviating the bound on the supremum $\|f\|_\infty$. The aim of the following paragraphs is to apply the above inequality and Lemma 4.33 to find a bound on (4.47). Omitting for the moment the supremum over the set $B(\bar{\theta})$ we compute for (4.47)

$$
\begin{aligned}
& \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant 1}} \left| E\left[f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) \right\} \mid \mathcal{Y}_T \right] - E\left[f\left\{ Z\sigma(\theta) \right\} \mid \mathcal{Y}_T \right] \right| \\
& \leqslant \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant 1}} \left| E\left[f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) \right\} \mid \mathcal{Y}_T \right] - E\left[f\left\{ Z\sigma_T(Y_{1:T},\theta) \right\} \mid \mathcal{Y}_T \right] \right| \\
& \quad + \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leqslant 1}} \left| E\left[f\left\{ Z\sigma_T(Y_{1:T},\theta) \right\} \mid \mathcal{Y}_T \right] - E\left[f\left\{ Z\sigma(\theta) \right\} \right] \right| \\
& \leqslant \sigma_T(Y_{1:T},\theta) \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \leqslant 1}} \left| E\left[f\left\{ S_T(\theta) \right\} \mid \mathcal{Y}_T \right] - E\left[f\left\{ Z \right\} \right] \right| + E\left[|Z| \right] \left| \sigma_T(Y_{1:T},\theta) - \sigma(\theta) \right|
\end{aligned}
$$

$$\leqslant \sigma_T(Y_{1:T},\theta) \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}}\leqslant 1}} \left| E\left[f\left\{S_T(\theta)\right\} \mid \mathcal{Y}_T\right] - E\left[f\left\{Z\right\}\right]\right| + \left(\frac{2}{\pi}\right)^{1/2} \left|\sigma_T(Y_{1:T},\theta) - \sigma(\theta)\right|, \tag{4.56}$$

We have already shown

$$\sup_{\theta\in B(\bar{\theta})} \left|\sigma_T^2(Y_{1:T},\theta) - \sigma^2(\theta)\right| = \sup_{\theta\in B(\bar{\theta})} \left|\sum_{t=1}^{T} \frac{\sigma^2\left(Y_t,\theta\right)}{T} - \sigma^2(\theta)\right| \xrightarrow{\mathbb{P}^Y} 0,$$

by the uniform law of large numbers (Lemma 4.34). Using $|\sqrt{a}-\sqrt{b}| \leqslant \sqrt{|a-b|}$, we have

$$\sup_{\theta\in B(\bar{\theta})} \left|\sigma_T(Y_{1:T},\theta) - \sigma(\theta)\right| \xrightarrow{\mathbb{P}^Y} 0.$$

For the first part of (4.56), by Lemma 4.33 applied conditionally on $\mathcal{Y}_T$

$$\sup_{\theta\in B(\bar{\theta})} \sigma_T(Y_{1:T},\theta) \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}}\leqslant 1}} \left| E\left[f\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\epsilon_T(Y_t,\theta)}{\sigma_T(Y_{1:T},\theta)}\right\} \mid \mathcal{Y}_T\right] - E\left[f\left\{Z\right\}\right]\right|$$

$$\leqslant 4 \sup_{\theta\in B(\bar{\theta})} \sigma_T(Y_{1:T},\theta) \sum_{t=1}^{T} E\left[\left\{\frac{\epsilon(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right\}^2 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}>1\right\}} \mid \mathcal{Y}_T\right] \tag{4.57}$$

$$+ 3 \sup_{\theta\in B(\bar{\theta})} \sigma_T(Y_{1:T},\theta) \sum_{i=1}^{T} E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|^3 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leqslant 1\right\}} \mid \mathcal{Y}_T\right]. \tag{4.58}$$

In order to control the $\sigma^2(Y_{1:T},\theta)$ term consider the set

$$A_T(\delta) = \left\{ y_{1:T} : \sup_{\theta\in B(\bar{\theta})} \left|\sigma_T^2(y_{1:T},\theta) - \sigma^2(\theta)\right| \leqslant \delta \right\}.$$

The uniform convergence of $\sigma_T^2(Y_{1:T},\theta)$ means that for any $\delta > 0$

$$\mathbb{P}^Y\left\{A_T(\delta)^{\complement}\right\} \to 0$$

as $T \to \infty$. Choosing $\delta > 0$ for any family of random variables $\gamma_T(Y_{1:T},\theta)$ we have

$$\mathbb{P}^Y\left(\left|\sup_{\theta\in B(\bar{\theta})} \gamma_T(Y_{1:T},\theta)\right| > \delta\right)$$

$$= \mathbb{P}^Y\left(\left\{\left|\sup_{\theta\in B(\bar\theta)}\gamma_T(Y_{1:T},\theta)\right| > \delta\right\}\cap A_T(\delta)\right) + \mathbb{P}^Y\left(\left\{\left|\sup_{\theta\in B(\bar\theta)}\gamma_T(Y_{1:T},\theta)\right| > \delta\right\}\cap A_T(\delta)^{\complement}\right)$$

where we have already shown

$$\mathbb{P}^Y\left[\left\{\left|\sup_{\theta\in B(\bar\theta)}\gamma_T(Y_{1:T},\theta)\right| > \eta\right\}\cap A_T(\delta)^{\complement}\right] \leqslant \mathbb{P}^Y\left\{A_T(\delta)^{\complement}\right\} \to 0. \tag{4.59}$$

Hence, for showing the convergence in probability for a random variable $\gamma_T(Y_{1:T},\theta)$ it suffices to ensure convergence on the set $A(\delta)$. On the set $A(\delta)$ we can estimate $\sigma_T^2\left\{Y_{1:T}(\omega),\theta\right\} \geqslant \sigma^2(\theta)-\delta$ for all θ. By continuity of $\sigma^2(\theta)$—and by shrinking $B(\bar\theta)$ if necessary—we further have $\sigma^2(\theta)\geqslant\sigma^2(\bar\theta)-\delta$ for all $\theta\in B(\bar\theta)$ and we get for (4.57), ignoring the constant for now

$$\begin{aligned}
&\sup_{\theta\in B(\bar\theta)}\sigma_T(Y_{1:T},\theta)\sum_{t=1}^{T}E\left[\left\{\frac{\epsilon(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right\}^2 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}>1\right\}} \mid \mathcal{Y}_T\right]\mathbf{1}_{A_T(\delta)}\\
&\leqslant \sup_{\theta\in B(\bar\theta)}\frac{1}{\sigma_T^{1+\Delta}(Y_{1:T},\theta)}\sum_{t=1}^{T}E\left[\left\{\frac{\epsilon(Y_t,\theta)}{\sqrt{T}}\right\}^{2+\Delta} 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}>1\right\}} \mid \mathcal{Y}_T\right]\mathbf{1}_{A_T(\delta)}\\
&\leqslant \sup_{\theta\in B(\bar\theta)}\frac{1}{\left\{\sigma^2(\theta)-\delta\right\}^{(1+\Delta)/2}}\frac{1}{T^{1+\Delta/2}}\sum_{t=1}^{T}E\left\{\left|\epsilon_T(Y_t,\theta)\right|^{2+\Delta}\mid\mathcal{Y}_T\right\}\mathbf{1}_{A_T(\delta)}\\
&\leqslant \frac{C}{\left\{\sigma^2(\bar\theta)-2\delta\right\}^{(1+\Delta)/2}T^{1+\Delta/2}}\sum_{t=1}^{T}\left\{g(Y_t)+1\right\}\xrightarrow{\mathbb{P}^Y}0
\end{aligned}$$

independently of θ by the Marcinkiewicz-Zygmund law of large numbers (Kallenberg 2006, Theorem 4.23). Together with (4.59) we can conclude that (4.57), vanishes in probability.

The second part, (4.58), can be controlled similarly via

$$\begin{aligned}
&\sigma_T(Y_{1:T},\theta)\sum_{i=1}^{T}E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|^3 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leqslant 1\right\}}\mid\mathcal{Y}_T\right]\mathbf{1}_{A_T(\delta)}\\
&\leqslant \frac{1}{\sigma_T^{1+\Delta}(Y_{1:T},\theta)}\sum_{t=1}^{T}E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}}\right|^{2+\Delta} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leqslant 1\right\}}\mid\mathcal{Y}_T\right]\mathbf{1}_{A_T(\delta)}\\
&\leqslant \frac{1}{\left\{\sigma^2(\theta)-\delta\right\}^{(1+\Delta)/2}}\sum_{t=1}^{T}E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}}\right|^{2+\Delta} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leqslant 1\right\}}\mid\mathcal{Y}_T\right]\mathbf{1}_{A_T(\delta)}\\
&\leqslant \frac{1}{\left\{\sigma^2(\bar\theta)-2\delta\right\}^{(1+\Delta)/2}}\frac{1}{T^{1+\Delta/2}}\sum_{t=1}^{T}E\left\{\left|\epsilon(Y_t,\theta)\right|^{2+\Delta}\mid\mathcal{Y}_T\right\}\mathbf{1}_{A_T(\delta)}
\end{aligned}$$

$$\leqslant \frac{C}{\left\{\sigma^2(\bar{\theta}) - 2\delta\right\}^{(1+\Delta)/2} T^{1+\Delta/2}} \sum_{t=1}^{T} \left\{g(Y_t) + 1\right\} \xrightarrow{\mathbb{P}^Y} 0,$$

which also does not depend on θ. A similar argument to the one used to conclude in the case of (4.57) suffices also in this case. □

Turning to part *b)*, we analyse $Z_T(\theta)$ under stationarity. Therefore we need to introduce the probability measure of the auxiliary variables under stationarity, i.e. the distribution of the auxiliary variables conditional on the current state θ. The conditional density is given by

$$\pi(u \mid \theta) = \frac{\pi(u, \theta)}{\pi(\theta)} = \pi(\theta)\frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m(u)/\pi(\theta) = \frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m(u)$$

which gives us the Radon-Nikodym derivative

$$\frac{d\pi(\cdot \mid \theta)}{dm} = \prod_{t=1}^{T} \frac{\hat{p}(y_t \mid \theta, u_t)}{p(y_t \mid \theta)} = \exp\left\{Z_T(\theta)\right\}$$

or alternatively

$$\begin{aligned}\prod_{t=1}^{T} \frac{\hat{p}(y_t \mid \theta, u_t)}{p(y_t \mid \theta)} &= \prod_{t=1}^{T} \left\{ \frac{\hat{p}(y_t \mid \theta, u_t) - p(y_t \mid \theta)}{p(y_t \mid \theta)} + 1 \right\} \\ &= \prod_{t=1}^{T} \left\{ \frac{\epsilon_T(y_t, \theta)}{\sqrt{T}} + 1 \right\}.\end{aligned}$$

The limiting distribution will now be Gaussian with a shifted mean, i.e. $\varphi(\cdot; \sigma^2(\theta)/2, \sigma^2(\theta))$. For $Z \sim \mathcal{N}(0, 1)$ we will make use of the following identity

$$E\left\{ f\left(Z\sigma + \frac{\sigma^2}{2} \right) \right\} = E\left\{ f\left(Z\sigma - \frac{\sigma^2}{2} \right) \exp\left(Z\sigma - \frac{\sigma^2}{2} \right) \right\}$$

for every bounded Lipschitz function f. The identity is not restricted to this case, but we will only consider bounded Lipschitz functions. Before we present the proof, we have the following useful result.

4.35 Proposition. *The Radon-Nikodym derivative is asymptotically uniformly bounded in its second moment,*

$$\limsup_{T\to\infty} \sup_{\theta \in B(\bar{\theta})} E\left[\prod_{t=1}^{T} \left\{ \frac{\epsilon_T(Y_t, \theta)}{\sqrt{T}} + 1 \right\}^2 \mid \mathcal{Y}_T \right] < \infty.$$

Proof. Using independence of $\left(U_{t,1:T}\right)_{t\geqslant 1}$ we compute for all $\theta \in B(\bar{\theta})$

$$\begin{aligned}\prod_{t=1}^{T} E\left\{\frac{\epsilon_T(Y_t,\theta)^2}{T}+2\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1 \mid \mathcal{Y}_T\right\} &= \prod_{t=1}^{T}\left\{\frac{\sigma^2(Y_t,\theta)}{T}+1\right\}\\ &\leqslant \exp\left\{\sum_{t=1}^{T}\frac{\sigma^2(Y_t,\theta)}{T}\right\}\\ &\leqslant \exp\left\{\sum_{t=1}^{T}\frac{C(1+g(Y_t))^{2/(2+\Delta)}}{T}\right\}\\ &\to \exp\left[CE\left\{\left(1+g\left(Y_1\right)\right)^{2/(2+\Delta)}\right\}\right]\end{aligned}$$

in $\mathbb{P}^Y$-probability, which is clearly finite by Assumption 4.4. □

In the following we denote E the expectation under m and $\tilde{E}$ the expectation under $\pi(\cdot \mid \theta)$. Using the Radon-Nikodym derivative, it is possible to relate the expectation of $\epsilon_T(y,\theta)^k$ under U at stationarity (conditional on θ) to the expectation under $U \sim m(\cdot)$ by

$$E_{U\sim\pi(\cdot|\theta)}\left\{\epsilon_T(y,\theta)^k\right\} = \frac{1}{\sqrt{T}}E_{U\sim m(\cdot)}\left\{\epsilon_T(y,\theta)^{k+1}\right\} + E_{U\sim m(\cdot)}\left\{\epsilon_T(y,\theta)^k\right\};$$

see Deligiannidis, Doucet and Pitt 2018, Lemma 4 for a proof. We are now able to prove the second part of Theorem 4.32.

Proof of Theorem 4.32, part b). Again we take $Z \sim \mathcal{N}(0,1)$ and use the same decomposition as before, but with all expectations replaced by $\tilde{E}$, the expectation at stationarity:

$$\begin{aligned}&\sup_{\theta\in B(\bar{\theta})} d_{\mathrm{BL}}\left[\bar{g}_T^{\omega}(\cdot \mid \theta), \varphi\left\{\cdot;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}\right]\\ &= \sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leqslant 1}}\left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2+\sum_{t=1}^{T}R_T(Y_t,\theta)\right\} \mid \mathcal{Y}_T\right]\right.\\ &\qquad \left.-\tilde{E}\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\right]\right|\\ &\leqslant \sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leqslant 1}}\left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2+\sum_{t=1}^{T}R_T(Y_t,\theta)-\frac{\sigma^2(\theta)}{2}+\frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right.\\ &\qquad \left.-\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)+\sum_{t=1}^{T}R_T(Y_t,\theta)-\frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right|\end{aligned} \tag{4.60}$$

$$+ \sup_{\theta\in B(\bar{\theta})} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| \tilde{E}\left[f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} \right\} \mid \mathcal{Y}_T \right] \right. \tag{4.61}$$

$$\left. - \tilde{E}\left[f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2} \right\} \mid \mathcal{Y}_T \right] \right|$$

$$+ \sup_{\theta\in B(\bar{\theta})} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}}\leqslant 1}} \left| \tilde{E}\left[f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2} \right\} \mid \mathcal{Y}_T \right] - \tilde{E}\left[f\left\{ Z\sigma(\theta) + \frac{\sigma^2(\theta)}{2} \right\} \right] \right|. \tag{4.62}$$

For (4.60) we have

$$(4.60) \leqslant \sup_{\theta\in B(\bar{\theta})} \tilde{E}\left[\min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \mid \mathcal{Y}_T \right].$$

An application of Cauchy-Schwartz yields

$$\begin{aligned}
&\tilde{E}\left[\min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \mid \mathcal{Y}_T \right] \\
&= E\left[\min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \prod_{t=1}^{T} \left\{ \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1 \right\} \mid \mathcal{Y}_T \right] \\
&\leqslant E\left[\min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right|^2 \right\} \mid \mathcal{Y}_T \right]^{1/2} E\left[\prod_{t=1}^{T} \left\{ \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1 \right\}^2 \mid \mathcal{Y}_T \right]^{1/2} \\
&\leqslant E\left[\min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \mid \mathcal{Y}_T \right]^{1/2} E\left[\prod_{t=1}^{T} \left\{ \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1 \right\}^2 \mid \mathcal{Y}_T \right]^{1/2}.
\end{aligned}$$

By Proposition 4.35

$$\limsup_{T\to\infty} \sup_{\theta\in B(\bar{\theta})} E\left[\prod_{t=1}^{T} \left\{ \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1 \right\}^2 \mid \mathcal{Y}_T \right]^{1/2} < \infty$$

and we have previously shown that

$$\sup_{\theta\in B(\bar{\theta})} E\left[\min\left\{ 1, \left| -\frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 + \frac{\sigma^2(\theta)}{2} \right| \right\} \mid \mathcal{Y}_T \right] \to 0.$$

As for the remainder (4.61) we argue analogously

$$\begin{aligned}
&\tilde{E}\left[\min\left\{1,\left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right] \\
&= E\left[\min\left\{1,\left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} \mid \mathcal{Y}_T\right] \\
&\leqslant E\left[\min\left\{1,\left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\}^2 \mid \mathcal{Y}_T\right]^{1/2} \cdot E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2} \\
&\leqslant E\left[\min\left\{1,\left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right]^{1/2} \cdot E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2}.
\end{aligned}$$

The first factor vanishes in probability as we have shown in the proof of Theorem 4.32(a), where as the second factor is bounded by Proposition 4.35.

For (4.62), note first that

$$E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} \mid \mathcal{Y}_T\right] = 1 \quad \text{and} \quad E\left[e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\right] = 1.$$

Hence, we can write

$$\begin{aligned}
&\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] \\
&= \tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} \mid \mathcal{Y}_T\right] E\left[e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\right] \\
&= \tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}} \mid \mathcal{Y}_T\right]
\end{aligned}$$

and similarly

$$\begin{aligned}
\tilde{E}\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\right] &= E\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\} e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\right] E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} \mid \mathcal{Y}_T\right] \\
&= E\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\} e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\} \mid \mathcal{Y}_T\right],
\end{aligned}$$

where we used that Z is independent of all other random variables in both cases. Using these

identities we obtain

$$\begin{aligned}
&\left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]-\tilde{E}\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\right]\right| \\
&\leqslant \left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]\right| \\
&\leqslant \left|E\left[\left(f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right)\right.\right. \\
&\qquad \left.\left.\times\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\mid \mathcal{Y}_T\right]\right| \\
&\leqslant \left|E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]^2\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right| \\
&\quad\times\left|E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\mid \mathcal{Y}_T\right]^{\frac{1}{2}}\right|.
\end{aligned}$$

We investigate the two factors of the product separately. First we use the fact that $\|f\|_\infty \leqslant 1$ when $\|f\|_{\mathrm{BL}} \leqslant 1$ (see (4.14)) and thus

$$\begin{aligned}
&\left|\sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leqslant 1}} E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]^2\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right| \\
&\leqslant \sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leqslant 1}}\left|E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right| \to 0
\end{aligned}$$

in $\mathbb{P}^Y$-probability as established in the previous part. For the second factor note that Z is independent of all other random variables and hence

$$\begin{aligned}
&\sup_{\theta\in B(\bar{\theta})}\left|E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\mid \mathcal{Y}_T\right]^{\frac{1}{2}}\right| \\
&=\sup_{\theta\in B(\bar{\theta})}\left|E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2\mid \mathcal{Y}_T\right]^{\frac{1}{2}} E\left[e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\right]^{\frac{1}{2}}\right|.
\end{aligned}$$

We know

$$\sup_{\theta\in B(\bar{\theta})} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 \mid \mathcal{Y}_T\right]^{\frac{1}{2}}$$

converges to a constant in $\mathbb{P}^Y$-probability and

$$\sup_{\theta\in B(\bar{\theta})} E\left[e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\right]^{\frac{1}{2}} = \sup_{\theta\in B(\bar{\theta})} \exp(\sigma(\theta)^2)^{1/2} < \infty.$$

□

4.11 Generalized Linear Mixed Models

4.11.1 Exponential Families and Random Effects

In this section we introduce a class of random effects models for which all assumptions required for Theorem 4.1 are satisfied. We analyse the latent variable model introduced in Section 4.5 for the popular class of generalized linear mixed models (see e.g McCulloch and Neuhaus 2005), where the observation density is of the form of an exponential family. We restrict attention here to the class of natural exponential family distributions, i.e. $T(y) = y$, with respect to the Lebesgue measure

$$p(y \mid \eta) = m(y)\exp\left\{\eta^{\mathsf{T}} y - A(\eta)\right\}, \tag{4.63}$$

where y is the natural sufficient statistic and η denotes the natural parameter, which will be set equal to the linear predictor in a generalized linear model. The function $m(y)$ is a base measure, which can be absorbed into the dominating measure. $A(\eta)$ is commonly referred to as the log-partition function and we assume that A is strictly convex and increasing in η so that the log-likelihood will be strictly concave. This assumption will be satisfied in the most common natural exponential family models including Poisson and Binomial models. In the following we will allow for multiple measurements for each group, which means we have one random effect associated with multiple observations. This corresponds to the logistic mixed model of Section 4.7. For the conditional exponential family with J repeated measurements $y_t = (y_{t,1},\dots,y_{t,J})^{\mathsf{T}}$ where $\eta_{t,j} = c_{t,j}^{\mathsf{T}}\beta + X_t, j = 1,\dots,J, t = 1,\dots,T$ and the random effects are centred Gaussian variables $X \sim \mathcal{N}(0,\tau^2)$ independent for each set of repeated measurements y. We will simplify the notation by dropping the subscript t as the importance

sampler for each t can be considered in isolation. Assume here that

$$g(y \mid x, \theta) = \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) y_j - A\{\eta_j(x)\}\right], \quad f(x \mid \theta) = \varphi(x; 0, \tau^2), \tag{4.64}$$

where $\eta_j(x) = c_j^\top \beta + x$ and c is a vector of covariates with corresponding parameter vector β. The (full) model likelihood for every observation is now given by

$$p(y, x \mid \theta) \propto \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) y_j - A\{\eta_j(x)\}\right] \varphi(x, 0, \tau^2).$$

Since X is unobserved, we are interested in the marginal likelihood

$$\begin{aligned} p(y \mid \theta) &= \int p(y, x \mid \theta) \mathrm{d}x \\ &= \int \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) y_j - A\{\eta_j(x)\}\right] \varphi(x, 0, \tau^2) \mathrm{d}x. \end{aligned}$$

Consequently, the likelihood of a set of observations $y_{1:T}$, with $y_i = (y_{i,1}, \ldots, y_{i,J})$ is

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} \int \prod_{j=1}^{J} m(y_{t,j}) \exp\left[\eta_{t,j}(x_t) y_{t,j} - A\{\eta_{t,j}(x_t)\}\right] \varphi(x_t, 0, \tau^2) \mathrm{d}x_t.$$

We list the log-partition function as well as it's first derivative $A'(x) = \partial_x A(x)$ (which will be important later) below together with the base measure.

Binomial. Denote n the number of trials, then

$$A(\eta) = n \log\left(1 + e^\eta\right), \quad A'(\eta) = \frac{n e^\eta}{1 + e^\eta}, \quad m(y) = \binom{n}{y}.$$

Poisson. For the Poisson family

$$A(\eta) = e^\eta, \quad A'(\eta) = e^\eta, \quad m(y) = \frac{1}{y!}.$$

4.11.2 *Asymptotic Posterior Normality*

This section establishes the Bernstein-von Mises theorem for priors having exponentially decaying tails. Denote $\Theta \subset \mathbb{R}^d$ a subset of the Euclidean space, where we take $d = 1$ without loss of generality. Consider the case of i.i.d. observations $Y_1, Y_2, \ldots$ drawn from a density $Y_i \sim f(\cdot \mid \bar{\theta})$, where $\bar{\theta} \in \Theta$ is assumed to be the 'true parameter'. The measure describing

the distribution of the data vector $Y_{1:T} = (Y_1, \ldots, Y_T)$ is written as $P_{T,\bar{\theta}}$. Writing $\pi(\theta)$ for the prior distribution we denote the posterior density as

$$\pi_T(\theta) = \pi(\theta \mid Y_{1:T}) = \frac{\prod_{i=1}^T f(y_i \mid \theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^T f(y_i \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

4.36 Theorem. *Let the experiment be differentiable in quadratic mean at $\bar{\theta}$ with non-singular Fisher information matrix $I_{\bar{\theta}}$, and suppose that for every $\varepsilon > 0$ there exist an increasing sequence of sets $K_1 \subset K_2 \subset \ldots$ with $\cup_{i=1}^\infty K_i = \Theta$ with K_T growing at rate T. Assume there exists a sequence of tests such that*

$$E(\phi_T) \to 0, \qquad \sup_{\{\|\theta-\bar{\theta}\|\geqslant\varepsilon\}\cap K_T} E_\theta^n\left(1-\phi_T\right) \to 0.$$

Furthermore, let the prior measure be absolutely continuous in a neighbourhood of $\bar{\theta}$ with a continuous positive density at $\bar{\theta}$ s.t. for T large enough, we have

$$\pi\left([-T,T]^\complement\right) \leqslant c_1 \exp\left(-c_2 T\right),$$

where c_1 and c_2 are positive constants. Then the corresponding posterior distributions satisfy

$$\int \left|\tilde{\pi}_T(h) - \varphi\left(h, \sqrt{T}\left(\hat{\theta}_T - \theta_0\right), I_{\bar{\theta}}^{-1}\right)\right| \mathrm{d}h \to 0 \tag{4.65}$$

in $P_{T,\bar{\theta}}$-probability where

$$\Delta_T(\bar{\theta}) = \frac{1}{\sqrt{T}} \sum_{i=1}^T \tilde{I}_{\bar{\theta}}^{-1} \frac{\partial \ell(\bar{\theta}, Y_i)}{\partial \theta}$$

and

$$\tilde{\pi}_T(h) = \frac{\pi_T(\bar{\theta} + h/\sqrt{T})}{T^{1/2}}$$

is a measure on $H = \left\{h = \sqrt{T}\left(\theta - \bar{\theta}\right) : \theta \in \Theta\right\}$.

