Estimates and Standard Errors for Ratios of Normalizing Constants from Multiple Markov Chains via Regeneration

Hani Doss[1] and Aixin Tan[2]

[1]Department of Statistics, University of Florida
[2]Department of Statistics, University of Iowa

Abstract

In the classical biased sampling problem, we have k densities $\pi_1(\cdot), \ldots, \pi_k(\cdot)$, each known up to a normalizing constant, i.e. for $l = 1, \ldots, k$, $\pi_l(\cdot) = \nu_l(\cdot)/m_l$, where $\nu_l(\cdot)$ is a known function and m_l is an unknown constant. For each l, we have an iid sample from π_l, and the problem is to estimate the ratios m_l/m_s for all l and all s. This problem arises frequently in several situations in both frequentist and Bayesian inference. An estimate of the ratios was developed and studied by Vardi and his co-workers over two decades ago, and since then there has been much subsequent work on this problem from many different perspectives. In spite of this, there are no rigorous results in the literature on how to estimate the standard error of the estimate. In this paper we present a class of estimates of the ratios of normalizing constants that are appropriate for the case where the samples from the π_l's are not iid sequences, but are Markov chains. We also develop an approach based on regenerative simulation for obtaining standard errors for the estimates of ratios of normalizing constants. These standard error estimates are valid for both the iid case and the Markov chain case.

Key words and phrases: Geometric ergodicity, importance sampling, Markov chain Monte Carlo, ratios of normalizing constants, regenerative simulation, standard errors.

Research supported by NSF Grants DMS-08-05860 and DMS-11-06395.

1 Introduction

The problem of estimating ratios of normalizing constants of unnormalized densities arises frequently in statistical inference. Here we mention three instances of this problem. In missing data (or latent variable) models, suppose that the data is X_{obs}, and the likelihood of the data is difficult to write down but X_{obs} can be augmented with a part X_{mis} in such a way that the likelihood for $(X_{\text{mis}}, X_{\text{obs}})$ is easy to write. In this case (using generic notation) we have $p_\theta(X_{\text{mis}} \mid X_{\text{obs}}) = p_\theta(X_{\text{mis}}, X_{\text{obs}})/p_\theta(X_{\text{obs}})$. The denominator, i.e. the likelihood of the observed data at parameter value θ, is precisely a normalizing constant. For the purpose of carrying out likelihood inference, if θ_1 is some reference value, knowledge of $\log\bigl(p_\theta(X_{\text{obs}})/p_{\theta_1}(X_{\text{obs}})\bigr)$ is equivalent to knowledge of $\log(p_\theta(X_{\text{obs}}))$: for these two functions the maximum occurs at the same point, and the negative second derivative at the maximum (i.e. the observed Fisher information) is the same.

A second example arises when the likelihood has the form $p_\theta(x) = g_\theta(x)/z_\theta$, where g_θ is a known function. This situation arises in exponential family problems, and except for the usual textbook examples, the normalizing constant is analytically intractable. If for some arbitrary point θ_1 we know the ratio z_θ/z_{θ_1}, then we would know $p_\theta(x)$ up to a multiplicative constant and, as before, this would be equivalent to knowing $p_\theta(x)$ itself. A third example arises in certain hyperparameter selection problems in Bayesian analysis. Suppose that we wish to choose a prior from the family $\{\pi_h,\, h \in \mathcal{H}\}$, where the π_h's are densities with respect to a dominating measure μ. For any $h \in \mathcal{H}$, the marginal likelihood of the data X when the prior is π_h is given by $m_h(X) = \int p_\theta(X)\pi_h(\theta)\, \mu(d\theta)$, i.e. it is the normalizing constant in the statement "the posterior is proportional to the likelihood times the prior." The empirical Bayes choice of h is by definition $\text{argmax}_h\, m_h(X)$. Suppose that h_1 is some arbitrary point in $\mathcal{H}$. As in the previous two examples, for the purpose of finding the empirical Bayes choice of h, knowing $m_h(X)/m_{h_1}(X)$ is equivalent to knowing $m_h(X)$. (One may also be interested in the closely related problem of estimating the posterior expectation of a function $f(\theta)$ when the hyperparameter is h, which is given by $E_h(f(\theta) \mid X) = \bigl(\int f(\theta)p_\theta(X)\pi_h(\theta)\, \mu(d\theta)\bigr)/m_h(X)$. Estimating $E_h(f(\theta) \mid X)$ as h varies is relevant in Bayesian sensitivity analysis. The scheme for doing this used in Buta and Doss (2011) does not involve estimating $m_h(X)$ itself and requires only estimating $m_h(X)/m_{h_1}(X)$ for some fixed $h_1 \in \mathcal{H}$.)

Now, estimation of a normalizing constant is generally a difficult problem; for example, the so-called harmonic mean estimator proposed by Newton and Raftery (1994) typically converges at a rate that is much slower than $\sqrt{n}$ (Wolpert and Schmidler, 2011). On the other hand, estimating a ratio of normalizing constants typically can be done with a $\sqrt{n}$-consistent estimator. To illustrate this fact, consider the second of the problems described above, and let μ be the measure with respect to which the p_θ's are densities. Suppose that $X_1, X_2, \ldots$ are a "sample" from p_{θ_1} (iid sample or ergodic Markov chain output). For the simple and well-known estimator $(1/n)\sum_{i=1}^n g_\theta(X_i)/g_{\theta_1}(X_i)$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g_\theta(X_i)}{g_{\theta_1}(X_i)} \xrightarrow{\text{a.s.}} \int \frac{g_\theta(x)}{g_{\theta_1}(x)} p_{\theta_1}(x)\, \mu(dx) = \frac{z_\theta}{z_{\theta_1}}, \tag{1.1}$$

and under certain moment conditions on the ratio $g_\theta(X_i)/g_{\theta_1}(X_i)$ and mixing conditions on the chain, the estimate on the left of (1.1) also satisfies a central limit theorem (CLT). In fact, in all the problems mentioned above, it is not necessary to estimate the normalizing constants themselves, and it is sufficient to estimate ratios of normalizing constants.

The estimator above does not work well if θ is not close to θ_1, or more precisely, if g_θ and g_{θ_1} are not close. It is better to choose $\theta_1, \dots, \theta_k$ appropriately spread out in the parameter space Θ, and on the left side of (1.1) replace g_{θ_1} with $\sum_{s=1}^k w_s g_{\theta_s}$, where $w_s > 0$, $s = 1, \dots, k$. The hope is that g_θ will be close to at least one of the g_{θ_s}'s, and so preclude having large variances. To implement this, *suppose* we know all the ratios $z_{\theta_s}/z_{\theta_t}$, $s, t \in \{1, \dots, k\}$, or equivalently, we know $z_{\theta_1}/z_{\theta_s}$, $s \in \{1, \dots, k\}$. In this case, if for each $l = 1, \dots, k$ there is available a sample $X_1^{(l)}, \dots, X_{n_l}^{(l)}$ from $g_{\theta_l}/z_{\theta_l}$, then letting $n = \sum_{l=1}^k n_l$ and $a_l = n_l/n$, we have

$$\begin{aligned}\sum_{l=1}^k \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{a_l g_\theta(X_i^{(l)})}{\sum_{s=1}^k a_s g_{\theta_s}(X_i^{(l)})(z_{\theta_1}/z_{\theta_s})} &\xrightarrow{\text{a.s.}} \sum_{l=1}^k \int \frac{a_l g_\theta(x)}{\sum_{s=1}^k a_s g_{\theta_s}(x)(z_{\theta_1}/z_{\theta_s})} \frac{g_{\theta_l}(x)}{z_{\theta_l}} \, \mu(dx) \\ &= \int \sum_{l=1}^k \frac{a_l g_{\theta_l}(x)/z_{\theta_l}}{\sum_{s=1}^k a_s g_{\theta_s}(x)(z_{\theta_1}/z_{\theta_s})} g_\theta(x) \, \mu(dx) = \frac{z_\theta}{z_{\theta_1}}.\end{aligned} \tag{1.2}$$

When compared with the estimate on the left side of (1.1), the estimate on the left side of (1.2) is accurate over a much bigger range of θ's. But to use it, it is necessary to be able to estimate the ratios $z_{\theta_1}/z_{\theta_s}$, $s \in \{1, \dots, k\}$, and it is this problem that is the focus of this paper.

We now state explicitly the version of this problem that we will deal with here, and we change to the notation that we will use for the rest of the paper. We have k densities $\pi_1, \dots, \pi_k$ with respect to the measure μ, which are known except for normalizing constants, i.e. we have $\pi_l = \nu_l/m_l$, where the ν_l's are known functions and the m_l's are unknown constants. For each l we have a Markov chain $\Phi_l = \{X_1^{(l)}, \dots, X_{n_l}^{(l)}\}$ with invariant density π_l, the k chains are independent, and the objective is to estimate all possible ratios m_i/m_j, $i \neq j$ or, equivalently, the vector

$$\boldsymbol{d} = (m_2/m_1, \dots, m_k/m_1).$$

When the samples are iid sequences, this is the biased sampling problem introduced by Vardi (1985), which contains examples that differ in character quite a bit from those considered here.

Suppose we are in the iid case, and consider the pooled sample $S = \big\{X_i^{(l)},\ i = 1, \dots, n_l,\ l = 1, \dots, k\big\}$. Let $x \in S$, and suppose that x came from the l^{th} sample. If we pretend that the only thing we know about x is its value, then the probability that x came from the l^{th} sample is

$$\frac{n_l \pi_l(x)}{\sum_{s=1}^k n_s \pi_s(x)} = \frac{a_l \nu_l(x)/m_l}{\sum_{s=1}^k a_s \nu_s(x)/m_s} := \lambda_l(x, \boldsymbol{m}), \tag{1.3}$$

where $\boldsymbol{m} = (m_1, \dots, m_k)$. Geyer (1994) proposed to treat the vector $\boldsymbol{m}$ as an unknown parameter and to estimate it by maximizing the quasi-likelihood function

$$L_n(\boldsymbol{m}) = \prod_{l=1}^k \prod_{i=1}^{n_l} \lambda_l(X_i^{(l)}, \boldsymbol{m}) \tag{1.4}$$

with respect to $\boldsymbol{m}$. Actually, there is a non-identifiability issue regarding L_n: for any constant $c > 0$, $L_n(\boldsymbol{m})$ and $L_n(c\boldsymbol{m})$ are the same. So we can estimate $\boldsymbol{m}$ only up to an overall multiplicative constant, i.e. we can estimate only $\boldsymbol{d}$. Accordingly, Geyer (1994) proposed to estimate $\boldsymbol{d}$ by maximizing $L_n(\boldsymbol{m})$ subject to the constraint $m_1 = 1$. (A more detailed discussion of the quasi-likelihood function (1.4) is given in Section 2.) In fact, the resulting estimate, $\hat{\boldsymbol{d}}$, was originally proposed by Vardi (1985), and studied further by Gill, Vardi and Wellner (1988), who showed that it is consistent and asymptotically normal, and established its optimality properties, all under the assumption that for each $l = 1, \dots, k$, $X_1^{(l)}, \dots, X_{n_l}^{(l)}$ is an iid sequence. Geyer (1994) extended the consistency and asymptotic normality result to the case where the k sequences $X_1^{(l)}, \dots, X_{n_l}^{(l)}$ are Markov chains satisfying certain mixing conditions. The estimate was rederived in Meng and Wong (1996), Kong et al. (2003), and Tan (2004) from completely different perspectives, all under the iid assumption.

As mentioned earlier, for the kinds of problems we have in mind the distributions π_l are analytically intractable, and estimates of the sort (1.1) or (1.2), and the estimate of $\boldsymbol{d}$ are applicable to a much larger class of problems if we are willing to use Markov chain samples instead of iid samples. Estimation of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$ is then difficult for two reasons. First, the estimate $\hat{\boldsymbol{d}}$ is obtained as the solution to a constrained optimization problem, and second, when the sequences $X_1^{(l)}, \dots, X_{n_l}^{(l)}$ are Markov chains instead of iid sequences, the asymptotic covariance matrix has a complex form and is difficult to estimate consistently.

The present paper deals with two issues. First, none of the authors cited above give consistent estimators of the variance, even in the iid case. (For the iid case, Kong et al. (2003) give an estimate that involves the inverse of a certain Fisher information matrix, but this formal calculation does not establish consistency of the estimate, or even the necessary CLT, nor do the authors make such claims.) As mentioned earlier, the problem of estimating the variance is *far* more challenging when the samples are Markov chains as opposed to iid sequences. In this paper we give a CLT for the vector $\hat{\boldsymbol{d}}$ based on regenerative simulation. The main benefit of this result is that it gives, essentially as a free by-product, a simple consistent estimate of the covariance matrix in the Markov chain setting. Second, the estimate obtained by the afore-mentioned authors is optimal in the case where the samples are iid. When the samples are Markov chains, the estimates are no longer optimal. We present a method for forming estimators which are suitable in the Markov chain setting. The regeneration-based CLT and estimate of the variance both apply to the class of estimators that we propose.

The rest of this paper is organized as follows. In Section 2 we use ideas from regenerative simulation to develop a CLT for $\hat{\boldsymbol{d}}$, and we show how our estimate of variance emerges as a by-product. In Section 3 we describe a class of estimators of $\boldsymbol{d}$ which are suitable when the samples are Markov chains, as opposed to iid samples, and we also propose a method for choosing an estimator from this class. In Section 4 we present a small study that illustrates the gains obtained from using an estimate of $\boldsymbol{d}$ designed for Markov chains, and we illustrate our methodology by showing how it can be used to estimate certain quantities of interest in the Ising model of statistical mechanics. The Appendix provides proofs of the three assertions made by the theorem in Section 2, namely strong consistency of $\hat{\boldsymbol{d}}$, the CLT for $\hat{\boldsymbol{d}}$, and strong consistency of the estimate of variance of $\hat{\boldsymbol{d}}$.

