From point estimation to Bayesian inference via dynamical systems

Gavin J. Gibson\textsuperscript{1} & Ben Hambly\textsuperscript{2}

February 7, 2014

\textsuperscript{1}School of Mathematical and Computer Sciences, The Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK

\textsuperscript{2}Mathematical Institute, University of Oxford, 24-29 St Giles, Oxford, OX1 3LB, UK

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

We elicit fundamental connections between point estimation and Bayesian, or Bayesian-like systems of inference using only principles which state that the class of acceptable systems for statistical inference should be closed under a certain data-augmentation process, and that the class be complete in a natural sense. Using these ideas we define refinement operators and investigate the systems of inference that correspond to their fixed points. In particular we characterise, in the 1-dimensional setting, fixed points obtained from systems of moment estimators, showing that these are Bayesian when the model lies in the exponential family and that the corresponding prior is Hartigan’s maximum-likelihood prior. 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 the maximum likelihood prior as the fixed point. A counter-example is given to show that, for non-regular models, a Bayesian fixed point does not generally arise from maximum-likelihood estimation. The results offer a new perspective on the relationship between Bayesian inference and classical point estimation whereby the former is constructed from the latter without direct reference to the Bayesian paradigm at the outset.
1 Introduction

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

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

where $p_{x_n}(\theta)$ is a density representing the belief about parameter $\theta$ given the outcome $x_n$. Systems of point estimators are obtained on setting $p_{x_n}(\theta) = \delta(\theta - \hat{\theta}(x_n))$. Bayesian systems have the property that $p_{x_n}(\theta) \propto \pi(\theta)\nu_\theta(x_n)$ for some prior density $\pi(\theta)$. For a given model, we denote by $\mathcal{X}$ the collection of all systems of inference. A system of inferences is essentially the same as the concept of an inversion as defined in [7].

This paper explores a novel, dynamical-systems approach to investigating the structure of $\mathcal{X}$ and comparing its constituent systems of inferences. Specifically we use a generalised data-augmentation principle introduced in [5], in order to define mappings, $\Psi$ and $\Phi$, called refinement operators in Section 2, from $\mathcal{X}$ to itself, which maps a given system of inferences to one which is preferable in a sense which we make explicit in Section 2. Interest then focuses on the fixed points of these operators. These have the property of being preferable to all systems in their domain of attraction and it may be argued that attention should be restricted to these fixed points when selecting appropriate statistical procedures. Moreover, a refinement operator induces a natural structure on $\mathcal{X}$ by partitioning it into the domains of attraction of the fixed points, with systems lying in distinct domains being mutually incomparable by our definition of preferability. This structure provides a means for exploring connections between approaches to inference and estimation. When the domain of attraction of a Bayesian fixed point contains a system of point estimators, then a correspondence between Bayesian and classical approaches is identified.

Connections between classical and Bayesian inference have been sought by identifying choices of prior distribution for $\theta$, so that the resulting posterior density satisfies certain classical criteria, at least asymptotically. Examples include reference priors (see [1, 2]), 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 ([9]) which attempts to assign equal prior probability to intervals of a given level of confidence. Of particular interest here is the maximum likelihood prior ([7, 8]). As shown in [8], this prior has the property that the Bayes estimator minimising expected KL divergence differs from the maximum likelihood estimator by order $n^{-3/2}$, in contrast to $n^{-1}$ for other choices of smooth prior. When $\theta$ is the canonical parameter in a distribution from the exponential family, the maximum-likelihood prior $\pi(\theta)$ for a model with density $\nu_\theta(y)$ satisfies

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

where

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

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

In the main result, Theorem 3.2, we characterise, for a broad class of 1-parameter models, those points fixed by \( \Psi \) whose domains of attraction contain a system of moment-based estimators. We show that the fixed points are Bayesian if and only if the model lies in the exponential family. In particular, a Bayesian analysis using the maximum-likelihood prior arises as the fixed point whose domain of attraction contains the system of maximum likelihood estimators, with other priors from the conjugate family arising in other cases. We also explore examples beyond the exponential family when the fixed point arising from a system of moment estimators is not Bayesian and show empirically that in some cases it may nevertheless approximate to a Bayesian analysis of an experiment in which only the sample mean is observed. In Section 5 we explore the generalisations of the main theorem to fixed points of \( \Phi \) 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.

Our analyses elicit connections that may be motivated in other ways. Nevertheless, we note the following distinctive features of the paper. Central to the approach is a concept of preferability of one inference over another with systems corresponding to fixed points being preferable to the initial system from which they are generated. Thus our approach facilitates critical comparisons between approaches to inference. Alternative approaches to identifying correspondences between classical and Bayesian analyses, such as the maximum-likelihood prior ([8]), may seek a Bayesian analysis from the outset and identify the prior according to some criterion for matching to the classical analysis. The approach of this paper does not explicitly seek to construct a Bayesian system of inferences. When one does arise then its Bayesian nature is a by-product of the construction, and the correspondence elicited between classical and Bayesian is arguably more fundamental.

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 observer. In what follows we do not attempt, therefore, to define validity in absolute terms. When the term valid is used, this should be interpreted as valid in the opinion of a given observer. Rather, we describe an approach for assessing the relative validity of procedures so that the preferability of one procedure over another can be assessed.

In [5] a generalised data augmentation principle was proposed and used to construct, or refine, inferences in the form of posterior-like summaries of belief. This asserts that the set of valid inferences for \( \theta \) should be closed under a data augmentation operation as described by Principle 2.1.
Augmentation Principle 2.1. Let $x_1$ and $x_2$ represent outcomes of two possible experiments on a system, parameterised by $\theta$, such that outcome $x_1 = f(x_2)$ for some function $f$ (i.e. the outcome of experiment $x_2$ completely specifies $x_1$). Suppose that $p_{x_1}(\theta)$ and $p_{x_2}(\theta)$ are inferences for $\theta$ that are considered to be valid by an observer given $x_1$ and $x_2$ respectively. Then the observer should also view $p_{x_1}^{(1)}(\theta)$ as valid, where

$$p_{x_1}^{(1)}(\theta) = \int \int p_{x_1}(\theta') \nu_\theta(x_2|x_1) p_{x_2}(\theta) dx_2 d\theta' ,$$

and $\nu_\theta(x_2|x_1)$ denotes the conditional density of $x_2$ given $x_1$ for the model with parameter $\theta'$.

We can use Principle 2.1 to define the concept of preferability. Accepting Principle 2.1, we can say if an observer views the pair of inferences $\{p_{x_1}, p_{x_2}\}$ as valid then they should also accept the validity of $\{p_{x_1}^{(1)}, p_{x_2}\}$, but not vice versa. Therefore we say that the latter pair of inferences is preferable to the former, in the sense that it is at least as assuredly valid as the former, in the opinion of any observer. Since the preferability of one system over another would be agreed by all observers that accept Principle 2.1, even those who view both systems as fatuous, then preferability as defined above is a property that can be determined objectively.

Now $p_{x_1}^{(1)} = p_{x_1} P$, where $P$ is the transition kernel of a Markov chain, called the generalised data augmentation chain, on the parameter space, in which updates to the current state $\theta(\cdot)$ are generated by first drawing $x_2 \sim \nu_{\theta(\cdot)}(x_2|x_1)$ and then drawing $\theta(\cdot+1) \sim p_{x_2}$. If this Markov chain is ergodic with stationary density, $\psi_{x_1}$, then $\psi_{x_1}$ should be a valid inference, so long as we allow the class of valid inferences to be complete. This motivates an additional principle from [5].

Completeness Principle 2.2. Suppose that $\{p_x^{(i)}(\theta) \mid i = 1, 2 \ldots\}$ denotes a sequence of valid inferences for $\theta$ given $x$ that tend in distribution to a distribution with density $\psi_x$ as $i \to \infty$. Then $\psi_x$ is also a valid inference. If, for some inference $p_x$ and every $i$, $p_x^{(i)}$ is preferable to $p_x$, then $\psi_x$ is preferable to $p_x$.

Accepting Principle 2.2, it follows that $\lim_{k \to \infty} p_{x_1} P^k$, when the limit exists, is preferable to $p_{x_1}$.

