From point estimation to Bayesian inference via dynamical systems

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

January 9, 2018

[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*

[*] *Corresponding author*

Running title: Dynamical Systems and Inference

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 one-dimensional setting, we characterise fixed points obtained from systems of moment estimators. These limiting inferences can be considered to be pseudo-Bayesian in that parameter densities combine a prior with a data-dependent pseudo-likelihood. They are precisely Bayesian when the model lies in the exponential family, with the usual conjugate prior arising as a by-product of the construction.

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 classical point estimation and Bayesian inference, whereby the former is generated from the latter without direct reference to the Bayesian paradigm, and motivate strategies for approximating Bayesian posteriors or constructing contrast-based pseudo-likelihoods. Approaches to generalising the results to higher-dimensional settings are discussed.

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

1 Introduction

Let $Y_1, Y_2, Y_3, \cdots \in \mathscr{Y} \subset \mathbb{R}$ denote a sequence of independently, identically distributed (i.i.d.) random variables drawn from a measure ν_θ with parameter $\theta \in \mathcal{K}$ where $\mathcal{K} \subseteq \mathbb{R}$. 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

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

where $\mathcal{P}(\mathcal{K})$ denotes the set of probability measures on $\mathcal{K}$ and p_{x_n} is a measure representing the belief about parameter θ given the outcome x_n. Systems of point estimators are obtained on setting $p_{x_n} = \delta_{\hat{\theta}(x_n)}$, where $\hat{\theta}(x_n)$ denotes the point estimate of θ calculated from data x_n and δ_x is the Dirac measure at x. Bayesian systems have the property that the measure $p_{x_n}(d\theta) \propto \pi(d\theta) \prod_{i=1}^{n} \nu_\theta(dy_i)$ for some prior measure $\pi(d\theta)$. 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 [12].

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 [9], in order to define mappings, Ψ and Φ, from $\mathscr{X}$ to itself, called *refinement operators* in Section 2, which map a given system of inferences to one which is preferable in a sense made 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 a prior distribution for θ such that the resulting posterior density satisfies certain classical criteria, at least asymptotically. Examples include reference priors (see [2, 3]), 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 [14] which attempts to assign equal prior probability to intervals of a given level of confidence. A decision-theoretic approach is taken by Hartigan [11] 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* [12, 10], which is the risk-matching prior corresponding to maximum-likelihood 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 one-dimensional models considered in this paper, the maximum-likelihood prior $\pi(\theta)$, for a model where the measure ν_θ has a density v_θ which is a C^2 function of θ, satisfies

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

where

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

and

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

These correspondences are constructed from a Bayesian starting point. 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, 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 arise from the construction.

In our main result, Theorem 3.4, we characterise, for a broad class of one-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 3.7 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 was 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. The proof of Theorem 3.4 will be given in Section 4 with some of the technical details postponed to the appendix.

In Section 5 we explore the generalisations of the main theorem to fixed points of Ψ 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, at least when observations are augmented with i.i.d. samples from ν_θ in the construction. Potential generalisations of the results to higher-dimensional settings are discussed in Section 6.

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. We will say that a system of inferences

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

is valid in the opinion of a given observer if they consider it appropriate, having observed $x_n = (y_1, ..., y_n)$, to represent their belief regarding θ via the measure p_{x_n} and regarding any future observation Y_{n+r}, independent of x_n given θ, as arising from the measure

$$q_{x_n}(dy_{n+r}) = \int \nu_\theta(dy_{n+r})p_{x_n}(d\theta).$$

In short, the system is valid if it can be used to form posterior-like distributions for parameters or predictive distributions for future observations analogous to Bayesian posteriors and predictive distributions.

If a system is valid for an observer then, informally, on observing x_n, they may be justified in using p_{x_n} to predict their inference on θ given further observations. The generalised data augmentation principle proposed in [9] states that such predictions themselves represent valid inferences. Formally, the generalised data augmentation principle [9] asserts that the set of all valid systems of inference for θ should be closed under a data-augmentation operation as described by Principle 2.1.

Augmentation Principle 2.1. *Let*

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

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

$$p^{(m)}_{x_n} = \int_{\mathcal{K}} \int_{\mathcal{Y}^m} p_{x_{n+m}} \prod_{i=1}^{m} \nu_{\theta'}(dy_{n+i})p_{x_n}(d\theta'). \tag{1}$$

For an observer who accepts Principle 2.1, the principle implies that, if Θ is valid, then so is Θ^m. However the converse does not hold; consequently the latter system may be considered more assuredly valid than the former. In this sense Θ^m is *preferable* to Θ.

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 further independent samples $y_{n+1}, ..., y_{n+m}$. Of course, when Θ is a Bayesian system of inferences, then Θ and Θ^m coincide. Our main interest will be in applying Principle 2.1 in other settings, to transform (or refine) a valid system of inferences into a preferable one in a systematic manner.

First note that $p_{x_n}^{(m)} = p_{x_n} P_{x_n,m,\Theta}$, where $P_{x_n,m,\Theta} : \mathcal{P}(\mathcal{K}) \to \mathcal{P}(\mathcal{K})$ is the transition kernel of $\{\theta_m(k), k \geq 0\}$, a Markov chain on $\mathcal{K}$ called the *generalised data-augmentation chain.* Updates to the current state $\theta_m(k)$ are generated by drawing $Y_{n+1}, ..., Y_{n+m}$ as i.i.d. samples from the measure $\nu_{\theta_m(k)}$, appending these to the observed x_n to form x_{n+m}, and then drawing $\theta_m(k+1)$ from the measure $p_{x_{n+m}}$. Applying Principle 2.1 sequentially, it follows that a valid system is obtained by replacing p_{x_n} with $p_{x_n} P^k_{x_n,m,\Theta}$ for any $k \in \mathbb{N}$. Moreover, if the generalised data-augmentation chain defined by $P_{x_n,m,\Theta}$ is ergodic with stationary measure, $\psi_{x_n}^{(m)}$, then replacing p_{x_n} with $\psi_{x_n}^{(m)}$ 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 [9].

Completeness Principle 2.2. *Suppose that* $\{\Theta_i, i = 1, 2 \ldots\}$, *where* $\Theta_i = \{p_{x_n,i} \in \mathcal{P}(\mathcal{K}) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$, *denotes a sequence of valid systems for which*

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

where convergence is in the sense of weak convergence of measures. Then

$$\Theta_\infty = \{\psi_{x_n} \in \mathcal{P}(\mathcal{K}) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$$

is also a valid system. If, for some system of inferences Θ *and every* i, Θ_i *is preferable to* Θ, *then* Θ_∞ *is preferable to* Θ.

We motivate Principles 2.1 and 2.2 from the perspective of coherence and the avoidance of a 'Dutch book' (a combination of bets leading to a guaranteed loss [13]). Suppose that two i.i.d observations Y_1, Y_2 from ν_θ are to be observed sequentially, after which the (currently unknown) value of θ will be revealed. A bookmaker (A) considers the system of inferences $\{p_{x_1}, p_{x_2}\}$ to be valid, where $x_1 = (y_1)$ and $x_2 = (y_1, y_2)$ but does not accept Principles 2.1 and 2.2. Suppose that A is obliged to accept any proposal from investor B for which A's expected loss is zero. Bets can be placed immediately after observing y_1 and again after observing y_2, prior to θ being revealed.

First y_1 is observed, so that A's belief regarding θ is now represented by $p_{x_1}(\theta)$. Suppose further that

$$p^{(1)}_{x_1}(S) = \int_{\mathcal{K}} \int_{\mathscr{Y}} p_{x_2}(S) \nu_{\theta'}(dy_2) p_{x_1}(d\theta')$$

differs from $p_{x_1}(S)$ for some subset $S \subset \mathcal{K}$, so that for some subset $\omega \subset S \subset \mathcal{K}$,

$$p_{x_1}(\omega) < p^{(1)}_{x_1}(\omega).$$

B proposes the following wager to be settled when θ is revealed.

Wager 1: B pays A £$p_{x_1}(\omega)$ in return for a pay-off of £1 if $\theta \in \omega$.

A accepts Wager 1 since, under $p_{x_1}(\theta)$, their expected loss is zero. B then proposes a further wager to A.

Wager 2: B pays A £$(1 - p^{(1)}_{x_1}(\omega))$ in return for a pay-off of £1 if $\theta \notin \omega$.

As A's expected loss for Wager 2 is positive under p_{x_1} they do not accept the bet as B is not offering a fair price. Note that Wager 2 and Wager 1 together constitute a 'Dutch Book', with A guaranteed to lose £$(p^{(1)}_{x_1}(\omega) - p_{x_1}(\omega))$. B now proposes the following wager.

Wager 3: B pays A £$(1 - p^{(1)}_{x_1}(\omega))$. On observing y_2, A pays B $q(y_2)$ where $q(y_2)$ is *A's* fair price for Wager 2 on observing y_2.

Since A's expectation of $q(Y_2)$ is £$(1 - p^{(1)}_{x_1}(\omega))$ they accept Wager 3. Now, B effectively holds an option to place Wager 2, once $Y_2 = y_2$ is observed, for their originally proposed price of £$(1 - p^{(1)}_{x_1}(\omega))$. A has accepted a Dutch book of bets with guaranteed loss of £$(p^{(1)}_{x_1}(\omega) - p_{x_1}(\omega))$. Had A utilised Principles 2.1 and 2.2 to replace $\{p_{x_1}(\theta), p_{x_2}(\theta)\}$ with the preferable system $\{\psi^1_{x_1}(\theta), p_{x_2}(\theta)\}$ (discussed before Principle 2.2) prior to negotiation with B then A would have been immune to this particular Dutch book. Therefore using Principles 2.1 and 2.2 to refine the initial system of inferences would have improved the coherence of A's system.

