Analysis of Single-Molecule Kinesin Assay Data by Hidden Markov Model Filtering

by

David Brian Walton

A Dissertation Submitted to the Faculty of the

Graduate Interdisciplinary Program
in Applied Mathematics

In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy

In the Graduate College

The University of Arizona

2 0 0 2

Get the official approval page
from the Graduate College
before your final defense.

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: _____

Acknowledgments

I'd like to thank everyone who has helped me through my education. Special thanks go to my advisor and mentor, Joe Watkins, who encouraged me in my studies of stochastic processes, and who provided an example of interdisciplinary and collaborative research. Also, I'd like to thank Koen Visscher for working with me so willingly, providing access to experimental data, and giving important feedback on a physical perspective to my research problem. I also am grateful for the general support of the faculty at The University of Arizona, particularly in the Mathematics and Physics departments, who worked individually with me. Some, with whom I interacted more extensively, include Johann Rafelski, Alain Goriely, Tom Kennedy, Greg Eyink, and Jan Wehr. Thanks to Jens Timmer for a timely outside review. Thanks to class mates and office mates for many helpful discussions, questions, and answers.

I'd also like to acknowledge those who provided a lot of background support. This includes the staff members with the Program in Applied Mathematics, who helped me focus on my research rather than on paperwork. The computer staff of the Mathematics department was very helpful in obtaining computational support, as was Juan Restrepo for the use of his cluster. I'd also like to thank Michael Tabor, the program head, for early mentoring and general support. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation. I'm grateful to the NSF (and, therefore, to the Congress and the citizens of the United States of America) for funding, as some of this research was also performed while I was a fellow in the Biology, Mathematics, and Physics IGERT initiative, also an NSF program. As a participant in the Biomathematics seminar and IGERT laboratory course, I'm grateful for the interactions with other students and scientists in a diverse academic setting for strengthening my views of interdisciplinary research.

Finally, I wish to thank my true support system. I'm particularly grateful to my family. My wife, Heather, was an ongoing pillar of strength and motivation; my first two sons, Michael and Nathaniel, who entered our family in Arizona, were a source of joy and a welcome release from academics. Our extended families, particularly our parents and grandparents, were also very supportive, both emotionally and temporally. I'm grateful to my God for answered prayers, spiritual strength, and a sense of ongoing peace in life, and to the local wards of the Church of Jesus Christ of Latter-day Saints for opportunities I had there to reach outside of myself by serving others.

To all—thank you.

Table of Contents

Table of Contents—*Continued*

List of Figures

List of Figures—*Continued*

List of Tables

Abstract

Observations of the position of a microscopic bead attached to a single kinesin protein moving along a microtubule contains detailed information about the position of the kinesin as a function of time, although this information remains obscured because of the fluctuations of the bead. The theory of hidden Markov models suggests a possible theoretical framework to analyze these data with an explicit stochastic model describing the kinesin cycle and the attached bead. We model the mechanical cycle of kinesin using a discrete time Markov chain on a periodic lattice, representing the microtubule, and model the position of the bead using an Ornstein-Uhlenbeck autoregressive process. We adapt the standard machinery of hidden Markov models to derive the likelihood of this model using a reference measure, and use the Expectation-Maximization (EM) algorithm to estimate model parameters. Simulated data sets indicate that the method does have potential to better analyze kinesin-bead experiments. However, analysis of the experimental data of Visscher et al. (1999) indicates that current data sets still lack the time resolution to extract significant information about intermediate states. Considerations for future experimental designs are suggested to allow better hidden Markov model analysis.

Analysis of Single-Molecule Kinesin Assay Data by Hidden Markov Model Filtering

David Brian Walton, Ph.D.
The University of Arizona, 2002

Director: Joseph C. Watkins

Observations of the position of a microscopic bead attached to a single kinesin protein moving along a microtubule contains detailed information about the position of the kinesin as a function of time, although this information remains obscured because of the fluctuations of the bead. The theory of hidden Markov models suggests a possible theoretical framework to analyze these data with an explicit stochastic model describing the kinesin cycle and the attached bead. We model the mechanical cycle of kinesin using a discrete time Markov chain on a periodic lattice, representing the microtubule, and model the position of the bead using an Ornstein-Uhlenbeck autoregressive process. We adapt the standard machinery of hidden Markov models to derive the likelihood of this model using a reference measure, and use the Expectation-Maximization (EM) algorithm to estimate model parameters. Simulated data sets indicate that the method does have potential to better analyze kinesin-bead experiments. However, analysis of the experimental data of Visscher et al. (1999) indicates that current data sets still lack the time resolution to extract significant information about intermediate states. Considerations for future experimental designs are suggested to allow better hidden Markov model analysis.

Chapter 1
Introduction

Kinesin, a motor protein originally discovered as a primary player in fast axonal transport (Vale et al., 1985a,b), has been actively studied to understand better how chemical energy is converted into mechanical energy. Kinesin uses energy released through the hydrolysis of adenosine triphosphate (ATP) to pull membrane-bound organelles along a microtubule, a filamentary and dynamic lattice structure made from $\alpha\beta$-tubulin dimers (Vale et al., 1985a). Kinesin moves vesicles and other membrane bound organelles toward the microtubule's fast-growing, or plus end, which in neurons corresponds with the anterograde direction (Hirokawa et al., 1991). The mechanical stepping remains tightly coupled to the chemical cycle, with a net gain of position of approximately 8.2 nm, the length of a single $\alpha\beta$-tubulin dimer, corresponding to a single ATP hydrolysis event (Schnitzer and Block, 1997; Hua et al., 1997; Coy et al., 1999) over a wide range of opposing forces and ATP concentrations (Visscher et al., 1999). In addition, structural analyses such as X-ray crystallography (Kozielski et al., 1997; Marx et al., 1998), electron paramagnetic resonance and cryo-electron microscopy (Rice et al., 1999), fluorescence microscopy (Vale et al., 1996; Rice et al., 1999; Sosa et al., 2001; Rosenfeld et al., 2001), and simulated annealing (Wriggers and Schulten, 1998; Xing et al., 2000) support the existence of different physical conformations for kinesin in different chemical states. However, the detailed relation between the chemical cycle and the conformational changes which lead to movement remains unknown.

As a processive motor, repeating many biochemical cycles prior to releasing from the microtubule, kinesin has been subject to a wide range of single-molecule studies (Gelles et al., 1988; Howard et al., 1989; Block et al., 1990; Svoboda et al., 1993;

Romberg and Vale, 1993; Kuo and Sheetz, 1993; Svoboda and Block, 1994b; Malik et al., 1994; Hunt et al., 1994; Meyhofer and Howard, 1995; Vale et al., 1996; Schnitzer and Block, 1997; Hua et al., 1997; Coy et al., 1999; Visscher et al., 1999; Schnitzer et al., 2000). Several of these (Block et al., 1990; Svoboda et al., 1993; Svoboda and Block, 1994b; Schnitzer and Block, 1997; Visscher et al., 1999; Schnitzer et al., 2000) have used an optical tweezers to manipulate microscopic glass beads with adsorbed kinesin which move along microtubules attached to a microscope slide coverslip. By controlling the concentration of kinesin, experimentalists can prepare beads with only a single kinesin molecule attached. When the attached kinesin, in the presence of ATP, binds to the microtubule, it undergoes its chemical and mechanical cycle, pulling the bead along. By measuring the position of the bead, experimentalists indirectly obtain information regarding the position of the kinesin molecule. Under regulated conditions, the noise of the bead position measurement is sufficiently small that individual stepping events can be directly visualized (Svoboda et al., 1993). Previous analysis of these assays have focused on determining the stall force, the average velocity, and the randomness, a measure of variation of kinesin movement about the mean velocity. However, analysis of the data has not yet been performed to incorporate the details of the measurement, such as through modeling the kinetic cycle of kinesin and the nature of the noisy observation.

