# Bayesian inference for Markov chains*

*Ruhr-Universität Bochum and Microsoft Research*

Peter Eichelsbacher† and Ayalvadi Ganesh ‡

**Abstract**

We consider the estimation of Markov transition matrices by Bayes' methods. We obtain large and moderate deviation principles for the sequence of Bayesian posterior distributions.

**MSC 2000 subject classification**: 60F10, 62M05

## 1 Introduction

The Bayesian approach to inference has seen a revival of popularity in applied statistics, motivated in part by increases in computational power. While there is a rich literature on the asymptotic properties of various estimators in classical statistics, the corresponding theoretical aspects of Bayes'

*Research supported by ARC Grant 880 from the Anglo-German foundation.

†Fakultät für Mathematik, Ruhr-Universität Bochum, NA 3/68, D–44780 Bochum, Germany, e-mail: peter.eichelsbacher@ruhr-uni-bochum.de

‡Microsoft Research, 7 J J Thomson Avenue, Cambridge CB3 0FB, UK, e-mail: ajg@microsoft.com.

methods seem to be less well studied. The most important property of any statistical procedure is its consistency and there is, indeed, a rich literature on the consistency of Bayes' posterior distributions. A good review of this work can be found in Ghosal *et al.* [5]. There does not appear to be much work on more refined asymptotics. Large and moderate deviations have been studied by Ganesh and O'Connell [4], and Eichelsbacher and Ganesh [2] respectively, for priors satisfying certain exchangeability properties. Central limit theorems have been obtained in specific cases by Le Cam [7] and Ibragimov and Has'minskii [6].

The results described above pertain to Bayesian inference for iid sequences of observations. Markov chains are among the simplest models for dependent sequences; nevertheless, they find an extremely rich and diverse set of applications in stochastic modeling. It is therefore natural to seek to extend the above results to the Markovian setting. Here, we have a sequence of observations of the states of a Markov chain whose transition probability matrix is unknown, and is to be inferred by a Bayesian procedure. In this paper, we consider an irreducible, discrete time Markov chain with a finite state space.

This model has been studied previously by Papangelou [8], who establishes a large deviation principle (LDP) for the sequence of Bayes' posteriors. This is much stronger than consistency, or even the assertion of exponential convergence of the posterior. The rate function for the LDP is the relative entropy function $H(P|\cdot)$, where $P$ is the transition matrix of the Markov chain. Relative entropy also appears as the rate function in Sanov's theorem for the pair empirical measure of a Markov chain, but in the form $H(\cdot|P)$. This interchange of arguments has been noted for iid sequences as well. It has a

simple intuitive explanation, as noted by Papangelou: in Sanov's theorem, we ask how likely it is that the empirical transition matrix is close to $Q$, given that the transition matrix is $P$, whereas in the Bayesian context we ask how likely it is that the transition matrix is close to $Q$ given that we observe an empirical transition matrix close to $P$.

Papangelou obtains the LDP for Markov chains whose order is unknown; the result when the order is known follows easily. The LDP for Markov chains of known order is derived independently by Paschalidis and Vassilaras [9], who give a nice application of their result to modeling telecommunication systems.

Our contributions in this paper are the following. We give independent proofs of both large and moderate deviation principles for Markov chains of known order, which we obtain as consequences of LDPs and moderate deviations principles (MDPs) for iid sequences. The study of moderate deviations for Bayesian inference of Markov transition functions is new, to the best of our knowledge. The rate function for the MDP is the Fisher information. This is the same rate function as arises in Sanov's theorem for the empirical measure of a Markov chain. This is analogous to the corresponding result for iid sequences and is in contrast to the large deviation setting, where the rate function is different from that in Sanov's theorem.

## 2 Statement of the main results

Let $\Omega$ be a finite set, and let $M_1(\Omega)$ denote the space of probability measures on $\Omega$. Let $X_1, X_2, \ldots$ be a Markov chain on $\Omega$ with irreducible transition

probability matrix $R = \{r(i,j) : i,j \in \Omega\}$, which is unknown. We consider the problem of inferring $R$ by a Bayesian procedure based on the observations $X_1, \ldots, X_{n+1}$. Note that $R \in M_1(\Omega)^\Omega$ and let $\mathcal{P} \in M_1(M_1(\Omega)^\Omega)$ denote the prior distribution on $R$. Let $\hat{P}_n$ denote the empirical transition probability matrix with entries

$$\hat{p}_n(i,j) = \frac{\sum_{k=1}^n 1(X_k = i, X_{k+1} = j)}{\sum_{k=1}^n 1(X_k = i)}, \quad i,j \in \Omega.$$

