Chapter 7
Models with Special Features

7.1 Introduction

Chapters 4 through 6 considered small noise large deviations of stochastic recursive equations, small noise moderate deviations for processes of the same type, and large deviations for the empirical measure of a Markov chain. These chapters thus consider models that are both standard and fairly general for each setting. In this chapter we consider discrete time models that are somewhat less standard, with the aim being to show how the weak convergence methodology can be adapted. The examples presented are just for illustrative purposes, and processes featuring other challenges are referenced at the end of the chapter.

We first consider occupancy models, which were originally introduced as simplified models for problems from physics. There are other interesting problems, such as the "coupon collector's problem" [165], that can be formulated in terms of occupancy models. In principle these problems can be treated using combinatorics. However, when the number of objects (e.g., distinct coupons) is large, combinatorial methods become numerically difficult, and large deviation approximations and related numerical methods can be more tractable. One can reformulate many occupancy problems in terms of Markov models of the type considered in Chap. 4, but owing to the fact that certain transition probabilities can be arbitrarily small the processes do not satisfy the conditions of that chapter. As will be discussed, the large deviation upper bound can be proved using essentially the same argument as in Chap. 4, but the lower bound requires a more careful analysis near points where the transition probabilities vanish [see Sect. 7.2.4]. A positive feature of these models is that for many occupancy-type problems one can solve to a fairly explicit degree for the optimal trajectories in variational problems that result from a large deviations analysis, and one can also construct explicit solutions to the related partial differential equations [see Sect. 7.2.5]. These, in turn, can be used to construct subsolutions for the accelerated Monte Carlo schemes discussed in Chaps. 14–17.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_7

The second class of models, discussed in Sect. 7.3, are discrete time recursive Markov models with two time scales. Such models and their continuous time counterparts occur in many applications, such as stochastic approximations [182] and chemical reaction networks [4]. Owing to a time scale separation, the large deviation properties of empirical measures are relevant, and these models can be analyzed using a combination of the arguments used for the small noise model of Chap. 4 and those applied in Chap. 6.

7.2 Occupancy Models

Occupancy problems center on the distribution of r balls that have been thrown into n urns. In the simplest scenario each ball is equally likely to fall in any of the urns, i.e., each ball is independently assigned to a given urn with probability $1/n$. In this case, we say that the urn model uses *Maxwell-Boltzmann* (MB) statistics. This model has been studied for decades and applied in diverse fields such as computer science, biology, and statistics. See [53, 154, 165] and the references therein. However, balls may also enter the urns in a nonuniform way. An important generalization is to allow the likelihood that the ball lands in a given urn to depend on its contents prior to the throw, as in *Bose-Einstein* (BE) and *Fermi-Dirac* (FD) statistics [129, 165, 219].

For MB statistics, many results have been obtained using "exact" methods. For example, combinatorial methods are used in [130] and methods that use generating functions are discussed in [165]. Although they do not directly involve approximations, the implementation of these methods can be difficult. For example, in combinatorial methods one has to deal with the difference of events using the inclusion-exclusion formula and the resulting computations can involve large errors. In the moment generating function approach in [165] similar difficulties occur. Large deviations approximations can give a useful alternative to both of these approaches. As we have discussed previously for other models, using large deviation theory one can often obtain useful qualitative insights. This is particularly true for problems of occupancy type, since in many cases variational problems involving the rate function can be solved explicitly.

In this section we consider a parametric family of models, of which the previously mentioned MB, BE and FD statistics are all special cases. We assume there are n urns and that $\lfloor Tn \rfloor$ balls are thrown into them (where $\lfloor s \rfloor$ denotes the integer part of s), and analyze the asymptotic properties as n goes to ∞. (In contrast with previous chapters we do not simplify notation by considering just the case $T = 1$. The reason for this is because there can be a link between the parameter that characterizes the particular statistics of the model and a limit on corresponding number of balls that may be thrown, which can constrain the value of T away from 1.) A typical problem of interest is to characterize the large deviation asymptotics of the empirical distribution after all the balls are thrown. For example, one may wish to estimate the probability that at most half of the urns are empty after all the balls are thrown. A direct analysis of this problem is hard, and instead we lift the problem to the process level and analyze the large deviation asymptotics at this process level.

Although we formulate occupancy models in terms of a stochastic recursive equation of the same general type as considered in Chap. 4, there are several interesting features, both qualitative and technical, that distinguish occupancy models from the processes studied in Chap. 4. The most significant of these as far as the proof is concerned are certain vanishing transition probabilities. A second very interesting feature which was commented on previously is that one can explicitly solve the variational problems that arise in the process level approximations. Such explicit formulas have many uses and add significantly to the practical value of the large deviation approximations.

In Sect. 7.2.1 the parametric family of occupancy problem is described in detail. A dynamical characterization of the occupancy model is given, and the representation for exponential integrals is stated. In Sect. 7.2.2 we prove the lower bound for the variational problem, which corresponds to the large deviation upper bound. Section 7.2.3 analyzes the rate function I, and proves properties that will allow us to deal with the technical difficulties of the vanishing transition probabilities. In Sect. 7.2.4, we prove the variational upper bound which corresponds to the large deviation lower bound. Finally, in Sect. 7.2.5 we give an explicit formula for the minimum of the rate function subject to a terminal constraint.

7.2.1 Preliminaries and Main Result

In this section, we formulate the problem of interest and state the LDP. The proof is given in sections that follow.

The general occupancy problem has the same structure as the MB occupancy problem, except that in the general problem urns are distinguished according to the number of balls contained therein. The full collection of models will be indexed by a parameter a. This parameter takes values in the set $(0, \infty] \cup \{-1, -2, \ldots\}$, and its interpretation is as follows. An urn is said to be of *category* i if it contains i balls. A ball is thrown in any given urn with probability proportional to $a + i$, where i denotes the category of the urn. In particular, suppose that a ball is about to be thrown, and that any two urns (labeled say A and B) are selected. Suppose that urn A is of category i, while B is of category j. Then the probability that the ball is thrown into urn A, conditioned on the state of all the urns and that the ball is thrown into either urn A or B, is

$$\frac{a+i}{(a+i)+(a+j)}.$$

When $a = \infty$ we interpret this to mean that the two urns are equally likely. Also, when $a < 0$ we use this ratio to define the probabilities only when $0 \le i \vee j \le -a$ and $i < -a$ or $j < -a$, so the formula gives a well defined probability. The probability that a ball is placed in an urn of category $-a$ is 0. Thus under this model, urns can only be of category $0, 1, \ldots, -a$, and we only throw balls into categories $0, 1, \ldots, -a-1$. Note that the case $a = 0$ is in some sense not interesting, in that as soon as there is an urn of category $j > 0$ all balls will be placed in that urn. Likewise the cases $a < 0$ but not an integer are hard to interpret.

In this setup, certain special cases are distinguished. The cases $a = 1$, $a = \infty$, $a \in -\mathbb{N}$ correspond to Bose-Einstein statistics, Maxwell-Boltzmann statistics, and Fermi-Dirac statistics, respectively.

Suppose that before we throw a ball there are already tn balls in all the urns, and further suppose that the occupancy state is $(x_0, x_1, \ldots, x_{J+})$. Here $x_i, i = 0, 1, \ldots, J$ denotes the fraction of urns that contain i balls, and x_{J+} denotes the fraction containing more than J balls. When $a < 0$, we will take $J = -a - 1$. The "un-normalized" or "relative" probability of throwing into a category i urn with $i \leq J$ is simply $(a + i)x_i$. Let us temporarily abuse notation, and let $x_{J+1}, x_{J+2}, \ldots$ denote the exact fraction in each category i with $i > J$. Since there are tn balls in the urns before we throw, $\sum_{i=0}^{\infty} i x_i = t$. Thus the (normalized and true) probability that the ball is placed in an urn that contains exactly i balls, $i = 0, 1, \ldots, J$, is $(a + i)x_i/(a + t)$, and the probability that the ball is placed in an urn that has more than J balls is $1 - \sum_{j=0}^{J}[(a + j)/(a + t)]x_j$.

An explicit construction of this process is as follows. To simplify, we assume the empty initial condition, i.e., all urns are empty. One can consider other initial conditions, with only simple notational changes in the results to be stated below. We introduce a time variable t that ranges from 0 to T. At a time t that is of the form l/n, with $0 \leq l \leq \lfloor nT \rfloor$ an integer, l balls have been thrown. Let $X^n(t) = (X_0^n(t), X_1^n(t), \ldots X_J^n(t), X_{J+}^n(t))'$ be the **occupancy state** at that time. As noted previously, $X_i^n(t)$ denotes the fraction of urns that contain i balls at time t, $i = 0, 1, \ldots, J$, and $X_{J+}^n(t)$ the fraction of urns that contain more than J balls. As usual, the definition of X^n is extended to all $t \in [0, T]$ not of the form l/n by piecewise linear interpolation. Note that for each t $X^n(t)$ is a probability vector in $\mathbb{R}^{J+2}$. Denoting $\Lambda \doteq \{0, 1, \ldots, J + 1\}$ and with an abuse of notation

$$\mathscr{P}(\Lambda) \doteq \left\{ x \in \mathbb{R}^{J+2} : x_i \geq 0, 0 \leq i \leq J + 1 \text{ and } \sum_{i=0}^{J+1} x_i = 1 \right\},$$

then for any $t \in [0, T]$, $X^n(t) \in \mathscr{P}(\Lambda)$. Thus X^n takes values in $\mathscr{C}([0, T] : \mathscr{P}(\Lambda))$. We equip $\mathscr{C}([0, T] : \mathscr{P}(\Lambda))$ with the usual supremum norm and on $\mathscr{P}(\Lambda)$ we take the $\mathscr{L}^1$-norm, which will be denoted by $\|\cdot\|_1$.

