Relative Entropy and Exponential Deviation Bounds for General Markov Chains

I. Kontoyiannis
Div of Applied Mathematics
& Dpt of Computer Science
Brown University
Providence, RI 02912, USA
Email: yiannis@dam.brown.edu

L.A. Lastras-Montaño
IBM TJ Watson Research Center
1101 Kitchawan Rd
Yorktown Heights, NY, 10598
Email: lastrasl@us.ibm.com

S.P. Meyn
Dept of Electrical & Computer Eng
& Coordinated Sciences Laboratory
University of Ill at Urbana-Champaign
Urbana, IL 61801, USA
Email: meyn@uiuc.edu

***Abstract*— We develop explicit, general bounds for the probability that the normalized partial sums of a function of a Markov chain on a general alphabet will exceed the steady-state mean of that function by a given amount. Our bounds combine simple information-theoretic ideas together with techniques from optimization and some fairly elementary tools from analysis. In one direction, we obtain a general bound for the important class of Doeblin chains; this bound is *optimal*, in the sense that in the special case of independent and identically distributed random variables it essentially reduces to the classical Hoeffding bound. In another direction, motivated by important problems in simulation, we develop a series of bounds in a form which is particularly suited to these problems, and which apply to the more general class of "geometrically ergodic" Markov chains.**

I. Introduction

A central computational problem in various scientific contexts is the computation of the expected value $E_\pi(F)$ of a function F under some probability distribution π. In practice it is often the case that, although π may be known explicitly, its computation is impossible for all practical purposes. For example, in areas such as Bayesian statistics, image processing, and statistical mechanics, this is the rule rather than the exception; see [2][12][16] and the references therein.

One of the most common solutions to this problem is the use of Markov Chain Monte Carlo (MCMC) techniques. There, the expectation of interest is estimated by the sample average

$$S_n = \frac{1}{n}\sum_{i=0}^{n-1} F(X_i), \tag{1}$$

where the sequence $\{X_n\ ;\ n \geq 0\}$ is a Markov chain which is known to have stationary distribution π. Usually, under appropriate conditions it can be easily verified that the Markov chain $\{X_n\}$ converges to its stationary distribution π, and by the law of large numbers we are assured that the sample averages in (1) will indeed converge to the expected value $E_\pi(F)$ of F under π, as $n \to \infty$. Moreover, the central limit theorem provides a rate for this convergence, stating that

$$\sqrt{n}\big[S_n - E_\pi(F)\big] \xrightarrow{\mathcal{D}} N(0,\sigma^2),$$

where $\sigma^2 = \lim_n \frac{1}{n}\text{Var}(S_n)$ is the asymptotic variance of F.

Such asymptotic results may be of limited use if we want to have some sort of guarantee that, after a certain number of steps in the simulation, the estimate S_n will indeed be close to $E_\pi(F)$. There are several well-studied approaches in the very extensive literature on this subject. Partly motivated by this discussion, here we consider the problem of providing simple, computable bounds for probabilities of "large deviations" type,

$$\Pr\Big\{\frac{1}{n}\sum_{i=0}^{n-1} F(X_i) \geq E_\pi(F) + \epsilon\Big\}, \tag{2}$$

for general classes of Markov chains $\{X_n\}$ and functions F.

Information-theoretic methods have been very influential in the development of asymptotic results as well as non-asymptotic inequalities for probabilities as in (2). When $\{X_n\}$ are of independent and identically distributed (i.i.d.) random variables, the combinatorial techniques based on the method of types have provided some of the simplest proofs as well as some of the strongest results [5][6][4]. But in the case of Markov chains they have been much less successful, their applicability limited essentially only to the more elementary case of finite state-space Markov chains; see [7] and the references therein. This difficulty reflects, to some extent, the intrinsic limitations of the information-theoretic methods, but it is also due to the much greater complexity of the field of large deviations for general Markov chains. For example, the elegant theorem of Sanov which holds in complete generality in the i.i.d. case, requires extremely strong assumptions in order to be translated to Markov chains; see Donsker and Varadhan's classic results [15], as well as the numerous counter-examples indicating that such strong assumptions are indeed necessary, e.g., [1][3].

