# Chapter 32

# Large Deviations for Markov Sequences

This chapter establishes large deviations principles for Markov sequences as natural consequences of the large deviations principles for IID sequences in Chapter 31. (LDPs for continuous-time Markov processes will be treated in the chapter on Freidlin-Wentzell theory.)

Section 32.1 uses the exponential-family representation of Markov sequences to establish an LDP for the two-dimensional empirical distribution ("pair measure"). The rate function is a relative entropy.

Section 32.2 extends the results of Section 32.1 to other observables for Markov sequences, such as the empirical process and time averages of functions of the state.

For the whole of this chapter, let $X_1, X_2, \ldots$ be a homogeneous Markov sequence, taking values in a Polish space $\Xi$, with transition probability kernel $\mu$, and initial distribution $\nu$ and invariant distribution $\rho$. If $\Xi$ is not discrete, we will assume that $\nu$ and $\rho$ have densities $n$ and $r$ with respect to some reference measure, and that $\mu(x, dy)$ has density $m(x, y)$ with respect to that same reference measure, for all $x$. (LDPs can be proved for Markov sequences without such density assumptions — see, e.g., Ellis (1988) — but the argument is more complicated.)

## 32.1 Large Deviations for Pair Measure of Markov Sequences

It is perhaps not sufficiently appreciated that Markov sequences form exponential families (Billingsley, 1961; Küchler and Sørensen, 1997). Suppose $\Xi$ is

discrete. Then

$$\begin{aligned}
\mathbb{P}\left(X_1^n = x_1^t\right) &= \nu(x_1)\prod_{i=1}^{t-1}\mu(x_i, x_{i+1}) && (32.1)\\
&= \nu(x_1)e^{\sum_{i=1}^{t-1}\log\mu(x_i,x_{i+1})} && (32.2)\\
&= \nu(x_1)e^{\sum_{x,y\in\Xi^2} T_{x,y}(x_1^t)\log\mu(x,y)} && (32.3)
\end{aligned}$$

where $T_{x,y}(x_1^t)$ counts the number of times the state $y$ follows the state $x$ in the sequence $x_1^t$, i.e., it gives the *transition counts*. What we have just established is that the Markov chains on $\Xi$ with a given initial distribution form an exponential family, whose natural sufficient statistics are the transition counts, and whose natural parameters are the logarithms of the transition probabilities.

(If $\Xi$ is not continuous, but we make the density assumptions mentioned at the beginning of this chapter, we can write

$$\begin{aligned}
p_{X_1^t}(x_1^t) &= n(x_1)\prod_{i=1}^{t-1} m(x_i, x_{i+1}) && (32.4)\\
&= n(x_1)e^{\int_{\Xi^2} dT(x_1^t)\log m(x,y)} && (32.5)
\end{aligned}$$

where now $T(x_1^t)$ puts probability mass $\frac{1}{n-1}$ at $x, y$ for every $i$ such that $x_i = x$, $x_{i+1} = y$.)

We can use this exponential family representation to establish the following basic theorem.

**Theorem 437** *Let $X_i$ be a Markov sequence obeying the assumptions set out at the beginning of this chapter, and furthermore that $\mu(x,y)/\rho(y)$ is bounded above (in the discrete-state case) or that $m(x,y)/r(y)$ is bounded above (in the continuous-state case). Then the two-dimensional empirical distribution ("pair measure") $\hat{P}_t^2$ obeys an LDP with rate $n$ and with rate function $J_2(\psi) = D\left(\psi\|\pi_1\psi\times\mu\right)$ if $\nu$ is shift-invariant, $J(\nu) = \infty$ otherwise.*

PROOF: I will just give the proof for the discrete case, since the modifications for the continuous case are straightforward (given the assumptions made about densities), largely a matter of substituting Roman letters for Greek ones.