Proof. The proof follows Van der Vaart (2000), Theorem 10.1, see also the lecture notes by Nickl (2012). We will show that it is enough to show convergence of the measures restricted on some arbitrarily large compact set. In order to do so, denote

$$P^C(A) = \frac{P(A \cap C)}{P(C)}$$

for any measurable set A the restriction of the probability measure P to the set C. Denote $h = \sqrt{T}\left(\theta - \bar{\theta}\right)$. We will write $P_{T,h}$ for the posterior distribution with data $Y_{1:T}$ and parameter $\bar{\theta} + h/\sqrt{T}(= \theta)$. Define the prior-weighted mixture measure over a set C as

$$P_{T,C} = \int P_{T,h}\tilde{\pi}_T^C(h)\mathrm{d}h.$$

The expectation with respect to $P_{T,C}$ is calculated as

$$E_{P_{T,C}}\left\{f(Y_{1:T})\right\} = \iint f(y_{1:T})\mathrm{d}P_{T,h}(y_{1:T})\tilde{\pi}_T^C(h)\mathrm{d}h.$$

For any sequence of sets A_T with $P_{T,\bar{\theta}}(A_T) \to 0$ it follows that $P_{T,B}(A_T) \to 0$ and vice versa, where B denotes a closed ball around 0. (Two measures with this relationship are called mutually contiguous.) This means that we can interchange convergence in probability under the measures $P_{T,B}$ and $P_{T,0}$. Let C now denote a ball of size M_T around 0 where $M_T \to \infty$ as $T \to \infty$. We can show that the total variation between distance between the posterior and the posterior restricted on the set C vanishes by estimating

$$\left\|\tilde{\pi}_T\left(B\right) - \tilde{\pi}_T^C\left(B\right)\right\|_{\mathrm{tv}} \leqslant 2\tilde{\pi}_T\left(C^{\complement}\right),$$

where $\|\cdot\|_{\mathrm{tv}}$ denotes the total variation norm. We will show that the left-hand side converges to zero under $P_{T,B}$ for B a closed ball around 0. We can now use the tests ϕ_T to bound

$$\begin{aligned}
E_{T,B}\left\{\tilde{\pi}_T(C^{\complement})\right\} &= E_{T,B}\left\{\tilde{\pi}_T\left(C^{\complement}\right)\left(1 - \phi_T + \phi_T\right)\right\} \\
&\leqslant E_{T,B}\left[\tilde{\pi}_T\left(C^{\complement}\right)\left(1 - \phi_T\right)\right] + E_{T,B}\left(\phi_T\right),
\end{aligned}$$

where $E_{T,B}\left(\phi_T\right) = o_{P_{T,B}}(1)$ by assumption. Now

$$\begin{aligned}
&E_{T,B}\left\{P_{H_T|Y_{1:T}}\left(C^{\complement}\right)\left(1 - \phi_T\right)\right\} \\
&= \int_B \int_{\mathbb{R}^T} \int_{C^{\complement}} \left(1 - \phi_T\right) \frac{\prod_{i=1}^T f(\bar{\theta} + g/\sqrt{T}, y_i)}{\int \prod_{i=1}^T f(\bar{\theta} + m/\sqrt{T}, y_i)\mathrm{d}\tilde{\pi}(m)} \prod_{i=1}^T f\left(\bar{\theta} + \frac{h}{\sqrt{T}}, y_i\right)\mathrm{d}y_i \frac{\mathrm{d}\tilde{\pi}(h)}{\tilde{\pi}(B)} \\
&= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} \int_{C^{\complement}} \int_{\mathbb{R}^T} \int_B \left(1 - \phi_T\right) \frac{\prod_{i=1}^T f(\bar{\theta} + \frac{h}{\sqrt{T}}, y_i)}{\int \prod_{i=1}^T f(\bar{\theta} + m/\sqrt{T}, y_i)\mathrm{d}\tilde{\pi}(m)} \mathrm{d}\tilde{\pi}(h)\mathrm{d}P_g^T(y)\mathrm{d}\tilde{\pi}^{C^{\complement}}(g) \\
&= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} E_{T,C^{\complement}}\left\{\tilde{\pi}_T(B)(1 - \phi_T)\right\}.
\end{aligned}$$

The upper bound is

$$\begin{aligned}
\frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} E_{C^{\complement}}^{T} \tilde{\pi}_T(B)(1-\phi_T) &= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} \int_{C^{\complement}} \tilde{\pi}_T(B)(1-\phi_T) P_h^T \frac{\mathrm{d}\tilde{\pi}(h \cap C^{\complement})}{\tilde{\pi}(C^{\complement})} \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} \int_{\mathbb{R}^d} \tilde{\pi}_T(B)(1-\phi_T) d\, P_h^T(y) d\tilde{\pi}(h \cap C^{\complement}) \\
&\leqslant \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h \cap C^{\complement}) \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) + \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h),
\end{aligned}$$

where $\tilde{K}_T = \{h = \sqrt{T}(\theta - \bar{\theta}) : \theta \in K_T\}$. For simplicity and without loss of generality we assume $K_T = [-T, T]$ in the following. By Van der Vaart (2000, Lemma 10.3) the tests converge exponentially fast so with $\theta = \bar{\theta} + h/\sqrt{T}, h = \sqrt{T}\left(\theta - \bar{\theta}\right), \mathrm{d}\theta = \mathrm{d}h/\sqrt{T}$

$$\begin{aligned}
\frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) &= \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta - \bar{\theta}\| \geqslant M_T/\sqrt{T}\} \cap K_T} E_\theta\left(1-\phi_T\right) \pi(\theta) \mathrm{d}\theta \\
&= \frac{1}{\tilde{\pi}(U)} \int_{\{\|\theta - \bar{\theta}\| \geqslant M_T/\sqrt{T}\} \cap K_T} E_\theta\left(1-\phi_T\right) \pi(\theta) \mathrm{d}\theta \\
&= \frac{1}{\tilde{\pi}(B)} \int_{\{D' \geqslant \|\theta - \bar{\theta}\| \geqslant M_T/\sqrt{T}\} \cap K_T} E_\theta\left(1-\phi_T\right) \pi(\theta) \mathrm{d}\theta \\
&\quad + \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta - \bar{\theta}\| \geqslant D'\} \cap K_T} E_\theta\left(1-\phi_T\right) \pi(\theta) \mathrm{d}\theta \\
&= c_2 \int_{\{D' \geqslant \|\theta - \bar{\theta}\| \geqslant M_T/\sqrt{T}\} \cap K_T} \exp\left(-DT \left\|\theta - \bar{\theta}\right\|^2\right) \mathrm{d}\theta \\
&\quad + \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta - \bar{\theta}\| \geqslant D'\} \cap K_T} \exp(-c_3 T) \pi(\theta) \mathrm{d}\theta \\
&\leqslant c_2 \int_{\{h : h \geqslant M_T\} \cap K_T} \exp\left(-DT \left\|\theta - \bar{\theta}\right\|^2\right) T^{1/2} \mathrm{d}\theta \\
&\quad + 2c_3 T^{1/2} \exp(-c_4 T),
\end{aligned}$$

where we used $\tilde{\pi}(B) \geqslant 1/(c_3 T^{1/2})$ for some constant c_3 because the prior is positive and continuous at $\bar{\theta}$. For the second part

$$\frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{c}} E_h(1-\phi_T) d\tilde{\pi}(h) \leqslant \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} \mathrm{d}\tilde{\pi}(h)$$

$$
\begin{aligned}
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} \pi(\bar{\theta} + h/\sqrt{T}) \mathrm{d}h \\
&\leqslant c_3 T^{1/2} \int_{K_T^{\complement}} \pi(\theta) \mathrm{d}\theta \\
&\leqslant c_3 T^{1/2} \cdot c_1 \exp\left(-c_2 T\right).
\end{aligned}
$$

As $T \to \infty$ we have

$$\|\tilde{\pi}_T - \tilde{\pi}_T^C\|_{\mathrm{tv}} \to 0$$

in $P_{T,B}$-probability and by contiguity also in $P_{T,\bar{\theta}}$.

Similarly, for a Gaussian distribution with means $\sup \left|\mu_T\right| < \infty$ and variance σ^2 we have

$$\left\| \mathcal{N}\left(\mu_T, \sigma^2\right) - \mathcal{N}^C\left(\mu_T, \sigma^2\right) \right\| \leqslant 2\mathcal{N}\left(\mu_T, \sigma^2\right)\left(C^{\complement}\right).$$

We know that $\Delta_{T,\bar{\theta}}$ is uniformly tight, i.e. for any $\varepsilon > 0$ there exists K such that $\sup_T P\left(\left|\Delta_{T,\bar{\theta}}\right| \leqslant K\right) = 1 - \varepsilon$. Hence, with probability $1 - \varepsilon$

$$\left\| \mathcal{N}\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) - \mathcal{N}^C\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) \right\| \leqslant 2\mathcal{N}\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)(C^{\complement})$$

by choosing M (the radius of C) sufficiently large. Hence, by the triangle inequality we have to show that

$$\int \left| \tilde{\pi}_T^C(h) - \varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) \right| \mathrm{d}h \to 0$$

in $P_{T,0}$-probability. Denoting $x^+ = \max\{0, x\}$

$$
\begin{aligned}
&\frac{1}{2} \int \left| \tilde{\pi}_T^C(h) - \varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) \right| \mathrm{d}h \\
&= \int \left(1 - \frac{\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{\tilde{\pi}_T^C(h)} \right)^+ \tilde{\pi}_T^C(h) \mathrm{d}h \\
&= \int \left(1 - \frac{\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) \int 1_C f_{T,g}^C(g) \pi(g) \mathrm{d}g}{1_C \tilde{\pi}_T^C(h)} \right)^+ \tilde{\pi}_T^C(h) \mathrm{d}h \\
&= \int \left(1 - \int \frac{1_C(g) f_{T,g}^C(g) \pi(g) \varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{1_C(h) f_{T,h}^C(h) \pi(h) \varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)} \varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) \mathrm{d}g \right)^+ \tilde{\pi}_T^C(h) \mathrm{d}h
\end{aligned}
$$

$$\leqslant \iint \left(1 - \frac{f^C_{T,g}(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)}{f^C_{T,h}(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)}\right)^+ \varphi^C\left(g; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right) \mathrm{d}g\tilde{\pi}^C_n(h)\mathrm{d}h$$

$$\leqslant \left\{\sup_{x\in C} \varphi^C\left(x; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)\right\} \iint \left(1 - \frac{f^C_{T,g}(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)}{f^C_{T,h}(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)}\right)^+ \mathrm{d}g\tilde{\pi}^C_T(h)\mathrm{d}h.$$

By dominated convergence it is enough to conclude that this quantity goes to 0 in

$$\begin{aligned}
P_{T,C}(\mathrm{d}y)\tilde{\pi}^C_T(\mathrm{d}h)\lambda_C(\mathrm{d}g) &= \int P_{T,x}(\mathrm{d}y)\tilde{\pi}^C_T(h)\mathrm{d}h\lambda_C(\mathrm{d}g) \\
&= \int \prod_{i=1}^T f\left(\theta + s/\sqrt{T}, y_i\right) \frac{\prod_{i=1}^T f\left(\theta + h/\sqrt{T}, y_i\right)\tilde{\pi}^C(h)\mathrm{d}h}{\int \prod_{i=1}^T f\left(\theta + u/\sqrt{T}, y_i\right)\tilde{\pi}^C(u)\mathrm{d}u}\mathrm{d}s\lambda_C(\mathrm{d}g) \\
&= \prod_{i=1}^T f\left(\theta + h/\sqrt{T}, y_i\right)\tilde{\pi}^C(h)\mathrm{d}h\lambda_C(\mathrm{d}g) \\
&= P_{T,C}(\mathrm{d}y)\tilde{\pi}^C(h)\mathrm{d}h\lambda_C(\mathrm{d}g)
\end{aligned}$$

probability. Under Theorem 7.2 in Van der Vaart (2000) mean-square differentiability of the likelihood implies that the likelihood ratio allows for the LAN (Van der Vaart 2000, Definition 7.14) expansion

$$\begin{aligned}
&\frac{\prod_{i=1}^T f(\theta + g/\sqrt{T}, y_i)}{\prod_{i=1}^T f(\theta + h/\sqrt{T}, y_i)} = \\
&= \prod_{i=1}^T \frac{f(\theta + g/\sqrt{T}, y_i)}{f(\theta, y_i)} \Big/ \prod_{i=1}^T \frac{f(\theta + h/\sqrt{T}, y_i)}{f(\theta, y_i)} \\
&= \exp\left(\frac{1}{\sqrt{T}}\sum_{i=1}^T g\ell'_\theta(y_i) - \frac{1}{2}g^2 I_\theta - \frac{1}{\sqrt{T}}\sum_{i=1}^T h\ell'_\theta(y_i) - \frac{1}{2}h^2 I_\theta + o_{P_\theta}(1)\right)
\end{aligned}$$

and thus as $T \to \infty$ and using continuity of the prior π at $\bar{\theta}$ we have

$$1 - \frac{f^C_{T,g}(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)}{f^C_{T,h}(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I^{-1}_{\bar{\theta}}\right)} \to 0$$

which yields the result. □

4.37 Remark. i) *The centring sequence $\Delta_{T,\theta}$ can be replaced by any best regular estimator. To see this note that following Van der Vaart (2000, Theorem 8.14) any best regular estimator, $\hat{\theta}_T$, satisfies the expansion*

$$\sqrt{T}(\hat{\theta}_T - \bar{\theta}) = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} \tilde{I}_{\bar{\theta}}^{-1} \frac{\partial \ell(\bar{\theta}, Y_i)}{\partial \theta} + o_{P_{T,\bar{\theta}}}(1)$$

and thus

$$\Delta_T(\bar{\theta}) - \sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right) \to 0$$

in P_{T,θ_0}-probability as $T \to \infty$. Since

$$\left\| \mathcal{N}\left(\Delta_{T,\bar{\theta}}, \tilde{I}_{\bar{\theta}}^{-1}\right) - \mathcal{N}\left\{\sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right), \tilde{I}_{\bar{\theta}}^{-1}\right\} \right\| \lesssim \left\| \sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right) - \Delta_{T,\bar{\theta}} \right\| \to 0$$

in probability.

ii) *Under regularity conditions Van der Vaart 2000, Theorem 5.39 the maximum likelihood estimator is best regular and can be used as a centring sequence following the argument in* i).

We will now apply this Bernstein-von Mises result to our exponential family models. Hence, consider again the likelihood contribution of every observation y,

$$p(y \mid \beta, \tau) = \int \prod_{j=1}^{J} m(y_j) \exp\left\{(c_j^\top \beta + x) y_j - A(c_j^\top \beta + x)\right\} \varphi(x, 0, \tau^2) \mathrm{d}x. \tag{4.66}$$

For simplicity we assume that the exogenous variables c_j are all identical and that Θ is a subset of $\mathbb{R}$. Let A be continuously differentiable (e.g. the Binomial and Poisson models introduced above). The prior can be easily chosen to fulfil the conditions of the updated Bernstein–von Mises theorem. The other conditions need further analysis. In order to show differentiability in quadratic mean it is sufficient to prove that the map $\theta \mapsto p(y \mid \theta)^{1/2}$ is continuously differentiable. By Lemma 7.6 in Van der Vaart (2000) we need to show that

$$\theta \mapsto p(y \mid \theta)^{1/2} = \left[\int m(y) \exp\left\{\left(c^\top \beta + x\right) \cdot y - A(c^\top \beta + x)\right\} \varphi(x, 0, \tau^2) \mathrm{d}x \right]^{1/2}$$

is continuously differentiable for all y. Firstly,

$$\frac{\partial}{\partial \theta} p(y \mid \theta)^{1/2} = \frac{1}{2 p(y \mid \theta)^{1/2}} \partial_\theta p(y \mid \theta).$$

It is easy to see that $\theta \mapsto p(y \mid \theta)$ and $\theta \mapsto \partial_\theta p(y \mid \theta)$ are continuous. The fisher information is well defined, continuous in θ and positive since

$$\begin{aligned} I_\theta &= E\left[\left\{\partial_\theta \log p(Y \mid \theta)\right\}^2\right] \\ &= \int \left\{\partial_\theta \log p(y \mid \theta)\right\}^2 p(y \mid \theta)\mathrm{d}y > 0 \end{aligned}$$

whenever $\partial_\theta \log p(y \mid \theta)$ is not identically 0 for all y. The multivariate case is more involved and treated for example in Mukerjee and Sutradhar (2002) for the Binomial and Poisson case. In order to ensure the existence of the tests consider $K_1 \subset K_2 \subset \ldots$ an increasing sequence of compact sets with $\cup_{i=1}^{\infty} K_i = \Theta$. Then, if the model is identifiable and continuous in total variation norm, Lemma 10.6 in Van der Vaart (2000), and a diagonal argument similar to that in the proof of Van der Vaart 2000, Lemma 10.6, ensures the existence of a sequence of estimators $\hat{\theta}_T$ such that $\sup_{\theta \in K_T} P_\theta(|\hat{\theta}_T - \theta| \geqslant \varepsilon) \to 0$ whence we have, see for example Nickl 2012, Lemmas 1,2 in Section 2.2.3,

$$E_{\bar{\theta}}(\phi_T) \to 0, \qquad \sup_{\{\|\theta - \bar{\theta}\| \geqslant \varepsilon\} \cap K_T} E_{T,\theta}\left(1 - \phi_T\right) \to 0.$$

Since our model has a density with respect to the Lebesgue measure continuity in total variation is trivially the case as we can write the total variation distance as

$$\|P_\theta - P_{\theta'}\|_{\mathrm{tv}} = \int \left|p(y \mid \theta) - p(y \mid \theta')\right| \mathrm{d}y.$$

Therefore, by Scheffé's lemma, continuity in the parameter already implies convergence of the integral and therefore continuity in the total variation distance. To conclude that our models are indeed identifiable it is enough to ensure that

i) the integral

$$E(Y) = E\left[A'(k + X)\right] = \int_{\mathbb{R}} A'(k + \tau x)\varphi(x; 0, 1)\mathrm{d}x < \infty,$$

for all k, τ and

ii) the equation

$$\frac{A'(c^T \beta_1 + \tau_1 x)}{\tau_1} = \frac{A'(c^T \beta_2 + \tau_2 x)}{\tau_2} \quad \text{for all } c \text{ and } x$$

has no solution,

see Labouriau (2014). These conditions are fulfilled for the Binomial case, $A'(\eta) = ne^{\eta}/(1 + e^{\eta})$, and Poisson case $A'(\eta) = e^{\eta}$.

4.11.3 *Importance Sampling with Univariate Random Effects*

We will now consider Assumption 4.3 in the context of generalized linear mixed models, which we will prove using Assumption 4.4 and Theorem 4.9. In the following we will first consider a univariate random effect and a Gaussian importance sampling proposal. This will include the example of Section 4.7. In addition we will show how fatter tails in the proposal affect the existence of moments by considering a univariate t-proposal. Recall that we are interested in bounds on

$$E^{Y}\left[\sup_{\theta\in B(\bar{\theta})} E^{X|Y}\left\{\overline{w}(Y,X,\theta)^{a}\right\}\right] = E^{Y}\left[\sup_{\theta\in B(\bar{\theta})} \frac{E^{X|Y}\{w(Y,X,\theta)^{a}\}}{p(Y \mid \theta)^{a}}\right], \tag{4.67}$$

where $a > 0$, $\theta = (\beta, \tau)$ and $B(\bar{\theta}) \subset \Theta$ denotes a closed ε-ball around $\bar{\theta}$. For additional clarity, we write E^{Y} and $E^{X|Y}$ for the expectations over Y and X given Y, respectively. Consider the Gaussian proposal centred at the mode

$$q(x \mid y) = \varphi(x; \hat{x}, \tau_q^2), \tag{4.68}$$

where τ_q^2 denotes the proposal variance and $\hat{x}$ is the mode of $h(x; y) = g(y \mid x)f(x)$ and fulfils the first order condition

$$\hat{x} = \tau^2\left\{S - \widetilde{A}'(\hat{x})\right\}, \tag{4.69}$$

where $\widetilde{A}'(x) = \sum_{j=1}^{J} A'(c_j^{\top}\beta + x)$ with $A'(z) = \partial_z A(z)$ and $S = \sum_{j=1}^{J} y_j$. For later convenience we define the unnormalized proposal density

$$\widetilde{q}(x; y) = \frac{q(x \mid y)}{q(\hat{x} \mid y)},$$

where $q(x \mid y)$ is the proposal density. For a symmetric proposal distribution centred at $\hat{x}$ the term $q(\hat{x} \mid y)$ is simply an inverse normalizing constant, which only involves the proposal parameters. For the Gaussian proposal

$$\widetilde{q}(x; y) = \exp\left\{-\frac{(x-\hat{x})^2}{2\tau^2}\right\}, \quad q(\hat{x} \mid y) = \frac{1}{(2\pi\tau_q^2)^{1/2}}. \tag{4.70}$$

Associated with this we introduce the modified weight which is defined as

$$\widetilde{w}(x,y) = \frac{g(y \mid x)f(x)}{g(y \mid \widehat{x})f(\widehat{x})}\frac{1}{\widetilde{q}(x;y)} = \frac{h(x;y)}{h(\widehat{x};y)}\frac{1}{\widetilde{q}(x;y)}, \tag{4.71}$$

where $h(x;y) = g(y \mid x)f(x)$. These weights are easier to work with as $\widetilde{w}(x,y) = 1$ when $x = \widehat{x}$. It is easily seen that

$$\widetilde{w}(x,y) = \frac{h(x;y)}{q(x \mid y)}\frac{q(\widehat{x} \mid y)}{h(\widehat{x};y)} = w(x,y)\frac{q(\widehat{x} \mid y)}{h(\widehat{x};y)},$$

so that the modified weights are proportional to the standard weights $w(x,y)$ as a function of x. We can recast the expectation (4.67) as

$$\begin{aligned} E^Y\left[\sup_{\theta\in B(\bar{\theta})} E^{X|Y}\left\{\overline{w}(Y,X,\theta)^a\right\}\right] &= E^Y\left[\sup_{\theta\in B(\bar{\theta})}\frac{E^{X|Y}\{w(X,Y,\theta)^a\}}{p(Y \mid \theta)^a}\right] \\ &= E^Y\left[\sup_{\theta\in B(\bar{\theta})}\frac{E^{X|Y}\left\{\widetilde{w}(X,Y,\theta)^a\right\}}{E^{X|Y}\left\{\widetilde{w}(X,Y,\theta)\right\}^a}\right]. \end{aligned} \tag{4.72}$$

The log-density of the observations is given by

$$\begin{aligned} \log g(y \mid x) &= \sum_{j=1}^{J}\{\log m(y_j) + y_j\eta_j - A(\eta_j)\} \\ &= \sum_{j=1}^{J}\{\log m(y_j) + y_j c_j^\top\beta + y_j x - A(c_j^\top\beta + x)\} \\ &= k(y) + x(J\overline{y}) - \widetilde{A}(x), \end{aligned}$$

where $k(y)$ represents constant values (which do not depend upon x), $J\overline{y} = \sum_{j=1}^{J} y_j$ and $\widetilde{A}(x) = \sum_{j=1}^{J} A(c_j^\top\beta + x)$. Hence, we get

$$\log h(x;y) = \log g(y \mid x)f(x) = c + x(J\overline{y}) - \widetilde{A}(x) - \frac{x^2}{2\tau^2}. \tag{4.73}$$

We will proceed by deriving bounds for the denominator and enumerator of (4.67) separately. We present the following lemma on the denominator without reference to the Gaussian proposal, because it holds for general proposal distribution.