Hidden Markov models provide a promising tool to analyze the single-molecule data according to their details while incorporating models to describe the kinesin mechanochemical cycle and the noisy bead observation. Briefly, a hidden Markov model is a stochastic model describing two related randomly evolving time sequences, or stochastic processes: a state process and an observation process. The state process is assumed to be a Markov process that is not directly observed. The observation process is generated from the state process, and yet obscures the true state because the observation includes independent random fluctuations. Hidden Markov models have been used successfully in a variety of areas including speech processing (Levinson

et al., 1983; Rabiner, 1989), DNA sequence analysis (Churchill, 1989), and ion channel analysis of neuron firing patterns (Chung et al., 1990; Fredkin and Rice, 1992), as well as a recent application to single-molecule studies of the motor protein myosin (Smith et al., 2001).

This dissertation describes an implementation of hidden Markov model filtering to analyze single-molecule kinesin-bead assay data. We focus on the experimental design and resulting data of Visscher et al. (1999). Given a number of data sets for each experimental condition under consideration, we wish to extract information from the data regarding the position of kinesin during its enzymatic cycle by analyzing the position of the bead. It will be necessary to determine the number of internal states supported by the data and to estimate various model parameters. We will model the mechanochemical cycle of kinesin as a Markov process, while we model the observed bead positions with an autoregressive noise process which depends on the kinesin position. Such a model allows the use of hidden Markov model filtering techniques. Simulated results indicate that such a model holds promise. Unfortunately, the experimental data were not quite adequate for the desired analysis, primarily due to an inability to identify the correct step-size of the underlying microtubule lattice.

This dissertation is organized as follows. This chapter will continue with more in depth background material regarding the biology of kinesin, the single-molecule bead assays, and the classical theory of hidden Markov models. The second chapter will focus on the development of the hidden Markov model which will be used to model the kinesin-bead system of the single-molecule assay, and will develop the theoretical framework for the filtering and analysis. The third chapter will focus on the effectiveness of the hidden Markov model filtering for simulated experimental data where the generating model is known. The fourth chapter will discuss the results of applying the hidden Markov model filtering techniques on the experimental kinesin data and the problems with the data arising in this analysis. Finally, we conclude in chapter five by discussing improvements in the experiments, in the modeling, and

in the algorithms so that valid experimental analysis can be effectively carried out in the future.

1.1 Biology and Chemistry

Kinesin has been identified as the protein primarily responsible for fast anterograde axonal transport in neurons (Vale et al., 1985a,b; Hirokawa et al., 1991). Neurons are comprised of three major regions: the cell body, the dendrites, and the axon. Electrical signals are received through the dendrites, directed to the cell body, and when the incoming signal is sufficiently strong, the neuron fires a new signal which is transmitted along the axon. Different neurons have axons of widely varying lengths, with some axons shorter than one millimeter and others longer than a meter. The cell body includes the nucleus, which contains the genetic material regulating protein assembly, and so protein assembly occurs in the cell body. In order for the axon to remain functional, these products must be carried outwardly along the axon, which movement is known as anterograde axonal transport. Axonal transport can be categorized into slow transport of cytosol ($\sim$1 mm/day) and fast transport of membrane bound organelles ($\sim$400 mm/day) (Vallee and Bloom, 1991). Diffusive processes involve time scales that grow quadratically with the distance traveled, and as such are incompatible with fast transport. Instead, fast axonal transport must be mediated by a directed motion, and this motion is performed by kinesin.

Kinesin travels along microtubules, which are typically composed of thirteen parallel tubulin protofilaments arranged to form an extended cylinder approximately 25 nm in diameter. Each tubulin protofilament has a lattice structure, made from alternating α- and β-tubulin, and the lattices of adjacent protofilaments have a slight offset. The $\alpha\beta$-tubulin dimer has a length of 8.2 nm, establishing the lattice spacing of the microtubule. Microtubules are naturally very dynamic molecules, with one end—the plus end—growing and shrinking much faster relative to the opposite end.

Kinesin is a directional motor protein, in that it always travels toward the plus end of the microtubule, which corresponds to the anterograde direction in axons. In addition, experimental evidence suggests that kinesin follows an individual protofilament of a microtubule, rather than crossing to other protofilaments (Ray et al., 1993).

Following the discovery of the motor protein kinesin in axonal transport, many related microtubule-associated motor proteins have also been identified. Dynein is a microtubule-associated motor protein which provides retrograde axonal transport (Schnapp and Reese, 1989), and which had been earlier associated with the movement of cilia and flagella (Gibbons, 1965). In addition, proteins closely related to kinesin, known as the kinesin superfamily of proteins, have been identified for a variety of different microtubule associated motor functions (Muresan, 2000), and some of which travel in opposite directions (Woehlke and Schliwa, 2000). One important kinesin-related protein is non-claret disjunctional (ncd), which is structurally homologous to kinesin, but which travels toward the minus end (Walker et al., 1990).

Conventional kinesin is a hetero-tetramer, meaning it is composed of four polypeptide chains which associate tightly to form the kinesin protein. Two identical heavy chains ($\sim$125 kDa each) form two globular heads which bind to the microtubule. These heavy chains join at a neck region, forming an extended alpha-helical coiled coil known as the stalk. Two light chains associate to the stalk and serve to bind the kinesin with membrane-bound organelles which will be transported along the microtubule. Each of the two heads contains a motor domain. The processive nature of kinesin—it remains bound to a microtubule for multiple cycles—has been experimentally linked to the presence of both heads (Young et al., 1998; Hancock and Howard, 1998). The more traditional interpretation of these results is that the two heads of kinesin alternately drive the motion, participating in a hand-over-hand action (Kuo et al., 1991), although the possibility of other mechanisms remain (Hua et al., 2002).

Kinesin acts as an ATPase, releasing stored energy through the hydrolysis of adenosine triphosphate (ATP). The globular head of the heavy chain includes an ATP

binding site, in addition to domains associated with microtubule binding. Movement by kinesin along a microtubule is driven by this ATPase cycle, converting the chemical energy stored in ATP into mechanical energy of directed motion. The presence of a microtubule enhances the ATPase rate by $\sim 10^3 - 10^4$ fold (Hackney, 1988). Furthermore, in the absence of attached cargo, the stalk region interacts with the head to inhibit motion (Friedman and Vale, 1999). Together, these two properties minimize the amount of futile hydrolysis—the hydrolysis of ATP without performing transport—reserving active hydrolysis for periods when transport of cellular products will result.

The chemical cycle of ATP hydrolysis is coupled to the mechanical motion of kinesin. The basic chemical cycle involves the enzyme-catalyzed breaking of a phosphate bond and the subsequent release of both an ortho-phosphate P_i and the resulting adenosine diphosphate (ADP):

$$K + ATP \leftrightarrow K \cdot ATP \leftrightarrow K \cdot ADP \cdot P_i \leftrightarrow K + ADP + P_i. \tag{1.1}$$

Experimental evidence through the use of crystallography (Kozielski et al., 1997; Marx et al., 1998), fluorescence and electron microscopy (Vale et al., 1996; Rice et al., 1999; Sosa et al., 2001; Rosenfeld et al., 2001) suggests that these stages correspond to various physical conformations of the involved kinesin attached to a microtubule. When ATP binds to kinesin, the kinesin adopts a rigid conformation with the detached head positioned in the direction of forward motion (Rice et al., 1999). When the phosphate is released but ADP remains attached, the kinesin enters a highly dynamic state (Rice et al., 1999; Sosa et al., 2001). Identifying the precise sequence of conformations and the exact relationships with the chemical cycle remains one of the puzzling mysteries of kinesin motion.

Although the details of the conformations in the mechanical cycle and the relationship with the chemical cycle remain uncertain, a number of relationships between the chemical and mechanical cycles have been determined. First of all, the coupling

between the two cycles is tight, with evidence suggesting that one 8.2 nm step corresponds exactly with one hydrolysis cycle of ATP (Schnitzer and Block, 1997; Hua et al., 1997; Coy et al., 1999) over a wide range of forces (Visscher et al., 1999). ADP-bound kinesin is weakly bound to the microtubule, while ATP-bound kinesin and empty kinesin have a strong association with the microtubule (Crevel et al., 1996; Rosenfeld et al., 1996). These different binding strengths between the kinesin and the microtubule contribute to the importance of the need for two heads, with one head staying in a state of high associativity while the other head dissociates according to the different chemical states of each head, leading to the high duty cycle of kinesin that corresponds to processivity.

