The index identifies definitions by bold page numbers, and includes the frequently used symbols.

> We often wish to make a comment, or expand on a particular topic, in such a way that the reader may feel free to skip the comment, but will know that it is there. This is how we do it: in small type, in a narrow paragraph.

0.3. Acknowledgments

This project spanned many more years than we had ever anticipated. In the course of those years we have had help from many, many people. Preeminent among them are Armand Makowski, Debasis Mitra, and S.R.S. Varadhan. Debasis was steadfast in his support: moral, financial, and technical. He believed in us when we weren't sure we believed in ourselves. This project would never have been done without him. And we would never have gotten into the field (it is not certain that there would be much of a field to get into!) without Professor Varadhan. It was a tremendous comfort to know that there was no technical point, however difficult or subtle, that could not be answered almost instantly by a simple visit to NYU. Armand Makowski not only introduced us, and not only is he responsible for stating that the world would benefit from lecture notes on queueing applications of large deviations, but he can also be held accountable for doing something about it. With the support of John Baras, Armand arranged a sabbatical at the Systems Research Center where a first draft of these notes was hammered out by AW, and provided a sabbatical at the Systems Research Center where, somewhat unexpectedly, most of the time of AS was devoted to this project.

There are many more people who have helped over the years. Robert J. Vanderbei was, for a time, a coauthor of the book, and one appendix still bears his sole authorship. Several "field tests" of these ideas were graciously hosted by Armand Makowski at the University of Maryland, College Park, by Elja Arjas and the Finnish Summer School, and by Sid Browne and the Columbia Business School and Department of Mathematics. Within Bell Labs and the Technion, our home institutions, it seems that nearly everyone had something to contribute. Notable among those were co-large deviants Ofer Zeitouni and Amir Dembo. Also, Marty Reiman was a constant source of technical wisdom, moral support (what do you mean you aren't done yet?), and was an invaluable asset in transportation and living accommodations. Howard Trickey was our accessible TeX wizard, and justified his title hands down. Thanks also to Andrew Trevorrow for long-distance TeX help.

There were many students and colleagues who gave suggestions and feedback on everything from typos to approach, from early drafts to the first printing of this book. They include also those attending several courses given at the Technion, as well as lectures delivered at AT&T. We are particularly grateful for comments from Laurence Baxter, Henri Casanova, Hong Chen, Bill Cleveland, Ed Coffman, Amir Dembo, Amanda Galtman, Leo Flatto, Ben Fox, Predrag Jelenkovic, Ariel Landau, Armand Makowski, Colin Mallows, Bill Massey, Jim Mazo, Debasis

Mitra, Marty Reiman, Emre Telatar, Stephen R.E. Turner, Yashan Wang, Phil Whiting, Ward Whitt, Paul Wright, Aaron Wyner, Ofer Zeitouni, and Li Zhang.

The editor-in-chief of this series Laurence Baxter did yeoman's work. Our editor John Kimmel amazed us by answering "yes" to every one of our requests, and promptly, too!

Typists Sue Pope and Lesley Price helped turn scribbled handwriting into beautiful TeX, quickly, accurately, and cheerfully.

This book was produced using TeX, with AS serving as local TeXpert, and was set in Times Roman, with MathTimes and other math fonts from Y&Y. The figures were drawn by AW using Canvas©, and according to the egalitarian tendencies of the authors, was set on Macintosh©, UNIX©, and various PC computers and clones.

I (AS) am delighted for this opportunity to acknowledge Armand Makowski for his role as colleague, collaborator, and catalyst in my professional life and, above all, to express my appreciation for his friendship.

And I (AW) am eternally grateful for my two mentors, Debasis Mitra and Raghu Varadhan. These two fine men have unselfishly nurtured me throughout this and other projects. I hope that in some way they can find some recompense in this volume.

Finally, our families, particularly our wives Shuli Cohen Shwartz and Judy Weiss, deserve thanks for putting up with us during all these years of labor. While we'll never know, it was probably as hard on them as having children; it was certainly longer and with less reward at the end. We promise we'll never do it again.

