A nonstandard form of the rate function for the occupation measure of a Markov chain

Paul Dupuis*
Division of Applied Mathematics
Brown University
Providence, RI 02912

Ofer Zeitouni †
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

March 17, 1995

Abstract We investigate, by means of an example, the large deviations principle for the empirical measure of a Markov chain when Feller continuity properties are not assumed. Using the weak convergence approach, we explicitly compute the resulting rate function, and find that it is not of the Donsker-Varadhan form.

1 Introduction

Let X_n denote a discrete time Markov chain, with Polish state space $\mathcal{X}$ and transition kernel $p(x, da)$, and let $L_n = n^{-1} \sum_{j=1}^n \delta_{x_j}$ denote its induced empirical measure (also called the occupation measure). One of the outstanding successes of the theory of large deviations has been the derivation, by Donsker and Vardahan [6], of a general large deviation principle for L_n, with the rate function given explicitly by the solution of a variational problem. Let $\mathcal{M}_1(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$ with the weak topology, and let $H(\mu_1|\mu_2)$ denote the relative entropy between probability measures μ_1, μ_2: $H(\mu_1|\mu_2) = \infty$ if μ_1 is not absolutely continuous with respect to μ_2, and $H(\mu_1|\mu_2) = \int \log((d\mu_1/d\mu_2)(x))\mu_1(dx)$ otherwise. Under appropriate conditions, it was proved in [6] that

$$-\inf_{\mu \in A^o} I(\mu) \leq \liminf_{n\to\infty} P(L_n \in A) \leq \limsup_{n\to\infty} P(L_n \in A) \leq -\inf_{\mu \in \bar{A}} I(\mu)\,,$$

where A^o and $\bar{A}$ denote, respectively, the interior and closure of a (measurable) set $A \subset \mathcal{M}_1(\mathcal{X})$. The rate function I is defined by

$$I(\mu) = \inf_{\{\pi(x,da):\int \mu(dx)\pi(x,da)=\mu(da)\}} H(\mu(dx)\pi(x,da)|\mu(dx)p(x,da))\,.$$

(We refer the reader to [3, 4, 7] for definitions and terminology related to the theory of large deviations).

*The work of this author was supported in part by NSF grant DMS–9403820.

†The work of this author was done while he visited MIT, under support from NSF grant 9302709–DMS.

Besides exponential tightness assumptions, which we will avoid in this paper by restricting $\mathcal{X}$ to be compact and hence forcing $\mathcal{M}_1(\mathcal{X})$ to be compact, this basic result is usually obtained under either an assumption of Feller continuity of the kernel $p(x, da)$, or assumptions on the uniformity of the kernel with respect to x, in the sense of the existence of a dominating measure (see [4, 2, 9, 3] for partial references). It has also been obtained under what can be thought of as a "with probability one Feller condition" in [7]. This assumption requires neither the Feller property nor a dominating measure, but instead assumes that the set of points where the Feller property fails is negligible in an appropriate sense. The rate function in this case is the same as under the Feller condition. On the other hand, under appropriate mixing hypotheses one can prove that a large deviations principle is valid, although in this case the rate function is not explicitly identified (see the exposition in [4] and, for weaker conditions, the paper [1]). This gap is a nontrivial gap, as the examples in [2, 5] amply demonstrate. Our goal in this paper is to further explore this gap by means of an example, which illustrates some new phenomena one should expect when Feller continuity is violated, even for very mixing chains. Our main vehicle is the weak convergence approach to large deviations [7], whose main advantage in the present context is that it allows one to understand intuitively how the Donsker-Varadhan rate function must be modified.

The example we give can easily be extended and generalized. However, our goal in this paper is to simply illustrate some of the possibilities, and in particular to examine the role that the Feller property plays in determining the form of the rate function. In the absence of any specific applications, we have forsaken the development of a general result.

2 The Example

We consider the Markov chain on the state space $\mathcal{X} = [0, 1]$, with initial position $X_0 = x \in [0, 1]$ and transition probability

$$p(x, da) = \begin{cases} U[0,1](da)\,, & x \in [1/2, 1] \\ U[\frac{x}{2} + \frac{1}{4}, \frac{3}{4} - \frac{x}{2}](da)\,, & x \in [0, 1/2)\,. \end{cases}$$

Here and in the sequel, $U[c, d]$ denotes the uniform distribution on $[c, d]$. As long as this process stays to the left of the point $1/2$, the distribution of the next step of the process looks more and more like a point mass concentrated on $1/2$. However, as soon as the process visits $[1/2, 1]$, the next step is distributed according to a uniform distribution. The Markov chain generated by $p(x, da)$ is readily seen to be strongly mixing, with a unique invariant measure. On the other hand, on an exponential scale, one can tilt the transition kernel when $x < 1/2$ in such a way that the Markov chain remains confined to the interval $[0, 1/2)$, and thereby force the empirical measure to converge to $\delta_{1/2}$. Since the cost of such tilting is (in exponential terms) finite, it is clear that such a possibility should be reflected in the LDP. More precisely, let $X_1, X_2, \ldots$ denote a realization of the Markov chain generated by $p(x, da)$, and let $L_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Our goal is to prove the following:

Theorem 1 *For $\mu \in \mathcal{M}_1(\mathcal{X})$, let $c \in [0, 1]$ be such that $\mu = c\delta_{1/2} + (1 - c)\nu$, where $\nu \in \mathcal{M}_1(\mathcal{X})$ satisfies $\nu(\{1/2\}) = 0$. Define*

$$I(\mu) = (1 - c) \inf_{\{\pi(x,da): \int \nu(dx)\pi(x,da) = \nu(da)\}} H(\nu(dx)\pi(x, da)|\nu(dx)p(x, da)) + c2\log 2\,. \qquad (1)$$

Then L_n satisfies the LDP with good rate function I.

