On risk rates and large deviations in finite Markov chain experiments

Peter Scheffel and Heinrich v. Weizsäcker

June 25, 1997

Abstract

The observation of an ergodic Markov chain asymptotically allows perfect identification of the transition matrix. In this paper we determine the rate of the information contained in the first n observations, provided the unknown transition matrix belongs to a known finite set. As an essential tool we prove new refinements of the large deviation theory of the empirical pair measure of finite Markov chains.
Keywords: Markov Chain, Entropy, Bayes risk, Large Deviations.

1 Introduction

The observation of an ergodic Markov chain asymptotically allows perfect identification of the transition matrix. In this paper we determine the rate of the information contained in the first n observations, provided the unknown transition matrix belongs to a known finite set.
In the case of iid. sequences and two parameters this question was in a certain sense answered by Chernoff [Che52]. Later Torgersen (e.g. [Tor81]) extended Chernoff's Theorem to a finite parameter set using the frame of abstract decision theory. In [KW93] it was shown that in finite parameter problems generally the decision theoretic rate and the Shannon theoretic rate coincide, where the former is measured by the minimal Bayes risk or equivalently by the deficiency distance to the most informative experiment and the latter is the rate of the entropy risk under the optimal ('reference') prior. This concept has been studied in Bayesian analysis starting with [Lin56] and [Ber79].
Our main results in this paper are stated in sections 3, 4 and 6. In analogy to Chernoff we show (Theorem 1) for the simple alternative of two irreducible transition matrices π_0, π_1 with the same zeroes that the rate of the risk R_n after n observations is given by

$$\lim_{n\to\infty} \sqrt[n]{R_n} = \inf_{0<t<1} \rho(\pi_t)$$

where ρ denotes the spectral radius and π_t is the nonnegative matrix with entries

$$\pi_t(i,j) = \pi_0(i,j)^t \, \pi_1(i,j)^{1-t}.$$

The proof in section 5 is based on a minimax argument and on the large deviation theorem for the empirical pair measure. Theorem 1 is a special case of Theorem 2 in section 4 which gives the asymptotic rate for general π_0 and π_1. In order to estimate the probability of paths which are possible for both Markov chains in this case we have to prove some refinements of the known

large deviation theorem for the empirical pair measure. These large deviation results are given in section 6. They are of independent interest. The main point there is to get lower bounds for sets which have empty interior in the usual topology of the stationary measures.
The formulation of Theorem 2 and of the crucial large deviation result Theorem 4 are based on the concept of λ-strings (introduced in section 4) which is related to the classification of states for general nonnegative but not necessarily stochastic matrices.

2 Notations

Let S be a finite set which will be fixed throughout this paper. $\mathbb{N}$ is the set of integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.We denote by $\mathcal{M}$ the set of all probability distributions on S^2, by $\mathcal{M}_s$ the set of all stationary distributions on S^2, i.e. those probability distributions $Q = (Q_{ij})_{i,j\in S}$ on S^2 whose two marginal distributions coincide. For $Q = (Q_{ij})_{i,j\in S} \in \mathcal{M}_s$ the marginal (i.e. the vector of row -or equivalently column- sums) is denoted by $(Q_i)_{i\in S}$. If $Q \in \mathcal{M}_s$ and $\pi = (\pi(i,j))_{i,j\in S}$ is a transition matrix the symbol $Q \times \pi$ denotes the element of $\mathcal{M}$ defined by $(Q \times \pi)_{ij} = Q_i \pi(i,j)$. For a nonnegative $S \times S-$matrix λ let $\mathcal{M}^\lambda := \{Q \in \mathcal{M} \,|\, Q \ll \lambda\}$.
We call an $S \times S$-matrix Q irreducible if there exists a set $I \subset S$ with $Q_{ij} = 0$ whenever $i \notin I$ or $j \notin I$, such that for all $i, j \in I$ there exists an $n \in \mathbb{N}$ with $Q^n_{ij} > 0$. By $\mathcal{M}_{s,i}$ we denote the set of all stationary irreducible probability distributions on S^2. According to this definition even the matrix

$$Q = \begin{pmatrix} * & * & 0 \\ * & * & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is irreducible. For transition matrices our definition of irreducibility is equivalent to the usual one since in this case there are no zero rows.
To every $Q \in \mathcal{M}_s$ there corresponds a unique law of a stationary Markov chain with state space S whose two dimensional marginal is Q. (Note that the law is unique whereas the rows of the transition matrix corresponding to states with $Q_i = 0$ are arbitrary.)
The 'entropy' of this stationary Markov chain is given by

$$\begin{aligned} H_s(Q) &= -\sum_{i,j\in S} Q_{ij} \log Q_{ij} + \sum_{i\in S} Q_i \log Q_i \\ &= -\sum_{i,j\in S} Q_{ij} \log \frac{Q_{ij}}{Q_i}. \end{aligned}$$

For $Q \in \mathcal{M}_s$ and a transition matrix π we also consider the relative entropy (Kullback-Leibler distance) of Q to $Q \times \pi$

$$\begin{aligned} H(Q|Q \times \pi) &= \sum_{i,j\in S} Q_{ij} \log \frac{Q_{ij}}{Q_i \pi_{ij}} \\ &= -H_s(Q) - \sum_{i,j\in S} Q_{ij} \log \pi_{ij} \end{aligned}$$

Remark 1:

i) By the representation of $H(Q|Q\times\pi)$ it is easily seen that for two matrices π_0 and π_1 the difference $H(Q|Q\times\pi_0) - H(Q|Q\times\pi_1)$ is a linear function in Q.

ii) Let $Q_1, Q_2 \in \mathcal{M}_s$ be concentrated on $I_1\times I_1$ resp. $I_2\times I_2$ with $I_1\cap I_2 = \emptyset$ and let $Q = \alpha_1 Q_1 + \alpha_2 Q_2$. Then

$$H(Q|Q\times\pi) = \alpha_1 H(Q_1|Q_1\times\pi) + \alpha_2 H(Q_2|Q_2\times\pi).$$

For $\omega\in\Omega = S^{\mathbb{N}_0}$ define $\hat{m}_n(\omega)\in\mathcal{M}$ by $\hat{m}_n(\omega)_{ij} = \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{\omega_k=i,\ \omega_{k+1}=j\}}(\omega)$ for $i,j\in S$. Note that $\hat{m}_n(\omega)\in\mathcal{M}_s$ if and only if $\omega_0=\omega_n$. The vector $\hat{m}_n(\omega)$ is the empirical distribution of the one step transitions of the Markov chain, it is called the empirical pair measure ([Ell85] (1.22)).

Let $\Theta = \{1,\dots,l\}$ be a finite parameter set. For $\theta\in\Theta$ let P_θ be a probability measure on $\Omega = S^{\mathbb{N}_0}$. Let $(X_n)_{n\in\mathbb{N}_0}$ be the coordinate process. For $\omega = (\omega_0,\dots,\omega_n)\in S^{n+1}$ we write $P_\theta^n(\omega) = P_\theta(X_0=\omega_0,\dots,X_n=\omega_n)$. If $\alpha=(\alpha_1,\dots,\alpha_l)$ is a prior distribution then P_α^n denotes the mixed distribution $\sum_{\theta\in\Theta}\alpha_\theta P_\theta^n$. We introduce the posterior distribution $\hat{\alpha}^n(\omega)$ by the equation

$$\hat{\alpha}^n(\omega) = \left(\frac{\alpha_\theta\, P_\theta^n(\omega)}{P_\alpha^n(\omega)}\right)_{\theta\in\Theta}.$$

Definition 1: *Let $(X_i)_{i\in\mathbb{N}}$ and P_θ, $\theta\in\Theta$ be as before. The entropy risk of α given $(X_0,\dots,X_n)$ is defined by*

$$H_n(\alpha|X) := \sum_{\omega\in S^{n+1}} P_\alpha^n(\omega)\left(-\sum_{\theta\in\Theta}\hat{\alpha}_\theta^n(\omega)\log\hat{\alpha}_\theta^n(\omega)\right).$$

Similarly

$$B_n(\alpha|X) := \sum_{\omega\in S^{n+1}} P_\alpha^n(\omega)\left(1-\max_{\theta=1,\dots,l}\hat{\alpha}_\theta^n(\omega)\right)$$

is called minimal Bayes risk given $X_0,\dots,X_n$.