2 A Regeneration-Based CLT and Variance Estimate

We begin by considering more carefully the quasi-likelihood function for $\boldsymbol{m}$ given by (1.4), and for the technical development it is much more convenient to work on the log scale. So define the vector $\boldsymbol{\zeta}$ by

$$\zeta_l = -\log(m_l) + \log(a_l), \qquad \text{for } l = 1, \ldots, k, \tag{2.1}$$

and rewrite (1.3) as

$$p_l(x, \boldsymbol{\zeta}) = \frac{\nu_l(x) e^{\zeta_l}}{\sum_{s=1}^k \nu_s(x) e^{\zeta_s}}, \qquad \text{for } l = 1, \ldots, k. \tag{2.2}$$

Clearly, $\boldsymbol{\zeta}$ determines and is determined by $(m_1, \ldots, m_k)$, and the log quasi-likelihood function for $\boldsymbol{\zeta}$ is

$$l_n(\boldsymbol{\zeta}) = \sum_{l=1}^k \sum_{i=1}^{n_l} \log\bigl(p_l(X_i^{(l)}, \boldsymbol{\zeta})\bigr). \tag{2.3}$$

In (2.1), $(m_1, \ldots, m_k)$ is an arbitrary vector with strictly positive components, i.e. m_l need not correspond to the normalizing constant for ν_l. We will use $\boldsymbol{\zeta}_{(t)}$ to denote the true value of $\boldsymbol{\zeta}$, i.e. the value it takes when the m_l's are the normalizing constants for the ν_l's. The non-identifiability issue now is that for any constant $c \in \mathbb{R}$, $l_n(\boldsymbol{\zeta})$ and $l_n(\boldsymbol{\zeta} + c 1_k)$ are the same (here, 1_k is the vector of k 1's), so we can estimate $\boldsymbol{\zeta}_{(t)}$ only up to an additive constant. Accordingly, with $\boldsymbol{\zeta}_0 \in \mathbb{R}^k$ defined by $[\boldsymbol{\zeta}_0]_l = [\zeta_{(t)}]_l - \bigl(\sum_{s=1}^k [\zeta_{(t)}]_s\bigr)/k$, Geyer (1994) proposed to estimate $\boldsymbol{\zeta}_0$ by $\hat{\boldsymbol{\zeta}}$, the maximizer of l_n subject to the linear constraint $\boldsymbol{\zeta}^\top 1_k = 0$, and thus obtain an estimate of $\boldsymbol{d}$.

The term $p_l(x, \boldsymbol{\zeta})$ in (2.2) has the appearance of a likelihood ratio, and in the denominator, the probability measure ν_s/m_s is given weight proportional to the length of the chain Φ_s. Now Gill et al.'s (1988) optimality result does not apply to the Markov chain case, in which the chains $\Phi_1, \ldots, \Phi_k$ mix at possibly different rates, and the a_s's should in some sense reflect the vague notion of "effective sample sizes" of the different chains. The optimal choice of the vector $\boldsymbol{a} = (a_1, \ldots, a_k)$ is very difficult to determine theoretically, and in Section 3 we describe an empirical method for choosing $\boldsymbol{a}$. Accordingly in (2.1) and henceforth, $\boldsymbol{a}$ will not necessarily be given by $a_l = n_l/n$, but will be an arbitrary probability vector satisfying the condition that $a_l > 0$ for $l = 1, \ldots, k$.

2.1 Regeneration and a Minorization Condition

We are interested in obtaining a standard error estimate for $\hat{\boldsymbol{\zeta}}$. To describe our approach, we first briefly review the available methods for estimating variances based on Markov chain output. Because $\hat{\boldsymbol{\zeta}}$ is a complicated estimate, we first discuss the much simpler case where we have a single Markov chain $X_1, X_2, \ldots$ on the measurable space $(\mathsf{X}, \mathcal{B})$, with invariant distribution π, $f \colon \mathsf{X} \to \mathbb{R}$ is a function, and we are interested in estimating the variance of $\bar{f}_n := n^{-1} \sum_{i=1}^n f(X_i)$. The commonly used approaches are those based on spectral methods, batching, and regeneration (see, e.g., Geyer, 1992; Mykland, Tierney and Yu, 1995; Jones et al., 2006). Among these three, the cleanest is the one based on regenerative simulation.

A *regeneration* is a random time at which a stochastic process probabilistically restarts itself. The "tours" made by the chain in between such random times are iid, and this fact makes much easier the asymptotic analysis of averages, and of statistics based on vectors of averages. In the discrete state space setting, if $x \in \mathsf{X}$ is any point to which the chain returns infinitely often, then the times of return to x form a sequence of regenerations. For most of the Markov chains used in MCMC algorithms, the state space is continuous, and there is no point to which the chain returns infinitely often with probability one. Even when the state space is discrete, regenerations based on returns to a point x, as described above, are often not useful, because if x has very small probability under the stationary distribution, then on average it will take a very long time to return to x. Fortunately, Mykland et al. (1995) provided a general technique for identifying a sequence of regeneration times $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ that is based on the construction of a *minorization condition*. This construction will be reviewed shortly, but we now briefly sketch how having a regeneration sequence $\{\tau_t\}_{t=0}^{\infty}$ enables us to construct a simple estimate of the standard error of $\bar{f}$. Define

$$Y_t = \sum_{i=\tau_{t-1}}^{\tau_t - 1} f(X_i) \quad \text{and} \quad T_t = \sum_{i=\tau_{t-1}}^{\tau_t - 1} 1 = \tau_t - \tau_{t-1}, \qquad t = 1, 2, \ldots,$$

and note that the pairs (Y_t, T_t) form an iid sequence. If we run the chain for ρ regenerations, then the total number of cycles, starting at τ_0, is given by $n = \sum_{t=1}^{\rho} T_t$. We may write $\bar{f}$ as

$$\frac{\sum_{i=1}^{n} f(X_i)}{n} = \frac{\sum_{t=1}^{\rho} Y_t}{\sum_{t=1}^{\rho} T_t} = \frac{\left(\sum_{t=1}^{\rho} Y_t\right)/\rho}{\left(\sum_{t=1}^{\rho} T_t\right)/\rho}. \tag{2.4}$$

Equation (2.4) expresses $\bar{f}$ as a ratio of two averages of iid quantities, and this representation enables us to use the delta method to obtain both a CLT for $\bar{f}$ and a simple standard error estimate for $\bar{f}$.

An outline of the argument is as follows. From (2.4) we see that as $\rho \to \infty$ (which implies that $n \to \infty$) we have

$$E_\pi(f(X)) \xleftarrow{\text{a.s.}} \frac{\sum_{i=1}^{n} f(X_i)}{n} = \frac{\left(\sum_{t=1}^{\rho} Y_t\right)/\rho}{\left(\sum_{t=1}^{\rho} T_t\right)/\rho} \xrightarrow{\text{a.s.}} \frac{E(Y_1)}{E(T_1)}, \tag{2.5}$$

where the convergence statement on the left follows from the ergodic theorem, and the convergence statement on the right follows from two applications of the strong law of large numbers. (In (2.5) the subscript π to the expectation indicates that $X \sim \pi$.) From (2.5) we obtain $E(Y_1) = E_\pi(f(X))E(T_1)$. Now the bivariate CLT gives

$$\rho^{1/2} \begin{pmatrix} \bar{Y} - E_\pi(f(X))E(T_1) \\ \bar{T} - E(T_1) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma_f), \tag{2.6}$$

where $\Sigma_f = \operatorname{Cov}\big((Y_1, T_1)^\top\big)$. The delta method applied to the function $h(y,t) = y/t$ gives the CLT

$$\rho^{1/2}\big(\bar{Y}/\bar{T} - E_\pi(f(X))\big) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2),$$

where $\sigma_f^2 = (\nabla h)^\top \Sigma_f \nabla h$ (and ∇h is evaluated at the vector of means in (2.6)). Moreover, it is straightforward to check that for the variance estimator

$$\hat{\sigma}_f^2 = \frac{\sum_{t=1}^{\rho} (Y_t - \bar{f} T_t)^2}{\rho \bar{T}^2},$$

we have $\hat{\sigma}_f^2 \xrightarrow{\text{a.s.}} \sigma_f^2$. The regularity conditions needed to make this argument rigorous are spelled out when we discuss the case of the more complicated estimator $\hat{\zeta}$ (Section 2.3 and the Appendix).

The argument above hinges on being able to arrive at a sequence of regeneration times, and whether these are useful depends on whether the sequence has the property that the length of the tours between regenerations is not very large. We now describe the minorization condition that can sometimes be used to construct useful regeneration sequences. Let $K(x, A)$ be the Markov transition distribution, and suppose that for each $x \in \mathsf{X}$, $K(x, \cdot)$ has density $k(x, \cdot)$ with respect to a dominating measure μ. The construction described in Mykland et al. (1995) requires the existence of a function $s \colon \mathsf{X} \to [0, 1)$, whose expectation with respect to π is strictly positive, and a probability density q with respect to μ, such that $k(\cdot, \cdot)$ satisfies

$$k(x, x') \geq s(x) q(x') \qquad \text{for all } x, x' \in \mathsf{X}.$$

This is called a minorization condition and, as we describe below, it can be used to introduce regenerations into the Markov chain driven by k. Define

$$r(x, x') = \frac{k(x, x') - s(x) q(x')}{1 - s(x)}.$$

Note that for fixed $x \in \mathsf{X}$, $r(x, x')$ is a density function in x'. We may therefore write

$$k(x, x') = s(x) q(x') + (1 - s(x)) r(x, x'),$$

which gives a representation of $k(x, \cdot)$ as a mixture of two densities, $q(\cdot)$ and $r(x, \cdot)$. This provides an alternative method of simulating from k. Suppose that the current state of the chain is X_n. We generate $\delta_n \sim \text{Bernoulli}(s(X_n))$. If $\delta_n = 1$, we draw $X_{n+1} \sim q$; otherwise, we draw $X_{n+1} \sim r(X_n, \cdot)$. Note that if $\delta_n = 1$, the next state of the chain is drawn from q, which does not depend on the current state. Hence the chain "forgets" the current state and we have a regeneration. To be more specific, suppose we start the Markov chain with $X_1 \sim q$ and then use the method described above to simulate the chain. Each time $\delta_n = 1$, we have $X_{n+1} \sim q$ and the process stochastically restarts itself; that is, the process regenerates.

In practice, simulating from r can be extremely difficult. Fortunately, Mykland et al. (1995), following Nummelin (1984, p. 62), noticed a clever way of circumventing the need to draw from r. Instead of making a draw from the conditional distribution of δ_n given x_n and then generating x_{n+1} given (δ_n, x_n), which would result in a draw from the joint distribution of (δ_n, x_{n+1}) given x_n, we simply draw from the conditional distribution of x_{n+1} given x_n in the usual way (i.e. using k), and then draw δ_n given (x_n, x_{n+1}). This alternative sampling mechanism yields a draw from the same joint density, but avoids having to draw from r. Moreover, given (x_n, x_{n+1}), δ_n has a Bernoulli distribution with success probability given simply by

$$P\big(\delta_n = 1 \,|\, x_n = x', x_{n+1} = x\big) = \frac{s(x') q(x)}{k(x \,|\, x')}.$$

2.2 A Quasi-Likelihood Function Designed for the Markov Chain Setting

As mentioned earlier, Geyer (1994) showed that when we take $a_j = n_j/n$, the maximizer of the log quasi-likelihood function defined by (2.3) (subject to the constraint $\boldsymbol{\zeta}^\top 1_k = 0$) is a consistent estimate of the true value $\boldsymbol{\zeta}_0$, and also satisfies a CLT, even when the k sequences $\{X_i^{(l)}\}_{i=1}^{n_l}$, $l = 1, \dots, k$ are Markov chains. But when the k sequences are Markov chains, the choice $a_j = n_j/n$ is no longer optimal, and for other choices of $\boldsymbol{a}$, the (constrained) maximizer of (2.3) is not necessarily even consistent. We will present a new log quasi-likelihood function which does yield consistent asymptotically normal estimates, and before doing this, we give a brief motivating argument.

Suppose that we are in the simple case where we have a parametric family $\{p_\theta, \, \theta \in \Theta\}$ and we observe data $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} p_{\theta_0}$ for some $\theta_0 \in \Theta$. Let $l_y(\theta) = \log(p_\theta(y))$, and let $Q(\theta) = E_{\theta_0}(l_Y(\theta))$. The fact that $\operatorname{argmax}_\theta Q(\theta) = \theta_0$ is well known (and easy to see via a short argument involving Jensen's inequality). The log likelihood function based on $Y_1, \dots, Y_n$ is $\sum_{i=1}^n l_{Y_i}(\theta)$. By the strong law of large numbers,

$$n^{-1} \sum_{i=1}^n l_{Y_i}(\theta) \xrightarrow{\text{a.s.}} Q(\theta) \qquad \text{for all } \theta \in \Theta, \tag{2.7}$$

and assuming sufficient regularity conditions, $\operatorname{argmax}_\theta n^{-1} \sum_{i=1}^n l_{Y_i}(\theta) \xrightarrow{\text{a.s.}} \operatorname{argmax}_\theta Q(\theta) = \theta_0$, i.e. the maximum likelihood estimator is consistent.

We now return to the present situation, in which for $l = 1, \dots, k$, $\{X_i^{(l)}\}_{i=1}^{n_l}$ is a Markov chain with invariant density π_l. Suppose we use $l_n(\boldsymbol{\zeta})$ given by (2.3), with $\boldsymbol{a}$ an arbitrary probability vector (i.e. $\boldsymbol{a}$ is not necessarily given by $a_j = n_j/n$), and let $Q(\boldsymbol{\zeta}) = E_{\boldsymbol{\zeta}_0}(l_n(\boldsymbol{\zeta}))$. The key condition

$$\operatorname*{argmax}_{\boldsymbol{\zeta}} Q(\boldsymbol{\zeta}) = \boldsymbol{\zeta}_0 \tag{2.8}$$

need not hold, and the constrained maximizer of $l_n(\boldsymbol{\zeta})$ may converge, but not to the true value.

With this in mind, suppose that $\boldsymbol{a}$ is an arbitrary probability vector with non-zero entries and define $w \in \mathbb{R}^k$ by

$$w_l = a_l \frac{n}{n_l}, \qquad l = 1, \dots, k. \tag{2.9}$$

The log quasi-likelihood function we will use is

$$\ell_n(\boldsymbol{\zeta}) = \sum_{l=1}^k w_l \sum_{i=1}^{n_l} \log\bigl(p_l(X_i^{(l)}, \boldsymbol{\zeta})\bigr) \tag{2.10}$$

instead of l_n given by (2.3) [note the slight change of notation from l to ℓ]. As will emerge in our proofs of consistency and asymptotic normality of the constrained maximizer of $\ell_n(\boldsymbol{\zeta})$, for this log quasi-likelihood function, the stochastic process (in $\boldsymbol{\zeta}$) $n^{-1}\ell_n(\boldsymbol{\zeta})$ converges almost surely to a function of $\boldsymbol{\zeta}$ which is maximized at $\boldsymbol{\zeta}_0$, a condition that plays the role of (2.7) and (2.8). Note that if $a_l = n_l/n$, then $w_l = 1$ and (2.10) reduces to (2.3).