4.38 Lemma. *Consider the exponential family model with repeated measurement* $j =$

$1, \ldots, J$ *and Gaussian random effects. For general proposal density* $q(x \mid y)$ *we have*

$$\frac{1}{E^X\left\{\widetilde{w}(X,y)\right\}} \leqslant \frac{(2\pi)^{1/2}}{C}(b+1),$$

where $b = \tau\widetilde{A}'(\widehat{x})$ *and* $C = q(\widehat{x} \mid y)(2\pi\tau^2)^{1/2}$.

of Lemma 4.38. For given observation y, the expectation of the rescaled weights is

$$\begin{aligned} E^X\left\{\widetilde{w}(X,y)\right\} &= \int \widetilde{w}(x,y)q(x \mid y)\mathrm{d}x \\ &= \int \frac{h(x;y)}{h(\widehat{x};y)}\frac{q(x \mid y)}{\widetilde{q}(x;y)}\mathrm{d}x \\ &= q(\widehat{x} \mid y)\int \frac{h(x;y)}{h(\widehat{x};y)}\mathrm{d}x. \end{aligned}$$

Write again $S = J\bar{y} = \sum_{j=1}^{J} y_j$. Since $\widetilde{A}$ is an increasing function we obtain for $x \leqslant \widehat{x}$,

$$\begin{aligned} \log h(x;y) - \log h(\widehat{x};y) &= -\widetilde{A}(x) + xS - \frac{1}{2}\frac{x^2}{\tau^2} + \widetilde{A}(\widehat{x}) - \widehat{x}S + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} \\ &\geqslant (x-\widehat{x})S - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} \\ &= R_2 - \frac{1}{2}\frac{\{x-\tau^2 S\}^2}{\tau^2}, \end{aligned}$$

where

$$R_2 = \frac{\tau^2 S^2}{2} - \widehat{x}S + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} = \frac{\tau^2\widetilde{A}'(\widehat{x})^2}{2},$$

by using the first order condition for the mode $\widehat{x} = \tau^2\{S - \widetilde{A}'(\widehat{x})\}$. Therefore

$$E^X\left\{\widetilde{w}(X,y)\right\} \geqslant q(\widehat{x} \mid y)\left(2\pi\tau^2\right)^{1/2}\exp\left\{\frac{\tau^2\widetilde{A}'(\widehat{x})^2}{2}\right\}\Phi\{-\tau\widetilde{A}'(\widehat{x})\}.$$

Consider the inequality due to Birnbaum (1942)

$$\frac{\exp(-b^2/2)}{1-\Phi(b)} < \left(\frac{\pi}{2}\right)^{1/2}\left\{b + (b^2+4)^{1/2}\right\}$$

Setting $b = \tau\widetilde{A}'(\widehat{x})$ and $C = q(\widehat{x} \mid y)(2\pi\tau^2)^{1/2}$ gives

$$\frac{1}{E^X\left\{\widetilde{w}(X,y)\right\}} \leqslant \frac{\exp(-b^2/2)}{C\{1-\Phi(b)\}}$$

$$\leqslant C^{-1}\left(\frac{\pi}{2}\right)^{1/2}\left\{b+(b^2+4)^{1/2}\right\}$$
$$\leqslant \frac{(2\pi)^{1/2}}{C}(b+1),$$

as $(b^2+4)^{1/2} \leqslant b+2$. □

Using Lemma 4.38 we can set

$$\sup_{\theta\in B(\bar{\theta})} \frac{1}{E^X\left\{\widetilde{w}(X,y,\theta)\right\}} \leqslant \sup_{\theta\in B(\bar{\theta})} \frac{1}{q(\widehat{x}\mid y)\tau}\left\{\tau\widetilde{A}'(\widehat{x})+1\right\}.$$

We will use this result in the following corollary.

4.39 Corollary. *Assume one of the following condition holds:*

(i) $\sup_x \widetilde{A}'(x) < \infty$,

(ii) $E^Y(Y^a) < \infty$ *and* $\sup_{\theta\in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

Then taking the expectation over Y, we have

$$E^Y\left[\sup_{\theta\in B(\bar{\theta})} \frac{1}{E^{X|Y}\{\widetilde{w}(X,Y)\}^a}\right] < \infty.$$

Proof. Applying Lemma 4.38 with $b = \tau\widetilde{A}'(\widehat{x})$ yields

$$E^Y\left[\sup_{\theta\in B(\bar{\theta})} \frac{1}{E^X\{\widetilde{w}(X,Y)\}^a}\right] \leqslant E^Y\left[\sup_{\theta\in B(\bar{\theta})}\left\{\frac{\tau\widetilde{A}'(\widehat{x})+1}{C\tau}\right\}^a\right], \tag{4.74}$$

where we write $C = q(\widehat{x}\mid y)$ which only involved parameters of the proposal distribution. The right-hand side of (4.74) is finite provided $E^Y\{\sup_{\theta\in B(\bar{\theta})}\widetilde{A}'(\widehat{x})^a\} < \infty$. This concludes the proof for *(i)*. For *(ii)* we need to control the function $\widetilde{A}'(\widehat{x})$. Therefore, it is useful to establish the behaviour of $\widetilde{A}'(\widehat{x})$ in terms of the random variables $y = (y_1, \ldots, y_J)$. Recall the first order condition (4.69)

$$\widehat{x} = \tau^2\{J\overline{y} - \widetilde{A}'(\widehat{x})\},$$

where the sufficient statistic is $S = J\overline{y} = \sum_{j=1}^J y_j$. It is easily established that $\widetilde{A}'(\widehat{x}) \leqslant \max\{\widetilde{A}'(0), J\overline{y}\}$. To see this note

$$\partial_x \log h(x;y) = J\overline{y} - \widetilde{A}'(x) - \frac{x}{\tau^2}. \tag{4.75}$$

The function $\widetilde{A}'(x)$ is monotonically increasing. If $\widetilde{A}'(0) < J\overline{y}$, then at $x = 0$, $\partial_x \log h(x; y) > 0$ and at $x = \widetilde{x}$, where $\widetilde{A}'(\widetilde{x}) = J\overline{y}$, $\partial_x \log h(x; y) < 0$ since $\widetilde{x} > 0$. Similarly, if $\widetilde{A}'(0) < J\overline{y}$ then at $x = 0$, $\partial_x \log h(x; y) < 0$ and at $x = \widetilde{x}$, $\partial_x \log h(x; y) > 0$. As a consequence, the mode of the concave function $\log h(x; y)$, $\widehat{x}$ is always between 0 and $\widetilde{x}$, where $\widetilde{A}'(\widetilde{x}) = J\overline{y}$. This yields $\widetilde{A}'(\widehat{x}) \leqslant \max\{\widetilde{A}'(0), J\overline{y}\}$ so that

$$
\begin{aligned}
E^Y\left\{\sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(\widehat{x})^a\right\} &\leqslant E^Y\left[\sup_{\theta\in B(\overline{\theta})} \max\{\widetilde{A}'(0), S\}^a\right] \\
&\leqslant E^Y\left[\max\left\{\sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(0), S\right\}^a\right] \\
&= \sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(0)^a \mathbb{P}^Y\left\{S < \sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(0)\right\} + \int_{\sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(0)}^{\infty} s^a \mathrm{d}F_S(s).
\end{aligned}
$$

The last quantity is finite whenever $\sup_{\theta\in B(\overline{\theta})} \widetilde{A}'(0) < \infty$ and $E^Y(Y^a) < \infty$. □

4.40 Remark (Examples with Gaussian proposal)**.** *If the proposal is a Gaussian centred at the mode $q(x \mid y) = \varphi(x; \widehat{x}, \tau_q^2)$ and C as defined in Lemma 4.38, then $C = \tau/\tau_q$. For the Binomial case, we know that $\sup_x \widetilde{A}'(x) < \infty$ and therefore condition (i) of the preceding Corollary 4.39 is fulfilled. For the Poisson case $\widetilde{A}'(x)$ is not bounded, but we can use the second part of the corollary. Note that $\widetilde{A}'(x)$ is continuous and therefore $\widetilde{A}'(0)$ can be bounded in a neighbourhood small enough. In addition, if the Poisson model is true, it is straightforward to establish that the moments $E(Y^a)$ exist for all $a > 0$ and we can therefore conclude by part* (ii).

Having established conditions to ensure

$$
E^Y\left[\sup_{\theta\in B(\overline{\theta})} E^{X|Y}\{\widetilde{w}(X, Y)\}^{-a}\right] < \infty
$$

we can bound (4.72) whenever there exists a constant $K < \infty$ such that

$$
\sup_{y\in\mathsf{Y}} \sup_{\theta\in B(\overline{\theta})} E^{X|Y}\left\{\widetilde{w}(X, y)^a\right\} < K.
$$

In the following we will provide conditions for Gaussian and t-distributed proposals.

4.41 Proposition. *Consider the Gaussian proposal* (4.68) *and some exponent* $a > 0$*. Then*

$$E^X\{\widetilde{w}(X, y)^a\} < \infty$$

if and only if $\tau_q^2 > \frac{(a-1)}{a}\tau^2$*, where* τ^2 *is the variance of the random effects term. If this condition is satisfied then*

$$E^X\{\widetilde{w}(X, y)^a\} \leqslant \left\{ \frac{a\tau_q^2 - (a-1)\tau^2}{\tau^2} \right\}^{-\frac{1}{2}},$$

independent of y*.*

Proof. For brevity we define the sum $S = J\overline{y} = \sum_{j=1}^J y_j$ and again have $\widetilde{A}(x) = \sum_{j=1}^J A(cj^\top\beta + x)$. Note that $x \mapsto \widetilde{A}(x)$ is convex and thus always dominates its chord

$$\widetilde{A}(x) \geqslant \widetilde{A}(\widehat{x}) + \widetilde{A}'(\widehat{x})(x - \widehat{x})$$

for any values $x, \widehat{x}$. Then the modified proposal form $\widetilde{q}(x; y)$ is given by (4.70), so

$$\begin{aligned}
\log \widetilde{w}(x, y) &= \log h(x; y) - \log h(\widehat{x}; y) - \log \widetilde{q}(x; y) \\
&= xS - \widetilde{A}(x) - \frac{1}{2}\frac{x^2}{\tau^2} \\
&\quad - \widehat{x}S + \widetilde{A}(\widehat{x}) + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \frac{1}{2}\frac{(x-\widehat{x})^2}{\tau_q^2}.
\end{aligned}$$

This is, by design, zero at $x = \widehat{x}$ and can be bounded as

$$\begin{aligned}
\log \widetilde{w}(x, y) &\leqslant \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \{S - \widetilde{A}'(\widehat{x})\}(x - \widehat{x}) - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{1}{2}\frac{(x-\widehat{x})^2}{\tau_q^2} \\
&= \frac{1}{2}(x - \widehat{x})^2 d,
\end{aligned}$$

by noting the first order condition that $\widehat{x}/\tau^2 = S - \widetilde{A}'(\widehat{x})$. The constant d is defined to be

$$d = \frac{1}{\tau_q^2} - \frac{1}{\tau^2},$$

and $d > 0$ if we choose $\tau_q^2 < \tau^2$. Hence

$$E^X\{\widetilde{w}(X, y)^a\} \leqslant E^X\left[\exp\left\{\frac{ad}{2}(X - \widehat{x})^2\right\}\right], \tag{4.76}$$

where again the expectation is with respect to $q(x \mid y) = \varphi(x \mid \widehat{x}, \tau_q^2)$. As $a > 0$, clearly the above expectation exists if $d \leqslant 0$ which would imply choosing $\tau_q^2 \geqslant \tau^2$. To obtain a precise condition we note that

$$\frac{(X - \widehat{x})^2}{\tau_q^2} \sim \chi_1^2.$$

Considering the moment generating function of the χ^2-distribution we know that the expectation (4.76) exists provided

$$ad\tau_q^2 = a\left(1 - \frac{\tau_q^2}{\tau^2}\right) < 1, \quad \text{i.e.} \quad \tau_q^2 > \frac{(a-1)}{a}\tau^2. \tag{4.77}$$

If this inequality holds, the moment generating function of the χ^2-distribution exists and we have

$$E^X\left[\exp\left\{\frac{ad}{2}(X - \widehat{x})^2\right\}\right] = \left(1 - ad\tau_q^2\right)^{-1/2}.$$

Finally we obtain

$$E^X\{\widetilde{w}(X, y)^a\} \leqslant \left\{1 - a\left(1 - \frac{\tau_q^2}{\tau^2}\right)\right\}^{-1/2} = \left\{\frac{a\tau_q^2 - (a-1)\tau^2}{\tau^2}\right\}^{-1/2}. \tag{4.78}$$

as required. □

Note that by the upper bound in Proposition 4.41 still depends on parameters via the variance term τ. However, since the dependence is continuous we can find an upper bound over any compact set. Thus, we have the simple corollary.

4.42 Corollary. *Under the conditions of Proposition 4.41 there exists a constant $K_1 < \infty$ such that*

$$\sup_{\theta \in B(\bar{\theta})} E^X\{\widetilde{w}(X, y, \theta)^a\} \leqslant K_1$$

independent of y.

We can summarize the results so far in the following theorem.

4.43 Theorem. *Consider the random effects model* (4.11) *and assume we have an importance sampling estimator with proposal distribution*

$$q(x \mid y) = \varphi(x; \widehat{x}, \tau_q^2)$$

and proposal variance $\tau_q^2 > \frac{(a-1)}{a}\tau^2$. *Assume additionally that either*

i) $\sup_x \widetilde{A}'(x) < \infty$ *or*

ii) $E^Y(Y^a) < \infty$ *and* $\sup_{\theta \in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

Then

$$E^Y \left[\sup_{\theta \in B(\bar{\theta})} E^X \left\{ \overline{w}(Y, X, \theta)^a \right\} \right] < \infty.$$

Proof. We have

$$\begin{aligned} E^Y \left[\sup_{\theta \in B(\bar{\theta})} E^X \left\{ \overline{w}(Y, X, \theta)^a \right\} \right] &= E^Y \left[\sup_{\theta \in B(\bar{\theta})} \frac{E^X \left\{ \widetilde{w}(X, Y, \theta)^a \right\}}{E^X \left\{ \widetilde{w}(X, Y, \theta) \right\}^a} \right] \\ &\leqslant E^Y \left[\sup_{\theta \in B(\bar{\theta})} \frac{K_1}{E^X \left\{ \widetilde{w}(X, Y, \theta) \right\}^a} \right] < \infty. \end{aligned}$$

where the first inequality is by Corollary 4.42 and the second by Corollary 4.39. □

For the logistic model of Section 7, $\widetilde{A}'(x)$ is bounded above by a constant. Indeed

$$\widetilde{A}'(x) = \sum_{j=1}^{J} A'(c_j^\top \beta + x) = \sum_{j=1}^{J} \frac{e^{c_j^\top \beta + x}}{1 + e^{c_j^\top \beta + x}}.$$

Hence, we know (see Remark 4.40) that

$$E^Y \left[\sup_{\theta \in B(\bar{\theta})} E^{X|Y} \left\{ \overline{w}(X, Y)^a \right\} \right] < \infty$$

for all a if we take, for example, $\tau_q^2 = \tau^2$. We note, however, that the proposal may not be particularly efficient as the proposal variance would ideally be made to be proportional to $1/J$, where J represents the number of observations associated with each latent variate. Hence, taking $\tau_q^2 = \tau^2$, for example, may be much too large as a choice for τ_q^2. This naturally leads to consideration of the t-distribution which has heavier tails, see for example Owen 2013, Chapter 9 and so controls the numerator term. We consider the t-distribution proposal centred at the mode, with scaling τ_q^2, so that $q(x \mid y) = t_\nu(x \mid \widehat{x}, \tau_q^2)$. For the t-proposal, we have

$$\widetilde{q}(x; y) = \left\{ 1 + \frac{(x - \widehat{x})^2}{\nu \tau_q^2} \right\}^{-(\nu+1)/2}, \quad q(\widehat{x} \mid y) = \frac{\sqrt{\nu\pi}\, \Gamma(\nu/2)\, \tau_q}{\Gamma\{(\nu+1)/2\}}. \tag{4.79}$$

We proceed in the same manner as in the Gaussian case. First we compute the bound from Lemma 4.38 for the t-distribution. Assume the proposal is a t-distribution centred at the mode $q(x \mid y, \theta) = t_\nu(x \mid \widehat{x}, \tau_q^2)$, then

$$C = \frac{\tau}{\tau_q}\sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$$

and thus

$$\frac{1}{E^X\left\{\widetilde{w}(X, y)\right\}} \leqslant \frac{\tau_q}{\tau}\sqrt{\frac{\nu}{2}}\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}(b+1).$$

4.44 Proposition. *For the target $h(x; y)$ of (4.73) with $q(x \mid y, \theta) = t_\nu(x \mid \widehat{x}, \tau_q^2)$ specified above we shall assume that the function $x \mapsto A(x)$ is a monotonically non-decreasing convex function. Then,*

$$E^{X|Y}\left\{\widetilde{w}(X, Y)^a\right\} \leqslant K_2^a,$$

where

$$K_2 = \left\{\frac{\tau^2}{\tau_q^2}\frac{(\nu+1)}{\nu}\right\}^{\frac{(\nu+1)}{2}} \exp\left\{\frac{\nu}{2}\left(\frac{\tau_q^2}{\tau^2} - 1 - \frac{1}{\nu}\right)\right\},$$

for $\tau_q^2 < \frac{(\nu+1)}{\nu}\tau^2$ and $K_2 = 1$ for $\tau_q^2 \geqslant \frac{(\nu+1)}{\nu}\tau^2$.

Unlike the Gaussian proposal above, the t-distributed proposal does not have any restriction on how small the variance τ_q^2 can be. This might be chosen, for example, according to the second derivative of $\log h(x; y)$ at 0 so that $\tau_q^{-2} = \tau^{-2} + \widetilde{A}''(0)$. This would reflect the influence of a large number of repeated observations, $\boldsymbol{J}$.

Proof of Proposition 4.44. Recall that $x \mapsto A(x)$ is convex and thus always dominates its chord

$$\widetilde{A}(x) \geqslant \widetilde{A}(\widehat{x}) + \widetilde{A}'(\widehat{x})(x - \widehat{x})$$

for any values $x, \widehat{x}$. For the modified log weight this yields

$$\begin{aligned}\log \widetilde{w}(x, y) &= \log h(x; y) - \log h(\widehat{x}; y) - \log \widetilde{q}(x; y)\\ &= xS - \widetilde{A}(x) - \frac{1}{2}\frac{x^2}{\tau^2}\\ &\quad - \widehat{x}S + \widetilde{A}(\widehat{x}) + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}\end{aligned}$$

$$\leqslant \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \{S - \widetilde{A}'(\widehat{x})\}(x - \widehat{x}) - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}.$$

We recall that $\widehat{x}/\tau^2 = S - \widetilde{A}'(\widehat{x})$. Hence

$$\log \widetilde{w}(x, y) \leqslant -\frac{(x-\widehat{x})^2}{2\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}.$$

Writing $\widetilde{x} = (x - \widehat{x})/\tau_q$ we obtain

$$\log \widetilde{w}(x, y) \leqslant -\frac{1}{2}\frac{\tau_q^2}{\tau^2}\widetilde{x}^2 + \frac{(\nu+1)}{2}\log\left(1 + \frac{\widetilde{x}^2}{\nu}\right).$$

The resulting symmetric function can be verified to be maximized at $\widetilde{x}^2 = (\nu+1)\tau^2/\tau_q^2 - \nu$, provided this expression is positive, otherwise the only maximising root is at $\widetilde{x} = 0$ and so $\log \widetilde{w}(x, y) \leqslant 0$. If the expression is positive we obtain an upper bound

$$\log \widetilde{w}(x, y) \leqslant -\frac{1}{2}\left\{(\nu+1) - \nu\frac{\tau_q^2}{\tau^2}\right\} + \frac{(\nu+1)}{2}\log\left\{\frac{(\nu+1)}{\nu}\frac{\tau^2}{\tau_q^2}\right\}.$$

□

4.45 Corollary. *Under the conditions of Proposition 4.44 there exists a constant $K_3 < \infty$ such that*

$$\sup_{\theta \in B(\overline{\theta})} E^X\{\widetilde{w}(X, y)^a\} \leqslant K_3$$

independent of y.

Proof. The constant in Proposition 4.44 depends on θ only through τ. Moreover, the upper bound in Proposition 4.44 is continuous in τ and thus can be bounded over the compact set $B(\overline{\theta})$. □

We can summarize the results regarding the t-distribution in the following theorem.

4.46 Theorem. *Consider the random effects model* (4.11) *and assume we have an importance sampling estimator with proposal distribution*

$$q(x \mid y) = t_\nu(x \mid \widehat{x}, \tau_q^2)$$

with $\tau_q^2 > 0$. *Assume additionally that either*

i) $\sup_x \widetilde{A}'(x) < \infty$ *or*

ii) $E(Y^a) < \infty$ *and* $\sup_{\theta \in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

Then

$$E^Y \left[\sup_{\theta \in B(\bar{\theta})} E^{X|Y} \left\{ \overline{w}(Y, X, \theta)^a \right\} \right] < \infty.$$

Proof. We can bound

$$\begin{aligned} E^Y \left[\sup_{\theta \in B(\bar{\theta})} E^{X|Y} \left\{ \overline{w}(Y, X, \theta)^a \right\} \right] &= E^Y \left[\sup_{\theta \in B(\bar{\theta})} \frac{E^{X|Y} \left\{ \widetilde{w}(X, Y, \theta)^a \right\}}{E^{X|Y} \left\{ \widetilde{w}(X, Y, \theta) \right\}^a} \right] \\ &\leqslant E^Y \left[\sup_{\theta \in B(\bar{\theta})} \frac{K_3}{E^{X|Y} \left\{ \widetilde{w}(X, Y, \theta) \right\}^a} \right] < \infty. \end{aligned}$$

where the first inequality is by Corollary 4.45 and the second by Corollary 4.39. □

Theorem 5 and Theorem 6 provide simple and verifiable conditions for Assumption 4.4 to hold in the case of generalized linear mixed models when using a Gaussian proposal or a t-distribution. We have established these conditions by formulating assumptions on the models and the proposal. The assumptions that are required for the model are fulfilled in the Binomial and Poisson cases as pointed out in Remark 4.40. Gaussian proposals require that the variance is large enough, namely

$$\tau_q^2 > \frac{1 + \Delta}{2 + \Delta} \tau^2,$$

where $0 < \Delta < 1$ corresponds to the quantity in Assumption 4.4. When one proposes from a t-distribution instead, no such restriction is required.