We shall assume without loss of generality that, for large enough $n$, the denominator above is positive for all $i \in \Omega$; we simply work with a set $\Omega$ consisting only of those states for which the assumption holds. We also define the empirical marginal distribution

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

Let $\mathcal{P}^n$ denote the posterior distribution corresponding to the prior $\mathcal{P}$ and the empirical distribution $\hat{P}_n$ (we do not make the dependence of $\mathcal{P}^n$ on $\hat{P}_n$ explicit in the notation). $\mathcal{P}^n$ is absolutely continuous with respect to $\mathcal{P}$ with Radon-Nikodym derivative on the support of $\mathcal{P}$ given by

$$\frac{d\mathcal{P}^n}{d\mathcal{P}}(Q) = \frac{\prod_{i,j\in\Omega} q(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)}}{\int_{M_1(\Omega)^\Omega} \prod_{i,j\in\Omega} r(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} d\mathcal{P}(R)} \tag{1}$$

with $Q = \{q(i,j) : i,j \in \Omega\}$.

**Assumption A**: Consider a given realization of the process $X_1, X_2, \ldots$. Suppose that, for this realization, $\hat{p}_n(i,j) \to p(i,j)$ and $\hat{\pi}_n(i) \to \pi(i)$ for all $i,j \in \Omega$, as $n \to \infty$. Suppose too that $p(i,j)$ is irreducible and that $\pi(i) > 0$ for all $i \in \Omega$.

It is not hard to verify that $\pi$ is the unique stationary distribution for the Markov chain with transition probability matrix $P = \{p(i,j) : i,j \in \Omega\}$.

Under the above assumption, the sequence of posterior distributions $\mathcal{P}^n$ obey large and moderate deviations principles as described below. In the remainder of this paper, $M_1(\Omega)^\Omega$ is equipped with its natural topology as a subset of a finite-dimensional Euclidean space.

**Theorem 1** *If assumption A holds for the observed realization of $X_1, X_2, \ldots$, then the sequence of posterior distributions $\mathcal{P}^n$ obeys an LDP in $M_1(M_1(\Omega)^\Omega)$ with the good rate function*

$$I(Q) = \begin{cases} \sum_{i,j\in\Omega} \pi(i)p(i,j)\log\frac{p(i,j)}{q(i,j)}, & \text{if } Q \in \text{supp } \mathcal{P}, \\ +\infty, & \text{otherwise.} \end{cases}$$

We note that the rate function is convex provided the support of the prior is convex. The above is a pointwise result on the set of all limit points of sequences of empirical distributions. Since the empirical distributions converge almost surely to the probability law of the Markov chain, the theorem implies an almost sure LDP, stated below.

**Corollary 1** *The sequence of posterior distributions $\mathcal{P}^n$ almost surely satisfies an LDP in $M_1(M_1(\Omega)^\Omega)$ with the rate function*

$$I(Q) = \begin{cases} \sum_{i,j\in\Omega} \rho(i)r(i,j)\log\frac{r(i,j)}{q(i,j)}, & \text{if } Q \in \text{supp } \mathcal{P}, \\ +\infty, & \text{otherwise,} \end{cases}$$

*where $R$ is the transition matrix and $\rho$ the stationary distribution of the Markov chain $\{X_i, i \in \mathbb{N}\}$.*

By the ergodic theorem for Markov chains, $\hat{\pi}_n$ and $\hat{P}_n$ converge almost surely to $\rho$ and $R$ respectively. Hence, the corollary is immediate from Theorem 1.

**Definitions** : Let $S^\Omega$ denote the $\Omega$-simplex,

$$S^\Omega = \{x \in \mathbb{R}^\Omega : x_i \geq 0 \; \forall \; i \in \Omega, \; \sum_{i \in \Omega} x_i = 1\}.$$

We identify $(S^\Omega)^\Omega$ with $M_1(\Omega)^\Omega$. We denote Lebesgue measure on $S^\Omega$ by $\lambda$ and we let $m = \lambda^\Omega$ denote Lebesgue measure on $(S^\Omega)^\Omega$. We say that $P \in M_1(\Omega)^\Omega$ is a regular point of the support of $\mathcal{P}$ if $p(i,j) > 0$ for all $i, j \in \Omega$ and if there is a neighbourhood of $P$ in $(S^\Omega)^\Omega$ on which $\mathcal{P}$ is absolutely continuous with respect to Lebesgue measure, $m$, on $(S^\Omega)^\Omega$, and if the density of $\mathcal{P}$ with respect to $m$ is bounded away from zero and infinity on this neighbourhood.