2.3 A CLT for the Estimate Designed for Markov Chains

We assume that for $l = 1, \dots, k$, chain l has Markov transition density $k_l(x, x')$ (with respect to some measure μ) which satisfies the minorization condition

$$k_l(x, x') \geq s_l(x) q_l(x') \qquad \text{for all } x, x' \in \mathsf{X} \tag{2.11}$$

for some density q_l and function $s_l \colon \mathsf{X} \to [0, 1)$ with $E_{\pi_l}(s_l(X)) > 0$, and that the chain has been run for ρ_l regenerations. Let $1 = \tau_0^{(l)} < \tau_1^{(l)} < \cdots < \tau_{\rho_l}^{(l)}$ denote the regeneration times of the l^{th} chain, and let $T_t^{(l)} = \tau_t^{(l)} - \tau_{t-1}^{(l)}$ be the length of the t^{th} tour of the l^{th} chain. So the length of the l^{th} chain, $n_l = T_1^{(l)} + \cdots + T_{\rho_l}^{(l)}$, is random. We will assume that $\rho_1, \dots, \rho_k \to \infty$ in such a way that $\rho_l / \rho_1 \to c_l \in (0, \infty)$, for $l = 1, \dots, k$. We will allow the vector $\boldsymbol{a}$ to depend on $\rho = (\rho_1, \dots, \rho_k)$, i.e. $\boldsymbol{a} = \boldsymbol{a}^{(\rho)}$ (although we will suppress this dependence in the notation except when this dependence matters), and we will make the minimal assumption that $\boldsymbol{a}^{(\rho)} \to \boldsymbol{\alpha}$ as $\rho_1, \dots, \rho_k \to \infty$, where $\boldsymbol{\alpha}$ is a probability vector with strictly positive entries. The extra generality is needed if we wish to choose $\boldsymbol{a}$ in a data-driven way (cf. Remark 3 of Section 3). The definitions of $\boldsymbol{\zeta}$ and $p_l(x, \boldsymbol{\zeta})$ given by (2.1) and (2.2), respectively, are still in force, $\boldsymbol{\zeta}_0$ is still the centered version of the true value of $\boldsymbol{\zeta}$, but now $\hat{\boldsymbol{\zeta}}$ is the constrained maximizer of the new log quasi-likelihood function (2.10). We will show that $\hat{\boldsymbol{\zeta}}$ is a consistent asymptotically normal estimate of $\boldsymbol{\zeta}_0$, and since $\boldsymbol{\zeta}_0$ determines and is determined by $\boldsymbol{d}$, this will produce a corresponding estimate $\hat{\boldsymbol{d}}$ of $\boldsymbol{d}$. Before proceeding, we mention the fact that difficulties arise if the supports of the distributions $\pi_1, \dots, \pi_k$ differ (the difficulties are pervasive: for the case where we have a continuum of distributions $\{\pi_\theta, \, \theta \in \Theta\}$, the simple estimate (1.1) is not even defined if π_θ is not absolutely continuous with respect to π_{θ_1}). So for the rest of this paper, we will assume that the k distributions $\pi_1, \dots, \pi_k$ are mutually absolutely continuous. We do not really need to make an assumption this strong, but the assumption is satisfied for all the classes of problems we are considering, and making it eliminates some technical issues.

In order to state our CLT for the vector $\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d})$, we need to define the quantities that go into the expression for the asymptotic variance. We first consider the vector $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$, whose covariance matrix is singular (since this vector sums to 0). The asymptotic distribution of $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$ involves the matrices B and Ω defined below. Let $\boldsymbol{\zeta}_{\boldsymbol{\alpha}}$ be the vector whose components are $[\boldsymbol{\zeta}_{\boldsymbol{\alpha}}]_l = -\log(m_l) + \log(\alpha_l)$, and let B be the $k \times k$ matrix given by

$$\begin{aligned} B_{rr} &= \sum_{j=1}^{k} \alpha_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})[1 - p_r(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})]\big), \qquad r = 1, \dots, k, \\ B_{rs} &= -\sum_{j=1}^{k} \alpha_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) p_s(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big), \qquad r, s = 1, \dots, k, \; r \neq s. \end{aligned} \tag{2.12}$$

We will be using the natural estimate defined by

$$\begin{aligned} \widehat{B}_{rr} &= \sum_{l=1}^{k} a_l \bigg(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \big[1 - p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})\big] \bigg), \qquad r = 1, \dots, k, \\ \widehat{B}_{rs} &= -\sum_{l=1}^{k} a_l \bigg(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \bigg), \qquad r, s = 1, \dots, k, \; r \neq s. \end{aligned} \tag{2.13}$$

Let

$$y_i^{(r,l)}(\boldsymbol{a}) = p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big), \qquad i = 1, \dots, n_l,$$
$$y_i^{(r,l)}(\boldsymbol{\alpha}) = p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big), \qquad i = 1, \dots, n_l, \tag{2.14}$$

and note that both $y_i^{(r,l)}(\boldsymbol{a})$ and $y_i^{(r,l)}(\boldsymbol{\alpha})$ have mean 0. Define

$$Y_t^{(r,l)}(\boldsymbol{a}) = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} y_i^{(r,l)}(\boldsymbol{a}), \quad \bar{Y}^{(r,l)}(\boldsymbol{a}) = \frac{1}{\rho_l}\sum_{t=1}^{\rho_l} Y_t^{(r,l)}(\boldsymbol{a}),$$
$$Y_t^{(r,l)}(\boldsymbol{\alpha}) = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} y_i^{(r,l)}(\boldsymbol{\alpha}), \quad \bar{Y}^{(r,l)}(\boldsymbol{\alpha}) = \frac{1}{\rho_l}\sum_{t=1}^{\rho_l} Y_t^{(r,l)}(\boldsymbol{\alpha}), \quad \text{and} \quad \bar{T}^{(l)} = \frac{1}{\rho_l}\sum_{t=1}^{\rho_l} T_t^{(l)}. \tag{2.15}$$

Let Ω be the $k \times k$ matrix defined by

$$\Omega_{rs} = \sum_{l=1}^{k} \frac{\alpha_l^2}{c_l} \frac{E\big(Y_1^{(r,l)}(\boldsymbol{\alpha}) Y_1^{(s,l)}(\boldsymbol{\alpha})\big)}{\big(E(T_1^{(l)})\big)^2}, \qquad r, s = 1, \dots, k, \tag{2.16}$$

To obtain an estimate $\widehat{\Omega}$, we let

$$Z_t^{(r,l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \quad \text{and} \quad \hat{\mu}_r^{(l)} = \frac{\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})}{n_l},$$

and define $\widehat{\Omega}$ by

$$\widehat{\Omega}_{rs} = \sum_{l=1}^{k} \frac{a_l^2}{c_l} \frac{1}{(\bar{T}^{(l)})^2} \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} \big(Z_t^{(r,l)} - T_t^{(l)}\hat{\mu}_r^{(l)}\big)\big(Z_t^{(s,l)} - T_t^{(l)}\hat{\mu}_r^{(l)}\big), \qquad r, s = 1, \dots, k. \tag{2.17}$$

The function $g\colon \mathbb{R}^k \to \mathbb{R}^{k-1}$ that maps $\boldsymbol{\zeta}_0$ into $\boldsymbol{d}$ is

$$g(\boldsymbol{\zeta}) = \begin{pmatrix} e^{\zeta_1 - \zeta_2} a_2/a_1 \\ e^{\zeta_1 - \zeta_3} a_3/a_1 \\ \vdots \\ e^{\zeta_1 - \zeta_k} a_k/a_1 \end{pmatrix}, \tag{2.18}$$

and its gradient at $\boldsymbol{\zeta}_0$ (in terms of $\boldsymbol{d}$) is

$$D = \begin{pmatrix} d_2 & d_3 & \dots & d_k \\ -d_2 & 0 & \dots & 0 \\ 0 & -d_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -d_k \end{pmatrix}. \tag{2.19}$$

We have $\boldsymbol{d} = g(\boldsymbol{\zeta}_0)$, and by definition $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$.

The theorem below has three parts, pertaining to the strong consistency of $\hat{\boldsymbol{d}}$, asymptotic normality of $\hat{\boldsymbol{d}}$, and a consistent estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$. For consistency we need only minimal assumptions on the Markov chains $\Phi_1, \ldots, \Phi_k$, namely the so-called basic regularity conditions (irreducibility, aperiodicity and Harris recurrence) that are needed for the ergodic theorem (Meyn and Tweedie, 1993, Chapter 17). CLTs and associated results always require a stronger condition, and the one that is most commonly used is geometric ergodicity. The theorem refers to the following conditions, which pertain to each $l = 1, \ldots, k$.

A1 The Markov chain $\{X_1^{(l)}, X_2^{(l)}, \ldots\}$ satisfies the basic regularity conditions.

A2 The Markov chain $\{X_1^{(l)}, X_2^{(l)}, \ldots\}$ is geometrically ergodic.

A3 The Markov transition density k_l satisfies the minorization condition (2.11).

For a square matrix C, $C^\dagger$ will denote the Moore-Penrose inverse of C.

Theorem 1 *Suppose that for each $l = 1, \ldots, k$, the Markov chain $\{X_1^{(l)}, X_2^{(l)}, \ldots\}$ has invariant distribution π_l.*

1. *Under A1, the log quasi-likelihood function* (2.10) *has a unique maximizer subject to the constraint $\boldsymbol{\zeta}^\top 1_k = 0$. Let $\hat{\boldsymbol{\zeta}}$ denote this maximizer, and let $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$. Then as $\rho_1 \to \infty$, $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$.*

2. *Under A1 and A2, as $\rho_1 \to \infty$,*

$$\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, W) \quad \text{where} \quad W = D^\top B^\dagger \Omega B^\dagger D. \tag{2.20}$$

3. *Assume A1–A3. Let $\widehat{D}$ be the matrix D in* (2.19) *with $\hat{\boldsymbol{d}}$ in place of $\boldsymbol{d}$, and let $\widehat{B}$ and $\widehat{\Omega}$ be defined by* (2.13) *and* (2.17)*, respectively. Then, $\widehat{W} := \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D}$ is a strongly consistent estimator of W.*

3 Choice of the Vector $\boldsymbol{a}$

As mentioned earlier, the log quasi-likelihood that has been proposed and studied in the literature involves the functions $p_l(x, \boldsymbol{\zeta})$ given by (2.2), which have the form

$$\frac{\frac{n_l}{n} \nu_l(x)/m_l}{\sum_{s=1}^k \frac{n_s}{n} \nu_s(x)/m_s}, \tag{3.1}$$

where in the denominator of (3.1), the probability density $\nu_s(x)/m_s$ is given weight proportional to the length of the s^{th} chain. Intuitively, one would want to replace n_s with the "effective sample size" for chain s, so that if chain s mixes slowly, the weight that is given to $\nu_s(x)/m_s$ is small. Unfortunately, there is really no such thing as an effective sample size because the effect of slow mixing varies quite a bit with the function whose mean is being estimated. Therefore, it is better to take a direct approach that involves replacing the vector $(n_1/n, \ldots, n_k/n)$ by a probability vector

$\boldsymbol{a}$, and choose $\boldsymbol{a}$ to minimize the variance of the resulting estimator. (It should be emphasized that the estimator is a complicated function of k chains.)

In more detail, we do the following. Let $\mathcal{S}_k = \{\boldsymbol{a} \in \mathbb{R}^k : a_1, \dots, a_k \geq 0 \text{ and } \sum_{s=1}^k a_s = 1\}$ be the k-dimensional simplex. For each $\boldsymbol{a} \in \mathcal{S}_k$, in (3.1) replace n_s/n by a_s and form the corresponding log quasi-likelihood function (see equation (2.10)), call it $\ell_n^{(\boldsymbol{a})}(\boldsymbol{\zeta})$. We let $\hat{\boldsymbol{\zeta}}_{\boldsymbol{a}}$ be the constrained maximizer of $\ell_n^{(\boldsymbol{a})}(\boldsymbol{\zeta})$, and let $\hat{\boldsymbol{d}}_{\boldsymbol{a}}$ be the corresponding estimate of $\boldsymbol{d}$. Let $W_{\boldsymbol{a}}$ be the covariance matrix of $\hat{\boldsymbol{d}}_{\boldsymbol{a}}$ given by Part 2 of Theorem 1, and let $\widehat{W}_{\boldsymbol{a}}$ be its estimate. We choose $\boldsymbol{a}$ to minimize $\operatorname{trace}(\widehat{W}_{\boldsymbol{a}})$ (this corresponds to the classical "A-optimal design").

It should be noted that we are able to carry out this optimization scheme precisely because Theorem 1 enables us to estimate $W_{\boldsymbol{a}}$. It is also worth noting that once we have constructed the k regeneration sequences $\tau_0^{(l)} < \tau_1^{(l)} < \cdots < \tau_{\rho_l}^{(l)}$, $l = 1, \dots, k$, these same sequences may be used in the computation of $\widehat{W}_{\boldsymbol{a}}$ for *all* $\boldsymbol{a} \in \mathcal{S}$.