In this work we combine simple information-theoretic ideas together with techniques from optimization and some fairly elementary tools from analysis, to obtain explicit, non-asymptotic bounds for the probabilities (2). In Section II we derive a natural information-theoretic bound stating that the probability in (2) is always bounded above by the exponential of an expression in terms of relative entropy. In Section III we specialize to the class of Doeblin chains and give an explicit bound, and in Section IV we show how information about the rate at which a finite-state chain converges to equilibrium can be used to obtain potentially better bounds.

Finally in Section V we briefly describe a series of associated extensions of these results and discuss their applications to more general simulation settings.

II. A General Information-Theoretic Bound

Consider a discrete-time Markov chain $\{X_n\ ;\ n \geq 0\}$ with values in the state space A. For simplicity we concentrate here in the case when A is countable. The distribution of $\{X_n\}$ is determined by its initial state $X_0 = x_0 \in A$ and its transition kernel $P(i,j) = P_{ij} = \Pr\{X_n = j | X_{n-1} = i\}$, $i, j \in A$. We write P_x for the distribution of the chain conditional on the initial state $X_0 = x$, and E_x for the corresponding expectation.

The relative entropy between two probability distributions P, Q on A is defined as usual by $H(P\|Q) = \sum_{i\in A} P_i \log \frac{P_i}{Q_i}$. Pinsker's inequality relates the relative entropy to the L^1-distance; for any P, Q,

$$H(P\|Q) \geq \frac{1}{2}\|P - Q\|^2, \tag{3}$$

where the L^1-norm is twice the total variation distance,

$$\|P - Q\| = 2 \sup_{E \subset A} |P(E) - Q(E)| = \sum_{i \in A} |P_i - Q_i|.$$

We begin with a simple and somewhat striking observation due to Csiszár [5]. It states that the probability of *any* event can be expressed as the exponential of a relative entropy. The proof is immediate from the definitions.

Lemma 1. (Csiszár's Lemma) Let p be an arbitrary probability distribution on any probability space, and E any event with $p(E) > 0$. Let $p|_E$ denote the conditional distribution $p|_E(\cdot) = p(\cdot \cap E)/p(E)$. Then:

$$-\log p(E) = H(p|_E \| p).$$

Next we obtain a general upper bound for the probability of deviations of the partial sums of $\{X_n\}$. Its proof, relying on little more than the above lemma and Jensen's inequality, is inspired by an argument used by Csiszár in the proof of [5, Theorem 1].

Proposition 1. For any function $F : A \to \mathbb{R}$ which is bounded above, any $c > 0$ and any initial condition $x \in A$, we have

$$-\log \mathsf{P}_x\Big\{\sum_{i=0}^{n-1} F(X_i) \geq nc\Big\} \geq (n-1)H(W\|W^1 \odot P),$$

where W is a bivariate distribution on $A \times A$ with marginals W^1 and W^2, $W^1 \odot P$ is simply the bivariate distribution $(W^1 \odot P)(x,y) = W^1(x)P(x,y)$, and where W^1, W^2 satisfy,

$$\|W^1 - W^2\| \leq \frac{2}{n-1} \quad \& \quad E_{W^1}(F) \geq c - \frac{a}{n-1},$$

where $a = \sup_{x\in A} F(x)$.

A more general version of Proposition 1 is given in Proposition 2 in Section V so we postpone its proof until then.

Remark. The classical extension of Sanov's theorem to Markov chains is Donsker and Varadhan's large deviations principle [15]. It states that, under appropriate conditions on the Markov chain and on F, as $n \to \infty$ we have,

$$-\log \mathsf{P}_x\Big\{\sum_{i=0}^{n-1} F(X_i) \geq nc\Big\} \approx n \inf_W H(W\|W^1 \odot P), \tag{4}$$

where the infimum is over all bivariate distributions W with marginals W^1 and W^2 that satisfy

$$W^1 = W^2 \;\;\&\;\; E_{W^1}(F) \geq c.$$

The above proposition gives a non-asymptotic version of the upper bound in (4), and offers an elementary "explanation" for the Donsker-Varadhan rate function.