Note: This printing incorporates all the corrections we accumulated during the first year the book was out. We thank our readers for reporting these mistakes, and our publisher for allowing us to make the changes. However, from the number of mistakes found so far, we know that more will be found. Please send comments to us at apdoo@research.bell-labs.com or adam@ee.technion.ac.il. You can obtain the latest errata sheet at any of the following locations:

http://cm.bell-labs.com/who/apdoo

http://www-ee.technion.ac.il/~adam

http://users.aol.com/apdoo

Chapter 1

Large Deviations of Random Variables

This chapter can be viewed as a guided tour through basic large deviations. Following a heuristic exposition, we derive large deviations estimates for random variables. We provide proofs when these provide insight, or are typical; otherwise, we provide references. The modern tools and approaches, especially those that have proved useful for the applications, are discussed in Chapter 2 and Appendix D.

The main results, Theorems 1.5, 1.10, and 1.22, as well as the computations of Examples 1.13–1.18 and Exercises 1.6, 1.17–1.25, will be used heavily throughout the book.

1.1. Heuristics and Motivation

Estimates of probabilities of rare events turn out to have an exponential form; i.e., these probabilities decrease exponentially fast as a function of the asymptotic parameter. To motivate the exponential form of the large deviations estimates, consider the following examples. Let $x_1, x_2, \ldots$ be a sequence of independent, identically distributed (i.i.d.) random variables with a common distribution function F and finite mean. Fix a number $a > \mathbb{E}x_1$. Now the probability that $x_1 + \cdots + x_n > na$ is clearly decreasing in n in a long-term sense, since by the (weak) law of large numbers

$$\mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \geq a\right) \to 0 \text{ as } n \to \infty.$$

The next question would be: How fast does this probability decrease? Let us perform some quick calculations. First, if for some integer k,

$$x_{jk+1} + \ldots + x_{(j+1)k} \geq ak \quad \text{for all} \quad j = 0, \ldots, (n/k) - 1,$$

then clearly $x_1 + \cdots + x_n \geq na$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \geq a\right) \\
&\qquad \geq \mathbb{P}\left(x_{jk+1} + \ldots + x_{(j+1)k} \geq ak \quad \text{for all} \quad j = 0, \ldots, (n/k) - 1\right) \\
&\qquad = \left(\mathbb{P}\left(x_1 + \ldots + x_k \geq ak\right)\right)^{n/k}
\end{aligned}$$

by independence. This immediately implies that the rate of convergence is at most exponential. On the other hand, for any positive θ, by Chebycheff's inequality

The lower bound in $\mathbb{R}^1$ was based on estimating the probability of the sample mean being in a small interval around the point where ℓ is smallest. A similar argument works in the d-dimensional case: we need to consider small neighborhoods, or balls, around the minimizing point.

Exercise 1.23. Generalize Exercise 1.7 to $\mathbb{R}^d$. ♠

For the upper bound, consider half-spaces of the form $H_{s,a} \stackrel{\triangle}{=} \{\vec{y} : \langle \vec{s}, \vec{y} \rangle \geq a\}$ for some $\vec{s} \in \mathbb{R}^d$ and $a > 0$. Then $\tilde{x}_i \stackrel{\triangle}{=} \langle \vec{s}, \vec{x}_i \rangle$ are i.i.d. and

$$\frac{\vec{x}_1 + \ldots + \vec{x}_n}{n} \in H_{s,a} \iff \frac{\tilde{x}_1 + \ldots + \tilde{x}_n}{n} \geq a.$$

Exercise 1.24. Assume ℓ is continuous and finite. Prove the upper bound for convex sets in $\mathbb{R}^d$. Extend the proof to finite unions of convex sets. Hint: use Chernoff's Theorem for $\tilde{x}_1, \tilde{x}_2, \ldots$. Note that ℓ is convex as explained below (1.4) so that $C_\varepsilon \stackrel{\triangle}{=} \{\vec{x} : \ell(\vec{x}) \leq J(C) - \varepsilon\}$ is convex, and has empty intersection with C. Therefore there is a half-space containing C that does not intersect C_ε. ♠

The following calculation will be used for our Poisson processes. It follows from (1.10)–(1.12).