Remarks

1. It is natural to ask whether in the Markov chain case our procedure gives rise to an optimal estimate of $\boldsymbol{d}$, and we now address this question. To keep the discussion as simple as possible, we consider the case $k = 2$. Let $\mathcal{B}$ be the set of all "bridge functions" $\beta \colon \mathsf{X} \to \mathbb{R}$ satisfying the conditions that $0 < |\int \beta(x)\pi_1(x)\pi_2(x)\,\mu(dx)| < \infty$ and $\beta(x) = 0$ when either $\pi_1(x) = 0$ or $\pi_2(x) = 0$. It is easy to see that when the two sequences $X_1^{(l)}, \dots, X_{n_l}^{(l)}$, $l = 1, 2$ are each iid, for any $\beta \in \mathcal{B}$, the estimate

$$\hat{d}_2 = \frac{n_1^{-1}\sum_{i=1}^{n_1} \beta(X_i^{(1)})\nu_2(X_i^{(1)})}{n_2^{-1}\sum_{i=1}^{n_2} \beta(X_i^{(2)})\nu_1(X_i^{(2)})}$$

is a consistent and asymptotically normal estimate of d_2. Meng and Wong (1996) show that within $\mathcal{B}$, the function for which the asymptotic variance is minimized is

$$\beta_{\text{opt,iid}}(x) = \big[s_1\nu_1(x) + s_2\nu_2(x)/d_2\big]^{-1},$$

where $s_j = n_j/n$, $j = 1, 2$. Because this function involves the unknown d_2, Meng and Wong (1996) propose an iterative scheme in which we start with, say, $\hat{d}_2^{(0)} = 1$, and at stage m, we form

$$\hat{d}_2^{(m+1)} = \frac{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} \frac{\nu_2(X_i^{(1)})}{s_1\nu_1(X_i^{(1)}) + s_2\nu_2(X_i^{(1)})/\hat{d}_2^{(m)}}}{\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2} \frac{\nu_1(X_i^{(2)})}{s_1\nu_1(X_i^{(2)}) + s_2\nu_2(X_i^{(2)})/\hat{d}_2^{(m)}}}.$$

They show that $\lim_{m\to\infty} \hat{d}_2^{(m)}$ exists and is exactly equal to the estimate considered by Geyer (1994), and so established an equivalence between the iterative bridge estimator and the estimate based on maximization of the log quasi-likelihood function.

When the sequences $X_1^{(l)}, \dots, X_{n_l}^{(l)}$, $l = 1, 2$ are Markov chains, the optimal bridge function has the form $\beta_{\text{opt,mcmc}}(x) = \beta_*(x)\beta_{\text{opt,iid}}(x)$, where the correction factor, $\beta_*(x)$, is the solution to a complicated Fredholm integral equation (Romero, 2003) and reflects the dependence structure of the two chains. In particular, for the case of Markov chains, the optimal bridge function need not have the form

$$\beta(x) = \left[t_1\nu_1(x) + t_2\nu_2(x)\right]^{-1}, \tag{3.2}$$

for any t_1, t_2. Unfortunately, β_* is very hard to identify, let alone estimate. To conclude, since our procedure is, effectively, searching within the class (3.2), it will not yield an optimal estimate in general, and instead should be viewed as a method for yielding estimates which are practically useful, even if not optimal.

2. A crude way to find $\hat{\boldsymbol{a}}_{\text{opt}} := \operatorname{argmin}_{\boldsymbol{a}} \operatorname{trace}(\widehat{W}_{\boldsymbol{a}})$ is to calculate $\operatorname{trace}(\widehat{W}_{\boldsymbol{a}})$ as $\boldsymbol{a}$ varies over a grid in $\mathcal{S}_k$ and then find the minimizing $\boldsymbol{a}$. This is inefficient and unnecessary, as there exist efficient algorithms for minimizing real-valued functions of several variables; see, e.g., Robert and Casella (2004, Chapter 5).

3. The vector $\hat{\boldsymbol{a}}_{\text{opt}}$ can be calculated from a small pilot experiment, after which new chains are run and used to form the log quasi-likelihood function $\ell_n^{(\hat{\boldsymbol{a}}_{\text{opt}})}(\boldsymbol{\zeta})$, from which we obtain $\hat{\boldsymbol{\zeta}}$ (and hence $\hat{\boldsymbol{d}}$).

4. If for each l, $X_1^{(l)}, \ldots, X_{n_l}^{(l)}$ is an iid sequence, then a regeneration occurs at each step. In this case, there is no need to estimate $\boldsymbol{a}$, since the optimal value is known to be $a_j = n_l/n$ (Meng and Wong, 1996). The w_l's in (2.9) reduce to 1, and the log quasi-likelihood function (2.10) reduces to exactly the log quasi-likelihood function used by Geyer (1994), so our estimate is exactly the estimate introduced by Vardi (1985), who worked in the iid setting.

4 Illustrations

Here we have two goals. In Section 4.1 we provide a simulation study to show the gains in efficiency that are possible if we use the method for choosing the weight vector $\boldsymbol{a}$ described in Section 3. Our illustration involves toy problems. The purpose of Section 4.2 is to demonstrate the applicability of our methodology, and we return to the second of the three classes of problems we discussed in Section 1, where we have a family of probability densities of the form $p_\theta(x) = g_\theta(x)/z_\theta$, which are intractable because the normalizing constant z_θ cannot be computed in closed form. Our focus here is a bit different, in that we are not interested in estimating the family z_θ, $\theta \in \Theta$; rather, we are now interested in estimating a family of expectations of the form $E_\theta(U(X))$, $\theta \in \Theta$, where U is a function, as well as estimating functions of these expectations. Our illustration is in the context of the Ising model of statistical physics, and we show how to estimate the internal energy and specific heat of the system as a function of temperature.

4.1 Gains in Efficiency When Using the Optimal Weight Vector $\boldsymbol{a}$

Recall that $\hat{\boldsymbol{a}}_{\text{opt}} = \operatorname{argmin}_{\boldsymbol{a}} \operatorname{trace}(\widehat{W}_{\boldsymbol{a}})$ is calculated from a small pilot experiment. Let $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ be the corresponding estimate of $\boldsymbol{d}$. Also, let $\hat{\boldsymbol{d}}_{\text{conv}}$ denote the estimate of $\boldsymbol{d}$ obtained when we use the conventional choice $a_j = n_j/n$. In this section we demonstrate through a simulation study that significant gains in efficiency are possible if we use $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ instead of $\hat{\boldsymbol{d}}_{\text{conv}}$ in situations where the Markov chains mix at different rates. We consider a very simple situation where $k = 2$, so that $\boldsymbol{d}$ is just d_2. We take π_1 and π_2 to be two t distributions, specifically $\pi_1 = t_{5,1}$ and $\pi_2 = t_{5,0}$, where $t_{r,\mu}$ denotes the t distribution with r degrees of freedom, centered at μ. The representation $\pi_l = \nu_l / m_l$ is taken to be trivial: $\nu_l = \pi_l$ and $m_l = 1$ for $l = 1, 2$. So $d_2 = m_2/m_1$ is known to be 1, but we proceed to estimate it as if we didn't know that fact.

In our simulations, chain 1 is an iid sequence from π_1. Chain 2 is an independence Metropolis-Hastings (IMH) chain with proposal density $t_{5,\mu}$. That is, if the current state of the chain is x, a proposal $Y \sim t_{5,\mu}$ is generated; the chain moves to Y with acceptance probability $\min\bigl\{[t_{5,0}(Y)t_{5,\mu}(x)]/[t_{5,0}(x)t_{5,\mu}(Y)], 1\bigr\}$, and stays at x with the remaining probability. We will let μ range over a fine grid in $(-3, 3)$. Note that when $\mu = 0$, the proposal is always accepted, and the chain is an iid sequence from $t_{5,0}$, but as μ moves away from 0 in either direction, proposals are less likely to be accepted, and the mixing rate of the chain is slower. It is simple to check that $\inf_x \bigl(t_{5,\mu}(x)/t_{5,0}(x)\bigr) > 0$, which implies that the IMH algorithm is uniformly ergodic (Mengersen and Tweedie, 1996, Theorem 2.1) and hence geometrically ergodic. Moreover, Mykland et al. (1995, Section 4.1) have shown that for IMH chains there is always a scheme for producing minorization conditions and regeneration sequences, and here we use the scheme they described.

Our simulation study is carried out as follows. For each value of μ, we conduct a pilot study to calculate $\hat{\boldsymbol{a}}_{\text{opt}}$, using the method described in Section 3. The pilot study is based on 1000 iid draws from π_1 and a number of regenerations of the IMH Markov chain for π_2 that gives a sample of approximately the same size. Then we run the main study, in which we form $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ (where $\hat{\boldsymbol{a}}_{\text{opt}}$ is obtained in the pilot study), and also form $\hat{\boldsymbol{d}}_{\text{conv}}$. The main study is 10 times as large as the pilot study. For each μ, the above is replicated 500 times, and from these replicates we form the sample variance of the $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$'s, the sample variance of the $\hat{\boldsymbol{d}}_{\text{conv}}$'s, and form the ratio, which we take as a measure of the efficiency of $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ vs. $\hat{\boldsymbol{d}}_{\text{conv}}$.

Figure 1 gives a plot of the estimate of $\operatorname{Var}(\hat{\boldsymbol{d}}_{\text{conv}}) / \operatorname{Var}(\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}})$ as μ varies over $(-3, 3)$, along with 95% confidence bands, valid pointwise (the bands are constructed via the delta method applied to the function $f(o, c) = o/c$). From the figure we see that, as expected, the efficiency is about 1 when μ is near 0. But it grows rapidly as μ moves away from 0 in either direction, reaching about 17 when μ is 3 or -3, and it is reasonable to believe that the efficiency is unbounded as $\mu \to \infty$ or $\mu \to -\infty$. Figure 2 provides a graphical description of the explanation. The figure gives a plot of $[\hat{\boldsymbol{a}}_{\text{opt}}]_1$, the first component of $\hat{\boldsymbol{a}}_{\text{opt}}$, as μ varies over $(-3, 3)$. When $\mu = 0$, the two chains are each iid sequences, and $\hat{\boldsymbol{a}}_{\text{opt}} \doteq (.5, .5)$, so that $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}} \doteq \hat{\boldsymbol{d}}_{\text{conv}}$. But when μ moves away from 0 in either direction, chain 2 mixes more slowly, and $[\hat{\boldsymbol{a}}_{\text{opt}}]_1$ increases towards 1, so that in the term (2.2) in our quasi-likelihood function, less weight is given to chain 2, which is why $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ is more efficient than is $\hat{\boldsymbol{d}}_{\text{conv}}$.

Of course, because the calculation of $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ requires a pilot study, the comparison above could be viewed as unfair. However, for $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ to perform well all that is required, both in theory and in practice, is that $\hat{\boldsymbol{a}}_{\text{opt}}$ consistently estimate $\operatorname{argmin}_{\boldsymbol{a}} \operatorname{Var}(\hat{\boldsymbol{d}}_{\boldsymbol{a}})$, and for this to occur all that is required is that the size of the pilot study increase to infinity. That is, the size of the pilot study can increase to infinity arbitrarily slowly when compared to the size of the main study so, asymp-

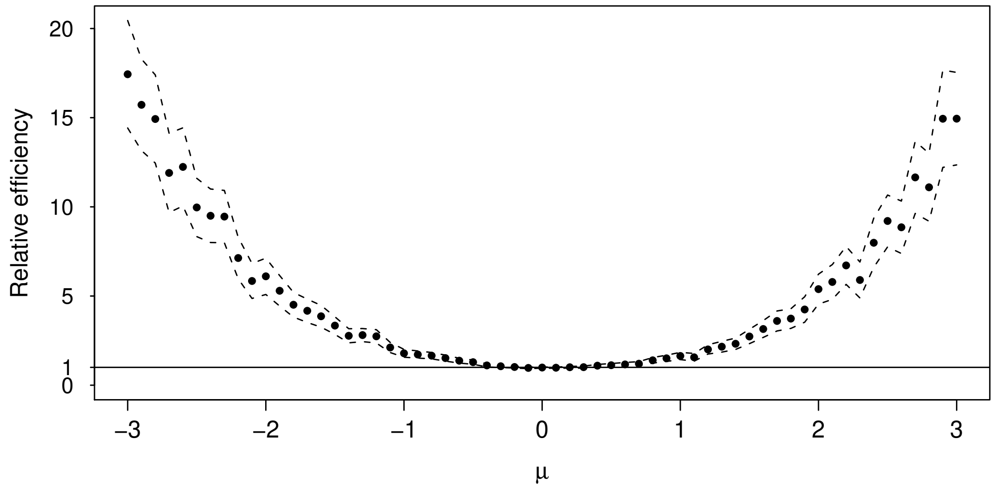

Figure 1: Estimated relative efficiency of $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ vs. $\hat{\boldsymbol{d}}_{\text{conv}}$, together with 95% confidence bands. As μ moves away from 0, the mixing rate of chain 2 slows, and the efficiency of $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ vs. $\hat{\boldsymbol{d}}_{\text{conv}}$ increases. The horizontal line at height 1 serves a reference line.

totically, the amount of time to compute $\hat{\boldsymbol{d}}_{\hat{\boldsymbol{a}}_{\text{opt}}}$ and $\hat{\boldsymbol{d}}_{\text{conv}}$ is the same.

4.2 Estimation of the Internal Energy and Specific Heat as Functions of Temperature in the Ising Model

We consider the Ising model on a $c \times c$ square lattice with periodic boundary conditions. That is, we have a graph (V, E) where V denotes the set of c^2 vertices of the lattice, and E denotes the set of $2c^2$ edges that connect nearest neighbors on the lattice. Vertices in the first and last rows are also considered neighbors, as are vertices in the first and last columns, so the graph resides on the torus. For each vertex $i \in V$, we have a random variable X_i taking on the values 1 and -1. The random vector $X = \{X_i,\, i \in V\}$ gives the state of the system, and the state space S contains 2^{c^2} states. For $x \in S$, let $H(x) = -\sum_{i \sim j} x_i x_j$, where the notation $i \sim j$ signifies that i and j are nearest neighbors. For each $\theta \in \Theta := [0, \infty)$, define a probability distribution p_θ on S by

$$p_\theta(x) = z_\theta^{-1} \exp[-\theta H(x)], \qquad x \in S,$$

where $z_\theta = \sum_{x \in S} \exp[-\theta H(x)]$ is the normalizing constant, called the partition function in the physics literature, and $\theta = 1/(\kappa T)$, where T is the temperature and κ is the Boltzmann constant. See, e.g., Newman and Barkema (1999, sec. 1.2) for an overview.