We use this idea to refine and compare systems of inferences. We say that $\Theta_1 = \{p_{x_n}(\theta)\mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n\}$ is preferable to $\Theta_2 = \{q_{x_n}(\theta)\mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n\}$ if $p_{x_n}(\theta)$ is preferable to $q_{x_n}(\theta)$, for all $n \in \mathbb{N}, x_n \in \mathcal{Y}^n$.

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

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

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

First denote by $P^{(0,m)}(x_n)$ the transition kernel of the generalised data augmentation Markov chain that arises when the observation $x_n$ is augmented by the next $m$ samples from the distribution. For this chain the state $\theta(\cdot)$ is updated by drawing $\theta(\cdot+m) \sim p_{x_n+m,0}$ where $x_n+m \sim \nu_{\theta(\cdot)}(x_{n+m}|x_n)$. We can now construct $\Theta_{i+1} = \{p_{x_n, i+1}(\theta)\mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n\}$ from $\Theta_i = \{p_{x_n,i}(\theta)\mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n\}$ recursively by setting

$$p_{x_n, i+1} = \lim_{m \to \infty} \lim_{k \to \infty} p_{x_n,i} [P^{(i,m)}(x_n)]^k, \quad (2)$$
where the convergence is in distribution. Suppose now that \( \lim_{i \to \infty} \Theta_i = \Theta_{\infty} \). Then \( \Theta_{\infty} \) is preferable to \( \Theta_i \), for all \( i \), is invariant under \( \Psi \), and is, in a natural sense, maximally preferable.

Denote by \( \mathcal{C} \subset \mathcal{X} \) the collection of those systems of inference \( \Theta_0 \) for which the limiting system \( \Theta_{\infty} \) exists, and denote by \( \mathcal{C}_F \subset \mathcal{C} \) the corresponding set of fixed points.

It is clear that any Bayesian system \( \Theta \), for which

\[
p_{x_n}(\theta) \propto \pi(\theta) \nu(\theta) \]

for some prior density \( \pi(\theta) \), lies in \( \mathcal{C}_F \). In this case the generalised data augmentation chain with transition kernel \( P(x_n, p_{x_{n+1}}) \) is a Gibbs sampler and the density \( p_{x_n} \) is invariant under it for any \( m > 0 \) and, hence under \( \Psi \). In general \( \mathcal{C}_F \) may contain non-Bayesian systems; therefore the property of invariance under \( \Psi \) may be seen as a weak form of coherence.

When the basin of attraction of a system \( \Theta \in \mathcal{C}_F \) fixed by \( \Psi \) contains a system of point estimators \( \Theta_0 \), a correspondence between classical and non-classical approaches follows. When \( \Theta \in \mathcal{C}_F \) is Bayesian, then a natural link is made between classical and Bayesian approaches. In the following section, we characterise for a general class of models the elements of \( \mathcal{C}_F \) whose basins include systems of moment-based point estimators. In particular, we will show that for the case of the 1-dimensional exponential family with the mean-value parameterisation, the maximum-likelihood prior of [8] can be obtained via this correspondence.

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

\[
p_{x_{n+1}} = \lim_{k \to \infty} \lim_{m \to \infty} p_{x_n} \left[ P(i, m) (x_n) \right]^{mk}.
\]

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 \( \Psi \) or \( \Phi \) is used; this is the case for the class of models considered in Theorem 3.2. At points in the paper, it will be convenient to work with the operator \( \Phi \) defined by (3).

### 3 Moment-based estimators and fixed points

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

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

1. The function \( \sigma^2(\theta) \) is locally Lipschitz in that for each \( U \in \mathbb{R}_+ \) there exists a constant \( K_U \) such that

\[
|\sigma^2(\theta) - \sigma^2(\theta')| \leq K_U |\theta - \theta'|, \quad \forall |\theta|, |\theta'| \leq U.
\]

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

\[
\sigma^2(\theta) \leq C_l (1 + \theta^2).
\]
3. There exists an $\epsilon > 0$ such that
\[
\int (x - \theta)^{2+\epsilon} \nu_\theta(x) \, dx < \infty.
\]

We now investigate those $\Theta \in \mathcal{Z}_F$ whose basins of attraction contain moment-based point estimators. The next result generalises [5], Example 2.3, which considered the special case of the Normal distribution.

**Theorem 3.2.** Suppose that $\nu_\theta$ satisfies Assumption 3.1 and let
\[
\Theta_0 = \{ p_{x_n,0}(\theta) = \delta(\theta - \bar{x}_n) \mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n \}.
\]

For $i = 1, 2, 3, \ldots$, let $c_i = 2 - 2^{-(i-1)}$. Then the systems $\Theta_i$, $i = 1, 2, 3, \ldots$ exist and are given by
\[
\Theta_i = \{ p_{x_n,i}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\{ \frac{2}{c_i} n f(\theta) \bar{x}_n - g(\theta)) \} \mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n \}
\]
where $f(\theta) = \int \sigma^{-2}(\theta) d\theta$ and $g(\theta) = \int \theta \sigma^{-2}(\theta) d\theta$. Moreover, the limiting system $\Theta_\infty$ is specified by
\[
\Theta_\infty = \{ p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\{ n f(\theta) \bar{x}_n - g(\theta)) \} \mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n \}
\]

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 diffusions whose stationary measures can be identified.

We now consider the conditions for $\Theta_\infty$ to be a Bayesian system, and the nature of the corresponding prior.

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

**Proof.** Clearly $\Theta_\infty$ is Bayesian only if the likelihood $\nu_\theta(x_n)$ satisfies
\[
\nu_\theta(x_n) = K_1(x_n)K_2(\theta) \exp\{ n f(\theta) \bar{x}_n - g(\theta)) \}
\]
identifying it as a member of the 1-parameter exponential family with mean value $\theta$ and sufficient statistic $\bar{x}_n$.

Conversely, suppose that $\nu_\theta(x)$ is a density from the 1-parameter exponential family with sufficient statistic $x$, mean $\theta$ and canonical parameter $a(\theta)$, then
\[
\nu_\theta(x) = K(x) \exp\{ a(\theta)x - c(\theta) \}
\]
From the score function $a'(\theta)x - c'(\theta)$ we obtain the information function $i(\theta) = \sigma^{-2}(\theta) = a'(\theta)$ implying that $a(\theta) = \int \sigma^{-2} d\theta$ and $c'(\theta) = a'(\theta)\theta$ in which case $c(\theta) = \int \theta \sigma^{-2}(\theta) d\theta$. It follows that $p_{x_n,\infty}(\theta) \propto \frac{1}{\sigma^2(\theta)} \exp\{ n(a(\theta)\bar{x}_n - c(\theta)) \}$; hence $\Theta_\infty$ 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 [8].

□
For the 1-parameter exponential family with mean-value parameterisation and sufficient statistic $\bar{x}_n$, Bayesian analyses with alternative priors from the conjugate family are obtained by specifying $\Theta_0$ appropriately in the construction. Given prior experience of a sample of size $k$ with mean value $a$, then a natural system of point estimators is

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

In this case $\Theta_\infty$ corresponds to a Bayesian analysis using the the prior

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

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

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

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

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

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

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

**Example 3.4.** The double exponential distribution has density given by

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

with mean given by $\theta$ and constant variance 2. In this case, Theorem 3.2 states that when

$$p_{x_n,0}(\theta) = \delta(\theta - \bar{x}_n)$$

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

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

**Example 3.5.** The Uniform(0, 2$\theta$) distribution has mean $\theta$ and variance $\sigma^2(\theta) = \frac{\theta^2}{3}$, and satisfies 3.1. In this case, Theorem 3.2 states that when

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

so that $p_{x_n,\infty}(\theta) \sim \text{IGamma}(3n + 1, 3n\bar{x}_n)$. We compare the density $p_{x_n,\infty}(\theta)$ with the Bayesian posterior density $\pi(\theta|\bar{x}_n)$ for the prior $\pi(\theta) \propto \sigma^{-2}(\theta) \propto \theta^{-2}.$
The likelihood $L(\theta; \bar{x}_n)$ is not convenient to work with directly being proportional to the $(n - 1)$-dimensional volume $V(A)$ of the set

