From point estimation to Bayesian inference via dynamical systems

Gavin J. Gibson[1] & Ben Hambly[2]

January 23, 2015

[1] *School of Mathematical and Computer Sciences, The Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK*

[2] *Mathematical Institute, University of Oxford, 24-29 St Giles, Oxford, OX1 3LB, UK*

Keywords: Bayesian inference, point estimation, generalised data augmentation, dynamical systems.

Abstract

Using only a simple principle that states that the class of valid systems for statistical inference should be closed under a certain data-augmentation process and complete in an obvious sense, we show how Bayesian and other systems of inferences can be generated in a direct manner from an initial system of point estimators. Using a generalisation of Gibbs sampling, we construct refinement operators that act on systems of inference to transform them into preferable systems. Interest then focuses on systems that are fixed by these operators. In the 1-dimensional setting, we characterise fixed points obtained from systems of moment estimators, showing that these are Bayesian when the model lies in the exponential family, with the usual conjugate prior arising as a by-product of the construction. In other cases, the limiting inferences are pseudo-Bayesian in that parameter densities combine a prior with a data-dependent pseudo-likelihood. We also show that, given sufficiently strong assumptions on the model, the construction, when applied to an initial system of maximum-likelihood estimators, leads to Bayesian inference with Hartigan's maximum likelihood prior as the fixed point, and consider further generalisations of this. A counter-example is given to show that, for non-regular models, a Bayesian fixed point may not arise from maximum-likelihood estimation. Inter alia, the results offer a new perspective on the relationship between Bayesian inference and classical point estimation whereby the former is generated from the latter without direct reference to the Bayesian paradigm.

1 Introduction

Let $y_1, y_2, y_3, \ldots$ denote a sequence of independently, identically distributed (i.i.d.) samples from a density $\nu_\theta(y)$ with $y \in \mathscr{Y}$ and parameter $\theta \in \mathcal{K}$ where $\mathscr{Y}, \mathcal{K} \subseteq \mathbb{R}$ and, for $n \in \mathbb{N}$ let $x_n = (y_1, ..., y_n)$ denote the outcome of an experiment that records the first n values. A *system of inferences* is defined as

$$\Theta = \{p_{x_n}(\theta) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

where $p_{x_n}(\theta)$ is a density representing the belief about parameter θ given the outcome x_n. Systems of point estimators are obtained on setting $p_{x_n}(\theta) = \delta(\theta - \hat{\theta}(x_n))$, where $\hat{\theta}(x_n)$ denotes the point estimate of θ calculated from data x_n. Bayesian systems have the property that $p_{x_n}(\theta) \propto \pi(\theta)\nu_\theta(x_n)$ for some prior density $\pi(\theta)$. For a given model, we denote by $\mathscr{X}$ the collection of all systems of inference. A system of inferences is essentially the same as the concept of an *inversion* as defined in Hartigan (1964).

This paper explores a novel, dynamical-systems approach to investigating the structure of $\mathscr{X}$ and comparing its constituent systems of inferences. Specifically we use a generalised data-augmentation principle introduced in Gibson et al. (2011), in order to define mappings, Ψ and Φ, called *refinement operators* in Section 2, from $\mathscr{X}$ to itself, which maps a given system of inferences to one which is preferable in a sense which we make explicit in Section 2. Interest then focuses on the fixed points of these operators. These have the property of being preferable to all systems in their domain of attraction and, arguably, attention should be restricted to these fixed points when selecting appropriate statistical procedures. Moreover, a refinement operator induces a natural structure on $\mathscr{X}$ by partitioning it into the domains of attraction of the fixed points, with systems lying in distinct domains being mutually incomparable by our definition of preferability. This structure provides a means for exploring connections between approaches to inference and estimation. When the domain of attraction of a Bayesian fixed point contains a system of point estimators, then a correspondence between Bayesian and classical approaches is identified.

Connections between classical and Bayesian inference have been sought by identifying choices of prior distribution for θ, so that the resulting posterior density satisfies certain classical criteria, at least asymptotically. Examples include reference priors (see Berger et al. (2009); Bernardo (1979)), which maximise, in the large-sample limit, the expected Kullback-Leibler (KL) distance between prior and posterior, making the data maximally informative in a natural sense. Another example is the Jeffreys prior (Jeffreys (1946)) which attempts to assign equal prior probability to intervals of a given level of confidence. A decision-theoretic approach is taken by Hartigan (Hartigan (2012)) where the notion of a risk-matching prior for an estimator is described, this being the prior for which the corresponding posterior Bayes estimator has the same risk to order n^{-2} as the given estimator. Of particular relevance here is the *maximum likelihood prior* (Hartigan (1964, 1998)), which is the risk-matching prior corresponding to maximum-likelhood estimation. When θ is the canonical parameter in a distribution from the exponential family, the maximum-likelihood prior is uniform on the parameter space. More generally, for the 1-dimensional models considered in this paper, the maximum-likelihood prior $\pi(\theta)$ for a model with density $\nu_\theta(y)$ satisfies

$$\frac{\partial \log \pi(\theta)}{\partial \theta} = \frac{a(\theta)}{i(\theta)},$$

where

$$a(\theta) = \mathbb{E}\left(\frac{\partial \log \nu_\theta(Y)}{\partial \theta}\frac{\partial^2 \log \nu_\theta(Y)}{\partial \theta^2}\right)$$

and

$$i(\theta) = \mathbb{E}\left(-\frac{\partial^2 \log \nu_\theta(Y)}{\partial \theta^2}\right).$$

These approaches establish correspondences by constructing an analysis which is Bayesian from the outset and which matches the classical analysis according to some external criterion. By contrast, our approach attempts to generate 'internally' from a system of point estimators new systems of inference which are invariant under certain data-augmentation operations. In some cases, namely when the initial estimators are essentially maximum-likelihood and sufficient regularity holds, the invariant systems generated are Bayesian and the Bayesian paradigm arises as a consequence, rather than a premise of the construction. On the other hand our results demonstrate that non-Bayesian invariant systems can be generated in this way.

In our main result, Theorem 3.2, we characterise, for a broad class of 1-parameter models, those points fixed by Ψ whose domains of attraction contain a system of moment-based estimators. These limiting inferences can be considered to be pseudo-Bayesian in the sense that the 'posterior' densities that arise are exhibited as a product of a data-independent function and data-dependent function, playing the respective roles of a prior and pseudo-likelihood. In Example 3.5 we give an example to show that that the limiting inference, when non-Bayesian, may nevertheless approximate a Bayesian analysis of an experiment in which only the sample mean were observed. For the models in the exponential family, given an initial system of maximum-likelihood estimators, a Bayesian analysis using the maximum-likelihood prior arises as the fixed point, with other priors from the conjugate family arising for other choices of initial estimators.

In Section 5 we explore the generalisations of the main theorem to fixed points of Φ arising from systems of maximum likelihood estimators. An argument is presented that suggests that the Bayesian analysis with the maximum-likelihood prior should be obtained as the fixed point given sufficiently strong regularity. Moreover, a counter-example based on the uniform distribution is included to demonstrate that the Bayesian limit does not arise in general.

2 Generalised data augmentation, validity and preferability

Throughout we take the view that the *validity* of any statistical procedure is a subjective judgement on the part of the user or observer. In what follows we do not attempt, therefore, to define validity in absolute terms. When the term *valid* is used, this should be interpreted as *valid in the opinion of a given observer.* We describe an approach that draws on the concept of the *relative* validity of procedures and the related concept of the *preferability* of one procedure to another.

In Gibson et al. (2011) a generalised data augmentation principle was proposed and used to construct, or refine, inferences in the form of posterior-like summaries of belief. This asserts that the set of all systems of inference for θ that are considered valid by a given observer, should be closed under a data-augmentation operation as described by Principle 2.1.

Augmentation Principle 2.1. *Let*

$$\Theta = \{p_{x_n}(\theta) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

denote a valid system of inferences. Then for any $n, m \in \mathbb{N}$, the system $\Theta^{n,m}$ is valid, where $\Theta^{n,m}$ is obtained from Θ by replacing $p_{x_n}(\theta)$ with

$$p_{x_n}^{(n,m)}(\theta) = \int \int p_{x_n}(\theta') \nu_{\theta'}(x_{n+m} | x_n) p_{x_{n+m}}(\theta) dx_{n+m} d\theta' \qquad (1)$$

and $\nu_{\theta'}(x_{n+m}|x_n)$ denotes the conditional density of x_2 given x_1 for the model with parameter θ' .

Principle 2.1 implies that, if Θ is valid, then so is $\Theta^{n,m}$ but not the converse. This leads us to define the notion of *preferability* as follows. We say that Θ_2 is preferable to Θ_1 if any observer who considers Θ_1 to be valid also considers Θ_2 to be valid. If we consider only observers who accept Principle 2.1 then the principle itself provides a mechanism for identifying inferences that are preferable to any given system.

Informally, Principle 2.1 states that a valid inference given x_n is obtained by taking a mixture of valid inferences based on x_{n+m}, in a manner analogous to Bayesian data augmentation, where the n samples in x_n are augmented by the next m samples in the sequence. Of course, when Θ is a Bayesian system of inferences, then Θ and $\Theta^{n,m}$ coincide. Our main interest will be in the application of Principle 2.1 more generally, to transform (or refine) a system of inferences into a preferable one. This leads to the refinement operators formulated using Principle 2.1 below.

First note that $p_{x_n}^{(n,m)} = p_{x_n}P$, where P is the transition kernel of a Markov chain, called the *generalised data-augmentation chain*, on the parameter space, in which updates to the current state $\theta^{(i)}$ are generated by first drawing $x_{n+m} \sim \nu_{\theta^{(i)}}(x_{n+m}|x_n)$ and then drawing $\theta^{(i+1)} \sim p_{x_{n+m}}$. Applying Principle 2.1 sequentially, it follows that a valid system is obtained by replacing p_{x_n} with $p_{x_n}P^k$ for any $k \in \mathbb{N}$. Moreover, if the generalised data-augmentation chain defined by P is ergodic with stationary density, ψ_{x_n}, then replacing p_{x_n} with ψ_{x_n} also yields a valid system of inferences, so long as we allow the class of valid inferences to be complete. This motivates an additional principle from Gibson et al. (2011).

Completeness Principle 2.2. *Suppose that* $\{\Theta^{(i)}, i = 1, 2\dots\}$ *denotes a sequence of valid systems for which*

$$\lim_{i\to\infty} p_{x_n}^{(i)} = \psi_{x_n}, n \in \mathbb{N}, x_n \in \mathscr{Y}^n,$$

Then

$$\Psi = \{\psi_{x_n}(\theta) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$$

is also a valid inference. If, for some system of inferences Θ and every i, $\Theta^{(i)}$ is preferable to Θ, then Ψ is preferable to Θ.

We now appeal to Principles 2.1 and 2.2 to formulate a *refinement operator*, Ψ, that can be applied to an initial system of inferences

$$\Theta_0 = \{p_{x_n,0}(\theta) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

to generate a sequence of systems $\{\Theta_i | i \in \mathbb{N}\}$ in which $\Theta_{i+1} = \Psi(\Theta_i)$ is preferable to Θ_i.

First denote by $P^{(0,m)}(x_n)$ the transition kernel of the generalised data-augmentation chain arising when the observation x_n is augmented by the next m samples from the distribution. For this chain the state $\theta^{(r)}$ is updated by drawing $\theta^{(r+1)} \sim p_{x_{n+m},0}$ where $x_{n+m} \sim \nu_{\theta^{(r)}}(x_{n+m}|x_n)$. We construct a new inference for x_n by considering the stationary distribution of the chain for each m, and then taking the limit of these stationary distributions as $m \to \infty$ in order to remove dependence on the particular choice of m. On performing this for each n in ascending order, appealing to Principle 2.2 as required, we generate the new system $\Theta_1 = \Psi(\Theta_0)$.

Generally, we construct $\Theta_{i+1} = \{p_{x_n,i+1}(\theta) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$ from $\Theta_i = \{p_{x_n,i}(\theta) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$ recursively by setting

$$p_{x_n,i+1} = \lim_{m \to \infty} \lim_{k \to \infty} p_{x_n,i}[P^{(i,m)}(x_n)]^k, \tag{2}$$

where the limits are taken in the sense of weak convergence. Suppose now that $\lim_{i\to\infty} \Theta_i = \Theta_\infty$. Then Θ_∞ is preferable to Θ_i, for all i, is invariant under Ψ, and is, in a natural sense, maximally preferable.