Exercise 1.25. Suppose y_{ij} are i.i.d. with $y_{ij} \stackrel{\mathcal{L}}{=} Pois(\lambda_j)$. Define

$$\vec{x}_i \stackrel{\triangle}{=} \sum_{j=1}^{J} y_{ij} \vec{e}_j .$$

Show that $\ell(\vec{a})$ defined in (1.10b) has the form

$$\ell(\vec{a}) \stackrel{\triangle}{=} \sup_{\vec{\theta}} \left(\langle \vec{\theta}, \vec{a} \rangle - g(\vec{\theta}) \right), \tag{1.13}$$

where $g(\vec{\theta}) = \log M(\vec{\theta})$ is given by

$$g(\vec{\theta}) = \sum_{j=1}^{J} \lambda_j \left(e^{\langle \vec{\theta}, \vec{e}_j \rangle} - 1 \right). \tag{1.14}$$

Consequently, Exercise 1.24 implies that

$$\lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P} \left(\left| \frac{\vec{x}_1 + \cdots + \vec{x}_n}{n} - \vec{a} \right| < \varepsilon \right) = -\ell(\vec{a}). \quad ♠$$

Let F be a closed set and define

$$I^*(F) = \inf_{y \in F} I(y) \quad \text{and} \quad \mathcal{O} \stackrel{\triangle}{=} \left\{x \in F : I(x) = I^*(F)\right\}. \tag{2.2}$$

If I is a good rate function, then $\mathcal{O}$ is not empty, so that the minimum $I^*(F)$ is attained in F. Therefore, if $I(x) > 0$ for all x in F, then $I^*(F) > 0$, so that the probability of this set decays exponentially fast. ♠

Exercise 2.4. There is at most one rate function I satisfying (2.1). That is, the rate function (if it exists) is unique. Hint: assume $I_1(x_0) > I_2(x_0)$ for some point x_0. By lower semicontinuity you can find a closed ball around x_0 and a positive δ so that $I_1(x) > I_2(x_0) - \delta$ for all x in that ball. Now use (2.1). ♠

Why the two separate statements for closed and open sets? Or, why separate upper and lower bounds? It is easy to see that the lower bound cannot hold for a general closed set, since it may be too small. Take a sequence $x_1, x_2, \ldots$ of Bernoulli ($\frac{1}{2}$) random variables (Example 1.15), and let $C \stackrel{\triangle}{=} \left\{\sqrt{1/2}\right\}$. Then $\mathbb{P}(z_n \in C) = \mathbb{P}(x_1 + \cdots + x_n = n/\sqrt{2}) = 0$, so that the left-hand side of (2.1l) equals $(-\infty)$, but from (1.9), $\ell\left(\sqrt{1/2}\right) < \infty$! On the other hand, for an open set G,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(z_n \in G) &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(z_n \in \overline{G}\right) \\ &\leq - \inf_{x \in \overline{G}} I(x).\end{aligned}$$

Now, if I is continuous (near the boundary of G), then $\inf_{x\in G} = \inf_{x\in\overline{G}}$, and the upper bound holds for the open set. But in general I may be only lower semicontinuous, and a discontinuity on the boundary of G may invalidate the upper bound. This leads to the following definition.

Definition 2.5. *A set S is called an I-continuity set for a rate function I if*

$$\inf_{x \in S^o} I(x) = \inf_{x \in \overline{S}} I(x).$$

The following is an immediate consequence of Definitions 2.2 and 2.5.

Exercise 2.6. If $z_1, z_2, \ldots$ satisfies a large deviations principle with rate function I and if S is an I-continuity set, then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(z_n \in S) = - \inf_{x \in S} I(x). \quad ♠$$

In general, "liminf" and "limsup" may be different, and a limit may not exist.

