CHAPTER 3

Metastability

Although it is intuitively clear what the term *metastable dynamical behavior* means, the mathematical definition of *metastability* is not a simple task. In fact, the term is defined in many ways in the literature. In this chapter we will review several of these definitions and related rigorous results for different systems. Only after discussing these different cases will we try to compile ways to define metastability. Our final definition, however, will have to wait until the end of the next chapter.

3.1. Hitting Times and Committors

In order to introduce the concept of metastability of a Markov process we will need several stopping times. Although stopping times rigorously have to be introduced relative to a filtration, we will omit the explicit introduction of this filtration in what follows. Our statements should be understood as relative to the filtration generated by the process.

We will mainly be concerned with questions like:

- How long does the process need to hit a certain set?
- How long does the process need to exit from a given set?
- Which of the sets A and B will be hit next by the process? What is the probability that A is hit next?

While the first two questions can be answered by computing expectation values or distributions of stopping times, the questions listed third require some basic notions of potential theory, which will be introduced next.

Let us first denote the *hitting time* for a set $A \subset \mathbb{X}$ conditional to the process being started in $X_0 = x$ by

$$\tau_x(A) = \inf\{t \geq 0 : X_t \in A\}.$$

In many cases one is mainly interested in the expectation value of the hitting time as a function of the starting point x,

$$m_A(x) = \mathbb{E}(\tau_x(A)),$$

which often is called the *mean first passage time*. Another very important quantity is the *committor function* associated with a pair of sets A and B [**67, 68**] (also called the "equilibrium potential" in [**7**]). The committor function is the probability of hitting B next if starting in x, i.e., the probability that starting in x, B is hit before A, which can be expressed by

$$q_{AB}(x) = \mathbb{P}(\tau_x(B) < \tau_x(A)). \tag{3.1}$$

Obviously, the committor function is mainly of interest if the two hitting times are almost surely finite as, e.g., in the case of an ergodic process with an invariant measure μ with $\mu(A), \mu(B) > 0$.

Both functions, mean first passage times as well as committors, can be considered as a special case of the so-called potential ϕ associated with two real-valued functions c and g on state space, a set D, and its complement $D^{\mathsf{c}} = \mathbb{X} \setminus D$,

$$\phi(x) = \mathbb{E}\left[\int_0^\tau c(X_t)dt + g(X_\tau)\mathbb{1}_{\tau<\infty} | X_0 = x\right], \tag{3.2}$$

where the expectation is relative to the law generated by the process and $\tau = \tau_x(D^{\mathsf{c}})$ is the hitting time of the complement of D. There are some restrictions to the choice of D. For example, for processes with $\mathbb{T} = \mathbb{R}$ and continuous paths, the process is stopped on the boundary ∂D of D, i.e., $X_\tau \in \partial D$, such that D should be an open set with a "nice" boundary. The functions c and g are regarded as *cost functions* so that $\int_0^\tau c(X_t)dt$ is the cost for "wandering around" in D, while $g(X_\tau)$ is the final cost when the process hits either D^{c} (discrete case) or the boundary ∂D (continuous case), and the potential ϕ can be interpreted as an *expected total cost*.

Many interesting quantities can be formulated as potentials [**19**]. For the moment we restrict our attention to the consideration of the mean first passage time and the committor function as special cases of potentials. To this end, first consider a set A, choose $D = A^{\mathsf{c}}$ so that $D^{\mathsf{c}} = A$, and set $c = \mathbb{1}$ the constant function and $g = 0$. Then the potential ϕ takes the form [**19**]

$$\phi(x) = \mathbb{E}\left[\int_0^{\tau_x(A)} c(X_t)dt\right] = \mathbb{E}[\tau_x(A)] = m_A(x).$$

Furthermore, if we choose two sets $A, B \subset S$ with $A \cap B = \emptyset$ and $\mathbb{P}(\tau_x(A) < \infty) = 1$, set $D = \mathbb{X} \setminus (A \cup B)$ such that $D^{\mathsf{c}} = A \cup B$, set $c = 0$, and choose $f = \mathbb{1}_B$, the indicator function of the set B, we arrive at

$$\begin{aligned}\phi(x) &= \mathbb{E}\big[f(X_{\tau_x(D^{\mathsf{c}})})\big] = \mathbb{E}\big[\mathbb{1}_B(X_{\tau_x(A\cup B)})\big] = \mathbb{P}\big[X_{\tau_x(A\cup B)} \in B\big] \\ &= \mathbb{P}[\tau_x(B) < \tau_x(A)] = q_{AB}(x).\end{aligned}$$

Potential theory provides the means for computing the potential ϕ as the solution of a *linear* problem. This will allow us to find linear equations for the mean first passage time and the committors. We will return to this later (Section 4.2) after the introduction of some additional tools.