Denote by $\mathscr{C} \subset \mathscr{X}$ the collection of those systems of inference Θ_0 for which the limiting system Θ_∞ exists, and denote by $\mathscr{C}_F \subset \mathscr{C}$ the corresponding set of fixed points.

It is clear that any Bayesian system Θ, for which

$$p_{x_n}(\theta) \propto \pi(\theta)\nu_\theta(x_n)$$

for some prior density $\pi(\theta)$, lies in $\mathscr{C}_F$. In this case the generalised data-augmentation chain with transition kernel $P(x_n, p_{x_{n+m}})$ is a Gibbs sampler and the density p_{x_n} is fixed by this kernel for any $m > 0$ and hence by Ψ. As we demonstrate, $\mathscr{C}_F$ contains non-Bayesian systems; therefore the property of invariance under Ψ may be seen as a weak form of coherence.

When the basin of attraction of a system $\Theta \in \mathscr{C}_F$ fixed by Ψ contains a system of point estimators Θ_0, a correspondence between classical and non-classical approaches follows. When $\Theta \in \mathscr{C}_F$ is Bayesian, then a natural link is made between classical and Bayesian approaches. In the following section, we characterise for a general class of models the elements of $\mathscr{C}_F$ whose basins include systems of moment-based point estimators. In particular, we will show that for the case of the 1-dimensional exponential family with the mean-value parameterisation, the maximum-likelihood prior of Hartigan (1998) can be obtained via this correspondence.

We define an alternative refinement operator, and corresponding constructions, by taking limits with respect to m and k in a different manner. Consider the new operator Φ for which

$$p_{x_n,i+1} = \lim_{k\to\infty}\lim_{m\to\infty} p_{x_n,i}[P^{(i,m)}(x_n)]^{mk}. \tag{3}$$

We may expect, given sufficiently strong conditions on the model, that the same sequence of systems of inference will arise from the above construction if Ψ or Φ is used; this is the case for the class of models considered in Theorem 3.2. At points in the paper, it will be convenient to work with the operator Φ defined by (3).

3 Moment-based estimators and fixed points

We retain the notation of the previous section and let $y_1, y_2, y_3, \ldots$ denote a sequence of i.i.d observations from a measure with density $\nu_\theta(y)$ with a 1-dimensional parameter $\theta \in (l, r)$ (where $l, r \in [-\infty, \infty]$) and let $x_n = (y_1, ..., y_n)$. We suppose that $\nu_\theta(y)$ has mean θ and variance $\sigma^2(\theta)$ and satisfies Assumption 3.1.

Assumption 3.1. *We assume that the following conditions hold:*

1. *The function σ^2 is locally Lipschitz continuous in that there is a constant K_U such that*

$$|\sigma^2(\theta) - \sigma^2(\theta')| \leq K_U |\theta - \theta'|, \;\; \forall |\theta|, |\theta'| < U.$$

2. *The function σ^2 satisfies a linear growth condition in that there exists a constant C_l, which we assume satisfies $C_l < \sqrt{2}n$, such that*

$$\sigma^2(\theta) \leq C_l(1 + \theta^2).$$

3. *There exists an $\epsilon > 0$ such that*

$$\int (x - \theta)^{2+\epsilon} \nu_\theta(x) dx < \infty.$$

We now investigate those $\Theta \in \mathscr{C}_F$ whose basins of attraction contain moment-based point estimators. The next result generalises Gibson et al. (2011), Example 2.3, which considered the special case of the Normal distribution. We write $f(\theta), g(\theta)$, for $\theta \in (l, r)$, for the indefinite integral of $\sigma^{-2}(\theta)$ and $\theta\sigma^{-2}(\theta)$ respectively. We note that by the Lipschitz continuity these functions are locally integrable at θ whenever $\sigma(\theta) > 0$.

Theorem 3.2. *Suppose that ν_θ satisfies Assumption 3.1 and let*

$$\Theta_0 = \{p_{x_n,0}(\theta) = \delta(\theta - \bar{x}_n) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

For $i = 1, 2, 3, \ldots,$ let $c_i = 2 - 2^{-(i-1)}$. Then the systems $\Theta_i,\ i = 1, 2, 3, ...$ exist and are given by

$$\Theta_i = \{p_{x_n,i}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\{\frac{2}{c_i} n(f(\theta)\bar{x}_n - g(\theta))\} \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

Moreover, the limiting system Θ_∞ is specified by

$$\Theta_\infty = \left\{ p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\{n(f(\theta)\bar{x}_n - g(\theta))\} \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n \right\}.$$

The proof is given in Section 4. It exploits the property that, as $m \to \infty$ the generalised data-augmentation chains that arise converge weakly to solutions to stochastic differential equations whose stationary measures can be identified.

We now consider the conditions for Θ_∞ to be a Bayesian system, and the nature of the corresponding prior.

Corollary 3.3. *Under Assumption 3.1, Θ_∞ is Bayesian if and only if $\nu_\theta(y)$ is a member of the 1-parameter exponential family (with the mean-value parameterisation) and sufficient statistic $\bar{x}_n$.*

Proof. Clearly Θ_∞ is Bayesian only if the likelihood $\nu_\theta(x_n)$ satisfies

$$\nu_\theta(x_n) = K_1(x_n) K_2(\theta) \exp\{n(f(\theta)\bar{x}_n - g(\theta))\}.$$

identifying it as a member of the 1-parameter exponential family with mean value θ and sufficient statistic $\bar{x}_n$.

Conversely, suppose that $\nu_\theta(x)$ is a density from the 1-parameter exponential family with sufficient statistic x, mean θ and canonical parameter $a(\theta)$, then

$$\nu_\theta(x) = K(x) \exp\{a(\theta)x - c(\theta)\}.$$

From the score function $a'(\theta)x - c'(\theta)$ we obtain the information function $i(\theta) = \sigma^{-2}(\theta) = a'(\theta)$ implying that $a(\theta) = \int \sigma^{-2} d\theta$ and $c'(\theta) = a'(\theta)\theta$ in which case $c(\theta) = \int \theta\sigma^{-2}(\theta) d\theta$.

It follows that $p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\left(n(a(\theta)\bar{x}_n - c(\theta)\right\}$; hence Θ_∞ represents a Bayesian analysis with prior density $\pi(\theta) \propto \sigma^{-2}(\theta)$. Note that $\pi(\theta) \propto \sigma^{-2}(\theta)$ induces a uniform measure on the canonical parameter $a(\theta)$. This corresponds to the maximum-likelihood prior distribution of Hartigan (1998). □

For the 1-parameter exponential family with mean-value parameterisation and sufficient statistic $\bar{x}_n$, Bayesian analyses with alternative priors from the conjugate family are obtained by specifying Θ_0 appropriately in the construction. Given prior experience of a sample of size k with mean value a, then a natural system of point estimators is

$$\Theta_0 = \{p_{x_n,0}(\theta) = \delta(\theta - \frac{n\bar{x}_n + ka}{n+k}) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

In this case Θ_∞ corresponds to a Bayesian analysis using the the prior

$$\pi(\theta) \propto \sigma^{-2}(\theta) \exp\{k(f(\theta)a - g(\theta)\},$$

and the correspondence between 'shrinkage' estimators and the choice of conjugate prior is obtained.

We now discuss distributions outside the 1-parameter exponential family. In this case, Theorem 3.2 demonstrates that $\mathscr{C}_F$ contains both Bayesian and non-Bayesian systems of inference that are fixed by Ψ. As in the exponential-family case, Theorem 3.2 predicts that the general system of estimators for which

$$p_{x_n,0}(\theta) = \delta(\theta - \frac{n\bar{x}_n + ka}{n+k})$$

lies in the basin of attraction of the fixed point for which

$$p_{x_n,\infty}(\theta) \propto \sigma^{-2}(\theta) \exp\{k(f(\theta)a - g(\theta)\} \times \exp\{n(f(\theta)\bar{x}_n - g(\theta))\}.$$

The first and second factors play roles analogous to a 'prior' density and a pseudo-likelihood respectively. In particular, the pseudo-likelihood $\exp\{n(f(\theta)\bar{x}_n - g(\theta))\}$ may be considered to approximate the true likelihood with one of exponential-family form. Since $\bar{x}_n$ is not generally sufficient for θ, then $p_{x_n,\infty}(\theta)$ may not coincide with $\pi(\theta|x_n)$ for any prior $\pi(\theta)$. Nevertheless, we might expect $p_{x_n,\infty}(\theta)$ to give a reasonable approximation to $\pi(\theta|\bar{x}_n)$ for some $\pi(\theta)$. We illustrate this in the following examples.

Example 3.4. The double exponential distribution has density given by

$$\nu_\theta(x) = \frac{1}{2}\exp\{-|x-\theta|\},\ x \in \mathbb{R}$$

with mean given by θ and constant variance 2. In this case, Theorem 3.2 states that when $p_{x_n,0}(\theta) = \delta(\theta - \bar{x}_n)$

$$p_{x_n,\infty}(\theta) \propto \exp\{-\frac{n}{4}(\bar{x}_n - \theta)^2\}.$$

Clearly, for large sample sizes, this is 'close' to a Bayesian analysis, with improper uniform prior, for an experiment recording the sample mean $\bar{x}_n$ since the likelihood $\nu_\theta(\bar{x}_n)$ can be approximated by the density of $N(\theta, \frac{2}{n})$.

Example 3.5. The Uniform(0, 2θ) distribution has mean θ and variance $\sigma^2(\theta) = \frac{\theta^2}{3}$, and satisfies Assumption 3.1. In this case, Theorem 3.2 states that when $p_{x_n,0}(\theta) = \delta(\theta - \bar{x}_n)$

$$\Theta_\infty = \{p_{x_n,\infty} \propto \theta^{-3n-2}\exp(-3n\bar{x}_n/\theta)|n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

so that $p_{x_n,\infty}(\theta) \sim$ IGamma$(3n+1, 3n\bar{x}_n)$. We compare the density $p_{x_n,\infty}(\theta)$ with the Bayesian posterior density $\pi(\theta|\bar{x}_n)$ for the prior $\pi(\theta) \propto \sigma^{-2}(\theta) \propto \theta^{-2}$.

The likelihood $L(\theta;\ \bar{x}_n)$ is not convenient to work with directly being proportional to the $(n-1)$-dimensional volume $V(A)$ of the set

$$A = \left\{(y_1, y_2, , ..., y_n) \in \mathbb{R}^n \mid \sum y_i = n\bar{x}_n\right\} \cap [0, 2\theta]^n.$$

Therefore we estimate $\pi(\theta|\bar{x}_n)$ using Gibbs sampling, treating the unobserved $y_1, ...y_n$ as additional unknown parameters.

From Figure 1 we see that $p_{x_n,\infty}(\theta)$ approximates the Bayesian posterior $\pi(\theta|\bar{x}_n)$ in the case where $n = 30$. Thus, although not precisely Bayesian, Θ_∞ represents a system which makes use of knowledge of the sample mean in an approximately Bayesian manner.

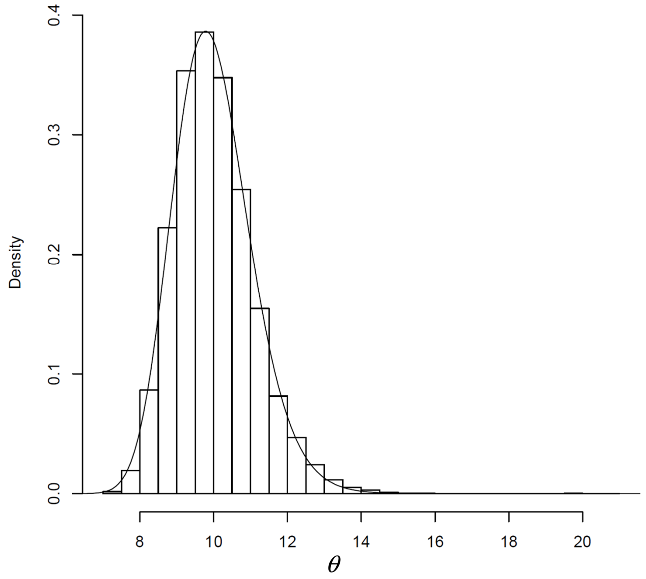

Figure 1: Comparison between $\pi(\theta|\bar{x}_n = 10)$ as estimated by Gibbs sampling and $p_{x_n,\infty}(\theta)$ for sample size $n = 30$.