We consider how the above arguments might be applied systematically in order to improve the coherence of a system of inferences more generally. We apply Principles 2.1 and 2.2 to formulate a *refinement operator*, Ψ, that can be applied to an initial system of inferences

$$\Theta_0 = \{p_{x_n,0} \in \mathcal{P}(\mathcal{K}) \mid n \in \mathbb{N}, x_n \in \mathscr{Y}^n\},$$

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

The generalised data-augmentation chain, $\{\theta_{0,m}(k), k \geq 0\}$ has transition kernel P_{x_n,m,Θ_0}, when the observation x_n is augmented by the next m samples. As above, the state $\theta_{0,m}(k)$ is updated by drawing $\theta_{0,m}(k+1)$ from the measure $p_{x_{n+m},0}$ where x_{n+m} is formed by augmenting the observed x_n by i.i.d. draws $Y_{n+1}, ..., Y_{n+m}$ from $\nu_{\theta_{0,m}(k)}$. We construct a preferable inference for x_n by taking the stationary distribution of the chain for each m, appealing to Principle 2.2, and then taking the limit of these stationary distributions as $m \to \infty$, again by Principle 2.2, to remove dependence on m. This is carried out for each n in ascending order to generate the new system $\Theta_1 = \Theta_0 \Psi$.

Generally, we construct

$$\Theta_{i+1} = \{p_{x_n,i+1} \in \mathcal{P}(\mathcal{K}) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\} = \Theta_i \Psi$$

from $\Theta_i = \{p_{x_n,i} \in \mathcal{P}(\mathcal{K}) | n \in \mathbb{N}, x_n \in \mathscr{Y}^n\}$ recursively by setting

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

where the limits are taken in the sense of weak convergence of measures. We denote by $\{\theta_{i,m}(k), k \geq 0\}$ the generalised data-augmentation chains that arise in the construction of Θ_{i+1} from Θ_i. Suppose now that $\lim_{i\to\infty} \Theta_i = \Theta_\infty$ exists. Then Θ_∞ is preferable to Θ_i, for all i. Moreover, in the situations that we consider here, Θ_∞ 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

$$p_{x_n}(d\theta) \propto \pi(d\theta) \prod_{i=1}^{n} \nu_\theta(y_i)$$

for some prior measure π, lies in $\mathscr{C}_F$. Here the generalised data-augmentation chain with transition kernel $P_{x_n,m,\Theta}$ is a Gibbs sampler and the measure p_{x_n} is fixed by this kernel for any $m > 0$ and, hence, by Ψ. As discussed later, $\mathscr{C}_F$ contains non-Bayesian systems. Therefore the property of invariance under Ψ may be seen as a weak form of coherence.

In the rest of the paper we will be particularly interested in systems $\Theta_\infty \in \mathscr{C}_F$ that arise as fixed points when Ψ acts on an initial system of point estimators Θ_0. When Θ_∞ is Bayesian, then a link is made between classical point estimation and Bayesian inference and we may consider how such linkages relate to connections established using decision theory and the construction of Bayes estimators. When Θ_∞ is non-Bayesian we can consider how it may relate to other non-Bayesian approaches to forming posterior-like distributions, or may approximate a Bayesian inference.

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. As a corollary, we show that for the case of the one-dimensional exponential family with the mean-value parametrisation, Bayesian analysis with the maximum-likelihood prior of [10] is obtained as the limiting system of inferences.

We define a second refinement operator, and corresponding constructions, by taking limits with respect to m and k in a different manner. If we first take the limit with respect to m and speed up the Markov chain, we may have weak convergence of the chain to a limiting stochastic process. If we then let time for the limiting process go to infinity we may see a stationary distribution. Thus, in a formal sense we can define the new operator Φ for which the measures comprising $\Theta_{i+1} = \Theta_i \Phi$ are given by

$$p_{x_n,i+1} = \lim_{t\to\infty} \lim_{m\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^{mt}. \tag{3}$$

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 3.4. At some points in the paper, it will be convenient to work with the operator Φ defined by (3). Figure 1 gives a schematic depiction of the operators Φ and Ψ and the way in which limits are taken in the respective cases.

3 Moment-based estimators and fixed points

We retain the notation of the previous section and let $\{Y_i : i \geq 1\}$, where $Y_i \in \mathcal{Y} \subset \mathbb{R}$, denote a sequence of i.i.d random variables drawn from a measure ν_θ with a one-dimensional parameter $\theta \in \mathcal{K} = (l, r)$ (where $l, r \in [-\infty, \infty]$) and we write $x_n = (y_1, ..., y_n)$ for the first n observations. As our parameter space is an interval, there is an increasing family $\{\mathcal{K}_u : u > 0\}$ of compact subsets of $\mathcal{K}$ with $\cup_u \mathcal{K}_u = \mathcal{K}$. For instance, if the boundaries are finite, we can set $\mathcal{K}_u = [l + 1/u, r - 1/u]$, when $u \geq 2/(r - l)$, and $\mathcal{K}_u = \emptyset$ otherwise. We will always assume that $\bar{x}_n \in \mathcal{K}$ to rule out possible degeneracies if our data lies on the boundary of the parameter space. We suppose that ν_θ has mean θ, is continuous in θ (in the sense that $\nu_\theta(A)$ is continuous in θ for each Borel set $A \subset \mathcal{K}$), and has variance

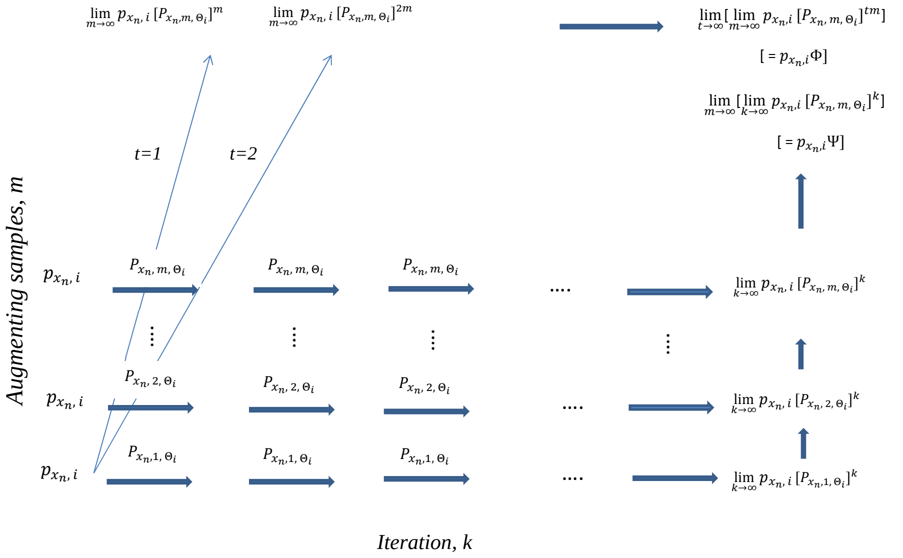
$\lim_{m\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^m$
$\lim_{m\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^{2m}$
$\lim_{t\to\infty}[\lim_{m\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^{tm}]$
$[= p_{x_n,i}\Phi]$
$\lim_{m\to\infty}[\lim_{k\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^k]$
$[= p_{x_n,i}\Psi]$
t=1
t=2
Augmenting samples, m
$p_{x_n,i}$
P_{x_n,m,Θ_i}
P_{x_n,m,Θ_i}
P_{x_n,m,Θ_i}
....
$\lim_{k\to\infty} p_{x_n,i}[P_{x_n,m,\Theta_i}]^k$
$P_{x_n,2,\Theta_i}$
$P_{x_n,2,\Theta_i}$
$P_{x_n,2,\Theta_i}$
$p_{x_n,i}$
....
$\lim_{k\to\infty} p_{x_n,i}[P_{x_n,2,\Theta_i}]^k$
$P_{x_n,1,\Theta_i}$
$P_{x_n,1,\Theta_i}$
$P_{x_n,1,\Theta_i}$
$p_{x_n,i}$
....
$\lim_{k\to\infty} p_{x_n,i}[P_{x_n,1,\Theta_i}]^k$
Iteration, k

Figure 1: Schematic depiction of operators Ψ and Φ, as applied to a measure $p_{x_n,i}$, highlighting the difference in how limits of measures are taken in the two cases. The strategy of the proof of Theorem 3.4 is to first characterise $p_{x_n,i}\Phi$ using the properties of limiting diffusions, before demonstrating the correspondence with $p_{x_n,i}\Psi$ via Theorem 4.2.

$\sigma^2(\theta)$. We will write $f(\theta) = \int \sigma^{-2}(\theta)d\theta$, $g(\theta) = \int \theta\sigma^{-2}(\theta)d\theta$ for $\theta \in (l,r)$ and for a fixed $c \in \mathcal{K}$ define

$$S(x) = \int_c^x \exp(-2n(\bar{x}_n f(\theta) - g(\theta))d\theta, \ \ \forall x \in \mathcal{K}.$$

We now give our main assumptions on the properties of the underlying measure.

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

1. *The function σ^2 is locally Lipschitz continuous in that for each $U > 0$ there is a constant K_U such that*

$$|\sigma^2(\theta) - \sigma^2(\theta')| \leq K_U|\theta - \theta'|, \ \ \forall\theta, \theta' \in \mathcal{K}_U.$$

2. *The function σ is non–degenerate and satisfies a linear growth condition, in that there exists a constant C_l such that*

$$0 < \sigma^2(\theta) \leq C_l(1+\theta^2), \ \ \forall \theta \in \mathcal{K}.$$

3. *For each $\mathcal{K}_u$, there exists an $\epsilon_u > 0$ such that*

$$\sup_{\theta \in \mathcal{K}_u} \int_{\mathcal{Y}} (x-\theta)^{2+\epsilon_u} \nu_\theta(x) dx < \infty.$$

4. *$S(l) = -\infty$ and $S(r) = \infty$.*

Remark 3.2. 1. We note that by the Lipschitz continuity and non-degeneracy ($\sigma(\theta) > 0$) the functions $f(\theta), g(\theta)$ are locally integrable.