1.2 Single-Molecule Bead Assays

In order to gain better insight into the nature of kinesin's mechanical cycle and its relationship to the chemical hydrolysis cycle, experimentalists have developed a variety of single-molecule experiments. The common objective of these experiments is to capture information about the individual events that combine to create the overall motion of kinesin. Some of the experiments include microtubule displacement by a single kinesin protein (Howard et al., 1989), the precise measurement of force exerted on a microtubule by a single kinesin (Meyhofer and Howard, 1995), fluorescence microscopy tracking of kinesin (Vale et al., 1996), and the use of kinesin-coated beads to simulate the pulling of cargo by kinesin along a microtubule (Block et al., 1990). We will focus on the single-molecule bead assays, as this dissertation focuses on analyzing the data from a particular set of experiments.

1.2.1 Experimental Setup

Although the data produced from single-molecule bead assays essentially depend only on the interactions of the kinesin with the microtubule and the bead with the optical

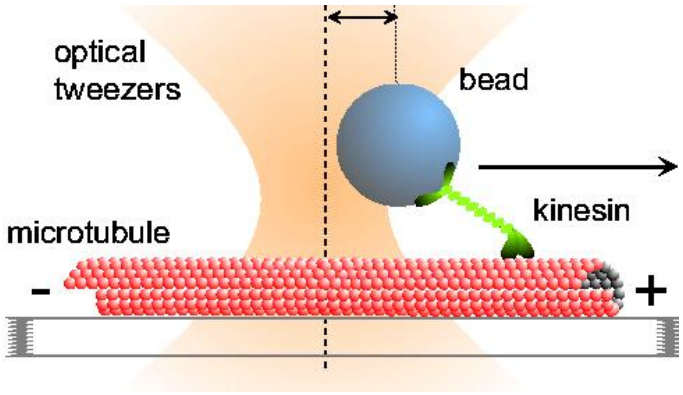


FIGURE 1.1. Illustration of the VSB experimental setup for single-molecule bead assays using optical tweezers. The distance between the position of the bead and the center of the trap is held constant through the use of a feedback loop. Courtesy: K. Visscher.

tweezers, we first briefly discuss some aspects of the preparation required in order to help the reader have a better understanding of the overall experimental setup (Block et al., 1990; Svoboda et al., 1993; Svoboda and Block, 1994b; Schnitzer and Block, 1997; Visscher et al., 1999). The essential idea is that a bead attached to kinesin is visible, whereas an individual kinesin molecule is not. Optical tweezers also treat the bead as a handle, which can move kinesin directly to a microtubule as well as exert forces on the kinesin as it moves. Figure 1.1 illustrates the basic components of the experiment.

Before the observations can begin, a flow cell must be prepared with microtubules and kinesin-coated beads. A flow cell is constructed from a microscope slide and a polylysine-treated coverslip, to which microtubules will adhere when flowed in. Microtubules, which are naturally unstable dynamic structures, are prepared and stabilized through the use of the drug taxol. The solution with microtubules is introduced to the flow cell quickly so that the microtubules will have a tendency to enter the channel straight and parallel to the channel. The excess solution is flushed out with a buffer, leaving several microtubules bound to the coverslip. The surface is subsequently blocked with the protein casein, which will prevent kinesin-coated beads from

sticking to it. Next, kinesin-coated beads must be prepared. Microscopic silica beads with a uniform diameter of 0.5 microns and a solution with kinesin are incubated together at a ratio of concentrations such that, on average, only one kinesin can move a bead at a time. The kinesin-bead solution is combined with an experimentally controlled concentration of ATP, and this solution is added to the flow cell, ready for observation.

We now turn to a discussion of the nature of the optical tweezers (Svoboda and Block, 1994a). The most basic ingredients are the microscope objective lens and a laser. When a laser beam is sent through a microscope objective lens with a high numerical aperture, the beam is very rapidly focused to a narrow waist, or diffraction limited spot. The electromagnetic field polarizes the dielectric beads when illuminated by the laser light, and the steep gradient in the intensity leads to a restoring force which pulls the bead to the region with the highest intensity. Near the center of the trap, the potential energy landscape is very well modeled as a harmonic potential. Thus, for displacements that are not too large, the trap will exert a restoring force that is proportional to the displacement, exactly like a standard spring obeying Hooke's law. This spring-like behavior is empirically observed to extend over $\sim$200 nm, allowing for well-calibrated force calculations based on the displacement of the bead from the center of the trap and the spring constant, or stiffness, of the trap. Outside of this linear regime, the force-displacement relationship becomes nonlinear so that at sufficiently large distances, the trap no longer even attracts the beads. Control mechanisms on the optical system can move the position of the tweezers in all three dimensions, allowing for precise manipulation of the bead in solution. By also incorporating a feedback system, the position of the trap can be controlled so that the displacement between the trap center and the position of the bead remains fixed, establishing a force clamp that maintains a constant load on the bead as it is pulled along a microtubule by kinesin (Visscher and Block, 1998).

Finally, with the tweezers providing a method to move beads to the vicinity of the

microtubules, as well as to exert calibrated forces on the beads, scientists can perform experiments where kinesin moves along a microtubule with a particular load and under regulated chemical conditions, such as the concentration of ATP. To quantify these motions, the position of the bead is recorded using laser tracking methods. The position of the bead, tethered by the extended stalk region of a kinesin protein, fluctuates subject to the forces exerted by the tweezers, the stalk, and the fluid medium. The stalk has elastic properties as well, with an elastic restoring force that is a nonlinear function of extension (Svoboda and Block, 1994b). One of the important benefits of using a force clamp is that by applying a constant force on the bead with minimal fluctuations, the tension on the kinesin also remains constant with correspondingly small fluctuations. As a result, the stalk similarly maintains a reasonably constant extension. Consequently, when a force clamp is used, an observed displacement in the position of the bead represents the same displacement in the position of kinesin—except for the nagging problem of thermal fluctuations by the bead in solution. Figure 1.2 shows a typical sample of the bead observation, coming from an experiment with a constant load of 3.5 pN and an ATP concentration of 100 μM.

1.2.2 Experimental Results and Previous Analysis

In 1999, Visscher, Schnitzer, and Block (VSB) first published their results of implementing a single-molecule bead assay for kinesin motility using a force clamp, as described above. This experiment tracked the position of the bead using laser tracking at a rate of approximately 20 kHz for updating the position of the trap to create the force clamp, and the position of the bead was recorded at the ten-fold slower rate of approximately 2 kHz. The design of the experiment included performing assays over a range of ATP concentrations ([ATP]) for three levels of the force, with [ATP] ranging from 1 μM to 2 mM and force levels of 1.05 pN, 3.59 pN, and 5.63 pN. In

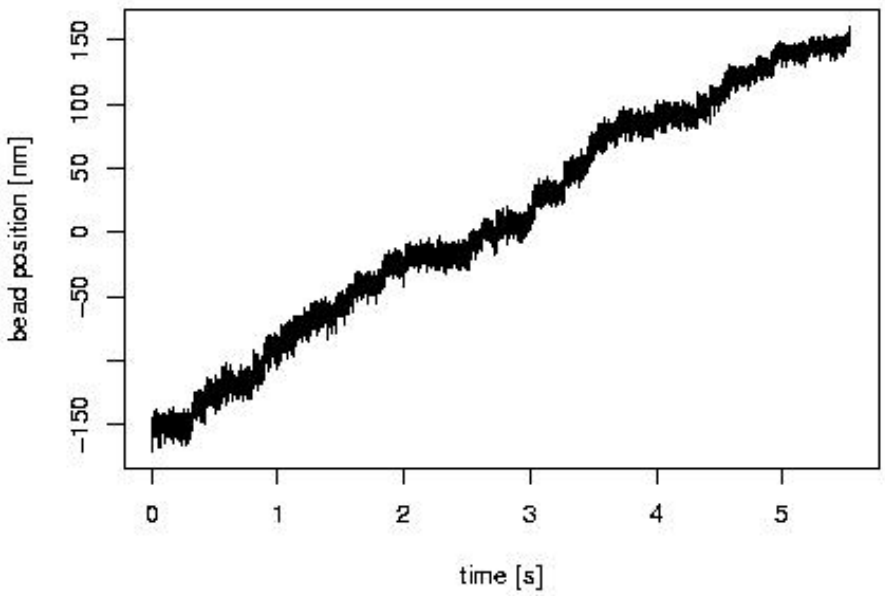


FIGURE 1.2.)
showing the bead position.] A typical data set for the bead position of a single-molecule bead assay. The data shown come from the VSB experiments, with a load of 3.5 pN and 100 μM [ATP].