Important to physicists are the internal energy of the system, defined by

$$I_\theta = E_{p_\theta}[H(X)], \qquad \theta \in \Theta,$$

and the specific heat, which is the derivative of the internal energy with respect to temperature, or equivalently,

$$C_\theta = -\kappa\theta^2 \frac{\partial I_\theta}{\partial \theta} = \kappa\theta^2 \big\{ E_{p_\theta}[H^2(X)] - \big(E_{p_\theta}[H(X)]\big)^2 \big\}, \qquad \theta \in \Theta,$$

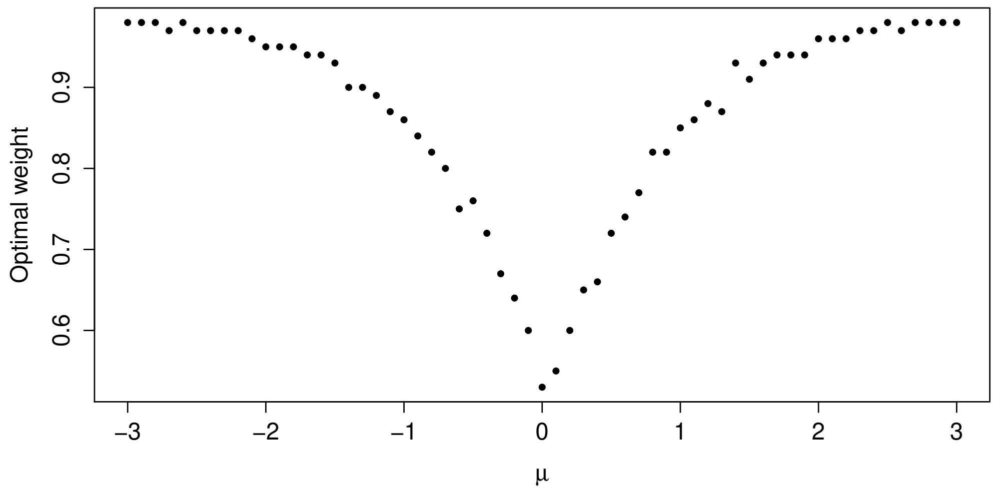

Figure 2: The points are the medians of the first component of $\hat{\boldsymbol{a}}_{\text{opt}}$, i.e. the weight assigned to sample 1 in the term (2.2) in our quasi-likelihood function, over the 500 replications at each μ. As μ moves away from 0, the weight given to the second (slower mixing) chain decreases to 0.

and interest is focused on how these quantities vary with θ. Because the size of the state space increases very rapidly as c increases, except for the case $c \leq 5$, the quantities above cannot be evaluated, and MCMC must be used. It is simple to implement a Metropolis-Hastings algorithm that randomly chooses a site, proposes to flip its spin, and accepts this proposal with the Metropolis-Hastings probability; however this algorithm converges very slowly. Swendsen and Wang (1987) proposed a data augmentation algorithm in which bond variables are introduced: if i and j are nearest neighbors and $X_i = X_j$, then with probability $1 - \exp(-\theta)$ an edge is placed between vertices i and j. This partitions the state space into connected components, and entire components are flipped. This algorithm converges far more rapidly than the single-site updating algorithm, and it is the algorithm we use here. Mykland et al. (1995, sec. 5.3) developed a simple minorization condition for the Swendsen-Wang algorithm, and we use it here to produce the regenerative chains that are needed to estimate the families $\{I_\theta,\, \theta \in \Theta\}$ and $\{C_\theta,\, \theta \in \Theta\}$ via the methods of this paper.

We now consider the problem of estimating the families $\{I_\theta,\, \theta \in \Theta\}$ and $\{C_\theta,\, \theta \in \Theta\}$, and as we will see, the issue of obtaining standard errors for our estimates is quite important. We are in the framework of the second of the three classes of problems mentioned in Section 1, and the two-step procedure given there, described in the present context, is as follows:

Step 1 We choose points $\theta_1, \dots, \theta_k$ appropriately spread out in the region of Θ of interest, and for $l = 1, \dots, k$, we run a Swendsen-Wang chain with invariant distribution p_{θ_l} for ρ_l regenerations. Using these k chains, we form $\hat{\boldsymbol{d}}$, the estimate of the vector $\boldsymbol{d}$, where $d_l = z_{\theta_l}/z_{\theta_1}$, $l = 2, \dots, k$.

Step 2 For each $l = 1, \dots, k$, we generate a new Swendsen-Wang chain with invariant distribution p_{θ_l} for R_l regenerations, and we use these new chains, together with the estimate $\hat{\boldsymbol{d}}$ produced in Step 1, to estimate I_θ and C_θ.

We now describe the details involved in Step 2. Denote the l^{th} sample (in Step 2) by $\{X_i^{(l)}, i = 1, \ldots, n_l\}$. For each $\theta \in \Theta$, define $g_\theta(x) = \exp[-\theta H(x)]$ for $x \in S$. Let

$$u(x) = \frac{g_\theta(x)}{\sum_{s=1}^k g_{\theta_s}(x)}, \quad v(x) = H(x)u(x), \quad \text{and} \quad z(x) = H^2(x)u(x),$$

and let

$$\hat{u}_n = \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} u(X_i^{(l)}), \quad \hat{v}_n = \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} v(X_i^{(l)}), \quad \text{and} \quad \hat{z}_n = \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} z(X_i^{(l)}).$$

(These quantities depend on θ, but this dependence is temporarily suppressed in the notation.) Using E_l to denote expectation with respect to p_{θ_l}, we have

$$\hat{I}_\theta := \frac{\hat{v}_n}{\hat{u}_n} \xrightarrow{\text{a.s.}} \frac{\sum_{l=1}^k d_l E_l(v(X))}{\sum_{l=1}^k d_l E_l(u(X))} = \frac{(z_\theta/z_{\theta_1}) \sum_{x \in S} H(x) p_\theta(x)}{(z_\theta/z_{\theta_1}) \sum_{x \in S} p_\theta(x)} = I_\theta$$

as $\rho_l \to \infty$ and $R_l \to \infty$ for $l = 1, \ldots, k$, where the convergence statement follows from ergodicity of the Swendsen-Wang chains and the fact that $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$. Similarly, we have

$$\widehat{C}_\theta := \kappa\theta^2 \left(\frac{\hat{z}_n}{\hat{u}_n} - \left(\frac{\hat{v}_n}{\hat{u}_n} \right)^2 \right) \xrightarrow{\text{a.s.}} C_\theta.$$

Furthermore, Theorem 2 of Tan, Doss and Hobert (2012) deals precisely with the asymptotic distribution of estimates of the form $\hat{I}_\theta$ and $\widehat{C}_\theta$, in the framework of regenerative Markov chains. This theorem, which relies on Theorem 1 of the present paper, states that if (i) both Stage 1 and Stage 2 chains satisfy A1–A3 of the present paper, (ii) for $l = 1, \ldots, k$ R_l/R_1 and ρ_l/ρ_1 converge to positive finite constants, and (iii) R_1/ρ_1 converges to a nonnegative finite constant, then $R_1^{1/2}(\hat{I}_\theta - I_\theta)$ and $R_1^{1/2}(\widehat{C}_\theta - C_\theta)$ have asymptotically normal distributions, and the theorem also provides regeneration-based consistent estimates of the asymptotic variances. These are the estimates we use in this section.

We will apply the approach described above in two situations. The first involves the Ising model on a square lattice small enough so that exact calculations can be done. This enables us to check the performance of our estimators and confidence intervals. The second involves the Ising model on a larger lattice, where calculations can be done only through Monte Carlo methods.

We first consider the Ising model on a 5×5 lattice, and we focus on the problem of estimating C_θ, the specific heat. Figure 3 was created using our methods. The left panel gives a plot of $\widehat{C}_\theta$, together with 95% confidence bands (valid pointwise), and a plot of the exact values. The right panel gives the standard error estimates for $\widehat{C}_\theta$. To create the figure, we used the approach described above, with $k = 5$ and $(\theta_1, \ldots, \theta_5) = (.3, .4, .5, .6, .7)$. For each $l = 1, \ldots, 5$, regenerative Swendsen-Wang chains of (approximate) length 10,000 were run for θ_l, based on which $\hat{\boldsymbol{d}}$ and $\widehat{W}$ from Theorem 1 were calculated. We then ran independent chains for the same five θ values, for as many iterations, to form estimates $\widehat{C}_\theta$ on a fine grid of θ values that range from .2 to 1 in increments of .01. The plot in the right panel was obtained from the formula in Theorem 2 of

Tan et al. (2012), and the exact values of C_θ were obtained using closed-form expressions from the physics literature.

We mention that Newman and Barkema (1999, sec. 3.7) also considered the problem of estimating the specific heat for the Ising model on a 5×5 lattice. They have a plot very similar to ours, but they produced it by running a separate Swendsen-Wang chain for each θ value on a fine grid, and each chain is used solely for the θ value under which it was generated. In contrast, our method requires only k Swendsen-Wang chains, where k is fairly small, and all chains are used to estimate C_θ for every θ. Here, we have considered a simple instance of the Ising model, the so-called one-parameter case. It is common to also consider the situation where there is an external magnetic field, in which case θ has dimension 2, and $p_\theta(x) \propto \exp\bigl(\theta_1 \sum_{i\sim j} x_i x_j + \theta_2 \sum_{i \in V} x_i\bigr)$. Running a separate Swendsen-Wang chain for each θ in a fine subgrid in dimension 2 becomes extremely time consuming, whereas our approach is easily still workable.

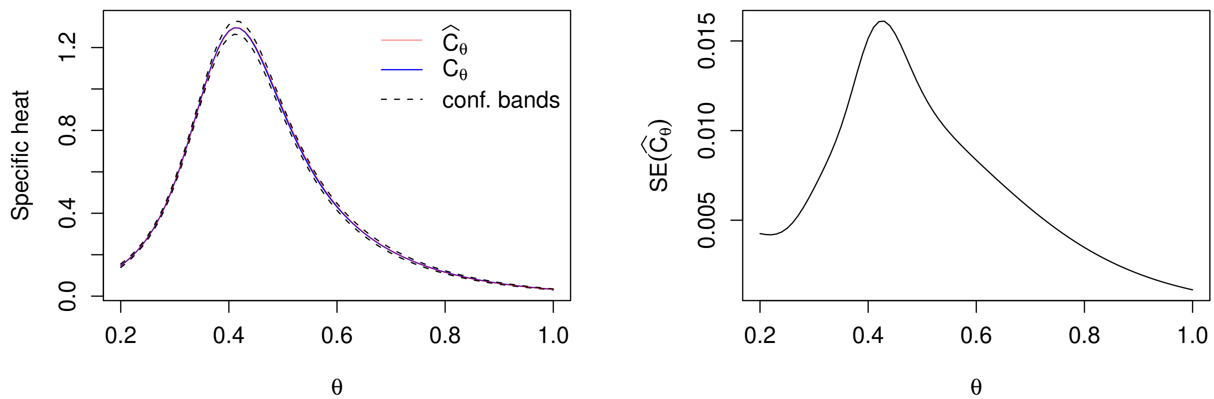

Figure 3: Estimation of the specific heat for the Ising model on a 5×5 lattice. Left panel gives a plot of the point estimates and a plot of the exact values, as θ varies. The two plots are visually indistinguishable. Also provided are 95% confidence bands. Right panel gives standard errors for $\widehat{C}_\theta$.

In our second example, we consider the Ising model on a 30×30 lattice, for which exact calculations of physical quantities are prohibitively expensive, and our interest is now on estimating the internal energy. The left panel of Figure 4 shows a plot of $\hat{I}_\theta$ vs. θ as θ ranges from $.35$ to 1.5 in increments of $.01$. To form the plot we carried out the two-step procedure discussed earlier, with $k = 5$ and reference points $(\theta_1, \dots, \theta_5) = (.65, .75, .85, .95, 1.05)$, and a sample size of $100{,}000$ for each chain in both steps. The left panel also shows 95% bands, valid pointwise, and the right panel shows the estimated standard errors. From the plot, we can see that the standard errors are much larger when $\theta < \theta_1 = .65$ than they are when $\theta \geq \theta_1$. The importance sampling estimates are not stable when we try to extrapolate below the lowest reference θ value, but we can go well above the highest reference value and still get accurate estimates. It is our ability to estimate SE's through regeneration that makes it possible for us to determine the range of θ's for which we have reliable estimates. In fact, this range depends in a complicated way on the reference points and the sample sizes, and even for the relatively simple case where $k = 1$, the range is not simply an interval centered at θ_1.

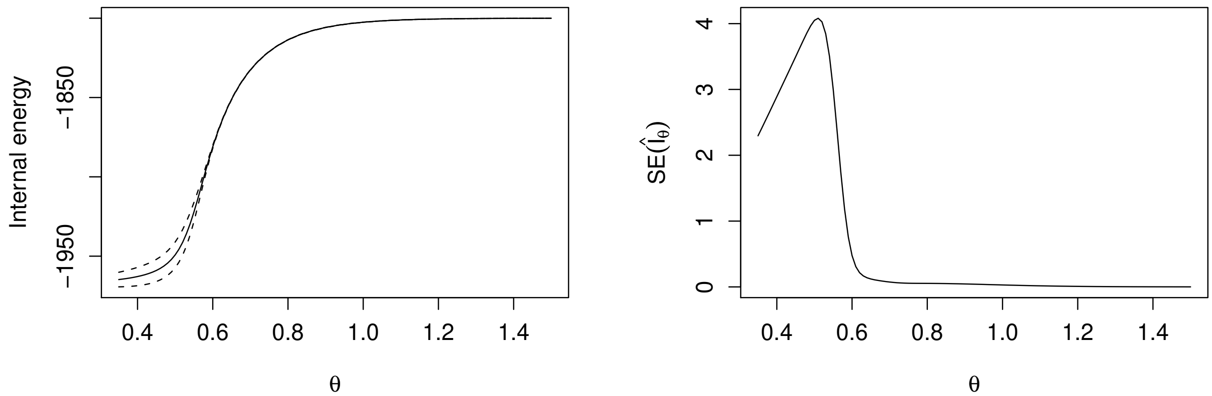

Figure 4: Estimation of the internal energy for the Ising model on a 30×30 lattice. Left panel gives estimated values, together with 95% confidence bands. Right panel gives the corresponding standard error estimates.

Appendix: Proof of Theorem 1

Proof of Consistency of $\hat{\boldsymbol{d}}$

We first work in the $\boldsymbol{\zeta}$ domain, and at the very end switch to the $\boldsymbol{d}$ domain. As mentioned earlier, in the standard textbook situation in which we have $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$ where $\theta_0 \in \Theta$, $l_n(\theta)$ is the log likelihood and $Q(\theta) = E_{\theta_0}(l_1(\theta))$, the classical proof of consistency (Wald, 1949) is based on the observation that $Q(\theta)$ is maximized at $\theta = \theta_0$, and that for each fixed θ, $l_n(\theta) \xrightarrow{\text{a.s.}} Q(\theta)$. The convergence may be non-uniform, and care needs to be exercised in showing that the maximizer of $l_n(\theta)$ converges to the maximizer of $Q(\theta)$. The present situation is simpler in that the log likelihood and its expected value are twice differentiable and concave, but is more complicated in that we have multiple sequences, they are not iid, and we have a non-identifiability issue, so that maximization is carried out subject to a constraint.