Remark 2:

i) In the case $\Theta=\{0,1\}$ of two parameters one has $2B_n\left((\frac{1}{2},\frac{1}{2})|X\right) = 2 - ||P_0^n - P_1^n||$ where $||\cdot||$ is the total variation.

ii) Observe that both $B_n(\alpha|X)$ and $H_n(\alpha|X)$ can be written as $\mathbb{E}(f(\hat{\alpha}^n))$ where f is a concave function and $(\hat{\alpha}^n)_{n\in\mathbb{N}}$ is a martingale, so that $(f(\hat{\alpha}^n))_{n\in\mathbb{N}}$ is a super-martingale. This implies that $B_n(\alpha|X)$ and $H_n(\alpha|X)$ are decreasing functions in n.

In order to be able to allow zero entries in the transition matrices it is convenient to equip $\mathcal{M}$ with two topologies, the topology τ of coordinatewise convergence and the following topology

τ_0 which is finer than τ : A set $U \subset \mathcal{M}$ is τ_0-open iff for every $Q \in U$ there is a $\delta > 0$ such that the set

$$U_\delta(Q) := \{P \in \mathcal{M} : P \ll Q \text{ and } |P_{ij} - Q_{ij}| < \delta, \text{ for all } i,j\}$$

is contained in U. Thus every point $Q \in \mathcal{M}$ has a τ_0-neighbourhood of points which have the same zeroes as Q.
The $\tau-$neighbourhood will be denoted by the letter V, i.e. $V_\delta(Q) := \{P \in \mathcal{M} : |P_{i,j} - Q_{i,j}| < \delta \text{ for all } i,j\}$. Further we set $V_\delta^\lambda(Q) := V_\delta(Q) \cap \mathcal{M}^\lambda$. So $U_\delta(Q) = V_\delta^Q(Q)$.

3 The risk rates

The following result is the starting point of this paper. It is a special case of the main result of [KW93].

Proposition 1: *Let $\alpha = (\alpha_1, \ldots, \alpha_l)$ be any strictly positive prior. Suppose that for every $\theta \in \Theta = \{1, \ldots, l\}$ the process $(X_i)_{i \in \mathbb{N}_0}$ forms under P_θ a Markov chain with the irreducible transition matrix π_θ where $\theta \mapsto \pi_\theta$ is injective. Then*

a) $\lim_{n\to\infty} B_n(\alpha|X) = \lim_{n\to\infty} H_n(\alpha|X) = 0$

b) $\lim_{n\to\infty} \frac{1}{n} \log H_n(\alpha|X) = \lim_{n\to\infty} \frac{1}{n} \log B_n(\alpha|X) = \lim_{n\to\infty} \frac{1}{n} \log B_n(\alpha'|X)$
for every other positive prior α', provided one of these limits exist.

c) Let $r(P_1, \ldots, P_l)$ denote the limit in b). Then $r(P_1, \ldots, P_l)$ is determined by the two parameter subexperiments as follows

$$r(P_1, \ldots, P_l) = max_{1 \le \vartheta \ne \eta \le l} r(P_\vartheta, P_\eta)$$

Proof: a) This follows e.g. from Proposition 4.1 in [KW93] and the ergodic theorem.
b) follows from Theorem 5.4 and Cor. 5.6 in [KW93].
c) follows also from Cor. 5.6 in [KW93]. *q.e.d.*

Remark 3: Proposition 1 a) shows that the rate $r(P_0, P_1)$ of a binary experiment may be computed either by the help of the minimal Bayes risk, or the entropy risk. Moreover by b) it is independent of the prior and according to c) the following Theorem has a straightforward extension to finitely many parameters.

Here is our main result in a special case.

Theorem 1: *Let (P_0, P_1) be a binary experiment of irreducible Markov chains with two different equivalent transition matrices π_0, π_1 and two initial distributions μ_0, μ_1 which are not mutually singular. For simplicity we write $K_\theta(Q) = H(Q|Q \times \pi_\theta)$ for $\theta = 0, 1$. Then for any strictly positive prior α the rate $r(P_0, P_1)$ is given by*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log B_n(\alpha|X) = \lim_{n\to\infty} \frac{1}{n} \log H_n(\alpha|X) &= -\inf_{Q \in \mathcal{M}_s} K_0(Q) \vee K_1(Q) \\ &= \inf_{0<t<1} \log \rho(\pi_t), \end{aligned}$$

where ρ denotes the spectral radius and $\pi_t = (\pi_0(i,j)^t \pi_1(i,j)^{1-t})_{i,j \in S}$.

Remark 4: In the special case of iid. observations the corresponding transition matrices π_0, π_1 have identical rows $\pi_k(i,j) = p_k(j)$. Then the matrix $\pi_t(i,j) = p_0(j)^{1-t} p_1(j)^t$ has the largest eigenvalue

$$\sum_{j \in S} p_0(j)^{1-t} p_1(j)^t$$

the infimum of which in t yields precisely Chernoff's rate ([Che52]).

4 Extensions

Theorem 1 will be shown by using large deviation results for the pair empirical measure which can be found for example in [DZ93]. Using the somewhat refined large deviation methods of section 6 it is possible to apply Theorem 1 to binary experiments of irreducible Markov chains whose transition matrices need no longer be equivalent but where π_t has to be irreducible for $0 < t < 1$.
Now what happens if π_t is no more irreducible? Can we still apply Theorem 1? Let us have a look at the following example.

Example 1: *Let $\varepsilon \geq 0$, $\frac{1}{2} \leq \delta < 1$ and consider the binary experiment induced by the pair*

$$\pi_0(\varepsilon) = \begin{pmatrix} 0 & 1-\delta & \delta \\ 1-\delta & \delta - \varepsilon & \varepsilon \\ \varepsilon & 1-\delta-\varepsilon & \delta \end{pmatrix}, \qquad \pi_1 = \begin{pmatrix} 1-\delta & \delta & 0 \\ \delta & 0 & 1-\delta \\ \delta & 0 & 1-\delta \end{pmatrix}$$

with arbitrary positive initial distribution. Then

$$\pi_t(\varepsilon) = \begin{pmatrix} 0 & (1-\delta)^t \delta^{(1-t)} & 0 \\ (1-\delta)^t \delta^{(1-t)} & 0 & \varepsilon^t (1-\delta)^{(1-t)} \\ \varepsilon^t \delta^{(1-t)} & 0 & \delta^t (1-\delta)^{(1-t)} \end{pmatrix}.$$

For $\varepsilon > 0$ the matrix $\pi_t(\varepsilon)$ is irreducible. So we can apply Theorem 1 according to our remark and obtain as risk rate $\inf_{0<t<1} \log \rho(\pi_t(\varepsilon))$. Now

$$\lim_{\varepsilon \to 0} \inf_{0<t<1} \log \rho(\pi_t(\varepsilon)) \leq \lim_{\varepsilon \to 0} \log \rho(\pi_{\frac{1}{2}}(\varepsilon)) = \log \rho(\pi_{\frac{1}{2}}(0))$$

which follows because the spectral radius depends continuously on the matrix, and on the other hand

$$\lim_{\varepsilon \to 0} \inf_{0<t<1} \log \rho(\pi_t(\varepsilon)) \geq \lim_{\varepsilon \to 0} \inf_{0<t<1} \log \rho(\pi_t(0)) = \log \rho(\pi_{\frac{1}{2}}(0))$$

since $\pi_t(\varepsilon) \geq \pi_t(0) \geq 0$ implies with the help of the Theorem of Perron-Frobenius (see e.g. [Sen81] Theorem 1.5) that $\rho(\pi_t(\varepsilon)) \geq \rho(\pi_t(0))$. We get

$$\lim_{\varepsilon \to 0} \inf_{0<t<1} \log \rho(\pi_t(\varepsilon)) = \frac{1}{2} \log \delta + \frac{1}{2} \log(1-\delta).$$

But either a direct computation or Theorem 2 shows that the rate of the Bayes risk according to the matrices $\pi_0(0)$ and π_1 equals $\log(1-\delta)$. This implies in particular that Theorem 1 does not hold for $\varepsilon = 0$, i.e. when π_t is no longer irreducible. It also shows a somewhat surprising discontinuity of the rate.