It will be convenient to work with the following dynamical representation. For $x \in \mathbb{R}^{J+2}$ and $t \in [0, -a1_{\{a<0\}} + \infty 1_{\{a>0\}})$ define the vector $\rho(t, x) \in \mathbb{R}^{J+2}$ by

$$\rho_k(t, x) = \frac{a + k}{a + t} x_k, \qquad \text{for } k = 0, 1, \ldots, J, \tag{7.1}$$

$$\rho_{J+1}(t, x) = 1 - \sum_{k=0}^{J} \frac{a + k}{a + t} x_k,$$

where, as before, when $a = \infty$ the fraction $(a + k)/(a + t)$ is taken to be 1. Then $\rho_\cdot(x, t)$ will play a role analogous to that of $\theta(\cdot|x)$ in Chap. 4 in identifying the conditional distribution of the increment of the process. Differences are that here

there is time dependence, and also that the increment is identified by (but not equal to) the k index in $\rho_k(t,x)$ [see (7.3)]. A straightforward calculation shows that if

$$x \in \mathscr{P}(\Lambda) \quad \text{and} \quad \sum_{k=0}^{J} k x_k \le t, \tag{7.2}$$

then $\rho_{J+1}(t,x) \ge 0$ and $\rho(t,x)$ is therefore a probability vector in $\mathbb{R}^{J+2}$, i.e., $\rho(t,x) \in \mathscr{P}(\Lambda)$. Also $\rho(t,x)$ is Lipschitz continuous in $(t,x) \in [0,T] \times \mathscr{P}(\Lambda)$, as long as $T < -a$ when $a \in -\mathbb{N}$. We then construct a family of independent random functions

$$\left\{ \nu_i^n(\cdot) : i = 0, 1, \ldots, \lfloor nT \rfloor - 1, \lfloor nT \rfloor \right\}$$

that take values in

$$\Lambda \doteq \{0, 1, \ldots, J+1\}$$

and with distributions

$$P\left\{ \nu_i^n(x) = k \right\} = \rho_k\left(i/n, x\right), \quad k \in \Lambda. \tag{7.3}$$

The mapping that takes an index $k \in \Lambda$ into a change in the occupancy numbers is

$$\gamma[k] = e_{k+1} - e_k, \quad 0 \le k \le K, \quad \gamma[J+1] = 0, \tag{7.4}$$

where for $j = 0, 1, \ldots, J+1$, e_j denotes the unit vector in $\mathbb{R}^{J+2}$ with 1 in the $j+1$th coordinate. Finally, we define X_l^n recursively by $X_0^n = (1, 0, \ldots, 0)' = e_0$ and

$$X_{i+1}^n = X_i^n + \frac{1}{n} \gamma[\nu_i^n\left(X_i^n\right)].$$

For the continuous time interpolation let $X^n(i/n) = X_i^n$, and for t not of the form i/n define $X^n(t)$ by piecewise linear interpolation. Observe that the conditional distribution of the increment $\left\{ \nu_i^n\left(X_i^n\right) \right\}$ is determined by $\rho\left(i/n, X_i^n\right)$. Thus the process X^n at the discrete times i/n is Markovian and will have the same distribution as the occupancy process described previously.

Define the $J+2$ by $J+2$ matrix

$$M \doteq \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

Let $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ be given with $\varphi_0(0) = 1$. Suppose there is a Borel measurable function $\theta : [0,T] \to \mathscr{P}(\Lambda)$ such that for all $t \in [0,T]$

$$\varphi(t) = \varphi(0) + \int_0^t M\theta(s)ds. \tag{7.5}$$

Note that $M\theta = \sum_{k=0}^{K} \gamma[k]\theta_k$ if $\theta \in \mathscr{P}(\Lambda)$. For $i = 0, 1, \ldots J$ (resp. $i = J+1$) we interpret $\theta_i(s)$ as the rate at which balls are thrown into urns that contain i balls (resp. greater than J balls) at time s. The rates $\theta(s)$ are unique in the sense that if another $\tilde{\theta} : [0,T] \to \mathscr{P}(\Lambda)$ satisfies (7.5) then $\tilde{\theta} = \theta$ a.e. on $[0,T]$. We call φ a **valid occupancy state process** if there exists $\theta : [0,T] \to \mathscr{P}(\Lambda)$ satisfying (7.5). In this case θ is called the **occupancy rate process** associated with φ. Using the observation that $\sum_{k=1}^{J+2}(k-1)M_{kj} = 1$ for all $j = 1, \ldots, J+1$, it is easy to check that if φ is valid then $\varphi(s)$ satisfies (7.2) with x replaced by $\varphi(s)$ and t by s, for all $s \in [0,T]$. This shows that $\rho(s, \varphi(s)) \in \mathscr{P}(\Lambda)$.

When two probability vectors θ and $\nu \in \mathscr{P}(\Lambda)$ appear in the relative entropy function, we interpret them as probability measures on $\{0, 1, \ldots, J, J+1\}$, and thus

$$R(\theta \,\|\, \nu) \doteq \sum_{k=0}^{J+1} \theta_k \log \frac{\theta_k}{\nu_k}.$$

As observed before, when φ is valid, $\rho(s, \varphi(s)) \in \mathscr{P}(\Lambda)$, so $R(\theta(s) \,\|\, \rho(s, \varphi(s)))$ is well defined for all $s \in [0,T]$. For such φ define

$$I(\varphi) \doteq \int_0^T R(\theta(s) \,\|\, \rho(s, \varphi(s)))ds. \tag{7.6}$$

If φ is not valid then define $I(\varphi) = \infty$.

As usual, representation formulas for exponential integrals will be used to prove the Laplace principle. The representation needed here is a special case of the one proved in Chap. 4, and we therefore just state the form of the representation. The controlled process $\bar{X}^n(t)$ is constructed as follows. The conditional distributions of controlled random integers $\{\bar{\nu}_i^n\}$ will be specified by a sequence $\{\bar{\mu}_i^n\}$ of controls. Each $\bar{\mu}_i^n$ is measurable with respect to the σ-algebra generated by $\{\bar{\nu}_j^n\}_{0,1,\ldots,i-1}$, and identifies the conditional distribution of $\bar{\nu}_i^n$. The controlled process is determined for t of the form j/n by $\bar{X}_0^n = e_0$ and

$$\bar{X}_{i+1}^n = \bar{X}_i^n + \frac{1}{n}\gamma[\bar{\nu}_i^n] \qquad \text{for } i = 0, 1, \ldots, \lfloor nT \rfloor,$$

with γ as in (7.4). The random quantities $\bar{X}_i^n$ and $\bar{\nu}_i^n$ are defined recursively in the order

$$\bar{X}_0^n, \bar{\nu}_0^n, \bar{X}_1^n, \bar{\nu}_1^n, \bar{X}_2^n, \ldots, \bar{X}_{\lfloor nT \rfloor + 1}^n,$$

We set $\bar{X}^n(i/n) = \bar{X}_i^n$ and use piecewise linear interpolation elsewhere.

Define

$$r_i^n(\{k\}) \doteq \rho_k\left(i/n, \bar{X}_i^n\right),$$

where $\rho\,(t, x)$ is given in (7.1).

Remark 7.1 As noted previously, there is an abuse of notation, in that we sometimes think of r_i^n as the probability vector with components $r_i^n(k)$ but at other times as the probability measure with values $r_i^n(\{k\})$. To reinforce the fact that certain probability measures are on the discrete set Λ, we write such measures with the differential dk. Also, note that k will appear both as a subscript, as in $\rho_k\left(i/n, \bar{X}_i^n\right)$, and as an argument, as in $r_i^n(\{k\})$.

Let L^n, $\bar{L}^n$, $\bar{\mu}^n$ and λ^n be measures on $\{0, 1, \ldots, J+1\} \times [0, T]$ defined as in Construction 4.4, except that λ^n uses $\rho.\left(i/n, \bar{X}_i^n\right)$ in place of $\theta(\cdot|\bar{X}_i^n)$, and the measures are of mass T and defined on subsets of $[0, T]$ rather than $[0, 1]$ in the second marginal. Specifically, for $A \subset \{0, 1, \ldots, J+1\}$ and $B \in \mathscr{B}([0, T])$,

$$L^n(A \times B) \doteq \int_B L^n(A|t)dt, \quad \bar{L}^n(A \times B) \doteq \int_B \bar{L}^n(A|t)dt, \tag{7.7}$$

$$\bar{\mu}^n(A \times B) \doteq \int_B \bar{\mu}^n(A|t)dt, \quad \lambda^n(A \times B) \doteq \int_B \lambda^n(A|t)dt, \tag{7.8}$$

where for $t \in [i/n, i/n + 1/n)$, $i = 0, \ldots, \lfloor nT \rfloor$

$$\begin{array}{ll} L^n(A|t) \doteq \delta_{\nu_i^n(X_i^n)}(A), & \bar{L}^n(A|t) \doteq \delta_{\bar{\nu}_i^n}(A), \\ \bar{\mu}^n(A|t) \doteq \bar{\mu}_i^n(A), & \lambda^n(A|t) \doteq r_i^n(A) \end{array} \tag{7.9}$$

The random measures L^n, $\bar{L}^n$, $\bar{\mu}^n$ and λ^n take values in the collection of nonnegative measures on $\mathscr{P}(\Lambda) \times [0, T]$ of total mass T. The topology used is the weak topology, where these measures are renormalized to have mass one, i.e., probability measures, and since $\mathscr{P}(\Lambda) \times [0, T]$ is compact this space is compact as well. If G is any bounded measurable function the space to $\mathbb{R}$, then

$$-\frac{1}{n}\log E\exp\left[-nG(L^n)\right] = \inf_{\{\bar{\mu}_i^n\}} E\left[G\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=0}^{\lfloor nT \rfloor} R\left(\bar{\mu}_i^n \,\|\, r_i^n\right)\right], \tag{7.10}$$

where the infimum is over all the admissible control sequences $\left\{\bar{\mu}_i^n\right\}$. Since

$$X^n(t) = e_0 + \int_0^t \gamma[k]L^n(dk \times ds), \tag{7.11}$$

this also gives a representation for functions of X^n: for any bounded and continuous $F : \mathscr{C}\left([0, T] : \mathscr{P}(\Lambda)\right) \to \mathbb{R}$,

$$-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]=\inf_{\{\bar{\mu}_i^n\}}E\left[F(\bar{X}^n)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]. \tag{7.12}$$

Here we have used the fact that (7.11) defines a measurable map that takes L^n to X^n, and let $\bar{X}^n$ be the image of $\bar{L}^n$ under that map.