For now let us get a first impression of how committor functions may look. To this end we consider diffusion molecular dynamics (2.12) in $\mathbb{X} = \mathbb{R}^2$ in the two different potential energy landscapes shown in Figure 3.1 (top panels). The first three-well energy landscape in the left top panel of Figure 3.1 is given by

$$\begin{aligned}V(x,y) = {} & 3e^{-x^2-(y-\frac{1}{3})^2} - 3e^{-x^2-(y-\frac{5}{3})^2} \\ & - 5e^{-(x-1)^2-y^2} - 5e^{-(x+1)^2-y^2} \\ & + \frac{2}{10}x^4 + \frac{2}{10}\left(y - \frac{1}{3}\right)^4\end{aligned} \tag{3.3}$$

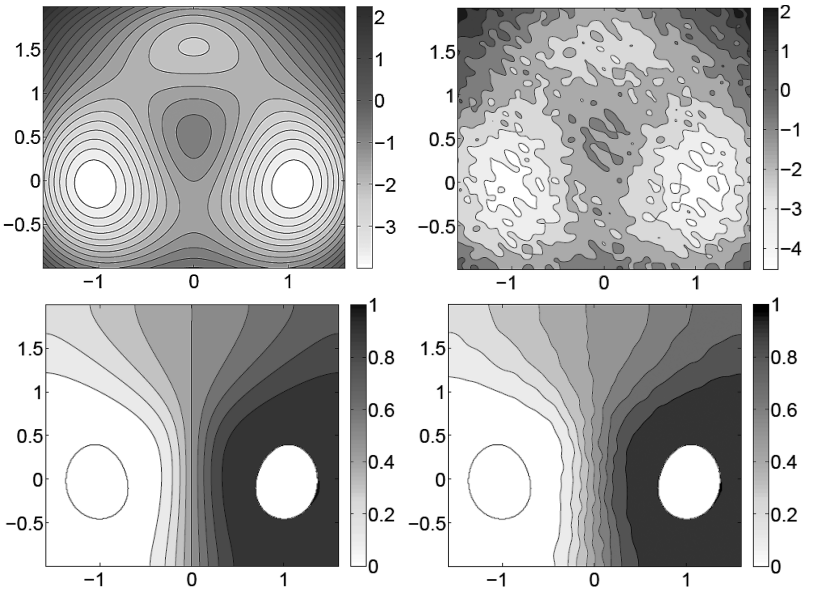

FIGURE 3.1. Top left panel: Three-well energy landscape V as described in the text. Top right panel: Rugged three-well energy landscape. Bottom panel: Committor functions q_{AB} for diffusion molecular dynamics with $\beta = 1.67$ and $\gamma = 1$ for the sets A (left main well) and B (right main well) for the three-well energy landscape (left) and the rugged three-well energy landscape (right). For details of the computations underlying the pictures, see [**65**].

and exhibits three wells, two of them being equally deep and the third one being significantly less deep. The second energy landscape in the right top panel of Figure 3.1 is a significantly rugged version of the first one where the ruggedness results from the addition of many local minima but in a form that still allows observation of the three original wells. The bottom panels of Figure 3.1 show the committor functions q_{AB} of the diffusion molecular dynamics process with $\gamma = 1$, $\beta = 1.67$, and $\sigma = \sqrt{2\gamma/\beta}$ in these two energy landscapes for the sets A and B that correspond to the two main wells of the energy landscapes. We observe that the committor for the smooth three-well energy landscape shown in the bottom left panel exhibits a smooth change from the value 0 in and close to A (left main well) and 1 close to B (right main well). The so-called isocommittor line, i.e., the contour set $q_{AB} = \frac{1}{2}$, on which the process is equally committed to A and B, is a straight line exactly between A and B. The committor for the rugged three-well energy landscape (bottom right value) looks like a rugged version of the first one.