As the example shows we have to find a more appropriate rate function in general. In the proof of Theorem 2 we will see that the Bayes risk $B_n(\alpha|X)$ gets a nonnegative contribution only from that subset of Ω on which $\hat{m}_n(\omega) \in \mathcal{M}^\pi$, for $\pi = \pi_0 \wedge \pi_1$, i.e. from those paths whose transitions are possible for both π_0 and π_1. Since π is in general no longer irreducible these paths ω have to spend a certain time in the irreducible blocks of π. To describe this process we introduce the concept of strings.
Let λ be a nonnegative $S \times S$-matrix. We write $i \underset{\lambda}{\rightsquigarrow} j$ (and say i leads to j) iff $\lambda^n(i,j) > 0$ for some $n \in \mathbb{N}$ and $i \underset{\lambda}{\sim} j$ (i communicates with j) iff $i = j$ or ($i \underset{\lambda}{\rightsquigarrow} j$ and $j \underset{\lambda}{\rightsquigarrow} i$). An equivalence class E of the relation $\underset{\lambda}{\sim}$ is called nontrivial if it contains at least two elements or $E = \{i\}$ with $\lambda(i,i) > 0$.

Definition 2: *We call a set $I \subset S$ a* **λ−string** *if the following holds: there exist a tuple $(E_1, \ldots, E_k)$ of nontrivial λ-equivalence classes with $I = E_1 \cup \ldots \cup E_k$, such that for every $l \in \{1, \ldots, k-1\}$ there are points $i_l \in E_l$, $j_{l+1} \in E_{l+1}$ with $i_l \underset{\lambda}{\rightsquigarrow} j_{l+1}$.*

Since no state $i \in E_{l+1}$ leads to a state $j \in E_l$ the order of $(E_1, \ldots, E_k)$ is uniquely determined. So a string is the union of a sequence of $\underset{\lambda}{\sim}$ -equivalence classes which can be visited successively by a single path. We want to make clear our definition by the following example.

Example 2:

i) Consider the following 7×7-matrix ($*$ denotes a positive entry)

$$\lambda = \begin{pmatrix} 0 & 0 & 0 & 0 & * & 0 & 0 \\ 0 & * & * & 0 & 0 & 0 & 0 \\ 0 & * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & 0 & * \\ 0 & 0 & 0 & * & * & 0 & 0 \\ * & 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The nontrivial $\underset{\lambda}{\sim}$ -equivalence classes are $E_1 = \{2,3\}, E_2 = \{4,5\}, E_3 = \{6\}$ ($\{1\}$ and $\{7\}$ would be a trivial classes). As λ-strings we get $E_1 \cup E_2 = \{2,3,4,5\}$ because $\lambda(3,4) > 0$, $E_3 \cup E_2 = \{6,4,5\}$ since $\lambda^2(6,5) = \lambda(6,1)\lambda(1,5) > 0$ and of course E_1, E_2 and E_3.

ii) Next we give an example of a matrix λ without any strings. Let

$$\lambda = \begin{pmatrix} 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & 0 & 0 & 0 \end{pmatrix}.$$

No string can exist since all paths end in the state 2 and cannot leave the state 2 anymore.

For $x \in S$ let $\mathcal{S}^\lambda(x) := \{I \subset S | I = E_1 \cup \ldots \cup E_k,$ is a $\lambda-$string, with $x \underset{\lambda}{\rightsquigarrow} y$ for some $y \in E_1\}$. For a measure μ on S let $\mathcal{S}^\lambda(\mu) = \{I | I \in \mathcal{S}^\lambda(x), \mu(x) > 0\}$. For $I \subset S$ we define further $\mathcal{M}_s^\lambda(I)$ as the set $\{Q \in \mathcal{M}_s \cap \mathcal{M}^\lambda | Q$ is concentrated on $I \times I\}$, $\mathcal{MS}^\lambda(x) := \cup\{\mathcal{M}_s^\lambda(I) | I \in \mathcal{S}^\lambda(x)\}$ and $\mathcal{MS}^\lambda(\mu) := \cup\{\mathcal{MS}^\lambda(x) | \mu(x) > 0\}$.

Remark 5: $\mathcal{MS}^\lambda(\mu)$ is the set of those stationary probability measures $Q \in \mathcal{M}$ which can be dominated by the empirical pair measure of a path of a Markov chain whose transitions are possible according to λ and which starts in a point of positive μ-measure. Of course $\mathcal{MS}^\lambda(\mu) = \emptyset$ if there are no λ-strings.
I is an element of $\mathcal{S}^\lambda(\mu)$ if there exist a path starting in x, $\mu(x) > 0$ whose transitions are allowed by λ and which reaches every $\underset{\lambda}{\sim}$ -equivalence class E_k in I.

For a λ-string $I = E_1 \cup \ldots \cup E_k$ we set $\lambda^I := \sum_{l=1}^k \lambda^{E_l}$ where λ^{E_l} is the $S \times S$ matrix which equals λ on $E_l \times E_l$ and is zero otherwise.

Now we are able to formulate our main result. The functions K_0 and K_1 are defined in Theorem 1.

Theorem 2: *Let (P_0, P_1) be a binary experiment of irreducible Markov chains with different transition matrices π_0, π_1 and initial distributions μ_0, μ_1. Let*

$$\pi := \pi_0 \wedge \pi_1 = (\min(\pi_0(i,j), \pi_1(i,j)))_{i,j \in S}$$

and

$$\mu := \mu_0 \wedge \mu_1 = (\min(\mu_0(i), \mu_1(i)))_{i \in S}.$$

Then the rate of the entropy risk is given by

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log H_n(\alpha|X) &= \lim_{n\to\infty} \frac{1}{n} \log B_n(\alpha|X) && (1) \\ &= -\inf_{Q \in \mathcal{MS}^\pi(\mu)} K_0(Q) \vee K_1(Q) && (2) \\ &= \max_{I \in \mathcal{S}^\pi(\mu)} \inf_{0<t<1} \log \rho(\pi_t^I) && (3) \end{aligned}$$

where $\pi_t(i,j) = \pi_0(i,j)^t \pi_1(i,j)^{1-t}$ and ρ denotes the spectral radius.

Example 3: *Let us consider the two matrices*

$$\pi_0 = \begin{pmatrix} * & * & 0 & * \\ 0 & * & * & 0 \\ * & * & 0 & * \\ * & 0 & * & 0 \end{pmatrix}, \qquad \pi_1 = \begin{pmatrix} 0 & * & * & 0 \\ * & 0 & 0 & * \\ * & * & * & 0 \\ * & * & 0 & * \end{pmatrix}.$$

Then $\pi = \pi_0 \wedge \pi_1$ is equivalent to the matrix λ in example 2 ii). Thus no π-string exists and therefore $\mathcal{MS}^\pi(\mu) = \mathcal{S}^\pi(\mu) = \emptyset$. This implies that both (2) and (3) are $-\infty$.
The Bayes risk measures how orthogonal the laws P_0^n and P_1^n are. Now since there is no path of length $n > 4$ which is possible for P_0^n and P_1^n the Bayes risk $B_n(\alpha|X)$ equals 0 for $n > 4$ and (1) is also $-\infty$.

Remark 6: It is easily seen that Theorem 1 is a special case of Theorem 2: In the setting of Theorem 1, π is concentrated on $S \times S$ and irreducible. So S itself is a π-equivalence class and for any μ one has $\mathcal{MS}^{\pi}(\mu) = \mathcal{M}_s^{\pi}(S) = \mathcal{M}_s^{\pi}$. Since $K_0(Q) = \infty = K_1(Q)$ on $\mathcal{M}_s \backslash \mathcal{M}_s^{\pi}$ we have $-\inf_{Q \in \mathcal{MS}^{\pi}(\mu)} K_0(Q) \vee K_1(Q) = -\inf_{Q \in \mathcal{M}_s} K_0(Q) \vee K_1(Q)$.

Example 4: *Consider example 1 for $\varepsilon = 0$. We get*

$$\pi_t(0) = \begin{pmatrix} 0 & (1-\delta)^t \, \delta^{(1-t)} & 0 \\ (1-\delta)^t \, \delta^{(1-t)} & 0 & 0 \\ 0 & 0 & \delta^t \, (1-\delta)^{(1-t)} \end{pmatrix}.$$

So we have two $\pi_t(0)$-strings: $E_1 = \{1,2\}$ and $E_2 = \{3\}$ and the rate according to Theorem 2 equals

$$max_{I \in \{E_1, E_2\}} \inf_{0<t<1} \log \rho(\pi_t^I) = log(1-\delta).$$

5 Proofs of Theorems 1 and 2

Theorem 1 is a special case of Theorem 2. So we need to prove only Theorem 2. However we indicate those points in the proof at which a more direct argument is sufficient for Theorem 1 in order to make clear the difficulties which arise by allowing zero entries in the transition matrices. These direct arguments are given in brackets [ ]. In particular the known large deviation result stated as Theorem 6 in section 6 is sufficient for the proof of Theorem 1.