III. Doeblin Chains

Next we go on to obtain a nontrivial large deviations bound by establishing a lower bound on the relative entropy appearing in Proposition 1. In this section we concentrate on what is probably the "nicest" class of general-alphabet Markov chains. These are the chains that converge to equilibrium exponentially fast, uniformly in the initial condition. This is formalized by requiring that $\{X_n\}$ has a unique stationary distribution π such that

$$d_n = \sup_x \|P^n(x,\cdot) - \pi\| \to 0 \quad \text{exponentially fast as } n \to \infty.$$

In fact this is equivalent to the seemingly weaker condition that there exists some $n \geq 1$ for which $d_n < 2$. An important and very useful feature of Doeblin chains is that, whether or not a given chain belongs to this class, can be readily verified via "Doeblin's minorization condition." This states that there exists an integer $m \geq 1$, an $\alpha > 0$ and a probability distribution ρ on A such that

$$P^m(x, E) \geq \alpha\rho(E), \quad \text{for all } x \in A,\ E \subset A. \tag{5}$$

See [13] for a detailed discussion. Additional motivation for considering Doeblin chains is given in Remark 3 below.

Theorem 1. Suppose the Markov chain $\{X_n\}$ satisfies the Doeblin condition (5) and has stationary distribution π. For any bounded function $F : A \to \mathbb{R}$ and any $\epsilon > 0$ we have,

$$\log \mathsf{P}_x\Big\{\sum_{i=0}^{n-1}[F(X_i) - E_\pi(F)] \geq n\epsilon\Big\} \leq -\frac{n-1}{2}\Big[\frac{\alpha}{m\overline{F}}\,\epsilon - \frac{3}{n-1}\Big]^2,$$

as long as $n \geq 1 + 3m\overline{F}/(\alpha\epsilon)$, where $\overline{F} = \sup_x |F(x)|$.

Remarks.

1) Note that, if $\{X_n\}$ is a sequence of i.i.d. random variables with common distribution π, then the Doeblin condition holds with $m = \alpha = 1$ and $\rho = \pi$, and the bound of Theorem 1 reduces to

$$-\frac{n-1}{2}\Big[\frac{\epsilon}{\overline{F}} - \frac{3}{n-1}\Big]^2.$$

This is essentially identical to the classical Hoeffding bound [9],

$$-\frac{n}{2}\Big(\frac{\epsilon}{\overline{F}}\Big)^2,$$

which is known to be tight in the i.i.d. case.

2) Theorem 1 improves upon a recent result of Glynn and Ormoneit [8] by a factor of 2 in the exponent. The proof technique of [8] is completely different, relying on martingale methods and a generalization of Hoeffding's original argument [9].[1]
3) Although Doeblin chains form a very restricted sub-class of all ergodic chains, it is perhaps the most natural class to consider in terms of large deviations properties. To see that, recall that Bryc and Dembo [3] have provided a counter-example of a stationary Doeblin chain for which the regular large deviations principle fails to hold with any rate function. Moreover, if it were possible to obtain meaningful exponential bounds as in the theorem above, with exponents that were independent of the initial condition, this would mean that the ergodic theorem would hold for all bounded functions uniformly in the initial condition, a fact which is known to *imply* that the chain is Doeblin [13].

To proceed with the proof we need to introduce some notation. We identify functions $f : A \to \mathbb{R}$ by the corresponding (infinite-dimensional) column vectors $f = (f_j)$ and probability measures μ by the corresponding row vectors $\mu = (\mu_i)$; similarly, an arbitrary finite signed measure μ on A is simply a row vector with finite L^1 norm, i.e., with $\sum_i |\mu_i| < \infty$. Recall that the transition kernel (P_{ij}) or any other (not necessarily positive) infinite matrix (Q_{ij}) acts on functions $f : A \to \mathbb{R}$ on the right and on signed measures μ on A on the left by:

$$(Qf)_i = \sum_j Q_{ij} f_j \quad \text{and} \quad (\mu Q)_j = \sum_i \mu_i Q_{ij}.$$

The operator norm of Q is defined as

$$|\!|\!|Q|\!|\!| = \sup_i \sum_j |Q_{ij}|,$$

and the convergence parameter d_n defined above can be expressed as $d_n = |\!|\!|P^n - \Pi|\!|\!|$, where the kernel Π corresponding to the stationary distribution π is defined by $\Pi_{ij} = \pi_j$.