If we reduce the temperature, e.g., by choosing $\gamma = 1$, $\beta = 6.67$, we get the committor function shown in Figure 3.2. While the committor for the smooth three-well potential (not shown) changes insignificantly, the committor for the rugged three-well potential looks severely rugged now.

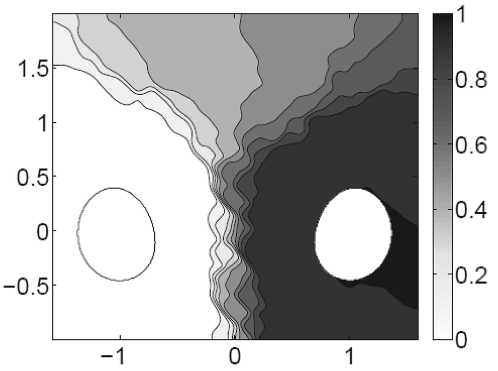

FIGURE 3.2. Committor functions q_{AB} for diffusion molecular dynamics with $\beta = 6.67$ and $\gamma = 1$ for the sets A (left main well) and B (right main well) for the rugged three-well energy landscape; compare [**65**].

We will discuss this example and a number of other ones below in more detail. However, the example just discussed contains a general warning that is not discussed very often: in rugged energy landscapes the committor functions between the main wells can exhibit rather complicated spatial structures, thus having rather rugged contour sets (especially for low temperatures).

The committor function q_{AB} has another interpretation that is at the core of the potential theoretic approach to metastability [**7, 19**]. In order to avoid unnecessary technical complications, let us restrict our consideration for the rest of this paragraph to the case of discrete state spaces $\mathbb{X}$ and some irreducible and aperiodic Markov process defined on $\mathbb{X}$ with transition function $p(x, y)$.

Next, let us consider the so-called Dirichlet form of two functions $f, g : \mathbb{X} \to \mathbb{R}$,

$$D(f, g) = \sum_{x,y \in \mathbb{X}} \mu(x)p(x, y) \cdot (f(x) - f(y)) \cdot (g(x) - g(y)). \tag{3.4}$$

Then, the committor function comes out to be the minimizer of the following variational problem: Let H_{AB} denote the set of all functions f on $\mathbb{X}$ such that $f|_A = 0$ and $f|_B = 1$ and such that $D(f, f) < \infty$. Then the committor function is the unique minimizer of

$$q_{AB} = \operatorname{argmin}_{f \in H_{AB}} D(f, f). \tag{3.5}$$

The minimum value of the Dirichlet form is called the *capacity* of the two sets A, B:

$$\operatorname{cap}(A, B) = \min_{f \in H_{AB}} D(f, f) = D(q_{AB}, q_{AB}). \tag{3.6}$$

3.1.1. Hitting Times and Metastability. Next, we will see how hitting times can be used to characterize the transition times between deep wells in an energy landscape. To this end let us consider the diffusive molecular dynamics model

(2.12) with potential energy landscape V, and choose $\gamma = 1$ and $\sigma = \sqrt{2\epsilon}$,

$$\dot{x}_t = -\nabla V(x_t) + \sqrt{2\epsilon}\,\dot{W}_t$$

such that the invariant measure has the form

$$\mu(\mathrm{d}x) = \frac{1}{Z}\exp\left(-\frac{1}{\epsilon}V(x)\right)\mathrm{d}x.$$

The energy barriers between the wells of the potential V are on the order $\mathcal{O}(1)$, while the average energy put into the system by the noise in a time interval of order $\mathcal{O}(1)$ is of order $\mathcal{O}(\epsilon)$. Therefore transitions from one well into the others are rare events and the basins of the wells intuitively are metastable sets of the dynamics. In order to be more precise we have to understand the relation between metastability and exit times from the wells. More precisely, we will compute explicit asymptotic expressions for the hitting time of the bottom of one well if the process is started in the bottom of another well. It is well-known that this hitting time scales like $\exp(\Delta/\epsilon)$ if Δ is the height of the energy barrier separating the two wells. However, almost all mathematically rigorous statements regarding this standard "Kramer's rule" consider the case of two minima only and are restricted to the one-dimensional case. We will review a rather general statement. To this end we need some additional quantities and some technical assumptions on V:

For any disjoint sets $A, B \subset \mathbb{X}$ define the height of the saddle between A and B by

$$\widehat{V}(A,B) = \inf_{\gamma\in\mathcal{P}(A,B)} \sup_{t\in[0,1]} V(\gamma(t)),$$

where $\mathcal{P}(A,B)$ denotes the set of all continuous paths γ in $\mathbb{X}$ with $\gamma(0) \in A$ and $\gamma(1) \in B$. Based on that, define the set of minimal points on these paths,

$$\mathcal{G}(A,B) = \{z \in \mathbb{X} : V(z) = \widehat{V}(A,B)\}.$$

Moreover, we denote by $\mathcal{P}_{\min}(A,B)$ the set of minimal paths from A to B,

$$\mathcal{P}_{\min}(A,B) = \{\gamma \in P(A,B) : \sup_{t\in[0,1]} V(\gamma(t)) = \widehat{V}(A,B)\},$$

and by $S(A,B)$ the *set of saddle points* as the maximal subset of $\mathcal{G}(A,B)$ such that for every $x \in S(A,B)$ there is a minimal path $\gamma \in \mathcal{P}_{\min}(A,B)$ that goes through x.

ASSUMPTION 3.1. Let the potential V be three times continuously differentiable and let it satisfy the following growth conditions:

$$\begin{aligned}\liminf_{x\to\infty} V(x) &= \infty,\\ \liminf_{x\to\infty} |\nabla V(x)| &= \infty,\\ \liminf_{x\to\infty} (|\nabla V(x)| - 2\Delta V(x)) &= \infty.\end{aligned}$$

Moreover, assume that V has finitely many minima $x \in \mathcal{M} = \{x_1, \dots, x_n\}$ and that for every minimum $x \in \mathcal{M}$ and any set $M \subset \mathcal{M}$ of other minima with $x \notin M$, the set of saddle points $S(\{x\}, M)$ contain exactly one point,

$$S(\{x\}, M) = \{z(x, M)\},$$

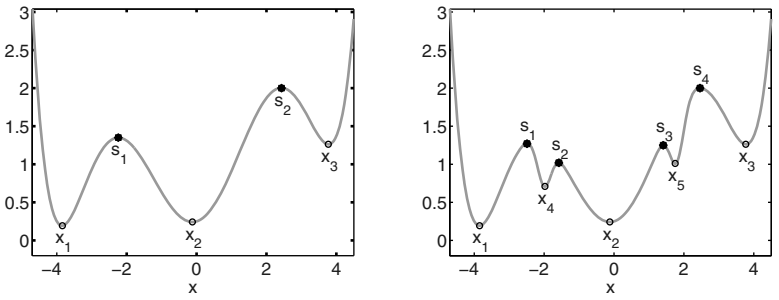

FIGURE 3.3. Illustration of Assumption 3.1 on minima and saddle points for two different example potentials.

and the Hessian of V in the minima x_i as well as in the saddle points $z(x, M)$ has only nonzero eigenvalues such that $\det(\nabla^2 V)$ does not vanish in the minima or in the saddle points.

This assumption gives a lot of structure to the sets of minima and essential saddle points of V but is rather generic for realistic potentials. It allows us to define the *depth of the wells* around the minimum $x \in \mathcal{M}$ regarding transitions to other minima from $M \subset \mathcal{M}$ with $x \in \mathcal{M}$ as

$$\Delta(x_k, M) = V(z(x, M)) - V(x_k).$$

Before we go on we should have a short look at Figure 3.3 for an illustration of the meaning of the objects introduced above. For the three-well potential shown in the left-hand panel of Figure 3.3, everything is easily understood and we get, e.g.,

$$z(x_1, \{x_2, x_3\}) = s_1, \quad z(x_2, \{x_1\}) = s_1 \quad z(x_3, \{x_1, x_2\}) = s_2.$$

For the five-well potential in the right panel things are just as easy and we get

$$z(x_1, \{x_2, x_3\}) = s_1, \quad z(x_2, \{x_1\}) = s_1, \quad z(x_3, \{x_1, x_2\}) = s_4,$$
$$z(x_4, \{x_1, x_2, x_3\}) = s_2, \quad z(x_5, \{x_1, x_2, x_3, x_4\}) = s_3, \quad z(x_5, \{x_3\}) = s_4.$$