A convention for the case $a \in -\mathbb{N}$. When $a \in -\mathbb{N}$ it is only possible to throw balls into the categories $0, 1, \ldots, -a-1$, and the only possible categories are $0, 1, \ldots, -a$. Thus if there are n urns there can at most be $-an$ balls thrown, and therefore $T \le -a$. When $T = -a$ all the urns have exactly $-a$ balls, which is not an interesting case to study. As a consequence, throughout this chapter we assume $T < -a$. Also, as was noted previously, because of the restriction on the possible categories we (without loss) assume that $J = -a-1$. Thus throughout this section for $a < 0$ we assume

$$T < -a, \quad J = -a-1. \tag{7.13}$$

7.2.2 Laplace Upper Bound

In this section, we prove the variational lower bound

$$\liminf_{n\to\infty}-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]\ge\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right],$$

which corresponds to the Laplace upper bound. By (7.12) it is enough to show that

$$\liminf_{n\to\infty}\inf_{\{\bar{\mu}_i^n\}}E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\ge\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

The upper bound is actually covered by the analysis of Chap. 4, since the occupancy model satisfies Condition 4.3 if one appends time as a state variable. However, for completeness we include the (short) proof here.

Recall the definitions in (7.8) and (7.9). Note that because relative entropy is nonnegative and $(\lfloor nT\rfloor+1)/n \ge T$,

$$\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\ge\int_0^T R\left(\bar{\mu}^n(\cdot|t)\,\|\lambda^n(\cdot|t)\right)dt. \tag{7.14}$$

As usual, we will need to understand conditions for tightness, and how the weak limits of $\bar{L}^n$, $\bar{\mu}^n$, λ^n and $\bar{X}^n$ are all related. As noted previously, tightness of the first three is automatic since they take values in a compact space. In addition, the process $\bar{X}^n$ takes values in a space of continuous trajectories that start at e_0 and which are Lipschitz continuous with the Lipschitz constant bounded by

$$\left\| \sup_{0\le t\le T} \int \gamma[k]\bar{L}^n(dk|t) \right\|_1 \le \sup_{k\in\Lambda} \|\gamma[k]\|_1 \le 2,$$

where $\|\cdot\|_1$ is the $\mathscr{L}^1$-norm on $\mathbb{R}^{J+2}$. Since the space of all such trajectories is also compact, $\{\bar{X}^n\}$ is also automatically tight. The relations between the limits can be determined using the same argument as in Lemma 4.12, save that $\rho_\cdot(t,x)$, which plays the role of $\theta(\cdot|x)$ in Chap. 4, is here time dependent, and whereas the dynamics of Chap. 4 take the form (4.1), if we consider $\rho_\cdot(t,x)$ as determining the noises that drive the system then these noises enter the system only after passing through $\gamma[\cdot]$ as in (7.11).

Rewritten for these differences, the analogue of Lemma 4.12 is as follows.

Lemma 7.2 *Consider the sequence $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)\}_{n\in\mathbb{N}}$ as defined in (7.7) and (7.8), and with $\bar{X}^n$ as in (7.11) with L^n replaced by $\bar{L}^n$. Let $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)\}$ denote a weakly converging subsequence, which for notational convenience we again label by n, with limit $(\bar{X}, \bar{L}, \bar{\mu}, \lambda)$. Then w.p.1 $\bar{L} = \bar{\mu}$, and $\bar{\mu}(dk \times dt)$ can be decomposed as $\bar{\mu}(dk|t)dt$, where $\bar{\mu}(dk|t)$ is a stochastic kernel on $\{0, 1, \ldots, J, J+1\}$ given $[0, T]$, and w.p.1 for all $t \in [0, T]$,*

$$\begin{aligned} \bar{X}(t) &= e_0 + \int_{\mathbb{R}^d\times[0,t]} \gamma[k]\bar{\mu}(dk \times ds) \\ &= e_0 + \int_{\mathbb{R}^d\times[0,t]} \gamma[k]\bar{\mu}(dk|s)ds. \end{aligned} \tag{7.15}$$

In addition, λ and $\bar{X}$ are related through

$$\lambda(\{k\} \times B) = \int_B \rho_k(t, \bar{X}(t))dt, \ k \in \{0, 1, \ldots, J+1\}, \ B \in \mathscr{B}([0, T]). \tag{7.16}$$

Theorem 7.3 *Define I by (7.6) for any of the occupancy models described in Sect. 7.2.1. If $F : \mathscr{C}([0, T] : \mathscr{P}(\Lambda)) \to \mathbb{R}$ is bounded and continuous, then*

$$\liminf_{n\to\infty} -\frac{1}{n} \log E \exp\left[-nF(X^n)\right] \ge \inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))} [I(\varphi) + F(\varphi)].$$

Proof Owing to the representation formula (7.10) it is enough to show that

$$\liminf_{n\to\infty} \inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{X}^n\right) + \frac{1}{n} \sum_{i=0}^{\lfloor nT\rfloor} R\left(\bar{\mu}_i^n \,\middle\|\, r_i^n\right) \right] \ge \inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))} [I(\varphi) + F(\varphi)]. \tag{7.17}$$

Consider any admissible sequence $\left\{\bar{\mu}_i^n\right\}$. Then (7.14) implies

$$E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\geq E\left[F\left(\bar{X}^n\right)+\int_0^T R\left(\bar{\mu}^n(\cdot\,|t)\,\|\lambda^n(\cdot|t)\right)dt\right]$$
$$=E\left[F\left(\bar{X}^n\right)+TR\left(\bar{\mu}^n/T\,\|\lambda^n/T\right)\right].$$

Applying Fatou's lemma and using the lower semicontinuity of relative entropy,

$$\begin{aligned}&\liminf_{n\to\infty}E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\\&\quad\geq\liminf_{n\to\infty}E\left[F\left(\bar{X}^n\right)+TR\left(\bar{\mu}^n/T\,\|\lambda^n/T\right)\right]\\&\quad\geq E\left[F\left(\bar{X}\right)+TR\left(\bar{\mu}/T\,\|\lambda/T\right)\right]\qquad(7.18)\\&\quad=E\left[F\left(\bar{X}\right)+\int_0^T R\left(\bar{\mu}(\cdot\,|t)\,\|\rho_\cdot(t,\bar{X}(t))\right)dt\right].\end{aligned}$$

If $\theta(s)=\sum_{k=0}^{J+1}e_k\bar{\mu}(\{k\}\,|s)$, then using $M\theta(s)=\sum_{k=0}^{J}\gamma[k]\theta_k(s)$ we see from Lemma 7.2 that $\bar{X}(t)=e_0+\int_0^t M\theta(s)ds$. Therefore by the definition (7.6) of the rate function $I(\varphi)$,

$$\int_0^T R\left(\bar{\mu}(\cdot\,|t)\,\|\rho_\cdot(t,\bar{X}(t))\right)dt=I(\bar{X}).$$

Thus (7.18) yields (7.17), and completes the proof of the Laplace upper bound. □

7.2.3 *Properties of the Rate Function*

In this section we prove important properties of the rate function, some of which will be used later on to prove the Laplace lower bound.

Theorem 7.4 *Let I be defined as in (7.6). Then for any $K\in[0,\infty)$ the level set $\{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda)):I(\varphi)\leq K\}$ is compact.*

Proof By adding time as a state variable we see that the occupancy model satisfies Condition 4.3 of Chap. 4. Thus the conclusion follows from Theorem 4.13.

Theorem 7.5 (ZERO COST TRAJECTORY) *For $t\in[0,T]$ let $f(t)\doteq\left(1+\frac{t}{a}\right)^{-a}$ when $a<\infty$ and $f(t)\doteq e^{-t}$ in the case $a=\infty$. Define*

$$\bar{\phi}_i(t)\doteq\frac{(-t)^i}{i!}f^{(i)}(t)\qquad\text{for } 0\leq i\leq J,$$

where $f^{(i)}(t)$ is the ith derivative of $f(t)$, and let $\bar{\phi}_{J+1}(t)\doteq 1-\sum_{i=0}^{J}\bar{\phi}_i(t)$. Then $I(\bar{\phi})=0$.