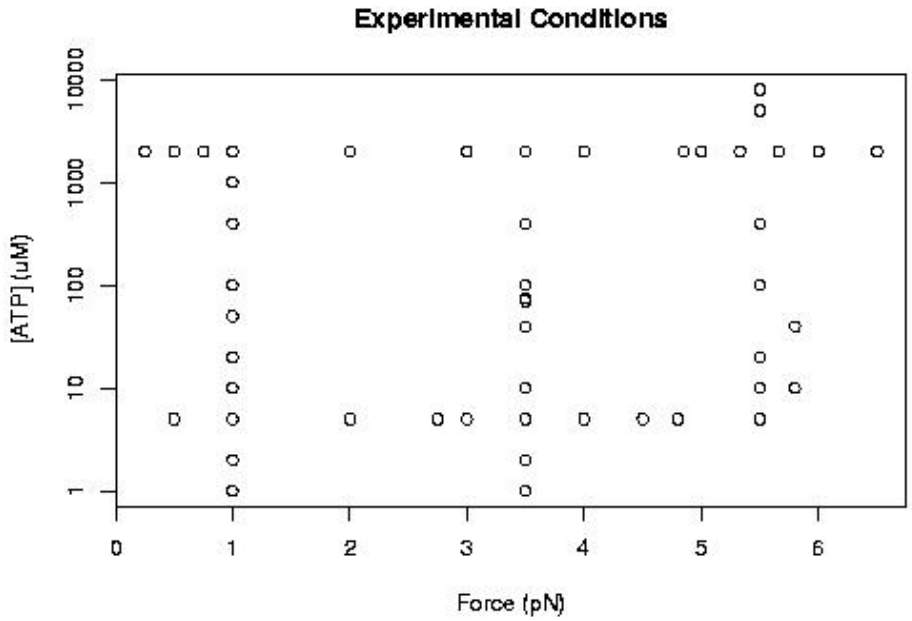


FIGURE 1.3. A scatter-plot illustrating the ATP concentration and externally applied load that characterize the experimental measurements of the VSB experiment for the determination of mean rate of kinesin motion.

addition, for the ATP concentrations of 5 μM and 2 mM, assays were performed for a number of additional intermediate force levels. The data were recorded for three main purposes: to determine how the rate of motion depends on the experimental conditions of ATP concentration and the opposing force, to measure more precisely the stall force of kinesin under different ATP concentrations, and to examine the processivity of kinesin by quantifying the rate of detachment (Schnitzer et al., 2000). The data collected to determine the rate of motion also was used to determine the variability of motion known as randomness, which will be more precisely defined later, and includes multiple recordings of kinesin moving along the microtubule for each experimental condition. Figure 1.3 provides a graphical representation of the experimental conditions at which measurements were recorded for the determination of the rate of motion.

Because the hidden Markov model analysis will focus on the dynamics of kinesin motion, we pay particular attention to the methods used to characterize the rate of motion. The first fundamental quantity for describing motion is the velocity at which kinesin moves along the microtubule. It is observed that kinesin does not move at a constant velocity, but instead persists at specific positions along the microtubule for a random amount of time and then makes a transition to another site. Accordingly, the velocity is actually an average velocity. Using the assumption that kinesin undergoes a chemical and mechanical cycle which ends in a state equivalent to the initial state, but offset by one site on the microtubule lattice, one can represent the random time required to complete this full cycle as the random variable T, with a mean cycle time of $E[T] = \tau$, where E represents the expectation operator. The theoretical mean velocity is defined as $v = d/\tau$, where $d \approx 8.2$ nm is the spacing of the lattice. The second fundamental quantity used to characterize the rate of motion, known as the randomness, characterizes the variation in motion. The randomness, r, a dimensionless parameter, is also defined in terms of the random variable T, as the ratio of the variance of T, σ_T^2, to the square of the mean (Svoboda et al., 1994; Schnitzer and Block, 1995):

$$r = \frac{\mathrm{Var}[T]}{E[T]^2} = \frac{\sigma_T^2}{\tau^2} \tag{1.2}$$

We note that this is actually just another name for the square of the coefficient of variation of the random variable T. Consequently, characterization of the mean velocity and the randomness parameter characterize the first two moments of the random time required to complete a mechanochemical cycle.

Due to the inherent difficulty in identifying individual steps in the presence of noise and discretized sampling, experimental calculations do not involve any direct estimates of the random time T. The data used for analyzing motion are composed of many files, each of which contains a record of a single kinesin run. Each such run consists of a sequence of a time and a corresponding bead position, with a time delay

between recordings of approximately 0.5 ms. To compute the average velocity for a single run, the simplest and most direct approach is to take the difference between the last and first observations and divide by the total time elapsed. However, apparently to deal with the variability in how long kinesin remains at a given position and the relatively short duration of individual runs, a filtered average is computed instead (Schnitzer and Block, 1997). For clarity, we let $y_1, \ldots, y_K$ represent the observations at the times $t_1, \ldots, t_K$, with a time spacing of Δt. For each integer $n \geq 1$, one computes the average n-lag pairwise displacements:

$$\overline{\Delta_n y} = \frac{1}{K-n} \sum_{i=1}^{K-n} (y_{i+n} - y_i). \tag{1.3}$$

The filtered average velocity $\bar{v}$ for a single run is computed by determining the slope of a line fit of these average pairwise displacements versus time:

$$\overline{\Delta_n y} = \bar{v} \cdot (n\Delta t) + \text{residuals}. \tag{1.4}$$

Finally, the overall average velocity, v, is the mean of the average velocities for all of the runs, $v = \langle \bar{v} \rangle$.

The randomness parameter is also experimentally computed based on an average variation in position, rather than from a direct estimation of the statistics of T (Schnitzer and Block, 1995). In particular, if we let $N(t)$ represent the number of complete cycles that have occurred up through time t, and each cycle takes an independent random time with a distribution identical to T, then $N(t)$ is a stochastic process with the following limit theorem for the randomness:

$$r = \frac{\text{Var}[T]}{E[T]^2} = \lim_{t \to \infty} \frac{\text{Var}[N(t)]}{E[N(t)]}. \tag{1.5}$$

The advantage to the expression in terms of $N(t)$ is that the number of cycles can be directly estimated in terms of the position of the bead, and that the noise involved in this estimation does not impact the limit. That is, if we represent the position of

the bead by $Y(t) = d \cdot N(t) + \xi(t)$, with $\xi(t)$ representing the noise, then we obtain

$$r = \lim_{t\to\infty} \frac{\mathrm{Var}[Y(t)]}{d\,E[Y(t)]}, \tag{1.6}$$

which can be estimated by replacing the theoretical variance and expectation with the ensemble averages based on the multiple runs which compose the data. Similar to the computation of the average velocity, this experimental computation has, in practice, been computed as a time-averaged quantity in terms of the time-lagged pairwise displacements (Schnitzer and Block, 1997). Writing

$$\widetilde{\Delta}_n y_i = y_{i+n} - y_i - v(n\Delta t), \tag{1.7}$$

the time-averaged mean squared variation from the expected displacement is computed for each run as

$$\overline{(\widetilde{\Delta}_n y)^2} = \frac{1}{K-n} \sum_{i=1}^{K-n} \left(\widetilde{\Delta}_n y_i\right)^2, \tag{1.8}$$

and, for each n, this is averaged over all runs. This time-averaged variation in displacement grows linearly in time,

$$\langle \overline{(\widetilde{\Delta}_n y)^2} \rangle = D \cdot (n\Delta t) + B + \text{residuals}. \tag{1.9}$$

using the fitted parameters D and v, the randomness can be estimated as

$$r = \frac{D}{d \cdot v}. \tag{1.10}$$

The VSB experiment succeeded in measuring the velocity and randomness over a wide range of experimental conditions. By considering the observed measurements for a particular concentration of ATP but different forces, the load-dependence of the velocity and the randomness were characterized. As the load increased, the velocity steadily decreased. When [ATP] = 5 μM, an ATP-limiting case, the velocity decreased approximately linearly with increasing load. For the ATP-saturating case of [ATP] = 2 mM, the velocity decreased with a more concave shape, and with a