We say that a positive sequence $\{b_n, n \in \mathbb{N}\}$ is regularly varying if $c(t) = \lim_{n\to\infty} b(\lfloor nt \rfloor)/b(n)$ exists for all $t > 0$. It is not hard to see that if the limit exists, then $c(t_1 t_2) = c(t_1)c(t_2)$ for all $t_1, t_2 > 0$, and so $c(t) = t^\alpha$ for some $\alpha \in \mathbb{R}$.

**Theorem 2** *Suppose Assumption A holds with $P$ a regular point of the support of $\mathcal{P}$. Let $(b_n)_{n\in\mathbb{N}}$ be a regularly varying sequence such that*

$$\frac{b_n}{n} \to 0, \quad \frac{b_n^2}{n} \to \infty \quad \text{as } n \to \infty, \tag{2}$$

*and suppose $|\hat{p}_n(i,j) - p(i,j)| = o(b_n/n)$, $|\hat{\pi}_n(i) - \pi(i)| = o(b_n/n)$ for all $i, j \in \Omega$. Then, for any open subset $G$ and closed subset $F$ of $M_0(\Omega)^\Omega$, we have*

$$\begin{aligned}
\liminf_{n\to\infty} \frac{n}{b_n^2} \log \mathcal{P}_n \left( P + \frac{b_n}{n} G \right) &\geq -\inf_{Q \in G} \sum_{i,j \in \Omega} \frac{\pi(i) q(i,j)^2}{p(i,j)}. \\
\limsup_{n\to\infty} \frac{n}{b_n^2} \log \mathcal{P}_n \left( P + \frac{b_n}{n} F \right) &\leq -\inf_{Q \in F} \sum_{i,j \in \Omega} \frac{\pi(i) q(i,j)^2}{p(i,j)}.
\end{aligned}$$

Here, $M_0(\Omega)$ is the set of signed measures $\lambda$ on $\Omega$ such that $\sum_{i\in\Omega}\lambda(i)=0$. As happened with the LDP, the above theorem implies an almost sure MDP but under stronger conditions on the sequence $b_n$.

**Corollary 2** *Suppose $R$, the transition matrix of $(X_i)_{i\in\mathbb{N}}$, is a regular point of the support of $\mathcal{P}$. Let $(b_n)_{n\in\mathbb{N}}$ be a regularly varying sequence such that*

$$\frac{b_n}{n}\to 0, \quad \frac{b_n^2}{n\log\log n}\to\infty \quad \text{as } n\to\infty.$$

*Then the conclusions of Theorem 2 hold $R$-almost surely with $\pi$ and $P$ replaced by $\rho$ and $R$ respectively, where $\rho$ is the stationary distribution corresponding to $R$.*

By the law of the iterated logarithm, $\sqrt{n/\log\log n}\|\hat{P}_n - R\|$ is almost surely bounded and so $\|\hat{P}_n - R\| = o(b_n/n)$ $R$-almost surely under the above assumptions on $b_n$. The same is true of $\|\hat{\pi}_n - \rho\|$. The corollary is now obvious.

**Remarks**: We have stated our results above in terms of first-order Markov transition functions in order to keep the notation simple. The statements and proofs carry over essentially unchanged to the setting of higher-order Markov chains, provided the order is known and the prior concentrates on transition functions of the correct order.

## 3 Proofs of the results

**Proof of Theorem 1**: Let $Z_1, Z_2, \ldots$ be an iid sequence in $\Omega^2$ with $\mathbf{P}(Z_1 = (i,j)) = \rho(i)r(i,j)$, $i,j\in\Omega$, where $R$ is an irreducible Markov transition matrix and $\rho$ its unique stationary distribution. Suppose the distribution,

$\rho \times R$, of this iid sequence is unknown and is to be inferred by a Bayesian procedure, based on the observed sequence $Z_1, \ldots, Z_n$. Define

$$\mu_n(i) = \frac{1}{n}\sum_{k=1}^{n} 1(Z_k(1) = i), \quad q_n(i,j) = \frac{1}{n\mu_n(i)}\sum_{k=1}^{n} 1(Z_k = (i,j)), \quad i,j \in \Omega,$$

where $Z_k(1)$ denotes the first component of $Z_k$. We assume that for large enough $n$, $\mu_n(i)$ is positive for all $i \in \Omega$ and so $Q_n = \{q_n(i,j) : i,j \in \Omega\}$ is well-defined.