Of course, in view of the results on the LDP for Markov chains described in the introduction, the interest in this particular Markov chain stems from the discontinuity of the statistics at $x = 1/2$. The explicit computation of the rate function, and the difference between the standard Donsker-Varadhan rate function and (1) are best explained in terms of the weak convergence approach to large deviations theory.

Proof of Theorem 1 : We recall the following known facts. A detailed discussion may be found in the book [7].

1. **Laplace principle** Assume that for each $f \in C_b(\mathcal{M}_1(\mathcal{X}))$,

   $$-\frac{1}{n}\log E_x e^{-nf(L_n)} \to_{n\to\infty} \inf_{\mu\in\mathcal{M}_1(\mathcal{X})} (I(\mu) + f(\mu)) \ ,$$

   and that $I(\cdot)$ is lower semicontinuous with respect to the weak topology on $\mathcal{M}_1(\mathcal{X})$. Then (c.f. [7, Theorem II.2.4]) L_n satisfies the LDP with the good rate function $I(\cdot)$.

2. **Representation formula** Let $\{\bar\nu^n_j(da|x,\ell), x \in \mathcal{X}, \ell \in \mathcal{M}_+(\mathcal{X}), j = 1,2,...,n\}$ denote a family of conditional probability distributions on $\mathcal{X}$, where $\mathcal{M}_+(\mathcal{X})$ denotes the space of non-negative Borel measures on $\mathcal{X}$. Define the Markov chain $\bar X^n_j$ by the relation $\bar X^n_0 = x$, $\bar P_x(\bar X^n_{j+1} \in A|\bar X^n_1, \dots, \bar X^n_j) = \bar\nu^n_j(A|\bar X^n_j, \bar L^n_j)$, where $\bar L^n_j = \frac{1}{n}\sum_{i=1}^{j}\delta_{\bar X^n_i}$, and $\bar L_n = \bar L^n_n$. Then (c.f. [7, Theorem V.2.2])

   $$-\frac{1}{n}\log E_x e^{-nf(L_n)} = \inf_{\{\bar\nu^n_j\}} \bar E_x\left(\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\nu^n_j(\cdot|\bar X^n_j, \bar L^n_j)|p(\bar X^n_j, \cdot)) + f(\bar L_n)\right) .$$

   Here $\bar E_x$ denotes expectation with respect to the measure $\bar P_x$. In the sequel we will not explicitly denote the dependence of $\bar\nu^n_j$ on $\bar X^n_j$ and $\bar L^n_j$.

For convenience, and since the proofs of the two parts are essentially different, we note that proving the LDP for L_n is equivalent to proving the lower semicontinuity of I and the following two bounds:

$$\liminf_{n\to\infty} \inf_{\{\bar\nu^n_j\}} \bar E_x\left(\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\nu^n_j(\cdot)|p(\bar X^n_j, \cdot)) + f(\bar L_n)\right) \geq \inf_{\mu\in\mathcal{M}_1(\mathcal{X})} (I(\mu) + f(\mu)) \ , \tag{2}$$

$$\limsup_{n\to\infty} \inf_{\{\bar\nu^n_j\}} \bar E_x\left(\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\nu^n_j(\cdot)|p(\bar X^n_j, \cdot)) + f(\bar L_n)\right) \leq \inf_{\mu\in\mathcal{M}_1(\mathcal{X})} (I(\mu) + f(\mu)) \ . \tag{3}$$

Note that under the lower semicontinuity of I the lower bound (2) is equivalent to the large deviations upper bound, whereas the upper bound (3) is equivalent to the large deviations lower bound [7].

As a preliminary step, we define an auxiliary Markov chain which will be needed later. Since $p(x, \{1/2\}) = 0$ for all x, $X_i \neq 1/2$ for all $i > 0$, w.p.1. Let $\mathcal{Z} = [0,1]^2$, let $\{X_i\}$ be as before, and define

$$Y_i = \begin{cases} 0\,, & X_{i-1} \geq 1/2 \\ \frac{X_i - 1/2}{X_{i-1} - 1/2} + \frac{1}{2}\,, & X_{i-1} < 1/2\,. \end{cases} \tag{4}$$