Proof We first assume $a \neq \infty$. It is easy to check that for any $0 \leq i < \infty$,

$$\frac{(-t)^i}{i!} f^{(i)}(t) \geq 0 \quad \text{and} \quad \sum_{i=0}^{\infty} \frac{(-t)^i}{i!} f^{(i)}(t) = 1. \tag{7.19}$$

Thus $\bar{\phi}$ as defined in the statement of the theorem is indeed a probability vector. It is also a continuously differentiable function and satisfies $\sum_{k=0}^{J} \bar{\phi}_k(t) \leq t$ for all $t \in [0, T]$. We will show that

$$\frac{d}{dt}\bar{\phi}(t) = M\rho(t, \bar{\phi}(t)). \tag{7.20}$$

If so, then the occupancy rate process $\bar{\theta}$ associated to $\bar{\phi}$ is $\rho(t, \bar{\phi}(t))$, and thus by the definition of rate function

$$I(\bar{\phi}) = \int_0^T R\left(\bar{\theta}(t) \,\|\, \rho(t, \bar{\phi}(t))\right) dt = 0.$$

To show (7.20) we calculate $\bar{\phi}_i(t) = \frac{(-t)^i}{i!} f^{(i)}(t)$ for $0 \leq i \leq J$ explicitly:

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i}.$$

Hence the derivative satisfies

$$\begin{aligned} \frac{d}{dt}\bar{\phi}_i(t) &= \frac{a+i-1}{a+t}\bar{\phi}_{i-1}(t) - \frac{a+i}{a+t}\bar{\phi}_i(t) \\ &= \rho_{i-1}(t, \bar{\phi}(t)) - \rho_i(t, \bar{\phi}(t)) = (M\rho(t, \bar{\phi}(t))_i, \end{aligned}$$

where ρ_{-1} is taken to be 0 and the second equality is due to the definition of $\rho(t, \bar{\phi}(t))$ in (7.1). The case of $\phi_{J+1}(t)$ follows on observing that

$$\frac{d}{dt}\bar{\phi}_{J+1}(t) = -\sum_{i=0}^{J} \frac{d}{dt}\bar{\phi}_i(t) = -\sum_{i=0}^{J} (M\rho(t, \bar{\phi}(t))_i = (M\rho(t, \bar{\phi}(t))_{J+1},$$

where the last identity is a consequence of the fact that $1^T M = 0$.

Next we consider the case when $a = \infty$. In this case $f(t) = e^{-t}$, and (7.20) is immediate on observing that $\bar{\phi}_i(t) = t^i e^{-t}/i!$ and so $\frac{d}{dt}\bar{\phi}_i(t) = \bar{\phi}_{i-1}(t) - \bar{\phi}_i(t)$. □

Lemma 7.6 *Let $\bar{\phi}$ be the zero-cost trajectory from Theorem 7.5. For every choice of the parameter a there exist $\delta > 0$ and $0 < K < \infty$ so that*

$$\bar{\phi}_i(t) \geq \delta t^K \tag{7.21}$$

for any $0 \leq i \leq J + 1$ and $t \in [0, T]$.

Proof Note that when $a > 0$, $0 \le i \le J$ and $0 \le t \le T$,

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i} \ge \frac{t^i}{J!} \left(1 + \frac{T}{a}\right)^{-a-J},$$

and because of (7.19) we have

$$\bar{\phi}_{J+1}(t) = 1 - \sum_{i=0}^{J} \bar{\phi}_i(t) \ge \frac{(-t)^{J+1}}{(J+1)!} f^{(J+1)}(t) \ge \frac{t^{J+1}}{(J+1)!} \left(1 + \frac{T}{a}\right)^{-a-J-1}.$$

Thus in this case, with $\bar{\delta} = \frac{1}{(J+1)!} \left(1 + \frac{T}{a}\right)^{-a-J-1}$,

$$\bar{\phi}_i(t) \ge \bar{\delta} t^i, \text{ for all } i = 0, 1, \ldots, J+1, \ t \in [0, T]. \tag{7.22}$$

For the case $a < 0$, by (7.13) $T < -a$ and $a = -J - 1$. If $0 \le i \le J$, then since $a + j \le -1$ for each $0 \le j \le J$,

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i} \ge \frac{t^i}{J!} \frac{1}{(-a)^i} \left(1 + \frac{t}{a}\right)^{-a-i}.$$

Moreover since $a < 0$, $t/a \in (-1, 0)$ for $t \in [0, T]$, and $-(a+i) \ge 1$, for each $i \le J$, $\left(1 + \frac{t}{a}\right)^{-a-i}$ is monotone decreasing in $t \in [0, T]$. Therefore

$$\bar{\phi}_i(t) \ge \frac{t^i}{J!} \left(-\frac{1}{a}\right)^i \left(1 + \frac{T}{a}\right)^{-a-i}.$$

For $\bar{\phi}_{J+1}(t)$ we have

$$\begin{aligned}
\bar{\phi}_{J+1}(t) &= 1 - \sum_{i=0}^{J} \bar{\phi}_i(t) \\
&\ge \frac{(-t)^{J+1}}{(J+1)!} f^{(J+1)}(t) \\
&= \frac{t^{J+1}}{(J+1)!} \frac{\prod_{j=0}^{J}(a+j)}{a^{J+1}} \\
&\ge \frac{t^{J+1}}{(J+1)!} \left(-\frac{1}{a}\right)^{J+1}.
\end{aligned}$$

Thus in this case (7.22) holds with $\bar{\delta} = (-a)^{-J-1}/(J+1)!$.

Finally, for the case $a = \infty$, using the fact that $\bar{\phi}_i(t) = t^i e^{-t}/i!$ for $i \le J$ and $\bar{\phi}_{J+1}(t) \ge t^{J+1}e^{-t}/(J+1)!$, we have that (7.22) holds with $\bar{\delta} = e^{-T}/(J+1)!$. The result now follows on taking $K = J+1$ and $\delta = \bar{\delta}(T^{-J-1} \wedge 1)$. □

For $f : [0,T] \to \mathbb{R}^{J+2}$, let $\|f\|_{\infty,T} \doteq \sup_{0\le t\le T} \|f(t)\|_1$, where $\|\cdot\|_1$ as before is the $\mathscr{L}^1$ norm on $\mathbb{R}^{J+2}$.

Lemma 7.7 *For a given value of a let the parameters δ and K be as in (7.21). Let $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ satisfy $I(\varphi) < \infty$. Then for any $\varepsilon > 0$ there exists $\varphi^\varepsilon \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ such that*

(a) $I(\varphi^\varepsilon) \le I(\varphi)$,
(b) $\|\varphi - \varphi^\varepsilon\|_{\infty,T} \le \varepsilon$,
(c) $\varphi_i^\varepsilon(t) \ge \varepsilon\delta t^K$ *for all* $t \in [0,T]$ *and* $i = 0, 1, \ldots J, J+1$.

Proof For any $\varepsilon > 0$ and $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$, let

$$\varphi^\varepsilon = (1-\varepsilon)\,\varphi + \varepsilon\bar{\varphi},$$

where $\bar{\varphi}$ is the zero cost trajectory from Theorem 7.5. Then $\varphi^\varepsilon \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$. From the definition of $\rho(t,x)$ in (7.1) it follows that $\rho(t,x)$ has the following linearity property in x. Suppose we are given $t \in [0,T]$ and $x, \tilde{x} \in \mathscr{P}(\Lambda)$ that satisfy (7.2). Then for any $\alpha \in [0,1]$, $\alpha x + (1-\alpha)\tilde{x}$ satisfies (7.2) and

$$\alpha\rho(t,x) + (1-\alpha)\rho(t,\tilde{x}) = \rho(t, \alpha x + (1-\alpha)\tilde{x}).$$

Hence recalling the definition of $I(\varphi)$ in (7.6) and the joint convexity of relative entropy, we find that $I(\varphi)$ is convex in φ. Therefore

$$I(\varphi^\varepsilon) \le (1-\varepsilon)\,I(\varphi) + \varepsilon I(\bar{\varphi}) = (1-\varepsilon)\,I(\varphi) \le I(\varphi).$$

Since $\|\varphi - \bar{\varphi}\|_{\infty,T} \le 2$

$$\|\varphi - \varphi^\varepsilon\|_{\infty,T} \le \varepsilon\,\|\varphi - \bar{\varphi}\|_{\infty,T} \le 2\varepsilon,$$

and also from Lemma 7.6, $\varphi_i^\varepsilon(t) \ge \varepsilon\bar{\varphi}_i(t) \ge \varepsilon\delta t^K$. □

The final theorem of this section is useful in proving the Laplace lower bound.

Definition 7.8 We call an occupancy path $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ a **good path** if $\varphi(0) = e_0$ and there exist constants $0 < \delta', K' < \infty$ so that $\varphi_i(t) \ge \delta' t^{K'}$ for $t \in [0,T]$ and $0 \le i \le J+1$.

Definition 7.9 We call an occupancy rate control $\theta : [0,T] \to \mathscr{P}(\Lambda)$ a **good control** if (i) there exist a finite number of intervals $[r_i, s_i]$, $1 \le i \le m$ so that $[0,T] = \cup_{i=1}^m [r_i, s_i]$, and $\theta(t)$ is a constant vector on each (r_i, s_i), (ii) there exists $0 < \sigma < T$ so that θ is "pure" on $[0,\sigma)$, in the sense that for any interval of constancy $(r,s) \subset [0,\sigma)$, there exists i, $0 \le i \le J+1$ such that $\theta_i(t) = 1$ for $t \in (r,s)$.