I. We first prove the first equality.

1. As noted above we may consider the minimal Bayes risk instead of the entropy risk. Also the rate does not depend on the particular prior. So we may assume $\alpha = (\frac{1}{2}, \frac{1}{2})$. By definition of the minimal Bayes risk and by the explicit form of the Markov chain probabilities one has

$$\begin{aligned} 2 \; B(\alpha | X_0, \ldots, X_n) &= 2 \sum_{\omega \in S^{n+1}} \left(\min_{\theta \in \{0,1\}} \hat{\alpha}_{\theta}^n(\omega) \right) P_{\alpha}^n(\omega) \\ &= 2 \sum_{\omega \in S^{n+1}} \min_{\theta \in \{0,1\}} \frac{1}{2} P_{\theta}^n(\omega) \\ &= P_0^n \Big(\omega : P_0^n(\omega) \leq P_1^n(\omega) \Big) + P_1^n \Big(\omega : P_1^n(\omega) < P_0^n(\omega) \Big). \end{aligned}$$

Let $\mathcal{B}_0^n$ and $\mathcal{B}_1^n$ be the two terms of the last expression. Then

$$\begin{aligned} \mathcal{B}_0^n &= P_0^n \Big(\omega : \frac{1}{n} \log P_0^n(\omega) \leq \frac{1}{n} \log P_1^n(\omega) \Big) \\ &= P_0^n \Big(\omega : \sum_{i,j \in S} \hat{m}_n(\omega)_{ij} \log \frac{\pi_0(i,j)}{\pi_1(i,j)} \leq \frac{1}{n} \log \frac{\mu_1(\omega_0)}{\mu_0(\omega_0)} \Big) \\ &= P_0^n \Big(\omega : K_1(\hat{m}_n(\omega)) - K_0(\hat{m}_n(\omega)) \leq \frac{1}{n} \log \frac{\mu_1(\omega_0)}{\mu_0(\omega_0)} \Big) \\ &= \sum_{x \in S_{\mu}} \mu_0(x) P_{0,x}^n \Big(\omega : K_1(\hat{m}_n(\omega)) - K_0(\hat{m}_n(\omega)) \leq \frac{1}{n} \log \frac{\mu_1(x)}{\mu_0(x)} \Big), \end{aligned}$$

where $S_\mu := \{s \in S \,|\, \mu(s) > 0\}$ and $P_{0,x}$ is P_0 conditioned on start in x, and similarly

$$\begin{aligned}\mathcal{B}_1^n &= P_1^n\Big(\omega : K_0(\hat{m}_n(\omega)) - K_1(\hat{m}_n(\omega)) \le \frac{1}{n}\log\frac{\mu_0(\omega_0)}{\mu_1(\omega_0)}\Big)\\ &= \sum_{x\in S_\mu}\mu_1(x)P_{1,x}^n\Big(\omega : K_0(\hat{m}_n(\omega)) - K_1(\hat{m}_n(\omega)) \le \frac{1}{n}\log\frac{\mu_0(x)}{\mu_1(x)}\Big).\end{aligned}$$

With these notations we have

$$\begin{aligned}\max\{\liminf\frac{1}{n}\log\mathcal{B}_0^n, \liminf\frac{1}{n}\log\mathcal{B}_1^n\} &\le \liminf\frac{1}{n}\log B_n(\alpha|X)\\ &\le \limsup\frac{1}{n}\log B_n(\alpha|X)\\ &\le \max\{\limsup\frac{1}{n}\log\mathcal{B}_0^n, \limsup\frac{1}{n}\log\mathcal{B}_1^n\}.\end{aligned}$$

We show that the first and the last term of this inequality (and therefore the Bayes risk) are equal to the righthand side of the first equality of the Theorem.

2. Let us first look at the lower bound which is more involved. For $\varepsilon > 0$ consider the relatively open subset $M_{-\varepsilon} = \{Q \in \mathcal{M}^\pi : K_1(Q) - K_0(Q) < -\varepsilon\}$ of $\mathcal{M}^\pi$. Theorem 4 yields

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mathcal{B}_0^n &\ge \sup_{\varepsilon>0}\liminf_{n\to\infty}\frac{1}{n}\log\sum_{x\in S_\mu}\mu(x)P_{0,x}^n(\hat{m}_n \in M_{-\varepsilon})\\ &\ge \sup_{\varepsilon>0} -\inf_{Q\in M_{-\varepsilon}\cap\mathcal{MS}^\pi(\mu)} K_0(Q)\\ &= -\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_1(Q)<K_0(Q)}} K_0(Q).\end{aligned}$$

For the last step note that the functions K_θ are continuous even for the coordinatewise topology on the set $\mathcal{M}^\pi$. By symmetry we obtain

$$\begin{aligned}&\liminf\frac{1}{n}\log B_n(\lambda|X)\\ \ge\ &\max\left(-\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_1(Q)<K_0(Q)}} K_0(Q),\ -\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_0(Q)<K_1(Q)}} K_1(Q)\right)\\ =\ &-\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_0(Q)\neq K_1(Q)}} K_0(Q)\vee K_1(Q)\\ =\ &\max_{I\in\mathcal{S}^\pi(\mu)} -\inf_{\substack{Q\in\mathcal{M}_s^\pi(I)\\ K_0(Q)\neq K_1(Q)}} K_0(Q)\vee K_1(Q). \qquad (4)\end{aligned}$$

⌈ For the proof of Theorem 1 one gets by the help of Theorem 6

$$\liminf\frac{1}{n}\log B_n(\lambda|X) \ge -\inf_{\substack{Q\in\mathcal{M}_s^\pi\\ K_0(Q)\neq K_1(Q)}} K_0(Q)\vee K_1(Q).$$

The condition $K_1(Q) \neq K_0(Q)$ can be omitted since for example P_0 is an element of $\mathcal{M}_s^\pi$ with $0 = K_0(P_0) \neq K_1(P_0)$. Thus by the convexity of $\mathcal{M}_s^\pi$ and the linearity of $K_0(Q) - K_1(Q)$ it is

possible to approach every element $Q \in \mathcal{M}_s^\pi$ by an element $Q_t = (1-t)Q + tP_0$ that fulfills the condition $K_1(Q_t) \neq K_0(Q_t)$. ⌉

Similarly we have to remove the condition $K_0(Q) \neq K_1(Q)$ in (4). For this we distinguish two kinds of strings $I \in \mathcal{S}^\pi(\mu)$. In the first case the set $\{Q \in \mathcal{M}_s^\pi(I) : K_0(Q) \neq K_1(Q)\}$ is nonempty. Then it is dense in the convex set $\mathcal{M}_s^\pi(I)$, since the difference $K_0(Q) - K_1(Q)$ is linear in Q. Therefore the condition can be removed by continuity.