As explained above, we expect that with a probability that is not negligible in an exponential scale, the empirical measure L_n will have a component that approximates an atom at the point 1/2. In order to understand how likely this phenomena is from the large deviation perspective, we must examine in some detail the behavior of the chain near this point. To do this we follow a time honored method in weak convergence theory, that is, we simply tack on to the state vector the quantity of interest, and study the joint distributions when we take limits. It is easy to see that conditioned on $X_{i-1} < 1/2$, the marginal law of Y_i is $U[0,1]$. Because the transition function $p(x, da)$ approaches a point mass as x tends to 1/2 from below, the detailed behavior about this point of the X_i process is obscured. The process Y_i magnifies the behavior when X_i is in a neighborhood of 1/2, and analysis of the corresponding variable $\bar{Y}_i^n$ will allow us to understand the details of how mass can pile up around this point. Define $f(x,a) = 0$ when $x \geq 1/2$ and $f(x,a) = (a-1/2)/(x-1/2) + 1/2$ when $x < 1/2$. Then in fact $\{Z_i\} = \{(X_i, Y_i)\}$ forms a Markov chain with transition probability

$$q((x,y), da \times db) = q(x, da \times db) = p(x, da)\delta_{f(x,a)}(db). \tag{5}$$

Proof of the inequality (2), and the large deviation upper bound. Let $\bar{\nu}_j^n(da)$ be a family of conditional laws as described above, and let $\bar{X}_j^n$, $j = 1, \ldots, n$, denote the associated Markov chain. Let $\bar{\alpha}_j^n(da \times db)$ denote an induced conditional law on $\mathcal{Z}$ defined by

$$\bar{\alpha}_j^n(da \times db) = \bar{\nu}_j^n(da)\delta_{f(x,a)}(db). \tag{6}$$

Let the random vector generated by $\bar{\alpha}_j^n(da \times db)$ be denoted by $\bar{Z}_{j+1}^n = (\bar{X}_{j+1}^n, \bar{Y}_{j+1}^n)$, $j = 1, \ldots, n$. Thus $\bar{P}_x(\bar{Z}_{j+1}^n \in A | \bar{X}_j^n, ..., \bar{X}_1^n) = \bar{\alpha}_j^n(A | \bar{X}_j^n, \bar{L}_j^n)$. Define the random measures $\bar{\mu}^n = n^{-1}\sum_{j=1}^n \delta_{\bar{Z}_j^n}$, $\bar{L}^n = n^{-1}\sum_{j=1}^n \delta_{\bar{X}_j^n}$, and $\bar{\beta}^n(dx \times da \times db) = n^{-1}\sum_{j=1}^n \delta_{\bar{X}_j^n}(dx)\bar{\alpha}_j^n(da \times db)$. (We will also omit in the sequel the conditioning argument in $\bar{\alpha}_j^n$.) Then by compactness of $\mathcal{Z}$, one has, at least on subsequences,

$$\bar{\beta}^n \rightarrow_{n\rightarrow\infty} \bar{\beta}\,, \quad \bar{\mu}^n \rightarrow_{n\rightarrow\infty} \bar{\mu}\,, \quad \text{and } \bar{L}^n \rightarrow_{n\rightarrow\infty} \bar{L}$$

where all convergence is in distribution with respect to weak convergence of probability measures.

The definitions above imply $\bar{\beta}^n(dx \times [0,1]^2) = \bar{L}^n(dx)$. Since the mapping that takes a measure into its first marginal is continuous, $\bar{\beta}(dx \times [0,1]^2) = \bar{L}(dx)$. An analogous argument shows that $\bar{\mu}(dx \times [0,1]) = \bar{L}(dx)$. The same argument as the one used to show the existence of regular conditional probabilities shows that with probability one there is a regular conditional distribution $\bar{\beta}(da \times db | x)$ such that $\bar{\beta}(dx \times da \times db) = \bar{L}(dx)\bar{\beta}(da \times db | x)$. We claim that:

Lemma 1

$$\int \bar{\mu}(dx \times dy)\bar{\beta}(da \times db|x) = \bar{\mu}(da \times db) \quad \text{a.s.},$$

and therefore

$$\int \bar{L}(dx)\bar{\beta}(da \times [0,1]|x) = \bar{L}(da) \quad \text{a.s.}$$

Proof of Lemma 1 : The proof is a rerun of the arguments in [7]. Let $g : \mathcal{Z} \to \mathbb{R}$ be a bounded measurable function. Note that with $\mathcal{F}_j^n = \sigma(\{\bar{X}_1^n, \ldots, \bar{X}_j^n\})$,

$$\bar{E}_x(g(\bar{Z}_{j+1}^n)|\mathcal{F}_j^n) = \int g(a,b)\bar{\alpha}_j^n(da \times db)\,.$$

Owing to the boundedness of g and the conditional independence of the summands,

$$\begin{array}{lcr} \displaystyle\int g(a,b)\bar{\mu}_n(da \times db) - \int\int g(a,b)\bar{\beta}^n(dx \times da \times db) & = & \\ \displaystyle\frac{1}{n}(g(\bar{Z}_0^n) - g(\bar{Z}_n^n)) + \frac{1}{n}\sum_{j=0}^{n-1}\left(g(\bar{Z}_{j+1}^n) - \bar{E}_x(g(\bar{Z}_{j+1}^n)|\mathcal{F}_j^n)\right) & \to_{n\to\infty} & 0\,, \end{array} \tag{7}$$

where the limit holds in probability. We conclude that for each such g, we have

$$\int g(a,b)\bar{\mu}(da \times db) = \int\int g(a,b)\bar{\beta}(dx \times da \times db) = \int\int g(a,b)\bar{\mu}(dx \times dy)\bar{\beta}(da \times db|x)$$

except on a set of measure zero. If we observe that the last equality holds with probability one for all g in a separating family of continuous functions, we obtain the first conclusion of the lemma. The second conclusion follows from the first by integrating out the b variable. □