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

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

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

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

4 Proof of Theorem 3.2

We retain the notation of earlier sections and consider the family of probability measures with density $\nu_0$ where the parameter space is a subset $\mathcal{K}$ of $\mathbb{R}$. Recall that $\theta$ is the distribution mean and $\sigma^2(\theta)$ the variance. Under the conditions of Assumption 3.1, we will show, using an induction argument, that the construction of Theorem 3.2 indeed converges to the system $\Theta_{\infty}$ given in the statement of the theorem. In Section 4.1 we consider the first step whereby $\Theta_1$ is constructed from $\Theta_0$, before considering the inductive step in Section 4.2. First we give some key auxiliary results required for the proof.
Suppose that we have already constructed the systems $\Theta_0, \ldots, \Theta_{i-1}$. Now fix $n$, and the observed sample $x_n$, and consider the Markov chain $\theta^{(k)}_{i,m}$ with transition kernel $P^{(i-1,m)}$ as described after Principle 2.2. We assume that, for all possible observed samples $x_n$, the sample mean $\bar{x}_n$ is constrained to lie in the allowable parameter space $K$. We first define a continuous-time process from $\theta^{(k)}_{i,m}$ by interpolation, setting $\theta^{m}_{i,0}(t) = \theta^{(k)}_{i,m}([mt])$, noting that $\theta^{(k)}_{i,m}$ and $\theta^{m}_{i,0}(t)$ are identical so far as the existence and nature of the stationary distribution is concerned.

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

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

$$
\begin{align*}
\frac{d\theta_i}{dt} &= n(\bar{x}_n - \theta_i)dt + \sigma_i(\theta_i)dW^i, \\
\theta_i(0) &= \xi
\end{align*}
$$

(4)

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

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

This result in essence enables one to demonstrate that the second refinement operator $\Phi$, introduced in Section 2, has $\Theta_\infty$ in Theorem 3.2 as a fixed point. With a little more work we can deduce the following version of Theorem 3.2 to show that $\Theta_\infty$ is the fixed point of $\Psi$ as required by the Theorem.

Theorem 4.2. The Markov chains $\theta^{m}_{i,0}$ have stationary distributions $\pi^{m}_{i,0}$, the diffusion process $\theta_i$ has a stationary distribution $\pi_i$ and

$$\pi^{m}_{i,0} \to \pi_i, \text{ weakly as } m \to \infty.$$

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

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

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

(2) The moments of $\theta_i(t)$ are bounded up to a level depending on $n$ in that there exist constants $C_k$ such that $E|\theta_i(t)|^k \leq C_k(1 \vee |\bar{x}_n|^k \vee |\xi|^k)$ for all $t \geq 0$ and $1 \leq k \leq (2n + C_1)/C_1$.

(3) If $\sqrt{2n} > C_1$, the stationary distribution of (4) exists and is given by

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

where

$$f_i(\theta) = \int \sigma_i^{-2}(\theta)d\theta, \quad g_i(\theta) = \int \theta\sigma_i^{-2}(\theta)d\theta.$$
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 3.2 the limiting system of inferences has a density given by

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

where

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

A key tool in establishing Theorem 4.1 is Corollary 7.4.2 of [4], which specifies conditions sufficient for the existence of a diffusion approximation to a sequence of Markov chains. We state a version of the result suited to our purposes. We denote by \( \mathcal{P}(\mathbb{R}) \) the set of probability measures on \( \mathbb{R} \).

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

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

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

\[ \mu_m(x) = m\mathbb{E}(Y_m(1) - x) \]

\[ \sigma_m^2(x) = m\mathbb{E}(Y_m(1) - x)^2 \]

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

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

(6)

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

(7)

and

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

(8)

Then \( X_m \) converges weakly to \( X \).

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

4.1 The case \( i = 1 \).

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

\[ \theta_{i,m}^{(k+1)} = \frac{n\bar{x}_n + m\bar{Y}_m^{(k)}}{n + m}, \]
where $Y_{m}^{(k)} = \frac{1}{m} \sum_{j=1}^{m} Y_{j}^{(k)}$, with $Y_{j}^{(k)}$ samples from the measure with density $\nu_{\theta_{i,m}^{(k)}}$. To simplify the notation we suppress the subscript $i$ and write $\theta_{1,m}^{(k)}$ as $\theta_{m}^{(k)}$. We now establish the conditions of Theorem 4.5 with $\theta_{m}^{(k)}$ in place of $Y_{m}^{(k)}$ and with the limiting diffusion process given by (3).

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

$$\theta_{m}^{(1)} = x + \frac{n}{n+m} (\bar{x}_{n} - x) + \frac{R_{m}(x)}{n+m},$$

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

$$\mu_{m}(x) = m \mathbb{E}(\theta_{m}^{(1)} - x) = \frac{nm}{n+m} (\bar{x}_{n} - x),$$

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

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

which establishes (6).

Also we have

$$\sigma_{m}^{2}(x) = m \mathbb{E}(\theta_{m}^{(1)} - x)^{2}$$

$$= m \mathbb{E} \left( \frac{n}{n+m} (\bar{x}_{n} - x) - \frac{1}{n+m} R_{m}(x) \right)^{2}$$

$$= m \left( \frac{n}{n+m} \right)^{2} \left( \bar{x}_{n} - x \right)^{2} + \frac{1}{(n+m)^{2}} \mathbb{E} R_{m}(x)^{2}$$

$$= \frac{mn^{2}}{(n+m)^{2}} (\bar{x}_{n} - x)^{2} + \left( \frac{m}{n+m} \right)^{2} \sigma^{2}(x).$$

Thus we have that $\sigma_{m}^{2}(x) \to \sigma^{2}(x)$ as $m \to \infty$. We establish our condition as

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

To handle the tail condition we observe that by Assumption 3.1 there is an $\epsilon > 0$ such that $\mathbb{E}(Y_{1} - x)^{2+\epsilon} < \infty$. Letting $p = 2 + \epsilon$ we see that

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

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

$$\mathbb{E}|R_{m}(x)|^{p} = \mathbb{E} \left| \sum_{i=1}^{m} (Y_{i}^{(0)} - x) \right|^{p}$$

$$\leq c_{p} \mathbb{E} \left| \sum_{i=1}^{m} (Y_{i} - x) \right|^{p/2}$$

$$\leq c_{p} m^{p/2-1} \mathbb{E} \sum_{i=1}^{m} |Y_{i} - x|^{p}$$

$$= c_{p} m^{p/2} \mathbb{E}|Y_{1}^{(0)} - x|^{p} = c_{p} m^{p/2}. \quad (9)$$
Thus we have, by Markov’s inequality and (9),

\[
m\mathbb{P}(|\theta_m(1) - x| \geq \epsilon) \leq \sup_{|x| < r} \frac{m}{|x|} \mathbb{E}[|\theta(1) - x|^p] \\
\leq \sup_{|x| < r} \left( \frac{2^{p-1}mn^p}{(n+m)p} |\bar{x}_{n} - x|^p + \frac{2^{p-1}m^{p/2+1}C_p}{(n+m)^p} \right) \\
\leq C_1 m^{1-p} + C_2 m^{1-p/2},
\]

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

Thus we have satisfied the conditions of Theorem 4.5.

**Proposition 4.6.** Under Assumption 3.1 the process \( \theta^n_1 \) converges weakly to \( \theta_1 \), the unique solution to

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

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

From Lemma 4.2 (2) we establish the form of \( \Theta_1 \) as given in Theorem 3.2.

### 4.2 The inductive step

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

The draw \( \theta^{(k+1)}_{i+1,m} \sim p_{i,x_{n+m}}(\theta) \) is a sample from the stationary distribution \( p_{i,x_{n+m}}(\theta) \) of the diffusion \( \theta_i^{n,m,k} \), given by (3) with \( n + m \) and \( \bar{x}_{n+m} \) replacing \( n \) and \( \bar{x}_n \) respectively. An approximate sample \( \theta_i^{n,m,k}(\tau_m) \) can be obtained by running the diffusion, with initial value \( \theta_i^{n,m,k}(0) = \bar{x}_{n+m} \), for a sufficiently long time \( \tau_m \).