In the second case

$$K_0(Q) = K_1(Q) \text{ for all } Q \in \mathcal{M}_s^\pi(I). \tag{5}$$

This holds e.g. in example 4 for $\delta = \frac{1}{2}$. In this case the estimate (4) is useless and we apply the large deviation Theorem 5 instead of Theorem 4. If $I = E_1 \cup \ldots \cup E_k$ Remark 1 ii) implies

$$-\inf_{Q \in \mathcal{M}_s^\pi(I)} K_0(Q) \vee K_1(Q) = -\inf_{Q \in \mathcal{M}_s^\pi(I)} K_0(Q) = \max_{k=1,\ldots,l} -\inf_{Q \in \mathcal{M}_s^\pi(E_l)} K_0(Q).$$

Therefore it is sufficient to prove for every class $E_l \subset S$ the inequality

$$\liminf_{n\to\infty} \frac{1}{n} \log B_n(\lambda|X) \geq -\inf_{Q \in \mathcal{M}_s^\pi(E_l)} K_0(Q).$$

Choose $\omega_0 \in S_\mu$ with $\pi^m(E_1|\omega_0) > 0$ for at least one $m \in \mathbb{N}$. Let $x \in E_l$ and $\tilde{\omega} = (\omega_0, \ldots, \omega_r = x)$ be a path of length $r+1$ with $P_0(\tilde{\omega}) \geq p > 0$ and $P_1(\tilde{\omega}) \geq p$. Without loss of generality we assume $r(K_1(\hat{m}_r(\tilde{\omega})) - K_0(\hat{m}_r(\tilde{\omega}))) \leq \log(\mu_1(\omega_0)/\mu_0(\omega_0))$. If the reverse inequality holds interchange the parameters $\theta = 0$ and $\theta = 1$ in the following. Then we have for all n by the linearity of $K_1 - K_0$

$$\begin{aligned}
\mathcal{B}_0^n &\geq \mu_0(\omega_0) P_{0,\omega_0}\left(K_1(\hat{m}_n) - K_0(\hat{m}_n) \leq \frac{1}{n}\log\frac{\mu_1(\omega_0)}{\mu_0(\omega_0)}\right) \\
&= \mu_0(\omega_0) P_{0,\omega_0}\left(\frac{r}{n}\Big(K_1 - K_0\Big)(\hat{m}_r(\tilde{\omega})) + \frac{n-r}{n}\Big(K_1 - K_0\Big)(\hat{m}_{n-r}(\omega_r, \ldots, \omega_n))\right. \\
&\qquad\qquad \left. \leq \frac{1}{n}\log\frac{\mu_1(\omega_0)}{\mu_0(\omega_0)}\right) \\
&\geq p\mu_0(\omega_0) P_{0,x}\Big(K_1(\hat{m}_{n-r}) - K_0(\hat{m}_{n-r}) = 0\Big) \\
&\geq p\mu_0(\omega_0) P_{0,x}\Big(\hat{m}_{n-r} \in \mathcal{M}_s^\pi(E_l)\Big) \qquad \text{with the help of (5)} \\
&\geq p\mu_0(\omega_0) P_{0,x}\Big(\hat{m}_{n-r} \in \mathcal{M}_s^\pi(E_l) \cap \mathcal{M}_{s,i}\Big).
\end{aligned}$$

Note that the risk $B_n(\lambda|X)$ and hence its logarithm is decreasing in n as a consequence of Jensen's inequality (see Remark 2). So we get for every d

$$\liminf_{n\to\infty} \frac{1}{n} \log B_n(\lambda|X) = \liminf_{n\to\infty} \frac{1}{nd+r} \log B_{nd+r}(\lambda|X). \tag{6}$$

The set $\mathcal{M}_s^\pi(E_l)\cap\mathcal{M}_{s,i}$ is relatively τ_0-open in $\mathcal{M}_{s,i}$. Thus by Theorem 5 there is a d such that

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{nd+r}\log B_{nd+r}(\lambda|X) &\geq \liminf_{n\to\infty}\frac{1}{nd}\log P_{0,x}\Big(\hat{m}_{nd}\in\mathcal{M}_s^\pi(E_l)\cap\mathcal{M}_{s,i}\Big)\\ &\geq -\inf_{\substack{Q\in\mathcal{M}_s^\pi(E_l)\cap\mathcal{M}_{s,i}\\ Q_x>0}}K_0(Q).\end{aligned}$$

Since for all $x\in E_l$ there is an r for which this holds we get from (6)

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log B_n(\lambda|X) &\geq -\inf_{Q\in\mathcal{M}_s^\pi(E_l)\cap\mathcal{M}_{s,i}}K_0(Q)\\ &= -\inf_{Q\in\mathcal{M}_s^\pi(E_l)}K_0(Q)\vee K_1(Q)\end{aligned}$$

for all $l=1,\ldots,k$. The last equality follows from the fact that $\mathcal{M}_s^\pi(E_l)$ is the closure of $\mathcal{M}_s^\pi(E_l)\cap\mathcal{M}_{s,i}$ in the topology of coordinatewise convergence and the assumption (5). This proves the lower estimate.

3. For the upper estimate we proceed in a similar but simpler way. Consider for $\varepsilon>0$ the closed set $M_\varepsilon=\{Q\in\mathcal{M}^\pi|K_1(Q)-K_0(Q)\leq\varepsilon\}$. Then we get by Theorem 4 and the continuity of K_0 and K_1

$$\limsup\frac{1}{n}\log\mathcal{B}_0^n\leq\lim_{\varepsilon\to0}\Big(-\inf_{M_\varepsilon\cap\mathcal{MS}^\pi(\mu)}K_0(Q)\Big)=-\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_1(Q)\leq K_0(Q)}}K_0(Q)$$

and the same with K_1 instead of K_0. By symmetry this yields

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log B_n(\lambda|X) &\leq \max\left(-\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_1(Q)\leq K_0(Q)}}K_0(Q),\ -\inf_{\substack{Q\in\mathcal{MS}^\pi(\mu)\\ K_0(Q)\leq K_1(Q)}}K_1(Q)\right)\\ &= -\inf_{Q\in\mathcal{MS}^\pi(\mu)}K_0(Q)\vee K_1(Q).\end{aligned}$$

⌈ Similarly one gets together with Theorem 6 for the Bayes risk of Theorem 1

$$\limsup_{n\to\infty}\frac{1}{n}\log B_n(\lambda|X)\leq-\inf_{Q\in\mathcal{M}_s^\pi}K_0(Q)\vee K_1(Q)=-\inf_{Q\in\mathcal{M}_s}K_0(Q)\vee K_1(Q)$$

since $K_0(Q)\vee K_1(Q)=\infty$ on $\mathcal{M}_s\backslash\mathcal{M}_s^\pi$. ⌉

So the first equality in the Theorem is proved.

II. For the proof of the second identity let us first discuss the case that the exponential risk rate equals $-\infty$. This means that $\mathcal{MS}^\pi(\mu)=\emptyset$ since $K_0\vee K_1<\infty$, on that set. Then there is no $I\in\mathcal{S}(\mu)$. This is equivalent to the assertion that the maximum in the second equality of the Theorem is taken over an empty set and hence is $-\infty$. Therefore we may assume in the following for some $I=E_1\cup\ldots\cup E_k\in\mathcal{S}^\pi(\mu)$ the set $\mathcal{M}_s^\pi(I)$ is not empty.
Observe that by the variational principle of relative entropy (see e.g. [DS89], p. 68) the functions $K_\theta:\mathcal{M}_s^\pi\longrightarrow[0,\infty)$ are convex, since they are suprema of convex functions. We

consider the compact convex set $\mathcal{M}_s^\pi(I)$. Note that $\mathcal{M}_s^\pi(I) \subset \mathcal{M}^\pi$ and hence K_0 and K_1 are finite and continuous on $\mathcal{M}_s^\pi(I)$. The map $k : [0,1] \times \mathcal{M}_s^\pi(I) \longrightarrow (-\infty, 0]$ given by $k(t,Q) = -(t\ K_0(Q) + (1-t)\ K_1(Q))$ is affine in t and concave in Q. Thus we can apply the minimax Theorem (e.g. [ET85],prop. VI.2.1) to get

$$\begin{aligned} -\inf_{\mathcal{M}_s^\pi(I)} K_0(Q) \vee K_1(Q) &= \sup_{\mathcal{M}_s^\pi(I)} \min(-K_0(Q), -K_1(Q)) \\ &= \sup_{\mathcal{M}_s^\pi(I)} \inf_{0\le t\le 1} k(t,Q) \\ &= \inf_{0\le t\le 1} \sup_{\mathcal{M}_s^\pi(I)} k(t,Q) \\ &= \inf_{0<t<1} \sup_{\mathcal{M}_s^\pi(I)} k(t,Q) \end{aligned}$$

where the last equality follows from the continuity of k. The proposition below applied to the matrices $\pi_t^{E_l}$ gives together with Remark 1 ii).