Exercise 2.7. Let $x_1, x_2, \ldots$ be i.i.d. Bernoulli ± 1 random variables and define $z_n \stackrel{\triangle}{=} n^{-1}(x_1 + \cdots + x_n)$. Find a set A that is not an I-continuity set. Find a set

S so that $n^{-1} \log \mathbb{P}(z_n \in S)$ possesses no limit. Hint: for A try an interval, for S a single point. ♠

The large deviations principle and the variational formulation imply something about the way rare events happen. If the variational problem has a unique optimal point x^*, then a point far from x^* cannot be optimal. Therefore, rare events happen by z_n staying close to x^*. Here is a precise statement.

Lemma 2.8. *Assume I is a good rate function on a metric space $(\mathcal{X}, d)$. Let F be a closed set and define $I^*(F)$ and $\mathcal{O}$ by (2.2). Then for each ε, there is a $\delta > 0$ so that $x \in F$ and $d(x^*, x) \geq \varepsilon$ for all x^* in $\mathcal{O}$ implies $I(x) \geq I^*(F) + \delta$. If $z_1, z_2, \ldots$ satisfies a large deviations principle with rate function I and if F is an I-continuity set, then*

$$\mathbb{P}\left(d(z_n, x^*) \geq \varepsilon \text{ for all } x^* \in \mathcal{O} \;\middle|\; z_n \in F\right) \leq e^{-n\delta + o(n)}.$$

Proof. Assume first that $\mathcal{O}$ contains exactly one point x^*. For a proof by contradiction, assume that there are points x_i in F with $I(x_i) \leq I(x^*) + 2^{-i}$ but $d(x^*, x) \geq \varepsilon$. Since the set $F \cap \left\{x : I(x) \leq I(x^*) + 1\right\}$ is compact, we can take a convergent subsequence of $\{x_i\}$ with limit x. Then $x \in F$ since F is closed, $d(x, x^*) \geq \varepsilon$ so that $x \neq x^*$, but, by lower semicontinuity, $I(x) \leq I(x^*)$, contradicting the definition of $\mathcal{O}$.

To establish the second claim note that from Bayes' Rule,

$$\begin{aligned}
\mathbb{P}\left(d(z_n, x^*) \geq \varepsilon \;\middle|\; z_n \in F\right) &\leq \frac{\mathbb{P}(d(z_n, x^*) \geq \varepsilon)}{\mathbb{P}(z_n \in F)} \\
&\leq e^{-n(I(x^*)+\delta)+o(n)} e^{nI(x^*)+o(n)} \\
&= e^{-n\delta+o(n)}.
\end{aligned} \tag{2.3}$$

The second inequality in (2.3) comes from the large deviations upper bound applied to the numerator, and the large deviations lower bound and Exercise 2.6 applied to the denominator. The proof for the case when $\mathcal{O}$ is not a single point is relegated to the exercise below. ■

Exercise 2.9. Extend the proof above to the case that $\mathcal{O}$ contains more than one point. ♠

We now give a formal statement of the large deviations principle in its standard, abstract form. We are given a sequence of probability spaces $(\mathcal{X}, \mathcal{F}, \{P_n\})$, where the P_n are all measures on $(\mathcal{X}, \mathcal{F})$. In terms of Definition 2.2, P_n is the "distribution" of z_n. We are also given a rate function I on $\mathcal{X}$.

Definition 2.10. *We say $\{P_n\}$ satisfies a large deviations principle with rate function I if, for every closed set $C \subset \mathcal{X}$, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leq -\inf_{x\in C} I(x), \tag{2.4u}$$

and for every open set $G \subset \mathcal{X}$, we have

Birth-Death processes.