Theorem 7.10 *For a good path $\varphi \in \mathscr{C}([0,T]:\mathscr{P}(\Lambda))$ assume $I(\varphi) < \infty$. Let δ', K' be the associated constants in the definition of a good path. For any $\varepsilon > 0$ there exists a good control θ^* and associated $\sigma > 0$ so that if φ^* is the occupancy path associated to θ^*, namely (7.5) holds with (φ,θ) replaced by (φ^*,θ^*), then there is $\delta'' \in (0,\infty)$ such that*

(a) $I(\varphi^*) \le I(\varphi) + \varepsilon$,
(b) $\|\varphi^* - \varphi\|_{\infty,T} \le \varepsilon$,
(c) *if $t < \sigma$ and $\theta_i^*(t) = 1$ then $\varphi_i^*(t) \ge \delta'' \sigma^{K'}$.*

Proof For a $\sigma \in (0,T)$ that will be specified later on, we construct a pure control $\theta^*(t)$, $t \in [0,\sigma)$ as follows. For $0 \le i \le J$ let $\theta_i^*(t) = 1$ if

$$\sum_{j=0}^{i} j\varphi_j(\sigma) + i\sum_{k=i+1}^{J+1}\varphi_k(\sigma) \le t < \sum_{j=0}^{i} j\varphi_j(\sigma) + (i+1)\sum_{k=i+1}^{J+1}\varphi_k(\sigma),$$

and let $\theta_{J+1}^*(t) = 1$ if

$$\sum_{j=0}^{J} j\varphi_j(\sigma) + (J+1)\varphi_{J+1}(\sigma) \le t < \sigma. \tag{7.23}$$

Observe that the component φ_i^* for $i > 0$ will increase only during the interval when $\theta_{i-1}^*(t) = 1$, and that it decreases to its final value while $\theta_i^*(t) = 1$. Observe also that $\varphi^*(\sigma) = \varphi(\sigma)$. Hence for $t < \sigma$, if $\theta_i^*(t) = 1$ then $\varphi_i^*(t) \ge \varphi_i^*(\sigma) \ge \delta'\sigma^{K'}$.

Now assume that $0 < a < \infty$. For i and t such that $t < \sigma$ and $\theta_i^*(t) = 1$,

$$\rho_i(t,\varphi^*(t)) = \frac{a+i}{a+t}\varphi_i^*(t) \ge \frac{a}{a+T}\delta'\sigma^{K'} = \delta''\sigma^{K'}, \tag{7.24}$$

where $\delta'' \doteq \frac{a}{a+T}\delta'$.

Recall that when $a < 0$ we assume without loss that $J = -a-1$, and that no balls are placed in urns that currently contain more than J balls. Thus $\rho_{J+1}(t,\phi(t)) = 0$ and consequently $\theta_{J+1}(t) = 0$ for all t. From (7.5) and recalling that $\sum_{j=0}^{J+1} jM_{(j+1),i} = 1$ for all $i = 1,\dots,J+1$, it follows that

$$\sum_{j=0}^{J+1} j\varphi_j(\sigma) = \sigma.$$

It then follows from (7.23) that $\theta_{J+1}^*(t) = 0$ for all $t \in [0,\sigma]$. For $0 \le i \le J$, we have, when $t < \sigma$ and $\theta_i^*(t) = 1$, that

$$\rho_i(t,\varphi^*(t)) \ge \frac{a+i}{a+t}\delta'\sigma^{K'} \ge \frac{a+J}{a+t}\delta'\sigma^{K'} \ge -\frac{1}{a}\delta'\sigma^{K'}.$$

Thus for $0 \le i \le J$, with $\delta'' = -a^{-1}\delta'$, $\rho_i(t, \varphi^*(t)) \ge \delta'' \sigma^{K'}$ when $\theta_i^*(t) = 1$ and $t \in [0, \sigma]$.

Finally, when $a = \infty$ we can choose $\delta'' = \delta'$ and (7.24) will hold. Thus in all cases (7.24) holds with some $\delta'' > 0$, that is independent of the choice of σ.

This completes the construction of θ^* and φ^* on $[0, \sigma)$. The lower bounds on the ρ_i and the fact that θ^* is pure on $[0, \sigma]$ imply

$$\int_0^\sigma R\left(\theta^*(t) \,\|\, \rho(t, \varphi^*(t))\right) dt \le -\sigma \log\left(\delta'' \sigma^{K'}\right).$$

Now choose $\sigma > 0$ small enough so that

$$-\sigma \log\left(\delta'' \sigma^{K'}\right) \le \varepsilon/2 \text{ and } \sup_{t \in [0,\sigma)} \left\|\varphi^*(t) - \varphi(t)\right\|_1 \le \varepsilon.$$

Note that the latter property can be satisfied by choosing σ sufficiently small since $\varphi(0) = \varphi^*(0) = e_0$ implies $\sup_{t\in[0,\sigma]} \|\varphi(t) - \varphi^*(t)\|_1 \le (J+1)|\varphi_0(0) - \varphi_0(\sigma)|$. Also, recall that under the construction $\varphi^*(\sigma) = \varphi(\sigma)$.

The construction of controls on $[\sigma, T]$ is easier. Let $\theta(t)$ be the rate process associated with $\varphi(t)$ by (7.5). For $N \in \mathbb{N}$ we partition $[\sigma, T]$ into N subintervals of length $c_N = (T - \sigma)/N$. For each s that $\sigma + l c_N \le s \le \sigma + (l+1) c_N$ where $0 \le l \le (N-1)$, let

$$\theta^{(N)}(s) = \frac{\int_{\sigma + l c_N}^{\sigma + (l+1) c_N} \theta(t) dt}{c_N}.$$

Let $\varphi^{(N)}$ be the occupancy path associated with $\theta^{(N)}$ over the interval $[\sigma, T]$, i.e.

$$\varphi^{(N)}(t) = \varphi^{(N)}(\sigma) + \int_\sigma^t M\theta^{(N)}(s) ds, \; t \in [\sigma, T], \; \varphi^{(N)}(\sigma) = \varphi(\sigma).$$

Then it is easy to check that $\varphi^{(N)}(t)$ coincides with $\varphi(t)$ on the "partition points" in $[\sigma, T]$, i.e., those points of the form $\{\sigma + l c_N : 0 \le l \le (N-1)\}$. Thus, since $\|\theta(t)\|_1 = 1$, for N large enough [e.g., $N > (T - \sigma)/\varepsilon$], $\sup_{t\in[\sigma,T]} \|\varphi^{(N)}(t) - \varphi(t)\|_1 \le \varepsilon$.

Because $\varphi(t)$ is good, when $t > \sigma$, we have $\varphi_i(t) \ge \delta' t^{K'} \ge \delta' \sigma^{K'} > 0$ for all $0 \le i \le J+1$. Therefore $\varphi(t)$ is uniformly bounded away from the boundary after time σ, and thus for sufficiently large N, so is $\varphi^{(N)}(t)$. This in particular says that for such N, $t \in [\sigma, T]$, $\rho_j(t, \varphi^{(N)}(t))$ is uniformly bounded away from 0 for $j = 0, \ldots, J+1$ when $a > 0$ and for $j = 0, \ldots, J$ when $a < 0$. In the latter case, both $\rho_{J+1}(t, \varphi^{(N)}(t))$ and $\theta_{J+1}^{(N)}(t)$ are identically 0.

As $N \to \infty$, $\theta^{(N)}(t)$ converges to $\theta(t)$ and $\varphi^{(N)}(t)$ converges to $\varphi(t)$ for a.e. $t \in [0, T]$. Using that ρ is bounded away from zero and $\theta^{(N)}$ is bounded above, by the dominated convergence theorem

$$\lim_{N\to\infty}\int_\sigma^T R(\theta^{(N)}(t)\,\|\rho(t,\varphi^{(N)}(t)))dt=\int_\sigma^T R\left(\theta(t)\,\|\rho(t,\varphi(t))\right)dt.$$

Now choose $N<\infty$ large enough so that the integrals differ by less than $\varepsilon/2$. Let θ^* be defined as it was previously on $[0,\sigma]$, and set it equal to $\theta^{(N)}$ on $[\sigma,T]$. Let φ^* denote the corresponding occupancy path over $[0,T]$. Then

$$\begin{aligned}I(\varphi^*)&=\int_\sigma^T R(\theta^{(N)}(t)\,\|\rho(t,\varphi^{(N)}(t)))dt+\int_0^\sigma R\left(\theta^*(t)\,\|\rho(t,\varphi^*(t))\right)dt\\&\le\int_\sigma^T R\left(\theta(t)\,\|\rho(t,\varphi(t))\right)dt+\varepsilon/2+\varepsilon/2\\&\le I(\varphi)+\varepsilon.\end{aligned}$$

This completes the proof. □

7.2.4 *Laplace Lower Bound*

Theorem 7.11 *Define I by (7.6) for any of the occupancy models described in Sect. 7.2.1. If $F:\mathscr{C}([0,T]:\mathscr{P}(\Lambda))\to\mathbb{R}$ is bounded and continuous, then*

$$\limsup_{n\to\infty}-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]\le\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

Proof According to (7.12), the theorem follows if

$$\limsup_{n\to\infty}\inf_{\{\bar\mu_i^n\}}\bar E\left[F\left(\bar X^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar\mu_i^n\,\|r_i^n\right)\right]\le\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

As was the case with Chap. 4, the main difficulty in the proof of the lower bound is that controls and controlled processes should be constructed so that the dominated convergence theorem can be used. Since vanishing transition probabilities can make relative entropy costs diverge some care is required, but the constructions of the last section will very carefully control the rates at which balls are put into urns of category i when r_i^n is small.