2. These conditions are satisfied by a wide range of distributions and are not particularly restrictive. This is discussed further in Remark 3.9.

3. The function S is the scale function for the limiting diffusion and the fourth condition ensures that the boundaries of the parameter space play no role.

We will need one further assumption to ensure that our generalized data augmentation chain is well behaved. We write ν_θ^{*m} for the m-fold convolution of the measure ν_θ with itself.

Assumption 3.3. *The Markov chain $\{\theta_m(k) : k \geq 0\}$ with transition function*

$$P(\theta_m(1) \in A | \theta_m(0) = \theta) = \nu_\theta^{*m}((n+m)A - n\bar{x}_n), \ \ \theta \in \mathcal{K}, A \subset \mathcal{K}$$

is ψ-irreducible and aperiodic for large enough m.

We note that an easy sufficient condition for this assumption is the existence of a density for the measure ν_θ or indeed that there exists a density for the m-fold convolution of ν_θ. It is also easy to see that discrete measures, such as the Poisson distribution, will also satisfy this condition.

We now investigate those $\Theta \in \mathscr{C}_F$ whose basins of attraction contain moment-based point estimators. The next result generalises [9], Example 2.3, which considered the special case of the Normal distribution with mean θ and unit variance. For this case, when the initial measures were specified as $p_{x_n,0}(\theta) = \delta_{\bar{x}_n}$, repeated application of Ψ resulted in a sequence of densities $p_{x_n,i}(\theta), i \geq 1$ where the density $p_{x_n,i}(\theta)$ was $N(\bar{x}_n, \frac{1-2^{-i}}{n})$ with $p_{x_n,\infty}(\theta)$ being $N(\bar{x}_n, \frac{1}{n})$ - the Bayesian posterior using an improper uniform prior on θ.

Theorem 3.4. *Suppose that ν_θ satisfies Assumptions 3.1 and 3.3. Let*

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

For $i = 1, 2, 3, \ldots,$ let $c_i = 2 - 2^{-(i-1)}$. Then, for $n > C_l/\sqrt{2}$ the measures $p_{x_n,i}$ in the systems Θ_i, $i = 1, 2, 3, ...$ exist, where $\Theta_i = \Theta_{i-1}\Psi$ is given by

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

Moreover, the limiting system Θ_∞ is specified by

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

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 3.5. *Under Assumption 3.1, Θ_∞ is Bayesian if and only if ν_θ has a density or mass function, v_θ, which is a member of the one-parameter exponential family with sufficient statistic $\bar{x}_n$.*

Proof. Clearly Θ_∞ is Bayesian only if the likelihood $v_\theta(x_n) = \prod_{i=1}^n v_\theta(y_i)$ satisfies

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

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

Conversely, if $v_\theta(x)$ is a density or mass function from the one-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 Fisher information function $i(\theta) = \sigma^{-2}(\theta) = a'(\theta)$ implying that $a(\theta) = \int \sigma^{-2}(\theta)d\theta = f(\theta)$ and $c'(\theta) = a'(\theta)\theta$ in which case $c(\theta) = \int \theta\sigma^{-2}(\theta)d\theta = g(\theta)$. It follows that $p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma^2(\theta)}\exp\left(n(f(\theta)\bar{x}_n - g(\theta)\right\}$; 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 [10]. □

Although the maximum-likelihood prior arises from the construction we note that its use may not be recommended due to its implication in paradoxes - such as the marginalisation paradox [5] - that can arise. Other constructions that yield Bayesian posteriors without the direct use of Bayes' Theorem, such as those based on fiducial arguments, e.g. [20] may result in alternative prior specifications such as the Jeffreys prior. Nevertheless, our construction can be generalised to yield a range of possible priors.

For the one-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

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

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

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

This demonstrates a 1-1 correspondence between systems of 'shrinkage' estimators (which estimate θ as a weighted average of a specified value and the observed sample mean) and Bayesian analyses using conjugate prior distributions.

We now discuss distributions outside the 1-parameter exponential family and for which the observation $\bar{x}_n$ is not sufficient for θ. In this case, Theorem 3.4 demonstrates that $\mathscr{C}_F$ must contain both Bayesian and non-Bayesian systems of inference that are fixed by Ψ. As in the exponential-family case, Theorem 3.4 predicts that the general system of estimators for which

$$p_{x_n,0} = \delta_{\frac{n\bar{x}_n+ka}{n+k}}$$

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

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

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 $p_{x_n,\infty}(d\theta)$ may not coincide with $\pi(d\theta|x_n)$ for any prior $\pi(d\theta)$. Nevertheless, it is plausible that $p_{x_n,\infty}(d\theta)$ may give a reasonable approximation to a Bayesian posterior distribution obtained from an experiment in which $\bar{x}_n$ is observed with corresponding likelihood $L(\theta;\bar{x}_n)$ - that is $\pi(d\theta|\bar{x}_n) \propto \pi(d\theta)L(\theta;\bar{x}_n)$ for some $\pi(d\theta)$. We illustrate this in the following examples.

Example 3.6. The double exponential distribution has density given by

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

with mean given by θ and constant variance 2. In this case, Theorem 3.4 states that when $p_{x_n,0} = \delta_{\bar{x}_n}$,

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

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

Example 3.7. 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 3.4 states that when $p_{x_n,0}(\theta) = \delta_{\bar{x}_n}$, writing $p_{x_n,\infty}(\theta)$ for the density

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

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

The likelihood $L(\theta;\ \bar{x}_n)$ is not convenient to work with directly being proportional to θ^{-n} multiplied by 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 2 we see that the density for $p_{x_n,\infty}$ 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.

Remark 3.8. The form of $p_{x_n,\infty}(d\theta)$ highlights a connection with an approach to approximate Bayesian inference using contrasts, where the 'true' data likelihood is replaced by $\exp(-nU(x_n, \theta))$ where U is some function of the data and the parameter, leading to a contrast-based posterior density proportional to $\pi(\theta)\exp(-U(x_n, \theta))$ [19]. In our case, the contrast which is implicitly constructed is

$$U(x_n, \theta) = g(\theta) - f(\theta)\bar{x}_n$$

which, when minimised with respect to θ, recovers the point estimate specified by Θ_0. The limiting inferences also share some similarity with those constructed using Gibbs posteriors [15] where the usual likelihood is replaced with $\exp(-n\alpha R(x_n, \theta))$ where α is a constant playing the role of a temperature and R denotes a risk measure. We further note that the form of $p_{x_n,\infty}(d\theta)$ is determined from the relationship between the sampling mean and variance as given by $\sigma^2(\theta)$. Thus limiting inferences will be robust to model mis-specification so long as this aspect of the model ν_θ is specified correctly.

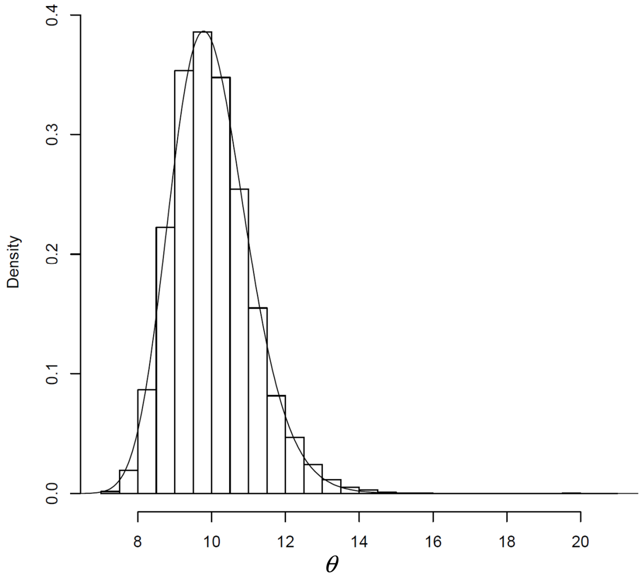

Figure 2: Comparison between $\pi(\theta|\bar{x}_n = 10)$ as estimated by Gibbs sampling and $p_{x_n,\infty}(\theta)$ for sample size $n = 30$.

Remark 3.9. We briefly discuss some of the conditions and assumptions used in Theorem 3.4.

1. The assumption that $\bar{x}_n \in \mathcal{K} = (l, r)$ may appear stringent. Nevertheless, for models

in which $\bar{x}_n \in [l, r]$ then it is automatic that if $\bar{x}_n \in (l, r)$ then so must be $\bar{x}_{n+m}$ for any outcome x_{n+m} obtained from x_n by augmenting it with m additional samples, and the constructions in Theorem 3.4 are valid so long as the *observed* sample mean $\bar{x}_n$ does not lie on the boundary of the parameter space. In instances where this does not hold, the nature of the resulting degeneracy of the distribution $p_{x_n,\infty}$ specified in Theorem 3.4 may be consistent with the behaviour of the Markov chains in the construction. For example, in the case of the Poisson distribution with $\bar{x}_n = 0$, $p_{x_n,\infty}(\theta) \propto \theta^{-1}e^{-n\theta}$ whose integral blows up on $(0, \epsilon)$. This is consistent with the behaviour of the level-m data augmentation chains in the construction of $p_{x_n,1}(.)$ from $p_{x_n,0}(.)$ which have an attracting state at $\theta = 0$ when $\bar{x}_n = 0$.

In cases where $\bar{x}_n$ may not lie in $[l, r]$ it may nevertheless be possible to modify the arguments used to prove Theorem 3.4 so that the result holds. Example 3.10 provides one such example.

2. The requirement that $n > C_l/\sqrt{2}$ places a restriction on the sample sizes for which the proof of Theorem 3.4 is valid. For many models this is satisfied for $n = 1$. In some situations where $C_l/\sqrt{2} \geq 2$, Theorem 3.4 may be extendable to all sample sizes through appropriate modification of the constructions. For example, when $\nu_\theta(.)$ lies in the exponential family the limiting inferences are Bayesian. Having obtained the (Bayesian) limiting inferences for $n > C_l/\sqrt{2}$ we apply the operator Ψ to the system

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

where $p^*_{x_n} = p_{x_n,0}$ if $n \leq C_l/\sqrt{2}$, and $p^*_{x_n} = p_{x_n,\infty}$ otherwise. It is now immediate that $\Theta^*\Psi$ is the limiting system from Theorem 3.4 with no restriction on the value of n since, for sufficiently large values of m, the chains in the construction will be standard Gibbs samplers.