Let $\rho(\cdot)$ denotes the uniform density on $[0,1]$. We next define

$$\tilde{q}(x, da \times db) = \begin{cases} q(x, da \times db)\,, & x \neq 1/2 \\ \delta_{1/2}(da)\rho(db)\,, & x = 1/2\,. \end{cases}$$

The transition function $\tilde{q}$ is essentially the "left continuous regularization" of q. This is the function that should be relevant in any calculation involving weak limits, since all the mass at $1/2$ in the limit must come from the left. Recall that $\bar{\mu}^n$ is the empirical distribution of the pair $(\bar{X}_i^n, \bar{Y}_i^n)$, and that $\bar{\mu}^n \to \bar{\mu}$ in distribution. We will see that the new term in the rate function can be expressed in terms of the relative entropy of the conditional law of the second variable in $\bar{\mu}$ (given that the first variable equals $1/2$) with respect to the distribution ρ. In order for the mass to accumulate at the point $1/2$, it will be necessary that this conditional distribution not equal ρ. In fact, we will see shortly that this measure must be entirely supported on $[0,1/2]$. Consequently, the contribution of this new term will be non-zero.

We now claim:

Lemma 2 *Assume that* $\sup_n \bar{E}_x H(\bar{\beta}^n|\bar{L}^n \otimes q) < K < \infty$. *Then*

$$\bar{L}^n(dx)q(x, da \times db) \to_{n\to\infty} \bar{L}(dx)\tilde{q}(x, da \times db)\,.$$

Proof of Lemma 2 : By the continuity of $q(x, da \times db)$ in x away from $x = 1/2$, it is obvious that one needs only check the behavior at $x = 1/2$. Let $A^\delta = [1/2 - \delta, 1/2 + \delta]$ and $C^\delta = A^\delta \times \{0\}$. Mass will "pile up" at the point $1/2$ if $\lim_{\delta\to 0} \limsup_{n\to\infty} \bar\mu^n(A^\delta \times [0,1]) > 0$. We now prove that, in probability,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \bar\mu^n(C^\delta) = 0\,. \tag{8}$$

According to equation (4), the set $\mathcal{X} \times \{0\}$ is given a positive mass by $(\bar X_i^n, \bar Y_i^n)$ if and only if $\bar X_i^n \geq 1/2$. Thus the last display implies that all the mass that piles up at $1/2$ comes from the left. Since $q(x, da \times db) \to \tilde q(1/2, da \times db)$ when $x \nearrow 1/2$, this will imply the lemma. It is easy to check that for all $x > 1/2$, $q(x, C^\delta) \leq 2\delta$. By using the Skorokhod representation [8], we can assume that $\bar\mu^n \to \bar\mu$ and $\bar\beta^n \to \bar\beta$ with probability one for purposes of calculating the limit of expectations involving these quantities. We see from equation (7) that

$$\frac{1}{n}\sum_{j=1}^{n} \bar\alpha_j^n(da \times db) = \bar\beta^n([0,1] \times da \times db) \to \bar\mu(da \times db)$$

w.p.1 in the weak topology. This implies that

$$\lim_{n\to\infty} \bar P_x\left(\frac{1}{n}\sum_{j=1}^{n} \bar\alpha_j^n(C^\delta) \geq \bar\mu(C^{\delta/2}) - \delta\right) = 1.$$

Hence if (8) does not hold, then for δ small enough there exist an N and positive constants γ_0, p_0 independent of δ such that for all $n > N$,

$$P\left(\frac{1}{n}\sum_{j=1}^{n} \bar\alpha_j^n(C^\delta) > \gamma_0\right) > p_0\,.$$

Let $N_R(n) = \{j = 1, 2, \ldots, n : \bar X_j^n > 1/2\}$. Then, for $n > N$,

$$P\left(|N_R(n)|/n > \gamma_0, \frac{1}{|N_R(n)|}\sum_{j\in N_R(n)} \bar\nu_j^n(C^\delta) > \gamma_0\right) > p_0\,.$$

Let $H(\mu(C^\delta)|\nu(C^\delta))$ denote the relative entropy between the restrictions of μ, ν to $C^\delta, (C^\delta)^c$. By the nonnegativity and convexity of H, for $n > N$,

$$\begin{aligned}
\bar E_x H(\bar\beta^n | \bar L^n \otimes q) &= \bar E_x\left[\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\alpha_j^n(\cdot)|q(\bar X_j, \cdot))\right] \\
&\geq \bar E_x\left[\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\alpha_j^n(C^\delta)|q(\bar X_j, C^\delta))\right] \\
&\geq \bar E_x\left[\frac{|N_R(n)|}{n} H\left(\frac{1}{|N_R(n)|}\sum_{j\in N_R(n)} \bar\alpha_j^n(C^\delta) \,\middle|\, \frac{1}{N_R(n)}\sum_{j\in N_R(n)} q(\bar X_j, C^\delta)\right)\right] \\
&\geq p_0 k \gamma_0 \log(\gamma_0/2\delta) - k,
\end{aligned}$$

where in the last inequality the boundedness below of the function $x \log x$ is used, and k is some constant independent of δ. Taking δ small enough, the last display contradicts the assumption $\sup_n \bar{E}_x H(\bar{\beta}^n | \bar{L}^n \otimes q) < K < \infty$. We conclude that equation (8) holds. □

We next consider an important property of the measures $\bar{\beta}(da \times db|x)$. The property is another expression of the fact that all the mass at 1/2 in the limit arrives from the left.