The first technical step is to establish a quantitative version of the following fact: If the chain $\{X_n\}$ converges to stationarity quickly and δ is a finite signed measure with total mass $\delta(A) = 0$, then $\|\delta(I - P)\|$ cannot be much smaller than $\|\delta\|$. This can be thought of as an upper bound on the operator norm of the fundamental kernel of the chain.

Lemma 2. If the chain $\{X_n\}$ satisfies the Doeblin condition (5), then for any finite signed measure δ such that $\delta(A) = 0$,

$$\|\delta(I - P)\| \ge \frac{\alpha}{m}\|\delta\|.$$

Proof: We first consider the case $m = 1$. Define the positive operators C and D by $C_{ij} = \alpha\rho_j$ and $D = P - C$. Observe that each row of D sums to $(1 - \alpha)$ so that $|\!|\!|D|\!|\!| = 1 - \alpha$, and also that we have $\delta C = 0$ and $\delta PC = 0$. Therefore, $\delta P^2 = \delta P(C + D) = \delta PD$, and hence

$$\begin{aligned}\|\delta(I - P^k)P^k\| &= \|\delta(I - P^k)D^k\| \\ &\le (1 - \alpha)^k \|\delta(I - P^k)\|. \end{aligned} \tag{6}$$

Now, for any k we have

$$(I - P^{2k}) = (I - P^k) + P^k(I - P^k),$$

so multiplying by δ on the left, taking norms, and using (6),

$$\begin{aligned}\|\delta(I - P^{2k})\| &\le \|\delta(I - P^k)\| + \|\delta P^k(I - P^k)\| \\ &\le (1 + (1 - \alpha)^k)\|\delta(I - P^k)\|.\end{aligned}$$

Applying this inductively we obtain,

$$\|\delta(I - P^{2^n})\| \le \|\delta(I - P)\| \prod_{k=0}^{n-1}(1 + (1 - \alpha)^{2^k})$$

so passing to the limit as $n \to \infty$ yields

$$\|\delta(I - \Pi)\| = \|\delta\| \le \|\delta(I - P)\| \prod_{k=0}^{\infty}(1 + (1 - \alpha)^{2^k}),$$

and this gives the required bound upon observing that that the above infinite product equals $1/\alpha$. To see this, observe that the nth partial product [from $k = 0$ to $k = (n - 1)$] consists of the first 2^{n-1} terms of the geometric series in $(1 - \alpha)$.

Finally, for the case $m > 1$, note that by the previous argument we have $\|\delta(I - P^m)\| \ge \alpha\|\delta\|$. To complete the proof it will suffice to show that for each k,

$$\|\delta(I - P)\| \ge \frac{1}{k}\|\delta(I - P^k)\|. \tag{7}$$

To that end, let $\beta_k = \|\delta(I - P^k)\|$. Since $|\!|\!|P|\!|\!| = 1$, we have

$$\begin{aligned}\beta_k &\ge \|\delta(I - P^k)P\| \\ &= \|\delta(I - P^{k+1}) - \delta(I - P)\| \\ &\ge \|\delta(I - P^{k+1})\| - \|\delta(I - P)\|\end{aligned}$$

i.e., $\beta_k \ge \beta_{k+1} - \beta_1$, implying that $\beta_{k+1} \le (k + 1)\beta_1$. This establishes our claim (7) and completes the proof. □

PROOF OF THEOREM 1. From Proposition 1, it suffices to show that

$$H(W\|W^1 \odot P) \ge \frac{1}{2}\Big[\frac{\alpha\epsilon}{m\overline{F}} - \frac{3}{n-1}\Big]^2, \tag{8}$$

whenever $W = (W_{ij})$ is a bivariate distribution satisfying

$$\|W^1 - W^2\| \le \frac{2}{n-1} \tag{9}$$

$$\text{and} \quad E_{W^1}(F) \ge \epsilon + E_\pi(F) - \frac{b}{n-1}. \tag{10}$$

Write $W_{ij} = W_i^1 Q_{ij}$ for some transition kernel Q. Applying Pinsker's inequality (3),

$$H(W\|W^1 \odot P) \ge \frac{1}{2}\Big[\sum_{ij} W_i^1 |P_{ij} - Q_{ij}|\Big]^2, \tag{11}$$