Thus we have an approximate Markov chain given by

\[
\theta^{(k+1)}_{i+1,m} = \theta^{(k)}_{i+1,m} + \frac{n}{n+m}(\bar{x}_n - \theta^{(k)}_{i+1,m}) + \frac{1}{m+n}R_m(k) + \int_0^{\tau_m} e^{-(n+m)(\tau_m-t)}\sigma_i(\theta^{(n,m,k)}_i(t))dW_t^k,
\]

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

For the one-step evolution of our Markov chain, from initial state \( x \), we have

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

where

\[
N_m(1)(\tau_m) = \int_0^{\tau_m} e^{-(n+m)(\tau_m-t)}\sigma_i(\theta^{(n,m,1)}_i(t))dW_t.
\]

We will write \( \theta_i^{n,m} \) for \( \theta_i^{n,m,1} \) and \( N_m(t) \) for \( N_m(1)(t) \). 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 \( 0 \leq t \leq \tau_m \). Abusing notation we continue to use \( \mathbb{E} \) both for expectation with respect to the probability measure governing the diffusion as well as that for the augmented sample. We note that the Brownian motion driving \( N_m(\tau_m) \) is independent of \( R_m(1) \) given
\( \theta^{(1)}_{i+1,m} \) so that we can treat the term in \( N_m \) 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 \( E N_m(t) = 0 \) and

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

(11)

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

\[
d\bar{\theta}_i = n(\bar{x}_n - \bar{\theta}_i)\, dt + \sigma(\bar{\theta}_i)\, dW.
\]

with \( \bar{\theta}_i(0) = \xi \in K \).

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

\[
\mu_{m,i}(x) = mE(\theta^{(1)}_{i+1,m} - x) = \frac{mn}{m+n}(\bar{x}_n - x).
\]

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,

\[
\sigma_{m,i+1}^2(x) = mE\left( \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{m+n} R_m(1) + N_m(\tau_m) \right)^2
\]

\[
= \frac{mn^2}{(n+m)^2}(\bar{x}_n - x)^2 + \frac{m}{(n+m)^2} E R_m^2(1) + mE N_m^2(\tau_m).
\]

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

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

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

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

\[
|\mathbb{E} m N_m^2(\tau_m) - \frac{\sigma_i^2(x)}{2}|
\]

\[
= \frac{m}{2(m+n)} \left( \mathbb{E} \int_0^{\tau_m} 2(n+m)e^{-2(n+m)(\tau_m-t)} \sigma_i^2(\theta_i^{n,m}(t))\, dt - \frac{m+n}{m} \sigma_i^2(x) \right)
\]

\[
\leq \frac{m}{2(n+m)} \int_0^{\tau_m} 2(n+m)e^{-2(n+m)(\tau_m-t)} E \left| \sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x) \right|\, dt
\]

\[
+ \frac{n}{2(m+n)} \sigma_i^2(x) + E e^{-2(n+m)\tau_m} \sigma_i^2(x).
\]

(12)

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

\[
E|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)| \leq K_U E (|\theta_i^{n,m}(t) - x|; |\theta_i^{n,m}(t)| < U)
\]

\[
+ E (|\sigma_i^2(\theta_i^{n,m}(t)) - \sigma_i^2(x)|; |\theta_i^{n,m}(t)| \geq U).
\]

(13)
We can estimate the first term on the right hand side using

\[ \theta_{i}^{n,m}(t) - x = \frac{n}{n + m}(\bar{x}_n - x) + \frac{1}{m + n}R_m(1) + N_m(t). \]

Taking expectations we have

\[
\mathbb{E}[\theta_{i}^{n,m}(t) - x] \leq \left| \frac{n}{n + m}(\bar{x}_n - x) \right| + \mathbb{E}\left[ \frac{1}{m + n}|R_m(1)| + \mathbb{E}|N_m(t)| \right] \\
\leq O\left(\frac{1}{m}\right) + \frac{1}{m + n}(\mathbb{E}R_m(1)^2)^{1/2} + (\mathbb{E}|N_m(t)|^2)^{1/2}.
\]

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

\[
\mathbb{E}N_m(t)^2 \leq \int_0^t e^{-2(m+n)(t-s)}\mathbb{E}(1 + \mathbb{E}(\theta_{i}^{n,m}(s))^2)ds \\
\leq \int_0^t e^{-2(m+n)(t-s)}\mathbb{E}C_1(1 + C_2(1 \vee \bar{x}_{n+m}^2))ds \\
\leq \frac{C_1(1 + C_2(1 \vee \bar{x}_{n+m}^2))}{m + n} = O(1/m).
\]

From this, and the expression for \( \mathbb{E}R_m(1)^2 \), we know that there is a further constant \( c \) such that

\[ \mathbb{E}[\theta_{i}^{n,m}(t) - x] \leq \frac{c}{\sqrt{m}}, \forall t \leq T. \]

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

\[
\mathbb{E} \left[ (\sigma_i^2(\theta_{i}^{n,m}(t)) - \sigma_i^2(x)) | |\theta_{i}^{n,m}(t)| \geq U \right] \leq (C_1 + \sigma_i^2(x))\mathbb{P}(|\theta_{i}^{n,m}(t)| \geq U) \\
\leq (C_1 + \sigma_i^2(x))U^{-p}\mathbb{E}[\theta_{i}^{n,m}(t)]^p + C_iU^{2-p}\mathbb{E}[\theta_{i}^{n,m}(t)]^p
\]

From the estimates for the diffusion started from \( \bar{x}_{n+m} \) in Lemma 4.3(2) and following (9) we have \( \mathbb{E}(\theta_{i}^{n,m}(t))^p \leq c_p\mathbb{E}(1 \vee |\bar{x}_{n+m}|)^p \leq c_p^{*}m^{-p/2} \) and hence

\[ \mathbb{E} \left[ (\sigma_i^2(\theta_{i}^{n,m}(t)) - \sigma_i^2(x)) | |\theta_{i}^{n,m}(t)| \geq U \right] = O(m^{-p/2}). \]

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

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

and we have the result.

To show the last condition of Theorem 4.5, as in the \( i = 1 \) case, we need a little more than second moments. For this we note that by Burkholder-Davis-Gundy and Hölder’s
inequality
\[
\mathbb{E} N_m(\tau_m)^p \leq c_p \mathbb{E} (N_m)^{p/2} \\
= c_p \mathbb{E} \left( \int_0^{\tau_m} e^{-2(n+m)(\tau_m-s)} \sigma^2(s) ds \right)^{p/2} \\
\leq c_p \left( \int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/(p-2)} ds \right)^{p/2-1} \left( \int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2} E\sigma^p_i(\theta_{i,n}^m(s)) ds \right) \\
= c_p \left( \frac{p-2}{p(n+m)} \right)^{p/2-1} \left( 1 - e^{-p(n+m)\tau_m/(p-2)} \right) \left( \int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2} E\sigma^p_i(\theta_{i,n}^m(s)) ds \right) \\
\leq c_p \left( \frac{p-2}{p(n+m)} \right)^{p/2-1} \int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2} E\sigma^p_i(\theta_{i,n}^m(s)) ds
\]

As we have a linear growth condition for \( \sigma \) and, by Assumption 3.1(3), the moments of the process \( \theta_{i,n}^m \) exist, at least up to \( p = 2 + \epsilon \), we have \( E\sigma^p_i(\theta_{i,n}^m(s)) \leq c_p E(1 \vee |\bar{x}_{n+m}|^p) \leq c_p' \) for all \( s \) and \( m \) using (9). Thus
\[
\int_0^{\tau_m} e^{-p(n+m)(\tau_m-s)/2} E\sigma^p_i(\theta_{i,n}^m(s)) ds \leq c_p' \frac{2}{p(m+n)} \left( 1 - e^{-p(n+m)\tau_m/2} \right),
\]
and hence for a constant \( C \) we have
\[
\mathbb{E} N_m(\tau_m)^p \leq \frac{C}{(m+n)^{p/2}}.
\]

Thus, by Markov’s inequality, we have
\[
\sup_{|x|<r} m \mathbb{P}(|\theta_i^{(1)} - x| > \epsilon) \leq \sup_{|x|<r} \frac{m \mathbb{E} |\theta_i^{(1)} - x|^p}{\epsilon^p} \\
= \sup_{|x|<r} \frac{m \mathbb{E} \left( \frac{n}{n+m} (\bar{x}_n - x) + \frac{1}{m+n} R_m(1) + N_m(\tau_m) \right)^p}{\epsilon^p} \\
\leq \sup_{|x|<r} \frac{c_p m \left( \frac{n^p}{(n+m)^p} (\bar{x}_n - x)^p + \frac{1}{(m+n)^p} \mathbb{E} R_m(1)^p + \mathbb{E} N_m(\tau_m)^p \right)}{\epsilon^p} \\
\leq \frac{C_1}{\epsilon^p m^{p-1}} + \frac{C_2}{\epsilon^p m^{p/2-1}} + \frac{C_3}{\epsilon^p m^{p/2-1}}.
\]

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

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

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

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

5.1 The regular case

We consider the application of refinement operators to systems of maximum-likelihood rather than moment estimators, noting the coincidence of the two in the exponential-family setting. In particular we ask whether the construction leads to a Bayesian analysis beyond the exponential-family case and, if so, whether the maximum-likelihood prior is recovered. We find it convenient to work with the operator $\Phi$ rather than $\Psi$ 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 $\Psi$ as was done via Theorem 4.2 and Lemma 4.3 when proving Theorem 3.2.