In most of the applications, the jumps of the process are only to the nearest neighbors. One such jump Markov process is the multidimensional birth-death process. We can loosely think of a d-dimensional birth-death process as representing the number of live creatures of d types. At each event-point, at most one individual of each type may be born or die, while the birth (death) rates depend on the current state. However, we may have a simultaneous birth-death case, where one individual from each of several types is born, and at the same time one individual from each of several other types dies. Other possible phenomena are mutation, coagulation, etc. Moreover, for this process the numbers of individuals of a type is not necessarily positive. Thus we have a finite number of jump directions $\vec{e}_1, \ldots, \vec{e}_m$, where the components of the vector $\vec{e}_i$ include 1 s, (-1) s, and 0 s only, and the rate takes the form

$$\lambda(x, y) = \begin{cases} \lambda_j(x) & \text{if } y = x + \vec{e}_j; \\ 0 & \text{otherwise} . \end{cases} \tag{4.8}$$

Most of the processes that arise in queueing applications are of this form. As the most basic example, the queue size of an $M/M/1$ queue is a birth-death process where arrivals are described by $e_1 = 1$, with rate $\lambda_1 = \lambda$, and departures by $e_2 = -1$, with rate $\lambda_2(x) = \mu \mathbf{1}[x > 0]$, where μ is the service rate.

A general jump Markov process need not be defined on a discrete state-space. For example, if $\mathcal{X} = \mathbb{R}^d$, we can construct such a process by making it jump from state x after an exponential time with mean $(\lambda(x))^{-1}$ and, when it jumps, have its next position distributed according to some distribution F_x. However, unless the distribution F_x is supported by at most a countable number of points, we lose the interpretation of this process as jumping according to the smallest exponential random variable with rate $\lambda(x, y)$. Definitions (4.5)–(4.6) still make sense, with sums in (4.5) replaced by integrals, and they are equivalent. In fact, it is easy to see that the same calculation establishes the equivalence for the general jump Markov case if the support of F_x is countable; we just have to invoke boundedness to interchange the order of limit and summation.

This trick of switching between interpretations of the process is fairly useful. What we have discovered is that, in some sense, only the first jump counts; after that the process "restarts." Let us specialize to processes of the type described in (4.8) and derive a general tool for approximations from one of the representations. So, consider a process with generator

$$L f(\vec{x}) = \sum_{i=1}^{k} \lambda_i(\vec{x}) \left(f(\vec{x} + \vec{e}_i) - f(\vec{x}) \right) . \tag{4.9}$$

Assume that the maximal rate of jump out of any state is finite:

$$\bar{\lambda} \stackrel{\triangle}{=} \sup_{\vec{x}} \sum_{i=1}^{k} \lambda_i(\vec{x}) < \infty.$$

Thus M_f is a martingale (in fact, it is constant). But this implies that L is indeed the generator of the process x. Since the generator determines the distribution of the process, we conclude that any process whose generator is a first-order differential operator is deterministic (except for possible random initial conditions), and Corollary 4.15 is established. ■

For the scaled Poisson process $z_n(t)$ we conclude that, under the appropriate technical conditions, the limit $z_\infty(t)$ of the process $N(t)$ scaled by n is a deterministic process. Since here $g \equiv \lambda$ the limit $z_\infty(t)$ must satisfy

$$\dot{z}_\infty(t) = \lambda \qquad \text{or} \qquad z_\infty(t) = z_\infty(0) + \lambda \cdot t \ .$$

For *fixed* t, we could obtain this result from the law of large numbers. However, our result is stronger—it deals with the convergence of the process. Therefore, in order to make this convergence result precise, we need to define a space of paths that is appropriate for the paths of the Poisson process, a notion of convergence in that space, and specify in what (probabilistic) sense the sequence converges. Using this technique one typically obtains "convergence in distribution" which, in this context, is called "weak convergence" (Definition A.87).