We will write ℓ_ρ instead of ℓ_n to remind ourselves that the ρ_l's are given and the n_l's are determined by these ρ_l's. Also, we will write $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ instead of $\ell_\rho(\boldsymbol{\zeta})$ when we need to note the dependence of $\ell_\rho(\boldsymbol{\zeta})$ on $\boldsymbol{X}$, where $\boldsymbol{X} = \big(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, \ldots, X_1^{(k)}, \ldots, X_{n_k}^{(k)}\big)$. We define the (scaled) expected log quasi-likelihood by

$$\lambda(\boldsymbol{\zeta}) = \sum_{l=1}^{k} a_l E_{\pi_l}\big(\log[p_l(X, \boldsymbol{\zeta})]\big).$$

As $\rho_l \to \infty$, we have $n_l \to \infty$, so $n_l^{-1} \sum_{i=1}^{n_l} \log\big(p_l(X_i^{(l)}, \boldsymbol{\zeta})\big) \xrightarrow{\text{a.s.}} E_{\pi_l}\big(\log[p_l(X, \boldsymbol{\zeta})]\big)$, and so

$$n^{-1}\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta}) \xrightarrow{\text{a.s.}} \lambda(\boldsymbol{\zeta}) \qquad \text{for all } \boldsymbol{\zeta}.$$

The structure of our proof is similar to that of Theorem 1 of Geyer (1994), and the outline of our proof is as follows. First, define $S = \{\boldsymbol{\zeta} : \boldsymbol{\zeta}^\top 1_k = 0\}$, and recall that $\hat{\boldsymbol{\zeta}}$ is defined to be a maximizer of $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ satisfying $\hat{\boldsymbol{\zeta}} \in S$.

1. We will show that for every $\boldsymbol{X}$, $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ is everywhere twice differentiable and concave in $\boldsymbol{\zeta}$.

2. We will show that $\lambda(\boldsymbol{\zeta})$ is finite, everywhere twice differentiable, and concave. We further show that its Hessian matrix is semi-negative definite, and that its only null eigenvector is 1_k.
3. We will show that $\nabla\lambda(\boldsymbol{\zeta}_0) = 0$.
4. We will note that the two steps above imply that $\boldsymbol{\zeta}_0$ is the unique maximizer of λ subject to the condition $\boldsymbol{\zeta}_0 \in S$.
5. We will argue that with probability one, for every $\boldsymbol{\zeta}$, $\nabla^2\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ is semi-negative definite, and 1_k is its only null eigenvector. This will show that $\hat{\boldsymbol{\zeta}}$ is the unique maximizer of $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ subject to $\hat{\boldsymbol{\zeta}} \in S$.
6. We will conclude that the convergence of $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ to $\lambda(\boldsymbol{\zeta})$ implies convergence of their maximizers that reside in S, that is, $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$.

We now provide the details.

1. The differentiability is immediate from the definition of ℓ_ρ (see (2.10)). To show concavity, it is sufficient to show that for every x, $\log\bigl(p_l(x, \boldsymbol{\zeta})\bigr)$ is concave in $\boldsymbol{\zeta}$. Now

$$\frac{\partial^2 \log\bigl(p_l(x, \boldsymbol{\zeta})\bigr)}{\partial \boldsymbol{\zeta}^2} = -(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top), \tag{A1}$$

where $\boldsymbol{p} = \bigl(p_1(x, \boldsymbol{\zeta}), \dots, p_k(x, \boldsymbol{\zeta})\bigr)^\top$. The matrix inside the parentheses on the right side of (A1) is the covariance matrix for the multinomial distribution with parameter $\boldsymbol{p}$, so this matrix is positive semi-definite.

2. First, $\lambda(\boldsymbol{\zeta})$ is finite because $\lambda(\boldsymbol{\zeta}) \le 0$, and

$$\begin{aligned}
-\lambda(\boldsymbol{\zeta}) &= \sum_{l=1}^k a_l E_{\pi_l}\left[\log\left(\frac{1}{p_l(X, \boldsymbol{\zeta})}\right)\right] \\
&= \sum_{l=1}^k a_l E_{\pi_l}\left[\log\left(1 + \sum_{s \ne l} \frac{\nu_s(X)}{\nu_l(X)} e^{\zeta_s - \zeta_l}\right)\right] \\
&\le \sum_{l=1}^k a_l E_{\pi_l}\left(\sum_{s \ne l} \frac{\nu_s(X)}{\nu_l(X)} e^{\zeta_s - \zeta_l}\right) \qquad (\log(1+a) < a \text{ for } a > 0) \\
&\le \sum_{l=1}^k a_l \sum_{s \ne l} e^{\zeta_s - \zeta_l} \int \frac{\nu_s(x)}{\nu_l(x)} \pi_l(x)\, \mu(dx) \\
&= \sum_{l=1}^k a_l \sum_{s \ne l} e^{\zeta_s - \zeta_l} \frac{m_s}{m_l} \int \frac{\pi_s(x)}{\pi_l(x)} \pi_l(x)\, \mu(dx) < \infty.
\end{aligned}$$

We now obtain the first and second derivatives of λ. By a standard argument involving the dominated convergence theorem, we can interchange the order of differentiation and integration. (If v is the vector of length k with a 1 in the r^{th} position and 0's everywhere else, then for any x, any $\boldsymbol{\zeta}$, and any $l \in \{1, \dots, k\}$, $\bigl[\log\bigl(p_l(x, \boldsymbol{\zeta} + v/m)\bigr) - \log\bigl(p_l(x, \boldsymbol{\zeta})\bigr)\bigr]m =$

$\partial \log\big(p_l(x, \boldsymbol{\zeta}_*)\big)/\partial \zeta_r$, where $\boldsymbol{\zeta}_*$ is between $\boldsymbol{\zeta} + v/m$ and $\boldsymbol{\zeta}$, and this partial derivative is uniformly bounded between -1 and 1.) So for $r = 1, \dots, k$, we have

$$\frac{\partial \lambda(\boldsymbol{\zeta})}{\partial \zeta_r} = \sum_{l=1}^{k} a_l E_{\pi_l}\left(\frac{\partial \log\big(p_l(X, \boldsymbol{\zeta})\big)}{\partial \zeta_r}\right) = a_r - \sum_{l=1}^{k} a_l E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta})\big). \tag{A2}$$

Consider the integrand on the right side of (A2), i.e. $p_r(X, \boldsymbol{\zeta})$. Its gradient is given by $\partial p_r/\partial \zeta_r = p_r - p_r^2$ and $\partial p_r/\partial \zeta_l = -p_r p_l$ for $l \neq r$, and these derivatives are uniformly bounded in absolute value by 1. Hence again by the dominated convergence theorem, we can interchange the order of differentiation and integration, and doing this gives

$$\begin{aligned}
-\frac{\partial^2 \lambda(\boldsymbol{\zeta})}{\partial \zeta_r^2} &= \sum_{l=1}^{k} a_l E_{\pi_l}\left(\frac{\partial p_r(X, \boldsymbol{\zeta})}{\partial \zeta_r}\right) = \sum_{l=1}^{k} a_l E_{\pi_l}[p_r(X, \boldsymbol{\zeta}) - p_r^2(X, \boldsymbol{\zeta})] \\
-\frac{\partial^2 \lambda(\boldsymbol{\zeta})}{\partial \zeta_s \partial \zeta_r} &= \sum_{l=1}^{k} a_l E_{\pi_l}\left(\frac{\partial p_r(X, \boldsymbol{\zeta})}{\partial \zeta_s}\right) = \sum_{l=1}^{k} a_l E_{\pi_l}[-p_r(X, \boldsymbol{\zeta}) p_s(X, \boldsymbol{\zeta})] \quad \text{for } s \neq r.
\end{aligned} \tag{A3}$$

Define the expectation operator $E_P = \sum_{l=1}^{k} a_l E_l$. From (A3) we have $-\nabla^2 \lambda(\boldsymbol{\zeta}) = E_P(J)$, where $J = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$, and as before $\boldsymbol{p} = \big(p_1(X, \boldsymbol{\zeta}), \dots, p_k(X, \boldsymbol{\zeta})\big)^\top$. As before, J is the covariance of the multinomial, so is positive semi-definite, and therefore so is $E_P(J)$.

We now determine the null eigenvectors of $\nabla^2 \lambda(\boldsymbol{\zeta})$ (which is $-E_P(J)$). If $\nabla^2 \lambda(\boldsymbol{\zeta}) u = 0$, then $u^\top [\nabla^2 \lambda(\boldsymbol{\zeta})] u = 0$, so $E_P(u^\top J u) = 0$. Since J is positive semi-definite, it has a square root, $J^{1/2}$. Hence $E_P(\|J^{1/2} u\|^2) = 0$, which implies $Ju = 0$ $[P]$-a.e. The condition $Ju = 0$ $[P]$-a.e. is expressed as

$$p_r(X, \boldsymbol{\zeta}) \big(\textstyle\sum_{l=1}^{k} p_l(X, \boldsymbol{\zeta}) u_l - u_r \big) = 0 \qquad [P]\text{-a.e. for } r = 1, \dots, k, \tag{A4}$$

and under our assumption that $\nu_1, \dots, \nu_k$ are mutually absolutely continuous, (A4) implies that $u_r = \sum_{l=1}^{k} p_l(X, \boldsymbol{\zeta}) u_l$ for $r = 1, \dots, k$. So $u_1 = \cdots = u_k$, i.e. $u \propto 1_k$.

3. To show that $\nabla \lambda(\boldsymbol{\zeta}_0) = 0$, we write

$$\begin{aligned}
\left.\frac{\partial \lambda(\boldsymbol{\zeta})}{\partial \zeta_r}\right|_{\boldsymbol{\zeta}_0} &= a_r - \sum_{l=1}^{k} a_l \int \frac{\nu_r(x) a_r/m_r}{\sum_{s=1}^{k} \nu_s(x) a_s/m_s} \pi_l(x)\, \mu(dx) \\
&= a_r - \int \frac{\sum_{l=1}^{k} a_l \pi_l(x)}{\sum_{s=1}^{k} a_s \nu_s(x)/m_s} \nu_r(x) a_r/m_r\, \mu(dx) \\
&= a_r - a_r \int \pi_r(x)\, \mu(dx) = 0.
\end{aligned}$$

4. For any $\boldsymbol{\zeta}$ satisfying $\boldsymbol{\zeta}^\top 1_k = 0$, we may write

$$\begin{aligned}
\lambda(\boldsymbol{\zeta}) &= \lambda(\boldsymbol{\zeta}_0) + (\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^\top \nabla \lambda(\boldsymbol{\zeta}_0) + \frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^\top \nabla^2 \lambda(\boldsymbol{\zeta}_*)(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0) \\
&= \lambda(\boldsymbol{\zeta}_0) + \frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^\top \nabla^2 \lambda(\boldsymbol{\zeta}_*)(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0),
\end{aligned}$$

where $\boldsymbol{\zeta}_*$ is between $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}_0$. If $\boldsymbol{\zeta} \neq \boldsymbol{\zeta}_0$, i.e. $\boldsymbol{\zeta} - \boldsymbol{\zeta}_0 \neq 0$, then since $(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^\top 1_k = 0$, $\boldsymbol{\zeta} - \boldsymbol{\zeta}_0$ cannot be a scalar multiple of 1_k. Hence by Step 2, $(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^\top \nabla^2 \lambda(\boldsymbol{\zeta}_*)(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0) < 0$.

5. Clearly $\nabla \ell_\rho(\boldsymbol{X}, \hat{\boldsymbol{\zeta}}) = 0$. The proof that (i) $\nabla^2 \ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ is semi-negative definite, (ii) the only null eigenvector of $\nabla^2 \ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ is 1_k, and (iii) $\hat{\boldsymbol{\zeta}}$ is the unique maximizer of $\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ subject to the constraint $\boldsymbol{\zeta} \in S$, is essentially identical to the proof of these assertions for $\lambda(\boldsymbol{\zeta})$.

6. Since $n^{-1}\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta}) \xrightarrow{\text{a.s.}} \lambda(\boldsymbol{\zeta})$ for each $\boldsymbol{\zeta}$, a.s. convergence occurs on a dense subset of S. Also, the functions involved are all concave in the entire space of $\boldsymbol{\zeta}$'s, hence are concave in S. Therefore, we have a.s. uniform convergence of $n^{-1}\ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta})$ to $\lambda(\boldsymbol{\zeta})$ on compact subsets of S. Under concavity, this is enough to imply $\operatorname{argmax}_{\boldsymbol{\zeta} \in S} \ell_\rho(\boldsymbol{X}, \boldsymbol{\zeta}) \xrightarrow{\text{a.s.}} \operatorname{argmax}_{\boldsymbol{\zeta} \in S} \lambda(\boldsymbol{\zeta})$, i.e. $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$.

Finally, to see that $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$, we write $\hat{\boldsymbol{d}} - \boldsymbol{d} = g(\hat{\boldsymbol{\zeta}}) - g(\boldsymbol{\zeta}_0) = \nabla g(\boldsymbol{\zeta}_*)^\top (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$, where $\boldsymbol{\zeta}_*$ is between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. The function g actually depends on $\boldsymbol{a}^{(\rho)}$, so depends on ρ, but the gradient $\nabla g(\boldsymbol{\zeta}_*)$ is bounded for large ρ because $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ and $\boldsymbol{a}^{(\rho)} \to \boldsymbol{\alpha}$. Therefore $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$.

Proof of Regeneration-Based CLT for $\hat{\boldsymbol{d}}$

We begin by considering $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$. As in the classical proof of asymptotic normality of maximum likelihood estimators, we expand $\nabla \ell_\rho$ at $\hat{\boldsymbol{\zeta}}$ around $\boldsymbol{\zeta}_0$, and using the appropriate scaling factor, we get

$$-\frac{\rho_1^{1/2}}{n}\big(\nabla \ell_\rho(\hat{\boldsymbol{\zeta}}) - \nabla \ell_\rho(\boldsymbol{\zeta}_0)\big) = -\frac{1}{n}\nabla^2 \ell_\rho(\boldsymbol{\zeta}_*)\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0), \tag{A5}$$

where $\boldsymbol{\zeta}_*$ is between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Consider the left side of (A5), which is just $\rho_1^{1/2} n^{-1} \nabla \ell_\rho(\boldsymbol{\zeta}_0)$, since $\nabla \ell_\rho(\hat{\boldsymbol{\zeta}}) = 0$. There are several nontrivial components to the proof, so we first give an outline.