4 Proof of Theorem 3.2

We retain the notation of earlier sections and consider the family of probability measures with density ν_θ where the parameter space is a subset $\mathcal{K}$ of $\mathbb{R}$. Recall that θ is the distribution mean and $\sigma^2(\theta)$ the variance. Under the conditions of Assumption 3.1, we will show, using an induction argument, that the construction of Theorem 3.2 indeed converges to the system Θ_∞ given in the statement of the theorem. In Section 4.1 we consider the first step whereby Θ_1 is constructed from Θ_0, before considering the inductive step in Section 4.2. First we give some key auxiliary results required for the proof.

Suppose that we have already constructed the systems $\Theta_0, ..., \Theta_{i-1}$. Now fix n, and the observed sample x_n, and consider the Markov chain $\theta_{i,m}^{(k)}$ with transition kernel $P^{(i-1,m)}$ as described after Principle 2.2. We assume that, for all possible observed samples x_n, the sample mean $\bar{x}_n$ is constrained to lie in the allowable parameter space $\mathcal{K}$. We first define a continuous-time process from $\theta_{i,m}^{(k)}$ by interpolation, setting $\theta_i^m(t) = \theta_{i,m}^{(\lfloor mt \rfloor)}$, noting that $\theta_{i,m}^{(k)}$ and $\theta_i^m(t)$ are identical so far as the existence and nature of the stationary distribution is concerned.

The main work in proving Theorem 3.2 lies in establishing the following theorem. We will write $\sigma_i(\theta) = \sqrt{2 - 2^{-(i-1)}}\sigma(\theta)$.

Theorem 4.1. *For $i = 1, 2, \ldots,$ under Assumption 3.1, there exists a unique solution $\theta_i = \{\theta_i(t); t \geq 0\}$ to the one dimensional stochastic differential equation*

$$\begin{aligned} d\theta_i &= n(\bar{x}_n - \theta_i)dt + \sigma_i(\theta_i)dW^i. \qquad (4) \\ \theta_i(0) &= \xi \end{aligned}$$

where W^i is a standard Brownian motion and $\xi \in \mathcal{K}$.

For each i, we have that the sequence of Markov chains θ_i^m, with $\theta_i^m(0) = \xi$, converges weakly to the diffusion process θ_i as $m \to \infty$.

Note that it is enough to have a weak solution to the equation (4) for our purposes. This result in essence enables one to demonstrate that the second refinement operator Φ, introduced in Section 2, has Θ_∞ in Theorem 3.2 as a fixed point. With a little more work we can deduce the following version of Theorem 3.2 to show that Θ_∞ is the fixed point of Ψ as required by the Theorem.

Theorem 4.2. *The Markov chains θ_i^m have stationary distributions π_i^m, the diffusion process θ_i has a stationary distribution π_i and*

$$\pi_i^m \to \pi_i, \quad \textit{weakly as } m \to \infty.$$

A consequence of Theorem 4.1 is that we can characterise the limiting systems arising from the generalised data-augmentation constructions by considering the properties of diffusion processes. Concerning these properties we have the following results.

Lemma 4.3. *(1) There exists a unique solution to the SDE (4), $\{\theta_i(t) : t \geq 0\}$ which can be written in integral form as*

$$\theta_i(T) = \bar{x}_n + (\xi - \bar{x}_n)e^{-nT} + \int_0^T e^{-n(T-t)}\sigma_i(\theta_i(t))dW_t. \tag{5}$$

(2) The moments of $\theta_i(t)$ are bounded up to a level depending on n in that there exist constants C_k such that $E|\theta_i(t)|^k \leq C_k(1 \vee |\bar{x}_n|^k \vee |\xi|^k)$ for all $t \geq 0$ and $1 \leq k \leq (2n + C_l)/C_l$.

(3) If $\sqrt{2}n > C_l$, the stationary distribution of (4) exists and is given by

$$p_i(\theta) \propto \frac{1}{\sigma_i(\theta)^2} \exp(2n(f_i(\theta)\bar{x}_n - g_i(\theta))),$$

where

$$f_i(\theta) = \int \sigma_i^{-2}(\theta)d\theta, \;\; g_i(\theta) = \int \theta\sigma_i^{-2}(\theta)d\theta.$$

The proof of this lemma can be found in the Appendix. Together Theorems 4.1, 4.2 and Lemma 4.3 lead to the following result.

Corollary 4.4. *In the construction of Theorem 3.2 the limiting system of inferences has a density given by*

$$p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma(\theta)^2} \exp(n(f(\theta)\bar{x}_n - g(\theta))),$$

where

$$f(\theta) = \int \sigma^{-2}(\theta)d\theta, \;\; g(\theta) = \int \theta\sigma^{-2}(\theta)d\theta.$$

A key tool in establishing Theorem 4.1 is Corollary 7.4.2 of Ethier and Kurtz (1986), which specifies conditions sufficient for the existence of a diffusion approximation to a sequence of Markov chains. We state a version of the result suited to our purposes. We denote by $\mathcal{P}(\mathbb{R})$ the set of probability measures on $\mathbb{R}$.

Theorem 4.5 (Ethier and Kurtz)**.** *Let $X = \{X(t); t \geq 0\}$ be a diffusion process satisfying the SDE*

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X(0) \sim \phi$$

where b is a continuous function and σ is also continuous and X_0 is drawn according to a measure $\phi \in \mathcal{P}(\mathbb{R})$. Let $Y_m = \{Y_m(k); k \geq 0\}$ be a discrete time Markov chain and set $X_m(t) = Y_m([mt])$. Let

$$\begin{aligned} \mu_m(x) &= m\mathbb{E}(Y_m(1) - x) \\ \sigma_m^2(x) &= m\mathbb{E}(Y_m(1) - x)^2 \end{aligned}$$

Suppose that the law of $X_m(0)$ converges weakly to ϕ and that for each $r > 0$ and $\epsilon > 0$ we have

$$\lim_{m\to\infty} \sup_{|x|<r} |\mu_m(x) - b(x)| = 0, \tag{6}$$

$$\lim_{m\to\infty} \sup_{|x|<r} |\sigma_m^2(x) - \sigma^2(x)| = 0, \tag{7}$$

and

$$\lim_{m\to\infty} \sup_{|x|<r} m\mathbb{P}(|Y_m(1) - x| \geq \epsilon) = 0. \tag{8}$$

Then X_m converges weakly to X.

We are now in a position to proceed with the inductive proof of Theorem 3.2.

4.1 The case $i = 1$.

We note that here and throughout the paper the notation c, c' will be used to denote arbitrary constants which may change from line to line, whereas labelled constants with an upper case C will be fixed. When our observation x_n is augmented by a further m samples we obtain a generalised data-augmentation chain with updates specified by

$$\theta_{1,m}^{(k+1)} = \frac{n\bar{x}_n + m\bar{Y}_m^{(k)}}{n+m},$$

where $\bar{Y}_m^{(k)} = \frac{1}{m}\sum_{j=1}^m Y_j^{(k)}$, with $Y_j^{(k)}$ samples from the measure with density $\nu_{\theta_{1,m}^{(k)}}$. To simplify the notation we suppress the subscript $i = 1$ and write $\theta_{1,m}^{(k)}$ as $\theta_m^{(k)}$. We now establish the conditions of Theorem 4.5 with $\theta_m^{(k)}$ in place of $Y_m(k)$ and with the limiting diffusion process given by (4).

Suppose now that $\theta_m^{(0)} = x$, then

$$\theta_m^{(1)} = x + \frac{n}{n+m}(\bar{x}_n - x) + \frac{R_m(x)}{n+m},$$

where $R_m(x) = \sum_{j=1}^m (Y_j^{(0)} - x)$, with $Y_j^{(0)}$ independent and identically distributed with mean x. Thus

$$\mu_m(x) = m\mathbb{E}(\theta_m^{(1)} - x) = \frac{nm}{n+m}(\bar{x}_n - x),$$

and hence with $\mu(x) = n(\bar{x}_n - x)$ we have

$$\sup_{|x|<r} |\mu_n(x) - \mu(x)| = \sup_{|x|<r} \left| \frac{n^2(\bar{x}_n - x)}{n+m} \right| \to 0, \text{ as } m \to \infty,$$

which establishes (6).

By construction we have

$$\begin{aligned}
\sigma_m^2(x) &= m\mathbb{E}(\theta_m^{(1)} - x)^2 \\
&= m\mathbb{E}\left(\frac{n}{n+m}(\bar{x}_n - x) - \frac{1}{n+m}R_m(x)\right)^2 \\
&= m\left(\left(\frac{n}{n+m}\right)^2 (\bar{x}_n - x)^2 + \frac{1}{(n+m)^2}\mathbb{E}R_m(x)^2\right) \\
&= \frac{mn^2}{(n+m)^2}(\bar{x}_n - x)^2 + \left(\frac{m}{n+m}\right)^2 \sigma^2(x).
\end{aligned}$$

Thus $\sigma_m^2(x) \to \sigma^2(x)$ as $m \to \infty$. We establish our condition (7) as

$$\sup_{|x|<r} |\sigma_m^2(x) - \sigma^2(x)| \leq \sup_{|x|<r} \left(\frac{mn^2}{(n+m)^2} (\bar{x}_n - x)^2 + \left(\frac{2mn + n^2}{(n+m)^2} \right) \sigma^2(x) \right) \to 0, \text{ as } m \to \infty.$$

To handle the tail condition (8) we observe that by Assumption 3.1 there is an $\epsilon > 0$ such that $\mathbb{E}(Y_1 - x)^{2+\epsilon} < \infty$. Letting $p = 2 + \epsilon$ we see that

$$\mathbb{E}|\theta^{(1)} - x|^p \leq 2^{p-1} \left(\left(\frac{n}{n+m} \right)^p |\bar{x}_n - x|^p + \frac{\mathbb{E}R_m(x)^p}{(n+m)^p} \right).$$

As $R_m(x)$ is the value at time m of a discrete martingale, we can apply the Burkholder-Davis-Gundy inequality to see that

$$\begin{aligned} \mathbb{E}|R_m(x)|^p &= \mathbb{E} \left| \sum_{i=1}^m (Y_i^{(0)} - x) \right|^p \\ &\leq c_p \mathbb{E} \left| \sum_{i=1}^m (Y_i - x)^2 \right|^{p/2} \\ &\leq c_p m^{p/2-1} \mathbb{E} \sum_{i=1}^m |Y_i - x|^p \\ &= c_p m^{p/2} \mathbb{E}|Y_1^{(0)} - x|^p = C_p m^{p/2}. \end{aligned} \tag{9}$$

Thus we have, by Markov's inequality and (9),

$$\begin{aligned} m\mathbb{P}(|\theta_m(1) - x| \geq \epsilon) &\leq \sup_{|x|<r} m \frac{\mathbb{E}|\theta^{(1)} - x|^p}{\epsilon^p} \\ &\leq \sup_{|x|<r} \left(\frac{2^{p-1} m n^p}{(n+m)^p} |\bar{x}_n - x|^p + \frac{2^{p-1} m^{p/2+1} C_p}{(n+m)^p} \right) \\ &\leq C_1 m^{1-p} + C_2 m^{1-p/2}, \end{aligned}$$

which tends to 0 as $m \to \infty$ since $p > 2$.

Thus we have satisfied the conditions of Theorem 4.5 and we have proved

Proposition 4.6. *Under Assumption 3.1 the process θ_1^m converges weakly to θ_1, the pathwise unique strong solution to*

$$d\theta_1 = n(\bar{x}_n - \theta_1)dt + \sigma(\theta_1)dW.$$

with $\theta_1(0) = \xi \in \mathcal{K}$.

From Lemma 4.3 (3) we have established the form of Θ_1 as given in Theorem 3.2.

4.2 The inductive step

To complete our induction we need to consider the general case. We assume that we have generated the system of inferences up to i. Again we fix n, and the observed sample x_n, and consider the Markov chain $\theta^{(k)}_{i+1,m}$ with transition kernel $P^{(i,m)}$ as described after Principle 2.2, where $\theta^{(k+1)}_{i+1,m} \sim p_{i,x_{n+m}}(\theta)$ with the augmented sample $x_{n+m} \sim \nu_{\theta^{(k)}_{i+1,m}}(x_{n+m}|x_n)$.