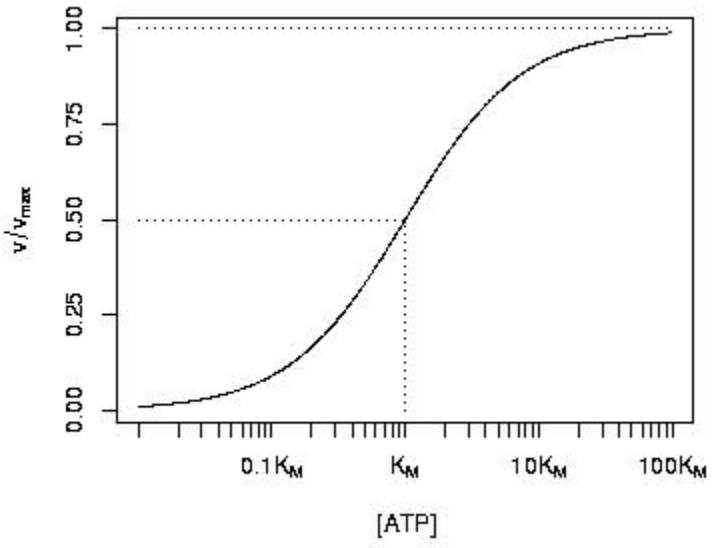


FIGURE 1.4. A velocity profile demonstrating a Michaelis-Menten relationship to ATP concentration. The velocity attains half the maximal velocity, $v_{\max}$, when $[\text{ATP}] = K_M$.

different stall force. Similarly, by considering a particular level of the force, the [ATP]-dependence of the velocity and randomness were also described. Under low loads, the dependence on [ATP] had been previously demonstrated to satisfy classical enzyme kinetics (Howard et al., 1989; Schnitzer and Block, 1997). That is, for low [ATP], the velocity grows proportionally to [ATP], but as the concentration increased, the velocity approaches an asymptotic limit $v_{\max}$. This relationship is well characterized by the Michaelis-Menten relationship,

$$v([\text{ATP}]) = \frac{v_{\max}[\text{ATP}]}{[\text{ATP}] + K_m}, \tag{1.11}$$

where the Michaelis-Menten constant K_m is the concentration at which the velocity is exactly half the asymptotic limit, as illustrated in Figure 1.4. The VSB experiment demonstrated, as anticipated, that the maximum velocity decreased as the load increased. However, contrary to many models' predictions, the Michaelis-Menten constant increased as the load increased. We will discuss more precisely these relationships in the next section in the context of theoretical models for kinesin in order

to give some possible explanations for these results.

1.3 Mathematical Models

The discovery of the directed motion of kinesin along a microtubule inspired a number of physical and mathematical models. Early models fell into two basic classes: Brownian ratchet models and simple mechanistic models. The Brownian ratchet models focused on the use of asymmetry to rectify thermal fluctuations, and typically attempted to model motor proteins such as kinesin through asymmetric and fluctuating energy interactions with a microtubule. The simple mechanistic models attempted to characterize motion by hypothesizing possible general mechanical structures, such as hinges and springs, and considering the motion of the motor protein passing through a sequence of states of these components. A third class of models represents somewhat of a hybrid of these earlier models, which we call mechanochemical models, although they actually were first proposed to explain the mechanics of myosin in muscle (Huxley, 1957; Huxley and Simmons, 1971). Mechano-chemical models incorporate a sequence of chemical and/or conformational states through which the motor protein progresses as a stochastic process, not necessarily attempting to identify specific mechanical features of these states but rather characterizing properties such as transition rates and partial displacement of the motor relative to a complete step. In this section, we discuss these models with a view of motivating the use of a Markov jump process to model kinesin motility.

A number of physical considerations motivated the development of Brownian ratchet models. First, the dimensions of the motor protein system necessitate the incorporation of thermal Brownian noise (Magnasco, 1993). All events occur at the molecular level, and thermal fluctuations keep everything moving. Energy barriers, which control the rates of transitions of binding and unbinding, are reasonably comparable with the thermal energy scale of room temperature where experiments are

performed, $k_BT = 4$ pN·nm. For example, the free energy released from the hydrolysis of ATP, the source of energy in the kinesin mechanical cycle, is ~ 31 kJ/mol, which corresponds to about $12\,k_BT$, although it should be emphasized that this is actually a standard free energy change which must be adjusted according to the concentrations of ATP, ADP and P_i in solution. Unfortunately, current experiments have not as yet characterized the ADP and phosphate concentrations, making a precise determination of available free energy unclear. In addition, by moving the distance of a single tubulin dimer, $d = 8.2$ nm, against observed opposing forces of greater than 5 pN, kinesin performs a comparable net amount of work. Second, most of the ratchet models utilize a potential energy that changes in time, relative to which a particle undergoes Brownian motion. The binding of kinesin to its tubulin track can, in principle, be described according to the potential energy. During the processes of binding ATP and its subsequent hydrolysis, the interaction changes between the kinesin and microtubule, which corresponds to a change in the potential energy (Astumian and Bier, 1994). Finally, a ratchet mechanism provides a reasonable method to allow directed motion at microscopic scales without violating the second law of thermodynamics (Feynman et al., 1963; Vale and Oosawa, 1990; Magnasco, 1993).

The basic mathematical model for a Brownian ratchet includes a particle, perhaps representing a motor protein, that undergoes motion in a potential energy landscape with thermal noise. If we let $x(t)$ represent the position of the particle at time t, and let $V(x)$ represent a periodic potential energy due to interactions between the particle and its track, then the Langevin equation describing the motion of the particle in the presence of thermal noise is given by

$$\gamma \frac{d}{dt} x(t) = -\frac{d}{dx} V(x(t)) + \sqrt{2\gamma k_B T}\, \xi(t), \tag{1.12}$$

where γ represents the viscous drag coefficient for the particle in solution and $\xi(t)$ is a standard white noise process. In the absence of any additional influences, this

system has a stationary distribution with a density given by

$$f(x) \propto \exp(-V(x)/k_B T), \tag{1.13}$$

and which satisfies detailed balance so that there is no net transport. However, by introducing an external fluctuation and by using an asymmetric potential V, the Brownian motion can be rectified to induce directed motion that can perform work. There are two classic methods to generate the external fluctuations, either by introducing an additional fluctuating force or by having the potential V fluctuate in time. We will briefly discuss simple examples of each of these (Astumian and Bier, 1994; Astumian, 1997).