The process we shall concentrate on throughout the rest of the book is a birth-death process with precisely the same scaling as in (4.13). So, let $\vec{x}(t)$ be a birth-death process with generator as defined in (4.5)–(4.8). The scaled process $\vec{z}_n$ is the jump Markov process such that

$$\vec{z}_n(0) \stackrel{\mathcal{L}}{=} \frac{1}{n}\vec{x}(0) \quad \text{and} \quad L_n f(\vec{x}) = \sum_{i=1}^{k} n\lambda_i(\vec{x}) \left(f\left(\vec{x} + \frac{\vec{e}_i}{n}\right) - f(\vec{x})\right). \quad (4.14)$$

Note that a generator defines the *distribution* of the process—not the paths. If the λ_j are constant (or more generally if $\lambda(n\vec{x}) = \lambda(\vec{x})$ for all $n \geq 1$), then we can define $\vec{z}_n$ by scaling time and space (as was done for the Poisson process). But this is not true for the general case; see also Chapter 5.

Exercise 4.16. Let $\vec{z}_n$ be defined through (4.14) and assume $\vec{z}_n(0) = 0$. Assume that $\log \lambda_i$ are bounded and Lipschitz continuous. Show that $\vec{z}_n \Rightarrow \vec{z}_\infty$, where

$$\vec{z}_\infty(0) = 0, \quad \text{and} \quad \frac{d\vec{z}_\infty(t)}{dt} = \sum_{i=1}^{k} \lambda_i(\vec{z}_\infty(t))\vec{e}_i .$$

Hint: use (4.12), write L as $L_n - (L - L_n)$, and bound the "error term," noting that f is smooth and bounded.

Generating martingales.

Using Theorem 4.12 we can generate other martingales. Our first goal is to derive (heuristically) the famous "exponential martingale" (4.16) below. We start with a Markov process x and its generator L. Fix a nice (smooth, bounded) function q and consider the two-dimensional process

$$z(t) \stackrel{\triangle}{=} \left\{ x_t, \exp\left(- \int_0^t q(x_s)\, ds \right) \right\} \ .$$

It is easy to see (at least formally) from Definition 4.6 that the generator $\tilde{L}$ for the process z is given by

$$\tilde{L} f(x, y) \stackrel{\triangle}{=} L f(x, 1) - q(x) \cdot \frac{d}{dy} f(x, y) . \tag{4.15}$$

[Use Taylor expansion near the initial point $(x, 0)$ and expand the exponent to a series.] Consider, for example, a function of the form $f(x) \cdot g(y)$, where $g(y) \equiv y$. Then

$$\tilde{L}\Big(f(x) \cdot g(y) \Big) = L f(x) - q(x) f(x) .$$

Now let $q \stackrel{\triangle}{=} L f / f$ (assume f is bounded away from 0). Then $\tilde{L}\Big(f(x) \cdot g(y) \Big) = 0$. Theorem 4.12 and the definition (4.10) now imply that

$$\begin{aligned} M_{f \cdot g}(t) &\stackrel{\triangle}{=} (f \cdot g)(z_t) - \int_0^t \tilde{L}\Big(f \cdot g \Big)(z_s)\, ds \\ &= f(x_t) \exp\left(- \int_0^t \left(\frac{L f(x_s)}{f(x_s)} \right) ds \right) \end{aligned} \tag{4.16}$$

is a martingale, called an exponential martingale associated with the process x. For a precise derivation of (4.16) see §B.4.

Exercise 4.17. Show that for a birth-death process (4.8), the exponential martingale corresponding to $f(x) = e^{\langle \theta, x \rangle}$ takes the form

$$M_t = \exp\left[\langle x_t, \theta \rangle - \int_0^t \sum_{i=1}^{k} \lambda_i(x_s) \left(e^{\langle \theta, \vec{e}_i \rangle} - 1 \right) ds \right] . \tag{4.17}$$

♠

Change of measure.

In the proof of Chernoff's Theorem we obtained the lower bound on the probability of a rare event by a change of measure. In other words, we constructed a different distribution, or measure $\tilde{\mathbb{P}}$, and wrote, for a set or event A

$$\mathbb{P}(A) = \int_A d\mathbb{P} = \int_A \left(\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}} \right) d\tilde{\mathbb{P}}.$$

$d\mathbb{P}/d\tilde{\mathbb{P}}$ is the Radon-Nikodym derivative; see §A.4. We would like to use the same idea for Poisson and related processes. Here is a heuristic derivation of a formula for $d\mathbb{P}/d\tilde{\mathbb{P}}$ when $\mathbb{P}$ and $\tilde{\mathbb{P}}$ represent Poisson processes. A rigorous treatment is given in §B.4.