[1] Note also that result of [8] upon which we improve is actually stated slightly incorrectly there; there should have been an extra factor of 2 in the estimate of the norm of the function g there, which translates to an extra factor of 1/2 in the exponent they finally obtain.

but

$$\begin{aligned} W^1(P-Q) &= W^1P - W^2 \\ &= W^1P - W^1 + \pi - \pi P - (W^2 - W^1) \\ &= \delta(I-P) - (W^2 - W^1), \end{aligned}$$

where $\delta = \pi - W^1$. Taking norms and using Lemma 2,

$$\begin{aligned} \sum_{ij} W^1_i |P_{ij} - Q_{ij}| &\geq \|W^1(P-Q)\| \\ &\geq \frac{\alpha}{m}\|\delta\| - \|W^1 - W^2\|. \qquad (12) \end{aligned}$$

But from the two assumptions (9) and (10) we have the bounds $\|W^1 - W^2\| \leq 2/(n-1)$ and

$$\begin{aligned} \epsilon &\leq E_{W^1}(F) - E_\pi(F) + \frac{b}{n-1} \\ &\leq |E_{W^1}(F) - E_\pi(F)| + \frac{\overline{F}}{n-1} \\ &\leq \|\delta\|\overline{F} + \frac{\overline{F}}{n-1}. \end{aligned}$$

Substituting these in (12) yields

$$\begin{aligned} \sum_{ij} W^1_i |P_{ij} - Q_{ij}| &\geq \frac{\alpha\epsilon}{m\overline{F}}\,\epsilon - \frac{1}{n-1}(2 + \alpha/m) \\ &\geq \frac{\alpha\epsilon}{m\overline{F}}\,\epsilon - \frac{3}{n-1}, \end{aligned}$$

and combining with (11) gives (8) as required. □

IV. Finite-State Chains

Every finite state chain which is ergodic – equivalently, irreducible and aperiodic – satisfies Doeblin's condition. But in many important special cases much more in known about the speed at which the chain converges to equilibrium, i.e., the rate at which $d_n \to 0$; see [2] for a starting point. In those cases it may be possible to get more accurate large deviations bounds by utilizing this knowledge. This is made precise in the following theorem. See [11][10] for earlier related work.

Theorem 2. Suppose $\{X_n\}$ is an ergodic finite-state chain (or, more generally, a Doeblin chain) and let d_n denote its L^1 convergence parameter as before. Then the series $d = \sum_{n\geq 0} d_n$ converges and for any bounded function $F : A \to \mathbb{R}$ and any $\epsilon > 0$ we have,

$$\begin{aligned} &\log \mathsf{P}_x \Big\{ \sum_{i=0}^{n-1} [F(X_i) - \pi(F)] \geq n\epsilon \Big\} \\ &\qquad \leq -\frac{n-1}{2}\Big[\frac{\epsilon}{d\overline{F}} - \frac{3}{n-1}\Big]^2, \end{aligned}$$

as long as $n \geq 1 + 3d\overline{F}/\epsilon$, where $\overline{F} = \sup_x |F(x)|$.

The proof of Theorem 2 is exactly analogous to that of Theorem 1, with Lemma 2 replaced by the following Lemma. Their proofs are omitted.

Lemma 3. Under the assumptions of Theorem 2 for any finite signed measure δ such that $\delta(A) = 0$ we have:

$$\|\delta(I-P)\| \geq \frac{1}{d}\|\delta\|.$$

V. Simulation

Here we very briefly sketch some related results that are motivated by problems in computer simulation. Recall [13] that an irreducible, aperiodic chain $\{X_n\}$ with values in the countable alphabet A is *geometrically ergodic* if there exists a function $V : A \to [1, \infty)$, positive constants δ, b and a finite set $S \subset A$ such that

$$E[V(X_1)|X_0 = x] \leq (1-\delta)V(x) + b\mathbb{I}_S(x). \qquad (13)$$

Although Doeblin chains are a subset of geometrically ergodic chains, in many applications where detailed quantitative results are sought, it is useful to find such a "Lyapunov function" V satisfying (13) even if the chain is Doeblin or finite-valued.