Lemma 3 *Assume that* $\sup_n \bar{E}_x H(\bar{\beta}^n | \bar{L}^n \otimes q) < K < \infty$, *and let* $\bar{\beta}^n \rightarrow \bar{\beta}$. *Then for some (random) probability measure* Ψ *on* $[0,1]$ *with* $\Psi((1/2,1]) = 0$, $\bar{\beta}(da \times db | 1/2) = \delta_{1/2}(da)\Psi(db)$ *w.p.1.*

Proof of Lemma 3 : We again invoke the Skorohod representation and assume that the convergences are all w.p.1. We have the following inequalities, each of which is explained below.

$$\begin{aligned}
\bar{\beta}(\{1/2\} \times [0,1] \times (1/2,1]) &= \bar{\beta}(\{1/2\} \times \{1/2\} \times (1/2,1]) \\
&\leq \lim_{\delta \rightarrow 0} \bar{\beta}([0,1] \times (1/2-\delta, 1/2+\delta) \times (1/2,1]) \\
&\leq \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \bar{\beta}^n([0,1] \times (1/2-\delta, 1/2+\delta) \times (1/2,1]) \\
&= \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \bar{\beta}^n([0,1] \times (1/2, 1/2+\delta) \times (1/2,1]) \\
&= \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \bar{\mu}^n((1/2, 1/2+\delta) \times (1/2,1]) \\
&\leq \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \bar{\mu}^n((1/2, 1/2+\delta) \times [0,1]) \\
&\leq \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \bar{\mu}^n(C^\delta) \\
&= 0.
\end{aligned}$$

The lower semicontinuity of $H(\cdot|\cdot)$, Lemma 2, and the assumption that $\sup_n \bar{E}_x H(\bar{\beta}^n | \bar{L}^n \otimes q) < K < \infty$ together imply $H(\bar{\beta}|\bar{L} \otimes \tilde{q}) < \infty$ and therefore $\bar{\beta} << \bar{L} \otimes \tilde{q}$ w.p.1. Since $\tilde{q}(1/2, da \times [0,1]) = \delta_{1/2}(da)$, we see that $\bar{\beta}(da \times db|1/2)$ takes the form $\delta_{1/2}(da)\Psi(db)$, which gives the first equality. The next inequality is obvious, and the second inequality follows from the weak convergence $\bar{\beta}^n \rightarrow \bar{\beta}$. The second equality is due to the fact that $q(x, [0,1/2] \times [1/2,1]) = 0$ for all $x \in [0,1]$. The third equality is due to (7), and all succeeding inequalities are obvious given the limit (8). The lemma now follows from the fact that $\bar{\beta}(da \times db|1/2)$ has the form $\delta_{1/2}(da)\Psi(db)$. □

We now put these facts together. Suppose that $\bar{L} = c\delta_{1/2} + (1-c)\nu$, where both ν and c may be random, $c \in [0,1]$, and $\nu(\{1/2\}) = 0$. Using the lower semicontinuity of H we obtain

$$\liminf_{n \rightarrow \infty} \bar{E}_x H(\bar{\beta}^n | \bar{L}^n \otimes q) \geq \bar{E}_x H(\bar{\beta}|\bar{L} \otimes \tilde{q}) = \bar{E}_x \int H(\bar{\beta}(da \times db|x) | \tilde{q}(x, da \times db)) \bar{L}(dx).$$

We consider the last integral over the sets $\{1/2\}$ and $[0,1] \backslash \{1/2\}$. According to Lemma 3 $\bar{\beta}(da \times db|x) = \delta_{1/2}(da)\Psi(db)$ w.p.1, where $\Psi(db)$ is supported on $[1/2,1]$ w.p.1. The infimum of $H(\cdot|\rho)$ over all such measures occurs at the measure $U[1/2,1]$, with relative entropy $H(U[1/2,1]\|\rho) = 2\log 2$. Thus

$$\int_{\{1/2\}} H(\bar{\beta}(da \times db|x) | \tilde{q}(x, da \times db)) \bar{L}(dx) \geq c 2 \log 2.$$

We next consider $x \neq 1/2$. Define the random transition kernel $\pi(x,da) = \bar{\beta}(da \times [0,1]|x)$. Then $\bar{\beta}(da \times db|x) << p(x,da)\delta_{f(x,a)}(db)$ implies $\bar{\beta}(da \times [0,1]|x) = \pi(x,da)\delta_{f(x,a)}(db)$ ($\bar{L}-$a.s., w.p.1). We can therefore w.p.1 write

$$\begin{aligned}\int_{[0,1]\setminus\{1/2\}} H(\bar{\beta}(da \times db|x)|\tilde{q}(x, da \times db))\bar{L}(dx) &= \int_{[0,1]\setminus\{1/2\}} H(\pi(x,da)|p(x,da))\bar{L}(dx) \\ &= (1-c)H(\nu(dx)\pi(x,da)|\nu(dx)p(x,da)).\end{aligned}$$