4.12 Further Simulation Studies

4.12.1 Toy example

We consider first a simple Gaussian latent variable model where

$$X_t \sim \mathcal{N}(\theta, 1), \qquad Y_t \mid X_t = x \sim \mathcal{N}(x, 1).$$

Here $X_t, (t = 1, \ldots, T)$ are assumed to be independent. In this case, the likelihood associated to T observations can be computed exactly as $p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} \varphi(y_t; \theta, 2)$. This makes it

an easy example to examine Assumption 4.1. The maximum likelihood estimator and Fisher information are given by

$$\hat{\theta}_T^{\omega} = \frac{1}{T}\sum_{t=1}^{T} Y_t, \qquad I_T(\theta) = I_T = \frac{T}{2}.$$

If we assign a zero mean Gaussian prior to θ of variance σ_0^2 then the posterior is also normal with mean μ_{post} and variance σ_{post}^2 given by

$$\mu_{\text{post}} = \left(\frac{1}{\sigma_0^2} + \frac{T}{2}\right)^{-1}\left(\frac{\sum_{t=1}^{T} Y_t}{2}\right), \qquad \sigma_{\text{post}}^2 = \left(\frac{1}{\sigma_0^2} + \frac{T}{2}\right)^{-1}.$$

Assume the data are arising from the model with true parameter value $\bar{\theta}$. It follows readily from Pinsker's inequality that the Bernstein-von Mises theorem holds for $\Sigma = 2$ as we have as $T \to \infty$

$$\int \left|\pi_T^{\omega}(\theta) - \varphi\left(\theta; \hat{\theta}_T^{\omega}, I_T^{-1}\right)\right| \mathrm{d}\theta = \int \left|\varphi\left(\theta; \mu_{\text{post}}, \sigma_{\text{post}}^2\right) - \varphi\left(\theta, \hat{\theta}_T^{\omega}, \frac{2}{T}\right)\right| \mathrm{d}\theta \xrightarrow{\mathbb{P}^Y} 0.$$

Hence this model fulfils Assumption 1. To estimate the likelihood we simulate data from the model with $\bar{\theta} = 0.5$ and use $\sigma_0^2 = 10^{10}$. The likelihood is estimated using importance sampling

$$\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \frac{1}{N}\sum_{i=1}^{N} \varphi(y_t - U_{t,i}; \theta, 1), \quad U_{t,i} \sim \mathcal{N}(0,1).$$

In order to prove that Assumption 4.3 is fulfilled we show the stronger Assumption 4.4, i.e. for some $\Delta > 0$

$$E\left[\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U, \theta)^{2+\Delta}\right\}\right] = E\left[\sup_{\theta \in B(\bar{\theta})} E\left\{\frac{\varphi(y - U; \theta, 1)^{2+\Delta}}{\varphi(y; \theta, 2)^{2+\Delta}}\right\}\right] < \infty$$

In a first step we compute for $a > 0$

$$\begin{aligned} E\left\{\frac{\varphi(y - U; \theta, 1)^a}{\varphi(y; \theta, 2)^a}\right\} &= \left(\frac{2^{a-1}}{\pi}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{a(y - x - \theta)^2}{2} + \frac{a(y-\theta)^2}{4} - \frac{x^2}{2}\right) \mathrm{d}x \\ &= \left(\frac{2^{a-1}}{\pi}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{2a(y - x - \theta)^2 - a(y-\theta)^2 + 2x^2}{4}\right) \mathrm{d}x. \end{aligned}$$

Completing the square yields

$$2a(y - x - \theta)^2 - a(y - \theta)^2 + 2x^2$$

$$= 2(a+1)\left(x - \frac{a}{(a+1)}(y-\theta)\right)^2 - \frac{a\,(a-1)}{a+1}(y-\theta)^2$$

and

$$E\left\{\frac{\varphi(y-U;\theta,1)^a}{\varphi(y;\theta,2)^a}\right\} = \left(\frac{2^a}{a+1}\right)^{1/2} \exp\left\{\frac{a\,(a-1)}{4(a+1)}(y-\theta)^2\right\}.$$

We now consider

$$\sup_{\theta\in B(\bar{\theta})} E\left\{\overline{w}(y,U,\theta)^a\right\} = \left(\frac{2^a}{a+1}\right)^{1/2} \exp\left\{\frac{a\,(a-1)}{4(a+1)} \sup_{\theta\in B(\bar{\theta})}\{(y-\theta)^2\}\right\}.$$

Now let us write

$$\begin{aligned}(y-\theta)^2 &= (y-\bar{\theta}+\bar{\theta}-\theta)^2 \\ &= (y-\bar{\theta})^2 + 2(y-\bar{\theta})(\bar{\theta}-\theta) + (\bar{\theta}-\theta)^2\end{aligned}$$

and consider $\theta \in B(\overline{\theta})$ corresponding to $\left|\theta - \overline{\theta}\right| \leqslant \varepsilon$, where $\varepsilon > 0$. It is clear then that $(y-\theta)^2$ is optimised over $B(\overline{\theta})$ at either $\theta = \overline{\theta} + \varepsilon$ or $\theta = \overline{\theta} + \varepsilon$. Let us denote $y_D = y - \overline{\theta}$ and $d = \overline{\theta} - \theta$ for simplicity so that

$$(y-\theta)^2 = y_D^2 + 2y_D d + d^2,$$

Then we consider an upper bound on this which is quadratic in y_D as

$$(1+\alpha)y_D^2 + (1+\varepsilon^2),$$

where we need to determine α to achieve bounding for all values $|d| \leqslant \varepsilon$. By symmetry of the left-hand side, we need only consider the supremum case $d = \varepsilon$ so that

$$(1+\alpha)y_D^2 + (1+\varepsilon^2) \geqslant y_D^2 + 2y_D\varepsilon + \varepsilon^2,$$

in which case, examining the roots of the resulting quadratic in y_D, it is required that $4\varepsilon^2 - 4\alpha \leqslant 0$, so $\alpha \geqslant \varepsilon^2$. Taking $\alpha = \varepsilon^2$ and using the bounding quadratic expression we obtain,

$$\begin{aligned}\sup_{\theta\in B(\bar{\theta})} E[\overline{w}(y,U,\theta)^a] &= \left(\frac{2^a}{a+1}\right)^{\frac{1}{2}} \exp\left\{\frac{a(a-1)}{4(a+1)} \sup_{\theta\in B(\bar{\theta})}\{(y-\theta)^2\}\right\} \\ &\leqslant \left(\frac{2^a}{a+1}\right)^{\frac{1}{2}} \exp\left\{\frac{a(a-1)}{4(a+1)}(y-\overline{\theta})^2(1+\varepsilon^2) + \frac{a(a-1)}{4(a+1)}(1+\varepsilon^2)\right\}\end{aligned}$$

$$= g(y; a).$$

So finally it is required that

$$E_Y\{g(y; a)\} = \int_{-\infty}^{\infty} g(y; a)\varphi(y; \overline{\theta}, 2)\mathrm{d}y < \infty,$$

for $a = 2 + \Delta$ for some $\Delta > 0$. The above integral is finite when

$$\frac{a(a-1)}{(a+1)}(1+\varepsilon^2) < 1.$$

Hence with $a = \Delta + 2$,

$$\varepsilon^2 < \frac{(3+\Delta)}{(2+\Delta)(1+\Delta)} - 1,$$

with the right-hand side always positive provided $\Delta < \sqrt{2} - 1$.

We apply the pseudo-marginal method to this model to demonstrate how our result can approximate its characteristics. For the Markov chain, we use a random walk proposal with variance equal to the inverse Fisher information I_T^{-1} scaled by $\ell = 2$. For each T, we run a pseudo-marginal chain for various N to sample the posterior for 250000 iterations as well as the limit Markov chain of kernel $\tilde{P}_{\ell,\sigma}$. In Table 4.4 we summarize the simulations results. As expected, we find that both the average acceptance probability and the integrated autocorrelation time for $f(\theta) = \theta$ of the pseudo-marginal algorithm converge to those of the limiting Markov chain as T increases.

4.12.2 Stochastic Lotka-Volterra Model

Assumption 4.3 is difficult to verify in state space models. To illustrate the applicability of our results beyond latent variable models we investigate here a stochastic kinetic Lotka-Volterra model arising in systems biology. Such models are used to describe interacting species in a predator and prey setting. In particular we consider the model with transition equations given by

$$\begin{aligned}
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 1, X_{2,t+h} - X_{2,t} = 0 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_1 x_{1,t} + o(h) \\
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = -1, X_{2,t+h} - X_{2,t} = 1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_2 x_{1,t} x_{2,t} + o(h) \\
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 0, X_{2,t+h} - X_{2,t} = -1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_3 x_{2,t} + o(h),
\end{aligned}$$

where $X_{1,t}$ and $X_{2,t}$ denotes the number of preys and predators at time $t \in [0, T]$. This

Data T	Particles N	$\hat{\sigma}$	$\widehat{\text{IAT}}$	$\hat{\text{pr}}_{\text{acc}}$	$\widehat{\text{IAT}}\left(\tilde{P}_{\ell=2,\sigma=\hat{\sigma}}\right)$	$\hat{\text{pr}}_{\text{acc}}\left(\tilde{P}_{\ell=2,\sigma=\hat{\sigma}}\right)$
$T = 20$	6	1·70	17·55	18·69%	31·25	15·32%
	8	1·44	12·34	23·14%	17·62	20·27%
	10	1·24	10·76	26·34%	12·44	24·25%
	12	1·12	8·98	28·78%	10·02	27·19%
$T = 30$	8	1·83	27·70	15·41%	46·57	13·17%
	11	1·47	16·32	20·24%	18·64	19·61%
	14	1·30	12·04	24·03%	12·74	23·29%
	17	1·16	10·85	26·68%	9·91	26·09%
$T = 50$	20	1·85	30·46	13·94%	41·53	13·10%
	30	1·48	18·59	19·58%	17·53	19·51%
	40	1·29	13·30	23·59%	11·63	23·34%
	50	1·16	10·51	26·86%	9·91	26·09%
$T = 100$	20	1·86	34·64	13·01%	41·04	12·81%
	30	1·51	17·98	19·15%	18·73	18·93%
	40	1·32	14·56	23·15%	13·59	22·99%
	50	1·16	10·51	26·33%	9·91	26·09%
$T = 200$	80	1·83	38·35	13·11%	46·57	13·17%
	120	1·52	20·65	18·90%	20·42	18·58%
	160	1·30	13·87	22·94%	12·74	23·29%
	200	1·17	11·15	26·07%	9·73	26·05%

Table 4.4: For T data and N particles: standard deviation $\hat{\sigma}$ of the log-likelihood estimator at $\bar{\theta}$, integrated autocorrelation time $\hat{\tau}$ and average acceptance probability $\hat{p}_{\text{acc}}$ for pseudo-marginal kernel with $\ell = 2$ and limiting kernel $\tilde{P}_{\ell=2,\hat{\sigma}}$.

model has been previously investigated, for example in (Andrieu, Doucet and Holenstein 2009) and (Wilkinson 2012). We assume independent gamma priors for the kinetic rate parameter vector $\beta = (\beta_1, \beta_2, \beta_3)$ with

$$\beta_1 \sim \Gamma(5, 5), \quad \beta_2 \sim \Gamma(1{\cdot}5, 10), \quad \beta_3 \sim \Gamma(3{\cdot}5, 5).$$

In our simulations we assume we are only able to observe predator and prey $X_t = (X_{1,t}, X_{2,t})$ at discrete equidistant time points with independent measurement error $Y_{i,t} = X_{i,t} + W_{i,t}$, $i = 1, 2$, $t = 0, \ldots, 50$ where $W_{i,t} \sim \mathcal{N}(0, 10^2)$. The artificial data have been generated using the Gillespie algorithm (Gillespie 1977) for the rate constants $\beta = (1, 0{\cdot}005, 0{\cdot}6)$.

In this context, it is difficult to develop standard MCMC algorithms to sample the posterior distribution while the pseudo-marginal algorithm can be easily applied as an unbiased estimate of the likelihood can be computed using a bootstrap particle filter; see, e.g., (Andrieu, Doucet and Holenstein 2009) and (Wilkinson 2012, Chapter 10). We use a multivariate Gaussian random walk proposal with scaling factor $\ell = 2{\cdot}17$ and covariance matrix close to the posterior covariance, which we estimated in a short preliminary run. This can efficiently implemented in R (R Core Team 2017) using the package **smfsb** (Wilkinson 2012) and the example code which can be found on the author's blog.

The algorithm is then run for 250000 iterations. We collect acceptance rate and computing time $\mathrm{CT}(N) = \mathrm{IAT}(N) \cdot N$ for a range of particles N, see Table 4.5. In practice we do not choose $\sigma\left(\bar{\theta}\right)$, but the number of particles, N, which is also displayed in Table 4.5. For comparison we also give an estimate of $\sigma\left(\bar{\theta}\right)$ for given N.

The computing time is optimized at $N = 225$ for all rates, β_1, β_2 and β_3. We estimate $\sigma\left(\bar{\theta}\right)$ to be $1{\cdot}44$, slightly above the results of Table 4.1 suggesting $\sigma = 1{\cdot}24$. The corresponding acceptance rate of $18{\cdot}57\%$ is in accordance with the one suggested by our theory, which for parameter dimension $d = 3$ yields an asymptotically optimal rate of around $19{\cdot}30\%$ ($\ell = 2{\cdot}17, \sigma = 1{\cdot}24$). We conjecture that the deviation from the results obtained in the limiting case are due to the fact that the posterior is not very concentrated around $\bar{\theta}$.

Sherlock et al. (2015) carry out Bayesian inference for a 5-dimensional stochastic Lotka-Volterra model using the pseudo-marginal algorithm based on a data set with $T = 50$ observations. The authors optimize over a grid of values for both σ and ℓ. Experimentally, it was found that the optimal standard deviation was $\sigma \approx 1{\cdot}45$ and the optimal tuning for the random walk achieved at $\ell = 2{\cdot}048$ with an associated optimal jumping rate of $15{\cdot}39\%$. This is slightly above our guidelines with the values $\hat{\sigma}_{\mathrm{opt}} = 1{\cdot}30$, $\hat{\ell}_{\mathrm{opt}} = 2{\cdot}17$ and $\mathrm{pr}_{\mathrm{acc}}(\hat{\sigma}_{\mathrm{opt}}, \hat{\ell}_{\mathrm{opt}}) = 17{\cdot}35\%$ obtained in Table 4.1.

Particles N	Acceptance Rate	CT(β_1)	CT(β_2)	CT(β_3)	$\hat{\sigma}(\bar{\theta})$
100	8·92%	7375	9035	7564	2·38
125	11·17%	6668	6717	6580	2·10
150	13·44%	5805	5903	6208	1·84
175	15·62%	5688	6137	6101	1·68
200	17·03%	5564	5632	5744	1·55
225	18·57%	5178	5452	5122	1·44
250	19·54%	6107	6958	5831	1·36
275	20·82%	5473	6087	5248	1·30
300	21·47%	6436	6340	5959	1·22
325	22·41%	5771	6586	6178	1·19
350	23·20%	6406	6234	6393	1·13

Table 4.5: Comparison of the computing time for different numbers of particles in the stochastic Lotka-Volterra model.

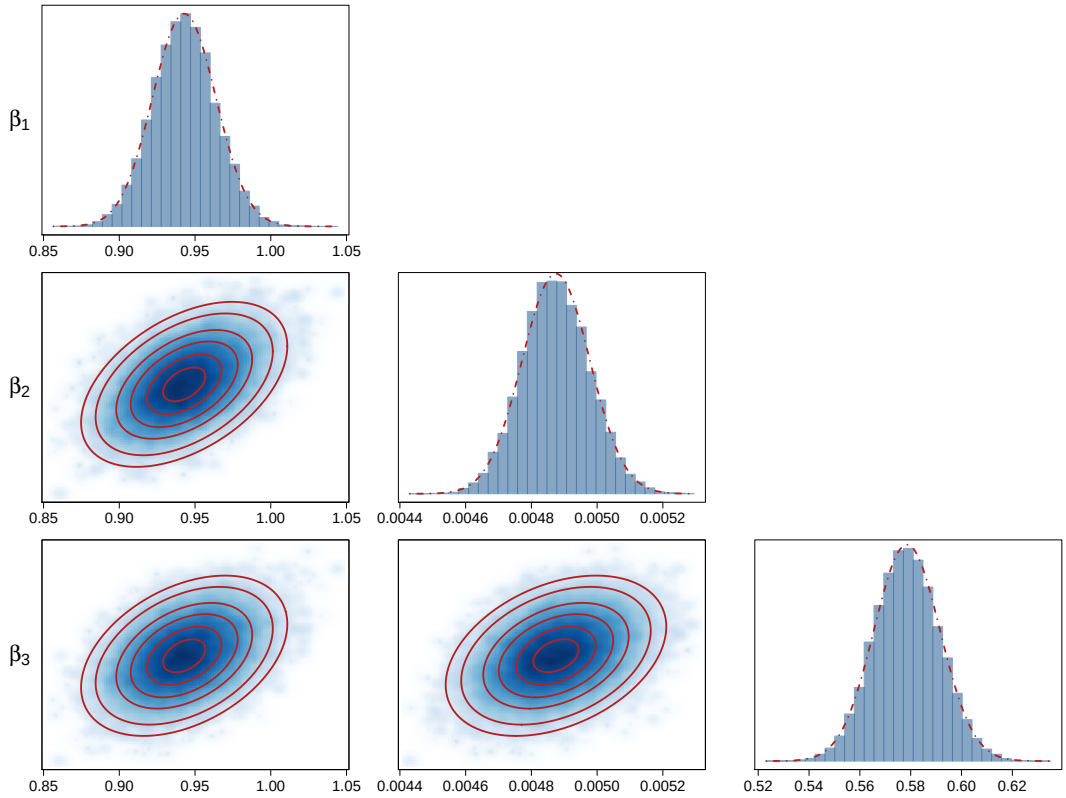


Figure 4.3: Histogram of marginal posterior $p(\beta_i \mid y_{1:T})$, $i = 1, 2, 3$ on the diagonal with Gaussian approximation (line) using sample mean and variance. In addition, we show density estimates of the projections to the plane. The ellipses indicate the contour lines of a Gaussian with sample mean and sample covariance matrix. It is clear from the plots that the posterior is very close to a Gaussian.

References

Andrieu, Christophe, Arnaud Doucet and Roman Holenstein (2009). 'Particle Markov chain Monte Carlo for efficient numerical simulation'. *Monte Carlo and quasi-Monte Carlo methods 2008*, pp. 45–60 (cit. on p. 133).

Bahr, Bengt von and Carl-Gustav Esseen (1965). 'Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leqq r \leqq 2$'. *Ann. Math. Statist.* 36.1, pp. 299–303. DOI: 10.1214/aoms/1177700291. URL: https://doi.org/10.1214/aoms/1177700291 (cit. on p. 96).

Berti, Patrizia, Luca Pratelli and Pietro Rigo (2006). 'Almost sure weak convergence of random probability measures'. *Stochastics* 78.2, pp. 91–97 (cit. on pp. 73, 75).

Billingsley, Patrick (1999). *Convergence of Probability Measures*. John Wiley & Sons (cit. on p. 78).

Birnbaum, Z. W. (1942). 'An Inequality for Mill's Ratio'. *Ann. Math. Statist.* 13.2, pp. 245–246. DOI: 10.1214/aoms/1177731611. URL: https://doi.org/10.1214/aoms/1177731611 (cit. on p. 120).

Borkar, Vivek S (1991). *Topics in Controlled Markov Chains*. Vol. 20. Longman Scientific & Technical UK (cit. on p. 74).

Castillo, Ismaël and Judith Rousseau (2015). 'Supplement to "A Bernstein–von Mises theorem for smooth functionals in semiparametric models."' DOI:10.1214/ 15-AOS1336SUPP (cit. on p. 77).

Chen, Louis HY, Larry Goldstein and Qi-Man Shao (2010). *Normal Approximation by Stein's Method*. Springer Science & Business Media (cit. on p. 93).

Crauel, Hans (2003). *Random Probability Measures on Polish Spaces*. CRC Press (cit. on pp. 71–72, 79).

Deligiannidis, George, Arnaud Doucet and Michael K Pitt (2018). 'The correlated pseudo-marginal method'. *J. R. Statist. Soc.* B 80.5, pp. 839–870 (cit. on pp. 91–92, 105).

Dudley, Richard M (2002). *Real Analysis and Probability*. Vol. 74. Cambridge University Press (cit. on pp. 71, 74).

Ethier, Stewart N and Thomas G Kurtz (2005). *Markov Processes: Characterization and Convergence*. Vol. 282. John Wiley & Sons (cit. on pp. 77–79).

Gillespie, D T (1977). 'Exact stochastic simulation of coupled chemical reactions.' *J. Phys. Chem.* 81.25, pp. 2340–2361 (cit. on p. 133).

Jennrich, R. I. (1969). 'Asymptotic properties of non-linear least squares estimators'. *Ann. Math. Statist.* 40.2, pp. 633–643 (cit. on p. 93).

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer-Verlag: New York (cit. on pp. 72, 79, 103).

Klenke, Achim (2013). *Probability Theory: a Comprehensive Course*. Springer Science & Business Media (cit. on p. 79).

Labouriau, Rodrigo (2014). 'A note on the identifiability of generalized linear mixed models'. *arXiv preprint arXiv:1405.0673* (cit. on p. 118).

McCulloch, Charles E and John M Neuhaus (2005). 'Generalized linear mixed models'. *Encyclopedia of Biostatistics* 4 (cit. on p. 109).

Mukerjee, Rahul and Brajendra C Sutradhar (2002). 'On the positive definiteness of the information matrix under the binary and Poisson mixed models'. *Ann. Instit. Statist. Math.* 54.2, pp. 355–366 (cit. on p. 117).

Nickl, Richard (2012). 'Statistical Theory'. *Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge* (cit. on pp. 111, 117).

Owen, Art B. (2013). *Monte Carlo Theory, Methods and Examples* (cit. on p. 125).

Pollard, David (2002). *A User's Guide to Measure Theoretic Probability*. Vol. 8. Cambridge University Press (cit. on p. 71).

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/` (cit. on p. 133).

Sherlock, Chris, Alexandre H Thiery, Gareth O Roberts, Jeffrey S Rosenthal et al. (2015). 'On the efficiency of pseudo-marginal random walk Metropolis algorithms'. *Ann. Statist.* 43.1, pp. 238–275 (cit. on p. 133).

Sweeting, TJ (1989). 'On conditional weak convergence'. *J. Theoret. Probab.* 2.4, pp. 461–474 (cit. on pp. 73, 75, 77–78).

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Cambridge University Press (cit. on pp. 111, 113, 115–117).

Wilkinson, D. J. (2012). *Stochastic Modelling for Systems Biology*. CRC Press (cit. on p. 133).

5

Epilogue: Large Sample Considerations for MCMC Algorithms

In this section we consider some further areas of research that follow naturally from the ideas introduced above. In doing so, we will start by considering the pseudo-marginal algorithm but we will also discuss other Markov chain Monte Carlo algorithms with noisy acceptance probabilities and implications of large sample asymptotics.

5.1 Large Sample Properties of the Pseudo-Marginal Algorithm

In the previous chapter we have shown—either empirically or by virtue of Theorem 4.1—that the properties of the limiting algorithm, e.g. the integrated autocorrelation time and the acceptance probability, can be well approximated as $T \to \infty$ by the respective characteristics of the limiting kernel started from stationarity.

A natural continuation of this analysis is to consider the question whether other characteristics, like qualitative convergence properties, can be studied by considering the limiting algorithm. It is easy to see, for example, that the limiting kernel of the pseudo-marginal algorithm is not geometrically ergodic. Too see this, note that while the classical random walk Metropolis algorithm targeting $\varphi(\theta, 0, \Sigma)$ with proposal $q(\theta, \theta') = \varphi(\mathrm{d}\theta', \theta, \ell^2\Sigma/d)$ is geometrically ergodic (see e.g. Jarner and Hansen 2000), it follows from Proposition 2.15 (see also Andrieu and Vihola 2015, Proposition 13) that the pseudo-marginal kernel $\tilde{P}$ is not as the support of e^Z when $Z \sim \varphi\left(\,\cdot\,; -\sigma^2/2, \sigma^2\right)$ is not bounded above and thus $\tilde{P}$ cannot have a spectral gap. By Theorem 2.8 this means that the limiting pseudo-marginal chain can also not be geometrically ergodic.

The question following this argument is how well this heuristic describes practical implementations of pseudo-marginal algorithms as we have not formally shown that convergence

properties, like geometric ergodicity, are inherited when passing through the limit $T \to \infty$. One possible way to establish this could be to consider conditions for the convergence of the spectral gap as $T \to \infty$. However, as can be seen from the presentation in Section 2.4 and Proposition 3.4 this is equivalent to showing convergence of $\tilde{P}_T \to \tilde{P}$ as $T \to \infty$ in the norm topology. Hence, proving such a result would require to prove uniform convergence of a large class of functions which is more challenging than the result in Section 4.9.

An inspection of the limiting transition kernel and the stationary distribution also opens up new questions about the asymptotic behaviour of pseudo-marginal algorithms in terms of the dimension, d. We have investigated the pseudo-marginal scheme for $T \to \infty$ for fixed parameter dimension d. However, in practice, increasing the number of data often means that more complex models with more parameters can be employed. Hence, an interesting avenue of research would be to consider the case where $d \to \infty$ as $T \to \infty$. Respective Bernstein–von Mises type theorems with growing parameter dimensions do exist, see e.g. Belloni and Chernozhukov (2009) and Ghosal (2000). Future research in this direction could include an attempt to use these results in order to merge the large sample asymptotics (Section 3.3 and Chapter 4) with the diffusion limit of Roberts, Gelman and Gilks (1997) and Sherlock et al. (2015). For the case of the random walk Metropolis–Hastings algorithm Belloni and Chernozhukov (2009) derive complexity results in such an asymptotic framework.