First, modify the representation of the probabilities in Eq. 32.3 slightly, so that it refers directly to $\hat{P}_t^2$ (as laid down in Definition 413), rather than to the transition counts.

$$\begin{aligned}
\mathbb{P}\left(X_1^t = x_1^t\right) &= \frac{\nu(x_1)}{\mu(x_t, x_1)}e^{t\sum_{x,y\in\Xi}\hat{P}_t^2(x,y)\log\mu(x,y)} && (32.6)\\
&= \frac{\nu(x_1)}{\mu(x_t, x_1)}e^{n\mathbf{E}_{\hat{P}_t^2}[\log\mu(X,Y)]} && (32.7)
\end{aligned}$$

Now construct a sequence of IID variables $Y_i$, all distributed according to $\rho$, the invariant measure of the Markov chain:

$$\mathbb{P}\left(Y_1^t = y_1^t\right) = e^{n\mathbf{E}_{\hat{P}_t^2}[\log\rho(Y)]} \quad (32.8)$$

The ratio of these probabilities is the Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_X}{d\mathbb{P}_Y}(x_1^t) = \frac{\nu(x_1)}{\mu(x_t, x_1)} e^{t\mathbf{E}_{\hat{P}_n^2}\left[t \log \frac{\mu(X,Y)}{\rho(Y)}\right]} \tag{32.9}$$

(In the continuous-$\Xi$ case, the derivative is the ratio of the densities with respect to the common reference measure, and the principle is the same.) Introducing the functional $F(\nu) = \mathbf{E}_\nu\left[\log \frac{\mu(X,Y)}{\rho(Y)}\right]$, the derivative is equal to $O(1)e^{tF(\hat{P}_t^2)}$, and our initial assumption amounts to saying that $F$ is not just continuous (which it must be) but bounded from above.

Now introduce $Q_{t,X}$, the distribution of the empirical pair measure $\hat{P}_t^2$ under the Markov process, and $Q_{t,Y}$, the distribution of $\hat{P}_t^2$ for the IID samples produced by $Y_i$. From Eq. 32.9,

$$\begin{aligned}
\frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \frac{1}{t}\log \int_B dQ_{t,X}(\psi) && (32.10)\\
&= \frac{1}{t}\log \int_B \frac{dQ_{t,X}}{dQ_{t,Y}} dQ_{t,Y}(\psi) && (32.11)\\
&= O\left(\frac{1}{t}\right) + \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.12)
\end{aligned}$$

It is thus clear that

$$\begin{aligned}
\liminf \frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \liminf \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.13)\\
\limsup \frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \limsup \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.14)
\end{aligned}$$

Introduce a (final) proxy random sequence, also taking values in $\mathcal{P}(()\,\Xi^2)$, call it $Z_t$, with $\mathbb{P}(Z_t \in B) = \int_B e^{tF(\psi)} dQ_{t,Y}(\psi)$. We know (Corollary 435) that, under $Q_{t,Y}$, the empirical pair measure satisfies an LDP with rate $t$ and good rate function $J_Y = D(\psi \| \pi_1\psi \otimes \rho)$, so by Corollary 416, $Z_t$ satisfies an LDP with rate $t$ and good rate function

$$J_F(\psi) = -(F(\psi) - J_Y(\psi)) + \sup_{\zeta \in \mathcal{P}(\Xi^2)} F(\zeta) - J_Y(\zeta) \tag{32.15}$$

A little manipulation turns this into

$$J_F(\psi) = D(\psi \| \pi_1\psi \otimes \mu) - \inf_{\zeta \in \mathcal{P}(\Xi^2)} D(\zeta \| \pi_1\zeta \otimes \mu) \tag{32.16}$$

and the infimum is clearly zero. Since this is the rate function $Z_t$, in view of Eqs. 32.13 and 32.14 it is also the rate function for $\hat{P}_n^2$, which we have agreed to call $J_2$. □

*Remark 1:* The key to making this work is the assumption that $F$ is bounded from above. This can fail if, for instance, the process is not ergodic, although usually in that case one can rescue the general idea by some kind of ergodic decomposition.

*Remark 2:* The LDP for the pair measure of an IID sequence can now be seen to be a special case of the LDP for the pair measure of a Markov sequence. The same is true, generally speaking, of all the other LDPs for IID and Markov sequences. Calculations are almost always easier for the IID case, however, which permits us to give explicit formulae for the rate functions of empirical means and empirical distributions unavailable (generally speaking) in the Markovian case.