The draw $\theta^{(k+1)}_{i+1,m} \sim p_{i,x_{n+m}}(\theta)$ is a sample from the stationary distribution $p_{i,x_{n+m}}(\theta)$ of the diffusion $\theta^{n,m,k}_i$, given by (4) with $n+m$ and $\bar{x}_{n+m}$ replacing n and $\bar{x}_n$ respectively. An approximate sample $\theta^{n,m,k}_i(\tau_m)$ can be obtained by running the diffusion, with initial value $\theta^{n,m,k}_i(0) = \bar{x}_{n+m}$, for a sufficiently long time τ_m. We note that

$$\bar{x}_{n+m} = \frac{n\bar{x}_n + m\bar{Y}_m(k)}{n+m},$$

where $\bar{Y}_m(k) = \sum_{i=1}^m Y_i$ and the Y_i are i.i.d. samples with mean $\theta^{(k)}_{i+1,m}$.

Thus we have an approximate Markov chain given by

$$\theta^{(k+1)}_{i+1,m} = \theta^{(k)}_{i+1,m} + \frac{n}{n+m}(\bar{x}_n - \theta^{(k)}_{i+1,m}) + \frac{1}{m+n} R_m(\theta^{(k)}_{i+1,m}) + \int_0^{\tau_m} e^{-(n+m)(\tau_m - t)} \sigma_i(\theta^{n,m,k}_i(t)) dW^k_t, \tag{10}$$

where the first two terms in the expression appeared in the previous case ($i = 1$).

For the one-step evolution of our Markov chain, from initial state x, we can write (10) as

$$\theta^{(1)}_{i+1,m} = x + \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{m+n} R_m(x) + N_m(0)(\tau_m),$$

where

$$N_m(0)(\tau_m) = \int_0^{\tau_m} e^{-(n+m)(\tau_m - t)} \sigma_i(\theta^{n,m,0}_i(t)) dW_t.$$

We will write $\theta^{n,m}_i$ for $\theta^{n,m,0}_i$ and $N_m(t)$ for $N_m(0)(t)$. We also note that $\exp((n+m)t)N_m(t)$ is a continuous local martingale and in the proof of the moment estimates in Lemma 4.3 we showed that it is in fact an L^2 bounded martingale under our assumption $0 \le t \le \tau_m$. Abusing notation we continue to use $\mathbb{E}$ both for expectation with respect to the probability measure governing the diffusion as well as that for the augmented sample. We note that the Brownian motion driving $N_m(\tau_m)$ is independent of $R_m(x)$ so that we can treat the term $N_m(\tau_m)$ separately. The quadratic variation process for $N_m(t)$ is given by

$$\langle N_m \rangle_t = \int_0^t e^{-2(n+m)(t-s)} \sigma_i^2(\theta_i^{n,m}(s)) ds.$$

Thus we note $\mathbb{E}N_m(t) = 0$ and

$$\mathbb{E}N_m(t)^2 = \mathbb{E}\langle N_m \rangle_t = \int_0^t e^{-2(n+m)(t-s)} \mathbb{E}\sigma_i^2(\theta_i^{n,m}(s)) ds. \tag{11}$$

Proposition 4.7. *Under Assumption 3.1 and for $\tau_m > \log m/2(m+n)$, the process $\theta_i^m(t)$ converges weakly to $\theta_i(t)$, the pathwise unique strong solution to*

$$d\theta_i = n(\bar{x}_n - \theta_i)dt + \sigma(\theta_i)dW.$$

with $\theta_i(0) = \xi \in \mathcal{K}$.

Proof. We establish the conditions of Theorem 4.5. Firstly the mean is given by

$$\mu_{m,i}(x) = m\mathbb{E}(\theta_{i+1,m}^{(1)} - x) = \frac{mn}{m+n}(\bar{x}_n - x).$$

This is the same as in the $i = 1$ case and it therefore satisfies condition (6).

For the variance we have by independence and the fact that R_m and N_m are mean 0,

$$\begin{aligned}\sigma_{m,i+1}^2(x) &= m\mathbb{E}\left(\frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{m+n}R_m(x) + N_m(\tau_m)\right)^2 \\ &= \frac{mn^2}{(n+m)^2}(\bar{x}_n - x)^2 + \frac{m}{(n+m)^2}\mathbb{E}R_m^2(x) + m\mathbb{E}N_m^2(\tau_m).\end{aligned}$$

Recall that $\sigma_{i+1}^2(x) = \sigma^2(x) + \frac{1}{2}\sigma_i^2(x)$. Thus we can write

$$|\sigma_{m,i}^2(x) - \sigma_{i+1}^2(x)| \leq \frac{n^2 m}{(n+m)^2}(\bar{x}_n - x)^2 + |\frac{m}{(m+n)^2}\mathbb{E}R_m(x)^2 - \sigma^2(x)| + |m\mathbb{E}N_m^2(\tau_m) - \frac{1}{2}\sigma_i^2(x)|.$$

From the calculations in the $i = 1$ case we can control the first two terms to show that they go to 0 as $m \to \infty$ on the region where $|x| < r$.

For the last term we need to do some work. Firstly we observe that

$$\begin{aligned}
&|\mathbb{E}mN_m^2(\tau_m) - \frac{\sigma_i^2(x)}{2}| \\
&= \frac{m}{2(m+n)}| \left(\mathbb{E}\int_0^{\tau_m} 2(n+m)e^{-2(n+m)(\tau_m - t)}\sigma_i^2(\theta_i^{n,m}(t))dt - \frac{m+n}{m}\sigma_i^2(x)\right) | \\
&\leq \frac{m}{2(n+m)}\int_0^{\tau_m} 2(m+n)e^{-2(n+m)(\tau_m-t)}\mathbb{E}\left|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)\right| dt \\
&\quad + \frac{n}{2(m+n)}\sigma_i^2(x) + \mathbb{E}e^{-2(n+m)\tau_m}\sigma_i^2(x). \qquad (12)
\end{aligned}$$

Now, as $\sigma_i^2(\theta)$ is locally Lipschitz, we have a constant K_U such that for $|x| < U$

$$\begin{aligned}
\mathbb{E}|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)| &\leq K_U\mathbb{E}\left(|\theta_i^{n,m}(t) - x|; |\theta_i^{n,m}(t)| < U\right) \\
&\quad + \mathbb{E}\left(|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; |\theta_i^{n,m}(t)| \geq U\right). \quad (13)
\end{aligned}$$

We can estimate the first term on the right hand side using

$$\theta_i^{n,m}(t) - x = \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{m+n}R_m(1) + N_m(t).$$

Taking expectations we have

$$\begin{aligned}
\mathbb{E}|\theta_i^{n,m}(t) - x| &\leq |\frac{n}{n+m}(\bar{x}_n - x)| + \mathbb{E}\frac{1}{m+n}|R_m(1)| + \mathbb{E}|N_m(t)| \\
&\leq O(\frac{1}{m}) + \frac{1}{m+n}(\mathbb{E}R_m(1)^2)^{1/2} + (\mathbb{E}|N_m(t)|^2)^{1/2}.
\end{aligned}$$

We observe that, by the linear growth condition and the bounds on $\mathbb{E}\theta_i^2$ from Lemma 4.3, and our sample for the diffusion starting at $\bar{x}_{n+m}$ we have

$$\begin{aligned}
\mathbb{E}N_m(t)^2 &\leq \int_0^t e^{-2(m+n)(t-s)}\mathbb{E}C_l(1 + \mathbb{E}(\theta_i^{n,m}(s))^2)ds \\
&\leq \int_0^t e^{-2(m+n)(t-s)}\mathbb{E}C_l(1 + C_2(1 \vee \bar{x}_{n+m}^2))ds \\
&\leq \frac{C_l(1 + c(1 \vee \bar{x}_n^2))}{m+n} = O(\frac{1}{m}),
\end{aligned}$$

where we have used that there is a constant such that $\mathbb{E}\bar{x}_{n+m}^2 \leq c\bar{x}_n^2$. From this, and the expression for $\mathbb{E}R_m(x)^2$, we know that there is a further constant c such that

$$\mathbb{E}|\theta_i^{n,m}(t) - x| \leq \frac{c}{\sqrt{m}}, \;\; \forall t \leq T.$$

For the second term on the right hand side of (13) we have by the linear growth bound, Hölder's and Markov's inequalities, that

$$\begin{aligned}\mathbb{E}\left(|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; |\theta_i^{n,m}(t)| \geq U\right) &\leq (C_l + \sigma_i^2(x))\mathbb{P}(|\theta_i^{n,m}(t)| \geq U) \\ &\quad + C_l\mathbb{E}\left(|\theta_i^{n,m}(t))|^2 I_{\{|\theta_i^{n,m}(t)| \geq U\}}\right) \\ &\leq (C_l + \sigma_i^2(x))U^{-p}\mathbb{E}|\theta_i^{n,m}(t)|^p + C_l U^{2-p}\mathbb{E}|\theta_i^{n,m}(t)|^p\end{aligned}$$

From the estimates for the diffusion started from $\bar{x}_{n+m}$ in Lemma 4.3(2) and following (9) we have $\mathbb{E}|\theta_i^{n,m}(t)|^p \leq c_p\mathbb{E}(1 \vee |\bar{x}_{n+m}|^p) \leq c_p' m^{-p/2}$ and hence

$$\mathbb{E}\left(|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; |\theta_i^{n,m}(t)| \geq U\right) = O(m^{-p/2}).$$

Thus, substituting into (12) and using our condition on τ_m we have $e^{-2(m+n)\tau_m} \leq 1/m$, which gives

$$\mathbb{E}|mN_m^2 - \frac{\sigma_i^2(x)}{2}| \leq c/m$$

and we have the result.

To show the last condition of Theorem 4.5, as in the $i = 1$ case, we need a little more than second moments. For this we note that by Burkholder-Davis-Gundy and Hölder's inequality

$$\begin{aligned}\mathbb{E}N_m(\tau_m)^p &\leq c_p\mathbb{E}\langle N_m\rangle_{\tau_m}^{p/2} \\ &= c_p\mathbb{E}\left(\int_0^{\tau_m} e^{-2(n+m)(\tau_m - s)}\sigma_i^2(\theta_i^{m,n}(s))ds\right)^{p/2} \\ &\leq c_p(\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/(p-2)}ds)^{p/2-1}\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \\ &= c_p(\frac{p-2}{p(n+m)})^{p/2-1}(1 - e^{-p(n+m)\tau_m/(p-2)})^{p/2-1}\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \\ &\leq c_p(\frac{p-2}{p(n+m)})^{p/2-1}\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds\end{aligned}$$

As we have a linear growth condition for σ and, by Assumption 3.1(3), the moments of the process $\theta_i^{m,n}$ exist, at least up to $p = 2 + \epsilon$, we have $\mathbb{E}\sigma^p(\theta_i^{m,n}(s)) \leq \mathbb{E}c_p(1 \vee |\bar{x}_{n+m}|^p) \leq c_p'$

for all s and m using (9). Thus

$$\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \leq c'_p\frac{2}{p(m+n)}(1-e^{-p(m+n)\tau_m/2}),$$

and hence for a constant C we have

$$\mathbb{E}N_m(\tau_m)^p \leq \frac{C}{(m+n)^{p/2}}.$$

Thus, by Markov's inequality, we have

$$\begin{aligned}
\sup_{|x|<r} m\mathbb{P}(|\theta_i^{(1)}-x|>\epsilon) &\leq \sup_{|x|<r}\frac{m\mathbb{E}|\theta_i^{(1)}-x|^p}{\epsilon^p} \\
&= \sup_{|x|<r}\frac{m\mathbb{E}\left(\frac{n}{n+m}(\bar{x}_n-x)+\frac{1}{m+n}R_m(x)+N_m(\tau_m)\right)^p}{\epsilon^p} \\
&\leq \sup_{|x|<r}\frac{c_p m\left(\frac{n^p}{(n+m)^p}(\bar{x}_n-x)^p+\frac{1}{(m+n)^p}\mathbb{E}R_m(x)^p+\mathbb{E}N_m(\tau_m)^p\right)}{\epsilon^p} \\
&\leq \frac{C_1}{\epsilon^p m^{p-1}}+\frac{C_2}{\epsilon^p m^{p/2-1}}+\frac{C_3}{\epsilon^p m^{p/2-1}}.
\end{aligned}$$

As $p>2$, this tends to 0 as $m\to\infty$ and we have the third condition of Theorem 4.5.

Proof of Theorem 4.1. We have established all the conditions of Theorem 4.5 and hence the weak convergence is proved. □

Proof of Theorem 4.2. In order to prove this theorem we need to show that the stationary distributions for the Markov chains converge to that of the diffusion.