5.2 The Correlated Pseudo-Marginal Algorithm

Motivated by the inefficiency of the limiting kernel (i.e. the nonexistence of a spectral gap) we want to discuss modifications and improvements to the pseudo-marginal algorithm. We begin by pointing out the sources of some of its weaknesses. Firstly, replacing the likelihood with an estimate produces a noisy version of the same quantity. If the estimator is not well controlled, the density could potentially be overestimated by a large amount. Since the estimate of the density is retained for the next iteration, subsequent proposals are very unlikely to be accepted. This leads to a well known behaviour of the Markov chain 'getting stuck'. Indeed, we have already seen that pseudo-marginal chains fail to inherit geometric ergodicity of a marginal algorithm when the noise in the estimator is not bounded (Proposition 2.15). The same argument also lead us to the conclusion that the limiting kernel (as $T \to \infty$) does not produce a geometrically ergodic Markov chain. In search for a remedy for this shortcoming, let us investigate again the way that the pseudo-marginal algorithm is constructed. In the standard implementation, the likelihood estimates in the Metropolis-Hastings ratio

$$\frac{\hat{\pi}(\theta')}{\hat{\pi}(\theta)} = \frac{\hat{p}(y \mid \theta', u')}{\hat{p}(y \mid \theta, u)} \frac{p(\theta')}{p(\theta)}$$

are constructed by drawing independent random variables u, u' to compute the estimators $\hat{p}(y \mid \theta', u')$ and $\hat{p}(y \mid \theta, u)$. As established before this corresponds to a standard Metropolis-Hastings algorithm on an extended state space with acceptance ratio

$$r(\theta, u; \theta', u') = \frac{p(\theta' \mid y, u')m(u')}{p(\theta \mid y, u)m(u)} \frac{q(\theta', \theta)m(u)}{q(\theta, \theta')m(u')},$$

where m denotes the density of the auxiliary random variables drawn in every iteration. Note that while $q(\theta, \theta')$ is a random walk proposal, the auxiliary variables behave like an independent sampler. However, using the formulation above it is clear, that $m(u)$ can be replaced by any m-reversible proposal, $k(u, u')$, say, without loosing the exactness of the algorithm. But how should k be chosen to achieve faster convergence? Consider again the 'stickiness' problem of the pseudo-marginal chain. When the variables drawn result in an estimate with a large weight, subsequent proposals are very unlikely to be accepted. However, we can 'slow down' the mixing of the MCMC sampler in the auxiliary space. The auxiliary variables are not our target for inference and slow movement in the auxiliary space means that the variance of the likelihood ratio estimate $\hat{p}(y \mid \theta', u') \big/ \hat{p}(y \mid \theta, u)$ is reduced, provided that the model is smooth in u and θ. This idea was first introduced by Deligiannidis, Doucet and Pitt (2018) and the resulting algorithm is termed the *correlated pseudo-marginal algorithm*. In practice, this algorithm is implemented by choosing m to be Gaussian. This can be done without loss of generality since in most computers random variables are generated from uniform distributions. To switch to Gaussian random variables simply draw $U \sim \mathcal{N}(0, 1)$, then $\Phi(U)$, where Φ is the standard normal cumulative distribution function, is uniformly distributed. We can then select $k(u, u')$ as a stationary Gaussian AR(1)-process defined by

$$u' = \varrho u + \sqrt{1 - \varrho^2} U, \qquad U \sim \mathcal{N}(0, 1)$$

with $0 \leqslant \varrho < 1$ the correlation coefficient between the Gaussian distributions. With the exception of the correlated proposal for the latent states, the algorithm is identical to the pseudo-marginal approach presented in Section 2.6. The authors show in experiments that high values of ϱ greatly improve the performance of pseudo-marginal schemes. As in the case of the pseudo-marginal algorithm an efficient implementation requires to select free tuning parameters. Similar to our approach in the previous section Deligiannidis, Doucet and Pitt (2018) derive a limiting kernel that describes the algorithm as $T \to \infty$ and transition kernel (in the notation of Chapter 4)

$$\tilde{P}(\tilde{\theta}; \mathrm{d}\tilde{\theta}') = \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\alpha(\tilde{\theta}, \tilde{\theta}') + \rho(\tilde{\theta})\delta_{(\tilde{\theta})}(\mathrm{d}\tilde{\theta}')$$

Dimension d	$\hat{\ell}_{\text{opt}}$	$\hat{\sigma}_{\text{opt}}$	$\text{CT}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$	$\text{pr}_{\text{acc}}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$
$d = 1$	2.34 (0.21)	2.29 (0.12)	2.51	15.00%
$d = 2$	2.37 (0.09)	2.46 (0.14)	3.77	11.36%
$d = 3$	2.54 (0.17)	2.54 (0.13)	4.98	9.27%
$d = 5$	2.48 (0.09)	2.48 (0.09)	7.56	9.26%
$d = 10$	2.55 (0.10)	2.69 (0.12)	13.81	6.92%
$d = 15$	2.56 (0.12)	2.77 (0.11)	12.49	6.54%
$d = 20$	2.61 (0.12)	2.92 (0.12)	12.29	5.59%

Table 5.1: Optimal values for scaling ℓ and noise σ and associated value of computing time and average acceptance probability: mean and standard deviation of the minimizers over 10 runs.

with

$$\alpha(\tilde{\theta}, \tilde{\theta}') = \int \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} e^{w}\right\} \varphi\left(\mathrm{d}w; -\kappa^2/2, \kappa^2\right),$$

where $\kappa = \kappa(\bar{\theta})$ (see e.g. Deligiannidis, Doucet and Pitt 2018, Theorem 4) and $\rho(\theta)$ is the corresponding rejection probability.

Deligiannidis, Doucet and Pitt (2018) consider the limiting kernel to theoretically justify the assumption that the noise in the logarithm of the likelihood estimator follows a log-normal distribution

$$Z \sim \mathcal{N}\left(-\frac{\kappa^2(\bar{\theta})}{2}, \kappa^2(\bar{\theta})\right)$$

independent of the current value of the chain. Under this assumption the authors derive an upper bound on the computing time that is subsequently optimized. In contrast, we will consider the tuning of the limiting kernel directly. We proceed analogously to Section 4.6.

Table 5.1 shows the optimal values for ℓ, σ on the grid for a selected values of the parameter dimension d. We can see that the optimal noise is around $\sigma = 2.3$ for $d = 1$ and increases continuously to almost $\sigma = 3$. Similarly, the optimal scaling value ℓ increases slowly from $\ell = 2.34$ to $\ell = 2.61$ for $d = 20$. It can be observed that these values keep the computing time relatively low, i.e. the increase in $\text{CT}(\ell, \sigma)$ as d grows is not as steep as in Table 4.1.

Comparing the results for the correlated pseudo-marginal algorithms to those of the standard pseudo-marginal algorithm in Table 4.1 (Section 4.6) we observe that correlating the auxiliary variables leads to drastic improvements. First, note that the optimal noise is higher (around double) than in the standard pseudo-marginal case while achieving a lower integrated autocorrelation time. This results in the correlated pseudo-marginal method achieving a much lower computing time than the standard implementation.

5.1 Remark. *Weak convergence results like the one in Chapter 4 for the pseudo-marginal algorithm and the result by Deligiannidis, Doucet and Pitt (2018) for the correlated pseudo-marginal algorithm do not necessarily imply convergence of the integrated autocorrelation time. Indeed, Deligiannidis, Doucet and Pitt (2018, Section 4.3) demonstrate that the number of particles, N, should grow at least at rate $\sqrt{T}$ to ensure that the integrated autocorrelation time remains bounded. In our simulation studies we found that the limiting kernel serves as a good approximation for the behaviour of the (correlated) pseudo-marginal method for values of N in the range of interest whereas the approximation becomes worse when N is chosen very small, see e.g. Figure 4.1.*

References

Andrieu, Christophe and Matti Vihola (2015). 'Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms'. *Ann. Appl. Probab.* 25.2, pp. 1030–1077 (cit. on p. 139).

Belloni, Alexandre and Victor Chernozhukov (2009). 'On the computational complexity of MCMC-based estimators in large samples'. *The Annals of Statistics* 37.4, pp. 2011–2055 (cit. on p. 140).

Deligiannidis, George, Arnaud Doucet and Michael K Pitt (2018). 'The correlated pseudo-marginal method'. *J. R. Statist. Soc.* B 80.5, pp. 839–870 (cit. on pp. 141–143).

Ghosal, Subhashis (2000). 'Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity'. *Journal of Multivariate Analysis* 74.1, pp. 49–68 (cit. on p. 140).

Jarner, Søren Fiig and Ernst Hansen (2000). 'Geometric ergodicity of Metropolis algorithms'. *Stochastic processes and their applications* 85.2, pp. 341–361 (cit. on p. 139).

Roberts, G.O., A. Gelman and W.R. Gilks (1997). 'Weak convergence and optimal scaling of random walk Metropolis algorithms.' *Ann. Appl. Probab.* 7, pp. 110–120 (cit. on p. 140).

Sherlock, Chris, Alexandre H Thiery, Gareth O Roberts, Jeffrey S Rosenthal et al. (2015). 'On the efficiency of pseudo-marginal random walk Metropolis algorithms'. *Ann. Statist.* 43.1, pp. 238–275 (cit. on p. 140).

Part III

Particle Filters with Intractable Weights

6

Inference for Hidden Markov Models

In the previous parts of this thesis we have mostly considered independent and identically distributed (*i.i.d*) data. Such assumptions are very popular since they often make mathematical analysis tractable or even just possible[1].

In this section we want to consider a model for dependent data. More precisely, we want to consider so-called *hidden Markov models*. Such models evolve according to a latent Markov chain $(X_t)_{t\in\mathbb{N}}$ taking value in a space X defined through the state transition distribution

$$X_t \mid (X_{t-1} = x) \sim f(\cdot \mid x) \quad \text{with} \quad t \geqslant 2 \quad \text{and } X_1 \sim \mu,$$

but are only observed through conditionally independent state observations $(Y_t)_{t\in\mathbb{N}}$ on Y with observation density

$$Y_t \mid (X_t = x) \sim g(\cdot \mid x) \quad t \geqslant 1.$$

This is a generalization of the *i.i.d*. model in Example 3.5 and Section 4.5 and is visualized in Figure 6.1. In contrast to the previous models, however, here we will not consider inference over some parameter, θ. Indeed, in many situation and due to the nature of the model interest lies in finding the (distribution of the) hidden state variables $X_{1:T}$ given some finite set of observations $Y_{1:T}$. This model also offers a Bayesian interpretations: define the prior

$$f(x_{1:T}) = \mu(x_1) \prod_{t=2}^{T} f(x_i \mid x_{i-1})$$

[1] In Section 4.5, for example, the independence of the model is crucial for our proof to work. No such result has been established yet for dependent data.

and the likelihood function

$$g(y_{1:T} \mid x_{1:T}) = \prod_{t=1}^{T} g(y_t \mid x_t).$$

The Bayesian posterior is then given by

$$p(x_{1:T} \mid y_{1:T}) = \frac{g(y_{1:T} \mid x_{1:T}) f(x_{1:T})}{p(y_{1:T})}$$

where

$$p(y_{1:T}) = \int g(y_{1:T} \mid x_{1:T}) f(x_{1:T}) \mathrm{d}x_{1:T}.$$

In general non-Gaussian and non-linear models the resulting posterior distributions are not tractable and cannot be computed analytically. Of course, (approximate) sampling from $p(x_{1:T} \mid y_{1:T})$ can be achieved through Markov chain Monte Carlo approaches as outlined in Section 2. However, due to their definition hidden Markov models are often used to model sequential data. Indeed, often only the distribution of the last state $p(x_T \mid y_{1:T})$ or an earlier state $p(x_t \mid y_{1:T}), 0 \leqslant t < T$ is of interest at all. In such cases observing a new state, y_{T+1}, means that Markov chain approaches have to restart. An algorithm that could be updated sequentially whenever new data is observed would be more desirable.

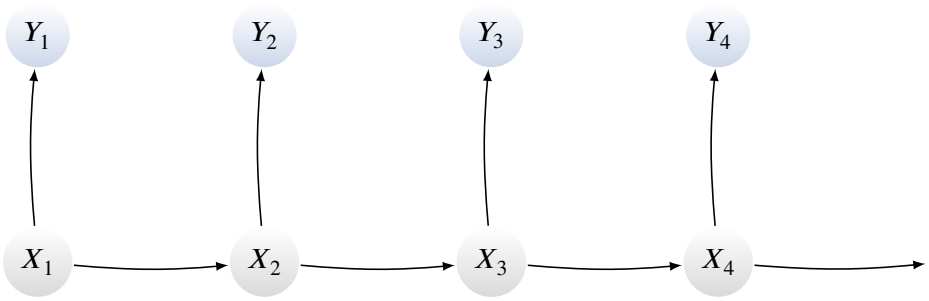


Figure 6.1: Example of a graph of a hidden Markov model. $Y_1, Y_2, \ldots$ (blue) are observed whereas $X_1, X_2, \ldots$ (grey) are hidden.

In the following we will consider sequential approaches using importance sampling that are efficient for online inference. The first approach is to use simple importance sampling and update the weights every time a value is observed.

6.1 Sequential Importance Sampling

As for many Bayesian settings, we are interested in computing expectations with respect to the posterior distribution

$$\mathcal{I}(h) = \int h(x_{1:t}) p(x_{1:t} \mid y_{1:t}) \mathrm{d}x_{1:t}$$

for $1 \leqslant t \leqslant T$.

Let us now consider an inference scheme that takes the sequential nature of the data into account, i.e. we assume that we first observe y_1, approximate $p(x_1 \mid y_1)$ and as we receive a new observation y_2 we update our inference scheme to approximate $p(x_{1:2} \mid y_{1:2})$.

Estimating the integrals with respect to the initial filtering distribution $p(x_1 \mid y_1)$ can be carried out with standard importance sampling estimators (e.g. Owen 2013, Chapter 9). Using a proposal distribution $X_1^i \sim q(\cdot \mid y_1), i = 1, \ldots, N$ we have for any test function $h : \mathsf{X} \to \mathbb{R}$

$$\hat{\mathcal{I}}_{\mathrm{PF}}(h) = \frac{\sum_{i=1}^N h(X_1^i) w_1(X_1^i)}{\sum_{i=1}^N w_1(X_1^i)}, \quad w_1(X_1^i) = \frac{g(y_1 \mid X_1^i)\mu(X_1^i)}{q(X_1^i \mid y_1)},$$

where we denote N the number of samples or *particles*. By the law large numbers $\hat{\mathcal{I}}_{\mathrm{PF}}(h) \to \mathcal{I}(h)$ almost surely as the number of particles $N \to \infty$ whenever $E\{|h(X)|w(X)\} < \infty$ for $X \sim p(x_1 \mid y_1)$. If now a second observation y_2 arises the joint importance weight is

$$\begin{aligned} w_{1:2}(X_1^i, X_2^i) &= \frac{g(y_1 \mid X_1^i)\mu(X_1^i)}{q(X_1^i \mid y_1)} \frac{g(y_2 \mid X_2^i) f(X_2^i \mid X_1^i)}{q(X_2^i \mid X_1^i, y_2)} \\ &= w_1(X_1^i) \cdot w_2(X_2^i). \end{aligned}$$

This factorization suggests that we can adjust the importance sampling weights when new observations arrive by just multiplying the marginal importance weight of the newest observation. Indeed, for general $1 < t \leqslant T$ we can always write

$$\begin{aligned} w_{1:t}(x_{1:t}) &\propto \frac{p(x_{1:t} \mid y_{1:t})}{q(x_{1:t})} \\ &\propto \frac{p(x_{1:(t-1)} \mid y_{1:(t-1)})}{q(x_{1:(t-1)})} \frac{g(y_t \mid x_t) f(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)} \\ &\propto w_{1:(t-1)}(x_{1:(t-1)}) w_t(x_t). \end{aligned}$$

The respective algorithm is called *sequential importance sampling*. Of course, the performance of this algorithm crucially depends on the characteristics of the proposal distribution

$q(x_t \mid x_{t-1}, y_t)$. The following result provides a guideline for the choice of the proposal.

6.1 Proposition. *The proposal $q^*(x_t \mid x_{t-1}, y_t)$ minimizing the one-step variance $\mathrm{Var}_q\left\{w(X_{1:t}) \mid X_{1:(t-1)}\right\}$ is given by*

$$q^*(x_t \mid x_{t-1}, y_t) = \frac{g(y_t \mid x_t) f(x_t \mid x_{t-1})}{p(y_t \mid x_t)}$$

where

$$p(y_t \mid x_{t-1}) = \int g(y_t \mid x_t) f(x_t \mid x_{t-1}) \mathrm{d}x_t$$

and the optimal weight is

$$w_t^*(x_{t-1}, x_t) = p(y_t \mid x_{t-1}).$$

Proof. See, for example, Doucet, Godsill and Andrieu (2000, Proposition 2). □

Unfortunately, the integral $p(y_t \mid x_{t-1}) = \int g(y_t \mid x_t) f(x_t \mid x_{t-1}) \mathrm{d}x_t$ is intractable in most cases. This intractability of the weights in the locally optimal proposal q^* prohibits the exact implementation of this algorithm, but it can provide a guide for good proposal strategies.

However, instead of prematurely optimizing this algorithm we want to point out that the sequential importance sampling scheme suffers from systemic flaws. The product nature means that many particles will have decaying weights (as they mostly consist of products of small positive numbers) and hence often one path $(X^i_{1:t}), t \leqslant T$ will dominate in terms of its weight. Practically, we will observe that very often the variance of the normalized weights

$$\mathrm{Var}\left\{\frac{w_t(x_{1:t})}{p(y_{1:t})}\right\} \lesssim k^t$$

for some constant $k > 1$ (see e.g. Doucet and Johansen 2009). *Resampling* is a popular approach to counter this problem and we will introduce this procedure now.

6.2 Sequential Importance Resampling

When a new state is proposed it might not align well with the observations and thus it might have very low importance sampling weight. The idea of resampling is to concentrate on the promising particles measured in terms of the computed weights. To this end, we need to introduce some new notation. We will now write $\widetilde{X}^i_t$ for the weighted sample, i.e. $\widetilde{X}^i_t \sim q(\cdot \mid X^i_{t-1}, y_t)$ and X^i_t the particles after resampling is performed.

Underlying the idea of resampling is the understanding that the set $\{w_i, \widetilde{X}^i_{1:T}; i = 1, \dots, N\}$ is a weighted approximation of $p(x_{1:T} \mid y_{1:T})$. In order to circumvent the decay of the weights

we wish to obtain an unweighted approximation $\{X_{1:T}^i; i = 1, \dots, N\}$. This can be achieved by using the weights as a sampling distribution and sample from the weights with replacement.

The algorithm proceeds as follows. After the first observation, y_1, is known we propose a set of particles $\widetilde{X}_1^i \sim q(\cdot \mid y_1), i = 1, \dots, N$ and subsequently compute the importance sampling weights

$$w_1(\widetilde{X}_1^i) = \frac{g(y_1 \mid \widetilde{X}_1^i)\mu(\widetilde{X}_1^i)}{q(\widetilde{X}_1^i \mid y_1)}, \quad i = 1, \dots, N.$$

However, in order to obtain an unweighted sample we perform the aforementioned resampling step. Sample

$$I_{1:N} \sim \mathsf{Mult}\left\{N; w_{1,1}, \dots, w_{1,N}\right\},$$

where we write $\mathsf{Mult}\left\{N; w_{1,1}, \dots, w_{1,N}\right\}$ to denote N independent samples from a categorical or multinomial distribution with weights proportional to $w_{1,1}, \dots, w_{1,N}$. Setting $X_1^i = \widetilde{X}_1^{I_i}, i = 1, \dots, N$ thus yields an approximate sample from $p(x_1 \mid y_1)$. This process is repeated every time we observe a new observation y_t; first we resample the current trajectory according to its weights $(w_{t-1,1}, \dots, w_{t-1,N})$, then we propose new values according a proposal distribution $\widetilde{X}_t^i \sim q(\cdot \mid X_{t-1}^i, y_t)$ and compute the next weights

$$w_t(X_{t-1}^i, \widetilde{X}_t^i) = \frac{g(\widetilde{X}_t^i \mid y_t) f(\widetilde{X}_t^i \mid X_{t-1}^i)}{q(\widetilde{X}_t^i \mid X_{t-1}^i, y_t)}.$$

The resulting algorithm is called a *particle filter* and is described briefly in the next chapter and is detailed in Algorithm 7.1. The optimal proposal is again the same as for sequential importance sampling (Proposition 6.1) and can be used to guide the proposal choice.

Of course, many extensions are possible. For example, the resampling step does not have to be carried out every iteration. A common approach is to monitor the *effective sample size* of the current particles and to trigger resampling as soon as this quantity decays below a certain pre-set threshold. We refer the reader to Doucet and Johansen (2009) and references therein. However, the majority of research in particle filters focuses on cases where the transition and observation densities are available analytically. In the following chapter, we want to consider a different direction and consider cases where the weights $w_t, 1 \leqslant t \leqslant T$ are not available analytically. We will demonstrate that resampling in those cases is still possible using so-called *Bernoulli races*, which are related to Bernoulli factories (see e.g. Keane and O'Brien 1994) and have been introduced in a simple form by Dughmi et al. (2017). The resulting article (Schmon, Deligiannidis and Doucet 2019) was published in the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019.

References

Doucet, Arnaud, Simon Godsill and Christophe Andrieu (2000). 'On sequential Monte Carlo sampling methods for Bayesian filtering'. *Statistics and computing* 10.3, pp. 197–208 (cit. on p. 150).

Doucet, Arnaud and Adam M Johansen (2009). 'A tutorial on particle filtering and smoothing: Fifteen years later'. *Handbook of nonlinear filtering* 12.656-704, p. 3 (cit. on pp. 150–151).

Dughmi, Shaddin, Jason D Hartline, Robert Kleinberg and Rad Niazadeh (2017). 'Bernoulli factories and black-box reductions in mechanism design'. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 158–169 (cit. on p. 151).

Keane, MS and George L O'Brien (1994). 'A Bernoulli factory'. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 4.2, pp. 213–219 (cit. on p. 151).

Owen, Art B. (2013). *Monte Carlo Theory, Methods and Examples* (cit. on p. 149).

Schmon, Sebastian M, George Deligiannidis and Arnaud Doucet (2019). 'Bernoulli Race Particle Filters'. *AISTATS* (cit. on p. 151).

7

Bernoulli Race Particle Filters

Published in the proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019

(with George Deligiannidis and Arnaud Doucet)

7.1 Introduction

OVER the last 25 years particle filters have become a standard tool for optimal estimation in the context of general state space hidden Markov models (HMMs) with applications in ever more complex scenarios. HMMs are described by a latent (unobserved) process $(X_t)_{t\in\mathbb{N}}$ taking values in a space X and evolving according to the state transition density

$$X_t \mid (X_{t-1} = x) \sim f(\cdot \mid x)$$

for $t \geqslant 2$ and $X_1 \sim \mu(\cdot)$. The states can be observed through an observation process $(Y_t)_{t\in\mathbb{N}}$, taking values in a space Y with observation density

$$Y_t \mid (X_t = x) \sim g(\cdot \mid x).$$

Particle filters (PFs) sequentially approximate the joint distributions

$$\begin{aligned}\pi_t\left(x_{1:t}\right) &= p(x_{1:t} \mid y_{1:t}) \\ &= \frac{g(y_1 \mid x_1)\mu(x_1)\prod_{k=2}^{t} g(y_k \mid x_k) f(x_k \mid x_{k-1})}{p(y_{1:t})} \\ &= \frac{g(y_1 \mid x_1)\mu(x_1)\prod_{k=2}^{t} \varpi\left(x_k \mid y_k, x_{k-1}\right)}{p(y_{1:t})}\end{aligned}$$

on the space X^t at time t by evolving a set of samples, termed *particles*, through time. We use here the shorthand notation $x_{j:k} = (x_j, \ldots, x_k)$. The quantity $p(y_{1:t})$ denotes the marginal likelihood of the model

$$p(y_{1:t}) = \int g(y_1 \mid x_1)\mu(x_1) \prod_{k=2}^{t} \varpi\left(x_k \mid y_k, x_{k-1}\right) \mathrm{d}x_{1:t} \tag{7.1}$$

which is usually intractable. The PF algorithm proceeds as follows. Given a set of particles $\{X_{t-1}^i, i = 1, \ldots, N\}$ at time $t-1$ the particles are propagated through $q(x_t \mid x_{t-1}, y_t)$. To adjust for the discrepancy between the distribution of the proposed states and $\pi_t\left(x_{1:t}\right)$ the particles are weighted by

$$\begin{aligned} w_t(x_{t-1}, x_t) &= \frac{\varpi\left(x_t \mid y_t, x_{t-1}\right)}{q(x_t \mid x_{t-1}, y_t)}, \quad t \geqslant 2, \\ w_1(x_1) &= \frac{g(y_1 \mid x_1)\mu(x_1)}{q(x_1 \mid y_1)}. \end{aligned} \tag{7.2}$$

A subsequent resampling step according to these weights ensures that only promising particles survive. Here we consider only multinomial resampling, where we write

$$I_{1:N} \sim \mathsf{Mult}\left\{N; w_1, \ldots, w_N\right\}$$

to denote the vector $I_{1:N}$ consisting of N samples from a multinomial distribution with probabilities proportional to $w_1, \ldots, w_N$. The algorithm is summarized in Algorithm 7.1.