We also note that the proof of Theorem 3.4 holds whenever $\sigma^2(\theta) \leq C_l(a+\theta^2)$ for any $a > 0$. To see this, we consider scaled observations $y^* = y/\sqrt{a}$ and parameter $\theta^* = \theta/\sqrt{a}$ noting that the variance for the measure $\nu^*_{\theta^*}$ satisfies $\sigma^{*2}(\theta^*) \leq C_l(1+\theta^{*2})$. Using quantile matching of updates to couple the generalised data augmentation chains in the construction of Theorem 3.4 for the unscaled system with the corresponding chains for the scaled system, we can apply Theorem 3.4 to the scaled system and deduce that Theorem 3.4 holds for the unscaled system. Details are not shown here. Thus, for Example 3.6, Theorem 3.4 holds for all sample sizes n despite the fact that $\sigma(\theta) = 2$ would strictly imply a minimum value of $C_l = 2$ requiring $n \geq 2$ in Theorem 3.4.

Example 3.10. The case of the Pareto distribution, where we try to estimate the tail parameter, provides a further example where assumptions may break down but the result of Theorem 3.4 remains valid.

Consider the setting where ν_θ has the Pareto density with a variance, so that for $\alpha > 2$ we have

$$\nu_\theta(dx) = \alpha x^{-\alpha-1}dx, \quad x \geq 1.$$

Parametrising by the mean gives $\alpha = \theta/(\theta-1)$ so that $\mathbb{E}(Y) = \theta$ and we have

$$\sigma^2(\theta) = \frac{\theta(\theta-1)^2}{2-\theta}, \quad \theta \in (1,2).$$

Note that this function only satisfies two of the conditions in Assumptions 3.1 in that it is locally Lipschitz and satisfies the moment condition. However it does not satisfy linear growth or the condition for the boundaries of the interval. The fact that both $S(1), S(2)$ are finite shows that they can be reached by the associated diffusion. Indeed they can be reached in finite time. In this case we do need to make an adjustment when considering the Markov chains as it is possible to choose a sample for which the parameter moves above 2 and hence we need to restrict our chain by setting it to be 2 in such a case. This ensures that 2 acts as a reflecting boundary for the diffusion. By modifying our arguments and using the explicit form of the variance we can still establish the weak convergence of the chains to the diffusion. This diffusion is unique up until the first exit time from $(1, 2)$. However by reflecting at the end points we can extend the solution uniquely to all times.

Using scale and speed measures for the diffusion shows that the boundaries are regular and there is a limiting stationary distribution for the diffusion which should be the limit of the stationary distributions for the Markov chains. The limiting system of inference is thus given by

$$\Theta_\infty = \{p_{x_n,\infty}(d\theta) \propto (2-\theta)\theta^{2n\bar{x}_n-1}(\theta-1)^{n-2-2n\bar{x}_n} \exp\left(-\frac{n(\bar{x}_n-1)}{\theta-1}\right) d\theta | n \in \mathbb{N},\ \ x_n \in \mathcal{Y}^n\}.$$

4 Proof of Theorem 3.4

We retain the notation of earlier sections and consider the family of probability measures ν_θ where the parameter space is a (not necessarily strict) 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 3.4 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. We begin by giving 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 $\{\theta_{i,m}(k) : k \geq 0\}$ with transition kernel $P_{x_n,m,\Theta_{i-1}}$ as described after Principle 2.2. We will assume that, for all possible observed samples x_n, the sample mean $\bar{x}_n$ will lie in the allowable parameter space $\mathcal{K}$. If this is not the case we may have degeneracies and we therefore avoid such situations. We first define a continuous-time process $\{\theta_i^m(t) : t \geq 0\}$ from the Markov chain $\{\theta_{i,m}(k) : k \geq 0\}$ by interpolation, setting $\theta_i^m(t) = \theta_{i,m}(\lfloor mt \rfloor)$, noting that the chain $\theta_{i,m}$ and the process θ_i^m are identical so far as the existence and nature of the stationary distribution are concerned.

The main work in proving Theorem 3.4 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 pathwise unique strong 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 would be enough to have a unique weak solution to the equation (4) for our purposes but our assumptions give us the existence of a strong solution. This result in essence enables one to demonstrate that the second refinement operator Φ, introduced in Section 2, has Θ_∞ in Theorem 3.4 as a fixed point. With a little more work we can deduce the following version of Theorem 3.4 to show that Θ_∞ is the fixed point of Ψ as required by the Theorem.

Theorem 4.2. *Under Assumptions 3.1 and 3.3 the Markov chains $\{\theta_{i,m}(k) : k \geq 0\}$ 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. *Under Assumption 3.1:*

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

$$\theta_i(t) = \bar{x}_n + (\xi - \bar{x}_n)e^{-nt} + \int_0^t e^{-n(t-s)}\sigma_i(\theta_i(s))dW_s, \quad t \geq 0. \tag{5}$$

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

(3) If $\sqrt{2}n > C_l$, there is a unique stationary distribution of (4) and it is given by the density function

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

where

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

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 3.4 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

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

A key tool in establishing Theorem 4.1 is Corollary 7.4.2 of [7], 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. For a stochastic process X we write $T_{\partial\mathcal{K}}(X) = \inf\{t \geq 0 : X(t) \notin \mathcal{K}\}$ for the exit time from $\mathcal{K}$. We will abuse notation by using the same symbol for an exit time when the time parameter is discrete. For the diffusion this time will coincide with the hitting time of the boundary $\partial\mathcal{K}$. For the discrete case it is the exit time from $\mathcal{K}$.

Theorem 4.5 (Ethier and Kurtz). *Let $X = \{X(t); 0 \leq t \leq T_{\partial\mathcal{K}}(X)\}$ be the diffusion process taking values in $\mathcal{K}$ satisfying the SDE*

$$dX(t) = b(X(t))dt + \sigma(X(t))dW(t), \;\; X(0) \sim \phi,$$

up until first exit from $\mathcal{K}$, where b is a continuous function and σ is also continuous and $X(0)$ is drawn according to a measure $\phi \in \mathcal{P}(\mathcal{K})$. Let $Y_m = \{Y_m(k); 0 \leq k \leq T_{\partial\mathcal{K}}(Y_m)\}$ be a discrete time Markov chain taking values in $\mathcal{K}$ with law $\mathbb{P}_m$ and set $X_m(t) = Y_m([mt])$. Let

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

where $\mathbb{E}_m^x$ denotes expectation for the Markov chain Y_m started from the point $x \in \mathcal{K}$.

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

$$\lim_{m\to\infty} \sup_{x\in\mathcal{K}_u} |\mu_m(x) - b(x)| = 0, \tag{6}$$

$$\lim_{m\to\infty} \sup_{x\in\mathcal{K}_u} |\sigma_m^2(x) - \sigma^2(x)| = 0, \tag{7}$$

and

$$\lim_{m\to\infty} \sup_{x\in\mathcal{K}_u} m\mathbb{P}_m(|Y_m(1) - x| \geq \epsilon) = 0. \tag{8}$$

Then X_m converges weakly to X.

We are now in a position to proceed with the inductive proof of Theorem 3.4.

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 $i = 1$ and our observation x_n is augmented by a further m samples we obtain a generalised data-augmentation chain with updates specified by

$$\theta_{1,m}(k+1) = \frac{n\bar{x}_n + m\bar{Y}_m^{(k)}}{n+m}, \quad k = 0, 1, 2, \ldots$$

where $\bar{Y}_m^{(k)} = \frac{1}{m}\sum_{j=1}^m Y_j^{(k)}$, with $Y_j^{(k)}$ samples from the measure $\nu_{\theta_{1,m}(k)}$. 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 for the Markov chain $\{\theta_m(k) : k \geq 0\}$ where the limiting diffusion process is 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

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

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

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

which establishes (6).

By construction we have

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

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

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

To handle the tail condition (8) we observe that by Assumption 3.1 for each $x \in \mathcal{K}_u$ there is an $\epsilon > 0$ such that, writing $\mathbb{E}$ for the expectation with respect to ν_x, $\mathbb{E}(Y_1^{(0)} - x)^{2+\epsilon} < \infty$. Letting $p = 2+\epsilon$ we see that for all $x \in \mathcal{K}_u$

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

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

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

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

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

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

Using this with Theorem 4.2 and Lemma 4.3 (3) we have established the form of Θ_1 as given in Theorem 3.4. Note that by Assumption 3.1(4) the boundaries of the domain $\mathcal{K}$ are natural and therefore cannot be reached in finite time, so $T_{\partial\mathcal{K}} = \infty$ almost surely.

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 $\{\theta_{i+1,m}(k) : k \geq 0\}$ with transition kernel P_{x_n,m,Θ_i} as described after Principle 2.2, where $\theta_{i+1,m}(k+1)$ is drawn from $p_{x_{n+m},i}$ where the augmented sample $x_{n+m} = (x_n, y_{n+1}, \ldots, y_{n+m})$ is obtained by drawing $(y_{n+1}, \ldots, y_{n+m})$ from $\nu_{\theta_{i+1,m}(k)}$.