Suppose that we observe a Poisson process with rate λ over a time interval $[0, T]$, and we see n events, at times $t_1 < t_2 < \ldots < t_n$ where $0 < t_1$ and $t_n < T$. From Definition 4.1, the probability that no event would occur in $[0, t_1]$ is $e^{-\lambda t_1}$.

the interval $[0, T]$ to J equal subintervals of size Δ. The probability that $\vec{z}_n(t)$ is close to $\vec{r}(t)$ for all $0 \le t \le T$ is approximated by the product of the probabilities that it remains close on each interval: it starts near $\vec{r}(j\Delta)$ and goes in the same direction as $\vec{r}$, namely in the direction $\vec{r}\,'(j\Delta)$ for each j. (Since we are being informal, in the following explanation we hope the reader doesn't mind a conditioning on $\vec{z}_n(js) = x$ being written as, e.g., $\mathbb{P}_{\vec{x}}(\sup_{js<t<T} \ldots)$. The notation $\mathbb{P}_{\vec{x}}$ usually means starting at time 0; here, we mean whatever time is appropriate.) In more detail, by the Markov property, for any $0 < s < T$,

$$\begin{aligned}
\mathbb{P}_{\vec{r}(0)} & \left(\sup_{0 \le t \le T} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right) \\
&= \mathbb{E}_{\vec{r}(0)} \left(\mathbf{1}\left[\sup_{0\le t\le s} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right] \mathbf{1}\left[\sup_{s\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right] \right) \\
&= \mathbb{E}_{\vec{r}(0)} \left(\mathbf{1}\left[\sup_{0\le t\le s} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right] \right. \\
&\qquad \left. \times\ \mathbb{E}_{\vec{z}_n(s)} \left(\mathbf{1}\left[|\vec{z}_n(s) - \vec{r}(s)| < \delta \right] \mathbf{1}\left[\sup_{s\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right] \right) \right) \\
&\le \mathbb{E}_{\vec{r}(0)} \mathbf{1}\left[\sup_{0\le t\le s} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right] \\
&\qquad \times \sup_{|\vec{x}-\vec{r}(s)|<\delta} \mathbb{E}_{\vec{x}} \mathbf{1}\left[\sup_{s\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < 2\delta \right].
\end{aligned}$$

Iterating on the J subintervals,

$$\begin{aligned}
\mathbb{P}_{\vec{r}(0)} \left(\sup_{0\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right) &\le \mathbb{P}_{\vec{r}(0)} \left(\sup_{0\le t\le \Delta} |\vec{z}_n(t) - \vec{r}(t)| < \delta \right) \\
&\quad \times \sup_{|\vec{x}-\vec{r}(\Delta)|<\delta} \mathbb{P}_{\vec{x}} \left(\sup_{\Delta\le t\le 2\Delta} |\vec{z}_n(t) - \vec{r}(t)| < 2\delta \right) \\
&\quad \vdots \\
&\quad \times \sup_{|\vec{x}-\vec{r}((J-1)\Delta)|<(J-1)\delta} \mathbb{P}_{\vec{x}} \left(\sup_{(J-1)\Delta\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < J\delta \right).
\end{aligned} \tag{5.6}$$

Since we assume that $\vec{z}_n$ and $\vec{r}$ are close at the beginning of each subinterval, we need to find the probability that a change $\Delta\vec{z}_n(t)$ in $\vec{z}_n(t)$ is about equal to the change $\Delta\vec{r}(t) \approx \vec{r}\,'(t)\Delta t$ in $\vec{r}(t)$. Now over a small interval of time, the jump rates $\lambda_i(\vec{z}_n(t))$ do not change much. Therefore, Proposition 4.5(ii) (extended to