In most applications, interest lies not in the distribution itself, but rather in the expectations

$$\mathcal{I}(h) = \int h(x_{1:t})\pi_t(x_{1:t}) dx_{1:t} \tag{7.3}$$

for some test function $h : \mathsf{X}^t \to \mathbb{R}$. Applying a PF we can estimate this integral, for a set of particle genealogies $\left\{X_{1:t}^i, i = 1, \ldots, N\right\}$, by taking

$$\widehat{\mathcal{I}}_{t,\mathrm{PF}}(h) = \frac{1}{N} \sum_{i=1}^{N} h\left(X_{1:t}^i\right). \tag{7.4}$$

When the PF weights (7.2) are not available analytically, standard resampling routines—steps 3 and 7 in Algorithm 7.1—cannot be performed. However, in many cases one might be able to construct an unbiased estimator of the resampling weights. Del Moral, Doucet and Jasra (2007), Fearnhead, Papaspiliopoulos and Roberts (2008), Liu and Chen (1998) and Rousset and Doucet (2006) show that using random but unbiased non-negative weights still yields

Algorithm 7.1 Particle filter with N particles

At time $t = 1$

1: Sample $\widetilde{X}_1^i \sim q(\cdot \mid y_1), \quad i = 1 \dots N$

2: Compute weights

$$w_{1,i} = w_t(\widetilde{X}_1^i), \quad i = 1, \dots, N$$

3: $I_{1:N} \sim \mathsf{Mult}\left\{N; w_{1,1}, \dots, w_{1,N}\right\}$

4: Set $X_1^i = \widetilde{X}_1^{I_i}, \quad i = 1. \dots, N$

At times $t \geqslant 2$

5: Sample $\widetilde{X}_t^i \sim q(\cdot \mid X_{t-1}^i, y_t), \quad i = 1, \dots, N$

6: Compute weights as in (7.2)

$$w_{t,i} = w_t(X_{t-1}^i, \widetilde{X}_t^i), \quad i = 1, \dots, N$$

7: $I_{1:N} \sim \mathsf{Mult}\{N; w_{t,1}, \dots, w_{t,N}\}$

8: For $i = 1, \dots, N$ set

$$X_{1:t}^i = \left(X_{1:(t-1)}^{I_i}, \widetilde{X}_t^{I_i}\right).$$

a valid algorithm. This can be easily seen by considering a standard particle filter on an extended space. Yet, this flexibility does not come without cost. Replacing the true weights with a noisy estimate increases the variance for Monte Carlo estimates of type (7.4).

Here we introduce a new resampling scheme that allows for multinomial resampling from the true intractable weights while just requiring access to non-negative unbiased estimates. Thus, any PF estimate of type (7.4) will have the same variance as if the true weights were known. This algorithm relies on an extension of recent work in Dughmi et al. (2017) where the authors consider an unrelated problem.

In the supplementary material we collect the proofs for Propositions 3, 4 and 5, Theorem 6 and additional simulation studies. Code to reproduce our results is available online[1].

7.2 Particle Filters with Intractable Weights

Particle filters for state space models with intractable weights rely on the observation that replacing the true weights with a non-negative unbiased estimate is equivalent to a particle filter on an extended space. In this section we introduce some examples of models where the weights are indeed not available analytically and review briefly how the random weight particle filter (RWPF) can be applied in these instances.

[1] URL: `http://www.stats.ox.ac.uk/~schmon/`.

7.2.1 *Locally optimal proposal*

Recall that in the setting of a PF for a state space model at time $t-1$ the proposal for t which leads to the minimum one step variance, termed the *locally optimal proposal* q^*, is

$$q^*(x_t \mid x_{t-1}, y_t) = \frac{g(y_t \mid x_t) f(x_t \mid x_{t-1})}{p(y_t \mid x_{t-1})},$$

see, e.g. Doucet, Godsill and Andrieu (2000, Proposition 2). Sampling from the locally optimal proposal is usually straightforward using a rejection sampler. The weights

$$\begin{aligned} w_t(x_{t-1}, x_t) &= p(y_t \mid x_{t-1}) \\ &= \int g(y_t \mid x_t) f(x_t \mid x_{t-1}) \mathrm{d}x_t, \end{aligned} \tag{7.5}$$

however, are intractable if the integral on the right-hand side does not have an analytical expression, thus prohibiting an exact implementation of this algorithm. Our algorithm will enable us to resample according to these weights.

7.2.2 *Partially observed diffusions*

Most research on particle filters with intractable weights has been carried out in the setting of partially observed diffusions, see e.g. Fearnhead, Papaspiliopoulos and Roberts (2008), which we will describe here briefly. For simplicity, consider the univariate diffusion[2] process of the form

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + \mathrm{d}B_t, \quad t \in [0, \mathcal{T}], \tag{7.6}$$

where $a : \mathbb{R} \to \mathbb{R}$ denotes the drift function and $\mathcal{T} \in \mathbb{R}$ is the time horizon. For a general diffusion constant speed as assumed in (7.6) can be obtained whenever the Lamperti transform is available. The diffusion is observed at discrete times $t_i, i = 1, \dots T$ with measurement error described by the density $g(y_{t_i} \mid x_{t_i})$. In this model the resampling weights are often intractable since for most diffusion processes the transition density $f_{\Delta t}(x_{t_i} \mid x_{t_{i-1}})$ over the interval with length $\Delta t = t_i - t_{i-1}$ is not available analytically. However, as shown in Beskos and Roberts (2005), Dacunha-Castelle and Florens-Zmirou (1986) and Rogers (1985) the transition density over an interval of length Δt can be expressed as

[2] It is not necessary to restrict oneself to univariate diffusions, see e.g. Fearnhead, Papaspiliopoulos and Roberts (2008).

$$f_{\Delta t}(y \mid x) = \varphi(y; x, \Delta t) \exp\left(A(y) - A(x)\right) \times \mathbb{E}\left[\exp\left(-\int_0^{\Delta t} \phi(W_s)\mathrm{d}s\right)\right] \tag{7.7}$$

where $\varphi(\,\cdot\,; x, \Delta t)$ is the density of a Normal distribution with mean x and variance Δt, $\left(W_s\right)_{s\in[0,\Delta t]}$ denotes a Brownian bridge from x to y and $\phi(x) = \left(a^2(x) + a'(x)\right)/2$ for a function $A : \mathbb{R} \to \mathbb{R}$ with $A'(x) = a(x)$. The expectation on the right-hand side in (7.7) is with respect to the Brownian bridge and is usually intractable. Thus, for a particle filter to work in this example one either needs to implement a Bootstrap PF, which can sample the diffusion (7.6) exactly (see, e.g. Beskos and Roberts 2005), or one constructs an unbiased estimator of the expectation on the right-hand side to employ the RWPF. Note that for a function g and $U \sim \mathrm{Unif}[0, t]$ we have

$$\mathbb{E}\left[\frac{g(W_U)}{\lambda} \mid W_s, 0 \leqslant s \leqslant t\right] = \int_0^t \frac{g(W_s)}{\lambda t}\mathrm{d}s.$$

Using this relationship we can use debiasing schemes such as the Poisson estimator (Beskos, Fearnhead et al. 2006) to find a non-negative unbiased estimator of the expectation of the exponential.

7.2.3 Random weight particle filter

Here we briefly review the RWPF and compare its asymptotic variance to that of a PF with known weights. Assume one does not have access to the weight $\varpi\left(x_t \mid y_t, x_{t-1}\right)$ in (7.2) but only to some non-negative unbiased estimate $\hat{\varpi}(x_t \mid y_t, x_{t-1}, U_t)$, where U_t are auxiliary variables sampled from some density m. Then we carry out the multinomial resampling in Algorithm 7.1 by taking $I_{1:N} \sim \mathsf{Mult}\{N; \hat{w}_{t,1}, \ldots, \hat{w}_{t,N}\}$, where $\hat{w}_{t,k}$ is defined as in (7.2) with $\varpi\left(x_t \mid y_t, x_{t-1}\right)$ replaced by $\hat{\varpi}(x_t \mid y_t, x_{t-1}, U_t)$. This is equivalent to a standard particle filter on the extended space targeting at time t the distribution

$$\bar{\pi}_t\left(x_{1:t}, u_{1:t}\right) \propto \prod_{k=1}^{t} \hat{\varpi}(x_k \mid y_k, x_{k-1}, u_k) m\left(u_k\right)$$

which satisfies

$$\bar{\pi}_t\left(x_{1:t}, u_{1:t}\right) = \pi_t(x_{1:t})\bar{\pi}_t(u_{1:t} \mid x_{1:t}) \tag{7.8}$$

with

$$\bar{\pi}_t(u_{1:t} \mid x_{1:t}) = \prod_{k=1}^{t} \frac{\hat{\varpi}(x_k \mid y_k, x_{k-1}, u_k) m\left(u_k\right)}{\varpi\left(x_k \mid y_k, x_{k-1}\right)}.$$

A PF targeting the sequence of distributions π_t directly can be interpreted as a Rao-Blackwellized PF (Doucet, Godsill and Andrieu 2000) of the PF on the extended space. Hence, the following is a direct consequence of (Chopin 2004, Theorem 3).

7.1 Proposition. *For any sufficiently regular*[3] *real-valued test function h, the exact weight PF (EWPF) and RWPF estimators of $\mathbf{1}(h)$ defined in (7.3) both satisfy a $\sqrt{N}$-central limit theorem with asymptotic variances satisfying $\sigma^2_{EWPF,h} \leqslant \sigma^2_{RWPF,h}$.*

7.3 Bernoulli Races

Assume now that the intractable weights can be written as follows

$$\begin{aligned} w_t(x_{t-1}, x_t) &= \frac{g(y_t \mid x_t) f\left(x_t \mid x_{t-1}\right)}{q\left(x_t \mid x_{t-1}, y_t\right)} \\ &= c\left(x_{t-1}, x_t, y_t\right) b\left(x_{t-1}, x_t, y_t\right), \end{aligned} \tag{7.9}$$

where $0 \leqslant b(x, x', y) \leqslant 1$ for all $x, x', y \in \mathsf{X}$, and that we are able to generate a coin flip Z with $\mathbb{P}(Z = 1) = b\left(x, x', y\right)$. We will assume that $c(x, x', y)$ in (7.9) is analytically available for any x, x', y. For brevity we will denote (7.9) as $w_i = c_i b_i$, for particles $i = 1, \ldots, N$, dropping for now the dependence on t.

The aim of this section is to develop an algorithm to perform multinomial sampling proportional to the weights w_i, that is, sample from discrete distributions of the form

$$p(i) = \frac{c_i b_i}{\sum_{k=1}^{N} c_k b_k}, \quad i = 1, \ldots N \tag{7.10}$$

where $c_1, \ldots, c_N$ denote fixed constants whereas the $b_1, \ldots, b_N$ are unknown probabilities. We assume we are able to sample coins $Z_i \sim \mathrm{Ber}(b_i), i = 1, \ldots, N$. In practice, the $c_1, \ldots, c_N$ are selected to ensure that $b_1, \ldots, b_N$ take values between 0 and 1. The Bernoulli race algorithm for multinomial sampling proceeds by first proposing from the distribution

$$\mathbb{P}\left(I = i\right) = \frac{c_i}{\sum_{k=1}^{N} c_k} \tag{7.11}$$

[3] We refer the reader to Chopin (2004) for the mathematical details.

and then sampling $Z_I \sim \text{Ber}(b_I)$. If $Z_I = 1$, return I, otherwise the algorithm restarts. Pseudo code describing this procedure is presented in Algorithm 7.2. If we take $c_j = 1$ for all $j = 1, \dots, N$ we recover the original Bernoulli race algorithm from Dughmi et al. (2017).

Algorithm 7.2 Bernoulli race

1: Draw $I \sim \mathsf{Mult}\{1; c_1, \dots, c_N\}$
2: Draw $Z \sim b_I$
3: **if** $Z = 1$ **then return** $I = I$
4: **else** go back to line 1
5: **end if**

7.2 Proposition. *Algorithm 7.2 samples from the distribution*

$$p(i) = \mathbb{P}(I = i \mid Z_I = 1) = \frac{c_i b_i}{\sum_{k=1}^{N} c_k b_k}.$$

Proof. Note that the probability of sampling $I = i$ and accepting is

$$\mathbb{P}(I = i, Z_i = 1) = \frac{b_i c_i}{\sum_k c_k}.$$

It follows that observing $Z_I = 1$ has probability $\mathbb{P}(Z_I = 1) = \sum_k b_k c_k / \sum_k c_k$. Now for any $i = 1, \dots N$

$$\begin{aligned} p(i) &= \mathbb{P}(I = i \mid Z_I = 1) \\ &= \frac{\mathbb{P}(I = i, Z_i = 1)}{\mathbb{P}(Z_I = 1)} \\ &= \frac{b_i c_i}{\sum_k c_k} \Big/ \frac{\sum_k b_k c_k}{\sum_k c_k} = \frac{b_i c_i}{\sum_k b_k c_k}. \end{aligned}$$

□

7.3.1 Efficient Implementation

This algorithm repeatedly proposes an a priori unknown amount of random variables from the multinomial distribution (7.11). Naïve implementations of multinomial sampling are of complexity $O(N)$ for one draw from the multinomial distribution. Standard resampling algorithms, however, can sample N random variables at cost $O(N)$ (see e.g. Hol, Schon and Gustafsson 2006) but in this case the number of samples needs to be known beforehand. We show here that Bernoulli resampling can still be implemented with an average of $O(N)$ operations.

To ensure that the Bernoulli resampling is fast we need cheap samples from the multinomial distribution with weights proportional to $c_1, \dots, c_N$. We can achieve this with the Alias method (Walker 1974, 1977) which requires $O(N)$ preprocessing after which we can sample with $O(1)$ complexity. Hence, the overall complexity depends on the number of calls to the Bernoulli factory per sample. Denote $C_{j,N}$ the number of coin flips that are required to accept a value for the jth sample, where $j = 1, \dots, N$. The random variable $C_{j,N}$ follows a geometric distribution with success probability

$$\rho_N = \mathbb{P}(Z_I = 1) = \frac{\sum_{k=1}^N c_k b_k}{\sum_{k=1}^N c_k}. \tag{7.12}$$

The expected number of trials until a value is accepted is then

$$\mathbb{E}\left[C_{j,N}\right] = \frac{1}{\rho_N} \quad (j = 1, \dots, N).$$

The complexity of the resampling algorithm depends on the values of the acceptance probabilities $b_1, \dots, b_N$. For example, if all probabilities are identical, that is, $b_1 = \dots = b_N = b$, we expect N/b coin flips, each of cost $O(1)$, which leads to overall order $O(N)$ complexity. In practice, however, the success probabilities of our Bernoulli factories are all different. Assuming all $b_j, j = 1, \dots, N$ are non-zero, the expected algorithmic complexity per particle is bounded by the inverse of smallest and largest Bernoulli probabilities. Let $\underline{b} = \min\{b_1, \dots, b_N\}$. Then,

$$\mathbb{E}[C_{j,N}] = \frac{\sum_k c_k}{\sum_k b_k c_k} \leqslant \frac{\sum_k c_k}{\underline{b} \sum_k c_k} = \frac{1}{\underline{b}}$$

and with $\overline{b} = \max\{b_1, \dots, b_m\}$

$$\mathbb{E}[C_{j,N}] \geqslant \frac{1}{\overline{b}}.$$

In practice, the complexity of the Bernoulli sampling algorithm depends on the behaviour of ρ_N as shown by the following central limit theorem.

7.3 Proposition. *Assume* $\lim_{N\to\infty} \rho_N =: \rho \in (0, 1)$. *Then we have the following central limit theorem for the average number of coin flips as* $N \to \infty$

$$\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^N C_{j,N} - \frac{1}{\rho_N}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\rho}{\rho^2}\right),$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

In particular, Proposition 7.3 implies that the run-time concentrates around N/ρ_N with fluctuations of order $\sqrt{N}$. Thus, the order of complexity depends on the order of ρ_N and we have complexity $O(N)$ if

$$\limsup_{N\to\infty} \left| \frac{\sum_{k=1}^N c_k}{\sum_{k=1}^N c_k b_k} \right| < \infty. \tag{7.13}$$

As an example consider the case where $c_1 = \ldots = c_N = c$ are all identical. Then at time t

$$\rho_{t,N} = \frac{1}{N} \sum_{k=1}^N w_{t,k}. \tag{7.14}$$

An instance of this setting is the locally optimal proposal presented in Section 7.2.1. It is well known that as $N \to \infty$ (7.14) will converge towards $p(y_t \mid y_{1:t-1})$ and indeed we see that the algorithm's run-time will concentrate around a quantity of order N.

7.4 Remark. *After the Alias table is constructed, the algorithm can be implemented in parallel. This can lead to considerable gains if the number of particles used in the particle filter is high. If* $c_1 = \ldots = c_N = c$*, or if the constants denote a distribution for which a table as in the Alias method is already implemented before program execution, the above algorithm can be implemented entirely in parallel. Murray, Lee and Jacob (2016) consider the case of multinomial resampling using a rejection sampler with uniform proposal on the set* $\{1, \ldots, N\}$ *with known weights and show that this algorithm has parallel complexity* $O(\log N)$*.*

7.3.2 Estimating the Probability of Stopping

In our later applications we will be interested in evaluating the success probability

$$\rho_N = \frac{\sum_{k=1}^N c_k b_k}{\sum_{k=1}^N c_k}.$$

Assume that we sample N independent realizations from a multinomial distribution using the Bernoulli race algorithm described above and that $C_{1,N}, \ldots, C_{N,N}$ are the geometric random variables that count the number of trials until the algorithm accepts a value and terminates. Then $\hat{\rho}_N^{\text{naive}} = 1/\bar{C}_N$, where $\bar{C}_N = \sum_{k=1}^N C_{k,N}/N$ is a consistent estimator of ρ since $\mathbb{E}(C_{i,N}) = 1/\rho_N$ for all $i = 1, \ldots, N$ and therefore by a weak law of large numbers $\bar{C}_N \to 1/\rho$ in probability and $1/\bar{C}_N \to \rho$ in probability by the continuous mapping theorem. Unfortunately, the estimator $\hat{\rho}_N^{\text{naive}}$ is not unbiased which would be desirable. However, this can

be remedied by constructing the minimum variance unbiased estimator (see Hogg, McVean and Craig 2005, Definition 7.1.1) for a geometric distribution. This result is well known, but we repeat it here for convenience.

7.5 Proposition. *For $N \in \mathbb{N}$ let $C_{1,N}, \ldots, C_{N,N}$ denote independent samples from a geometric distribution with success probability ρ_N, then the minimum variance unbiased estimator for ρ_N is*

$$\hat{\rho}_N^{\text{mvue}} = \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}. \tag{7.15}$$

In order to understand the asymptotic behaviour of the estimator (7.15) we also provide the following central limit theorem.

7.6 Proposition. *For $N \in \mathbb{N}$ let $C_{1,N}, C_{2,N}, \ldots$ denote independent samples from a geometric distribution with success probability ρ_N, then*

$$\sqrt{N}\left(\hat{\rho}_N^{\text{mvue}} - \rho_N\right) \xrightarrow{d} \mathcal{N}\left(0, (1-\rho)\,\rho^2\right).$$

7.4 Bernoulli Race Particle Filter

7.4.1 Algorithm Description

We now consider the application of the Bernoulli race algorithm to particle filter methods. The Bernoulli race can be employed as a multinomial resampling algorithm, $\texttt{Mult}\{w_1, \ldots, w_N\}$, in Algorithm 7.1. However, the clear advantage is that this algorithm can be implemented even if the true weights are not analytically available, but we do have access to non-negative unbiased estimates. A $[0,1]$-valued unbiased estimator $\hat{b}$ for b can be converted into an unbiased coin flip by noting that $\mathbb{P}\left(V \leqslant \hat{b}\right) = b$, where $V \sim \text{Unif}[0,1]$ (see e.g. Lemma 2.1 in Łatuszyński et al. 2011).

We will refer to such a particle filter as a Bernoulli race particle filter (BRPF).

7.4.2 Likelihood estimation

Even though the Bernoulli race resampling scheme enables us to resample according to the true weights, the normalizing constant or marginal likelihood (7.1) remains intractable. We show here how we can still obtain an unbiased estimator for the normalizing constant. We

first recall that in particle filters an estimator of the normalizing constant is obtained by

$$\hat{p}(y_{1:T}) = \prod_{t=1}^{T} \hat{p}(y_t \mid y_{1:t-1}) = \prod_{t=1}^{T} \frac{1}{N} \sum_{k=1}^{N} w_{t,k}. \tag{7.16}$$

This estimator is well-known to be unbiased (see Del Moral 2004, Chapter 7). If the weights are not available this estimator cannot be employed. Fortunately, the quantity (7.16) comes up naturally when running the BRPF. Recall that the probability for the Bernoulli race to stop at a given iteration, i.e. to accept a value, is

$$\mathbb{P}\left(C_{j,N} = 1\right) = \frac{\sum_{k=1}^{N} c_{t,k} b_{t,k}}{\sum_{k=1}^{N} c_{t,k}} = \frac{\frac{1}{N}\sum_{k=1}^{N} w_{t,k}}{\frac{1}{N}\sum_{k=1}^{N} c_{t,k}}.$$

Thus, conditional on the weights $w_{t,k}, k = 1, \dots, N$, an unbiased estimator of (7.16) is given by virtue of (7.15):

$$\hat{\rho}_{N,T} = \prod_{t=1}^{T} \frac{1}{N} \sum_{k=1}^{N} c_{t,k} \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}. \tag{7.17}$$

7.7 Theorem. *The estimator* (7.17) *is unbiased for* $p(y_{1:T})$*, i.e.* $\mathbb{E}\left[\hat{\rho}_{N,T}\right] = p(y_{1:T})$.

7.5 Applications

7.5.1 Locally optimal proposal

We start with a Gaussian state space model. In linear Gaussian state space models the Kalman Filter can be used to analytically evolve the system through time. Nevertheless, the aim here is a proof of concept of the BRPF. We assume the latent variables follow the Markov chain

$$X_t = aX_{t-1} + V_t$$

where we take $a = 0.8$ and we observe these hidden variables through the observation equation

$$Y_t = X_t + W_t$$

with initialization $X_0 \sim \mathcal{N}(0, 5)$ and $V_t \sim \mathcal{N}(0, 5)$, $W_t \sim \mathcal{N}(0, 5)$. In this particular instance the locally optimal proposal is available analytically, but in most practical scenarios this will not be the case. For this reason sampling from the locally optimal proposal is implemented using a rejection sampler that proposes from the state equation. Coin flips for the weights

(7.5) can be obtained by sampling from the model $\xi_t \sim f(\cdot \mid x_{t-1})$ and computing

$$Z_t = 1\left\{U \leqslant \exp\left(-\frac{(y_t - \xi_t)^2}{10}\right)\right\}, \quad U \sim \text{Unif}[0,1].$$

This leads to the choice $c_{t,1} = \ldots = c_{t,N} = 1/\sqrt{10\pi}$ and

$$b_{t,k} = \int \exp\left(-\frac{(y_t - x)^2}{10}\right) f(x \mid x_{t-1}) \mathrm{d}x_{t-1}.$$

Note that $c_{t,k}$ is defined such that

$$\sup_{x,x',y} b_t(x, x', y) = 1$$

Such a choice ensures that the acceptance probability in the Bernoulli race, Algorithm 7.2, is as large as possible. This can lead to a considerable speedup when using the Bernoulli resampling algorithm.

Complexity and Run-time

Figure 7.1 shows the run-time for RWPF and the BRPF for the Gaussian state space model. The BRPF clearly has linear complexity in the number of particles. In a sequential implementation the Bernoulli race (orange) performs worse than the classical resampling scheme (blue), but the difference vanishes when implementing a parallel version of the Bernoulli race (green, with 32 cores). This demonstrates the speedup due to parallel sampling of the coin flips. Since we need many Bernoulli random variables each of which is computationally cheap this algorithm lends itself to a implementation using GPUs to further improve the performance of the Bernoulli race.