We have established that the limit as $m \to \infty$ of the chain is a diffusion process and that the i-th system of inferences is obtained from the stationary distribution of the diffusion process θ_i. In order to generate the $k+1$-th sample from $p_{x_{n+m},i}(\theta)$ we take $\theta_i^{n,m,k}$, a copy of the diffusion process given by (4) with $n+m$ and $\bar{x}_{n+m}$ replacing n and $\bar{x}_n$ respectively. Thus the draw $\theta_{i+1,m}(k+1)$ from $p_{x_{n+m},i}$ is a sample from the stationary distribution $p_{x_{n+m},i}(\theta)$ of the diffusion $\theta_i^{n,m,k}$. We will show that it is enough to use an approximate sample $\theta_i^{n,m,k}(\tau_m)$, obtained by running the diffusion with initial value

$\theta_i^{n,m,k}(0) = \bar{x}_{n+m}$, for a sufficiently long time τ_m (determined in Proposition 4.9). This gives us a sequence of approximate chains for each (m, τ_m), where if we let $\tau_m = \infty$ we recover the exact sampling distribution. We also note that

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

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

Thus we have our approximate Markov chain and the transition to the new state can be expressed as

$$\begin{aligned}\theta_{i+1,m}(k+1) &= \bar{x}_{n+m} + \int_0^{\tau_m} e^{-(n+m)(\tau_m - t)}\sigma_i(\theta_i^{n,m,k}(t))dW_t^k,\\ &= \theta_{i+1,m}(k) + \frac{n}{n+m}(\bar{x}_n - \theta_{i+1,m}(k)) + \frac{1}{m+n}R_m(\theta_{i+1,m}(k))\\ &\qquad + \int_0^{\tau_m} e^{-(n+m)(\tau_m - t)}\sigma_i(\theta_i^{n,m,k}(t))dW_t^k,\end{aligned}$$

where the first three terms in the expression appeared in the previous case ($i = 1$) and for each k, W^k is an independent Brownian motion. The last term gives our approximate sample from the stationary distribution.

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

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

where

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

We will write $\theta_i^{n,m}$ for $\theta_i^{n,m,0}$ 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 will 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

$$\mathbb{E}N_m(t)^2 = \mathbb{E}\langle N_m \rangle_t = \int_0^t e^{-2(n+m)(t-s)} \mathbb{E}\sigma_i^2(\theta_i^{n,m}(s))ds. \tag{10}$$

Lemma 4.7. *For the augmented sample for the chain started from x, for each $p \geq 2$ there exists a constant C_p such that*

$$\mathbb{E}|\bar{x}_{n+m}|^p \leq C_p \left(|x|^p + \frac{m^{p/2}(1 \vee |\bar{x}_n|^p \vee |x|^p)}{(n+m)^p} \right).$$

Proof. This follows the proof for the $i = 1$ case. We have

$$\begin{aligned}
\mathbb{E}|\bar{x}_{n+m}|^p &= \mathbb{E}\left| x + \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{m+n} R_m(x) \right|^p \\
&\leq c_p|x|^p + c_p\mathbb{E}\left| \frac{n}{n+m}(\bar{x}_n - x) \right|^p + c_p\mathbb{E}\left| \frac{1}{m+n}|R_m(x)| \right|^p \\
&\leq c_p|x|^p + c_p' \frac{|\bar{x}_n|^p \vee |x|^p}{(m+n)^p} + c_p\mathbb{E}\left(\frac{1}{m+n}|R_m(x)| \right)^p .
\end{aligned}$$

From the argument for (9) we have that $\mathbb{E}|R_m(x)|^p \leq cm^{p/2}$ for all $p \geq 1$ and this gives the result. □

Lemma 4.8. *Under Assumption 3.1, for each $1 \leq p \leq 2 + \epsilon$, there is a constant c_p such that*

$$\mathbb{E}|N_m(t)|^p \leq c_p \frac{1 \vee |\bar{x}_n|^p \vee |x|^p}{(n+m)^{p/2}}, \quad \forall t \geq 0, \quad m \geq 1.$$

Proof. For this we note that by Burkholder-Davis-Gundy and Hölder's inequality, when $p \geq 2$,

$$\begin{aligned}
\mathbb{E}|N_m(t)|^p &\leq c_p \mathbb{E}\langle N_m\rangle_t^{p/2} \\
&= c_p \mathbb{E}\left(\int_0^t e^{-2(n+m)(t-s)}\sigma_i^2(\theta_i^{m,n}(s))ds\right)^{p/2} \\
&\leq c_p \left(\int_0^t e^{-p(n+m)(t-s)/(p-2)}ds\right)^{p/2-1}\int_0^t e^{-p(n+m)(t-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \\
&= c_p \left(\frac{p-2}{p(n+m)}\right)^{p/2-1}(1-e^{-p(n+m)t/(p-2)})^{p/2-1}\int_0^t e^{-p(n+m)(t-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \\
&\leq c_p \left(\frac{p-2}{p(n+m)}\right)^{p/2-1}\int_0^t e^{-p(n+m)(t-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds. \qquad (11)
\end{aligned}$$

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

$$\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s)) \leq c_p(1 \vee \mathbb{E}|\bar{x}_{n+m}|^p), \ \ \forall s \geq 0, \ \ m \geq 1. \qquad (12)$$

By Lemma 4.7,

$$\mathbb{E}|\bar{x}_{n+m}|^p \leq c(1 \vee |\bar{x}_n|^p \vee |x|^p), \ \ \forall m \geq 1.$$

Thus

$$\int_0^t e^{-p(n+m)(t-s)/2}\mathbb{E}\sigma_i^p(\theta_i^{m,n}(s))ds \leq c_p' \frac{2}{p(m+n)}(1-e^{-p(m+n)t/2})(1 \vee |\bar{x}_n|^p \vee |x|^p),$$

and hence, replacing this in (11), for a constant C we have

$$\mathbb{E}|N_m(t)|^p \leq \frac{C(1 \vee |\bar{x}_n|^p \vee |x|^p)}{(m+n)^{p/2}}.$$

For $1 \leq p < 2$ we can follow a similar, slightly easier argument where we replace the use of Hölder's inequality with the concavity of the function $x^{p/2}$ to enable us to bring the expectation inside the integral in a similar calculation to that leading to (11). □

Proposition 4.9. *Under Assumption 3.1 and for $\tau_m > \frac{\log m}{2(m+n)}$, $n > C_l/\sqrt{2}$, the process $\{\theta_i^m(t) : t \geq 0\}$ converges weakly to $\{\theta_i(t) : t \geq 0\}$, the pathwise unique strong solution to*

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

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

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

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

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,

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

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

$$|\sigma^2_{m,i+1}(x) - \sigma^2_{i+1}(x)| \leq \frac{n^2m}{(n+m)^2}(\bar{x}_n - x)^2 + |\frac{m}{(m+n)^2}\mathbb{E}R_m(x)^2 - \sigma^2(x)| + |m\mathbb{E}N^2_m(\tau_m) - \frac{1}{2}\sigma^2_i(x)|.$$

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 $x \in \mathcal{K}_u$.

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

$$\begin{aligned} &|\mathbb{E}mN^2_m(\tau_m) - \frac{\sigma^2_i(x)}{2}| \\ &\quad = \frac{m}{2(m+n)}|\left(\mathbb{E}\int_0^{\tau_m} 2(n+m)e^{-2(n+m)(\tau_m - t)}\sigma^2_i(\theta^{n,m}_i(t))dt - \frac{m+n}{m}\sigma^2_i(x)\right)| \\ &\quad \leq \frac{m}{2(n+m)}\int_0^{\tau_m} 2(m+n)e^{-2(n+m)(\tau_m - t)}\mathbb{E}\left|\sigma^2_i(\theta^{n,m}_i(t)) - \sigma^2_i(x)\right| dt \\ &\qquad\quad + \frac{n}{2(m+n)}\sigma^2_i(x) + e^{-2(n+m)\tau_m}\sigma^2_i(x). \end{aligned} \tag{13}$$

Now, by Asssumption 3.1(1), as $\sigma^2_i(x)$ is locally Lipschitz, we have for a fixed $x \in \mathcal{K}$ that there is a constant K_U such that for $x \in \mathcal{K}_U$,

$$\begin{aligned} \mathbb{E}|\sigma^2_i(\theta^{n,m}_i(t)) - \sigma^2_i(x)| &\leq K_U\mathbb{E}\left(|\theta^{n,m}_i(t) - x|; \theta^{n,m}_i(t) \in \mathcal{K}_U\right) \\ &\qquad + \mathbb{E}\left(|\sigma^2_i(\theta^{n,m}_i(t)) - \sigma^2_i(x)|; \theta^{n,m}_i(t) \notin \mathcal{K}_U\right). \end{aligned} \tag{14}$$

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

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

Taking p-th moments and using previous estimates, we have

$$\mathbb{E}|\theta_i^{n,m}(t) - x|^p \quad \leq \quad c_p|\frac{n}{n+m}(\bar{x}_n - x)|^p + c_p\mathbb{E}(\frac{1}{m+n}|R_m(1)|)^p + \mathbb{E}|N_m(t)|^p$$

Under our assumptions, by Lemma 4.8, and the expression for $\mathbb{E}|R_m(x)|^p$ in (9), there will be a constant c such that

$$\mathbb{E}|\theta_i^{n,m}(t) - x|^p \leq \frac{c}{m^{p/2}}, \ \ \forall t \geq 0. \tag{15}$$

Thus for $p = 1$ we have the required bound.