1. We show that each element of the vector $n^{-1}\nabla \ell_\rho(\boldsymbol{\zeta}_0)$ can be represented as a linear combination of mean 0 averages of functions of the k chains plus a vanishingly small term.
2. Using Step 1 above, we obtain a regeneration-based CLT for the scaled score vector, via a considerably more involved version of the method we used in Section 2.1: we show that $\rho_1^{1/2} n^{-1} \nabla \ell_\rho(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$, where Ω is given by (2.16).
3. We argue that $-n^{-1}\nabla^2 \ell_\rho(\boldsymbol{\zeta}_*) \xrightarrow{\text{a.s.}} B$ and that $\big(-n^{-1}\nabla^2 \ell_\rho(\boldsymbol{\zeta}_*)\big)^\dagger \xrightarrow{\text{a.s.}} B^\dagger$, where B is defined in (2.12), using ideas in Geyer (1994).
4. We conclude that $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, B^\dagger \Omega B^\dagger)$.
5. We note the relationships $\boldsymbol{d} = g(\boldsymbol{\zeta}_0)$ and $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$, where g was defined by (2.18), and apply the delta method to obtain the desired result.

We now provide the details.

1. We start by considering $n^{-1}\nabla \ell_\rho(\boldsymbol{\zeta}_0)$. For $r = 1, \dots, k$, we have

$$\begin{aligned}\frac{\partial \ell_\rho(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= w_r \sum_{i=1}^{n_r} \big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0)\big) - \sum_{\substack{l=1\\ l\neq r}}^{k} w_l \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) \\ &= w_r \sum_{i=1}^{n_r} \Big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0) - \big[1 - E_{\pi_r}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big]\Big) \\ &\qquad - \sum_{\substack{l=1\\ l\neq r}}^{k} w_l \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] + e,\end{aligned} \tag{A6}$$

where

$$e = w_r \sum_{i=1}^{n_r} \big[1 - E_{\pi_r}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] - \sum_{\substack{l=1\\ l\neq r}}^{k} w_l \sum_{i=1}^{n_l} E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big). \tag{A7}$$

We claim that $e = 0$. To see this, note that from (A7) we have

$$e = w_r n_r - \sum_{l=1}^{k} w_l n_l E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big) = w_r n_r - \sum_{l=1}^{k} w_l n_l \frac{a_r}{a_l} E_{\pi_r}\big(p_l(X, \boldsymbol{\zeta}_0)\big). \tag{A8}$$

The last equality in (A8) holds because

$$\begin{aligned}E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big) &= \int \frac{\nu_r(x) e^{[\zeta_0]_r}}{\sum_{s=1}^k \nu_s(x) e^{[\zeta_0]_s}} \pi_l(x)\, \mu(dx) = \int \frac{\nu_r(x) a_r/m_r}{\sum_{s=1}^k \nu_s(x) a_s/m_s} \pi_l(x)\, \mu(dx) \\ &= \int \frac{\nu_l(x) a_r/m_l}{\sum_{s=1}^k \nu_s(x) a_s/m_s} \pi_r(x)\, \mu(dx) = \frac{a_r}{a_l} E_{\pi_r}\big(p_l(X, \boldsymbol{\zeta}_0)\big).\end{aligned}$$

Therefore, using the fact that $w_l n_l a_r / a_l = w_r n_r$, we get

$$e = w_r n_r - w_r n_r \textstyle\sum_{l=1}^k E_{\pi_r}\big(p_l(X, \boldsymbol{\zeta}_0)\big) = w_r n_r - w_r n_r E_{\pi_r}\big(\sum_{l=1}^k p_l(X, \boldsymbol{\zeta}_0)\big) = 0.$$

We summarize: Because $e = 0$, (A6) can be used to view $n^{-1}\partial \ell_\rho(\boldsymbol{\zeta}_0)/\partial \zeta_r$ as a linear combination of mean 0 averages of functions of the k chains. To express these averages in terms of iid quantities, we first recall the definitions of $y_i^{(r,l)}(\boldsymbol{a})$, $Y_t^{(r,l)}(\boldsymbol{a})$, $\bar{Y}^{(r,l)}(\boldsymbol{a})$, and $\bar{T}^{(l)}$, given in (2.14) and (2.15), and multiplying by the scaling factor $\rho_1^{1/2} n^{-1}$, we rewrite (A6) as

$$\begin{aligned}
\frac{\rho_1^{1/2}}{n}\frac{\partial \ell_\rho(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= -\frac{\rho_1^{1/2}}{n}\sum_{l=1}^k w_l \sum_{i=1}^{n_l}\bigl[p_r(X_i^{(l)},\boldsymbol{\zeta}_0) - E_{\pi_l}\bigl(p_r(X,\boldsymbol{\zeta}_0)\bigr)\bigr] \\
&= -\sum_{l=1}^k \frac{\rho_1^{1/2} n_l}{n} w_l \frac{1}{n_l}\sum_{t=1}^{\rho_l} Y_t^{(r,l)}(\boldsymbol{a}) \\
&= -\sum_{l=1}^k \frac{\rho_1^{1/2} n_l}{n} w_l \frac{\sum_{t=1}^{\rho_l} Y_t^{(r,l)}(\boldsymbol{a})}{\sum_{t=1}^{\rho_l} T_t^{(l)}} \\
&= -\sum_{l=1}^k \left[\Bigl(\frac{\rho_1}{\rho_l}\Bigr)^{1/2} \frac{n_l}{n} w_l\right]\left[\rho_l^{1/2}\frac{\bar{Y}^{(r,l)}(\boldsymbol{a})}{\bar{T}^{(l)}}\right] \\
&= -\sum_{l=1}^k \left[\Bigl(\frac{\rho_1}{\rho_l}\Bigr)^{1/2} a_l\right]\left[\rho_l^{1/2}\frac{\bar{Y}^{(r,l)}(\boldsymbol{a})}{\bar{T}^{(l)}}\right]. \qquad \text{(A9)}
\end{aligned}$$

2. We now apply a more complex and more rigorous version of the argument we used in Section 2.1. We note the following: (i) the k chains are geometrically ergodic by Assumption A2; (ii) since $p_r(x,\boldsymbol{\zeta}) \in (0,1)$ for all x and all $\boldsymbol{\zeta}$, $E_{\pi_l}\bigl(|y_1^{(r,l)}(\boldsymbol{a})|^{2+\epsilon}\bigr) < \infty$ for some $\epsilon > 0$ (in fact for any $\epsilon > 0$); and (iii) by (2.14) the mean of $Y_t^{(r,l)}(\boldsymbol{a})$ is 0. The usual CLT for iid sequences does not apply to the sequence $Y_1^{(r,l)}(\boldsymbol{a}),\dots,Y_{\rho_l}^{(r,l)}(\boldsymbol{a})$ because $\boldsymbol{a} = \boldsymbol{a}^{(\rho)}$ is allowed to change with ρ, so the distribution of $Y_t^{(r,l)}(\boldsymbol{a})$ changes with ρ. Since r and l are now fixed and play no important role, while the dependence of $\boldsymbol{a}$ on ρ now needs to be noted we will write $y_i(\boldsymbol{a}^{(\rho)})$ instead of $y_i^{(r,l)}(\boldsymbol{a})$, $Y_t(\boldsymbol{a}^{(\rho)})$ instead of $Y_t^{(r,l)}(\boldsymbol{a})$, etc. We really have a triangular array of random variables, and we will apply the Lindeberg-Feller version of the CLT.

   We first need to show that $E\bigl([Y_t(\boldsymbol{a}^{(\rho)})]^2\bigr) < \infty$. (This condition is nontrivial because $Y_t(\boldsymbol{a}^{(\rho)})$ is the sum of a random number of terms.) Note that since $p_r(x,\boldsymbol{\zeta}) \in (0,1)$, $|y_i(\boldsymbol{a}^{(\rho)})| \le 1$, and therefore,

$$|Y_t(\boldsymbol{a}^{(\rho)})| \le T_t^{(l)}. \qquad \text{(A10)}$$

   Theorem 2 of Hobert, Jones, Presnell and Rosenthal (2002) states that $E\bigl[(T_t^{(l)})^2\bigr] < \infty$ under geometric ergodicity. So $E\bigl([Y_t(\boldsymbol{a}^{(\rho)})]^2\bigr) < \infty$, and we may form the triangular array whose ρ_l^{th} row consists of the variables $U_1(\boldsymbol{a}^{(\rho)}),\dots,U_{\rho_l}(\boldsymbol{a}^{(\rho)})$, where

$$U_t(\boldsymbol{a}^{(\rho)}) = \frac{Y_t(\boldsymbol{a}^{(\rho)})}{\bigl(\sum_{s=1}^{\rho_l}\operatorname{Var}[Y_s(\boldsymbol{a}^{(\rho)})]\bigr)^{1/2}}.$$

   Clearly, $E[U_t(\boldsymbol{a}^{(\rho)})] = 0$ and $\sum_{t=1}^{\rho_l}\operatorname{Var}[U_t(\boldsymbol{a}^{(\rho)})] = 1$.

   The Lindeberg Condition is that for every $\eta > 0$,

$$\sum_{t=1}^{\rho_l} E\bigl([U_t(\boldsymbol{a}^{(\rho)})]^2 I(|U_t(\boldsymbol{a}^{(\rho)})| > \eta)\bigr) \xrightarrow{\text{a.s.}} 0 \qquad \text{as } \rho_l \to \infty,$$

and this is equivalent to the condition

$$E\left[\frac{[Y_1(\boldsymbol{a}^{(\rho)})]^2}{\operatorname{Var}[Y_1(\boldsymbol{a}^{(\rho)})]} I\left(\frac{|Y_1(\boldsymbol{a}^{(\rho)})|}{\left(\rho_l \operatorname{Var}[Y_1(\boldsymbol{a}^{(\rho)})]\right)^{1/2}} > \eta\right)\right] \to 0 \qquad \text{as } \rho_l \to \infty. \tag{A11}$$

To see (A11) note that as $\rho_l \to \infty$, by the assumption that $\boldsymbol{a}^{(\rho)} \to \boldsymbol{\alpha}$ where all the components of $\boldsymbol{\alpha}$ are strictly positive and dominated convergence, we have

$$Y_1(\boldsymbol{a}^{(\rho)}) \xrightarrow{\text{a.s.}} Y_1(\boldsymbol{\alpha}).$$

By (A10), we have $[Y_t(\boldsymbol{a}^{(\rho)})]^2 \leq (T_t^{(l)})^2$, and $E\big[(T_t^{(l)})^2\big] < \infty$ by Theorem 2 of Hobert et al. (2002). Therefore, $E\big([Y_t(\boldsymbol{a}^{(\rho)})]^2\big) \to E\big([Y_t(\boldsymbol{\alpha})]^2\big)$ by (A10) and dominated convergence, and we also have $E[Y_t(\boldsymbol{a}^{(\rho)})] \to E[Y_t(\boldsymbol{\alpha})]$, so that $\operatorname{Var}[Y_t(\boldsymbol{a}^{(\rho)})] \to \operatorname{Var}[Y_t(\boldsymbol{\alpha})]$. Since $I\big[|Y_1(\boldsymbol{a}^{(\rho)})| > \big(\rho_l \operatorname{Var}[Y_1(\boldsymbol{a}^{(\rho)})]\big)^{1/2}\eta\big] = 0$ for large ρ, (A11) follows by dominated convergence.

The Lindeberg-Feller theorem (together with the fact that $\bar{T}^{(l)} \xrightarrow{\text{a.s.}} E(T_1^{(l)})$) now states that the term in the second set of brackets in (A9) has an asymptotic normal distribution, with mean 0, and variance $E\big([Y_1^{(r,l)}(\boldsymbol{\alpha})]^2\big)/\big(E(T_1^{(l)})\big)^2$. The term in the first set of brackets converges to $\alpha_l c_l^{-1/2}$. Since the k chains are independent, we conclude that

$$\frac{\rho_1^{1/2}}{n}\frac{\partial \ell_\rho(\boldsymbol{\zeta}_0)}{\partial \zeta_r} \xrightarrow{d} \mathcal{N}(0, \Omega_{rr}) \qquad \text{as } \rho_1 \to \infty,$$

where Ω was defined in (2.16). But by the Cramér-Wold Theorem, we obtain the more general statement involving the asymptotic distribution of the entire gradient vector. The argument is standard and gives

$$\frac{\rho_1^{1/2}}{n}\nabla \ell_\rho(\boldsymbol{\zeta}_0) \xrightarrow{d} N(0, \Omega) \qquad \text{as } \rho_1 \to \infty.$$

3. Now, referring to (A5), denote the matrix $-n^{-1}\nabla^2 \ell_\rho(\boldsymbol{\zeta}_*)$ by B_ρ. We have

$$\begin{aligned}
[B_\rho]_{rr} &= \sum_{l=1}^k a_l \left(\frac{1}{n_l}\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_*)\big[1 - p_r(X_i^{(l)}, \boldsymbol{\zeta}_*)\big]\right), \qquad r = 1, \dots, k, \\
[B_\rho]_{rs} &= -\sum_{l=1}^k a_l \left(\frac{1}{n_l}\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_*) p_s(X_i^{(l)}, \boldsymbol{\zeta}_*)\right), \qquad r, s = 1, \dots, k,\ r \neq s,
\end{aligned} \tag{A12}$$

and for later use also define $B_\rho^{(\boldsymbol{\alpha})}$ by

$$\begin{aligned}
[B_\rho^{(\boldsymbol{\alpha})}]_{rr} &= \sum_{l=1}^k a_l \left(\frac{1}{n_l}\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big[1 - p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big]\right), \qquad r = 1, \dots, k, \\
[B_\rho^{(\boldsymbol{\alpha})}]_{rs} &= -\sum_{l=1}^k a_l \left(\frac{1}{n_l}\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) p_s(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})\right), \qquad r, s = 1, \dots, k,\ r \neq s.
\end{aligned} \tag{A13}$$