Recall that by Lemma 1, $\bar{L}$ is an invariant measure for π. Because of the definition $\pi(x,da) = \beta(da \times [0,1]|x)$ and the equality $\beta(da \times db|1/2) = \delta_{1/2}(da)\Psi(db)$, we see $\pi(1/2,da) = \delta_{1/2}(da)$. Hence π decomposes $\mathcal{X}$ into $[0,1/2) \cup (1/2,1]$ and $\{1/2\}$. When combined with the fact that $\bar{L}$ is invariant under π and $\nu(\{1/2\}) = 0$, this implies that both $\delta_{1/2}$ and ν are invariant under π. The definition (1) then implies

$$(1-c)H(\nu(dx)\pi(x,da)|\nu(dx)p(x,da)) + cH(\Psi|\rho) \geq I(\bar{L}),$$

w.p.1. It follows from (5) and (6) and the convexity of the relative entropy function that

$$\frac{1}{n}\sum_{j=1}^{n-1} H(\bar{\nu}_j^n(da)|p(\bar{X}_j^n, da)) = \frac{1}{n}\sum_{j=1}^{n-1} H(\bar{\alpha}_j^n(da \times db)|q(\bar{X}_j^n, da \times db)) \geq H(\bar{\beta}^n|\bar{L}^n \otimes q).$$

Hence the continuity of f and the last five displays allow one to calculate

$$\liminf_{n\to\infty} \inf_{\{\bar{\nu}_j^n\}} \bar{E}_x\left(\frac{1}{n}\sum_{j=1}^{n-1} H(\bar{\nu}_j^n(\cdot)|p(\bar{X}_j^n,\cdot)) + f(\bar{L}_n)\right) \geq \bar{E}_x\left(I(\bar{L}) + f(\bar{L})\right) \geq \inf_{\mu\in\mathcal{M}_1(\mathcal{X})} (I(\mu) + f(\mu)),$$

which completes the proof of (2).

Proof of the lower semicontinuity of $I(\cdot)$. We next prove that I defined on $\mathcal{M}_1(\mathcal{X})$ in (1) is lower semicontinuous. As always with the weak convergence approach, the proof of lower semicontinuity is essentially a deterministic version of the proof of the large deviation upper bound. Let $\mu^n \to \mu = (1-c)\nu + c\delta_{1/2}$ with $\nu(\{1/2\}) = 0$, and for simplicity assume that $\mu^n(\{1/2\}) = 0$. (The general case poses only notational difficulties). Assume that $I(\mu^n) \leq K$, i.e.,

$$\begin{aligned} K &\geq \inf_{\{\pi(x,da):\int \mu^n(dx)\pi(x,da)=\mu^n(da)\}} H(\mu^n(dx)\pi(x,da)|\mu^n(dx)p(x,da)) \qquad (9) \\ &= \inf_{\{\pi(x,da):\int \mu^n(dx)\pi(x,da)=\mu^n(da)\}} H(\mu^n(dx)\pi(x,da)\delta_{f(x,a)}(db)|\mu^n(dx)p(x,da)\delta_{f(x,a)}(db)).\end{aligned}$$

(We recall the definition $f(x,a) = 0$ if $x \geq 1/2$ and $f(x,a) = (a-1/2)/(x-1/2) + 1/2$ if $x < 1/2$.)

By essentially the same argument as in the proof of Lemma 2, one sees that

$$\mu^n(dx)p(x,da)\delta_{f(x,a)}(db) \to_{n\to\infty} \mu(dx)\tilde{q}(x, da \times db).$$

Next let π^n be an approximate minimizer in (9). By this we mean that for some sequence $\epsilon_n \to_{n\to\infty} 0$ we choose $\pi^n(x, da)$ to satisfy

$$\begin{aligned} \inf_{\{\pi(x,da):\int \mu^n(dx)\pi(x,da)=\mu^n(da)\}} \quad & H(\mu^n(dx)\pi(x,da)|\mu^n(dx)p(x,da)) \\ & \geq H(\mu^n(dx)\pi^n(x,da)|\mu^n(dx)p(x,da)) - \epsilon_n \, , \end{aligned}$$

where $\int \mu^n(dx)\pi^n(x,da) = \mu^n(da)$. Let $\theta^n(dx \times da) = \mu^n(dx)\pi^n(x,da)$. By compactness, at least along a subsequence we will have $\theta^n(dx \times da) \to_{n\to\infty} \theta(dx \times da) = \mu(dx)\pi(x,da)$, with $\theta(A \times [0,1]) = \theta([0,1] \times A) = \mu(A)$ for all measurable $A \subset [0,1]$. Thus μ is invariant under π. Moreover, again by compactness, $\theta_n(dx \times da)\delta_{f(x,a)}(db) \to_{n\to\infty} \bar{\beta}(dx \times da \times db)$ for some probability measure $\bar{\beta}$. By the lower semicontinuity of $H(\cdot|\cdot)$,

$$\liminf_{n\to\infty} H(\theta_n(dx \times da)\delta_{f(x,a)}(db)|\mu^n(dx)p(x,da)\delta_{f(x,a)}(db)) \geq H(\bar{\beta}(dx \times da \times db)|\mu(dx)\tilde{q}(x, da \times db)) \, .$$