For the second term on the right hand side of (14) we have that, as $x \in \mathcal{K}_U$, for $\theta_i^{n,m}(t) \notin \mathcal{K}_U$ there is a $u > 0$ such that $|\theta_i^{n,m}(t) - x| > u$. By the linear growth bound on σ, Hölder's and Markov's inequalities, we have

$$\begin{aligned} \mathbb{E}\left(|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; \theta_i^{n,m}(t) \notin \mathcal{K}_U\right) \quad &\leq \quad \mathbb{E}\left(|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; |\theta_i^{n,m}(t) - x| \geq u\right) \\ &\leq \quad C(1 + x^2 + \sigma_i^2(x))\mathbb{P}(|\theta_i^{n,m}(t) - x| \geq u) \\ &\qquad +C'\mathbb{E}\left(|\theta_i^{n,m}(t)) - x|^2 I_{\{|\theta_i^{n,m}(t) - x| \geq u\}}\right) \\ &\leq \quad (C(1 \vee |x|^2)u^{-p} + C'u^{2-p})\mathbb{E}|\theta_i^{n,m}(t) - x|^p, \end{aligned}$$

for $2 < p < 2 + \epsilon$. Using (15) we have, for such a p, that

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

Thus, substituting into (13) and using our condition on τ_m we have $e^{-2(m+n)\tau_m} \leq 1/m$, which gives

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

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. By Markov's inequality, and an application of Lemma 4.8, we have

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

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 show that under our assumptions there exist invariant distributions for the data augmentation chains. Our estimates on the moments in Lemma 4.3 shows that these distributions are tight and hence there is a limit stationary distribution for the diffusion. By the uniqueness of the invariant measure for the diffusion we recover the weak convergence of the whole sequence of stationary distributions to this limit.

The details are given using Theorem A.4 and Corollary A.5 in the appendix. □

Finally we note that these results combine to give the proof of Theorem 3.4.

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 3.4. We will also make some stronger assumptions than in the moment-estimator case.

In this section we assume the existence of a density for our measure and hence consider the model $\nu_\theta(dy) = v_\theta(y)dy$ and let $l(\theta, y) = \log v_\theta(y)$, $i(\theta) = \mathbb{E}_Y\left(-\frac{\partial^2 l}{\partial \theta^2}\right)$, $a(\theta) = \mathbb{E}_Y\left(\frac{\partial^2 l}{\partial \theta^2}\frac{\partial l}{\partial \theta}\right)$, and $c(\theta) = \mathbb{E}_Y\left(-\frac{\partial^3 l}{\partial \theta^3}\right)$. Suppose that ν_θ satisfies regularity conditions that allow the interchange of the order of integration with respect to y and differentiation with respect to θ. We can then easily verify the identity

$$\frac{\partial i(\theta)}{\partial \theta} + a(\theta) + c(\theta) = 0. \tag{16}$$

Now let

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

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 3.4 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. [4]), 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

$$d\theta_1 = \frac{1}{i^2(\theta_1)} \left(a(\theta_1) + \frac{c(\theta_1)}{2} + l_n'(\theta_1)i(\theta_1) \right) dt + \sqrt{\frac{1}{i(\theta_1)}} dB. \tag{17}$$

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 $\theta_{1,m}$ satisfies

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

By [6] Theorem (4.53) conditions for the existence and uniqueness of a weak solution are that $i(x) < \infty$ for $x \in \mathcal{K}$ and the quantities $\frac{a(x)}{i(x)} + \frac{c(x)}{2i(x)} + l'_n(x)$ and $i(x)$ are locally integrable in $\mathcal{K}$.

Theorem 5.2. *1. Under Assumption 5.1, the interpolated chain θ^m_1 converges weakly to θ_1.*

2. The associated system of inferences, obtained from the stationary measure of the diffusion in (17), *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 3.4 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 (16) this can be written as

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

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 $\gamma_j = 1/(1-2^{-j}), j = 1, 2, \dots$, 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 $\{\theta_{j,m}(k) : k \geq 0\}$ and the continuous-time interpolation $\{\theta_j^m(t) : t \geq 0\}$ determined by setting $\theta_j^m(t) = \theta_{j,m}(\lfloor mt \rfloor)$. 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 [8] 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

$$\mathbb{E}(\theta - \hat{\theta}_m | y_1, ..., y_m) = \left[-\frac{\partial^2 l_m}{\partial \theta^2}\right]_{\hat{\theta}_m}^{-2} \left(\frac{1}{2}\left[\frac{\partial^3 l_m}{\partial \theta^3}\right]_{\hat{\theta}_m} - \left[\frac{\partial^2 l_m}{\partial \theta^2}\right]_{\hat{\theta}_m} \left[\frac{\partial \log \pi_L}{\partial \theta}\right]_{\hat{\theta}_m}\right) + o(1/m) \tag{18}$$

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

$$\mathbb{E}(\theta - \hat{\theta}_m | y_1, ..., y_m) = \frac{1}{\gamma_{j-1}} \left[-\frac{\partial^2 l_m}{\partial \theta^2}\right]_{\hat{\theta}_m}^{-2} \left(\frac{1}{2}\left[\frac{\partial^3 l_m}{\partial \theta^3}\right]_{\hat{\theta}_m} - \left[\frac{\partial^2 l_m}{\partial \theta^2}\right]_{\hat{\theta}_m} \left[\frac{\partial \log \pi_L}{\partial \theta}\right]_{\hat{\theta}_m}\right) \tag{19}$$

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

Started from the point x we discern two components to the increment $\theta_{j,m}(1)-x$ - one given by $\hat{\theta}_m - x$ and the other arising when we sample $\theta_{j,m}(1)$ from a density proportional to $\pi_L(\theta)L_m(\theta)^{\gamma}_{j-1}$. We approximate expectations over $y_1, ..., y_m$ in (19) 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

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

and

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

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

$$d\theta_j = \frac{1}{i^2(\theta_j)} \left(a(\theta_j) + \frac{c(\theta_j)(\gamma_{j-1}+1)}{2\gamma_{j-1}} + \frac{i(\theta_j)}{\gamma_{j-1}} \frac{\partial \log \pi_L}{\partial \theta_j} \right) dt + \sqrt{\frac{\gamma_{j-1}+1}{\gamma_{j-1} i(\theta_j)}} dB. \qquad (20)$$

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 weak solution.

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

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

Theorem 5.4. *Under Assumption 5.3, for $j \geq 2$, the process θ_j^m converges weakly to θ_j, the solution to* (20)*. 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 (16). 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

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

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

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

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

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 (17), 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 may not arise when Ψ is applied to an initial system based on maximum-likelihood estimation if the model is not sufficiently regular. Consider again the uniform distribution of Example 3.7 reparameterised for convenience so that $\nu_\theta(y) = \theta^{-1}, 0 < y < \theta$ and the system of (maximum likelihood) estimators

$$\Theta_{MLE} = \{p_{x_n,i}(\theta) = \delta_{x_{(n)}} |\, n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\},$$

where $x_{(n)}$ is the maximum of the observations. Now Θ_{MLE} is trivially a fixed point of Ψ. Therefore consider a more general system of MLE-based estimators of the form

$$\Theta_{\mathbf{a}} = \{p_{x_n,i}(\theta) = \delta_{a_n x_{(n)}} |\, n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\},$$

with $a_n = \frac{n+1}{n}$ giving an unbiased system of 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 [17].

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 $a \to 0$.

Set $\Theta_0 = \Theta_{\mathbf{a}}$ and consider the construction of $\Theta_1 = \Theta_0\Psi$. 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 $\{\theta_{0,m}(k) : k \geq 0\}$ when we augment x_n with m additional observations. For this chain, for $k = 1, 2, 3, \ldots$

$$\theta_{0,m}(k) = a_{n+m} \sup\{1, \eta_k \theta_{0,m}(k-1)\},$$

where $\{\eta_k\}$ are i.i.d. Beta(m, 1), and $\eta_k\theta_m(k-1)$ represents the supremum of the m additional samples imputed during the update process. The corresponding chain for $\lambda_{0,m} = \log\theta_{0,m}$ has update

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

where the $\{\xi_k\}$ are i.i.d. Exp(m). Set $c_m = m\log a_{n+m}$, $\nu_{0,m}(k) = m\lambda_{0,m}(k) - c_m$ and write the update as

$$\nu_{0,m}(k) = \sup\{0, \nu_{0,m}(k-1) + \zeta_m(k)\} \tag{21}$$

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

$$S_r = \sum_{i=1}^{r} \zeta_m(i), \tag{22}$$

It follows from standard results that $\nu_{0,m} \sim \sup_{r\geq 1} S_r = M^{(m)}$ where $\nu_{0,m}$ denotes the stationary distribution of the Markov chain (21).

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 $\{\nu_{0,m}(k)\}$ and $\{\lambda_{0,m}(k)\}$. Therefore we must assume that $c_m < 1$ for all but finitely many m for $\Theta_0 = \Theta_{\mathbf{a}} \in \mathscr{C}$. 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, $\nu_{0,m}(k) \to \nu_{0,m}$ in distribution as $k \to \infty$, where $\nu_{0,m}$ is stochastically dominated by some random variable, τ, independent of m. It is then immediate that $\lambda_{0,m}(k) \to \lambda_{0,m}$ and that $\lambda_{0,m}$ must tend to 0 in probability as $m \to \infty$.

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

$$(1 - c_m)M^{(m)} \sim (1 - c_m)\nu_{0,m} \to M$$

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

Now consider the large-m behaviour of

$$\lambda_{0,m} \sim \frac{c_m}{m} + \frac{\nu_{0,m}}{m} \sim \frac{c_m}{m} + \frac{(1 - c_m)\nu_{0,m}}{m(1 - c_m)},$$

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 $\lambda_{0,m} \to 0$ weakly.

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

$$\lambda_{0,m} \to \mathrm{Exp}(2(n + \mu)).$$

In either case (1) or (2)(a) our system of inferences (for the log-transformed parameter) is $\Lambda^{(1)} = \{p_{x_n,i}(\lambda) = \delta(\lambda) \mid n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\}$ with corresponding system for θ, for the case of general $x_{(n)}$, given by

$$\Theta_1 = \{p_{x_n,1}(\theta) = \delta_{x_{(n)}} \mid n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\} = \Theta_{\mathbf{MLE}}.$$

Thus a single application of Ψ maps $\Theta_0 = \Theta_{\mathbf{a}}$ to the fixed point $\Theta_{\mathbf{MLE}}$.