Under the above assumption the following statement holds [**7**]:

THEOREM 3.2. *Let Assumption* 3.1 *hold, let* $B(y)$ *denote a ball of radius* ϵ *around the minimum* $y \in \mathcal{M}$*, select a minimum* x*, and let* $M_x = \{y_1, \dots, y_k\}$ *be the subset of minima such that* $V(y_j) \le V(x)$*,* $y_j \neq x$*. Set*

$$S_k = \bigcup_{j=1}^{k} B(y_j),$$

and assume that ϵ *is small enough such that* $\operatorname{dist}(z(x, M_x), S_k) > 0$*. Then the hitting time* $\sigma_x(S_k)$ *for the process started in* x *satisfies*

$$\mathbb{E}[\sigma_x(S_k)] = C \exp\Big(\frac{1}{\epsilon}\Delta(x, M_x)\Big)(1 + \mathcal{O}(\epsilon^{1/2}|\log \epsilon|)),$$

with a prefactor

$$C = \frac{2\pi}{|\hat{\lambda}|} \frac{\sqrt{|\det(\nabla^2 V(z(x, M_x)))|}}{\sqrt{|\det(\nabla^2 V(x))|}},$$

where $\hat{\lambda}$ denotes the negative eigenvalue of the Hessian of V at the saddle point $z(x, M_x)$. Furthermore, one finds that the exit time $\sigma_x(S_k)$ is asymptotically exponentially distributed.

For an energy landscape with n minima, this theorem gives us statements about $n-1$ exponentially scaled hitting times. It does not give us a statement about, e.g., the hitting time of the process for the second lower minimum if started in the lowest minimum. For the five-well potential from Figure 3.3, for example, four hitting times are characterized: (1) from (the vicinity of) the minimum x_2 to (the vicinity of) the global minimum x_1 across $s_2 = z(x_2, \{x_1\})$, (2) from x_4 to x_2 (the process will always hit x_2 before x_1, which is special in the one-dimensional case) across $s_2 = z(x_4, \{x_1, x_2\})$, (3) from x_5 to x_2 across $s_3 = z(x_5, \{x_1, x_2, x_4\})$, and (4) from x_3 to x_2 across $s_4 = z(x_3, \{x_1, x_2, x_4, x_5\})$.

REMARK 3.3. The set S_k in Theorem 3.2 can be extended to a closed set D that is a union of connected, closed subsets D_k, $k = 1, \dots, m$, such that

(i) $\bigcup_{j=1}^{k} B(y_j) \subset D$,
(ii) all D_k, $k = 1, \dots, m$, contain at least one of the minima y_j, and
(iii) $\operatorname{dist}(z(x, M_x), D) > \delta > 0$ for some δ independent of ϵ,

then the hitting time $\sigma_x(D)$ from D also satisfies the statement of Theorem 3.2. The statement in [**7**] should be considered carefully since condition (ii) is missing.

3.2. Exit Times and Exit Rates

The *exit time* σ from a set $A \subset \mathbb{X}$ is often defined as the hitting time of the complement of A, i.e.,

$$\sigma_x(A) = \tau_x(A^{\mathsf{C}}).$$

While this definition is sufficient in the discrete case, we will use a more sophisticated one in the continuous case: Let $A \subset \mathbb{X}$ be some connected open subset and consider some point $x \in A$ and define

$$\varrho_x(A) = \inf\left\{ t \geq 0 : \int_0^t \mathbb{1}_{A^{\mathsf{c}}}(X_s)\, \mathrm{d}s > 0 \right\}, \tag{3.7}$$

which measures only exits that happen for some nonnull time interval neglecting exit events that are "singular" in time.

For later use let us consider the asymptotic decay of the *distribution of exit times* [**46, 100**]

$$F_x(s) = \mathbb{P}[\varrho_x(A) \geq s].$$

While for small values of s the function F_x may show complicated behavior, it asymptotically may decay almost exponentially, at least under certain well-established conditions. The decay rate of F_x can best be expressed by means of the

conditional exit-time distribution

$$F_x(s,t) = \mathbb{P}[\varrho_x(A) \geq s + t \mid \varrho_x(A) \geq t]$$

for $s, t \geq 0$, which describes the tail of the distribution, for which the exit time is larger than the so-called waiting time t. The decay rate is equal to Γ if the conditional distribution decays exponentially with rate $\Gamma > 0$, i.e.,

$$F_x(s,t) \propto \exp(-\Gamma s), \tag{3.8}$$

for $s \geq 0$ and $t \geq 0$.