Now suppose we are simulating a geometrically ergodic Markov chain $\{X_n\}$ with stationary distribution π, and we plan to estimate the expected value $E_\pi(F)$ of a given function F which is is dominated by V (in the sense that $|F(x)| \leq CV(x)$ for all $x \in A$, for some constant C). We select a finite set of states $B \subset A$ for which we can find an accurate lower bound on the probability $\pi(B)$, and define

$$U(x) = V(x) - E[V(X_1)|X_0 = x], \quad x \in A;$$

observe that the mean of U under π equals zero.

Our main result here is an explicit exponential upper bound for the following conditional probability, for any positive ϵ, u:

$$\begin{aligned} \mathsf{P}_x \Bigg\{ \sum_{i=0}^{n-1} [F(X_i) - E_\pi(F)] \geq n\epsilon \;\Bigg|\; \Big|\sum_{i=0}^{n-1} U(X_i)\Big| \leq u, \\ X_{n-1} \in B \Bigg\} \end{aligned}$$

The idea is that we sequentially calculate the values of the partial sums of F and of U. If it at some point it turns out that the partial sums of U are absolutely bounded by u and that at that point the chain is in some state in B, then our bound gives a precise qualitative guarantee on the probability that the partial sums of interest are within ϵ of $E_\pi(F)$. If that guarantee is satisfactory, we stop; if not, we continue and repeat the above process.

It is somewhat remarkable that it is possible to extablish a large deviations upper bound at this level of generality, since even the standard asymptotic large deviations principle may well fail for unbounded functions of geometrically ergodic chains, as for example for the simple nearest neighbor random walk on the nonnegative integers with $F(x) = x$.

The proof of this inequality proceeds in two steps. First we obtain an information-theoretic upper bound for the probability of interest in terms of relative entropy; this is done in Proposition 2 below. Then we apply Pinsker's inequality as in the beginning of the proof of Theorem 1 above, and we bound the resulting expression from below by calculating a tight lower bound on its minimum. This is achieved by constructing an appropriate linear program and considering its dual. Similar tools are used to obtain worst-case exponential bounds for constrained problems in the i.i.d. case in [14].

Proposition 2. Let $F_1, F_2, \ldots, F_m$ be an arbitrary finite collection of function from A to $\mathbb{R}$ and B be an arbitrary subset of A. For any initial condition $x \in A$ and nonnegative constants $c_1, c_2, \ldots, c_m$, we have,

$$-\log \mathsf{P}_x\Big\{\sum_{i=0}^{n-1} F_j(X_i) \geq nc_j \;\text{ for all } j, \quad \text{and } X_{n-1} \in B\Big\} \\ \geq (n-1)H(W\|W^1 \odot P),$$

where W is a bivariate distribution on $A \times A$ whose marginals W^1 and W^2 satisfy,

$$\begin{aligned} \|W^1 - W^2\| &\leq \frac{2}{n-1} \\ \text{and } E_{W^1}(F_j) &\geq c_j - \frac{b_j}{n-1} \quad \text{for all } j, \end{aligned}$$

where $b_j := \sup_{x \in B} F_j(x)$.

Proof: Fix an initial state $x \in A$. Let p denote the measure on A^n induced by the distribution of X_0^{n-1} conditioned on $\{X_0 = x\}$, write E for the event of interest, $E = \big\{\sum_{i=0}^{n-1} F_j(X_i) \geq nc_j \text{ for all } j, \text{ and } X_{n-1} \in B\big\}$, and let μ_n denote the conditional measure $\mu_n = p|_E$. From Lemma 1,

$$-\log \mathsf{P}_x\Big\{\cdots\Big\} \geq H(\mu_n\|p). \tag{14}$$

Writing μ_n and p as products of conditional distributions,

$$\begin{aligned} \mu_n(x_0^{n-1}) &= \mu_1(x_0)\mu_2(x_1|x_0)\cdots\mu_n(x_{n-1}|x_0^{n-2}) \\ p(x_0^{n-1}) &= \delta_x(x_0)P(x_1|x_0)\cdots P(x_{n-1}|x_{n-2}), \end{aligned}$$

the relative entropy in (14) can be expanded,

$$H(\mu_n\|p) = \sum_{i=1}^{n-1}\sum_{x_0^{i-1}} H\Big(\mu_{i+1}(\cdot|x_0^{i-1})\|P(\cdot|x_{i-1})\Big)\mu_i(x_0^{i-1}).$$