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

$$\Theta_1 = \{p_{x_n,1}(\theta) \propto \theta^{-2n}\theta^{-2\mu-1} \mid n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\}. \tag{23}$$

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

$$\Theta_1 = \{p_{x_n,1}(\theta) \propto \sigma(\theta)^{-2}L(\theta; x_n)^2 \mid n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\},$$

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 Θ_1 to $\Theta_{\mathbf{MLE}}$. Working in the λ parameterisation we apply Ψ to the system

$$\Lambda_1 = \{p_{x_n,1}(\lambda) \propto 2(n+\mu)e^{-2(n+\mu)(\lambda - \log x_{(n)})}, \lambda > x_{(n)} \mid n \in \mathbb{N}, x_n \in \mathbb{R}_0^n\},$$

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

$$\lambda_{1,m}(k) = \sup\{\eta_{1,k}, \lambda_{1,m}(k-1) - \eta_{2,k} + \eta_{1,k}\},$$

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 $\lambda_{1,m}(0)$, 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 $\{\zeta_m(k) : k \geq 0\}$. The result is then immediate.

6 Discussion

In this paper we have shown how the generalised data augmentation principles introduced in [9] can be applied to elicit connections between classical point estimation and Bayesian (or Bayesian-like) inference where the latter is constructed from the former using refinement operators. This contrasts with other approaches to constructing 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 with fixed points of refinement operators representing maximally preferable inferences. Our results show that, in the 1-dimensional case, 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.

We consider how the results of the paper might be extended to higher-dimensional models. One natural generalisation of the moment-based constructions (Theorem 3.4) concerns the case where the samples remain one-dimensional but $\boldsymbol{\theta} \in \mathcal{K} \subset \mathbb{R}^d$ parameterises the sampling model in terms of the first d moments of $v_\theta(dy)$ and the initial system of point estimators Θ_0 is formed from measures, $p_{x_n} = \delta_{\mathbf{s}(x_n)}$ where $\mathbf{s}(x_n) = (s_1(x_n), ..., s_d(x_n))$ and $s_i(x_n)$ denotes the i^{th} sample moment of $x_n = (y_1, ..., y_n)$. Under regularity conditions that guarantee $\mathbf{s}(x_n) \in \mathcal{K}$ for any observed x_n, then the generalisation of the construction of the refinement operator Ψ to the d-dimensional setting is clear. The data augmentation chain will have for $i = 1$

$$\boldsymbol{\theta}_{1,m}(k+1) = \frac{ns(x_n) + \sum_{j=1}^{m} \mathbf{Y}_j^{(k)}}{n+m},$$

where the $\mathbf{Y}_j^{(k)} = (Y_j^{(k)}, (Y_j^{(k)})^2, ..., (Y_j^{(k)})^d)$ and the values $Y_j^{(k)}$ are i.i.d. draws from the distribution $\nu_{\boldsymbol{\theta}_{1,m}(k)}$. By rewriting this and applying a multidimensional version of the technique in Section 4 we will obtain a limiting SDE of the form

$$d\boldsymbol{\theta}_1 = n(\bar{\mathbf{x}}_n - \boldsymbol{\theta}_1)dt + \Sigma(\boldsymbol{\theta}_1)dW,$$

where W is a d-dimensional Brownian motion and $A(\boldsymbol{\theta}_1) = \Sigma(\boldsymbol{\theta}_1)\Sigma(\boldsymbol{\theta}_1)^T$ is the covariance matrix of $(Y, Y^2, ..., Y^d)$ under $\nu_{\boldsymbol{\theta}}$. Then our putative stationary distribution would be the solution to the PDE

$$\mathcal{A}^* p_1 = -\sum_j \frac{\partial}{\partial \theta_j} \left(n\left((\bar{\mathbf{x}}_n)_j - \theta_j\right) p_1\right) + \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \left(A_{jk}(\boldsymbol{\theta}_1) p_1\right) = 0.$$

Challenges inherent in generalising Theorem 3.2 to this setting include the identification of appropriate conditions on the moments of the distribution $\Sigma(\boldsymbol{\theta}_1)$. We note a potential link between the results of this paper and Approximate Bayesian Computation (ABC) [1]. Under this approach parameters in models with intractable likelihoods are inferred by replacing the observed data x_n with a (possibly mutivariate) summary statistic $T(x_n)$ and exploring the posterior $\pi(\boldsymbol{\theta}|T(x_n))$. The need to compute a likelihood is avoided by drawing samples of T from its sampling distribution given $\boldsymbol{\theta}$ and comparing with the observed $T(x_n)$. Our methods may have potential for constructing appropriate approximations to $\pi(\boldsymbol{\theta}|T(x_n))$ without the need for extensive simulation (*cf* Example 3.7) and this may be worthy of further investigation.

More straightforward may be the derivation of higher-dimensional results that demonstrate that Bayesian inferences arise as limiting, preferable systems for initial systems of maximum likelihood estimators to provide an alternative motivation for the use of

Bayesian procedures to that provided by the Bernstein Von-Mises Theorem. The latter justifies Bayesian procedures in terms of insensitivity of posteriors with respect to prior choice and consistency of estimates derived from the posterior as the sample size increases while the justification in terms of preferability would apply to any sample size. In the settings we consider, imputed observations are sequences of i.i.d samples from the model $\nu_\theta(y)$. This approach reflects classical asymptotic analysis and elicits connections with concepts, such as the maximum-likelihood prior, developed using classical asymptotics. Extending this approach to higher-dimensional models would however be challenging given the need to understand the distributions of MLEs in large-sample settings and to develop multi-dimensional diffusion approximations to the generalised data-augmentation chains arising. Alternatively we can note that imputed observations arise from 'thought experiments'. These can be designed in an arbitrary manner so long as the model for the observed data, $x_{obs} \sim \nu_{\boldsymbol{\theta}}(.)$ where $\theta \in \mathbb{R}^d$, is preserved. Now consider the sequence of experiments $x_n, n = 1, 2, ...$ where $x_n = (x_{obs}, y_1, ..., y_{n-1})$ and $y_1, ..., y_{n-1}$ is a random sample (independent of x_{obs}) from a multivariate normal distribution MVN$(\boldsymbol{\theta}, I_d)$. Starting from an initial system $\boldsymbol{\Theta_0}$ of maximum likelihood estimators and applying the operator Ψ recursively it should be feasible to demonstrate, subject to modest conditions on the likelihood $L(\boldsymbol{\theta}; x_0)$, that a Bayesian limiting system is reached, thanks to the simplifications arising from the normality of the augmenting data when demonstrating diffusion limits.

The results of the paper offer 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.

Acknowledgements

The authors are grateful to Seva Shneer, Stan Zachary, Jennie Hansen and Andrew Cairns for useful discussions in the course of this research and to the Associate Editor and two anonymous referees for their constructive comments on an earlier version of this paper.

A Proofs of Lemma 4.3 and Theorem 4.2

We recall the scale and speed measure approach to one-dimensional diffusions. The scale function for our diffusion satisfying (4) is given by fixing a $c \in \mathcal{K}$ and setting

$$S(x) = \int_c^x \exp(-2\int_c^y \frac{n(\bar{x}_n - z)}{\sigma^2(z)} dz) dy, \forall x \in \mathcal{K}.$$

By our definitions of f, g this is the same form as given before the statement of Assumption 3.1. The speed measure is then given by

$$m(dx) = \frac{2dx}{\sigma^2(x)S'(x)}.$$

We now recall Lemma 4.3 and provide a proof.

Lemma A.1. *Under Assumptions 3.1 we have:*

(1) There exists a pathwise unique strong solution to the SDE (4), $\{\theta_i(t) : t \geq 0\}$ which

can be written in integral form as

$$\theta_i(t) = \bar{x}_n + (\xi - \bar{x}_n)e^{-nt} + \int_0^t e^{-n(t-s)}\sigma_i(\theta_i(s))dW_s.$$

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

(3) *If $\sqrt{2}n > C_l$, there is a unique stationary distribution of (4) 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) The local Lipschitz condition for σ^2 can be used to establish that σ satisfies the condition of [7] Theorem 5.3.8. As $\sigma > 0$ we have for $\theta, \theta' \in \mathcal{K}_U$

$$\begin{aligned} |\sigma(\theta) - \sigma(\theta')|^2 &\le |\sigma(\theta) - \sigma(\theta')|\left(\sigma(\theta) + \sigma(\theta')\right) \\ &\le |\sigma^2(\theta) - \sigma^2(\theta')| \\ &\le \tilde{K}_U|\theta - \theta'| \end{aligned}$$

where $\tilde{K}_U$ is a finite constant. This gives pathwise uniqueness by [7] Theorem 5.3.8.

We recall [6] Theorem (4.53). We consider $\mathcal{N} := \{x \in \mathbb{R} : \sigma(x) = 0\}$ and observe that, if $\sigma(x) = 0$ for $x \notin \mathcal{K}$, and as f, g are locally integrable, we have the conditions for the theorem. Thus existence of a weak solution follows if $\mathcal{S} := \{x \in \mathbb{R} : \int_{x-}^{x+} \sigma^{-2}(y)dy = \infty\} \subset \mathcal{N}$. 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. Coupled with pathwise uniqueness we have the existence of strong solutions up until exit from $\mathcal{K}$.

The Assumption 3.1(4) is the definition for the boundaries of $\mathcal{K}$ to be natural, so the exit time is infinite almost surely, see [16] Section 5.5. Hence we have a pathwise unique strong solution for all time.

It is a simple exercise to establish the integral form.