In the appendix we use the Lyapunov function technique to show that the chains and the diffusions have stationary distributions $\pi_{i,m}$ and π_i and they are geometrically and exponentially ergodic respectively. The fact that the stationary distributions $\pi_{i,m}$ converge to π_i is a consequence of the weak convergence we have already established plus the geometric ergodicity. The proofs can be found in Theorem A.3 and Corollary A.4. □

5 Fixed points arising from maximum-likelihood estimation

5.1 The regular case

We consider the application of refinement operators to systems of maximum-likelihood rather than moment estimators, noting the coincidence of the two in the exponential-family setting. In particular we ask whether the construction leads to a Bayesian analysis beyond the exponential-family case and, if so, whether the maximum-likelihood prior is recovered. We find it convenient to work with the operator Φ rather than Ψ and to consider the fixed points of the former as the limiting - and maximally preferable - systems of inference. This avoids the need to derive any correspondence between these and the fixed points of Ψ as was done via Theorem 4.2 and Lemma 4.3 when proving Theorem 3.2. We will also make some stronger assumptions than in the moment-estimator case.

Retaining the notation from earlier sections, consider the model $\nu_\theta(y)$ and let $l(\theta, y) = \log \nu_\theta(y)$, $i(\theta) = \mathbb{E}_Y\left(-\frac{\partial^2 l}{\partial \theta^2}\right)$, $a(\theta) = \mathbb{E}_Y\left(\frac{\partial^2 l}{\partial \theta^2}\frac{\partial l}{\partial \theta}\right)$, and $c(\theta) = \mathbb{E}_Y\left(-\frac{\partial^3 l}{\partial \theta^3}\right)$. Suppose that ν_θ satisfies regularity conditions that allow change of order of integration with respect to y and differentiation with respect to θ. We can then easily verify the identity

$$\frac{\partial i(\theta)}{\partial \theta} + a(\theta) + c(\theta) = 0. \tag{14}$$

Now let

$$\Theta_0 = \{p_{x_n,0}(\theta) = \delta(\theta - \hat{\theta}(x_n)) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

where $\hat{\theta}(x_n)$ denotes the maximum-likelihood estimate. For observations $x_n = (y_1, ..., y_n)$ denote by $l_n(\theta)$ and $L_n(\theta)$ the resulting log-likelihood and likelihood respectively. Consider the data-augmentation chain $\{\theta_{1,m}^{(k)} \mid k = 0, 1, 2, ...\}$ arising in the construction of Θ_1 from Θ_0, where m is large. Denote the log-likelihood function for the additional m samples, generated when updating $\theta_{1,m}^{(k)}$, by $l_m(\theta)$ which is maximised by $\hat{\theta}_m$. As in the proof of Theorem 3.2 we are interested in the case where $\theta_1^m(t) = \theta_{1,m}^{(\lfloor mt \rfloor)}$ converges to a diffusion process as $m \to \infty$ and where $p_{x_n,1}(\theta)$ can then be derived as the stationary distribution of this process. We identify a candidate for this limiting diffusion by considering the increment $\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)}$ in the augmented chain. From standard results on the asymptotic mean, variance and normality of maximum-likelihood estimators for sufficiently regular models, (see e.g. Cox and Snell (1968)), we have

$$\mathbb{E}(\hat{\theta}_m - \theta_{1,m}^{(k)}) = \frac{1}{i^2(\theta_{1,m}^{(k)})m}\left(a(\theta_{1,m}^{(k)}) + \frac{c(\theta_{1,m}^{(k)})}{2}\right) + o(1/m)$$

and

$$\mathbb{E}(\hat{\theta}_m - \theta_{1,m}^{(k)})^2 = \frac{1}{i(\theta_{1,m}^{(k)})m} + o(1/m).$$

Since

$$\theta_{1,m}^{(k+1)} - \hat{\theta}_m = l_n'(\theta_{1,m}^{(k)})i(\theta_{1,m}^{(k)}) + o(1/m),$$

the form of the candidate limiting diffusion is given by

$$d\theta_1 = \frac{1}{i^2(\theta_1)}\left(a(\theta_1) + \frac{c(\theta_1)}{2} + l_n'(\theta_1)i(\theta_1)\right)dt + \sqrt{\frac{1}{i(\theta_1)}}dB. \tag{15}$$

We assume the following conditions hold.

Assumption 5.1.

1. *The stochastic differential equation*

$$d\theta_1 = \frac{1}{i^2(\theta_1)}\left(a(\theta_1) + \frac{c(\theta_1)}{2} + l_n'(\theta_1)i(\theta_1)\right)dt + \sqrt{\frac{1}{i(\theta_1)}}dB,$$

where B is a Brownian motion and $\theta_1(0) = \xi \in \mathcal{K}$, has a unique solution.

2. *The Markov chain $\theta_{1,m}^{(k)}$ satisfies*

$$\begin{aligned}\mathbb{E}(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)}) &= \frac{1}{i^2(\theta_{1,m}^{(k)})m}\left(a(\theta_{1,m}^{(k)}) + \frac{c(\theta_{1,m}^{(k)})}{2} + l_n'(\theta_{1,m}^{(k)})i(\theta_{1,m}^{(k)})\right) + o(1/m) \\ \mathbb{E}(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)})^2 &= \frac{1}{i(\theta_m^{(k)})m} + o(1/m) \\ \mathbb{E}(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)})^{2+\epsilon} &\leq \frac{C}{m^{1+\epsilon'}}\end{aligned}$$

Theorem 5.2. *1. Under Assumption 5.1, the interpolated chain θ_1^m converges weakly to θ_1.*

2. The associated system of inferences, obtained from the stationary measure of the diffusion in (15)*, is given by*

$$\Theta_1 = \{p_{x_n,1}(\theta) \propto \pi(\theta)L_n(\theta;x_n)^2 \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$$

where $\frac{\partial}{\partial\theta}\log\pi = \frac{a(\theta)}{i(\theta)}$.

Proof. Under Assumption 5.1 we can satisfy the conditions of Theorem 4.5 and hence we can deduce the $i = 1$ case, in a manner analogous to the Proof of Theorem 3.2 given in Section 4. To verify the form of Θ_1 we solve the associated Fokker-Planck equation to show that the stationary measure, $p(\theta)$, satisfies

$$p(\theta) \propto i(\theta)\exp\left(2\int \frac{a(\theta)}{i(\theta)} + \frac{c(\theta)}{2i(\theta)} + l_n'(\theta)d\theta\right),$$

From (14) this can be written as

$$\begin{aligned}p(\theta) &\propto i(\theta)\exp\left(2\int \frac{a(\theta)}{2i(\theta)} - \frac{i'(\theta)}{2i(\theta)} + l_n'(\theta)d\theta\right) \\ &\propto \exp\left(\int \frac{a(\theta)}{i(\theta)}d\theta\right)L_n^2(\theta).\end{aligned}$$

It follows that

$$\Theta_1 = \{p_{x_n,1}(\theta) \propto \pi(\theta)L_n^2(\theta;x_n) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

where $\frac{\partial}{\partial\theta}\log\pi = \frac{a(\theta)}{i(\theta)}$, so that $\pi(\theta)$ is the maximum-likelihood prior. □

We proceed to the inductive step. Suppose now that

$$\Theta_{j-1} = \{p_{x_n,j}(\theta) \propto \pi(\theta)L_n^{\gamma_{j-1}}(\theta; x_n) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

where $\gamma_i = 1/(1-2^{-i}), i = 1, 2, \ldots$, and consider the construction of Θ_j from Θ_{j-1}. Let, x_n, $L_n(\theta)$ and $l_n(\theta)$ be as above and let $\pi_L = \pi(\theta)L_n^{\gamma_{j-1}}(\theta)$.

Consider the chain $\theta_{j,m}^{(k)}$ and the continuous-time interpolation $\theta_j^m(t) = \theta_{j,m}^{(\lfloor mt \rfloor)}$. We seek the form of a limiting diffusion for this process and consider, therefore, the increment to $\theta_{j,m}^{(k)}$ when m is large. As before $\hat{\theta}_m$ denotes the MLE for θ given the m additional samples generated using the current value $\theta_{j,m}^{(k)}$, and $L_m(\theta)$ denotes the likelihood function for these samples. We appeal to standard results regarding the asymptotic Bayesian posterior distribution of θ about the MLE, $\hat{\theta}_m$ for regular models.

From Chapter 5 of Ghosh (1994) it follows that, using prior density $\pi_L(\theta)$, and given observations $y_1, ..., y_m$, then, the posterior $\pi(\theta|y_1, ..., y_m) \propto \pi_L(\theta)L_m(\theta)$ satisfies

$$\mathbb{E}(\theta - \hat{\theta}_m|y_1, ..., y_m) = \left[-\frac{\partial^2 l_m}{\partial\theta^2}\right]_{\hat{\theta}_m}^{-2} \left(\frac{1}{2}\left[\frac{\partial^3 l_m}{\partial\theta^3}\right]_{\hat{\theta}_m} - \left[\frac{\partial^2 l_m}{\partial\theta^2}\right]_{\hat{\theta}_m}\left[\frac{\partial \log \pi_L}{\partial\theta}\right]_{\hat{\theta}_m}\right) + o(1/m) \tag{16}$$

and has variance $\left[-\frac{\partial^2 l_m}{\partial\theta^2}\right]^{-1} + o(1/m)$. Replacing $L_m(\theta)$ with $L_m(\theta)^{\gamma_{j-1}}$, so that $\hat{\theta}_m$ is unaffected, the corresponding expectation for the 'posterior' with density proportional to $\pi_L(\theta)L_m(\theta)^{\gamma_{j-1}}$ becomes

$$\mathbb{E}(\theta - \hat{\theta}_m|y_1, ..., y_m) = \frac{1}{\gamma_{j-1}}\left[-\frac{\partial^2 l_m}{\partial\theta^2}\right]_{\hat{\theta}_m}^{-2} \left(\frac{1}{2}\left[\frac{\partial^3 l_m}{\partial\theta^3}\right]_{\hat{\theta}_m} - \left[\frac{\partial^2 l_m}{\partial\theta^2}\right]_{\hat{\theta}_m}\left[\frac{\partial \log \pi_L}{\partial\theta}\right]_{\hat{\theta}_m}\right) \tag{17}$$

with approximate variance is $\frac{1}{\gamma_{j-1}}\left[-\frac{\partial^2 l_m}{\partial\theta^2}\right]^{-1}$.

We discern two components to the increment $\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)}$ - one given by $\hat{\theta}_m - \theta_{j,m}^k$ and the other arising when we sample $\theta_{j,m}^{k+1}$ from a density proportional to $\pi_L(\theta)L_m(\theta)_{j-1}^{\gamma}$. We approximate expectations over $y_1, ..., y_m$ in (17) by setting the derivatives of l_m to their expected values at $\theta = \theta_{j,m}^{(k)}$, and combine the two increments to show that

$$\mathbb{E}(\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)}) = \frac{1}{i^2(\theta_{j,m}^{(k)})} \left(a(\theta_{j,m}^{(k)}) + \frac{c(\theta_{j,m}^{(k)})(\gamma_{j-1}+1)}{2\gamma_{j-1}} + \frac{i(\theta_{j,m}^{(k)})}{\gamma_{j-1}} \left[\frac{\partial \log \pi_L}{\partial \theta} \right]_{\theta_{j,m}^{(k)}} \right) \frac{1}{m} + o(1/m)$$

and

$$\mathrm{Var}(\theta_{j,m}^{(k+1)} | \theta_{j,m}^{(k)}) = \frac{\gamma_{j-1}+1}{i(\theta_{j,m}^{(k)})\gamma_{j-1}} \frac{1}{m} + o(1/m).$$

We can now discern the candidate for the limiting diffusion to be

$$d\theta_j = \frac{1}{i^2(\theta_j)} \left(a(\theta_j) + \frac{c(\theta_j)(\gamma_{j-1}+1)}{2\gamma_{j-1}} + \frac{i(\theta_j)}{\gamma_{j-1}} \frac{\partial \log \pi_L}{\partial \theta_j} \right) dt + \sqrt{\frac{\gamma_{j-1}+1}{\gamma_{j-1} i(\theta_j)}} dB. \tag{18}$$

We make the following assumptions for $j \geq 2$.

Assumption 5.3.

1. *The stochastic differential equation*

$$d\theta_j = \frac{1}{i^2(\theta_j)} \left(a(\theta_j) + \frac{c(\theta_j)(\gamma_{j-1}+1)}{2\gamma_{j-1}} + \frac{i(\theta_j)}{\gamma_{j-1}} \frac{\partial \log \pi_L}{\partial \theta_j} \right) dt + \sqrt{\frac{\gamma_{j-1}+1}{\gamma_{j-1} i(\theta_j)}} dB,$$

where B is a Brownian motion and $\theta_j(0) = \xi \in \mathcal{K}$, has a unique solution.