Let $\pi$ be any probability distribution on $\Omega$ such that $\pi(i) > 0$ for all $i \in \Omega$, and let $\delta_\pi \in M_1(M_1(\Omega))$ denote unit mass at $\pi$. We take our prior on $\rho \times R$, the distribution of $Z_1$, to be $\delta_\pi \times \mathcal{Q}$, where $\mathcal{Q}$ is some probability distribution on $M_1(\Omega)^\Omega$. A simple calculation using Bayes' theorem shows that the posterior based on $Z_1, \ldots, Z_n$ is $\delta_\pi \times \mathcal{Q}^n$ where $\mathcal{Q}^n$ is absolutely continuous with respect to $\mathcal{Q}$ and has density

$$\frac{d\mathcal{Q}^n}{d\mathcal{Q}}(\nu) = \frac{\prod_{i,j\in\Omega} \nu(i,j)^{n\mu_n(i)q_n(i,j)}}{\int_{M_1(\Omega)^\Omega} \prod_{i,j\in\Omega} \lambda(i,j)^{n\mu_n(i)q_n(i,j)} d\mathcal{Q}(\lambda)} \tag{3}$$

at $\nu \in \text{supp } \mathcal{Q}$.

Suppose $\mu_n \to \pi$ and $Q_n \to Q \in \text{supp } \mathcal{Q}$ as $n \to \infty$, for a given realization of $Z_1, Z_2, \ldots$. Since $(Z_k)_{k\in\mathbb{N}}$ is an iid sequence, it follows from [3, Theorem 1] that the sequence of posterior distributions $\delta_\pi \times \mathcal{Q}^n$ satisfies an LDP with rate function $I$ given by

$$I(\lambda \times R) = \begin{cases} h(\pi \times Q | \lambda \times R), & \text{if } \lambda \times R \in \text{supp } \delta_\pi \times \mathcal{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

In other words, the rate function is $h(\pi \times Q | \pi \times R)$ if $\lambda = \pi$ and $R$ is in the support of $\mathcal{Q}$, and infinite otherwise. Here $h$ is the relative entropy function for probability measures,

$$h(\pi \times Q | \pi \times R) = \sum_{i,j\in\Omega} \pi(i)q(i,j) \log \frac{\pi(i)q(i,j)}{\pi(i)r(i,j)} = \sum_{i,j\in\Omega} \pi(i)q(i,j) \log \frac{q(i,j)}{r(i,j)}.$$

Thus, for any open subset $A$ of $M_1(\Omega)^\Omega$, we have the lower bound,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathcal{Q}^n(A) = - \inf_{R \in A \cap \text{supp } \mathcal{Q}} \sum_{i,j\in\Omega} \pi(i) q(i,j) \log \frac{q(i,j)}{r(i,j)},$$

and a corresponding upper bound for closed sets.

With the identification $Z_n = (X_n, X_{n+1})$, it follows from (1) and (3) that, if $\mathcal{P} = \mathcal{Q}$, $\mu_n = \hat{\pi}_n$ and $\hat{P}_n = Q_n$, then $\mathcal{P}^n = \mathcal{Q}^n$. Consequently, if $\hat{\pi}_n \to \pi$ and $\hat{P}_n \to P$ as $n \to \infty$, we immediately have for any open subset $A$ of $M_1(\Omega)^\Omega$ that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathcal{P}^n(A) = - \inf_{Q \in A \cap \text{supp } \mathcal{P}} \sum_{i,j\in\Omega} \pi(i) p(i,j) \log \frac{p(i,j)}{q(i,j)}.$$

Likewise, we obtain a corresponding upper bound for closed sets. This completes the proof of the theorem. ■

**Proof of Theorem 2**: The reason that we can't follow the same approach to prove the MDP is that the MDP for iid sequences was proved under certain 'regularity' conditions on the prior. A prior of the form $\delta_\pi \times \mathcal{Q}$ which concentrates on a single point in one of the co-ordinates does not satisfy these conditions. The proof now proceeds through a sequence of lemmas.