**Corollary 438** *The minima of the rate function $J_2$ are the invariant distributions.*

PROOF: The rate function is $D(\psi\|\pi_1\psi\otimes\mu)$. Since relative entropy is $\geq 0$, and equal to zero iff the two distributions are equal (Lemma 360), we get a minimum of zero in the rate function iff $\psi = \pi_1\psi\otimes\mu$, or $\psi = \rho^2$, for some $\rho\in\mathcal{P}(\Xi)$ such that $\rho\mu = \rho$. Conversely, if $\psi$ is of this form, then $J_2(\psi) = 0$. □

**Corollary 439** *The empirical distribution $\hat{P}_t$ obeys an LDP with rate $t$ and good rate function*

$$J_1(\psi) = \inf_{\zeta\in\mathcal{P}(\Xi^2):\pi_1\zeta=\psi} D(\zeta\|\pi_1\zeta\otimes\mu) \tag{32.17}$$

PROOF: This is a direct application of the Contraction Principle (Theorem 410), as in Corollary 415. □

*Remark:* Observe that if $\psi$ is invariant under the action of the Markov chain, then $J_1(\psi) = 0$ by a combination of the preceding corollaries. This is good, because we know from ergodic theory that the empirical distribution converges on the invariant distribution for an ergodic Markov chain. In fact, in view of Lemma 361, which says that $D(\psi\|\rho)\geq\frac{1}{2\ln 2}\|\psi-\rho\|_1^2$, the probability that the empirical distribution differs from the invariant distribution $\rho$ by more than $\delta$, in total variation distance, goes down like $O(e^{-t\delta^2/2})$.

**Corollary 440** *If Theorem 437 holds, then time averages of observables, $A_t f$, obey a large deviations principle with rate function*

$$J_0(a) = \inf_{\zeta\in\mathcal{P}(\Xi^2):\ \mathbf{E}_{\pi_1\zeta}[f(X)]} D(\zeta\|\pi_1\zeta\otimes\mu) \tag{32.18}$$

PROOF: Another application the Contraction Principle, as in Corollary 415. □

*Remark:* Observe that if $a = \mathbf{E}_\rho[f(X)]$, with $\rho$ invariant, then the $J_0(a) = 0$. Again, it is reassuring to see that large deviations theory is compatible with ergodic theory, which tells us to expect the almost-sure convergence of $A_t f$ on $\mathbf{E}_\rho[f(X)]$.

**Corollary 441** *If $X_i$ are from a Markov sequence of order $k+1$, then, under conditions analogous to Theorem 437, the $k+1$-dimensional empirical distribution $\hat{P}_t^{k+1}$ obeys an LDP with rate $t$ and good rate function*

$$D(\nu\|\pi_{k-1}\nu\otimes\mu) \tag{32.19}$$

PROOF: An obvious extension of the argument for Theorem 437, using the appropriate exponential-family representation of the higher-order process. □

Whether all exponential-family stochastic processes (Küchler and Sørensen, 1997) obey LDPs is an interesting question; I'm not sure if anyone knows the answer.

## 32.2 Higher LDPs for Markov Sequences

In this section, I assume without further comment that the Markov sequence $X$ obeys the LDP of Theorem 437.

**Theorem 442** *For all $k \geq 2$, the finite-dimensional empirical distribution $\hat{P}_t^k$ obeys an LDP with rate $t$ and good rate function $J_k(\psi) = D\left(\psi \| \pi_{k-1}\psi \otimes \mu\right)$, if $\psi \in \mathcal{P}\left(\Xi^k\right)$ is shift-invariant, and $= \infty$ otherwise.*

PROOF: The case $k = 2$ is just Theorem 437. However, if $k \geq 3$, the argument preceding that theorem shows that $\mathbb{P}\left(\hat{P}_t^k \in B\right)$ depends only on $\pi_2 \hat{P}_t^k$, the pair measure implied by the $k$-dimensional distribution, so the proof of that theorem can be adapted to apply to $\hat{P}_t^k$, in conjunction with Corollary 435, establishing the LDP for finite-dimensional distributions of IID sequences. The identification of the rate function follows the same argument, too. □

**Theorem 443** *The empirical process distribution obeys an LDP with rate $t$ and good rate function $J_\infty(\psi) = d(\psi \| \rho)$, with $\rho$ here standing for the stationary process distribution of the Markov sequence.*

PROOF: Entirely parallel to the proof of Theorem 436, with Theorem 442 substituting for Corollary 435. □

Consequently, any continuous function of the empirical process distribution has an LDP.