For any $\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))$ such that $I(\varphi)<\infty$, Lemma 7.7 and Theorem 7.10 imply that for any $\varepsilon>0$ there exists (φ^*,θ^*) with the properties described in Theorem 7.10. Since F is continuous on $\mathscr{C}([0,T]:\mathscr{P}(\Lambda))$, we only need to show that there exists a sequence of admissible controls $\left\{\bar\mu_i^n\right\}$ so that

$$\limsup_{n\to\infty}\bar E\left[F\left(\bar X^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar\mu_i^n\,\|r_i^n\right)\right]\le I(\varphi^*)+F(\varphi^*).$$

The latter inequality will follow if we can find a sequence of admissible $\left\{\bar{\mu}_i^n\right\}$ such that

$$\limsup_{n\to\infty} \bar{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor} R\left(\bar{\mu}_i^n \,\|\, r_i^n\right)\right] \le I(\varphi^*), \tag{7.25}$$

and such that if $\bar{X}^n$ is the occupancy process constructed under $\left\{\bar{\mu}_i^n\right\}$ then for any small $b > 0$

$$\limsup_{n\to\infty} \bar{P}\left\{\left\|\bar{X}^n - \varphi^*\right\|_{\infty,T} > b\right\} = 0. \tag{7.26}$$

In other words, $\bar{X}^n$ converges to φ^* in probability.

To prove the desired inequalities (7.25) and (7.26) we need to construct the proper $\left\{\bar{\mu}_i^n\right\}$. Recall that $\left\{\bar{\mu}_i^n\right\}$ can depend in any measurable way on the "past," and so we could, in principle, use such information in constructing the controls. However, as seen previously for certain problems of this type we can construct the controls without reference to the controlled process (i.e., "open loop" controls). Let θ^* be the good control as described in Theorem 7.10. We know that θ^* is piecewise constant and pure up to time $\sigma > 0$. From property (c) in Theorem 7.10, we also know that before time σ, if $\theta_i^*(t) = 1$ then both $\rho_i(t, \varphi^*(t))$ and $\varphi_i^*(t)$ are greater than a fixed value $\zeta > 0$ (for all $i \le J+1$ when $a > 0$ and for all $i \le J$ when $a < 0$). Using part (b) of Theorem 7.10 we can also assume for the same value of ζ that both $\rho_i(t, \varphi^*(t))$ and $\varphi_i^*(t)$ are greater than ζ for all $t \in [\sigma, T]$ (and again for all $i \le J+1$ when $a > 0$ and for all $i \le J$ when $a < 0$).

Although the limit trajectory stays away from the boundary after time σ, there is no guarantee that the random process $\bar{X}^n$ is uniformly bounded away. In order to handle this possibility, we use a stopping time argument similar to one used in [109].

Let (l_n/n) be the minimum of the first time such that for some i, $\bar{X}_i^n(l_n/n) \le \zeta/2$ and $\theta_i^*(l_n/n) > 0$, and the fixed deterministic time $\lfloor nT \rfloor / n$. This is the first time the random process is close to the boundary, with the possibility of a large contribution to the total cost [note that when $\theta_i^*(l_n/n) = 0$ there is no contribution to the cost regardless of the value of $\bar{X}_i^n(l_n/n)$]. The control $\left\{\bar{\mu}_i^n\right\}$ is then defined by

$$\bar{\mu}_i^n(\{k\}) = \begin{cases} \theta_k^*(i/n) & \text{if } i \le l_n \\ \rho_k\left(i/n, \bar{X}^n(i/n)\right) & \text{if } i > l_n. \end{cases}$$

Prior to the stopping time, we use exactly what θ^* suggests, and after the stopping time we follow the law of large number trajectory (and therefore incur no additional cost).

Now we apply Lemma 7.2. Thus given any subsequence there is convergence along a further subsequence as indicated in the theorem, with limit $\left(\bar{X}, \bar{\mu}\right)$. Using the standard argument by contradiction, it will be enough to prove (7.25) and (7.26) for this convergent processes. Let $\tau^n = (l_n/n) \le T$. Note that because the applied controls are pure, the process $\bar{X}^n(t)$ is deterministic prior to σ, and also that prior to this time, the time derivatives of $\bar{X}^n(t)$ and $\varphi^*(t)$ are piecewise constant. In fact,

the two derivatives are identical except possibly on a bounded number of intervals each of length less than $1/n$ [the points where they may disagree are all located with within distance $1/n$ of the endpoints of the intervals of constancy of $\dot{\varphi}^*(t)$]. Thus for large n we cannot have $\tau^n < \sigma$. Since the range of τ^n is the bounded set $[0, T]$, we can also assume τ^n converges in distribution to a limit τ, and without loss we assume the convergence is along the same subsequence. Since $\tau^n \geq \sigma$ for large n we have $\tau \geq \sigma$ w.p.1.

Suppose that $\tau < T$. It is easy to check that the limit control processes w.p.1 satisfies, for a.e. $t \in [0, T]$,

$$\bar{\mu}(\{k\} \,|\, t) = \begin{cases} \theta_k^*(t) & \text{if } t \leq \tau \\ \rho_k\left(t, \bar{X}(t)\right) & \text{if } t > \tau \end{cases}.$$

Owing to the definition of τ^n, $\tau < T$ implies $\bar{X}_i(\tau) \leq \zeta/2$ for some $i \in \Lambda$ (although $\varphi_i^*(t) \geq \zeta$ when $t \in [\sigma, T]$). We use that $\bar{\mu}(\{k\} \,|t) = \theta_k^*(t)$ when $t \leq \tau$ and that $\theta^*(t)$ is deterministic. As shown in Theorem 7.2, $\left(\bar{X}, \bar{\mu}\right)$ satisfies (7.15) for $t \in [0, \tau]$. Thus for $t \in [0, \tau]$, $\bar{X}(t) = \varphi^*(t)$ w.p.1. This gives a contradiction since

$$\bar{X}_i(\tau) \leq \zeta/2 < \zeta \leq \varphi_i^*(\tau).$$

Therefore $\tau = T$, and thus for all $t \in [0, T]$, $\bar{X}(t) = \varphi^*(t)$ w.p.1. This also proves that the weak limit of the random processes $\bar{X}^n$ is indeed φ^*, which implies (7.26). To prove (7.25), we use the weak convergence, the continuity of the map $(x, y) \mapsto x \log(x/y)$ on $[0, \infty) \times (0, \infty)$ and the dominated convergence theorem to obtain

$$\limsup_{n\to\infty} \bar{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor nT \rfloor} R\left(\bar{\mu}_i^n \,\|r_i^n\right)\right] = \int_0^T R\left(\theta^*(t)||\rho(t, \varphi^*(t))\right) dt = I(\varphi^*).$$

This completes the proof. □

7.2.5 *Solution to Calculus of Variations Problems*

In the previous sections we identified the process level large deviation rate function (7.6) for a class of occupancy problems. The large deviation principle for the process at a given fixed time can then be expressed in terms of the solution to a calculus of variations problem. In most cases this calculus of variations problem will not have a closed form solution. However, for the class of occupancy models studied here it can be identified with the solution to a related finite dimensional minimization problem. This latter problem can be solved by the standard Lagrange multiplier method, which is easily implemented numerically. In this section we give the precise statement of this equivalence. We mention two results. The first gives the minimum of the rate function subject to a terminal constraint, and the second gives the minimum of the

sum of the rate function plus a cost that is affine in the terminal location. The explicit formulas generalize ones obtained in [109] for the special case of MB statistics. The techniques used are quite different, based as they are on dynamic programming and control theory rather than methods from the calculus of variations. When combined with methods for accelerated Monte Carlo as discuss in later chapters, these explicit solutions allow one to obtain not just large deviation approximations but also accurate approximations to nonasymptotic quantities. Proofs are not given, but interested readers can find the details in [266].

7.2.5.1 Problem Formulation

Suppose the current occupancy state is $x \in \mathscr{P}(\Lambda)$ and that t is the number of balls per urn among all categories. If $y_i, i = 0, 1, \ldots, J, J+1, \ldots$ are the fraction in category i, then $x_i = y_i$ for $i \leq J$ and

$$t = \sum_{k=0}^{\infty} k y_k.$$

Note that $t \geq \sum_{k=0}^{J+1} k x_k$.

In previous sections we considered the large deviation analysis for just the case of the initial condition where all urns are empty. To use dynamic programming, one must introduce the analogue of the rate function that is suitable for general initial times and states. The set of possible states for a given t [i.e., ones that can be reached starting from $(1, 0, \ldots, 0)$ at $t = 0$] depends on both t and a, which leads to the following definition.

Definition 7.12 Define $\mathscr{D}_a$, the **feasible domain** for the occupancy model with parameter a, as follows:

- when $a > 0$,

$$\mathscr{D}_a \doteq \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : x_{J+1} > 0 \text{ and } t \geq \sum_{i=0}^{J+1} i x_i \right\} \bigcup \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : x_{J+1} = 0 \text{ and } t = \sum_{i=0}^{J} i x_i \right\};$$

- and when $a < 0$ and $J = -a - 1$,

$$\mathscr{D}_a \doteq \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : t = \sum_{i=0}^{J+1} i x_i \right\}.$$

As before, when $a < 0$ we restrict to $T < -a$. In the first case the second set in the union reflects the fact that when $x_{J+1} = 0$ the number of balls thrown is exactly $\sum_{i=0}^{J} i x_i$, and similarly for the second case.