**Lemma 1** *Under the assumptions of Theorem 2, there are constants $c_1$, $c_2$ in $(0, \infty)$ such that*

$$c_1 n^{-|\Omega|(|\Omega|-1)/2} \leq \int_{M_1(\Omega)^\Omega} \exp(-nH(\hat{P}_n|R)) d\mathcal{P}(R) \leq c_2 n^{-|\Omega|(|\Omega|-1)/2}.$$

The relative entropy between Markov transition functions is defined as

$$H(\hat{P}_n|R) = \sum_{i,j\in\Omega} \hat{\pi}_n(i) \hat{p}_n(i,j) \log \frac{\hat{p}_n(i,j)}{r(i,j)},$$

where $\hat{\pi}_n$ is the stationary distribution corresponding to the transition matrix $\hat{P}_n$, which is assumed to be irreducible.

*Proof*: In the following, $c$ and $k$ will denote generic positive constants, not necessarily the same at each occurrence. We fix $\alpha, \delta > 0$ and define the sets

$$\begin{aligned} \mathcal{A}^n &= \{R \in M_1(\Omega)^\Omega : |r(i,j) - \hat{p}_n(i,j)| < \alpha n^{-1/2} \ \forall \ i,j \in \Omega\}, \\ \mathcal{A}^n_\delta &= \{R \in M_1(\Omega)^\Omega : |r(i,j) - \hat{p}_n(i,j)| < \delta \ \forall \ i,j \in \Omega\}. \end{aligned}$$

Now, for $R \in \mathcal{A}^n$, we have

$$H(\hat{P}_n|R) = -\sum_{i,j\in\Omega} \hat{\pi}_n(i)\hat{p}_n(i,j) \log\left[1 + \frac{r(i,j) - \hat{p}_n(i,j)}{\hat{p}_n(i,j)}\right] =$$

$$-\sum_{i,j\in\Omega}\left[\hat{\pi}_n(i)r(i,j) - \hat{\pi}_n(i)\hat{p}_n(i,j) - \hat{\pi}_n(i)\frac{(r(i,j) - \hat{p}_n(i,j))^2}{2\hat{p}_n(i,j)}\right] + O\left(n^{-\frac{3}{2}}\right).$$

But $\sum_{i,j} \hat{\pi}_n(i)r(i,j) = \sum_{i,j} \hat{\pi}_n(i)\hat{p}_n(i,j) = 1$, and so it follows from the above that there is a positive constants $k$ such that $nH(\hat{P}_n|R) \leq k$ for all $R \in \mathcal{A}^n$ and $n$ sufficiently large. Since we assumed that $\mathcal{P} \geq cm$ on $\mathcal{A}^n$ for some $c > 0$ (here $m$ denotes Lebesgue measure on $M_1(\Omega)^\Omega$), we have

$$\begin{aligned} \int_{M_1(\Omega)^\Omega} \exp(-nH(\hat{P}_n|R))d\mathcal{P}(R) &\geq \int_{\mathcal{A}^n} ce^{-k} dm \\ &= ce^{-k} \text{ vol}(\mathcal{A}^n) \ = \text{const } n^{-|\Omega|(|\Omega|-1)/2}, \end{aligned}$$

which completes the proof of the lower bound. We also obtain from the non-negativity of relative entropy and the assumption that $\mathcal{P} \leq cm$ on $\mathcal{A}^n$ for some $c > 0$ that,

$$\int_{\mathcal{A}^n} \exp(-nH(\hat{P}_n|R))d\mathcal{P}(R) \leq c \text{ vol}(\mathcal{A}^n) \ = \ \text{const. } n^{-|\Omega|(|\Omega|-1)/2}. \qquad (4)$$

Recall that the total variation distance between two probability measures $p$ and $q$ on $\Omega$ is defined as $d_{TV}(p,q) = \sum_{x\in\Omega} |p(x) - q(x)|/2$. Now, by a

well-known inequality of Csiszar and Kullback (see, e.g., [1, Ex. 6.2.17]),

$$\begin{aligned} H(\hat{P}_n|R) &= \sum_{i\in\Omega} \hat{\pi}_n(i) h(\hat{P}_n(i,\cdot)|R(i,\cdot)) \\ &\geq 2\sum_{i\in\Omega} \hat{\pi}_n(i) d_{TV}(\hat{P}_n(i,\cdot)|R(i,\cdot))^2 \\ &\geq 2\min_{i\in\Omega} \hat{\pi}_n(i) \max_{i\in\Omega} d_{TV}(\hat{P}_n(i,\cdot)|R(i,\cdot))^2 \\ &\geq \frac{\min_{i\in\Omega} \hat{\pi}_n(i)}{|\Omega|} \max_{i\in\Omega} d_{TV}(\hat{P}_n(i,\cdot)|R(i,\cdot)) \sum_{i,j\in\Omega} |r(i,j) - \hat{p}_n(i,j)|. \end{aligned}$$