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

Retaining the notation from earlier sections, consider the model $\nu_0(y)$ and let $l(\theta, y) = \log \nu_0(y)$, $i(\theta) = \mathbb{E}_Y \left( -\frac{\partial^2 l}{\partial \theta^2} \right)$, $a(\theta) = \mathbb{E}_Y \left( \frac{\partial^2 l}{\partial \theta \partial \theta} \right)$, and $c(\theta) = \mathbb{E}_Y \left( -\frac{\partial l}{\partial \theta} \right)$. Suppose that $\nu_0$ satisfies regularity conditions that allow change of order of integration with respect to $y$ and differentiation with respect to $\theta$. We can then easily verify the identity

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

Now let

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

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 $\Theta_1$ from $\Theta_0$, where $m$ is large. Denote the log-likelihood function for the additional $m$ samples, generated when updating $\theta_{1,m}^{(k)}$, by $l_m(\theta)$ which is maximised by $\hat{\theta}_m$. As in the proof of Theorem 3.2 we are interested in the case where $\theta_{1,m}^{(\infty)}(t) = \theta_{1,m}^{(int)}$ 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. [3]), 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

$$d\theta_1 = \frac{1}{i(\theta_1)} \left( a(\theta_1) + \frac{c(\theta_1)}{2} + l_n(\theta_1)i(\theta_1) \right) dt + \sqrt{\frac{1}{i(\theta_1)}} dB. \quad (15)$$
We assume the following conditions hold.

**Assumption 5.1.**

1. The stochastic differential equation
   \[ d\theta_1 = \frac{1}{i^2(\theta_1)} \left( a(\theta_1) + \frac{c(\theta_1)}{2} + l_n'(\theta_1)i(\theta_1) \right) dt + \frac{1}{i(\theta_1)} dB, \]
   where \( B \) is a Brownian motion and \( \theta_1(0) = \xi \in \mathcal{K} \), has a unique solution.

2. The Markov chain \( \theta_{1,m}^{(k)} \) satisfies
   \[ E(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)}) = \frac{1}{i^2(\theta_{1,m}^{(k)})m} \left( a(\theta_{1,m}^{(k)}) + \frac{c(\theta_{1,m}^{(k)})}{2} + l_n'(\theta_{1,m}^{(k)})i(\theta_{1,m}^{(k)}) \right) + o(1/m) \]
   \[ E(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)})^2 = \frac{1}{i(\theta_{1,m}^{(k)})m} + o(1/m) \]
   \[ E(\theta_{1,m}^{(k+1)} - \theta_{1,m}^{(k)})^{2+\epsilon} \leq C \frac{1}{m^{1+\epsilon}} \]

**Theorem 5.2.**

1. Under Assumption 5.1, the interpolated chain \( \theta_{1,m} \) converges weakly to \( \theta_1 \).

2. The associated system of inferences, obtained from the stationary measure of the diffusion in (15), is given by
   \[ \Theta_1 = \{ p_{x_n,1}(\theta) \propto \pi(\theta)L_n^{2}(\theta; x_n) \mid n \in \mathbb{N}, x_n \in \mathcal{Y}_n \} \]
   where \( \partial \frac{\partial}{\partial \theta} \log \pi = \frac{a(\theta)}{i(\theta)} \), so that \( \pi(\theta) \) is the maximum-likelihood prior.
We proceed to the inductive step. Suppose now that
\[ \Theta_{j-1} = \{ p_{m,j}(\theta) \propto \pi(\theta) L_{m}^{\gamma_{j-1}}(\theta; x_n) \mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n \}, \]
where \( \gamma_i = 1/(1 - 2^{-i}), i = 1, 2, \ldots \), and consider the construction of \( \Theta_j \) from \( \Theta_{j-1} \). Let, \( x_n, L_n(\theta) \) and \( l_n(\theta) \) be as above and let \( \pi_L = \pi(\theta) L_n^{\gamma_{j-1}}(\theta) \).

Consider the chain \( \theta_{j,m}^{(k)} \) and the continuous-time interpolation \( \theta_{j,m}^{(m)}(t) = \theta_{j,m}^{([mt])} \). 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 \( \theta \) 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 \( \theta \) about the MLE, \( \hat{\theta}_m \) for regular models.

From Chapter 5 of [6] it follows that, using prior density \( \pi_L(\theta) \), and given observations \( y_1, \ldots, y_m \), then, the posterior \( \pi(\theta|y_1, \ldots, y_m) \propto \pi_L(\theta)L_m(\theta) \) satisfies
\[ \mathbb{E}(\theta - \hat{\theta}_m|y_1, \ldots, y_m) = \left[ -\frac{\partial^2 \log \pi_L(\theta)}{\partial \theta^2}\right]_{\hat{\theta}_m}^{\hat{\theta}_m} \left( \frac{1}{2} \left[ \frac{\partial \log \pi_L(\theta)}{\partial \theta}\right]_{\hat{\theta}_m} \right) + O(1/m) \]
and has variance \( \left[ -\frac{\partial^2 \log \pi_L(\theta)}{\partial \theta^2}\right]_{\hat{\theta}_m}^{-1} + O(1/m) \). 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
\[ \mathbb{E}(\theta - \hat{\theta}_m|y_1, \ldots, y_m) = \frac{1}{\gamma_{j-1}} \left[ -\frac{\partial^2 \log \pi_L(\theta)}{\partial \theta^2}\right]_{\hat{\theta}_m}^{\hat{\theta}_m} \left( \frac{1}{2} \left[ \frac{\partial \log \pi_L(\theta)}{\partial \theta}\right]_{\hat{\theta}_m} \right) + O(1/m) \]
with approximate variance is \( \frac{1}{\gamma_{j-1}} \left[ -\frac{\partial^2 \log \pi_L(\theta)}{\partial \theta^2}\right]_{\hat{\theta}_m}^{-1} \).

We discern two components to the increment \( \theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)} \): one given by \( \hat{\theta}_m - \theta_{j,m}^{(k)} \) and the other arising when we sample \( \theta_{j,m}^{(k+1)} \) from a density proportional to \( \pi_L(\theta)L_m(\theta)^{\gamma_{j-1}} \). We approximate expectations over \( y_1, \ldots, y_m \) in (17) by setting the derivatives of \( l_m \) to their expected values at \( \theta = \theta_{j,m}^{(k)} \), and combine the two increments to show that
\[ \mathbb{E}(\theta_{j,m}^{(k+1)} - \theta_{j,m}^{(k)}) = \frac{1}{i^2(\theta_{j,m}^{(k)})} \left( a(\theta_{j,m}^{(k)}) + c(\theta_{j,m}^{(k)}) \frac{\gamma_{j-1} + 1}{2\gamma_{j-1}} + i(\theta_{j,m}^{(k)}) \frac{\partial \log \pi_L(\theta)}{\partial \theta}\right) \frac{1}{m} + O(1/m) \]
and
\[ \text{Var}(\theta_{j,m}^{(k+1)}|\theta_{j,m}^{(k)}) = \frac{\gamma_{j-1} + 1}{i(\theta_{j,m}^{(k)})\gamma_{j-1}} \frac{1}{m} + O(1/m). \]
We can now discern the candidate for the limiting diffusion to be
\[ d\theta_j = \frac{1}{i^2(\theta_j)} \left( a(\theta_j) + c(\theta_j) \frac{\gamma_{j-1} + 1}{2\gamma_{j-1}} + i(\theta_j) \frac{\partial \log \pi_L(\theta)}{\partial \theta_j}\right) dt + \sqrt{\frac{\gamma_{j-1} + 1}{\gamma_{j-1}i(\theta_j)}} dB. \]
We make the following assumptions for \( j \geq 2 \).