$$\begin{aligned} \sup_{\mathcal{M}_s^\pi(I)} k(t,Q) &= \sup_{\mathcal{M}_s^\pi(I)} -\sum_{i,j\in I} Q_{ij}\left(\log\frac{Q_{ij}}{Q_i} - t\log\pi_0(i,j) - (1-t)\log\pi_1(i,j)\right) \\ &= \sup_{\mathcal{M}_s^\pi(I)} \left(H_s(Q) + \sum_{i,j\in I} Q_{ij}\log\pi_t(i,j)\right) \\ &= \max_{l=1,\dots,k} \sup_{\mathcal{M}_s^\pi(E_l)} \left(H_s(Q) + \sum_{i,j\in I} Q_{ij}\log\pi_t(i,j)\right) \\ &= \max_{l=1,\dots,k} \log\rho(\pi_t^{E_l}) \\ &= \log\rho(\pi_t^I). \end{aligned}$$

Maximizing over all $\pi-$strings $I \in \mathcal{S}^\pi(\mu)$ we arrive at

$$\begin{aligned} -\inf_{Q\in\mathcal{MS}^\pi(\mu)} K_0(Q) \vee K_1(Q) &= \max_{I\in\mathcal{S}^\pi(\mu)} -\inf_{\mathcal{M}_s^\pi(I)} K_0(Q) \vee K_1(Q) \\ &= \max_{I\in\mathcal{S}^\pi(\mu)} \inf_{0<t<1} \log\rho(\pi_t^I). \end{aligned}$$

⌈ The same calculation applied to the rate of Theorem 1 gives

$$-\inf_{Q\in\mathcal{M}_s} K_0(Q) \vee K_1(Q) = \inf_{0<t<1} \log\rho(\pi_t). \qquad \rceil$$

This completes the proof. q.e.d.

In the proof of the last equality in the Theorem we made use of the following variational representation of the spectral radius of a nonnegative matrix. The proof in Ellis ([Ell85], p. 284), Theorem IX.4.4 is applicable even though the statement of the Theorem there is formally weaker.

Proposition 2: *Let $(\pi(i,j))_{i,j\in I}$ be a real nonnegative irreducible matrix. Then*

$$\log\rho(\pi) = \sup\{H_s(Q) + \sum Q_{ij}\log\pi(i,j)\}$$

where the sup *is taken over all $Q \in \mathcal{M}_s$ with $Q_{ij} = 0$ if $\pi(i,j) = 0$. The* sup *is attained at a matrix of the form $(u_i\pi(i,j)v_j)$ where u and v are left and right eigenvectors of π. In particular the optimal Q has the same zero entries as π.*

6 Large Deviations for the Empirical Pair Measure of Finite State Markov Chains

In this section we prove some refinements of the large deviation Theorem for the empirical pair measure of finite Markov chains (for the standard result see e.g. [DZ93]). The main point is to get lower bounds for sets which have empty interior in the standard topology of $\mathcal{M}$. As a first step (Theorem 3) we consider open sets in the finer topology τ_0, which was introduced in section 2. This implies the large deviation Theorem 4 based on the λ-strings introduced in section 4. Moreover in Theorem 5 we get large deviation estimates for the probability that the empirical pair measure is in certain subsets of the set of stationary measures.

Theorem 3: *Let P^π be the law of a Markov chain with the irreducible transition matrix π. Let $U \subset \mathcal{M}^\pi$ be open in the topology τ_0. Then for all $x \in S$*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq - \inf_{Q \in U_x \cap \mathcal{M}_{s,i}} H(Q|Q \times \pi)$$

where $U_x = \{Q \in U \,|\, Q_x > 0\}$.

Proof: Let $Q \in U_x \cap \mathcal{M}_{s,i}$ and let $\eta > 0$. Let P^q be the law of the ergodic Markov chain which corresponds to Q. Choose $\delta > 0$ such that $U_\delta(Q) \subset U$ which is possible since U is τ_0-open. Applying either the Large Deviation Theorem for the standard topology (see Theorem 6) or simply the classical coding Theorem of Shannon Theory it is also possible to choose δ small enough such that

$$\liminf_{n\to\infty} \frac{1}{n} \log P_x^q(\hat{m}_n \in V_\delta(Q)) \geq -\eta \tag{7}$$

where $V_\delta(Q)$ is the δ-neighbourhood of Q in the standard topology. Since however every $Q' \in \mathcal{M}$ with $P_x^q(\hat{m}_n = Q') > 0$ for some n is absolutely continuous with respect to Q we get

$$P_x^q(\hat{m}_n \in V_\delta(Q)) = \sum_{Q' \in U_\delta(Q)} N_{n,x}(Q') \prod_{i,j} q_{ij}^{nQ'_{ij}} \tag{8}$$

where $q_{ij} = \frac{Q_{ij}}{Q_i}$ and $N_{n,x}(Q') = \#\{\omega \in S^{n+1} : \hat{m}_n(\omega) = Q', \omega_0 = x\}$. The number of nonzero terms in the sum grows only polynomially in n as $n \longrightarrow \infty$. Therefore (7) implies

$$\liminf_{n\to\infty} \max_{Q' \in U_\delta(Q)} \left(\frac{1}{n} \log N_{n,x}(Q') + \sum_{i,j} Q'_{ij} \log q_{ij} \right) \geq -\eta. \tag{9}$$

Replacing in (8) the transition matrix (q_{ij}) by π we have

$$\begin{aligned} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U_\delta(Q)) &\geq \max_{Q' \in U_\delta(Q)} \left(\frac{1}{n} \log N_{n,x}(Q') + \sum_{i,j} Q'_{ij} \log \pi_{ij} \right) \\ &\geq \left(\max_{Q' \in U_\delta(Q)} \left(\frac{1}{n} \log N_{n,x}(Q') \right) \right) + \sum_{i,j} Q_{ij} \log \pi_{ij} - \eta \end{aligned}$$

if δ is sufficiently small. Then together with (9) and the fact that $Q \ll \pi$ we get

$$\liminf_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U_\delta(Q)) \geq \sum_{i,j} Q_{ij} \log \frac{q_{ij}}{\pi_{ij}} - 2\eta.$$

Since $U_\delta(Q) \subset U$ this completes the proof. *q.e.d.*

In the proof of the next Theorem we need the following straightforward fact.

Lemma 1: *Let $Q \in \mathcal{M}_s$. Then there exists a unique representation $Q = \sum_l \alpha_l Q_l$ with $Q_l \in \mathcal{M}_{s,i}$.*

Proof: This Lemma reflects the general ergodic decomposition of stationary processes. Here is a direct proof.
For $Q \in \mathcal{M}_s$ we can write $Q_{i\,j} = \mu_i \pi_{i\,j}$, where $(\pi_{i\,j})_{i,j \in S}$ is a transition matrix and $(\mu_i)_{i \in S}$ a stationary distribution according to the matrix π, defined by $\mu_i = \sum_j Q_{i\,j} = \sum_j Q_{j\,i}$. Let P denote a Markov chain with initial distribution μ and transition matrix π. The renewal Theorem tells us that for transient states i we have

$$P_j(X_n = i) \stackrel{n \to \infty}{\longrightarrow} 0 \qquad \text{for every } j$$

So, since

$$\mu_i = P(X_n = i) = \sum_j \mu_j P_j(X_n = i)$$

holds for every n, it follows that $\mu_i = 0$ for all transient states i. Thus $Q_{i\,j} = 0 = Q_{j\,i}$ for i transient and all j. This means that Q is a block matrix consisting of irreducible blocks. So we can write $Q = \sum_l \alpha_l Q_{|E_l \times E_l}$ with $\alpha_l = \sum_{r,s \in E_l} Q_{r\,s}$ and $Q_{|E_l \times E_l} \in \mathcal{M}_{s,i}$ *q.e.d.*

The following Theorem shows the use of strings in connection with large deviations.