2. *The Markov chain $\theta_{j,m}^{(k)}$ satisfies*

$$\begin{aligned} \mathbb{E}(\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)}) &= \frac{1}{i^2(\theta_{j,m}^{(k)})} \left(a(\theta_{j,m}^{(k)}) + \frac{c(\theta_{j,m}^{(k)})(\gamma_{j-1}+1)}{2\gamma_{j-1}} + \frac{i(\theta_{j,m}^{(k)})}{\gamma_{j-1}} \left[\frac{\partial \log \pi_L}{\partial \theta} \right]_{\theta_{j,m}^{(k)}} \right) \frac{1}{m} \\ &\quad + o(1/m) \\ \mathbb{E}(\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)})^2 &= \frac{\gamma_{j-1}+1}{i(\theta_{j,m}^{(k)})\gamma_{j-1}} \frac{1}{m} + o(1/m) \\ \mathbb{E}(\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)})^{2+\epsilon} &\leq \frac{C}{m^{1+\epsilon'}} \end{aligned}$$

Theorem 5.4. *Under Assumption 5.3, for $j \geq 2$, the process θ_j^m converges weakly to θ_j, the solution to* (18)*. The associated system Θ_j exists and is given by*

$$\Theta_j = \{ p_{x_n,j}(\theta) \propto \pi(\theta) L_n(\theta; x_n)^{\gamma_j} \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n \}$$

and the limiting system is given by

$$\Theta_\infty = \{p_{x_n,\infty}(\theta) \propto \pi(\theta) L_n(\theta; x_n) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$$

Proof. Since Assumption 5.3 implies that the conditions of Theorem 4.5 hold, it suffices to confirm that form of the stationary density, $p(\theta)$, which is given by

$$\begin{aligned}
\frac{\partial \log p}{\partial \theta} &= \frac{i'(\theta)}{i(\theta)} + \frac{2a(\theta)\gamma_{j-1}}{(\gamma_{j-1}+1)i(\theta)} + \frac{c(\theta)}{i(\theta)} + \frac{2}{\gamma_{j-1}+1}\frac{\partial \log \pi_L}{\partial \theta} \\
&= \frac{i'(\theta)}{i(\theta)} + \frac{2a(\theta)\gamma_{j-1}}{(\gamma_{j-1}+1)i(\theta)} + \frac{c(\theta)}{i(\theta)} + \frac{2}{(\gamma_{j-1}+1)}\left(\frac{a(\theta)}{i(\theta)} + \gamma_{j-1} l_n'(\theta)\right) \\
&= \frac{a(\theta)}{i(\theta)} + \frac{2\gamma_{j-1} l_n'(\theta)}{\gamma_{j-1}+1}
\end{aligned}$$

by (14). It follows that

$$\Theta_j = \{p_{x_n,i}(\theta) \propto \pi(\theta) L_n(\theta; x_n)^{\frac{2\gamma_{j-1}}{(\gamma_{j-1}+1)}} \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

and the result follows since $\frac{2\gamma_{j-1}}{(\gamma_{j-1}+1)} = \gamma_j$. In the limit we obtain

$$\Theta_\infty = \{p_{x_n,\infty}(\theta) \propto \pi(\theta) L_n(\theta; x_n) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

□

We note that alternative priors to the maximum-likelihood could be obtained in the limiting system by initialising the construction with a 'bias-adjusted' system of the form

$$\Theta_0 = \{p_{x_n,0}(\theta) = \delta(\theta - \hat{\theta}(x_n) - \frac{b(\hat{\theta}(x_n))}{n}) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}.$$

The construction can be treated as above on replacing $l'(\theta)$ with a term $l'(\theta) + i(\theta)b(\theta)$ when specifying the drift term in diffusions such as (15), leading to a limiting system in which

$$p_{x_n,\infty}(\theta) \propto \pi_b(\theta) L_n(\theta; x_n),$$

where $b(\theta)$ and $\pi(\theta)$ are related by

$$\frac{\partial}{\partial\theta}\log\pi_b = \frac{a(\theta)}{i(\theta)} + i(\theta)b(\theta).$$

5.2 An irregular model

We give an example to show that a Bayesian limiting system does not arise from the construction if the model is not sufficiently regular.

Consider again the uniform distribution of Example 3.5 reparameterised for convenience so that $\pi(y|\theta) = \theta^{-1}, 0 < y < \theta$ and the system of (maximum likelihood) estimators

$$\Theta_{MLE} = \{p_{x_n,i}(\theta) = \delta(\theta - x_{(n)})|\; n \in \mathbb{N}\}.$$

Now Θ_{MLE} is trivially a fixed point of Ψ. Therefore consider a more general system of estimators of the form

$$\Theta_{\mathbf{a}} = \{p_{x_n,i}(\theta) = \delta(\theta - a_n x_{(n)})|\; n \in \mathbb{N}\},$$

with $a_n = \frac{n+1}{n}$ giving, for example, a system of unbiased MLE-based estimators. We now investigate whether a Bayesian limit arises when Ψ is applied to $\Theta_{\mathbf{a}}$ for suitably chosen $\mathbf{a} = (a_1, a_2, ...)$. For simplicity we restrict attention to sequences $\mathbf{a}$ for which $\lim_{n\to\infty} a_n = 1$, so that the system of estimators is consistent. Subject to this assumption we make the following claim:

(i) If, for all $n \geq 1$, $m \log a_{n+m} < 1$ for all but finitely many m then

$$\Theta_\infty = \lim_{k\to\infty} \Theta_{\mathbf{a}}\Psi^k = \Theta_{MLE}.$$

(ii) Otherwise Θ_∞ does not exist.

Proof. We will make use of the following standard result on random walks from Kingman (1962).

Lemma 5.5. *Let $X_i, i = 1, 2, 3...$ denote a sequence of i.i.d. random variables with mean 0 and variance 1, such that X_1 has an exponential moment. Let $S_r^{(a)} = \sum_{i=1}^r X_i - ra$, where $a > 0$, and let $M^{(a)} = \sup_{r\geq 1} S_r^{(a)}$. Then*

$$\lim_{a\to 0} \Pr(aM^{(a)} > z) = e^{-2z},$$

so that $aM^{(a)}$ converges weakly to an Exp(2) distribution as $a \to 0$.

Now consider the construction of $\Theta_{\mathbf{a}}^{(1)} = \Theta_{\mathbf{a}}\Psi$. Fix n, suppose we have data x_n and suppose without loss of generality that $x_{(n)} = \max(y_1, .., y_n) = 1$. Consider the generalised data augmentation chain $\{\theta_m^{(k)} \mid k = 0, 1, 2, ...\}$ when we augment x_n with m additional observations. For this chain, for $k = 1, 2, 3, \ldots$

$$\theta_m^{(k)} = a_{n+m} \sup\{1, \eta_k \theta_m^{(k-1)}\},$$

where $\{\eta_k\}$ are i.i.d. Beta$(m, 1)$, and $\eta_k\theta_m^{(k-1)}$ represent the supremum of the m additional samples imputed during the update process. The corresponding chain for $\lambda = \log\theta$ has update

$$\begin{aligned}\lambda_m^{(k)} &= \log a_{n+m} + \sup\{0, \lambda_m^{(k-1)} - \xi_k\} \\ &= \sup\{\log a_{n+m}, \lambda_m^{(k-1)} - \xi_k + \log a_{n+m}\}\end{aligned}$$

where the $\{\xi_k\}$ are i.i.d. Exp(m) . Set $c_m = m \log a_{n+m}$, $\nu_m^{(k)} = m\lambda_m^{(k)} - c_m$ and write the update as

$$\nu_m^{(k)} = \sup\{0, \nu_m^{(k-1)} + \zeta_m^{(k)}\} \tag{19}$$

where the $\zeta_m^{(k)} = c_m - m\xi_k$, are i.i.d. with mean and variance $c_m - 1$ and 1 respectively.

Now let

$$S_r = \sum_{i=1}^r \zeta_i, \tag{20}$$

It follows from standard results that $\nu_m \sim M^{(m)} = \sup_{r\geq 1} S_r$ where ν_m denotes the stationary distribution of the Markov chain (19).

If $c_m \geq 1$ then the random walk S_r is not positive recurrent and proper stationary distributions do not exist for the Markov chains $\{\nu_m^{(k)}\}$ and $\{\lambda_m^{(k)}\}$. Therefore we must assume that $c_m < 1$ for all but finitely many m for $\Theta_{\mathbf{a}}$ to lie in $\mathscr{C}$. Part (ii) of the claim follows. Assuming this condition we identify distinct cases according to the limiting behaviour of c_m.

1. If $\lim_{m\to\infty} c_m < 1$ then, for sufficiently large m, $\nu_m^{(k)} \to \nu_m$ where ν_m is stochastically dominated by some random variable, τ, independent of m. It is then immediate that $\lambda_m^{(k)} \to \lambda_m$ and the λ_m must tend to 0 in probability as $m \to \infty$.

2. Suppose now $\lim_{m\to\infty} c_m = 1$. By 5.5 it follows that, as $m \to \infty$,

$$(1 - c_m)M^{(m)} \sim (1 - c_m)\nu_m \to M$$

where $M \sim \mathrm{Exp}(2)$.

Now consider the large-m behaviour of

$$\lambda_m \sim \frac{c_m}{m} + \frac{\nu_m}{m} \sim \frac{c_m}{m} + \frac{(1-c_m)\nu_m}{m(1-c_m)},$$

which in turn is determined by that of $m(1 - c_m)$. Writing this as

$$m(1-c_m) = \frac{m}{m+n}\left(n + m(1 - (m+n)\log a_{n+m})\right),$$

we see that $\lim_{m\to\infty} m(1-c_m) = n + \lim_{m\to\infty} m(1 - m\log a_m) \geq n$. There are two cases to consider

(a) If $\lim_{m\to\infty} m(1 - m\log a_m) = \infty$, then $\lambda_m \to 0$ weakly.

(b) If $\lim_{m\to\infty} m(1 - m\log a_m) = \mu$, where $0 \leq \mu < \infty$, then

$$\lambda_m \to \mathrm{Exp}(2(n+\mu)).$$

In either case (1) or (2)(a) our system of inferences (for the log-transformed parameter) is $\Lambda^{(1)} = \{p_{x_n,i}(\lambda) = \delta(\lambda)\}$ with corresponding system for θ, for the case of general $x_{(n)}$, given by

$$\Theta_{\mathbf{a}}^{(1)} = \{p_{x_n,i}(\theta) = \delta(\theta - x_{(n)})|\, n \in \mathbb{N}\} = \Theta_{\mathbf{MLE}}.$$

Thus a single application of Ψ maps $\Theta_{\mathbf{a}}$ to the fixed point $\Theta_{\mathbf{MLE}}$.

In case (2)(b) it can be shown that

$$\Theta_{\mathbf{a}}^{(1)} = \{p_{x_n,1}(\theta) \propto \theta^{-2n}\theta^{-2\mu-1}\}. \tag{21}$$

In particular, if $a_n = \frac{n+1}{n}$, then $\mu = 1/2$ and

$$\Theta_{\mathbf{a}}^{(1)} = \{p_{x_n,1}(\theta) \propto \sigma(\theta)^{-2} L(\theta; x_n)^2\},$$

as was the case for models in the exponential family.

However, any hopes of ultimate convergence to a Bayesian solution are dashed by further applications of Ψ. We show that for case (2)(b) Ψ maps $\Theta_{\mathbf{a}}^{(1)}$ to $\Theta_{\mathbf{MLE}}$. Working in the λ parameterisation we apply Ψ to the system

$$\Lambda_{\mathbf{a}}^{(1)} = \{p_{x_n,1}(\lambda) \propto 2(n+\mu)e^{-2(n+\mu)(\lambda - \log x_{(n)})}, \lambda > x_{(n)}\},$$

Assuming $x_{(n)} = 1$, the level-m data-augmentation chain is defined by

$$\lambda_m^{(k)} = \sup\{\eta_{1,k}, \lambda_m^{(k-1)} - \eta_{2,k} + \eta_{1,k}\},$$

where the $\{\eta_{1,k}\}$ are i.i.d. $\text{Exp}(2(m+n+\mu)$ and the $\{\eta_{2,k}\}$ are i.i.d. $\text{Exp}(m)$. For common initial value $\lambda_m^{(0)}$, this process is stochastically dominated by a process

$$\zeta_m^{(k)} = \sup\{\eta'_{1,k}, \zeta_m^{(k-1)} - \eta_{2,k} + \eta'_{1,k}\},$$

where the $\{\eta'_{1,k}\}$ are i.i.d. $\text{Exp}(2m)$. It is clear that $\{m\zeta_m^{(k)}\}$ has the same proper stationary distribution for all m, so $\zeta_m \to 0$ in distribution as $m \to \infty$ where ζ_m follows the stationary distribution of the unscaled chain $\{\zeta_m^{(k)}\}$. The result is then immediate.