**Assumption 5.3.**
1. The stochastic differential equation

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

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

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

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

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

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

and the limiting system is given by

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

**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

\[
\begin{align*}
\frac{\partial \log p}{\partial \theta} &= \frac{i'(\theta)}{i(\theta)} + \frac{2a(\theta)\gamma_{j-1}}{(\gamma_{j-1} + 1)i(\theta)} + \frac{c(\theta)}{\gamma_{j-1} + 1} + \frac{2}{i(\theta)} \frac{\partial \log \pi_L}{\partial \theta} \\
&= \frac{i'(\theta)}{i(\theta)} + \frac{2a(\theta)\gamma_{j-1}}{(\gamma_{j-1} + 1)i(\theta)} + \frac{c(\theta)}{\gamma_{j-1} + 1} + \frac{2}{i(\theta)} \left( a(\theta) + \gamma_{j-1} |\theta| \right) \\
&= \frac{a(\theta)}{i(\theta)} + \frac{2\gamma_{j-1} |\theta|}{\gamma_{j-1} + 1}
\end{align*}
\]

by (14). It follows that

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

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

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

\[ \square \]

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

\[ \Theta_0 = \{p_{x_n,0}(\theta) = \delta(\theta - \hat{\theta}(x_n)) - b(\hat{\theta}(x_n)) \mid n \in \mathbb{N}, x_n \in \mathcal{Y}^n\}. \]
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 the diffusions such as (15), leading to a limiting system in which
\[ p_{x_n,\infty}(\theta) \propto \pi_b(\theta)L_n(\theta; x_n), \]
where $b(\theta)$ and $\pi(\theta)$ are related by
\[ \frac{\partial}{\partial \theta} \log \pi_b = \frac{a(\theta)}{i(\theta)} + i(\theta)b(\theta). \]

### 5.2 An irregular model

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

Consider again the uniform distribution of Example 3.5 reparameterised for convenience so that $\pi(y|\theta) = \theta - 1$, $0 < y < \theta$ and the system of (maximum likelihood) estimators
\[ \Theta_{MLE} = \{p_{x_n,i}(\theta) = \delta(\theta - x_n)\mid n \in \mathbb{N}\}. \]
Now $\Theta_{MLE}$ is trivially a fixed point of $\Psi$. Therefore consider a more general system of estimators of the form
\[ \Theta_a = \{p_{x_n,i}(\theta) = \delta(\theta - a_n x_n)\mid n \in \mathbb{N}\}, \]
with $a_n = \frac{n+1}{n}$ giving, for example, a system of unbiased MLE-based estimators. We now investigate whether a Bayesian limit arises when $\Psi$ is applied to $\Theta_a$ for suitably chosen $a = (a_1, a_2, \ldots)$. For simplicity we restrict attention to sequences $a$ 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
\[ \Theta_\infty = \lim_{k \to \infty} \Theta_a \Psi^k = \Theta_{MLE}. \]

(ii) Otherwise $\Theta_\infty$ does not exist.

**Proof.** We will make use of the following standard result on random walks from [10].

**Lemma 5.5.** Let $X_i, i = 1, 2, 3, \ldots$ 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
\[ \lim_{a \to 0} \Pr(a M^{(a)} > z) = e^{-2z}, \]
so that $aM^{(a)}$ converges weakly to an $\text{Exp}(2)$ distribution as $a \to 0$. 

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

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

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

\[
\lambda_m^{(k)} = \log a_{n+m} + \sup \{0, \lambda_m^{(k-1)} - \xi_k\} = \sup \{\log a_{n+m}, \lambda_m^{(k-1)} - \xi_k + \log a_{n+m}\}
\]

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

\[
\nu_m^{(k)} = \sup \{0, \nu_m^{(k-1)} + \zeta_m^{(k)}\}
\]

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

Now let

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

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

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

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

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

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

where \( M \sim \text{Exp}(2) \).

Now consider the large-\( m \) behaviour of

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

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

\[
m(1 - c_m) = \frac{m}{m+n} (n + m(1 - (m + n) \log a_{n+m}))
\]

we see that \( \lim_{m \to \infty} m(1 - c_m) = n + \lim_{m \to \infty} m(1 - m \log a_m) \geq n \). There are two cases to consider.
(a) If \( \lim_{m \to \infty} m(1 - m \log a_m) = \infty \), then \( \lambda_m \to 0 \) weakly.

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

\[
\lambda_m \to \text{Exp}(2(n + \mu)).
\]

In either case (1) or (2)(a) our system of inferences (for the log-transformed parameter) is

\[
\Lambda^{(1)} = \{p_{x_n,i}(\lambda) = \delta(\lambda)\}
\]

with corresponding system for \( \theta \), for the case of general \( x(n) \), given by

\[
\Theta^{(1)}_a = \{p_{x_n,i}(\theta) = \delta(\theta - x(n)) \mid n \in \mathbb{N}\} = \Theta_{\text{MLE}}.
\]

Thus a single application of \( \Psi \) maps \( \Theta_a \) to the fixed point \( \Theta_{\text{MLE}} \).

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

\[
\Theta^{(1)}_a = \{p_{x_n,1}(\theta) \propto \theta^{-2\mu} - 2\mu - 1\}.
\]

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

\[
\Theta^{(1)}_a = \{p_{x_n,1}(\theta) \propto \sigma(\theta)^{-2} L(\theta; x_n)^2\},
\]

as was the case for models in the exponential family.

However, any hopes of ultimate convergence to a Bayesian solution are dashed by further applications of \( \Psi \). We can show that for case (2)(b) \( \Psi \) maps \( \Theta^{(1)}_a \) to \( \Theta_{\text{MLE}} \). This can be most easily seen by applying \( \Phi \) to the system

\[
\Lambda^{(1)} = \{p_{x_n,1}(\lambda) \propto 2(n + \mu)e^{-2(n+\mu)(\lambda \log x(n))}, \lambda > x(n)\},
\]

For large \( m \), increments in the data augmentation chain are essentially composed of a decrease with distribution \( \text{Exp}(m) \) counterbalanced by an increase with distribution \( \text{Exp}(2(m + n + \mu)) \). When we scale by \( m \) in the manner of (19) the expectation and variance of the increments tend to \(-1/2\) and \(5/4\) respectively as \( m \to \infty \), resulting in a proper stationary distribution for the scaled chain and forcing \( \lambda_m \to 0 \) in distribution where \( \lambda_m \) follows the stationary distribution of the unscaled chain.

### 6 Discussion

In this paper we have shown how the generalised data augmentation principles introduced in [5] can be applied to elicit connections between classical point estimation and Bayesian (or Bayesian-like) inference where the latter is constructed from the former by repeated application of a refinement operator based on the principles. This contrasts with other approaches to defining connections, for example, by defining point estimators from Bayesian analyses using decision-theoretic ideas. A key notion in our treatment is that of preferability of one system of inferences over another and of fixed points of the refinement operators representing maximally preferable inferences. Our results show that for sufficiently regular models parameterised by their mean, the limiting systems of inferences is Bayesian when the model lies in the exponential family and otherwise takes the form of a pseudo-Bayesian analysis in which the true model likelihood is replaced by one with an exponential-family form. More generally, subsequent investigations suggest
that limiting systems derived from initial systems of maximum-likelihood estimators correspond to Bayesian inference using Hartigan’s maximum-likelihood prior specification, given sufficiently strong regularity conditions on the model.