The simplest example of a fluctuating force ratchet involves an additive, homogeneous force that alternates between two opposite values, $\pm\Delta F$. That is, we consider an expanded system that incorporates a fluctuating state $z(t) = \pm 1$ so that the Langevin equation becomes

$$\gamma \frac{d}{dt} x(t) = -\frac{d}{dx} V(x(t)) + \sqrt{2\gamma k_B T}\, \xi(t) + \Delta F \cdot z(t). \tag{1.14}$$

At a given time t, the effect of the additional force is that the potential energy landscape has been tilted with a slope of $\pm\Delta F$. As a simple example of an asymmetric potential, consider the piecewise linear potential energy illustrated in Figure 1.5. Because the repeating peaks on this potential energy function have the same height, a particle in equilibrium undergoing Brownian motion in this potential will have the same rate to cross in either direction. The addition of a homogeneous force superimposes a tilt on the potential, which alternately raises and lowers the energy barrier by a fixed amount on each side, and these models are sometimes referred to as a rocking ratchet. For each of the fixed tilts, the stationary distribution corresponds to a particle having an exponentially faster rate to cross by diffusion the lowered barrier than the raised barrier, leading to a net velocity in the direction of the superimposed force. The asymmetry in the potential energy leads to a larger change in energy

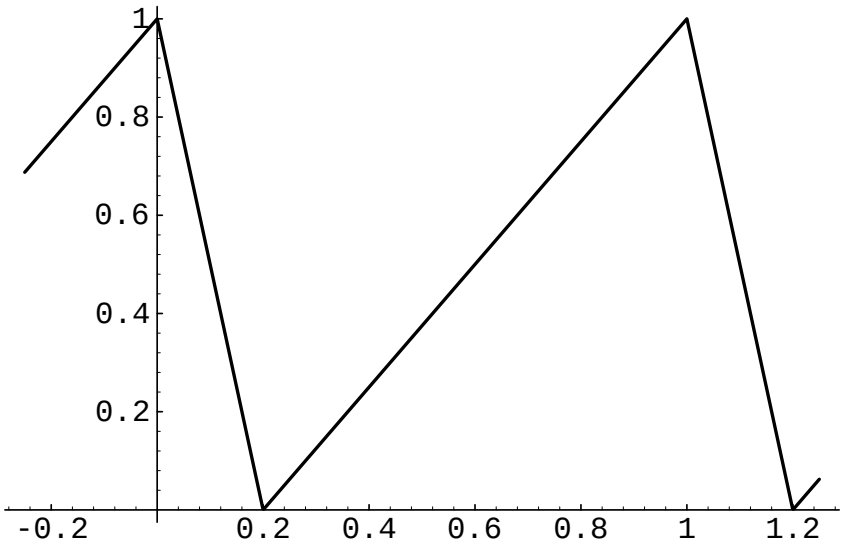


FIGURE 1.5. Periodic piecewise linear, asymmetric potential. The peaks all have $V = E_0$ and the valleys all have $V = 0$.

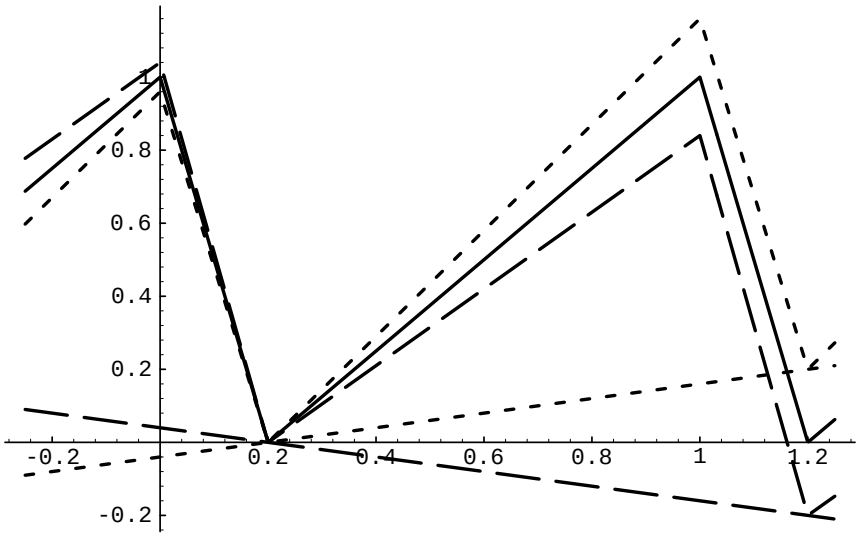


FIGURE 1.6. Plot of the potential in a rocking ratchet model. The fluctuating homogeneous reduces energy barrier the most along the longer, but less steep edge, leading to exponentially more likely transitions over that barrier. Net motion goes to the right.

barrier for the peak further away from the potential energy minimum. As a result, a given force applied in this direction leads to a greater average velocity than the same force applied in the opposite direction (see Figure 1.6). By fluctuating between these two forces at a slow enough rate, the overall average force will be zero, yet the particle will exhibit an average velocity moving in the direction of the peak further away from the minimum. The presence of an additional constant force opposing motion, perhaps caused by a load, reduces the amount the barrier is lowered, but if the force is not too large, the motion will continue, moving against this force and performing work. Note that Brownian motion is essential to this behavior, to allow for the escape over the potential energy barrier. Other variations of the fluctuating force ratchet include continuously varying forces (Magnasco, 1993), as well as additive colored noise (Millonas and Dykman, 1994). The fluctuating force ratchet, however, does not describe a system compatible with motor proteins such as kinesin. In particular, the non-equilibrium fluctuations in the force are spatially homogeneous, whereas the fluctuations for motor proteins need to be localized to the interaction between the motor and its track.

Fluctuating potential ratchets (Astumian and Bier, 1994; Astumian, 1997) overcome the localization problem, because instead of superimposing a global change in the energy landscape, the potential energy itself changes, possibly only in a region local to the position of the motor. A prototype of the fluctuating potential arises by considering the same initial potential energy landscape as shown in Figure 1.5. In this base state, the particle will localize near the minimum of the potential energy, in that the most likely configurations are for the particle at or near the minimum. Occasional transitions over the boundaries will occur, though they will happen at the same rate for each direction. Consider, then, the effect of the potential energy suddenly being turned off, leaving the particle free to undergo standard Brownian motion. The probability density of a particle undergoing Brownian motion with a well-defined initial condition is simply a Gaussian density with a center at the initial

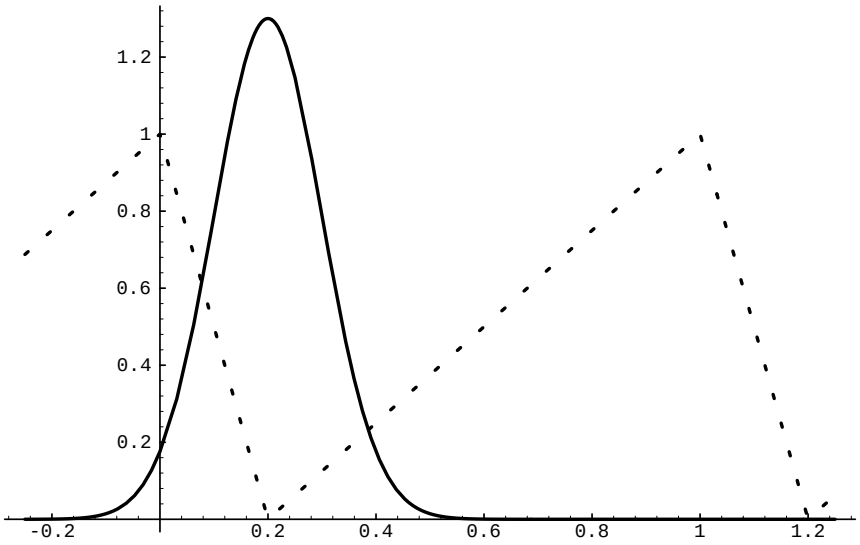


FIGURE 1.7. Plot of the potential in a flashing ratchet model. A particle undergoing Brownian motion which starts at the minimum of the potential energy function (dashed line) has a Gaussian probability density which spreads in time. The period to the left has a greater probability than the period to the right, so the net motion is to the left. Note that the probability of moving to the left is still less than 1/2.

position with a variance that grows proportionally with the time. In this case, the asymmetry in the potential energy function places the initial starting point for the Brownian motion to be closer to one of the barriers than the other. If the potential energy is turned back on, then the probability of being trapped in the region which is closer to the starting minimum will be higher than the probability of being trapped in the region which is farther away, as illustrated in Figure 1.7. By alternately turning the potential on and off, hence the name of a flashing potential model, the particle will alternately move to the energy minimum followed by Brownian motion, leading to a net motion in the direction of the peak nearest the minimum. The presence of an opposing force introduces a tilt to the potential energy landscape, and when the potential is turned off, the particle will undergo a constant drift in the direction of the force. However, if the force is not too large and the time that the potential is turned off remains sufficiently short, then spread of the density will still favor the site