When aiming at a definition of decay rates for entire *subsets*, there are two problems. First, the relation (3.8) will only hold for very special Markov processes; see [**46**]. Second, we have to expect that the decay rate depends on the starting point, i.e., $\Gamma = \Gamma_x$. As we will see later, there are specific sets for which there is a time t such that the decay rate Γ of $F_x(s,t)$ is asymptotically exponential and independent of x for all $x \in A$. In this case, we will call $\Gamma = \Gamma(A)$ the *exit rate* from the set A.

3.2.1. Exit Rates and Metastability. The approach presented herein is based on the fact that there exists a subset C for which the decay rate is basically independent for all states $x \in C$. In a more general setting but for a specific class of dynamical systems including, e.g., the case of diffusion molecular dynamics, we are able to assign a so-called exit rate $\Gamma = \Gamma(C)$ to an entire subset $C \subset \mathbb{X}$ rather than to single points $x \in \mathbb{X}$, thus circumventing the two above-mentioned problems with (3.8). These exit rates may be thought of as some generalization of decay rates; see Appendix B, in particular Theorem B.1.

Figure 3.4 gives a first, still rough illustration of the situation. There we again consider diffusive molecular dynamics ($\gamma = 1$, $\sigma = 0.7071$) in the three-well potential of Figure 3.3. The exit-time distribution is shown for the (metastable) set $D = (-\infty, -2.074]$, whose choice will be described in extensive detail later and for two different initial states, one being the leftmost (and global) minimum x_1 of the potential, the other one being located almost at the right boundary of the set. For small exit times the two distributions are quite different. The reason simply is that the process with initial state close to the boundary exits from the set rather frequently in short time, but with a certain probability it first relaxes to the minimum x_1 before exiting from there. Thus the tail of the exit-time distributions should be the same. And in fact, despite the fact that the exit-time distributions are very different for small exit times, they clearly show the same exponential decay rate for very long exit rates, $\Gamma(D) = 0.0056$. What Figure 3.4 does *not* show is that the complement of D, $D^{\mathsf{C}} = (-2.074, \infty)$, exhibits *exactly the same* small exit rate, $\Gamma(D^{\mathsf{C}}) = 0.0056$.

What we have learned so far regarding the asymptotic scaling of exit times and exit rates seems to indicate that the exit rate from any open subset of set D should be larger than the exit rate from D; i.e., the decay of the exit-time distribution is faster. Therefore, we may call a subset $B \subset \mathbb{X}$ metastable with exit rate $\Gamma(B)$ if

$$\Gamma(A) > \Gamma(B) \text{ for all open, connected sets } A \subset B,\ A \neq B. \tag{3.9}$$