Now, if $R \notin \mathcal{A}^n$, then $\max_{i\in\Omega} d_{TV}(\hat{P}_n(i,\cdot)|R(i,\cdot)) \geq \alpha n^{-1/2}/2$. Moreover, $\pi(i) > 0$ for all $i$ by the assumption that $P$ is irreducible, and so $\min_{i\in\Omega} \hat{\pi}_n(i)$ is bounded away from zero for all $n$ large enough. Hence, $nH(\hat{P}_n|R) \geq k\sqrt{n}\sum_{i,j\in\Omega} |r(i,j) - \hat{p}_n(i,j)|$. Since $\mathcal{P} \leq cm$ on $\mathcal{A}_\delta^n$ for some $c > 0$ and small enough $\delta$ by assumption, it follows that

$$\begin{aligned} &\int_{\mathcal{A}_\delta^n \setminus \mathcal{A}^n} e^{-nH(\hat{P}^n|R)} d\mathcal{P}(R) \leq \int_{\mathcal{A}_\delta^n} c e^{-k\sqrt{n}\sum_{i,j\in\Omega} |r(i,j)-\hat{p}_n(i,j)|} dm(R) \\ &\leq c \prod_{i\in\Omega} \prod_{j=1}^{|\Omega|-1} \int_{-\delta}^{\delta} e^{-k\sqrt{n}|x|} dx \;\leq\; c n^{-|\Omega|(|\Omega|-1)/2}. \end{aligned} \tag{5}$$

Finally, if $R \notin \mathcal{A}_\delta^n$, then the total variation distance, $d_{TV}(R(i,\cdot), \hat{P}_n(i,\cdot))$ exceeds $\delta$ for some $i \in \Omega$, and so $H(\hat{P}_n|R) \geq k\delta^2$. Hence,

$$\int_{M_1(\Omega)^\Omega \setminus \mathcal{A}_\delta^n} \exp(-nH(\hat{P}_n|R)) d\mathcal{P}(R) \leq c e^{-kn\delta^2}. \tag{6}$$

It is clear from (4,5,6) that the upper bound in the lemma holds. Together with the lower bound established earlier, this completes the proof of the lemma. ∎

A similar argument shows that there are constants $c_1$, $c_2 > 0$ such that

$$c_1 n^{-|\Omega|(|\Omega|-1)/2} \leq \int_{M_1(\Omega)^\Omega} \exp(-nH(\hat{P}_n|R)) dm(R) \leq c_2 n^{-|\Omega|(|\Omega|-1)/2}, \tag{7}$$

where $m$ denotes Lebesgue measure on $M_1(\Omega)^\Omega$.

We can rewrite (1) as

$$\frac{d\mathcal{P}^n}{d\mathcal{P}}(Q) = \frac{\prod_{i,j\in\Omega} q(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)}}{\int_{M_1(\Omega)^\Omega} \prod_{i,j\in\Omega} r(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} dm(R)} \frac{\int_{M_1(\Omega)^\Omega} \exp[-nH(\hat{P}_n|R)] dm(R)}{\int_{M_1(\Omega)^\Omega} \exp[-nH(\hat{P}_n|R)] d\mathcal{P}(R)}.$$

Hence, by Lemma 1 and (7), there are constants $c_1$, $c_2$ such that

$$c_1 \frac{d\mathcal{P}^n}{d\mathcal{P}}(Q) \leq \frac{\prod_{i,j\in\Omega} q(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)}}{\int_{M_1(\Omega)^\Omega} \prod_{i,j\in\Omega} r(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} dm(R)} \leq c_2 \frac{d\mathcal{P}^n}{d\mathcal{P}}(Q).$$

Using the assumption of regularity of $P$ once more, we obtain for any measurable $A \subseteq M_0(\Omega)^\Omega$ and large enough $n$ that

$$c_1 \mathcal{P}^n \left( P + \frac{b_n}{n} A \right) \leq \tilde{\mathcal{P}}^n \left( P + \frac{b_n}{n} A \right) \leq c_2 \mathcal{P}^n \left( P + \frac{b_n}{n} A \right) \tag{8}$$

where

$$\tilde{\mathcal{P}}^n(B) := \frac{\int_B \prod_{i,j\in\Omega} r(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} dm(R)}{\int_{M_1(\Omega)^\Omega} \prod_{i,j\in\Omega} r(i,j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} dm(R)}. \tag{9}$$

The claim of Theorem 2 is now immediate from the following lemma.