The results of the paper do not lead to new statistical methodology but, rather, a fresh perspective on established approaches to inference. It is arguably surprising that, by applying principles that require only that the class of acceptable inferences be closed under a certain data augmentation operation and that it be complete in a natural sense, the Bayesian paradigm can be constructed from a starting point that considers only point estimators. The property that a system of inferences is invariant under $\Psi$ or $\Phi$, 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 $\Phi$ or $\Psi$, then these fixed points may nevertheless be strongly coherent Bayesian systems when their basin of attraction contains the system of maximum-likelihood point estimators. There are several natural extensions to the ideas of the paper that would be worthy of investigation, the most obvious of which is to generalise the results to models beyond the 1-dimensional case.

**Acknowledgements**

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

**A Appendix: Proofs of Lemma 4.3 and Theorem 4.2**

We first recall Lemma 4.3

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

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

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

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

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

where

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

**Proof.** (1) The local Lipschitz condition and the linear growth condition of Assumption 3.1 ensure the existence of a pathwise unique strong solution for all time for a one-dimensional SDE. It is a simple exercise to establish the integral form.
(2) In order to show this we first need to establish some crude moment bounds to ensure that the stochastic integral in the integral representation for \( \theta \) is a martingale. The stochastic integral is a local martingale and thus if we define the stopping times \( T_m := \inf\{ t : |\theta(t) - \bar{x}_n| > m \} \) we have for \( k \geq 2 \) using (5),

\[
\phi_t^{m,k} := E[e^{n(t/T_m)} (\theta(t \wedge T_m) - \bar{x}_n)|^k = E|(\theta_t(0) - \bar{x}_n) + \int_0^t e^{ns} \sigma_i(\theta_i(s))dW_i|^k.
\]

Applying the Burkholder-Davis-Gundy inequality, Hölder’s inequality and the linear growth condition on \( \sigma_i \) we see that

\[
\phi_t^{m,k} \leq 2^{k-1}|\theta_t(0) - \bar{x}_n|^k + 2^{k-1}E|\int_0^{t \wedge T_m} e^{2ns} \sigma_i(\theta_i(s))^2 ds|^k
\]

\[
\leq c_k + 2^{k-1}c_kE|\int_0^{t \wedge T_m} e^{2ns} \sigma_i(\theta_i(s))^2 ds|^{k/2}
\]

\[
\leq c_k + c_k^' E T^{k/2 - 1} \int_0^{t \wedge T_m} e^{-kns} \sigma^k(\theta(s))ds
\]

\[
\leq c_k^' + c_k T^{k/2 - 1} E |\int_0^t e^{kn(s \wedge T_m)}(C_k^' + C_k(\theta(s \wedge T_m) - \bar{x}_n))^k ds |
\]

\[
\leq c_k T^{k/2 - 1} c_k T^{k/2 - 1} \int_0^t \phi_s^{m,k} ds.
\]

A simple application of Gronwall’s inequality gives that for \( 0 \leq t \leq T \)

\[
\phi_t^{m,k} \leq c_k T^{k/2 - 1} \exp(c_k T^{k/2 - 1} t) \leq c_k T^{k/2 - 1} \exp(c_k T^{k/2}).
\]

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

\[
E|\theta(t)|^k \leq c_k T^{k/2 - 1} \exp(C_k T^{k/2}), \quad 0 \leq t \leq T.
\]

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

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

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

\[
E(\int_0^{t \wedge T_m} \theta^{k-1}\sigma(\theta)dW)^2 \leq E(\int_0^{t \wedge T_m} \theta^{2k-2}\sigma^2(\theta)ds)
\]

\[
\leq E(\int_0^{t \wedge T_m} C_1 \theta^{2k-2} + C_1 \theta^{2k} ds)
\]

\[
\leq\int_0^t C_1 E \theta^{2k-2} + C_1 E \theta^{2k} ds
\]

\[
\leq c_k T^{k} e^{c T^k}
\]

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

\[
\phi_t^k = \phi_0^k + E \int_0^t (kn(\bar{x}_n - \theta(s))\theta(s)^{k-1} + \frac{1}{2}k(k - 1)\theta(s)^{k-2}\sigma^2(\theta(s)))ds.
\]
We proceed by induction noting that $\phi^0_t = 1$ and $E\theta_t = \bar{x}_n$ so that using $k = 2$ and the linear growth bound we have
\[
\phi^2_t \leq (C_l - 2n) \int_0^t \phi^2_s ds + \left(2n\bar{x}_n^2 + C_l\right)t.
\]
Assume that $n$ is large enough so that $2n > C_l$, then in differential form we have
\[
\frac{d\phi^2_t}{dt} \leq (C_l - 2n)\phi^2_t + 2n\bar{x}_n^2 + C_l,
\]
this can be solved to get
\[
\phi^2_T \leq \frac{2n\bar{x}_n^2 + C_l}{2n - C_l} + \left(\xi^2 - \frac{2n\bar{x}_n^2 + C_l}{2n - C_l}\right) e^{-(2n-C_l)T}.
\]
Thus we have the uniform bound for all $T > 0$,
\[
\phi^2 \leq C_2(1 \vee \xi^2 \vee \bar{x}_n^2).
\]
For the general case, using the linear growth of $\sigma$, we have
\[
\phi^k_t \leq |\xi|^k + \int_0^t (kn\bar{x}_n \phi^{k-1}_s + \frac{1}{2}k(k-1)C_l - kn)\phi^k_s + \frac{1}{2}k(k-1)C_l|\phi^{k-2}_s|ds.
\]
(23)
Now assume that $\phi^p_t \leq C_p(1 \vee |\xi|^p \vee |\bar{x}_n|^p)$ for all $p \leq k - 1$ and $t > 0$. Using this in (23) we get
\[
\phi^k_t \leq |\xi|^k + kC_{k-1}n\bar{x}_n(1 \vee |\xi|^{k-1} \vee |\bar{x}_n|^{k-1})t + \int_0^t \left(\frac{1}{2}(k-1)C_l - n\right)\phi^k_s ds + C_lC_{k-2}(1 \vee |\xi|^{k-2} \vee |\bar{x}_n|^{k-2})t.
\]
For $n > \frac{1}{2}(k-1)C_l$ we have, by solving the associated differential inequality, that there is a $C_k$ such that
\[
\phi^k_t \leq C_k(1 \vee |\xi|^k \vee |\bar{x}_n|^k),
\]
provided that $(k-1)C_l < 2n$ as required.
(3) The generator of the diffusion is
\[
Af = n(\bar{x}_n - \theta)\frac{\partial f}{\partial \theta} + \frac{1}{2}\sigma_i^2(\theta)\frac{\partial^2 f}{\partial \theta^2}.
\]
Under the condition that $\sqrt{2C_l} < 2n$ we can use the Lyapunov function technique to show that we have convergence to a stationary distribution at an exponential rate. Let $V(x) = 1 + x^2$, then
\[
AV(x) = 2n(\bar{x}_n - x) + \sigma_i^2(x).
\]
Using the linear growth bound $\sigma_i^2(x) \leq \sqrt{2C_l}V(x)$ for all $i$, we have
\[
AV(x) = -(2n - \sqrt{2C_l})V(x) + 2n(\bar{x}_n x + 1).
\]
Hence, if $\beta = (2n - \sqrt{2C_l})/2$, $b = (n^2(1 + \bar{x}_n^2) - C^2_l/2)/\beta$, we have
\[
AV(x) = -\beta V(x) + bIC,
\]
with $C = \{x : |x - \frac{n2\bar{x}_n}{n}| \leq \sqrt{\frac{b}{2}}\}$. By [11] 20.3.2, this condition gives the existence of, and exponential convergence to, the stationary distribution $p_i$. 

To find the stationary distribution we just need to solve
\[ A^* p_i = -\frac{\partial}{\partial \theta} (n(x_n - \theta)p_i) + \frac{1}{2} \frac{\partial^2}{\partial \theta^2} (\sigma^2(\theta)p_i) = 0. \]  
(24)

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

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

**Corollary A.2.** The stationary distribution of the SDE has moments of order up to $\frac{2n + C_1}{C_1}$.

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.