Theorem 4: *Let P^π be the law of a Markov chain with the irreducible transition matrix π. Let λ be a matrix with $\lambda \ll \pi$, $x \in S$.*
a) If U is a relatively open subset of $\mathcal{M}^\lambda$ in the topology of coordinatewise convergence then

$$\liminf_{n \to \infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq - \inf_{U \cap \mathcal{MS}^\lambda(x)} H(Q|Q \times \pi).$$

b) If $U \subset \mathcal{M}^\lambda$ is closed in the topology of coordinatewise convergence then

$$\limsup_{n \to \infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \leq - \inf_{U \cap \mathcal{MS}^\lambda(x)} H(Q|Q \times \pi).$$

Proof: First of all we want to show what happens if $\mathcal{MS}^\lambda(x) = \emptyset$. If this happens, no λ-possible path ω longer than $\#S$ with $\hat{m}_n \in U$ exists since otherwise ω would visit at least one state a several times, which means that this state would belong to a λ-equivalence class and $\mathcal{MS}^\lambda(x)$ could not be empty. Therefore $P_x^\pi(\hat{m}_n \in U) = 0$ if $n > \#S$ and lim inf and lim sup are both equal to $-\infty$. Thus the Theorem is true in this case since the infimum over an empty set is ∞.

a) Let $U \subset \mathcal{M}^\lambda$ be relatively open. Choose $Q \in U \cap \mathcal{M}_s^\lambda(I)$ where $I = E_1 \cup \cdots \cup E_k$ is some λ-string in $\mathcal{S}^\lambda(x)$ and let $\eta > 0$. We want to show that

$$\liminf_{n \to \infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq -H(Q|Q \times \pi) - \eta.$$

Since Q is stationary it has a finite representation $Q = \sum_{l=1}^{L} \alpha_l Q_l$ with $Q_l \in \mathcal{M}_{s,i}$ according to the preceeding lemma.
Choose the sets J_l minimal such that Q_l is concentrated on $J_l \times J_l$. Each J_l is contained in **exactly** one of the $\underset{\lambda}{\sim}$ -equivalence classes E_r because if J_l meets E_{r_1} and E_{r_2} then due to the irreducibility of Q_l a Q_l- Markov chain could pass between E_{r_1} and E_{r_2} in both directions with positive probability in contradiction to $Q \ll \lambda$ and the definition of λ-strings. In particular $L \geq k$.
Since the E_i form a $\lambda-$string and $\lambda \ll \pi$ there are numbers $m \in \mathbb{N}$ and $p > 0$ such that $P_x^{\pi}(X_m \in J_1, \hat{m}_m \in \mathcal{M}^{\lambda}) \geq p$ and

$$P_i^{\pi}(X_m \in J_{l+1}, \hat{m}_m \in \mathcal{M}^{\lambda}) \geq p \tag{10}$$

whenever $i \in J_l$ and $l \in \{1, \ldots, L-1\}$.

Applying Theorem 3 we can choose $\delta > 0$ such that for every $l \in \{1, \cdots, L\}$ and every $z \in J_l$ we have

$$\liminf \frac{1}{n} \log P_z^{\pi}\left(\hat{m}_n \in U_{\delta}(Q_l)\right) \geq -H(Q_l | Q_l \times \pi) - \frac{\eta}{2L}. \tag{11}$$

In the following we assume that n is a sufficiently large integer. We can choose for $l \in \{1, \cdots, L\}$ indices $t_{n,l} \in \{1, \cdots, n\}$ such that

$$\lim_{n \to \infty} \frac{t_{n,l}}{n} = \alpha_l \tag{12}$$

and

$$\sum_{l=1}^{L} t_{n,l} \leq n - Lm. \tag{13}$$

We partition $\{0, \ldots, n\}$ into L segments ($\{s_{n,l} - m, \ldots, s_{n,l} + t_{n,l} - 1\}$ for $l = 1, \ldots, L-1$ and $\{s_{n,L} - m, \ldots, s_{n,L} + t_{n,L} = n\}$) by setting $s_{n,1} = m$ and $s_{n,l} = \sum_{s=1}^{l-1}(t_{n,s} + m) + m$ for $l \in \{2, \cdots, L\}$. Consider the empirical pair measure $\hat{m}_l^n$ which is induced by the first $t_{n,l}$ steps after $s_{n,l}$, i.e.

$$(\hat{m}_l^n(\omega))_{i,j} := \frac{1}{t_{n,l}} \sum_{s=s_{n,l}}^{s_{n,l}+t_{n,l}-1} \chi_{i,j}(\omega_s, \omega_{s+1}).$$

Moreover let B_l be the set of all paths for which the first part of the l-th segment is λ-possible, i.e.

$$B_l = \left\{\omega \; : \quad \hat{b}_l(\omega) := \left(\frac{1}{m} \sum_{s=s_{n,l}-m}^{s_{n,l}-1} \chi_{i,j}(\omega_s, \omega_{s+1})\right)_{i,j \in S} \in \mathcal{M}^{\lambda}\right\}.$$

Consider the set $\bigcap_{l=1}^{L} B_l \cap \{\hat{m}_l^n \in U_{\delta}(Q_l)\}$. For every element of this set the empirical pair measure is given by

$$\hat{m}_n(\omega) = \sum_{l=1}^{L} \left(\frac{t_{n,l}}{n} \hat{m}_l^n(\omega) + \frac{m}{n} \hat{b}_l(\omega)\right).$$

Because of (12) and (13) we have uniformly on that set

$$\lim_{n \to \infty} |\hat{m}_n(\omega) - \sum_{l=1}^{L} \alpha_l \hat{m}_l^n(\omega)| = 0. \tag{14}$$

Moreover we know that if $\hat{m}_l^n(\omega) \in U_\delta(Q_l)$ for every l, then for sufficiently large n

$$|\sum_{l=1}^{L} \alpha_l \hat{m}_l^n(\omega) - \sum_{l=1}^{L} \alpha_l Q_l| < \delta.$$

Thus we have together with (14)

$$|\hat{m}_n(\omega) - Q| < \delta$$

for all $\omega \in \bigcap_{l=1}^{L} B_l \cap \{\hat{m}_l^n \in U_\delta(Q_l)\}$. Since the transitions of every such path ω are possible with respect to λ (this is clear in the sets J_l since there $\hat{m}_l^n(\omega) \ll Q_l \ll \lambda$ and outside those sets it is guaranteed by the condition B_l) and because $Q \in U$ and U is a relatively open subset of $\mathcal{M}^\lambda$ in the topology of coordinatewise we have for sufficiently small δ

$$\liminf \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq \liminf \frac{1}{n} \log P_x^\pi\Big(\bigcap_{l=1}^{L} B_l \cap \{\hat{m}_l^n \in U_\delta(Q_l)\}\Big). \tag{15}$$

In order to estimate the probability of this intersection let i be any element of the set J_{l-1} with $P_x^\pi(X_{s_{n,l}-m} = i) > 0$. Then

$$\begin{aligned} & P_x^\pi\Big(B_l, \{\hat{m}_l^n \in U_\delta(Q_l)\} | X_{s_{n,l}-m} = i\Big) \\ = & \sum_{z \in J_l} P^\pi\Big(B_l, X_{s_{n,l}} = z | X_{s_{n,l}-m} = i\Big) \cdot P^\pi(\hat{m}_l^n \in U_\delta(Q_l) | X_{s_{n,l}} = z) \\ \geq & P_i^\pi(\hat{m}_m \in \mathcal{M}^\lambda, X_m \in J_l) \inf_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l)). \end{aligned}$$

Now let $A_{l-1} \in \sigma\{X_0, \ldots, X_{s_{n,l}-m}\}$ be such that $X_{s_{n,l}-m} \in J_{l-1}$ P_x^π-a.s. on A_{l-1}. From (10) we can now conclude that

$$\begin{aligned} & P_x^\pi\Big(B_l, \{\hat{m}_l^n \in U_\delta(Q_l)\} | A_{l-1}\Big) \\ \geq & \sum_{i \in J_{l-1}} P_x^\pi(X_{s_{n,l}-m} = i | A_{l-1}) \cdot p \cdot \inf_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l)) = p \inf_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l)). \end{aligned}$$

By induction we get

$$P_x^\pi\Big(\bigcap_{l=1}^{L} B_l \cap \{\hat{m}_l^n \in U_\delta(Q_l)\}\Big) \geq p^L \prod_{l=1}^{L} \inf_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l))$$

and using (11),(12) and (15)

$$\begin{aligned} & \liminf \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq \sum_{l=1}^{L} \liminf \frac{t_{n,l}}{n} \frac{1}{t_{n,l}} \log \min_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l)) \\ = & \sum_{l=1}^{L} \alpha_l \liminf \frac{1}{t_{n,l}} \log \min_{z \in J_l} P_z^\pi(\hat{m}_{t_{n,l}} \in U_\delta(Q_l)) \geq -\sum_{l=1}^{L} \alpha_l H(Q_l | Q_l \times \pi) - \eta \\ = & -H(\sum_{l=1}^{L} \alpha_l Q_l | (\sum_{l=1}^{L} \alpha_l Q_l) \times \pi) - \eta = -H(Q | Q \times \pi) - \eta\,. \end{aligned}$$