Consider a valid occupancy process $\varphi \in \mathscr{C}([t, T] : \mathscr{P}(\Lambda))$ with $\varphi(t) = x$ and $(x, t) \in \mathscr{D}_a$. Making the dependence on (x, t) explicit, the rate function $I(x, t; \varphi)$ for such paths can be written

$$I(x, t; \varphi) \doteq \int_t^T R(\theta(s) \| \rho(s, \varphi(s))) \, ds,$$

where

$$\varphi(s) = \varphi(t) + \int_t^s M\theta(r) dr$$

and

$$\rho_k(s, y) \doteq \frac{a + k}{a + s} y_k, \qquad k = 0, 1, \ldots, J, \qquad \rho_{J+1}(s, y) \doteq 1 - \sum_{k=0}^{J} \rho_k(s, y).$$

The relevant calculus of variations problem for a point in the feasible domain is

$$O(x, t; \omega) \doteq \inf_{\substack{\varphi \in \mathscr{C}([t,T]:\mathscr{P}(\Lambda)) \\ \varphi(t)=x, \varphi(T)=\omega}} I(x, t; \varphi). \tag{7.27}$$

The formula for the finite dimensional minimization problem requires some notation. For all $a \in \mathbb{R}$, $a \neq 0$ and $i \in \mathbb{N}$, let

$$\binom{a}{i} \doteq \frac{\prod_{j=0}^{i-1}(a - j)}{i!}$$

and $\binom{a}{0} = 1$. Note that if $a \in \mathbb{N}$ and $i > a$ then $\binom{a}{i} = 0$, and that if $a \notin \mathbb{N} \cup \{0\}$, then $\binom{a}{i} \neq 0$. We will use the fact that if $a \in \mathbb{R}$ and $|z| < 1$ then the binomial expansion

$$(1 + z)^{-a} = \sum_{i=0}^{\infty} \binom{-a}{i} z^i$$

is valid, and if $-a \in \mathbb{N}$ then the sum contains only a finite number of nonzero terms and is valid for all $z \in \mathbb{R}$.

For $i \in \mathbb{N} \cup \{0\}$ and $a > 0$, $s \geq 0$ or $a \in -\mathbb{N}$, $0 \leq s \leq -a$, define

$$Q_i^a(s) \doteq \left(-\frac{s}{a}\right)^i \binom{-a}{i} \left(1 + \frac{s}{a}\right)^{-a-i}.$$

When $a = 0$ we use the limiting values

$$Q_0^0(s) = 1, \;\; Q_i^0(s) = 0$$

for all $i \in \mathbb{N}$ and $s \geq 0$. One can check that $\{Q_i^a(s)\}_{i=0}^{\infty}$ is a probability vector for any choice of (a, s) as above.

Denote $\pi^k = \{\pi_0^k, \pi_1^k, \dots\}$ for all $0 \leq k \leq J+1$, where π_i^k represents the probability of throwing i additional balls into the kth category. Denote $\pi = (\pi^0, \pi^1, \dots, \pi^{J+1})$. For any given $x \in \mathscr{P}(\Lambda)$, we say $\pi = (\pi^0, \pi^1, \dots, \pi^J, \pi^{J+1}) \in \mathscr{F}(x, t; \omega, T)$ if

$$\sum_{j=0}^{\infty} \pi_j^k = 1, \quad 0 \leq k \leq J+1, \quad \sum_{k=0}^{J+1} x_k \sum_{j=0}^{\infty} j\pi_j^k = T - t, \tag{7.28}$$

and

$$\omega_i = \sum_{k=0}^{i} x_k \pi_{i-k}^k, \quad 0 \leq i \leq J, \qquad \omega_{J+1} = 1 - \sum_{k=0}^{J} \omega_k. \tag{7.29}$$

We will use $\omega \doteq x \times \pi$ as shorthand for the last display. Roughly speaking, if $\{x_k\}_{k=0}^{J+1}$ is the occupancy state at time instant t and π_i^k represents the probability of throwing i additional balls over the interval $[nt, nT]$ into the kth category, then ω_i gives the average fraction of category i urns at time nT.

A terminal point ω is **feasible** (for the given initial time and condition) if $\mathscr{F}(x, t; \omega, T)$ is not empty.

Now we are ready to state the theorem. For $s > 0$ let $P(s)$ denote the Poisson distribution with parameter s, and if $s = 0$ let $P(s)$ denote the probability measure on $\{0, 1, \dots\}$ with mass one on $\{0\}$. The proof of the representation can be found in [266].

Theorem 7.13 (Explicit Formula for the Rate Function) *Consider an initial condition $(x, t) \in \mathscr{D}_a$, and a feasible terminal condition ω. If $a \in (0, \infty)$, then for $x_{J+1} > 0$ let*

$$\tau(x, t) \doteq \frac{(t - \sum_{k=0}^{J} k x_k)}{x_{J+1}}$$

(so that $\tau(x, t)$ is the average number of balls per urn distributed in the $J+$ categories for the initial condition (x, t)) and if $x_{J+1} = 0$ let $\tau(x, t) = 0$. Then the quantity $O(x, t; \omega)$ defined in (7.27) has the representation

$$O(x, t; \omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J} x_k R\left(\pi^k \,\middle\|\, Q^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) \right.$$
$$\left. + x_{J+1} R\left(\pi^{J+1} \,\middle\|\, Q^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right)\right].$$

If $a \in -\mathbb{N}$ with $J = -a - 1$ then $\tau(x, t) = J+1$, and

$$O(x,t;\omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J+1} x_k R\left(\pi^k \left\| Q^{a+k}\left(\frac{a+k}{a+t}(T-t) \right) \right) \right].$$

In the final case of $a = \infty$, we have

$$O(x,t;\omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J+1} x_k R\left(\pi^k \| P(T-t) \right) \right].$$

Although these minimization problems as stated appear to be infinite dimensional, they can in fact be reduced to finite dimensional problems. This is because if π^k is the minimizer, then π_j^k takes a prescribed form for $j > J$. In fact, all π_j^k can be represented in terms of no more than $J + 3$ Lagrange multipliers [119, 266].

Theorem 7.13 gives the minimal cost to move from one point in the feasible domain to another. For the construction of accelerated Monte Carlo schemes it is useful to know how to construct subsolutions to the related Hamilton-Jacobi-Bellman (HJB) equation with various terminal conditions. This can often be done by approximating general terminal conditions from below by a special class of terminal conditions, such as those involving affine costs (see the examples in Chap. 17). Such a result is stated in Proposition 7.14, and in fact Theorem 7.13 is shown to be a consequence of Proposition 7.14 by approximating the function equal to 0 when $x = \omega$ and ∞ elsewhere from below by affine functions.

7.2.5.2 The Hamilton-Jacobi-Bellman Equation

In this section we assume $a < \infty$, noting that the Maxwell-Boltzmann case ($a = \infty$) can easily be obtained as a limit. See [119, 266] for further discussion.

The calculus of variations problem (7.27) has a natural control interpretation, where $\theta(s)$ is the control, $\dot{\varphi}(s) = M\theta(s)$ are the dynamics, $R(\theta(s) \| \rho(s, \varphi(s)))$ is the running cost and $g(x) = \infty 1_{\{\omega\}^c}(x)$ is the terminal cost. It is expected that if we define

$$V(x,t) \doteq \inf_{\varphi \in \mathscr{C}([t,T]:\mathscr{P}(\Lambda)),\varphi(t)=x} \left[\int_t^T R(\theta(s)||\rho(s,\varphi(s)))\, ds + g(\varphi(T)) \right], \tag{7.30}$$

then $V(x,t)$ is a weak-sense solution [14] to the HJB equation

$$W_t(x,t) + \mathbb{H}(DW(x,t),x,t) = 0,$$

and terminal condition

$$W(x,T) = \infty 1_{\{\omega\}^c}(x).$$

Here the Hamiltonian $\mathbb{H}(p,x,t)$ is defined by

$$\mathbb{H}(p,x,t) \doteq \inf_{\theta \in \mathscr{P}(\Lambda)} [\langle p, M\theta \rangle + R(\theta \,\|\rho(t,x))]$$

and W_t and DW denote the partial derivative with respect to t and gradient in x, respectively. Note that by the representation formula Proposition 2.2, the infimum in the definition of $\mathbb{H}(p,x,t)$ can be evaluated, yielding

$$W_t(x,t) = \log\left(\sum_{k=0}^{J} x_k \left(\frac{a+k}{a+t}\right) e^{\left(W_{x_k}(x,t) - W_{x_{k+1}}(x,t)\right)} + \left(\frac{a+\tau(x,t)}{a+t}\right) x_{J+1}\right)$$

plus the terminal condition $W(x,T) = g(x)$, where $W_{x_k}(x,t)$ is the partial derivative and $\tau(x,t)$ is as in Theorem 7.13.

A class of problems that are of interest in applications are those with a terminal condition of the form

$$g(x) = \infty 1_{\mathscr{A}^c}(x),$$

where $\mathscr{A}$ is some convex set. Such terminal conditions usually yield only a weak-sense solution, and not a classical-sense C^1 solution to the HJB equation. However, as mentioned previously it is possible for the purposes of design of accelerated Monte Carlo to approximate these terminal conditions from below in terms of affine terminal conditions. In the next result we state a representation for the calculus of variations problem with affine terminal cost $g(\omega) = \langle l, \omega \rangle + b$. The representation turns out to be the unique classical sense solution to the corresponding PDE. To simplify, we first observe that W is a solution of just the PDE alone (i.e., without the terminal condition) if and only if $W + c$ is a solution for any real number c. Since ω is a probability vector, it suffices to prove the representation when $l_{J+1} = 0$ and $b = 0$. We also recall the definition (7.29).

Proposition 7.14 *Consider* $(x,t) \in \mathscr{D}_a$ *and* $g(\omega) = \langle l, \omega \rangle$, *where* $l \in \mathbb{R}^{J+2}$ *and* $l_{J+1} = 0$. *Define* V *by (7.30) and*

$$U(x,t) \doteq \min_{\pi \in \mathscr{F}(x,t;T)} \left[\sum_{k=0}^{J} x_k R\left(\pi^k \,\middle\|\, Q^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) + x_{J+1} R\left(\pi^{J+1} \,\middle\|\, Q^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right) + g(x \times \pi)\right]$$

where $\pi \in \mathscr{F}(x,t;T)$ *means that* π *satisfies the constraints in (7.28). Then* $V(x,t) = U(x,t)$.