Efficiency

As alluded to earlier, if Bernoulli resampling is performed, the variance for any Monte Carlo estimate (7.4) will be the same as if the true weights were known and one applies standard multinomial resampling. From Proposition 7.1 it follows that the asymptotic variance of any Monte Carlo estimate of type (7.4) will be smaller when applying a BRPF over a RWPF. While the variance for functions (7.4) in the BRPF coincides with the standard particle filter we do not have access to the same estimator for the normalising constant and instead need to use the methods from Section 7.4.2.

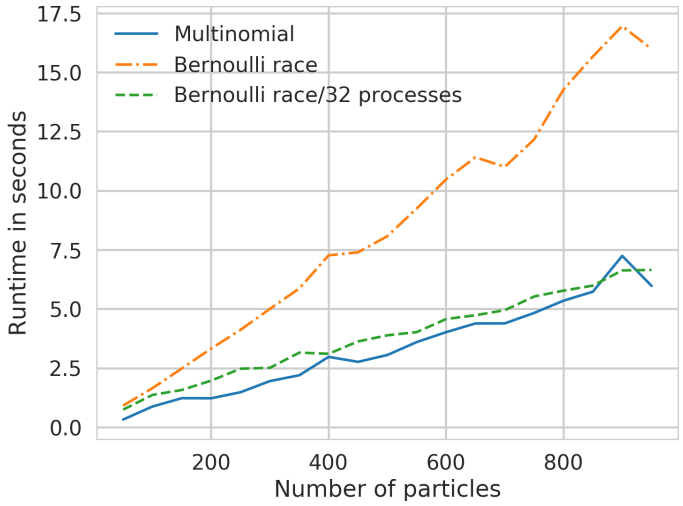


Figure 7.1: Comparison of the run-time for RWPF and BRPF for Gaussian state space model. For BRPF we show a sequential and a parallel implementation with 32 cores. For all algorithms we show wall-clock time for number of particles N.

For these reasons we will compare the performance of the BRPF with the RWPF for a set of different test functions $h : \mathsf{X}^T \to \mathbb{R}$ by comparing the variance of PF estimates (7.4). We then study the estimation of the normalizing constant separately. For the simulation we consider the functions

$$h_1(x_{1:T}) = \frac{1}{T}\sum_{t=1}^{T} x_t, \qquad h_2(x_{1:T}) = \|x_{1:T}\|_2,$$
$$h_3(x_{1:T}) = x_T, \qquad h_4(x_{1:T}) = \left(x_T - \bar{x}_T\right)^2,$$

where $\bar{x}_T = \sum_{i=1}^{N} x_T^i / N$. The PF approximations (7.4) using h_3 and h_4 are estimators for the quantities $\mu_T = \int x_T p(\mathrm{d}x_T \mid y_{1:T})$ and $\int \left(x_T - \mu_T\right)^2 p(\mathrm{d}x_T \mid y_{1:T})$.

All estimates are based on 100 runs of each particle filter using $N = 100$ particles. The results are collected in Table 7.1. We denote the standard deviation of the test functions under the BRPF as σ_{BRPF} and σ_{RWPF} for the RWPF. All measures indicate a reduction in the standard deviation. For the estimator of the function h_1, the standard deviation is reduced by 26% when compared to the RWPF.

We now investigate the estimates for the normalizing constant. Table 7.2 shows the standard deviation of the (log-)normalizing constant of a Gaussian state space model for $T = 50$ time steps. In this setting it is not obvious why the Bernoulli race resampling estimate should outperform the estimate provided by the RWPF. In our case however, we find that the BRPF

Table 7.1: For test functions $h_i, \dots i = 1, \dots, 4$ the standard deviation for RWPF and BRPF over 100 iterations.

Test function	σ_{RWPF}	σ_{BRPF}	$\sigma_{\text{BRPF}}/\sigma_{\text{RWPF}}$
h_1	0.17	0.12	0.74
h_2	0.93	0.78	0.84
h_3	0.19	0.19	0.96
h_4	0.42	0.40	0.94

Table 7.2: Comparison of the normalizing constant estimate for different implementations of the particle filter.

	$\text{sd}(\log \hat{p}(y_{1:T}))$
BRPF	0.55
RWPF	0.66

performs better.

7.5.2 *Partially Observed Diffusion*

We use the sine diffusion, a commonly used example in the context of partially observed diffusions, see e.g. Fearnhead, Papaspiliopoulos and Roberts (2008), given by the stochastic differential equation (SDE)

$$\mathrm{d}X_s = \sin(X_s)\mathrm{d}t + \mathrm{d}B_s, \quad s \in [0, 15].$$

Here, $(B_s)_{s\in[0,15]}$ denotes a Brownian motion and the drift function in (7.6) is $a(x) = \sin(x)$. Consequently, $A(x) = -\cos(x)$ and with $\phi(x) = \left(\sin(x)^2 + \cos(x)\right)/2$, the transition density is

$$\begin{aligned} f_{\Delta t}(x, y) = \varphi(y; x, \Delta t) \exp\left(-\cos(y) + \cos(x)\right) \\ \times \mathbb{E}\left[\exp\left(-\int_0^{\Delta t} \phi(W_s)\mathrm{d}s\right)\right]. \end{aligned} \tag{7.18}$$

We observe the state of the SDE through zero mean Gaussian noise with standard deviation 5 yielding weights

$$w(x_{t-1}, x_t, y_t) = \frac{\varphi(y_t; x_t, 5^2) f_{\Delta t}(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)},$$

As the proposal $q(\cdot \mid x_{t-1}, y_t)$ we take one step of the Euler–Maruyama scheme.

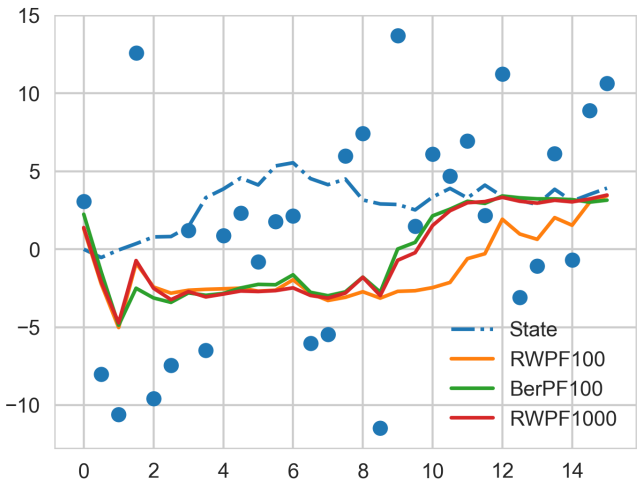


Figure 7.2: Tracking performance for different particle filters; RWPF with 100 particles (RWPF100) and 1000 particles (RWPF1000) as well as a particle filter using Bernoulli resampling with 100 particles (BerPF100).

The weights are intractable because of the expectation on the right-hand side in (7.18). The construction of unbiased coin flips can often rely on similar techniques as the construction of unbiased estimators. We can construct an unbiased estimator for the transition density in the following way. Fix a Brownian bridge $\left(W_s\right)_{s\in[0,\Delta t]}$ starting at x and finishing at y. Sample $\kappa \sim \text{Pois}(\lambda\Delta t)$, $U_1 \ldots, U_\kappa \sim U[0, \Delta t]$ then the Poisson estimator (Beskos, Fearnhead et al. 2006) is given by

$$\hat{P}(\kappa, U_1 \ldots, U_\kappa) = \exp\left\{(\lambda - c)\,\Delta t\right\} \prod_{i=1}^{\kappa} \frac{\left\{c - \phi\left(W_{U_i}\right)\right\}}{\lambda},$$

where c is chosen such that the estimator is non-negative. The Poisson estimator is unbiased

$$\mathbb{E}\left[\hat{P}(\kappa, U_1 \ldots, U_\kappa)\right] = \mathbb{E}\left[\exp\left(-\int_0^{\Delta t} \phi(W_s)\mathrm{d}s\right)\right]. \tag{7.19}$$

For the purposes of implementing the BRPF, we need to construct a coin flip with success

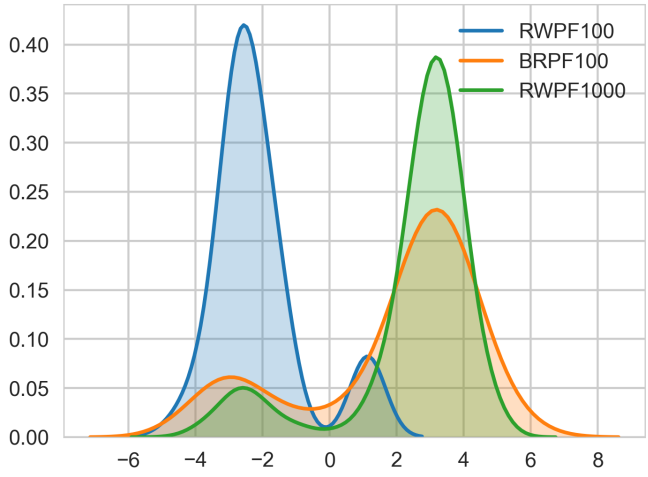


Figure 7.3: Kernel density estimate of $p(x_{10} \mid y_{1:10})$ based on particle approximations from a RWPF with 100 particles (RWPF) a BRPF with 100 particles (Bernoulli) and a RWPF with 1000 particles.

probability proportional to (7.19). We use the probability generating function approach which works analogously to the Poisson estimator. Sample $\kappa \sim \text{Pois}(\lambda\Delta t), V_i \sim \text{Unif}[0,1], i = 1, \dots \kappa$, then

$$Z = \prod_{i=1}^{\kappa} 1\left\{V_i \leqslant \frac{\left\{c - \phi\left(W_{U_i}\right)\right\}}{\lambda}\right\}$$

has success probability

$$\exp\left\{(c-\lambda)\Delta t\right\} \mathbb{E}\left[\exp\left(-\int_0^{\Delta t} \phi(W_s)\mathrm{d}s\right)\right].$$

Hence, to implement the BRPF we can choose

$$\begin{aligned} c_{t,k} &= \frac{\varphi(y_t; x_t, 5^2)\varphi(x_t; x_{t-1}, \Delta t)}{\varphi(y_t; x_{t-1} + \Delta t \sin(x_{t-1}), \Delta t)} \\ &\quad \times \exp\left(-\cos(x_t) + \cos(x_{t-1}) - (c-\lambda)\Delta t\right) \\ b_{t,k} &= \exp\left\{(c-\lambda)\Delta t\right\} \mathbb{E}\left[\exp\left(-\int_0^{\Delta t} \phi(W_s^k)\mathrm{d}s\right)\right], \end{aligned}$$

where $(W_s^k)_{s\in[0,\Delta t]}$ denotes a Brownian bridge from x_{t-1}^k to x_t^k.

Table 7.3: Comparison of the normalizing constant estimate for different implementations of the particle filter.

Test function	σ_{RWBF}	σ_{BRPF}	$\sigma_{\text{BRPF}}/\sigma_{\text{RWPF}}$
h_1	1.41	1.27	0.91
h_2	1.61	1.58	0.98
h_3	1.26	0.64	0.50
h_4	1.09	0.87	0.80

Table 7.4: Comparison of the normalizing constant estimate for different implementations of the particle filter.

	$\text{sd}(\log \hat{p}(y_{1:T}))$
RWPF	3.83
BRPF	3.11
Bootstrap	3.13

Complexity and Run-time

As in the previous example, we observe the BRPF to be of order $O(N)$ and slower than the RWPF. However, implementing the Bernoulli race in parallel yields almost the same performance in terms of run-time. The details can be found in the supplementary material.

Efficiency

One run for the RWPF and the BRPF is shown in Figure 7.2, where we show the true state of the SDE as well as the noisy observations and the particle filter approximations. In this scenario precise state estimation is hampered by the high noise and resulting partial multimodality of the filtering distribution as shown in Figure 7.3. For $N = 100$, Figure 7.2 shows that the BRPF performs much better as it finds the true state earlier. The RWPF finds this trajectory only when the number of particles is increased (here we show also the case $N = 1000$). With the same test functions as above we compare both algorithms in Table 7.3. We observe gains for all functions, with the most significant gain for the conditional mean, h_3. Again, the Bernoulli race will ordinarily be slower, but most of the difference in run-time vanishes when the Bernoulli race is implemented in parallel. Further details are provided in the supplementary material.

As a final test we use both particle filters for estimating the (log-)normalizing constant. The results are listed in Table 7.4. For comparison we also employ a bootstrap particle filter using the exact algorithm (Beskos and Roberts 2005) to propose from the model. We observed this implementation to be slower than the other two. The BRPF estimate (7.17) outperforms the RWPF and the bootstrap PF.

7.6 Conclusion

We have introduced the idea of Bernoulli races to resampling in particle filtering, utilizing the equivalence of unbiased estimation and the construction of unbiased coin-flips. This algorithm provides the first resampling scheme to allow for exact implementation of multinomial resampling when the weights are intractable, but unbiased estimates are available. We have shown that our algorithm has a complexity of order N, like standard multinomial resampling, and demonstrated its advantages over alternative methods in a variety of settings. In doing so we have focused our attention to the resampling step. Further gains using our algorithm could be obtained by considering auxiliary particle filters and to resample only if the effective sample size drops beyond a certain threshold. We expect both to positively impact the performance of the BRPF.

References

Beskos, Alexandros, Paul Fearnhead, Omiros Papaspiliopoulos and Gareth O Roberts (2006). 'Exact and Efficient Likelihood inference for Discretely Observed Diffusion Processes (with discussion).' *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 68.3, pp. 333–382 (cit. on pp. 157, 167).

Beskos, Alexandros and Gareth O Roberts (2005). 'Exact simulation of diffusions'. *The Annals of Applied Probability* 15.4, pp. 2422–2444 (cit. on pp. 156–157, 170).

Chopin, Nicolas (2004). 'Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference'. *The Annals of Statistics* 32.6, pp. 2385–2411 (cit. on p. 158).

Dacunha-Castelle, Didier and Danielle Florens-Zmirou (1986). 'Estimation of the coefficients of a diffusion from discrete observations'. *Stochastics: An International Journal of Probability and Stochastic Processes* 19.4, pp. 263–284 (cit. on p. 156).

Del Moral, Pierre (2004). *Feynman-Kac Formulae*. Springer, pp. 47–93 (cit. on p. 163).

Del Moral, Pierre, Arnaud Doucet and Ajay Jasra (2007). 'Sequential Monte Carlo for Bayesian Computation'. *Bayesian Statistics*. Ed. by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West. 8, pp. 1–34 (cit. on p. 154).

Doucet, Arnaud, Simon Godsill and Christophe Andrieu (2000). 'On sequential Monte Carlo sampling methods for Bayesian filtering'. *Statistics and computing* 10.3, pp. 197–208 (cit. on pp. 156, 158).

Dughmi, Shaddin, Jason D Hartline, Robert Kleinberg and Rad Niazadeh (2017). 'Bernoulli factories and black-box reductions in mechanism design'. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 158–169 (cit. on pp. 155, 159).

Fearnhead, Paul, Omiros Papaspiliopoulos and Gareth O Roberts (2008). 'Particle filters for partially observed diffusions'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, pp. 755–777 (cit. on pp. 154, 156, 166).

Hogg, Robert V, Joseph W McVean and Allen T Craig (2005). *Introduction to Mathematical Statistics*. 6th edition. Pearson Education, Upper Saddle River, New Jersey (cit. on p. 162).

Hol, Jeroen D, Thomas B Schon and Fredrik Gustafsson (2006). 'On resampling algorithms for particle filters'. *IEEE Nonlinear Statistical Signal Processing Workshop*, pp. 79–82 (cit. on p. 159).

Łatuszyński, Krzysztof, Ioannis Kosmidis, Omiros Papaspiliopoulos and Gareth O Roberts (2011). 'Simulating events of unknown probabilities via reverse time martingales'. *Random Structures & Algorithms* 38.4, pp. 441–452 (cit. on p. 162).

Liu, Jun S and Rong Chen (1998). 'Sequential Monte Carlo methods for dynamic systems'. *Journal of the American Statistical Association* 93.443, pp. 1032–1044 (cit. on p. 154).

Murray, Lawrence M, Anthony Lee and Pierre E Jacob (2016). 'Parallel resampling in the particle filter'. *Journal of Computational and Graphical Statistics* 25.3, pp. 789–805 (cit. on p. 161).

Rogers, LCG (1985). 'Smooth Transition Densities for One-Dimensional Diffusions'. *Bulletin of the London Mathematical Society* 17.2, pp. 157–161 (cit. on p. 156).

Rousset, M and A Doucet (2006). 'Discussion of Beskos et al.: 'Exact and Efficient Likelihood inference for Discretely Observed Diffusion Processes'''. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 68, pp. 374–375 (cit. on p. 154).

Walker, Alastair J (1974). 'New fast method for generating discrete random numbers with arbitrary frequency distributions'. *Electronics Letters* 10.8, pp. 127–128 (cit. on p. 160).

– (1977). 'An efficient method for generating discrete random variables with general distributions'. *ACM Transactions on Mathematical Software (TOMS)* 3.3, pp. 253–256 (cit. on p. 160).

Supplementary Material to Bernoulli Race Particle Filters

7.7 Proofs

When we say that T is a geometric random variable with parameter, or success probability, p we mean that

$$\mathbb{P}(T \geqslant k) = (1-p)^k, \qquad k \geqslant 1.$$

7.8 Proposition. *Assume* $\lim_{N\to\infty} \rho_N =: \rho \in (0,1)$. *Then we have the following central limit theorem for the average number of coin flips as* $N \to \infty$

$$\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^{N} C_{j,N} - \frac{1}{\rho_N}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\rho}{\rho^2}\right),$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

Proof. The random variables $Z_{i,N} = \left(C_{i,N} - 1/\rho_N\right)/\sqrt{N}$ form a triangular array

$$\left\{Z_{i,N}; i = 1, \dots N, N \in \mathbb{N}\right\}$$

of row wise independent random variables. Define $s_N^2 = \sum_{i=1}^{N} \operatorname{var}\left(Z_{i,N}\right) = \left(1-\rho_N\right)/\rho_N^2$ and notice that $s_N^2 \to 1-\rho \in (0,1)$ as $N \to \infty$. In addition one can easily show that $\mathbb{E}[|Z_{1,N}|^3]$ are bounded uniformly in N and therefore

$$\begin{aligned}\frac{1}{s_N^3}\sum_{i=1}^{N}\mathbb{E}\left[Z_{i,N}\right] &= \frac{\rho_N^{3/2}}{(1-\rho_N)^3}\frac{N}{N^{3/2}}\mathbb{E}\left[\left|Z_{1,N}\right|^3\right] \\ &\leqslant \frac{C}{\sqrt{N}} \to 0, \qquad \text{as } N \to \infty,\end{aligned}$$

that is the *Lyapunov condition* is satisfied and therefore $N \to \infty$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(C_{i,N} - \frac{1}{\rho_N} \right) \xrightarrow{d} \mathcal{N}\left(0, s^2\right),$$

where $s^2 = \lim_{N\to\infty} s_N^2 = (1-\rho)/\rho^2$. □

7.9 Proposition. *For $N \in \mathbb{N}$ let $C_{1,N}, \ldots, C_{N,N}$ denote independent samples from a geometric distribution with success probability ρ_N, then the minimum variance unbiased estimator for ρ_N is*

$$\hat{\rho}_N^{\text{mvue}} = \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}. \tag{7.20}$$

Proof. This is a straightforward application of the Lehmann-Scheffé Theorem (Hogg, McVean and Craig 2005, Theorem 7.4.1). Take the unbiased estimator $1\left\{C_{1,N} = 1\right\}$, where $1\{A\}$ denotes the indicator function of the set A, and condition on the complete and sufficient statistic (of the coin flips) $\sum_{k=1}^{N} C_{k,N}$. A straightforward calculation yields

$$\begin{aligned}
\mathbb{E}\left[1\left\{C_{1,N} = 1\right\} \mid \sum_{k=1}^{N} C_{k,N} = n \right] &= \frac{\mathbb{P}\left(C_1 = 1, \sum_{k=2}^{N} C_{k,N} = n-1 \right)}{\mathbb{P}\left(\sum_{k=1}^{N} C_{k,N} = n \right)} \\
&= \frac{p\binom{n-2}{N-2} p^{N-1}(1-p)^{n-N}}{\binom{n-1}{N-1} p^{N}(1-p)^{n-N}} \\
&= \frac{N-1}{n-1}
\end{aligned}$$

and thus

$$\hat{\rho}_N^{\text{mvue}} = \mathbb{E}\left[1\left\{C_{1,N} = 1\right\} \mid \sum_{k=1}^{N} C_{k,N} \right] = \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}.$$ □

7.10 Proposition. *For $N \in \mathbb{N}$ let $C_{1,N}, C_{2,N}, \ldots$ denote independent samples from a Geometric distribution with success probability ρ_N, then*

$$\sqrt{N}\left(\hat{\rho}_N^{\text{mvue}} - \rho_N \right) \xrightarrow{d} \mathcal{N}\left(0, (1-\rho)\,\rho^2\right).$$

Proof. By Proposition 7.3 we have convergence

$$\sqrt{N}\left(\frac{1}{N} \sum_{i=1}^{N} C_{i,N} - \frac{1}{\rho_N} \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\rho}{\rho^2}\right).$$

In addition,

$$\left|\sqrt{N}\left(\frac{\sum_{i=1}^{N} C_{i,N} - 1}{N-1} - \frac{1}{\rho_N}\right) - \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} C_{i,N} - \frac{1}{\rho_N}\right)\right| \to 0$$

almost surely, so

$$\sqrt{N}\left(\frac{\sum_{i=1}^{N} C_{i,N} - 1}{N-1} - \frac{1}{\rho_N}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\rho}{\rho^2}\right).$$

By the δ-Method with $g(x) = 1/x$, $|g'(x)| = 1/x^2$ we obtain

$$\begin{aligned}\sqrt{N}\left(g\left(\frac{\sum_{i=1}^{N} C_{i,N} - 1}{N-1}\right) - g\left(\frac{1}{\rho_N}\right)\right) &= \sqrt{N}\left(\frac{N-1}{\sum C_{i,N} - 1} - \rho_N\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, \frac{1-\rho}{\rho^2}\cdot\rho^4\right) = \mathcal{N}\left(0, (1-\rho)\rho^2\right). \square\end{aligned}$$

7.11 Theorem. *The estimator*

$$\hat{\rho}_{N,T} = \prod_{t=1}^{T} \frac{1}{N}\sum_{k=1}^{N} c_t^k \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}.$$

is unbiased for $p(y_{1:T})$, i.e.

$$\mathbb{E}\left[\hat{\rho}_{N,T}\right] = p(y_{1:T}).$$

Proof. The main argument is that the standard proof for unbiasedness using backward induction can be adapted to include our estimator using the geometric random variables $C_{1,N}, \ldots, C_{N,N}$. Denote by $X, \overline{X}$ the random variables before and after resampling respectively. Then we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} w_T^k \mid X_{T-1}^{1:N}\right] \\ &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} g(y_T \mid X_T^i) \mid X_{T-1}^{1:N}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\int g(y_T \mid x_T^i) p(x_T^i \mid X_{T-1}^{1:N}) \mathrm{d}x_T^i \\ &= \frac{1}{N}\sum_{i=1}^{N}\int g(y_T \mid x_T^i)\sum_{j=1}^{N} f(x_T^i \mid X_{T-1}^j)\frac{g(y_{T-1} \mid X_{T-1}^j)}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^k)}\mathrm{d}x_T^i\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\int g(y_T \mid x_T^i) f(x_T^i \mid X_{T-1}^j)\frac{g(y_{T-1} \mid X_{T-1}^j)}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^k)}\mathrm{d}x_T^i \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p(y_T \mid X_{T-1}^j)\frac{g(y_{T-1} \mid X_{T-1}^j)}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^k)} \\
&= \frac{\sum_{j=1}^{N} p(y_{T-1:T} \mid X_{T-1}^j)}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^k)}.
\end{aligned}
$$