6 Discussion

In this paper we have shown how the generalised data augmentation principles introduced in Gibson et al. (2011) can be applied to elicit connections between classical point estimation and Bayesian (or Bayesian-like) inference where the latter is constructed from the former by repeated application of a refinement operator based on the principles. This contrasts with other approaches to defining connections, for example, by defining point estimators from Bayesian analyses using decision-theoretic ideas. A key notion in our treatment is that of preferability of one system of inferences over another and of fixed points of the refinement operators representing maximally preferable inferences. Our results show that for sufficiently regular models parameterised by their mean, the limiting systems of inferences is Bayesian when the model lies in the exponential family and otherwise takes the form of a pseudo-Bayesian analysis in which the true model likelihood is replaced by one with an exponential-family form. More generally, subsequent investigations suggest that limiting systems derived from initial systems of maximum-likelihood estimators correspond to Bayesian inference using Hartigan's maximum-likelihood prior specification, given sufficiently strong regularity conditions on the model.

The results of the paper do not lead to new statistical methodology but, rather, a fresh perspective on established approaches to inference. It is arguably surprising that, by applying principles that require only that the class of acceptable inferences be closed under a certain data augmentation operation and that it be complete in a natural sense, the Bayesian paradigm can be constructed from a starting point that considers only point estimators. The property that a system of inferences is invariant under Ψ or Φ, though not necessarily by the transition kernels in the finite-m data-augmentation chains involved in the formulation of these operators, may be seen as a weak form of coherence. Our results show that when we attempt to construct weakly coherent systems by seeking fixed points of Φ or Ψ, then these fixed points may nevertheless be strongly coherent Bayesian systems when their basin of attraction contains the system of maximum-likelihood point estimators. There are several natural extensions to the ideas of the paper that would be worthy of investigation, the most obvious of which is to generalise the results to models beyond the 1-dimensional case.

Acknowledgements

The authors are grateful to Seva Shneer, Stan Zachary, Jennie Hansen and Andrew Cairns for useful discussions in the course of this research.

A Proofs of Lemma 4.3 and Theorem 4.2

We first recall Lemma 4.3

Lemma A.1. *(1) There exists a unique solution to the SDE (4), $\{\theta_i(t) : t \geq 0\}$ which can be written in integral form as*

$$\theta_i(T) = \bar{x}_n + (\xi - \bar{x}_n)e^{-nT} + \int_0^T e^{-n(T-t)}\sigma_i(\theta_i(t))dW_t.$$

(2) The moments of $\theta_i(t)$ are bounded up to a level depending on n in that there exist constants C_k such that $E|\theta_i(t)|^k \leq C_k(1 \vee |\xi|^k \vee |\bar{x}_n|^k$ for all $t \geq 0$ and $1 \leq k \leq (2n + C_l)/C_l$.

(3) If $\sqrt{2}n > C_l$, the stationary distribution of (4) is given by

$$p_i(\theta) \propto \frac{1}{\sigma_i(\theta)^2}\exp(2n(f_i(\theta)\bar{x}_n - g_i(\theta))),$$

where

$$f_i(\theta) = \int \sigma_i^{-2}(\theta)d\theta, \;\; g_i(\theta) = \int \theta\sigma_i^{-2}(\theta)d\theta, \;\; \theta \in (l, r).$$

Proof. (1) By Engelbert and Schmidt (1991) Theorem (4.53) the existence of a unique weak solution follows in our setting if

$$\mathcal{N} := \{x \in \mathbb{R} : \sigma(x) = 0\} = \mathcal{S} := \{x \in \mathbb{R} : \int_{x-}^{x+} \sigma^{-2}(y)dy = \infty\}.$$

As σ^2 is (Lipschitz) continuous we see that for any point x such that $\sigma^2(x) > 0$ we have that σ^{-2} is locally integrable at x and hence $\mathcal{S} \subset \mathcal{N}$ giving the existence of a weak solution.

We also see that if $x \in \mathcal{N}$, then by the Lipschitz condition, for any open set G containing x there is a K_G such that

$$\int_G \sigma^{-2}(y)dy = \int_G \frac{1}{|\sigma^2(y) - \sigma^2(x)|} \geq \int_G \frac{1}{K_G}|y - x|^{-1}dy = \infty.$$

Thus $\mathcal{N} \subset \mathcal{S}$ and we have the uniqueness in law of the solution.

It is a simple exercise to establish the integral form.

(2) In order to show this we first need to establish some crude moment bounds to ensure that the stochastic integral in the integral representation for θ is a martingale. The stochastic integral is a local martingale and thus if we define the stopping times $T_m := \inf\{t : |\theta(t) - \bar{x}_n| > m\}$ we have for $k \geq 2$ using (5),

$$\phi_t^{m,k} := E|e^{n(t\wedge T_m)}\left(\theta(t \wedge T_m) - \bar{x}_n\right)|^k = E|(\theta_i(0) - \bar{x}_n) + \int_0^t e^{ns}\sigma_i(\theta_i(s))dW_s|^k.$$

Applying the Burkholder-Davis-Gundy inequality, Hölder's inequality and the linear growth condition on σ_i we see that

$$\begin{aligned}
\phi_t^{m,k} &\leq 2^{k-1}|\theta_i(0) - \bar{x}_n|^k + 2^{k-1}E|\int_0^{t\wedge T_m} e^{2ns}\sigma_i(\theta_i(s))^2 ds|^k \\
&\leq c_k + 2^{k-1}c'_k E|\int_0^{t\wedge T_m} e^{2ns}\sigma_i(\theta_i(s))^2 ds|^{k/2} \\
&\leq c_k + C'_k E T^{k/2-1}\int_0^{t\wedge T_m} e^{-kns}\sigma^k(\theta(s))ds \\
&\leq c'_k + c_k T^{k/2-1}E\int_0^t e^{kn(s\wedge T_m)}(C'_k + C_k(\theta(s\wedge T_m) - \bar{x}_n)^k)ds \\
&\leq c''_k T^{k/2-1}e^{nkt} + c_k T^{k/2-1}\int_0^t \phi_s^{m,k}ds.
\end{aligned}$$

A simple application of Gronwall's inequality gives that for $0 \leq t \leq T$

$$\phi_t^{m,k} \leq c_k'' T^{k/2-1} \exp(c_k T^{k/2-1} t) \leq c_k'' T^{k/2-1} \exp(c_k T^{k/2}).$$

As this bound is independent of m we can apply the dominated convergence theorem and let $m \to \infty$ to see that for $k \geq 2$

$$E|\theta(t)|^k \leq c_k' T^{k/2-1} \exp(C_k T^{k/2}), \;\; 0 \leq t \leq T.$$

Equipped with this we can improve the estimates on the moments. Using Ito's formula we have

$$d\theta^k = (kn(\bar{x}_n - \theta)\theta^{k-1} + \frac{1}{2}k(k-1)\theta^{k-2}\sigma^2(\theta))dt + k\theta^{k-1}\sigma(\theta)dW.$$

Applying the moment estimates above we can see that for $0 \leq t \leq T$

$$\begin{aligned} E(\int_0^{t\wedge T_m} \theta^{k-1}\sigma(\theta)dW)^2 &\leq E\int_0^{t\wedge T_m} \theta^{2k-2}\sigma^2(\theta)ds \\ &\leq E\int_0^{t\wedge T_m} C_l\theta^{2k-2} + C_l\theta^{2k}ds \\ &\leq \int_0^t C_l E\theta^{2k-2} + C_l E\theta^{2k}ds \\ &\leq c_k T^k e^{cT^k} \end{aligned}$$

independent of m. Again letting $m \to \infty$ we see that the stochastic integral term is a true martingale and hence we have the following expression for the moments $\phi_t^k = E|\theta(t)|^k$,

$$\phi_t^k = \phi_0^k + E\int_0^t (kn(\bar{x}_n - \theta(s))\theta(s)^{k-1} + \frac{1}{2}k(k-1)\theta(s)^{k-2}\sigma^2(\theta(s)))ds. \qquad (22)$$

We proceed by induction noting that $\phi_t^0 = 1$ and $E\theta_t = \bar{x}_n$ so that using $k = 2$ and the linear growth bound we have

$$\phi_t^2 \leq (C_l - 2n)\int_0^t \phi_s^2 ds + (2n\bar{x}_n^2 + C_l)t.$$

Assume that n is large enough so that $2n > C_l$, then in differential form we have

$$\frac{d\phi^2}{dt} \leq (C_l - 2n)\phi^2 + 2n\bar{x}_n^2 + C_l,$$

this can be solved to get

$$\phi_T^2 \leq \frac{2n\bar{x}_n^2 + C_l}{2n - C_l} + \left(\xi^2 - \frac{2n\bar{x}_n^2 + C_l}{2n - C_l}\right) e^{-(2n-C_l)T}.$$

Thus we have the uniform bound for all $T > 0$,

$$\phi^2 \leq C_2(1 \vee \xi^2 \vee \bar{x}_n^2).$$

For the general case, using the linear growth of σ, we have

$$\phi_t^k \leq |\xi|^k + \int_0^t (kn\bar{x}_n\phi_s^{k-1} + (\frac{1}{2}k(k-1)C_l - kn)\phi_s^k + \frac{1}{2}k(k-1)C_l\phi_s^{k-2}ds. \qquad (23)$$

Now assume that $\phi_t^p \leq C_p(1 \vee |\xi|^p \vee |\bar{x}_n|^p)$ for all $p \leq k-1$ and $t > 0$. Using this in (23) we get

$$\phi_t^k \leq |\xi|^k + kC_{k-1}n\bar{x}_n(1 \vee |\xi|^{k-1} \vee |\bar{x}_n|^{k-1})t + \int_0^t k(\frac{1}{2}(k-1)C_l - n)\phi_s^k ds + C_lC_{k-2}(1 \vee |\xi|^{k-2} \vee |\bar{x}_n|^{k-2})t.$$

For $n > \frac{1}{2}(k-1)C_l$ we have, by solving the associated differential inequality, that there is a C_k such that

$$\phi_t^k \leq C_k(1 \vee |\xi|^k \vee |\bar{x}_n|^k).$$

Thus we have the general case provided that $(k-1)C_l < 2n$ as required.

(3) The generator of the diffusion is

$$\mathcal{A}f = n(\bar{x}_n - \theta)\frac{\partial f}{\partial \theta} + \frac{1}{2}\sigma_i^2(\theta)\frac{\partial^2 f}{\partial \theta^2}.$$

Under the condition that $\sqrt{2}C_l < 2n$ we can use the Lyapunov function technique to show that we have convergence to a stationary distribution at an exponential rate. Let $V(x) = 1 + x^2$, then

$$\mathcal{A}V(x) = 2n(\bar{x}_n - x)x + \sigma_i^2(x).$$

Using the linear growth bound $\sigma_i^2(x) \leq \sqrt{2}C_lV(x)$ for all i, we have

$$\mathcal{A}V(x) = -(2n - \sqrt{2}C_l)V(x) + 2n(\bar{x}_nx + 1).$$

Hence, if $\beta = (2n - \sqrt{2}C_l)/2$, $b = (n^2(1+\bar{x}_n^2) - C_l^2/2)/\beta$, we have

$$\mathcal{A}V(x) = -\beta V(x) + bI_{\mathcal{C}},$$

with $\mathcal{C} = \{x : |x - \frac{n\bar{x}_n}{\beta}| \leq \sqrt{\frac{b}{\beta}}\}$. By Meyn and Tweedie (2009) 20.3.2, this condition gives the existence of, and exponential convergence to, the stationary distribution p_i.

To find the stationary distribution we just need to solve

$$\mathcal{A}^* p_i = -\frac{\partial}{\partial\theta}(n(\bar{x}_n - \theta)p_i) + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}(\sigma_i^2(\theta)p_i) = 0. \tag{24}$$

We can check that the solution as given satisfies equation (24). □

As a consequence of the moment estimates for the diffusion and the fact that it will converge to a stationary distribution, we have immediately that

Corollary A.2. *The stationary distribution of the SDE has moments of order up to* $\frac{2n+C_l}{C_l}$.