7.3 Two Scale Recursive Markov Systems with Small Noise

In this section we consider a discrete time stochastic dynamical system in which there are two components to the state. One of the components evolves at a slower time scale then the other, and this scale separation is determined by the parameter

that also scales the size of the noise. Such systems include many models arising in queuing theory and communication systems [18, 35, 182], where they are called Markov-modulated processes.

We are interested in studying the large deviation behavior of the slow component (though one could also study the joint large deviation properties of the slow component and a time dependent empirical measure of the fast process). The main result of the section is Theorem 7.17, which establishes the LDP for the slow component. The proof, which is left to the reader, combines techniques from Chaps. 4 and 6. We begin by describing the model in precise terms.

7.3.1 Model and Assumptions

Let S be compact metric space and let $p(\xi, d\varsigma)$ be a probability transition kernel on S. We assume that the kernel satisfies the Feller and the transitivity properties from Chap. 6, namely Conditions 6.2 and 6.3. The fast component of the Markov chain will be governed by this kernel. The slow component is described through a stochastic kernel $\theta(dy|x, \xi)$ on $\mathbb{R}^d$ given $\mathbb{R}^d \times S$. We suppose as given a probability space that supports iid random vector fields $\{v_i(x, \xi), i \in \mathbb{N}_0, (x, \xi) \in \mathbb{R}^d \times S\}$, with the property that for any $(x, \xi) \in \mathbb{R}^d \times S$ $v_i(x, \xi)$ has distribution $\theta(\cdot|x, \xi)$. We also suppose as given an S-valued Markov chain $\{\Xi_i\}_{i\in\mathbb{N}_0}$ on this probability space with transition kernel $p(\xi, d\varsigma)$ and with $\Xi_0 = \xi_0 \in S$. The sequence $\{\Xi_i\}$ will be the fast component, and is independent of $\{v_i\}$. The stochastic process describing the evolution of the slow component is then given by

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n, \Xi_{i+1}), \quad X_0^n = x_0.$$

Thus $\{X_i^n\}$ is a stochastic dynamical system with small noise, though the distribution of the noise depends on both X_i^n and the modulating process Ξ_i. The evolution of X_i^n, being scaled by $1/n$, is slow relative to that of Ξ_i. As in Chap. 4 this discrete time process is interpolated into continuous time according to

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n].$$

We are interested in the large deviation properties of the sequence $\{X^n\}_{n\in\mathbb{N}}$ of $\mathscr{C}([0, 1] : \mathbb{R}^d)$-valued random variables.

We impose the following analogues of Conditions 4.3 and 4.7 from Chap. 4. For $(x, \xi) \in \mathbb{R}^d \times S$ and $\alpha \in \mathbb{R}^d$ define

$$H(x, \xi, \alpha) \doteq \log E e^{\langle \alpha, v_i(x,\xi)\rangle}.$$

Condition 7.15 (a) *For each* $\alpha \in \mathbb{R}^d$ $\sup_{(x,\xi)\in\mathbb{R}^d\times S} H(x, \xi, \alpha) < \infty$.

(b) *The mapping* $(x, \xi) \mapsto \theta(\cdot|x, \xi)$ *from* $\mathbb{R}^d \times S$ *to* $\mathscr{P}(\mathbb{R}^d)$ *is continuous in the topology of weak convergence.*

Condition 7.16 *For each* $(x, \xi) \in \mathbb{R}^d \times S$, *the convex hull of the support of* $\theta(\cdot|x, \xi)$ *is* $\mathbb{R}^d$.

7.3.2 Rate Function and the LDP

We next introduce the rate function for $\{X^n\}$. For $\mu \in \mathscr{P}(S)$ define $A(\mu)$ as in Sect. 6.3 [see 6.6]:

$$A(\mu) \doteq \left\{\gamma \in \mathscr{P}(S^2) : [\gamma]_1 = [\gamma]_2 = \mu\right\}.$$

Also, as in Chap. 6, given $\mu \in \mathscr{P}(S)$, let $(\mu \otimes p)(dx \times dy)$ denote the probability measure on S^2 given by $\mu(dx)p(x, dy)$. Let I_1 denote the rate function I in Theorem 6.6:

$$I_1(\mu) = \inf_{\gamma \in A(\mu)} R\left(\gamma \,\|\, \mu \otimes p\right), \ \mu \in \mathscr{P}(S).$$

Define $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ by

$$L(x, \beta) \doteq \inf\left[\int_S R(\nu(\cdot|\xi)\|\theta(\cdot|x, \xi))\mu(d\xi) + I_1(\mu) : \int_{S\times\mathbb{R}^d} y\nu(dy|\xi)\mu(d\xi) = \beta\right],$$

where the infimum is over $\mu \in \mathscr{P}(S)$ and stochastic kernels ν on $\mathscr{P}(\mathbb{R}^d)$ given S. The definition of the local rate function involves two changes of distribution and the associated relative entropy costs. The first switches the distribution of transitions of $\{\Xi_i\}$ from $p(\xi, d\varsigma)$ to $q(\xi, d\varsigma)$, where $[\mu \otimes q]_2 = \mu$. Since X^n moves only a small distance over a small interval in continuous time, it is the invariant distribution μ of q which affects the evolution of the controlled analogue of X^n. Thus if we shift from the invariant distribution of p to μ, then we must pay a cost of $I_1(\mu)$ per unit time. Once this is done, as in Chap. 4 the distribution of the noises $v_i(x, \xi)$ can be perturbed away from $\theta(\cdot|x, \xi)$ to $\nu(\cdot|\xi)$, but one must pay a relative entropy cost. The overall cost to track a velocity β minimizes these two costs.

Recall that $\mathscr{AC}_{x_0}([0, 1] : \mathbb{R}^d)$ denotes the subset of $\mathscr{C}([0, 1] : \mathbb{R}^d)$ consisting of all absolutely continuous functions satisfying $\phi(0) = x_0$. The rate function for $\{X^n\}$ is given as follows. Let

$$I(\phi) = \int_0^1 L(\phi(s), \dot{\phi}(s))ds \quad \text{if } \phi \in \mathscr{AC}_{x_0}([0, 1] : \mathbb{R}^d),$$

and in all other cases $I(\phi) = \infty$.

The following theorem states the LDP for $\{X^n\}$.

Theorem 7.17 *Suppose that Conditions 6.2, 6.3, 7.15 and 7.16 are satisfied. Then I is a rate function and $\{X^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle on $\mathscr{C}([0,1]:\mathbb{R}^d)$ with rate function I, uniformly for initial conditions in compact sets.*

7.3.3 Extensions

We have considered the simplest form of a two scale system in discrete time, and in particular under assumptions such that a straightforward combination of the methods from Chaps. 4 and 6 can be applied to complete the proof. The model can in principle be extended in several directions, under various sets of additional assumptions. For example, as in Chap. 6 the compactness of S can be replaced by a condition on the existence of a suitable Lyapunov function. Likewise the condition on the support of the transition kernel $\theta(\cdot|x,\xi)$, Condition 7.16, can be replaced by a Lipschitz type condition of a similar form as Condition 4.8. Finally, for the model considered here the evolution of the fast component did not depend on the state of the slow variable. This condition can be relaxed to allow for a fully coupled system. See [42] for sufficient conditions in a continuous time setting and [94] for a discrete time system.

7.4 Notes

An overview of occupancy models and their applications can be found in [165]. The first paper to consider the large deviation properties of an occupancy model appears to be [109], which was motivated by the problem of sizing switches in optical communications. In [109] the LDP for the MB model is obtained, and the rate function exhibited in more-or-less explicit form. The arguments in Sect. 7.2 are based on those used in [266], though as in previous chapters the presentation here first studies the large deviation properties of an empirical measure and then obtains those for the process.

As was discussed in Sect. 7.2, the most difficult part of the analysis is in dealing with parts of the state space where rates go to zero, which produces singular behaviors in the local rate function. The are many other classes of models in applied probability where transition probabilities vanish (or in their continuous time analogues jump rates vanish), including models from queueing and related stochastic networks [231], chemical reaction networks, and random graphs [23]. A positive feature of this collection of problems (one that is emphasized in Sect. 7.2) is that the associated variational problems have explicit or nearly explicit solutions.

The main difficulties are typically in the proof of the large deviation lower bound, and the approach used in this chapter involves a careful analysis of the local rate function to construct controls that can be used to establish the lower bound. For the corresponding continuous time models, one can sometimes represent the process

as the solution to a stochastic differential equation driven by one or more Poisson random measures. In this case one might ask if the perspective of Sects. 3.1 and 3.3, which exploits the fact that the mapping from the noise model (Brownian motion or Poisson random measure) into the state variable is "nearly continuous" could be used. This turns out to be possible, as described for example in [23].

The second model of this chapter is a stochastic recursive system with two time scales. Models of this type appear in many different areas of application, and general references include [171, 259]. One of the first papers to consider the large deviation properties of processes of this general sort is Freidlin [138]. Continuous time analogues of such two time scale systems have also been well studied (see [42] and references therein). Related and very challenging problems involve systems where the averaging is with respect to an "environment" variable rather than time, e.g., a stochastic differential equation where the drift is itself random or periodic and ergodic in an appropriate sense. An example of how weak convergence methods can be used to account for such averaging in a relatively simple setting appears in [111].