**Theorem A.3.** For each $i \in \mathbb{N}$, the Markov chain $\{\theta_{i,m}^{(k)}, k = 0, 1, 2, \ldots\}$ converges to a unique stationary distribution $\pi_{i,m}$. There exists $r_i < 1$ and $R_i < \infty$ such that for any Borel set $A$

\[ \sup_x \frac{|\mathbb{P}^{\theta_{i,m}^{(m)}}(x) - \pi_{i,m}(A)|}{1 + x^2} \leq R_i r_i^m. \]

**Proof.** We give the proof for the $i = 1$ case and then discuss the extensions required for the general case. In order to establish the positive recurrence we use the Lyapunov function technique. Let $V(x) = 1 + x^2$. From [11], Chapter 15, we have convergence to stationarity if there exists a petite set $C$ as well as a $\beta, b > 0$ such that

\[ \Delta V(x) := \mathbb{E}^\theta V(\theta^{(1)}_m) - V(x) \leq -\beta V(x) + bI_C. \]

From our above estimates and linear growth assumption we have

\[ \mathbb{E}^\theta V(\theta^{(1)}) = 1 + \mathbb{E} \left( x + \frac{n(x_n - x)}{n + m} + R_m(x) \right)^2 \]

\[ = V(x) + 2x \frac{n(x_n - x)}{n + m} + \mathbb{E} \left( \frac{n(x_n - x)}{n + m} + R_m(x) \right)^2 \]

\[ = V(x) + \frac{2n(x_n - x)(n + m) + n^2(x_n - x)^2 + m\sigma^2(x)}{(n + m)^2} \]

\[ \leq V(x) - \frac{n^2 + (2n - C)m}{(n + m)^2} x^2 + \frac{2nm|x_n|}{(n + m)^2} x + \frac{n^2 + mC}{(n + m)^2} \]

\[ \Delta V(x) \leq -\alpha x^2 + \beta x + \gamma \]

with

\[ \alpha = \frac{n^2 + (2n - C)m}{(n + m)^2}, \quad \beta = \frac{2nm|x_n|}{(n + m)^2}, \quad \gamma = \frac{n^2|x_n|^2 + mC}{(n + m)^2}. \]

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

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

A simple calculation gives

\[ \Delta V(x) \leq -\frac{1}{2} \alpha V(x) + (\gamma + \frac{1}{2} \alpha + \frac{\beta^2}{2\alpha})I_{\{|x - \frac{\beta}{\alpha}| < \sqrt{\frac{\alpha}{\beta} + 1} + \frac{\beta^2}{\alpha^2}\}}. \]
It is easy to see that if the chain has a transition kernel with full support the set
\[ \mathcal{C} = \{ x : |x - \frac{\beta}{\alpha}| < \sqrt{\frac{2\gamma}{\alpha} + 1 + \frac{\beta^2}{\alpha^2}} \} \]
is petite for a suitable multiple of Lebesgue measure.

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 $\alpha, \beta, \gamma$ 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
\[ \Delta_m V(x) \leq \frac{1}{2}\alpha_0 V(x) + (\gamma_0 + \frac{1}{2}\alpha_0 + \frac{\beta^2}{2\alpha_0})I_\mathcal{C}. \]

We can now apply the result on $V$-geometric ergodicity, [11] Theorem 16.0.1, to deduce the estimate, with coefficients independent of $m$.

The general case where $i > 1$ is a simple extension of the $i = 1$ case. We first observe that, by the linear growth condition and the bounds on $E\theta^2$ from Lemma 4.3, we have
\[
\mathbb{E}N_m(\tau_m)^2 \leq \mathbb{E} \int_0^{\tau_m} e^{-2(n+m)(\tau_m-s)}C_i^2(\theta_i^{n,m}(s))ds \\
\leq \mathbb{E} \int_0^{\tau_m} e^{-2(n+m)(\tau_m-s)}C_i(1 + \mathbb{E}(\theta_i^{n,m}(s))^2)ds \\
\leq \frac{C_i(1 + c_2\bar{x}_{n+m}^2)}{2(n+m)} \\
\leq \frac{C_i(1 + c_2\mathbb{E}(x + \frac{n}{n+m}(\bar{x}_n - x) + \frac{1}{n+m}R_m(x))^2)}{2(n+m)}
\]

Thus we keep the same Lyapunov function and from our above estimates and linear growth assumption, we obtain the following uniform control
\[
\mathbb{E}^x V(\theta_i^{(1)}) = 1 + \mathbb{E} \left( x + \frac{n(\bar{x}_n - x) + R_m(x)}{n+m} + N_m(\tau_m) \right)^2 \\
= 1 + \mathbb{E} \left( x + \frac{n(\bar{x}_n - x) + R_m(x)}{n+m} \right)^2 + \mathbb{E} \left( N_m(\tau_m) \right)^2 \\
\leq 1 + \frac{C_i}{2(m+n)} + \left( 1 + \frac{C_i c_2}{2(m+n)} \right) \mathbb{E} \left( x + \frac{n(\bar{x}_n - x) + R_m(x)}{n+m} \right)^2 \\
= V(x) + \frac{C_i}{2(m+n)} + \left( 1 + \frac{C_i c_2}{2(m+n)} \right) \frac{2nx(\bar{x}_n - x)(n+m) + n^2(\bar{x}_n - x)^2 + m\sigma^2(x)}{(n+m)^2} \\
+ \frac{C_i c_2 x^2}{2(m+n)} \\
\Delta^{(i)} V(x) \leq -\alpha_i x^2 + \beta_i x + \gamma_i 
\]
with \[
\alpha_i = \frac{2n - C_i(1 + \frac{c_2}{2})}{m} + O\left( \frac{1}{m^2} \right), \quad \beta_i = \left( 1 + \frac{C_i c_2}{2(m+n)} \right) \frac{2nm\bar{x}_n}{(n+m)^2},
\]
\[ \gamma_i = \frac{C_1}{2(m+n)} + \left(1 + \frac{C_1c_2}{2(m+n)}\right) \frac{n^2x^2 + mC}{(n+m)^2}. \]

Thus, incorporating the time change, and provided \(2n > C, C_1 = C_1(1 + \frac{1}{2}c_2),\) we have that \(\alpha_i, \beta_i, \gamma_i > 0,\) independent of \(m,\) and

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

The same calculations as before give

\[ \Delta_m^{(i)} V(x) \leq -\frac{1}{2} \alpha_i V(x) + (\gamma_i + \frac{1}{2} \alpha_i + \frac{\beta^2}{2\alpha_i}) I_{c_i}, \]

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

\[
\square
\]

**Corollary A.4.** The sequence of stationary distributions \(\pi_{i,m}\) for the sped up Markov chains \(\theta_m\) converges to \(\pi_i,\) the stationary distribution for the solution \(\theta_i\) to the SDE.

**Proof.** All we need to show now is that the sequence \(\pi_m\) converges weakly to \(\pi,\) the stationary distribution for our limit stochastic differential equation. In order to show this consider a Borel set \(A.\) By stationarity

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

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

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

By the geometric ergodicity in Theorem A.3 there exists \(1 < r_i\) and \(R_i < \infty\) such that

\[ \sup_x \left| \mathbb{P}^x(\theta_i^{[mt]}(t) \in A) - \pi_i,m(A) \right| 1 + x^2 \leq R_i r_i^{-t}. \]

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

\[
|\pi_i(A) - \pi_i,m(A)| \leq \left| \int \pi_i(dx) \mathbb{P}^x(\theta_i(t) \in A) - \int \pi_i(dx) \mathbb{P}^x(\theta_{i,m}(t) \in A) \right| + \left| \int \pi_i(dx) \mathbb{P}^x(\theta_{i,m}(t) \in A) - \pi_{i,m}(A) \right| \\
= \epsilon/2 + 1 - \pi(K_i) + \sup_{x \in K_i} \left| \mathbb{P}^x(\theta_{i,m}(t) \in A) - \pi_{i,m}(A) \right| \\
= \epsilon/2 + \epsilon/4 + \sup_{x \in K_i} \left| \mathbb{P}^x(\theta_i^{[mt]}(t) \in A) - \pi_{i,m}(A) \right| \\
= 3\epsilon/4 + (1 + K^2)R_i r_i^{-t}. \]

Thus we can take \(t\) large enough to ensure that have that for all \(m > m_1\)

\[ |\pi_i(A) - \pi_i,m(A)| < \epsilon. \]

As this holds for each \(A\) we have weak convergence of the measures. \[
\square
\]
References