Here we have used the linearity of remark 1 ii). This proves part a) of the theorem.
b) Let $U \subset \mathcal{M}^\lambda$ and $\delta > 0$ be given. Let ω be a path such that $\hat{m}_n(\omega) \in U$ and $\omega_0 = x$. Then $\lambda(\omega_s, \omega_{s+1}) > 0$ for $s = 0, \ldots, n-1$. This implies that there are a λ-string $I = E_1 \cup \ldots \cup E_k \in \mathcal{S}^\lambda(x)$ and numbers $0 = s_1, \ldots, s_{k+1} = n - m$, with $m \in \mathbb{N}$ fixed, such that $\omega_{s_r+m}, \ldots, \omega_{s_{r+1}} \in E_r$ for $r = 1, \ldots, k$. Let (like in the proof of part a) $\hat{m}_r^n$ denote the empirical pair measure corresponding to the part $\omega_{s_r+m}, \ldots, \omega_{s_{r+1}}$ of the path. Since E_r is a $\underset{\lambda}{\sim}$ -equivalence class this part of the path can be extended by $\#S$ steps to a closed loop of λ-possible steps in E_r. The empirical pair measure Q_r of this loop is stationary and concentrated on $E_r \times E_r$. Hence it is an element of $\mathcal{MS}^\lambda(x)$ and it differs from $\hat{m}_r^n$ at most by $\frac{\#S}{n} < \delta$ if $n \geq n_0$ where n_0 can be chosen independently of ω. Thus the measure $\hat{m}_n(\omega)$ differs at most by δ from a suitable convex combination of the Q_r. This combination is also in $\mathcal{MS}^\lambda(x)$. Hence for sufficiently large n

$$P_x^\pi(\hat{m}_n \in U) = P_x^\pi(\hat{m}_n \in U_x \cap V_\delta(\mathcal{MS}^\lambda(x))).$$

Therefore by Theorem 6 we get

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) &\leq \inf_{\delta>0} - \frac{\inf}{U_x \cap V_\delta(\mathcal{MS}^\lambda(x))} H(Q|Q \times \pi) \\ &\leq \inf_{\delta>0} - \frac{\inf}{U \cap V_\delta(\mathcal{MS}^\lambda(x))} H(Q|Q \times \pi) \\ &= - \inf_{U \cap \mathcal{MS}^\lambda(x)} H(Q|Q \times \pi), \end{aligned}$$

since $\mathcal{MS}^\lambda(x)$ is closed and the rate function is continuous on that set. *q.e.d.*

Similarly one gets new bounds if the set U is contained in $\mathcal{M}_{s,i}$. The set of irreducible measures is τ_0 open in contrast to the set $\mathcal{M}_{s,i}$. Therefore the estimates of the next result do not follow directly from Theorem 3. Since the empirical pair measure can be stationary only if the path has completed a loop the probability in the following result is nonzero only on a periodic set of times. Therefore the liminf has to be taken along a suitable subsequence of the integers.

Theorem 5: *Let P^π be the law of a Markov chain with irreducible transition matrix π.*
a) Let U be a relatively τ_0-open subset of $\mathcal{M}_{s,i}$ and let d be the least common multiple of the set of periods $\{d_Q : Q \in U\}$. Then with $U_x = \{Q \in U \,|\, Q_x > 0\}$ we have

$$\liminf_{n\to\infty} \frac{1}{nd} \log P_x^\pi(\hat{m}_{nd} \in U) \geq - \inf_{U_x} H(Q|Q \times \pi).$$

b) Let U be a subset of $\mathcal{M}_{s,i}$. Then

$$\limsup_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \leq - \inf_{U_x} H(Q|Q \times \pi).$$

Proof: a) Let $Q \in U_x$ be such that $P_x^\pi(\hat{m}_n(\omega) = Q) > 0$ for some ω and n and $\eta > 0$. Then according to Theorem 3 there is some $\delta > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U_{\frac{\delta}{2}}(Q)) \geq -H(Q|Q \times \pi) - \eta.$$

Choose $m \in \mathbb{N}$ and $I \subset S$ such that Q^{md} consists of strictly positve blocks. Then every path $\omega_0, \ldots, \omega_{nd}$ with $\hat{m}_{nd}(\omega) \in U_{\frac{\delta}{2}}(Q)$ can be extended to a closed path ω' by md additional $\pi-$possible steps. Then $Q' := \hat{m}_{(n+m)d}(\omega') \in \mathcal{M}_{s,i}$ and $|Q' - m_{nd}(\omega)| < \frac{\delta}{2}$ for sufficiently large n. Therefore $Q' \in U_\delta(Q) \cap \mathcal{M}_{s,i} \subset U$. If p is the minimal probability of these extensions we get

$$\liminf_{n\to\infty} \frac{1}{nd} \log P_x^\pi(\hat{m}_{nd} \in U) = \liminf_{n\to\infty} \frac{1}{(n+m)d} \log P_x^\pi(\hat{m}_{(n+m)d} \in U)$$

$$\geq \liminf_{n\to\infty} \frac{1}{nd} \log P_x^\pi(\hat{m}_{nd} \in U_{\frac{\delta}{2}}(Q))\, p \geq -H(Q|Q\times\pi) - \eta$$

proving part a).

For b) note that the usual large deviation result (Theorem 6) below implies

$$\limsup_{n\to\infty} P_x^\pi(\hat{m}_n \in U) = \limsup_{n\to\infty} P_x^\pi(\hat{m}_n \in U_x \cap \mathcal{M}_\pi)$$

$$\leq -\inf_{U_x \cap \mathcal{M}^\pi} H(Q|Q\times\pi) = -\inf_{U_x} H(Q|Q\times\pi)$$

where the last equality uses that $\mathcal{M}^\pi$ is the set where the rate function is finite and that this function is continuous on $\mathcal{M}^\pi$. *q.e.d.*

The following Theorem will be given without proof. It can be found as remark (a) to Theorem 3.1.13 in [DZ93]

Theorem 6: *Let P^π be the law of an irreducible Markov chain with nonnegative transition matrix π and $x \in S$.*
a) For a relatively $\mathcal{M}^\pi$-open set $U \subset \mathcal{M}^\pi$ we have

$$\liminf_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in U) \geq -\inf_{U\cap\mathcal{M}_s} H(Q|Q\times\pi).$$

b) For a closed set $C \subset \mathcal{M}^\pi$ it holds

$$\limsup_{n\to\infty} \frac{1}{n} \log P_x^\pi(\hat{m}_n \in C) \leq -\inf_{C\cap\mathcal{M}_s} H(Q|Q\times\pi).$$

7 Conclusion

We considered in this paper a finite family of laws of irreducible Markov chains on the same finite state space with pairwise different transition matrices. We obtained the rate at which these laws become singular as the number of observations increases. We use the fact that this rate can be expressed as the rate alternatively of the entropy risk or of the minimal Bayes risk. The main reason why this rate can be computed for Markov chains is the existence of a sufficient statistic namely the empirical pair measure which can be treated by large deviation theory for each parameter. In the case the transition matrices are strictly positive we can use known results from the theory of large deviatons, whereas for the general case we refine the corresponding classical results. The ideas of this paper will be extended to more general random processes.

References

[Ber79] J.M. Bernardo. Reference Posterior Distributions for Bayesian Inference. *J. R. Statist. Soc., Ser. B*, 41:113–147, 1979.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[DS89] J.-D. Deuschel and D.W. Stroock. *Large Deviations.* Academic Press, Boston etc., 1989.

[DZ93] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett Publishers, Boston etc., 1993.

[Ell85] R.J. Ellis. *Entropy, Large Deviations, and Statistical Mechanics.* Springer-Verlag, New York etc., 1985.

[ET85] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems.* North Holland, American Elsevier, Amsterdam etc., 1985.

[KW93] J. Krob and H.v. Weizsäcker. The rate of information gain in experiments with a finite parameter set. *submitted*, 1993.

[Lin56] D.V. Lindley. On a Measure of the Information provided by an Experiment. *Ann. Math. Stat.*, 27:986–1005, 1956.

[Sen81] E. Seneta. *Non-negative Matrices and Markov Chains.* Springer Verlag, New York, 1981.

[Tor81] E.N. Torgersen. Measures of Information based on Comparison with total Information and with total Ignorance. *Ann. of Statistics*, 9:638–657, 1981.