**Lemma 2** *Let $\tilde{\mathcal{P}}^n$ be defined as in (9). For any open subset $G$ and closed subset $F$ of $M_0(\Omega)^\Omega$, we have under the assumptions of Theorem 2 that*

$$\begin{aligned} \liminf_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}_n \left( P + \frac{b_n}{n} G \right) &\geq -\inf_{Q\in G} \sum_{i,j\in\Omega} \frac{\pi(i) q(i,j)^2}{p(i,j)}. \\ \limsup_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}_n \left( P + \frac{b_n}{n} F \right) &\leq -\inf_{Q\in F} \sum_{i,j\in\Omega} \frac{\pi(i) q(i,j)^2}{p(i,j)}. \end{aligned}$$

*Proof*: Let $B = B_1 \times \cdots \times B_{|\Omega|}$ be an open rectangle in $M_1(\Omega)^\Omega$. Since $m$ is the product measure $\lambda^\Omega$, where $\lambda$ is Lebesgue measure on the $\Omega$-simplex, $S^\Omega$, we can rewrite (9) as

$$\tilde{\mathcal{P}}^n(B) = \prod_{i\in\Omega} \frac{\int_{B_i} \prod_{j\in\Omega} x(j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} d\lambda(\mathbf{x})}{\int_\Omega \prod_{j\in\Omega} x(j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} d\lambda(\mathbf{x})}.$$

For $i \in \Omega$ and $B \subseteq M_1(\Omega)$, define

$$\tilde{\mathcal{P}}_i^{n\hat{\pi}_n(i)}(B) = \frac{\int_B \prod_{j\in\Omega} x(j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} d\lambda(\mathbf{x})}{\int_\Omega \prod_{j\in\Omega} x(j)^{n\hat{\pi}_n(i)\hat{p}_n(i,j)} d\lambda(\mathbf{x})},$$

so that

$$\tilde{\mathcal{P}}^n(B) = \prod_{i\in\Omega} \tilde{\mathcal{P}}_i^{n\hat{\pi}_n(i)}(B_i). \tag{10}$$

Here, $\tilde{\mathcal{P}}_i^{n\hat{\pi}_n(i)}$ is the posterior distribution for transitions out of state $i$ when the prior is Lebesgue measure on $M_1(\Omega)$ and there are $n\hat{\pi}_n(i)$ observations, a fraction $\hat{p}_n(i,j)$ of which are transitions to state $j$. Note that the transitions out of each state $i$ are iid. Thus, for each $i \in \Omega$ and any open set $A_i \subseteq M_0(\Omega)$, we have by [2, Theorem 1] that,

$$\lim_{n\to\infty} \frac{n\hat{\pi}_n(i)}{b^2_{n\hat{\pi}_n(i)}} \log \tilde{\mathcal{P}}_i^{n\hat{\pi}_n(i)} \left( P(i,\cdot) + \frac{b_{n\hat{\pi}_n(i)}}{n\hat{\pi}_n(i)} A_i \right) = -\inf_{q\in A_i} \sum_{j\in\Omega} \frac{q(j)^2}{p(i,j)}. \tag{11}$$

Here, we have used the continuity of $q \mapsto \sum_j q(j)^2/p(i,j)$ to replace the upper and lower bounds in the statement of the MDP with an equality for the limit.

By the assumption that $b_n$ is a regularly varying sequence, and that $\hat{\pi}_n(i) \to \pi(i)$ as $n \to \infty$, we obtain that as $n \to \infty$,

$$\frac{b_{n\hat{\pi}_n(i)}}{b_n} \to c(\pi(i)) \quad \text{and so} \quad \frac{b_{n\hat{\pi}_n(i)}}{n\hat{\pi}_n(i)} \Big/ \frac{b_n}{n} \to \frac{c(\pi(i))}{\pi(i)}. \tag{12}$$

Thus, for arbitrary $\delta > 0$, we have for large enough $n$ that

$$\frac{b_{n\hat{\pi}_n(i)}}{n\hat{\pi}_n(i)} A_i^{-\delta} \subseteq \frac{b_n}{n} \frac{c(\pi(i))}{\pi(i)} A_i \subseteq \frac{b_{n\hat{\pi}_n(i)}}{n\hat{\pi}_n(i)} A_i^{\delta}. \tag{13}$$