(2) In order to show the moment bounds we first need to establish some crude estimates 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 $\kappa \geq 2$ using (5),

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

Applying the Burkholder-Davis-Gundy inequality, Hölder's inequality and the linear growth condition on σ_i we see that for $0 \leq t \leq T$ for a fixed $T > 0$,

$$\begin{aligned}
\phi_t^{M,\kappa} &\leq 2^{\kappa-1}|\theta_i(0) - \bar{x}_n|^\kappa + 2^{\kappa-1}E|\int_0^{t\wedge T_M} e^{ns}\sigma_i(\theta_i(s))dW_s|^\kappa \\
&\leq c_\kappa + 2^{\kappa-1}c'_\kappa E|\int_0^{t\wedge T_M} e^{2ns}\sigma_i(\theta_i(s))^2 ds|^{\kappa/2} \\
&\leq c_\kappa + C'_\kappa E T^{\kappa/2-1}\int_0^{t\wedge T_M} e^{\kappa ns}\sigma^\kappa(\theta(s))ds \\
&\leq c'_\kappa + c_\kappa T^{\kappa/2-1}E\int_0^t e^{\kappa n(s\wedge T_M)}(C'_\kappa + C_\kappa(\theta(s\wedge T_M) - \bar{x}_n)^\kappa)ds \\
&\leq c''_\kappa T^{\kappa/2-1}e^{n\kappa t} + c_\kappa T^{\kappa/2-1}\int_0^t \phi_s^{M,\kappa}ds.
\end{aligned}$$

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

$$\phi_t^{M,\kappa} \leq c''_\kappa T^{\kappa-2}\exp(c_\kappa T^{\kappa/2-1}t) \leq c''_\kappa T^{\kappa-2}\exp(c_\kappa T^{\kappa/2}).$$

As this bound is independent of M we can apply the dominated convergence theorem and let $M \to \infty$ to see that, by modifying the constants, for $\kappa \geq 2$

$$E|\theta(t)|^\kappa \leq c'_\kappa T^{\kappa-2}\exp(C_\kappa T^{\kappa/2}), \;\; 0 \leq t \leq T.$$

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

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

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

$$\begin{aligned} E(\int_0^{t\wedge T_M} \theta^{\kappa-1}\sigma(\theta)dW)^2 &\le E\int_0^{t\wedge T_M} \theta^{2\kappa-2}\sigma^2(\theta)ds \\ &\le E\int_0^{t\wedge T_M} C_l\theta^{2\kappa-2} + C_l\theta^{2\kappa}ds \\ &\le \int_0^t C_l E\theta^{2\kappa-2} + C_l E\theta^{2\kappa}ds \\ &\le c_\kappa T^{2\kappa-2}e^{cT^\kappa} \end{aligned}$$

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 $\phi_t^\kappa = E|\theta(t)|^\kappa$,

$$\phi_t^\kappa = \phi_0^\kappa + E\int_0^t (\kappa n(\bar{x}_n - \theta(s))\theta(s)^{\kappa-1} + \frac{1}{2}\kappa(\kappa-1)\theta(s)^{\kappa-2}\sigma^2(\theta(s)))ds. \tag{24}$$

We will proceed by induction noting that $\phi_t^0 = 1$ and $E\theta_t = \bar{x}_n + (\xi - \bar{x}_n)e^{-nt}$ and so that using $\kappa = 2$ and the linear growth bound we have

$$\phi_t^2 \le \phi_0^2 + (C_l - 2n)\int_0^t \phi_s^2 ds + \int_0^t (2n\bar{x}_n^2 + C_l + 2n\bar{x}_n(\xi - \bar{x}_n)e^{-ns})ds.$$

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

$$\frac{d\phi^2}{dt} \le (C_l - 2n)\phi^2 + (2n\bar{x}_n^2 + C_l + 2n\bar{x}_n(\xi - \bar{x}_n)e^{-nt}).$$

By taking $\psi_t = e^{(2n-C_l)t}\phi_t^2$, we have

$$\frac{d\psi}{dt} = e^{(2n-C_l)t}\left(\frac{d\phi^2}{dt} + (2n - C_l)\phi^2\right) \le e^{(2n-C_l)t}(2n\bar{x}_n^2 + C_l + 2n\bar{x}_n(\xi - \bar{x}_n)e^{-nt}).$$

Integrating gives

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

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

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

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

$$\phi_t^\kappa \leq |\xi|^\kappa + \int_0^t (\kappa n \bar{x}_n \phi_s^{\kappa-1} + (\frac{1}{2}\kappa(\kappa-1)C_l - \kappa n)\phi_s^\kappa + \frac{1}{2}\kappa(\kappa-1)C_l\phi_s^{\kappa-2} ds. \qquad (25)$$

Now assume that $\phi_t^p \leq C_p(1 \vee |\xi|^p \vee |\bar{x}_n|^p)$ for all $p \leq \kappa - 1$ and $t > 0$. Using this in (25) we get

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

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

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

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

(3) Under the condition that $\sqrt{2}C_l < 2n$ we can see that we have finite moment bounds independent of T. Thus there will exist a stationary distribution. There is a finite invariant measure when the speed measure is integrable and it is proportional to the speed measure. This gives the result. □

Remark A.2. We can also obtain the invariant measure by solving an ODE. The generator of the diffusion acting on a function $u \in$Dom$(\mathcal{A})$ is given by

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

The stationary distribution then has a density p_i which is the solution to

$$\mathcal{A}^* p_i = -\frac{d}{d\theta}(n(\bar{x}_n - \theta)p_i) + \frac{1}{2}\frac{d^2}{d\theta^2}(\sigma_i^2(\theta)p_i) = 0. \qquad (26)$$

We can check that the solution as given satisfies equation (26).

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.3. *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. In fact we show a much stronger result than required for our main theorem in that we also establish geometric ergodicity.

Theorem A.4. *Under Assumption 3.3 for each* $i, m \in \mathbb{N}$*, the Markov chain* $\{\theta_{i,m}(k) : k \geq 0\}$ *is ergodic with a unique stationary distribution* π_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 begin with the case $i = 1$.

In order to establish the positive recurrence we use the Lyapunov function technique. Let $V(x) = 1 + x^2$. From our earlier estimates and linear growth assumption we have

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

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

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

where

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

By our Assumption 3.3 the chain $\theta_{1,m}$ is ψ-irreducible and hence there exists at least one petite set. In fact by [18] Proposition 5.5.5, there exists a sequence $\{\mathcal{C}_j\}_j$ of petite sets such that $\cup_j \mathcal{C}_j = \mathcal{K}$. Hence our interval $\mathcal{C}$ is contained in some petite set $\mathcal{C}_{j^*}$. The geometric drift condition of [18] (V4) is satisfied with the petite set $\mathcal{C}_{j^*}$. Thus as the chain is also aperiodic we will have existence of a unique invariant measure and geometric convergence towards it.

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 [18] 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. For this we note that the Markov chain transitions are selected from the density determined by the limiting distribution of the diffusion arising in the case $i - 1$. We know that this has a density with respect to Lebesgue measure for all parameter choices and hence the chain is Lebesgue-irreducible and will be a T-chain as defined in [18] Chapter 6. In this case we know that every compact set is petite.

We keep the same Lyapunov function and, using Lemma 4.8, and the fact that $\sigma_i(x) \leq \sqrt{2}\sigma(x)$, we have a minor modification of the $i = 1$ case in that

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

where, for suitable positive constants,

$$\alpha_i = \frac{2n - c_1}{m} + O(\frac{1}{m^2}), \;\; \beta_i = \frac{c_2}{m} + O(\frac{1}{m^2}), \;\; \gamma_i = \frac{c_3}{m} + O(\frac{1}{m^2}).$$

Thus, incorporating the time change, and provided $2n > c_1$, we have that $\alpha_i, \beta_i, \gamma_i > 0$, independent of m (and i), 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

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

for a petite set $\mathcal{C}_i$ and we can proceed along exactly the same lines as the $i = 1$ case to deduce the result as there is no dependence on m. □

Corollary A.5. *The sequence of stationary distributions π_i^m for the sped up Markov chains $\theta_{i,m}$ converges to π_i, the stationary distribution for the solution θ_i to the SDE.*

Proof. Consider the Markov chain $\theta_{i,m}$ started according to π_i^m. By assumption this is a stationary Markov chain and we have for any finite collection of times $k_1, \ldots, k_j, l \in \mathbb{N}$ that

$$\mathbb{P}(\theta_{i,m}(k_1) \in A_1, \ldots, \theta_{i,m}(k_j) \in A_j) = \mathbb{P}(\theta_{i,m}(k_1 + l) \in A_1, \ldots, \theta_{i,m}(k_j + l) \in A_j). \quad (27)$$

Now we know that the sped up chains $\theta_{i,m}(t)$ converge weakly to a limit diffusion θ_i as $m \to \infty$. Hence, using this scaling, and the convergence of the finite dimensional distributions, if $mk_i \to t_1, \dots mk_j \to t_j, ml \to s$, then applying the convergence to both sides of (27), we have

$$\mathbb{P}(\theta_i(t_1) \in A_1, \dots, \theta_i(t_j) \in A_j) = \mathbb{P}(\theta_i(t_1 + s) \in A_1, \dots, \theta_i(t_j + s) \in A_j).$$

Hence the diffusion is also stationary. As the diffusion has a unique stationary measure we must have that the stationary distributions for the chain converge to it. □

References

[1] Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate bayesian computation in population genetics. *Genetics 162*, 2025–2035.

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

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

[4] Cox, D. R. and E. Snell (1968). A general definition of residuals. *J. Roy. Statist. Soc., Series B 30*, 248–275.

[5] Dawid, A. P., M. Stone, and J. V. Zidek (1973). Marginalization paradoxes in bayesian and structural inference. *J. Roy. Statist. Soc., Series B 35*, 189–233.

[6] 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.

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

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

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

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

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

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

[13] Heath, D. and W. Sudderth (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist. 6*, 333–345.

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

[15] Jiang, W. and M. A. Tanner (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist. 36*, 2207–2231.

[16] Karatzas, I. and S. E. Shreve (1991). *Brownian motion and stochastic calculus, 2nd Edition.* Springer Verlag.

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

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

[19] Soubeyrand, S., F. Carpentier, N. Desassis, and J. Chadoeuf (2009). Inference with a contrast-based posterior distribution and application in spatial statistics. *Statist. Meth. 6*, 466–477.

[20] Veronese, P. and E. Melilli (2015). Fiducial and confidence distributions for real exponential families. *Scand. J. Statist. 42*, 471–484.