If $c = \mu(\{1/2\}) > 0$, then $\bar{\beta}(\{1/2\}, da \times db)$ must be absolutely continuous with respect to $\delta_{1/2}(da)$ $\rho(db)$. Hence, $\bar{\beta}(\{1/2\} \times da \times db) = c\delta_{1/2}(da)\Psi(db)$ where Ψ is a probability measure. By essentially the same argument as that used in the proof of Lemma 3, $\Psi((1/2, 1]) = 0$. It follows by convexity that

$$H(\bar{\beta}(dx \times da \times db)|\mu(dx)\tilde{q}(x, da \times db)) \geq (1 - c)H(\nu(dx)\pi(x,da)|\nu(dx)p(x,da)) + cH(\Psi|\rho) \, .$$

Now since θ equals the first two marginals of $\bar{\beta}$, $\bar{\beta}(\{1/2\} \times da \times db) = \delta_{1/2}(da)\Psi(db)$ implies $\pi(1/2, da) = \delta_{1/2}(da)$. Now we use the facts that μ is invariant under π and that $\nu(\{1/2\}) = 0$ to conclude that ν is also invariant under π. According to the definition (1), this implies

$$(1 - c)H(\nu(dx)\tilde{\pi}(x,da)|\nu(dx)p(x,da)) + cH(\Psi|\rho) \geq I(\mu),$$

and the lower semicontinuity follows.

Proof of the inequality (3), and the large deviation lower bound. Let $f \in C_b(\mathcal{M}_1(\mathcal{X}))$ be given. We assume that the right hand side of (3) is finite, for otherwise there is nothing to prove. Let $\mu = c\delta_{1/2} + (1 - c)\nu$ be a minimizer in the right hand side of (3), with $\nu(\{1/2\}) = 0$ and $I(\mu) < \infty$. It is easy to check that ν is absolutely continuous with respect to Lebesgue measure. Indeed, suppose that $\pi(x, da)$ has ν as an invariant measure and that $H(\nu(dx)\pi(x,da)|\nu(dx)p(x,da)) < \infty$. Then $\pi(x, \cdot)$ is absolutely continuous with respect to $p(x, \cdot)$, and hence also Lebesgue measure, for $\nu-$almost every x. If the Lebesgue measure of A is zero, then $\nu(A) = \int \pi(x, A)\nu(dx) = 0$, which shows that ν is absolutely continuous with respect to Lebesgue measure.

To simplify the notation, we assume $1 > c \geq 0$. The general case is quite similar. For $\delta > 0$ let $\pi_\delta(x, dy)$ be a transition kernel such that ν is invariant under π_δ and

$$H(\nu(dx)\pi_\delta(x,da)|\nu(dx)p(x,da)) \leq (I(\mu) - 2c\log 2) / (1 - c) + \delta.$$

A slight difficulty is that ν is not necessarily an ergodic measure for $\pi_\delta(x, da)$, and hence the Markov chain generated by the latter might not have empirical measure converging to ν. Consider the family of transition kernels and measures ν^ε and π_δ^ε defined by

$$\nu^\varepsilon(da) = (1 - \varepsilon)\nu(da) + \varepsilon\nu^*(da),$$

$$\nu^\varepsilon(dx)\pi^\varepsilon_\delta(x,da) = (1-\varepsilon)\nu(dx)\pi_\delta(x,da) + \varepsilon\nu^*(dx)p(x,da),$$

where $\varepsilon \in [0,1]$ and ν^* is the unique invariant measure of $p(x,da)$. Note that because the x marginals of both sides of the last equality are equal, such a π^ε_δ is well defined for Lebesgue-almost all x, and the definition is extended to all x in such a way as to make $\pi^\varepsilon_\delta(x,da)$ dominate $p(x,da)$ for all x. The following facts are easily shown. For full details the interested reader can consult [7, Lemma IX.6.3]. The first item is true because $p(x,da)$ is ergodic, while the second follows from the convexity of $H(\cdot|\cdot)$.

- For each $\varepsilon > 0$, the kernel $\pi^\varepsilon_\delta(x,da)$ is ergodic, with unique invariant measure ν^ε.
- For each $\varepsilon \in [0,1]$, $H(\nu^\varepsilon(dx)\pi^\varepsilon_\delta(x,da)|\nu^\varepsilon(dx)p(x,da)) \leq H(\nu(dx)\pi_\delta(x,da)|\nu(dx)p(x,da))$.