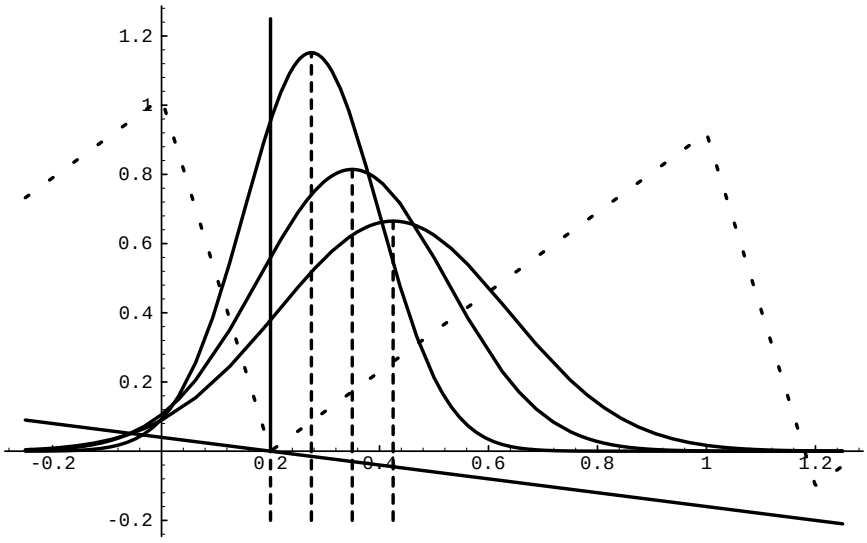


FIGURE 1.8. With a superimposed force, the Gaussian density in a flashing ratchet model drifts to the right and spreads symmetrically. The spread still favors the particle changing sites to the left compared to the right for short flashing times.

in the direction of the nearest peak, as shown in Figure 1.8.

Fluctuating or flashing potential ratchet models provide an intriguing possible model for motor proteins, although there are a number of serious challenges. Typically, the transition in the potential energy corresponds to an event in the chemical cycle of ATP hydrolysis (Peskin et al., 1994; Jülicher et al., 1997). In fact, as mentioned earlier, when kinesin binds ATP or has no ligand, the microtubule-kinesin interaction becomes strong, whereas bound ADP leads to a weak interaction. Consequently, modeling the transitions in chemical states as corresponding to different potential energies makes good physical sense. However, in this simplistic model, because the probability of diffusing to either side with the potential off is, by symmetry, at most 1/2, this model requires an average of greater than two cycles for each step in the direction of motion, which is counter to the tightly-coupled behavior of kinesin. This weakness can be overcome by introducing additional states (Jülicher et al., 1997; Astumian and Derényi, 1999) and by incorporating different potential energies for each state rather than simply alternating between two states (Kolomeisky and Widom, 1998).

However, a greater challenge is that dealing with the entire potential energy function becomes unwieldy. All present ratchet models use simple potential energies, while the true interaction between kinesin and the microtubule remains elusive. So while it might be argued that a sufficiently complicated flashing potential ratchet model might explain the motion of kinesin, the state of science is insufficient to pursue this approach directly.

Another class of models for kinesin, mechanical or power-stroke models, attempt to simplify the overall structure of the protein into a few simple mechanical parts, such as hinges, springs, rods and levers (Peskin and Oster, 1995; Duke and Leibler, 1996). In these models, energy stored from binding events or ATP hydrolysis is transferred through springs and levers to provide a physical power stroke which propels the motor forward. Note that because of the overdamped conditions that exist in the molecular domain, the propulsion does not impart a momentum that can carry the motor forward, but instead serves to move part of the protein to a forward position. Most of these models also directly incorporate the presence of both heads of kinesin linked by a spring or a hinge, with one of the heads undergoing hydrolysis and a power stroke and moving the other head into the presence of the next binding site. Paralleling some of the ideas of the Brownian ratchet models, the power stroke provides an asymmetry, and typically the second head which has been positioned forward must then wait for Brownian motion to complete the step. The effect of an opposing force is typically included as a shift in equilibrium positions through linkage in the springs joining the heads. Each of the dynamic mechanical components must be characterized through phenomenological parameters, such as spring constants, lever arms, and power-stroke amplitudes. Typically, these parameters are chosen in order to fit reported quantities such as the force-velocity curves and randomness. The intended advantages of these models are that they propose specific behaviors in the physical proteins that might generate the motion. That is, they attempt to model conformational changes as simple mechanical changes. The greatest challenge for this approach is that the number

of parameters is high relative to the quantities to which they are fit.

A third class of models, mechanochemical models combine several of the advantages of the flashing potential ratchet models and the mechanical models. As in the flashing potential models, mechanochemical models hypothesize that the motor protein has a number of different chemical states, with random transitions between these states (Astumian, 1997). However, instead of attempting to describe the transition in terms of a potential energy landscape, the transitions are characterized by transition rates. Most commonly, these transitions are assumed to satisfy the properties of a Markov process, although there has been some work to generalize to arbitrary waiting times (Fisher and Kolomeisky, 1999a,b). Mechano-chemical models with a single cycle are sometimes called tightly-coupled models, as a step forward would correspond precisely with one completion of the cycle. As in the mechanical models, conformational changes can be incorporated by associating each state with a specific displacement. Early tightly-coupled models for kinesin (Leibler and Huse, 1991) selected specific chemical states and corresponding kinesin-microtubule interactions. In fact, mechanical models might be viewed as being built around a tightly-coupled model, with the addition of attempting to model the power-stroke and the coupling between the two different heads (Duke and Leibler, 1996).

Most recently, models have attempted to incorporate the dependence on an external force, motivated by the single-molecule experiments. In order to accommodate a force, thermodynamics requires the model to account for reversibility. Consider a mechanochemical cycle in which all of the load-dependence occurs within a single reversible transition between two states and which involves a displacement of length ℓ. Let $k_f(F)$ be the forward transition rate, which depends on the applied force, and let $k_r(F)$ be the load-dependent reverse transition rate. A detailed balance condition dictates the force dependence on transition rates that involve physical displacement

(Qian, 1997), according to the relationship

$$\frac{k_f(F)}{k_r(F)} = \frac{k_f(0)}{k_r(0)} e^{-F\ell/k_B T}. \tag{1.15}$$

That is, the forward and backward rates adjust in order to account for the conversion of free energy $F\ell$ into work. However, this does not account for transition rates that may be affected indirectly due to an external force, such as strains in the protein that might alter the molecular dynamics. Nevertheless, a number of different models have incorporated external forces by implementing a detailed balance condition as in Equation 1.15.

Astumian and Derényi (AD) developed a flashing ratchet model with four different potential energy landscapes corresponding to four chemical states (1999). Under the assumption that the physical transitions between different states, such as for Brownian motion over a barrier, occur faster than the chemical transitions which drive the potential energy landscape changes, this model was recast as a Markov jump process. This simplification breaks down when waiting times are not exponentially distributed, such as when a comparable deterministic time for a mechanical transition is added to a typical chemical transition. However, by casting the problem as a Markov process, the machinery available for Markov processes can be applied to make the analysis of the model explicit. Load dependence was introduced by imposing a detailed balance condition on diffusive steps, where an imposed force had the effect of altering the energy barriers. Explicitly, transitions which involved a displacement ϵd, where ϵ represents the fraction of a complete step of size d which occurs during that transition, had a suppression of the forward rate by the factor $f = e^{-\epsilon F d/2k_B T}$, and reverse transition rates were enhanced by the factor f^{-1}. This representation implements the detailed balance condition $f/f^{-1} = e^{-\epsilon F d/k_B T}$ by dividing the influence equally in the forward and reverse direction.

In a follow-up paper to the single-molecule kinesin experiments, Schnitzer et al. (2000) proposed another model (SVB) based on a load-dependent composite state.