Letting μ^i denote the one-dimensional marginal of μ_n corresponding to x_{i-1}, and $\mu^{i,i+1}$ denote two-dimensional marginal corresponding to (x_{i-1}, x_i), we can expand $H(\mu_n\|p)$ as

$$\sum_{i=1}^{n-1}\sum_{x_0^{i-1}} H\Big(\mu_{i+1}(\cdot|x_0^{i-1})\|P(\cdot|x_{i-1})\Big)\mu_i(x_0^{i-2}|x_{i-1})\mu^i(x_{i-1})$$

which, using the joint convexity of the relative entropy in its two arguments, is bounded below by

$$\begin{aligned} &\sum_{i=1}^{n-1}\sum_{x_{i-1}} H\Big(\mu_{i+1}(\cdot|x_{i-1})\|P(\cdot|x_{i-1})\Big)\mu^i(x_{i-1}) \\ = &\sum_{i=1}^{n-1} H\Big(\mu^{i,i+1}\|\mu^i \odot P\Big). \end{aligned}$$

Using the joint convexity of H again,

$$\begin{aligned} H(\mu_n\|p) &\geq (n-1)\sum_{i=1}^{n-1}\frac{1}{n-1}H\Big(\mu^{i,i+1}\|\mu^i \odot P\Big) \\ &\geq (n-1)H(W\|W^1 \odot P), \end{aligned}$$

where the bivariate measure W and its first marginal W^1 are,

$$W = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu^{i,i+1} \quad \text{and} \quad W^1 = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu^i.$$

This combined with (14) gives the required bound, and it only remains to verify that W satisfies the stated properties. Indeed, since the second marginal of W is $W^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu^{i+1}$, their difference is $W^1 - W^2 = \frac{\mu^1 - \mu^n}{n-1}$, and since the L^1-norm is bounded by 2 it follows that $\|W^1 - W^2\| \leq 2/(n-1)$.

Finally, by the definition of W^1 and the event E, for any j we have that $E_{W^1}(F_j)$ is given by

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^{n-1} E_{\mu^i}(F_j) &= \mathsf{E}_x\Big[\frac{1}{n-1}\sum_{i=1}^{n-1} F_j(X_{i-1})\Big|E\Big] \\ = \quad &\frac{n}{n-1}\mathsf{E}_x\Big[\frac{1}{n}\sum_{i=0}^{n-1} F_j(X_i)\Big|E\Big] - \mathsf{E}_x\Big[\frac{F_j(X_{n-1})}{n-1}\Big|E\Big] \\ \geq \quad &c - \frac{b_j}{n-1}, \end{aligned}$$

□

Acknowledgments

I.K. was supported in part by a Sloan Foundation Research Fellowship and by NSF grant #00-73378-CCR. S.P.M. was supported in part by the grant NSF ITR #00-85929. We also wish to thank Andrew Barron for a stimulating conversation on the large deviations properties of Markov chains.

References

[1] J.R. Baxter, N.C. Jain, and S.R.S. Varadhan. Some familiar examples for which the large deviation principle does not hold. *Comm. Pure Appl. Math.*, 44(8-9):911–923, 1991.

[2] P. Brémaud. *Markov chains*, volume 31 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1999. Gibbs fields, Monte Carlo simulation, and queues.

[3] W. Bryc and A. Dembo. Large deviations and strong mixing. *Ann. Inst. H. Poincaré Probab. Statist.*, 32(4):549–569, 1996.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. J. Wiley, New York, 1991.

[5] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.

[6] I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, 1998. Information theory: 1948–1998.

[7] I. Csiszár, T.M. Cover, and B.S. Choi. Conditional limit theorems under Markov conditioning. *IEEE Trans. Inform. Theory*, 33(6):788–801, 1987.

[8] P.W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.*, 56(2):143–146, 2002.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[10] C.A. León and F. Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 2004.

[11] P. Lezaud. Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 1998.

[12] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[13] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[14] C. Pandit. *Robust Statistical Modeling Based On Moment Classes With Applications to Admission Control, Large Deviations and Hypothesis Testing*. PhD thesis, Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2004.

[15] S.R.S. Varadhan. *Large Deviations and Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1984.

[16] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*. Springer-Verlag, Berlin, 1995.