Here $A_i^\delta = \{\mathbf{x} \in M_0(\Omega) : \inf_{\mathbf{y}\in A_i} d(x,y) < \delta\}$, where $d(\cdot,\cdot)$ denotes Euclidean distance, and $A_i^{-\delta}$ is the interior of the complement of $(A_i^c)^\delta$, where $A_i^c$ denotes the complement of $A_i$. Letting $\delta \to 0$ and using the continuity of

$q \mapsto \sum_j q(j)^2/p(i,j)$, we have from (11,12,13) that

$$\lim_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}_i^{n\hat{\pi}_n(i)} \left( P(i,\cdot) + \frac{b_n}{n} A_i \right)$$

$$= -\frac{c(\pi(i))^2}{\pi(i)} \inf_{q \in \frac{\pi(i)}{c(\pi(i))} A_i} \sum_{j\in\Omega} \frac{q(j)^2}{p(i,j)} = -\inf_{q\in A_i} \sum_{j\in\Omega} \frac{\pi(i) q(j)^2}{p(i,j)}. \quad (14)$$

Let $A = A_1 \times \ldots \times A_{|\Omega|}$ be an open rectangle in $M_0(\Omega)^\Omega$. It follows from (10) and (14) that

$$\lim_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}^n \left( P + \frac{b_n}{n} A \right) = -\inf_{Q\in A} \sum_{i,j\in\Omega} \frac{\pi(i) q(i,j)^2}{p(i,j)}.$$

Since the open rectangles constitute a base for the topology on $M_0(\Omega)^\Omega$, we obtain a *weak* MDP using [1, Theorem 4.1.11]. In other words, the upper bound in the statement of the lemma holds for all compact $F$, while the lower bound holds for all open $G$. In order to strengthen this result to a *full* MDP, we need to establish exponential tightness, which we do below.

Fix $\alpha \in \mathbb{R}$. For $i \in \Omega$, let

$$K_i = \left\{ q \in M_0(\Omega) : \sum_{j\in\Omega} \frac{q(j)^2}{p(i,j)} \le \alpha \right\}.$$

Clearly, $K = K_1 \times \cdots \times K_\Omega$ is compact. Defining

$$A^i = \{Q \in M_0(\Omega)^\Omega : Q(i,\cdot) \in K_i^c\},\ i \in \Omega,$$

where $K_i^c$ denotes the complement of $K_i$, we see that $K^c \subseteq \cup_{i\in\Omega} A^i$ and, by (10) and (14),

$$\lim_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}^n \left( P + \frac{b_n}{n} A^i \right) \le -\alpha\pi(i), \quad i \in \Omega.$$

Therefore,

$$\lim_{n\to\infty} \frac{n}{b_n^2} \log \tilde{\mathcal{P}}^n \left( P + \frac{b_n}{n} K^c \right) \le -\alpha \min_{i\in\Omega} \pi(i).$$

Since $\pi(i) > 0$ for all $i \in \Omega$ under the assumptions of the lemma, and $\alpha$ is arbitrary, the above establishes exponential tightness of $\{\mathcal{P}^n, n \in \mathbb{N}\}$. By [1, Lemma 1.2.18], this implies the *full* MDP, and concludes the proof of the lemma and of Theorem 2. ■

## References


[1] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer New-York, 1998.

[2] P. Eichelsbacher and A. J. Ganesh, Moderate deviations for Bayes posteriors, to appear in *Scand. J. Stat.*, 2002.

[3] A. J. Ganesh and N. O'Connell, An inverse of Sanov's theorem, *Stat. and Prob. Letters* 42, 201–206, 1999.

[4] A. J. Ganesh and N. O'Connell, A large deviation principle for Dirichlet posteriors, *Bernoulli*, 6(6), 1021–1034, 2000.

[5] S. Ghosal, J. K. Ghosh and R. V. Ramamoorthi, *Consistency issues in Bayesian nonparametrics*, in Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri (Subir Ghosh ed.), Marcel Dekker, 1998.

[6] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York, 1981.

[7] L. M. Le Cam, Convergence of estimates under dimensionality restrictions, *Ann. Statist.*, 22: 38–53, 1973.

[8] F. Papangelou, Large deviations and the Bayesian estimation of higher-order Markov transition functions, *J. Appl. Prob.* 33, 18–27. 1996.

[9] I. Ch. Paschalidis and S. Vassilaras, On estimating buffer overflow probabilities under Markov-modulated inputs, *Proc. 37th Annual Allerton Conference*, 1999.