From (A12) we can check that

$$B_\rho 1_k = 0, \tag{A14}$$

and because $1_k^\top \hat{\boldsymbol{\zeta}} = 0$ and $1_k^\top \boldsymbol{\zeta}_0 = 0$, we have

$$\begin{pmatrix} B_\rho \\ \frac{1_k^\top}{\sqrt{k}} \end{pmatrix} \rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = \begin{pmatrix} \frac{\rho_1^{1/2}}{n} \nabla \ell_\rho(\boldsymbol{\zeta}_0) \\ 0 \end{pmatrix}.$$

Hence

$$\left(B_\rho^\dagger, \tfrac{1_k}{\sqrt{k}}\right) \begin{pmatrix} B_\rho \\ \frac{1_k^\top}{\sqrt{k}} \end{pmatrix} \rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = \left(B_\rho^\dagger, \tfrac{1_k}{\sqrt{k}}\right) \begin{pmatrix} \frac{\rho_1^{1/2}}{n} \nabla \ell_\rho(\boldsymbol{\zeta}_0) \\ 0 \end{pmatrix}. \tag{A15}$$

Now from (A14) and the spectral decomposition of the symmetric matrix B_ρ, we have $B_\rho^\dagger B_\rho = I_k - 1_k 1_k^\top / k$, so (A15) becomes

$$\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = B_\rho^\dagger \frac{\rho_1^{1/2}}{n} \nabla \ell_\rho(\boldsymbol{\zeta}_0).$$

We now study the asymptotic behavior of $B_\rho^\dagger$. From (A13), the fact that $w_l = a_l n / n_l$ by definition, and ergodicity, we have

$$\begin{aligned}
[B_\rho^{(\boldsymbol{\alpha})}]_{rr} &= \sum_{l=1}^k a_l \biggl(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) \bigl[1 - p_r(X_i^{(l)}, \boldsymbol{\zeta}_0)\bigr] \biggr) \xrightarrow{\text{a.s.}} B_{rr}, \\
[B_\rho^{(\boldsymbol{\alpha})}]_{rs} &= -\sum_{l=1}^k a_l \biggl(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) p_s(X_i^{(l)}, \boldsymbol{\zeta}_0) \biggr) \xrightarrow{\text{a.s.}} B_{rs}, \qquad r \neq s.
\end{aligned}$$

The first part of Theorem 1 states that $\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0 \xrightarrow{\text{a.s.}} 0$ as $\rho_1 \to \infty$. Now since all partial derivatives (with respect to $\boldsymbol{\zeta}$) of terms of the form $p_r(x, \boldsymbol{\zeta})(1 - p_r(x, \boldsymbol{\zeta}))$ or $p_r(x, \boldsymbol{\zeta}) p_s(x, \boldsymbol{\zeta})$ are uniformly bounded by 1 in absolute value, we see that $[B_\rho]_{rs} - [B_\rho^{(\boldsymbol{\alpha})}]_{rs} = O(\|\boldsymbol{\zeta}_* - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1)$ a.s. for all r and s, and conclude that $B_\rho \xrightarrow{\text{a.s.}} B$. (Here, $\|v\|_1$ denotes the L_1 norm of a vector $v \in \mathbb{R}^k$.) Similarly, $[\widehat{B}]_{rs} - [B_\rho^{(\boldsymbol{\alpha})}]_{rs} = O(\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1)$ a.s. for all r and s, so $\widehat{B} \xrightarrow{\text{a.s.}} B$. Furthermore, from the spectral decomposition of B_ρ and B, and the fact that $B_\rho 1_k = 0$ and $B 1_k = 0$, we have

$$B_\rho^\dagger = \Bigl(B_\rho + \frac{1}{k} 1_k 1_k^\top\Bigr)^{-1} - \frac{1}{k} 1_k 1_k^\top \qquad \text{and} \qquad B^\dagger = \Bigl(B + \frac{1}{k} 1_k 1_k^\top\Bigr)^{-1} - \frac{1}{k} 1_k 1_k^\top, \tag{A16}$$

showing that $B_\rho^\dagger \xrightarrow{\text{a.s.}} B^\dagger$.

4. The convergence statement $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, B^\dagger \Omega B^\dagger)$ now follows immediately.
5. Finally, we write $\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) = \rho_1^{1/2}(g(\hat{\boldsymbol{\zeta}}) - g(\boldsymbol{\zeta}_0)) = \nabla g(\boldsymbol{\zeta}_*)^\top \rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$, where $\boldsymbol{\zeta}_*$ is between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Since $\nabla g(\boldsymbol{\zeta}_*)^\top \xrightarrow{\text{a.s.}} D$, the desired result (2.20) now follows.

Proof of Consistency of the Estimate of the Asymptotic Covariance Matrix

In the proof of the first part of Theorem 1, we showed that $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ and $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$. Hence, $\widehat{D} \xrightarrow{\text{a.s.}} D$. In the proof of the second part of Theorem 1 we showed that $\widehat{B} \xrightarrow{\text{a.s.}} B$. Using the spectral representation of $\widehat{B}$ and of B (see (A16)), we see that this entails $\widehat{B}^\dagger \xrightarrow{\text{a.s.}} B^\dagger$.

To complete the proof, we need to show that $\widehat{\Omega} \xrightarrow{\text{a.s.}} \Omega$. Consider the expressions for Ω and $\widehat{\Omega}$ given by (2.16) and (2.17), respectively. Since $\boldsymbol{a} \to \boldsymbol{\alpha}$ and $\bar{T}^{(l)} \xrightarrow{\text{a.s.}} E(T_1^{(l)})$, to show that $\widehat{\Omega} \xrightarrow{\text{a.s.}} \Omega$, we need only show that

$$\frac{1}{\rho_l}\sum_{t=1}^{\rho_l}\big(Z_t^{(r,l)} - \hat{\mu}_r^{(l)}T_t^{(l)}\big)\big(Z_t^{(s,l)} - \hat{\mu}_r^{(l)}T_t^{(l)}\big) \xrightarrow{\text{a.s.}} E\big(Y_1^{(r,l)}(\boldsymbol{\alpha})Y_1^{(s,l)}(\boldsymbol{\alpha})\big). \tag{A17}$$

Now, the left side of (A17) is an average of quantities that involve $Z_t^{(r,l)}$ and $\hat{\mu}_r^{(l)}$, which themselves are a sum and an average, respectively, of a function that involves the random quantity $\hat{\boldsymbol{\zeta}}$. At the risk of making the notation more cumbersome, we will now write $Z_t^{(r,l)}(\hat{\boldsymbol{\zeta}})$ instead of $Z_t^{(r,l)}$ and $\hat{\mu}_r^{(l)}(\hat{\boldsymbol{\zeta}})$ instead of $\hat{\mu}_r^{(l)}$. Our plan is to introduce $\Omega_\rho^{(\boldsymbol{\alpha})}$, a version of $\widehat{\Omega}$ in which $\hat{\boldsymbol{\zeta}}$ is replaced by the non-random quantity $\boldsymbol{\zeta}_{\boldsymbol{\alpha}}$, and show that (i) $\Omega_\rho^{(\boldsymbol{\alpha})} \xrightarrow{\text{a.s.}} \Omega$ and (ii) $\widehat{\Omega} - \Omega_\rho^{(\boldsymbol{\alpha})} \xrightarrow{\text{a.s.}} 0$. To this end, let

$$Z_t^{(r,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \quad \text{and} \quad \hat{\mu}_r^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) = \frac{\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})}{n_l},$$

and note that by definition

$$Z_t^{(r,l)}(\hat{\boldsymbol{\zeta}}) = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \quad \text{and} \quad \hat{\mu}_r^{(l)}(\hat{\boldsymbol{\zeta}}) = \frac{\sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})}{n_l}.$$

Define the $k \times k$ matrices Ψ, $\widehat{\Psi}$, and $\Psi_\rho^{(\boldsymbol{\alpha})}$ by

$$\begin{aligned}
\Psi_{rs} &= E\Big[\big\{Z_1^{(r,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) - T_1^{(l)}E_{\pi_l}[p_r(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})]\big\}\big\{Z_1^{(s,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) - T_1^{(l)}E_{\pi_l}[p_s(X, \boldsymbol{\zeta}_{\boldsymbol{\alpha}})]\big\}\Big], \\
\widehat{\Psi}_{rs} &= \frac{1}{\rho_l}\sum_{t=1}^{\rho_l}\big(Z_t^{(r,l)}(\hat{\boldsymbol{\zeta}}) - T_t^{(l)}\hat{\mu}_r^{(l)}(\hat{\boldsymbol{\zeta}})\big)\big(Z_t^{(s,l)}(\hat{\boldsymbol{\zeta}}) - T_t^{(l)}\hat{\mu}_s^{(l)}(\hat{\boldsymbol{\zeta}})\big), \\
[\Psi_\rho^{(\boldsymbol{\alpha})}]_{rs} &= \frac{1}{\rho_l}\sum_{t=1}^{\rho_l}\big(Z_t^{(r,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) - T_t^{(l)}\hat{\mu}_r^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big)\big(Z_t^{(s,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) - T_t^{(l)}\hat{\mu}_s^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}})\big).
\end{aligned}$$

Note that Ψ_{rs} is simply the right side of (A17). Here, Ψ_{rs} is the population-level quantity (which we wish to estimate), $\widehat{\Psi}_{rs}$ is the empirical estimate of this quantity, and $[\Psi_\rho^{(\boldsymbol{\alpha})}]_{rs}$ is an "intermediate" or bridging quantity, used only in our proof. We will show that (i) $\Psi_\rho^{(\boldsymbol{\alpha})} \xrightarrow{\text{a.s.}} \Psi$ and (ii) $\widehat{\Psi} - \Psi_\rho^{(\boldsymbol{\alpha})} \xrightarrow{\text{a.s.}} 0$.

To show that $\Psi_\rho^{(\boldsymbol{\alpha})} \xrightarrow{\text{a.s.}} \Psi$, we first express $\widehat{\Psi}_{rs}$ as a sum of four averages. That the four averages converge to their respective population counterparts follows from the ergodic theorem, together with the fact that $E\big[(T_1^{(l)})^2\big] < \infty$.

To show that $\widehat{\Psi}_{rs} - [\Psi_\rho^{(\boldsymbol{\alpha})}]_{rs} \xrightarrow{\text{a.s.}} 0$, we express $\widehat{\Psi}_{rs} - [\Psi_\rho^{(\boldsymbol{\alpha})}]_{rs}$ as the sum of four differences of averages, and show that each of these converges almost surely to 0. Consider the first difference, which is

$$D_1 := \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} \big[Z_t^{(r,l)}(\hat{\boldsymbol{\zeta}}) Z_t^{(s,l)}(\hat{\boldsymbol{\zeta}}) - Z_t^{(r,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) Z_t^{(s,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big]. \tag{A18}$$

The expression inside the brackets in (A18) is equal to

$$D_{1t} := \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \sum_{j=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \big[p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_j^{(l)}, \hat{\boldsymbol{\zeta}}) - p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) p_s(X_j^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big], \tag{A19}$$

and because all partial derivatives with respect to $\boldsymbol{\zeta}$ of functions of the form $p_r(x, \boldsymbol{\zeta}) p_s(y, \boldsymbol{\zeta})$ are uniformly bounded by 1 in absolute value, the expression inside the brackets in (A19) is bounded by $\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1$. Since there are $(T_t^{(l)})^2$ summands in the double sum in (A19), $|D_{1t}| < (T_t^{(l)})^2 \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1$, and from the fact that $E\big[(T_1^{(l)})^2\big] < \infty$ we now see that $D_1 \xrightarrow{\text{a.s.}} 0$.

The second difference is

$$D_2 := \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} \big[Z_t^{(r,l)}(\hat{\boldsymbol{\zeta}}) T_t^{(l)} \hat{\mu}_s^{(l)}(\hat{\boldsymbol{\zeta}}) - Z_t^{(r,l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) T_t^{(l)} \hat{\mu}_s^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big]. \tag{A20}$$

The expression inside the brackets in (A20)

$$D_{2t} := T_t^{(l)} \frac{1}{n_l} \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \sum_{j=1}^{n_l} \big[p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_j^{(l)}, \hat{\boldsymbol{\zeta}}) - p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) p_s(X_j^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big],$$

and reasoning as we did for the case of the first difference, we have $|D_{2t}| < T_t^{(l)} \cdot T_t^{(l)} \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1$, which implies that $D_2 \xrightarrow{\text{a.s.}} 0$. The third difference is handled in a similar way.

The fourth difference is

$$D_4 := \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} \big[(T_t^{(l)})^2 \hat{\mu}_r^{(l)}(\hat{\boldsymbol{\zeta}}) \hat{\mu}_s^{(l)}(\hat{\boldsymbol{\zeta}}) - (T_t^{(l)})^2 \hat{\mu}_r^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \hat{\mu}_s^{(l)}(\boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big]. \tag{A21}$$

The expression inside the brackets in (A21) is

$$D_{4t} := (T_t^{(l)})^2 \frac{1}{n_l^2} \sum_{i=1}^{n_l} \sum_{j=1}^{n_l} \big[p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_j^{(l)}, \hat{\boldsymbol{\zeta}}) - p_r(X_i^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) p_s(X_j^{(l)}, \boldsymbol{\zeta}_{\boldsymbol{\alpha}}) \big],$$

and we have $|D_{4t}| < (T_t^{(l)})^2 \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_{\boldsymbol{\alpha}}\|_1$, from which we conclude that $D_4 \xrightarrow{\text{a.s.}} 0$.

References

Buta, E. and Doss, H. (2011). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. *Annals of Statistics* **39** 2658–2685.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 473–511.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, Department of Statistics, University of Minnesota.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16** 1069–1112.

Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society,* Series B **65** 585–618.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6** 831–860.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* **24** 101–121.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, London.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association* **90** 233–241.

Newman, M. and Barkema, G. (1999). *Monte Carlo Methods in Statistical Physics*. Oxford University Press.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, London.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods (Second Edition)*. Springer-Verlag, New York.

Romero, M. (2003). *On Two Topics with no Bridge: Bridge Sampling with Dependent Draws and Bias of the Multiple Imputation Variance Estimator*. Ph.D. thesis, University of Chicago.

Swendsen, R. and Wang, J. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters* **58** 86–88.

Tan, A., Doss, H. and Hobert, J. P. (2012). Honest importance sampling with multiple Markov chains. Tech. rep., Department of Statistics, University of Florida.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association* **99** 1027–1036.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13** 178–203.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20** 595–601.

Wolpert, R. L. and Schmidler, S. C. (2011). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica* (in press); preprint at `http://ftp.stat.duke.edu/WorkingPapers/10-19.html`.