Now, considering the case where we estimate the weights using the geometric random variables, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} c_T^k \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T} - 1} \mid X_{T-1}^{1:N}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} c_T^k \mathbb{E}\left\{\frac{N-1}{\sum_{k=1}^{N} C_{k,N,T} - 1} \mid X_T^{1:N}, \overline{X}_{T-1}^{1:N}\right\} \mid X_{T-1}^{1:N}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} c_T^k \frac{\sum_{k=1}^{N} w_T^k}{\sum_{k=1}^{N} c_T^k} \mid X_{T-1}^{1:N}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} w_T^k \mid X_{T-1}^{1:N}\right] \\
&= \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid x_{T-1}^k)}{\sum_{k=1}^{N} g(y_{T-1} \mid x_{T-1}^k)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} c_{T-1}^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T-1} - 1}\right)\cdot\left(\frac{1}{N}\sum_{j=1}^{N} c_T^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T} - 1}\right) \mid X_{T-2}^{1:N}\right] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\left(\frac{1}{N}\sum_{j=1}^{N} c_{T-1}^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T-1} - 1}\right) \\
&\qquad \times \left(\frac{1}{N}\sum_{j=1}^{N} c_T^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T} - 1}\right) \mid X_{T-2:T-1}^{1:N}, \overline{X}_{T-2}^{1:N}, C_{1:N,N,T-1}\Bigg] \mid X_{T-2}^{1:N}\Bigg] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} c_{T-1}^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T-1} - 1}\right) \right. \\
&\qquad \left. \times \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} c_T^j \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T} - 1}\right) \mid X_{T-2:T-1}^{1:N}, \overline{X}_{T-2}^{1:N}, C_{1:N,N,T-1}\right] \mid X_{T-2}^{1:N}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} c_{T-1}^{j} \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T-1} - 1}\right) \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k})}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k})} \mid X_{T-2}^{1:N}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} c_{T-1}^{j} \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N,T-1} - 1}\right) \mid X_{T-1}^{k}, \overline{X}_{T-2}^{k}\right] \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k})}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k})} \mid X_{T-2}^{1:N}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} w_{T-1}^{k}\right) \cdot \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k})}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k})} \mid X_{T-2}^{1:N}\right].
\end{aligned}
$$

For the final expectation, we can calculate

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} w_{T-1}^{k} \cdot \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k})}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k})} \mid X_{T-2}^{1:N}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k}) \frac{\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k})}{\sum_{k=1}^{N} g(y_{T-1} \mid X_{T-1}^{k})} \mid X_{T-2}^{1:N}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} p(y_{T-1:T} \mid X_{T-1}^{k}) \mid X_{T-2}^{1:N}\right] \\
&= \frac{1}{N}\sum_{k=1}^{N} \int p(y_{T-1:T} \mid x_{T-1}^{k}) p(x_{T-1}^{k} \mid X_{T-2}^{1:N}) dx_{T-1}^{k} \\
&= \frac{1}{N}\sum_{k=1}^{N} \int p(y_{T-1:T} \mid x_{T-1}^{k}) \frac{\sum_{j=1}^{N} g(y_{T-2} \mid X_{T-2}^{j}) f(x_{T-1}^{k} \mid X_{T-2}^{j})}{\sum_{j=1}^{N} g(y_{T-2} \mid X_{T-2}^{j})} dx_{T-1}^{k} \\
&= \sum_{j=1}^{N} p(y_{T-1:T} \mid X_{T-2}^{j}) \frac{g(y_{T-2} \mid X_{T-2}^{j})}{\sum_{k=1}^{N} g(y_{T-2} \mid X_{T-2}^{k})} = \frac{\sum_{j=1}^{N} p(y_{T-2:T} \mid X_{T-2}^{j})}{\sum_{j=1}^{N} g(y_{T-2} \mid X_{T-2}^{j})}.
\end{aligned}
$$

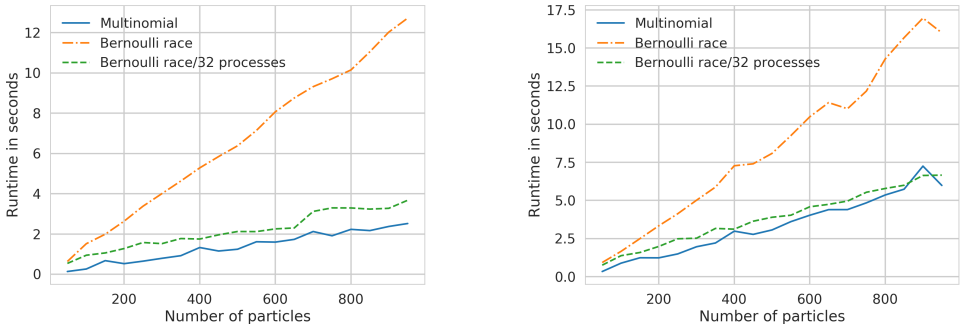
Repeated application of these steps yields

$$
\begin{aligned}
\mathbb{E}\left[\prod_{t=1}^{T} \frac{1}{N}\sum_{k=1}^{N} c_{t}^{k} \cdot \frac{N-1}{\sum_{k=1}^{N} C_{k,N} - 1}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} g(y_1 \mid X_1^k) \frac{\sum_{j=1}^{N} p(y_{1:T} \mid X_1^j)}{\sum_{j=1}^{N} g(y_1 \mid X_1^j)}\right] \\
&= p(y_{1:T}). \qquad \square
\end{aligned}
$$

7.8 Further Details to the Applications

7.8.1 Locally optimal proposal: run-time

We conjecture that the advantage of the BRPF over the RWPF grows as the state transition get computationally more expensive. We will demonstrate that on a toy example.



(a) Simulating from known Gaussian. **(b)** Rejection Sampler.

Figure 7.4: Comparing the run-time for different implementation for the proposal $q(\cdot \mid x, y)$. In case (a) the Gaussian proposal is sampled using a naïve implementation. In (b) we use a rejection sampler proposing from the state transition.

First note that in case of the Gaussian state space model, the locally optimal proposal,

$$\begin{aligned} q^*(x_t \mid y_t, x_{t-1}) &\propto g(y_t \mid x_t) f(x_t \mid x_{t-1}) \\ &\propto \varphi\left(x_t; \frac{1}{2}(a x_{t-1} + y_t), 2.5^2\right) \end{aligned}$$

is known analytically and is Gaussian. Therefore, in this special case the rejection sampler can be avoided. This will not usually be the case, hence we use the rejection sampler for our simulation. However, this will significantly speed up the state transition. In Figure 7.4 we compare the run-time of both, the RWPF and BRPF for different numbers of particles. In scenario (a) we use the cheap transition. Here the RWPF is very efficient. However, in scenario (b), which we also show in the main paper, where sampling from the transition is more expensive, we can implement the BRPF with the same run-time as the RWPF.

7.8.2 *Sine diffusion: run-time*

Here we compare the run-time for the RWPF and BRPF for the sine diffusion state space model. As pointed out in the main text, the BRPF is slower when implemented sequentially, but we observe in Figure 7.5 that the difference almost completely vanishes when use a parallel implementation with 16 processes.

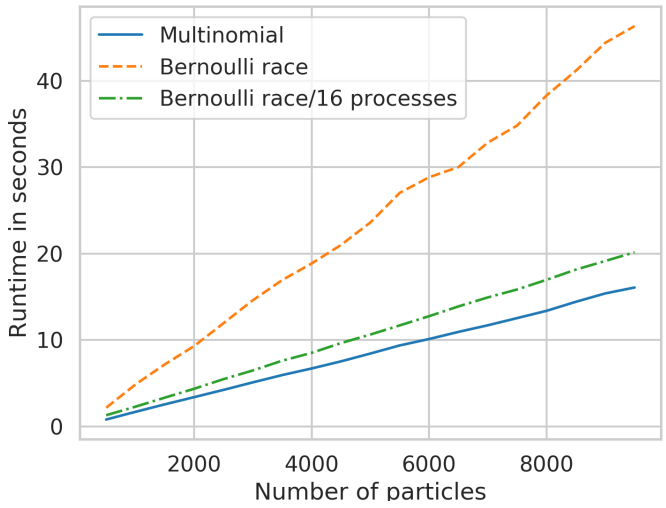


Figure 7.5: Comparison of the run-time for RWPF and BRPF for sine diffusion state space model. For BRPF we show a sequential and a parallel implementation with 16 cores. For all algorithms we show wall-clock time for number of particles N.

7.9 Cox Process inference

To further demonstrate the range of possible application and for further illustration of the Bernoulli race particle filter, we consider another application of the BRPF to a Cox process whose intensity function is governed by a Gaussian process that is normalized through a sigmoid function. The underlying Gaussian process is modelled as a Gauss–Markov process, allowing us to perform inference sequentially as done by Li and Godsill (2018) based on models of Adams, Murray and MacKay (2009) and Christensen, Murphy and Godsill (2012).

7.9.1 Likelihood and Estimation

We assume the latent Gaussian process to evolve according to the Ornstein–Uhlenbeck process

$$\mathrm{d}X_t = AX_t\mathrm{d}t + h\mathrm{d}B_t \qquad t \in [0, \mathcal{T}], \tag{7.21}$$

where $(B_t)_{t\in[0,\mathcal{T}]}$ denotes a Brownian motion from 0 to $\mathcal{T}$. We will describe the structural assumptions on A and h in more detail below. Denote the observed data by $s_{1:n} \subset [0, \mathcal{T}]$, where n is the number of observations. We model the data by a Cox process with intensity function

$$\lambda(t) = \sigma(X_{1,t}) = \frac{\lambda_{\max}}{1 + \exp(-X_{1,t})} = \lambda_{\max}\frac{\exp(X_{1,t})}{1 + \exp(X_{1,t})},$$

where $(X_{1,t})_{t\in[0,\mathcal{T}]}$ is the first coordinate of the SDE (7.21) introduced before. The likelihood function of the Cox process conditional on the intensity is Poisson with

$$p(s_{1:n} \mid \lambda, \mathcal{T}) = \exp\left(-\int_0^{\mathcal{T}} \lambda(s)\mathrm{d}s\right)\prod_{i=1}^{n}\lambda(s_i). \tag{7.22}$$

The likelihood is clearly intractable due to the integral which involves the latent Gaussian process. To implement a particle filter for sequential inference we need to find an unbiased coin with probability proportional to the integral in (7.22). For any interval $[t_0, t_1] \subset [0, \mathcal{T}]$ we can find an estimator for $\exp(-\int_{t_0}^{t_1}\lambda(s)\mathrm{d}s)$ by observing that for $U \sim \mathrm{Unif}[t_0, t_1]$ and a fixed process $X = (X_t)_{t\in[t_0,t_1]}$ we have

$$\begin{aligned}\mathbb{E}\left[\lambda(U) \mid X = x\right] &= \int_{t_0}^{t_1}\frac{\lambda(s)}{t_1 - t_0}\mathrm{d}s \\ &= \frac{1}{t_0 - t_1}\int_{t_0}^{t_1}\frac{1}{1+\exp(-x_s)}\mathrm{d}s\end{aligned}$$

and hence, if $V \sim \mathrm{Unif}[0, 1]$ independent of all other random variables, we obtain

$$P\left(V \leqslant \lambda(U) \mid X = x\right) = \int_{t_0}^{t_1}\frac{\lambda(s)}{t_1 - t_0}\mathrm{d}s.$$

In order to generate a coin involving the exponential, we can now use the *probability generating function* approach by noting that for $\kappa \sim \mathrm{Pois}(\Lambda)$ the coin

$$P = \prod_{k=1}^{\kappa} P_i, \quad P_i \sim \mathrm{Ber}(p),$$

has success probability $\exp(\Lambda(p-1))$, see e.g. (Dughmi et al. 2017). Now sample $K \sim \mathrm{Pois}(\lambda_{\max}(t_1 - t_0))$, then an unbiased coin flip for

$$\exp\left(-\int_{t_0}^{t_1}\lambda(s)\mathrm{d}s\right)$$

can be generated using

$$Z = \prod_{i=1}^{K} 1\left\{V_i \leqslant \frac{\lambda_{\max} - \lambda(U_i)}{\lambda_{\max}}\right\},$$

where $1\{A\}$ denotes the indicator function of the set A. This coin indeed has the correct success probability as can be seen by

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\prod_{i=1}^{K} 1\left\{V_i \leqslant \frac{\lambda_{\max} - \lambda(U_i)}{\lambda_{\max}}\right\} \mid X = x\right] \\
&= e^{-\lambda_{\max}(t_1 - t_0)} \sum_{k=0}^{\infty} \frac{\left(\lambda_{\max}(t_1 - t_0)\right)^k}{k!} \prod_{i=1}^{k} \mathbb{E}\left[1\left\{V_i \leqslant \frac{\lambda_{\max} - \lambda(U_i)}{\lambda_{\max}}\right\} \mid K = k, X = x\right] \\
&= e^{-\lambda_{\max}(t_1 - t_0)} \sum_{k=0}^{\infty} \frac{\lambda_{\max}^k \left(t_1 - t_0\right)^k}{k!} \left(\int_{t_0}^{t_1} \frac{\lambda_{\max} - \lambda(s)}{\lambda_{\max}(t_1 - t_0)} \mathrm{d}s\right)^k \\
&= e^{-\lambda_{\max}(t_1 - t_0)} \sum_{k=0}^{\infty} \frac{\left(\int_{t_0}^{t_1} \lambda_{\max} - \lambda(s) \mathrm{d}s\right)^k}{k! \lambda_{\max}^k} \\
&= e^{-\lambda_{\max}(t_1 - t_0)} e^{\lambda_{\max}(t_1 - t_0)} \exp\left(-\int_{t_0}^{t_1} \lambda(s) \mathrm{d}s\right) \\
&= \exp\left(-\int_{t_0}^{t_1} \lambda(s) \mathrm{d}s\right).
\end{aligned}$$

To implement the Bernoulli race particle filter we sample from the prior, which can be done exactly, as seen in the next section. As we are sampling from the prior the particle weights just depend on the observation likelihood. Suppose the particle filter is at current time t_0 and we propose to move to $t_1 = t_0 + \Delta t$. Let (X^k_{1,t_1}), $k = 1, \ldots, N$ denote a proposed particle path. Then the weight is given by

$$g(s_{i\,:\,s_i \in [t_0, t_1]} \mid X^k) = \exp\left(-\int_{t_0}^{t_1} \lambda^k(s) \mathrm{d}s\right) \prod_{i\,:\,s_i \in [t_0, t_1]} \lambda(s_i^k),$$

where

$$\lambda^k(s) = \frac{\exp(X^k_{1,s})}{1 + \exp(X^k_{1,s})}.$$

In the notation of Section 4 we have

$$\begin{aligned}
c^k_{t_1} &= \prod_{i\,:\,s_i \in [t_0, t_1]} \lambda^k(s_i), \\
b^k_{t_1} &= \exp\left(-\int_{t_0}^{t_1} \lambda^k(s) \mathrm{d}s\right).
\end{aligned}$$

To implement the estimator $\hat{Z}$, we need to be able to simulate a bridge from the stochastic differential equation (7.21). Since the process at hand is analytically tractable, exact sampling from the bridge is straightforward, see e.g. Bladt, Finch and Sørensen (2016) using the quantities computed below.

7.9.2 Ornstein–Uhlenbeck Prior

In this section we provide further details on the assumptions on the Gaussian process and show how we can sample the state transition. We assume the prior distribution on the latent space is given by the Ornstein–Uhlenbeck process (7.21), that is

$$\mathrm{d}X_t = AX_t\mathrm{d}t + h\mathrm{d}B_t \qquad t \in [0, \mathcal{T}].$$

For A and h we take the values

$$A = \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix}, \quad h = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}$$

where θ is a negative real value. The process can be written as a differential equation with random velocity following a mean reverting real-valued Ornstein–Uhlenbeck process,

$$\begin{pmatrix} \mathrm{d}X_{1,t} \\ \mathrm{d}X_{2,t} \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix} \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ \sigma \end{pmatrix} \mathrm{d}W_t.$$

This can be written as

$$\begin{aligned} \mathrm{d}X_{1,t} &= X_{2,t}\mathrm{d}t \\ \mathrm{d}X_{2,t} &= \theta X_{2,t}\mathrm{d}t + \sigma\mathrm{d}B_t. \end{aligned}$$

This is a linear Gaussian system and thus the solution can be derived analytically, see e.g. Christensen, Murphy and Godsill (2012). The discretized system at time $s > r$ can be simulated exactly by sampling

$$X_s \mid (X_r = x) \sim \mathcal{N}\left(e^{A(s-r)x}, e^{A(s-r)}Q(r,s)\left(e^{A(s-r)}\right)^{\mathsf{T}}\right),$$

where

$$Q(r,s) = \int_r^s \exp(-At)hh^{\mathsf{T}}\exp(-At)^{\mathsf{T}}\mathrm{d}t.$$

In order to implement the Ornstein–Uhlenbeck process we need the quantities

$$\exp(At) = \sum_{k=0}^{\infty} t^k \frac{A^k}{k!}.$$

Note that

$$A^2 = \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix} = \begin{bmatrix} 0 & \theta \\ 0 & \theta^2 \end{bmatrix}$$

$$\begin{aligned} A^k = A \cdot A^{k-1} &= \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix} \begin{bmatrix} 0 & \theta^{k-2} \\ 0 & \theta^{k-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \theta^{k-1} \\ 0 & \theta^k \end{bmatrix} \end{aligned}$$

Hence, we can compute the matrix exponential analytically

$$\begin{aligned} \exp(At) &= \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \begin{bmatrix} 0 & \theta^{k-1} \\ 0 & \theta^k \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sum_{k=1}^{\infty} \frac{t^k \theta^k}{k!} \begin{bmatrix} 0 & 1/\theta \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & \exp(\theta t)/\theta - 1/\theta \\ 0 & \exp(\theta t) - 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \exp(\theta t)/\theta - 1/\theta \\ 0 & \exp(\theta t) \end{bmatrix}. \end{aligned}$$

This gives us

$$\begin{aligned} \exp(-At)h &= \begin{bmatrix} 1 & \exp(-\theta t)/\theta - 1/\theta \\ 0 & \exp(-\theta t) \end{bmatrix} \begin{bmatrix} 0 \\ \sigma \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma}{\theta} \exp(-\theta t) - \sigma/\theta \\ \sigma \exp(-\theta t) \end{bmatrix} \end{aligned}$$

and

$$\exp(-At) h h^\top \exp(-At)$$

$$= \begin{bmatrix} \frac{\sigma}{\theta}\exp(-\theta t) - \sigma/\theta \\ \sigma\exp(-\theta t) \end{bmatrix} \begin{bmatrix} \frac{\sigma}{\theta}\exp(-\theta t) - \sigma/\theta & \sigma\exp(-\theta t) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\sigma^2}{\theta^2}\left(\exp(-\theta t) - 1\right)^2 & \frac{\sigma^2}{\theta}\left(\exp(-2\theta t) - \exp(-\theta t)\right) \\ \frac{\sigma^2}{\theta}\left(\exp(-2\theta t) - \exp(-\theta t)\right) & \sigma^2\exp(-2\theta t) \end{bmatrix}.$$

Thus

$$\begin{aligned} Q(r,s) &= \int_r^s \exp(-At)hh^\mathsf{T}\exp(-At)^\mathsf{T}\mathrm{d}t \\ &= \begin{bmatrix} \int_r^s \sigma^2\left(\exp(-\theta t) - 1\right)^2\mathrm{d}t & \int_r^s \frac{\sigma^2}{\theta}\left(\exp(-2\theta t) - \exp(-\theta t)\right)\mathrm{d}t \\ \int_r^s \sigma^2\left(\exp(-2\theta t) - \exp(-\theta t)/\theta\right)\mathrm{d}t & \int_r^s \sigma^2\exp(-2\theta t)\mathrm{d}t \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma^2}{\theta^3}\left(-2\theta r + e^{-2\theta r} - 4e^{-\theta r} - e^{-2\theta s} + 4e^{-\theta s} + 2\theta s\right) & \frac{\sigma^2}{2\theta^2}\left(e^{-2r\theta} - e^{-2s\theta}\right) - \frac{\sigma^2}{\theta^2}\left(e^{-r\theta} - e^{-s\theta}\right) \\ \frac{\sigma^2}{2\theta^2}\left(e^{-2r\theta} - e^{-2s\theta}\right) - \frac{\sigma^2}{\theta^2}\left(e^{-r\theta} - e^{-s\theta}\right) & \frac{\sigma^2}{2\theta}\left(e^{-2r\theta} - e^{-2s\theta}\right) \end{bmatrix} \end{aligned}$$

The covariance matrix for the system transition

$$\mathrm{Cov}(r,s) = e^{A(s-r)}Q(r,s)\left(e^{A(s-r)}\right)^\mathsf{T}$$

can thus be computed analytically.

7.9.3 Simulation

We simulate a non-homogeneous Poisson process in the interval $[0, 50]$ using the intensity function

$$\lambda(s) = 2\exp(-s/15) + \exp\left(-((s-25)/10)^2\right),$$

see e.g. Adams, Murray and MacKay (2009) and Li and Godsill (2018). We run the Bernoulli race particle filter with the above specification and 30 particles to find the intensity function on the interval $[0, 50]$ which is divided into 10 equispaced intervals. The result of one run is shown in Figure 7.6. We can see that the Bernoulli race particle filter is approximating the true intensity function.

References

Adams, Ryan Prescott, Iain Murray and David JC MacKay (2009). 'Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities'. *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp. 9–16 (cit. on pp. 178, 183).

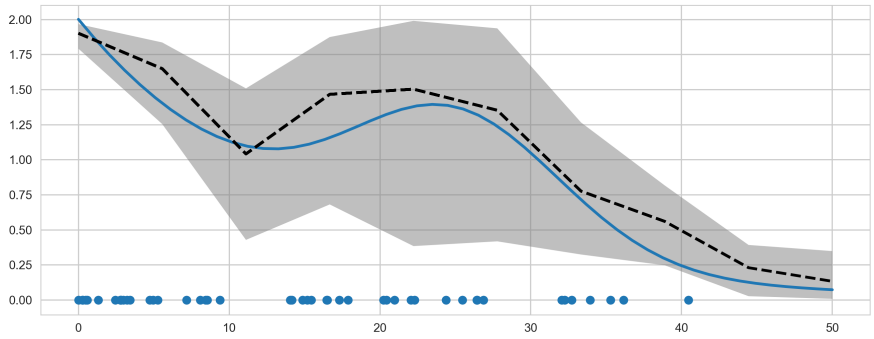


Figure 7.6: Result of one run of the Bernoulli race particle filter using 30 particles. Arrival times of the Poisson process are shown as dots on the abscissa. The solid blue line shows the true intensity function and the dashed line shows the mean of the particle approximation at times $t_1, t_2, \ldots$. The shaded area highlights the 10% and 90% quantile of the particle approximation.

Bladt, Mogens, Samuel Finch and Michael Sørensen (2016). 'Simulation of multivariate diffusion bridges'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.2, pp. 343–369 (cit. on p. 181).

Christensen, Hugh L, James Murphy and Simon J Godsill (2012). 'Forecasting high-frequency futures returns using online Langevin dynamics'. *IEEE Journal of Selected Topics in Signal Processing* 6.4, pp. 366–380 (cit. on pp. 178, 181).

Dughmi, Shaddin, Jason D Hartline, Robert Kleinberg and Rad Niazadeh (2017). 'Bernoulli factories and black-box reductions in mechanism design'. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 158–169 (cit. on p. 179).

Hogg, Robert V, Joseph W McVean and Allen T Craig (2005). *Introduction to Mathematical Statistics*. 6th edition. Pearson Education, Upper Saddle River, New Jersey (cit. on p. 173).

Li, Chenhao and Simon J Godsill (2018). 'Sequential Inference Methods for Non-Homogeneous Poisson Processes with State-Space Prior'. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2856–2860 (cit. on pp. 178, 183).

8

Epilogue: Further Research on Bernoulli Race Particle Filters

We have introduced a resampling scheme for particle filters that can be used even when the precise weights cannot be computed. Our approach is related to the theory of Bernoulli factories (see e.g. Keane and O'Brien 1994). We draw upon the concept of Bernoulli races introduced by Dughmi et al. (2017) which can work very well, but suffers from poor performance if the coin flips are not carefully constructed. For example, if the problem is such that the coin flip probabilities are very low, the algorithm will take very long to accept a coin flip. Even worse, if the probability of acceptance is below computer precision, the algorithm would be unable to terminate successfully. Our extension of the approach by Dughmi et al. (2017) can help to circumvent this issue by allowing the construction of an efficient proposal that leads to a high acceptance probability. This way, the above problems can be alleviated in many scenarios.

Another avenue not pursued in our research is to fully leverage the parallel structure of the independent proposals and coin flips that are needed for resampling. While we have used multiple CPUs to demonstrate the effect of a parallel implementation, the true benefit would lie in a GPU implementation. The latter are famously efficient for a large load of small parallel tasks. Such an implementation could increase the performance of our method significantly.