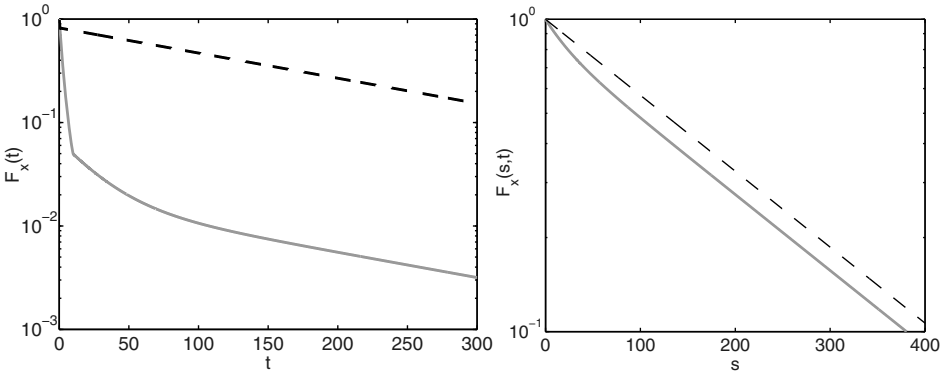

FIGURE 3.4. Dependence of the exit-time distribution $F_x(s)$ on the exit time s for the set $D = (-\infty, -2.074]$ and diffusive molecular dynamics ($\gamma = 1$, $\sigma = 0.7071$) in the three-well potential in Figure 3.3. Left hand panel: Distribution $F_x(t)$ in a semilogarithmic plot for $0 \leq t \leq 300$ for two different initial states $x = -3.84$ (the leftmost minimum of the potential; dashed line) and $x = -2.2$ (close to the right boundary of the set D; solid line). Right hand panel: Conditional distribution $F_x(s,t)$ in semilogarithmic plot for $t = 80$ versus s for two different initial states $x = -3.84$ (dashed line) and $x = -2.2$ (solid line). The asymptotic decay rate of both distributions $F_x(s,t)$ is estimated as $\Gamma(D) = 0.0056$.

3.3. Metastable Full Partitions and Almost Invariance

Since we are aiming at the construction of MSMs we should be interested in full partitions using metastable sets. Therefore, we may call a full partition $A_1, \ldots, A_m$ metastable if all A_k are metastable. As we will see, there are metastable full partitions such that $\Gamma(A_k) = \Gamma$ for all $k = 1, \ldots, m$, as, e.g., in the last example where we considered the decomposition D, D^{c}. We will see later that there are different metastable full partitions such that we would be interested in finding the "best" one. In the context of exit rates this "best" metastable full partition into m sets could be the one that, for example, minimizes

$$\max_{k=1,\ldots,m} \Gamma(A_k).$$

The literature contains another concept for the *optimal metastable full partition* for which a timescale T has to be selected, and we consider the probability

$$p(A, A) = \mathbb{P}_\mu(X_T \in A \mid X_0 \in A),$$

which measures the probability that we find the process in the set A at time $t = T$ after having started in A at time $t = 0$. For a metastable set this probability will be close to 1 if the timescale T is large but not as large as the transition times between metastable sets. Therefore a full partition $A_1, \ldots, A_m$ is called metastable

with respect to timescale T if

$$\sum_{k=1}^{m} p(A_k, A_k) \approx m. \tag{3.10}$$

Thus the *optimal* metastable full partition into m sets is given by

$$(A_1^*, \dots, A_n^*) = \operatorname{argmax}_{(A_1,\dots,A_n)} \sum_{k=1}^{m} p(A_k, A_k). \tag{3.11}$$

Finally, let us introduce another term from the literature: An invariant set cannot be left by the process; i.e., for an invariant set A we have $p(A, A) = 1$. Whenever a metastable full partition satisfies (3.10), each of its sets A_k satisfies

$$p(A_k, A_k) \approx 1,$$

which means they can be considered as *almost invariant*.

3.4. Definitions of Metastability

The intuitive understanding of metastability is connected to the metastability of a subset A of state space under the dynamics in the following sense: A is not invariant (stability on a timescale $t = \infty$) but is almost invariant in the sense that the process does not exit from A over a timescale t that is much longer than some "normal" reference timescale t_f for the dynamical fluctuations of the system. The choice of t_f depends on the system and will be explained below; let us assume it has been selected properly.

So far we have discussed three different aspects of metastability:

Large hitting times: The process is metastable if we find (small) sets with the property that the hitting time for one of these sets starting from another one is very large compared to t_f, i.e., $\mathbb{E}(\tau_x(A)) \gg t_f$; this is the situation in Theorem 3.2.

Small exit rates: Metastability of some set A is characterized by the property that the distribution of exit times from the set is asymptotically exponential with some small rate $\Gamma(A) \ll t_f^{-1}$.

Metastable full partition: A full partition $A_1, \dots, A_m$ is called metastable if

$$\sum_{k=1}^{m} \mathbb{P}_\mu(X_{t_f} \in A_k \mid X_0 \in A_k) \approx m.$$

Then each of its sets A_k is almost invariant with respect to timescale t_f.

The small sets in the large hitting time approach do not form a full partition of state space such that the relation to MSMs is not obvious based on what we have discussed so far. In contrast, the small exit rates approach allows for full partition into sets with small exit rates while the metastable full-partition approach seems to be tailored to allow the construction of MSMs. We will finally see that the three approaches can be composed into one concerted approach to the construction of MSMs; in order to achieve this composition, we need additional mathematical tools.