We now give the estimates needed to establish that each Markov chain in the sequence has a stationary distribution and that these converge to the stationary distribution for the diffusion. We assume the generalized data augmentation chain is suitably irreducible.

Theorem A.3. *For each $i, m \in \mathbb{N}$, the Markov chain $\{\theta_{i,m}^{(k)}, k = 0, 1, 2, \dots\}$ is ergodic with a unique stationary distribution $\pi_{i,m}$. There exists $r_i < 1$ and $R_i < \infty$ independent of m such that for any Borel set A and all $t > 0$*

$$\sup_x \frac{|\mathbb{P}^x(\theta_{i,m}^{(\lfloor mt \rfloor)} \in A) - \pi_{i,m}(A)|}{1 + x^2} \leq R_i r_i^t.$$

Proof. We give the proof for the $i = 1$ case and then discuss the extensions required for the general case. In order to establish the positive recurrence we use the Lyapunov function technique. Let $V(x) = 1 + x^2$. From Meyn and Tweedie (2009), Chapter 15, we have convergence to stationarity if there exists a petite set $\mathcal{C}$ as well as a $\beta, b > 0$ such that

$$\Delta V(x) := \mathbb{E}^x V(\theta_m^{(1)}) - V(x) \leq -\beta V(x) + bI_{\mathcal{C}}.$$

From our above estimates and linear growth assumption we have

$$\begin{aligned}
\mathbb{E}^x V(\theta^{(1)}) &= 1 + \mathbb{E}\left(x + \frac{n(\bar{x}_n - x)}{n+m} + \frac{R_m(x)}{n+m}\right)^2 \\
&= V(x) + 2x(\frac{n(\bar{x}_n - x)}{n+m}) + \mathbb{E}\left(\frac{n(\bar{x}_n - x)}{n+m} + \frac{R_m(x)}{n+m}\right)^2 \\
&= V(x) + \frac{2nx(\bar{x}_n - x)(n+m) + n^2(\bar{x}_n - x)^2 + m\sigma^2(x)}{(n+m)^2} \\
&\leq V(x) - \frac{n^2 + (2n - C)m}{(n+m)^2}x^2 + \frac{2nm\bar{x}_n}{(n+m)^2}x + \frac{n^2\bar{x}_n^2 + mC}{(n+m)^2} \\
\Delta V(x) &\leq -\alpha x^2 + \beta x + \gamma
\end{aligned}$$

with

$$\alpha = \frac{n^2 + (2n - C)m}{(n+m)^2}, \quad \beta = \frac{2nm|\bar{x}_n|}{(n+m)^2}, \quad \gamma = \frac{n^2\bar{x}_n^2 + mC}{(n+m)^2}.$$

Thus, provided $2n > C$, we have that $\alpha > 0$ and

$$\Delta V(x) \leq -\frac{1}{2}\alpha V(x) + \gamma + \beta x - \frac{1}{2}\alpha x^2 + \frac{1}{2}\alpha.$$

A simple calculation gives

$$\Delta V(x) \leq -\frac{1}{2}\alpha V(x) + (\gamma + \frac{1}{2}\alpha + \frac{\beta^2}{2\alpha})I_{\{|x - \frac{\beta}{\alpha}| < \sqrt{\frac{2\gamma}{\alpha} + 1 + \frac{\beta^2}{\alpha^2}}\}}.$$

It is easy to see that if the chain has a transition kernel with full support the set

$$\mathcal{C} = \{x : |x - \frac{\beta}{\alpha}| < \sqrt{\frac{2\gamma}{\alpha} + 1 + \frac{\beta^2}{\alpha^2}}\}$$

is petite for a suitable multiple of Lebesgue measure on a subset of $\mathcal{C}$. If the chain has support on a discrete subset $\mathcal{D}_m$ of $\mathbb{R}$, then the chain will be petite for a suitable multiple of the discrete uniform measure on $\mathcal{D}_m \cap \mathcal{C}$.

We now note that the Markov process $\theta_m(t)$ is the original chain sped up by a factor m. Thus, its generator $\Delta_m = m\Delta$ and as $\mathcal{C}$ is invariant under the time change, we have

$$\Delta_m V(x) \leq -\frac{1}{2}m\alpha V(x) + m(\gamma + \frac{1}{2}\alpha + \frac{\beta^2}{2\alpha})I_{\mathcal{C}}.$$

By the definition of α, β, γ we see that, for our sped up process, we have the existence of constants $\alpha_0, \beta_0, \gamma_0 > 0$, independent of m, such that

$$\Delta_m V(x) \leq -\frac{1}{2}\alpha_0 V(x) + (\gamma_0 + \frac{1}{2}\alpha_0 + \frac{\beta_0^2}{2\alpha_0})I_{\mathcal{C}}.$$

We can now apply the result on V-uniform ergodicity in Meyn and Tweedie (2009) Theorem 16.0.1, to deduce the estimate, with coefficients independent of m.

The general case where $i > 1$ is a simple extension of the $i = 1$ case. We first observe that, by the linear growth condition and the bounds on $\mathbb{E}\theta^2$ from Lemma 4.3, we have

$$\begin{aligned}
\mathbb{E}N_m(\tau_m)^2 &\leq \mathbb{E}\int_0^{\tau_m} e^{-2(n+m)(\tau_m - s)}\sigma_i^2(\theta_i^{n,m}(s))ds \\
&\leq \mathbb{E}\int_0^{\tau_m} e^{-2(n+m)(\tau_m - s)}C_l(1 + \mathbb{E}(\theta_i^{n,m}(s))^2)ds \\
&\leq \mathbb{E}\frac{C_l(1 + c_2\bar{x}_{n+m}^2)}{2(n+m)} \\
&\leq \frac{C_l(1 + c_2\mathbb{E}(x + \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{n+m}R_m(x))^2)}{2(n+m)}
\end{aligned}$$

Thus we keep the same Lyapunov function and from our above estimates and linear growth assumption, we obtain the following uniform control

$$\begin{aligned}
\mathbb{E}^x V(\theta_i^{(1)}) &= 1 + \mathbb{E}\left(x + \frac{n(\bar{x}_n - x)}{n+m} + \frac{R_m(x)}{n+m} + N_m(\tau_m)\right)^2 \\
&= 1 + \mathbb{E}\left(x + \frac{n(\bar{x}_n - x)}{n+m} + \frac{R_m(x)}{n+m}\right)^2 + \mathbb{E}\left(N_m(\tau_m)\right)^2 \\
&\leq 1 + \frac{C_l}{2(m+n)} + (1 + \frac{C_l c_2}{2(m+n)})\mathbb{E}\left(x + \frac{n(\bar{x}_n - x)}{n+m} + \frac{R_m(x)}{n+m}\right)^2 \\
&= V(x) + \frac{C_l}{2(m+n)} + \left(1 + \frac{C_l c_2}{2(m+n)}\right)\frac{2nx(\bar{x}_n - x)(n+m) + n^2(\bar{x}_n - x)^2 + m\sigma^2(x)}{(n+m)^2} \\
&\quad + \frac{C_l c_2 x^2}{2(m+n)} \\
\Delta^{(i)}V(x) &\leq -\alpha_i x^2 + \beta_i x + \gamma_i
\end{aligned}$$

with

$$\alpha_i = \frac{2n - C_l(1 + \frac{1}{2}c_2)}{m} + O(\frac{1}{m^2}), \quad \beta_i = (1 + \frac{C_l c_2}{2(m+n)})\frac{2nm\bar{x}_n}{(n+m)^2},$$

$$\gamma_i = \frac{C_l}{2(m+n)} + \left(1 + \frac{C_l c_2}{2(m+n)}\right) \frac{n^2 \bar{x}_n^2 + mC}{(n+m)^2}.$$

Thus, incorporating the time change, and provided $2n > C_* = C_l(1 + \frac{1}{2}c_2)$, we have that $\alpha_i, \beta_i, \gamma_i > 0$, independent of m, and

$$\Delta_m^{(i)} V(x) \leq -\frac{1}{2}\alpha_i V(x) + \gamma_i + \beta_i x - \frac{1}{2}\alpha_i x^2 + \frac{1}{2}\alpha_i.$$

The same calculations as before give

$$\Delta_m^{(i)} V(x) \leq -\frac{1}{2}\alpha_i V(x) + (\gamma_i + \frac{1}{2}\alpha_i + \frac{\beta_i^2}{2\alpha_i}) I_{\mathcal{C}_i},$$

and we can proceed along exactly the same lines to deduce the result as there is no dependence on m. □

Corollary A.4. *The sequence of stationary distributions $\pi_{i,m}$ for the sped up Markov chains θ_m converges to p_i, the stationary distribution for the solution θ_i to the SDE.*

Proof. All we need to show now is that the sequence π_m converges weakly to π, the stationary distribution for our limit stochastic differential equation. In order to show this consider A, a Borel set of $\mathbb{R}$. By stationarity, for any time t,

$$\pi_i(A) = \int \pi_i(dx) \mathbb{P}^x(\theta_i(t) \in A).$$

By the weak convergence of the processes for any t we have for $\epsilon/2$ that there exists an m_1 such that for $m > m_1$,

$$|\mathbb{P}^{\pi_i}(\theta_{i,m}(t) \in A) - \mathbb{P}^{\pi_i}(\theta_i(t) \in A)| < \epsilon/2.$$

By the geometric ergodicity in Theorem A.3 there exists $1 < r_i$ and $R_i < \infty$ independent of m such that for all t

$$\sup_x \frac{|\mathbb{P}^x(\theta_i^{(\lfloor mt \rfloor)} \in A) - \pi_{i,m}(A)|}{1 + x^2} \leq R_i r_i^{-t}.$$

We can now put these pieces together to prove our result. Let $\mathcal{K}_\epsilon = \{x : |x| < K\}$ where K is chosen such that $\pi(\mathcal{K}_\epsilon) = 1 - \epsilon/4$. Then

$$\begin{aligned}
|\pi_i(A) - \pi_{i,m}(A)| &\leq \left| \int \pi_i(dx)\mathbb{P}^x(\theta_i(t) \in A) - \int \pi_i(dx)\mathbb{P}^x(\theta_{i,m}(t) \in A) \right| \\
&\qquad + \left| \int \pi_i(dx)\mathbb{P}^x(\theta_{i,m}(t) \in A) - \pi_{i,m}(A) \right| \\
&= \epsilon/2 + 1 - \pi(\mathcal{K}_\epsilon) + \sup_{x \in \mathcal{K}_\epsilon} |\mathbb{P}^x(\theta_{i,m}(t) \in A) - \pi_{i,m}(A)| \\
&= \epsilon/2 + \epsilon/4 + \sup_{x \in \mathcal{K}_\epsilon} |\mathbb{P}^x(\theta_i^{\lfloor mt \rfloor} \in A) - \pi_{i,m}(A)| \\
&= 3\epsilon/4 + (1 + K^2)R_i r_i^{-t}.
\end{aligned}$$

We now take t large enough to ensure that have that for all $m > m_1$

$$|\pi_i(A) - \pi_{i,m}(A)| < \epsilon.$$

As this holds for each A we have weak convergence of the measures. □

References

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. *Ann. Statist. 37*, 905–938.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *J. Roy. Statist. Soc. B 41*, 113–147.

Cox, D. R. and E. Snell (1968). A general definition of residuals. *J. Roy. Statist. Soc., Series B 41*, 248–275.

Engelbert, H. J. and W. Schmidt (1991). Strong Markov continuous local martingales and solutions of one-dimensional stochastic differential equations. III. *Math. Nachr. 151*, 149–197.

Ethier, S. and T. Kurtz (1986). *Markov Processes: Characterization and Convergence.* John Wiley and Sons.

Ghosh, J. K. (1994). Higher order asymptomatics. Volume 4 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* IMS.

Gibson, G. J., G. Streftaris, and S. Zachary (2011). Generalised data augmentation and posterior inferences. *J. Statist. Plann. and Inference 141*, 156–171.

Hartigan, J. (1998). The maximum likelihood prior. *Ann. Statist. 26*, 2083–2103.

Hartigan, J. (2012). Asymptotic admissability of priors and elliptic differential equations. *IMS Collections 8*, 117–130.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist. 35*, 836–845.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A 186*, 453–461.

Kingman, J. (1962). On queues in heavy traffic. *J. Roy. Statist. Soc., Series B 24*, 383–392.

Meyn, S. and R. Tweedie (2009). *Markov chains and stochastic stability.* Cambridge University Press.