We now fix $\varepsilon > 0$ such that $d(\nu,\nu^\varepsilon) \leq \delta$. If we choose the measures $\bar\nu^n_j(\cdot)$ to equal $\pi^\varepsilon_\delta(\bar X^n_j,\cdot)$, then the L^1 ergodic theorem implies that there exists an N such that for all $x \in [0,1]$ and all $n > (1-c)N$,

$$d\left(\bar E_x\left(\frac{1}{n}\sum_{j=1}^n \delta_{\bar X^i_j}\right), \nu^\varepsilon\right) < \delta\,,$$

and moreover

$$\bar E_x\left(\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\nu^n_j(\cdot)|p(\bar X^n_j,\cdot))\right) \leq \delta + H(\nu(dx)\pi_\delta(x,da)|\nu(dx)p(x,da))\,.$$

Next, let

$$\bar p(x,da) = \begin{cases} U[0,1/2]\,, & x \in [1/2,1] \\ U[\frac{x}{2}+1/4,1/2]\,, & x \in [0,1/2)\,. \end{cases}$$

If we choose the measures $\bar\nu^n_j(\cdot)$ to equal $\bar p(\bar X^n_j,\cdot)$, then by increasing N if need be, we obtain for all $x \in [0,1]$ and all $n > (1-c)N$

$$\bar P_x\left(d\left(\frac{1}{n}\sum_{j=1}^n \delta_{\bar X^i_j}, \delta_{1/2}\right) > \delta\right) = 0\,,$$

and the cost

$$\bar E_x\left(\frac{1}{n}\sum_{j=0}^{n-1} H(\bar\nu^n_j(\cdot)|p(\bar X^n_j,\cdot))\right) = 2\log 2\,.$$

We next define a time inhomogeneous Markov chain. If j is of the form $j = k(N+2)+\ell$, where $\ell \in \{1,\ldots,N+2\}$ and $k = 0,1,\ldots$, then we set

$$\bar\nu^n_j(da) = \begin{cases} \begin{array}{ll} U[1/2,3/4-\bar X^n_j/2] & \text{if } \bar X^n_j < 1/2, \\ U[1/2,1] & \text{if } \bar X^n_j \geq 1/2, \end{array} & \ell = 0, & \text{phase } 1, \\ U[0,1], & \ell = 1, & \text{phase } 2, \\ \pi^\varepsilon_\delta(\bar X^n_j,da)\,, & \ell = 2,\ldots,2+(1-c)N\,, & \text{phase } 3\,, \\ \bar p(\bar X^n_j,da)\,, & \ell = 2+(1-c)N,\ldots,N+1\,, & \text{phase } 4. \end{cases}$$

(We have assumed cN to be an integer, the modification required if it is not is straightforward). During phase 1, the chain is positioned somewhere in $[1/2, 1]$ at a cost of $2\log 2$. During phase two, the chain is redistributed according to $U[0,1]$ at zero cost. Let η_k^n denote the occupation measure of the chain during the third phase of the kth cycle:

$$\eta_k^n = \frac{1}{(1-c)N} \sum_{\ell=2}^{2+(1-c)N} \delta_{\bar{X}^n_{k(N+2)+\ell}}.$$

Because phase 2 guarantees that during each cycle phase 3 is started from the uniform distribution, the random variables $\{\eta_k^n, k = 1, ..., [n/(N+2)]\}$ are independent and identically distributed, where $[a]$ denotes the integer part of a. By the law of large numbers,

$$\limsup_{n\to\infty} d\left(\frac{1}{[n/(N+2)]} \sum_{k=1}^{[n/(N+2)]} \eta_k^n, \nu^\varepsilon\right) \leq \limsup_{n\to\infty} d\left(\frac{1}{[n/(N+2)]} \sum_{k=1}^{[n/(N+2)]} \eta_k^n, E\eta_1^n\right) + d(E\eta_1^n, \nu^\varepsilon) < \delta$$

in probability. The chain spends a period of length cN in phase 4 of the kth cycle, during which the empirical measure is controlled by $\bar{p}$ and hence converges to within δ of $\delta_{1/2}$, w.p.1.

By taking N large, we can guarantee that the contributions due to phases 1 and 2 to $\bar{L}_n$ are smaller than δ in the total variation norm. Thus for sufficiently large N, $\limsup_{n\to\infty} d(\bar{L}_n, \mu) \leq 3\delta$ in probability. When combined with the given bounds on the costs, we obtain

$$\limsup_{n\to\infty} \bar{E}_x \left(\frac{1}{n} \sum_{j=0}^{n-1} H(\bar{\nu}_j^n(\cdot)|p(\bar{X}_j^n, \cdot)) + f(\bar{L}_n) \right) \leq I(\mu) + \delta + 2\log 2/N + \sup_{\bar{\mu}:d(\bar{\mu},\mu)\leq 3\delta} f(\bar{\mu})\,.$$

Inequality (3) follows since $N < \infty$ and $\delta > 0$ are arbitrary. This completes the proof of the theorem. □

References

[1] W. Bryc and A. Dembo, Large deviations and strong mixing, *preprint*, 1993.

[2] A. de Acosta, Large deviations for empirical measures of Markov chains, *J. Theo. Prob.*, 3 (1990), pp. 395–431.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, 1993.

[4] J.D. Deuschel and D.W. Stroock, *Large Deviations*, Academic Press, 1989.

[5] I. Dinwoodie, Identifying a large deviation rate function, *Ann. Probab* 21 (1993), pp. 216–231.

[6] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time-III. *Comm. Pure Appl. Math.* 24 (1976), pp. 389–461.

[7] P. Dupuis and R.S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1995.

[8] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[9] P. Ney and E. Nummelin, Markov additive processes II. Large deviations. *Ann. Probab.* 15 (1987), pp. 593– 609.