The motivation for the model was to account for the load-dependence of the observed velocity, as well as to account for the processivity measurements reported in this same paper. Starting with a typical tightly-coupled model based on the standard ATP hydrolysis cycle, the SVB model introduces a composite state when ATP binds with one head of the kinesin. Thus, as ATP binds kinesin, kinesin undergoes an isomerization with two possible configurations, undergoing rapid transitions between these states. If the isomerization is rapid enough, the kinesin can be treated as existing in an equilibrium between the two states. Load-dependence for the model arises through the load-dependence of the equilibrium distribution between these states.

Rather than parametrize all of the transition rates, the SVB model summarizes the effect of all transition rates through the use of a binding rate k_b and a catalytic rate k_{cat}. The binding rate k_b represents the rate at which kinesin binds ATP which will actually undergo hydrolysis, and thus incorporates both the rate of binding ATP as well as subsequent release, relative to proceeding through the remainder of the hydrolysis cycle. The catalytic rate k_{cat} represents the rate at which kinesin with an irreversibly bound ATP molecule will complete the remainder of the hydrolysis cycle, and thus includes effects from all of the remaining transition rates. Allowing both the catalytic rate and effective binding rate to depend on the external force, the functional dependence of velocity on load and [ATP] can then be written

$$v(F, [\text{ATP}]) = \frac{d \cdot k_{\text{cat}}(F)[\text{ATP}]}{[\text{ATP}] + k_{\text{cat}}(F)/k_b(F)}, \tag{1.16}$$

so that the asymptotic maximal velocity and Michaelis-Menten constant can be written $v_{\max}(F) = d \cdot k_{\text{cat}}(F)$ and $K_M(F) = k_{\text{cat}}(F)/k_b(F)$. In order that the maximum velocity decrease as the opposing force increases, k_{cat} must decrease. In order that the Michaelis-Menten constant increase, the effective binding rate k_b must decrease faster than k_{cat}.

Motivated by the assumption of an underlying potential energy landscape, the

SVB model proposes that these two isomer states, which represent two conformational states with shifted positions, correspond to two local minima of the potential energy at each of the isomer positions and separated by a relatively small energy barrier. The resulting tilt of the energy landscape due to an external force shifts the equilibrium between these two isomers in the direction of the force. Because of the assumed fast equilibrium, detailed balance is implemented through the load dependence of the equilibrium constant for this isomerization K according to $K = K_0 e^{-\epsilon F d/k_B T}$, where ϵ represents the fraction of a complete lattice step that corresponds to a transition between the two isomer states. Consequently, the presence of the external force increases the fraction of time spent in the first isomer configuration, which will increase the effective rate of ATP unbinding and thus decrease k_b as the force increases. In addition, the corresponding decrease of the second configuration occupancy leads to a reduction in the catalytic rate. All other transitions are assumed to have no load dependence. The overall effects of the various transition rates and load are combined to obtain the expressions for $k_{\rm cat}$ and k_b:

$$k_{\rm cat}(F) = \frac{k_{\rm cat}^0}{p_{\rm cat} + q_{\rm cat} e^{\epsilon F d/k_B T}}, \tag{1.17}$$

$$k_b(F) = \frac{k_b^0}{p_b + q_b e^{\epsilon F d/k_B T}}, \tag{1.18}$$

where p, q, and k^0 parametrize these rates. Such a parametrization leads to good global fits to the velocity data of the single-molecule VSB experiments (Visscher et al., 1999), and can also be adapted to account for the processivity results.

Another approach to modeling load-dependence, but without explicitly referencing an underlying potential energy landscape, has been recently developed by Fisher and Kolomeisky (Fisher and Kolomeisky, 1999a,b; Kolomeisky and Fisher, 2000; Fisher and Kolomeisky, 2001). In this framework, the motor protein moves along a lattice, passing through N internal states associated with each lattice site. The simplest scheme within the framework is a standard kinetic scheme, where the motor proceeds through the states in sequence, with the transition out of the last state returning to

the first state, but with the lattice site increasing by one. Labeling the internal states $1, 2, \ldots, N$, the following kinetic diagram represents the dynamics:

$$1_l \underset{v_2}{\overset{u_1}{\rightleftharpoons}} 2_l \underset{v_3}{\overset{u_2}{\rightleftharpoons}} \cdots \underset{v_N}{\overset{u_{N-1}}{\rightleftharpoons}} N_l \underset{v_1}{\overset{u_N}{\rightleftharpoons}} 1_{l+1}, \tag{1.19}$$

where the subscript specifies the site on the lattice, and u and v represent the transition rates between sites. Unlike earlier models, the framework does not try to assign a chemical interpretation for each of the states, preferring instead to let the state represent some generic stage through which the motor must pass, except for the first state which is explicitly assigned the state immediately prior to ATP binding. Consequently, the first forward transition u_1 in the absence of load (indicated by the superscript of 0) is modeled in terms of [ATP] by

$$u_1^0 = k_0[\text{ATP}]. \tag{1.20}$$

In addition, the final reverse transition is modeled as having a non-linear [ATP]-dependence as

$$v_1^0 = k_0'[\text{ATP}]/(1 + [\text{ATP}]/c_0)^{1/2}. \tag{1.21}$$

In addition, non-Markovian structure was discussed by theoretically replacing the transition rates with arbitrarily distributed waiting times. The Markov structure would be recovered with exponentially distributed waiting times. Alternative models within this framework provide for additional properties such as side branches as well as death states that mightcorrespond to irreversible detachment of the motor from the microtubule. However, we focus on the standard kinetic scheme for clarity.

Coupling the opposing force into the model occurs, again, through the use of multiplicative factors on the transition rates that enforce detailed balance. However, unlike Astumian and Derényi's implementation, the displacement of a particular transition is not explicitly stated. Writing the external force as F and the lattice size as

d, the forward and backward transitions, respectively, take the functional forms

$$u_k = u_k^0 e^{-\vartheta_k^+ Fd/k_BT} \tag{1.22}$$

$$v_k = v_k^0 e^{+\vartheta_k^- Fd/k_BT}. \tag{1.23}$$

Again, considering the ratio of corresponding forward and backward transitions between states k and $k+1$, we observe

$$\frac{u_k}{v_{k+1}} = \frac{u_k^0}{v_{k+1}^0} e^{-(\vartheta_k^+ + \vartheta_{k+1}^-)Fd/k_BT}. \tag{1.24}$$

Consequently, the sum $\epsilon_k = \vartheta_k^+ + \vartheta_{k+1}^-$ might be interpreted as an effective partial step-size by comparison with Equation 1.15. Although a corresponding physical displacement may be suggested by this form, it is not necessarily implied, as strains in the protein may account for some load dependence. We will return to this point in the next chapter. We note that the use of both ϑ^+ and ϑ^- allows the effect of the load to be distributed unequally between the forward and backward transitions. Essentially, this implementation assumes that the load dependence can be reasonably approximated by an exponential of a linear function of the force.

Using the parametrization for transition rates and their dependence on load, Fisher and Kolomeisky (FK) used this model to fit (Fisher and Kolomeisky, 2001) the experimental single-molecule results (Visscher et al., 1999; Schnitzer et al., 2000). By varying the parameters $u, v, \vartheta^+, \vartheta^-$, they found global fits for the experimentally determined [ATP]-velocity, force-velocity relationships, and stall forces, as well as randomness. By including side branches and death states, they also fit the processivity results. When $N = 2$, they were able to fit the velocity results and stall forces reasonably well. However, the randomness dependencies required more detailed models. One approach they used to account for randomness was to introduce a non-exponential waiting time for the second state. However, they also expanded the model to $N = 4$, which results in good agreement with the published experimental results for both velocity and randomness. Interpreting the detailed balance parameters

$\vartheta^{\pm}$ as corresponding directly to a physical size of a substep, the predicted effective displacement arising from ATP binding is 1.8 nm for the two-state model and 2.1 nm for the four-state model.

Recalling the SVB model, it seems as though the composite state model might be considered as a particular case of the FK model. In both cases, the underlying models have Markov structure and incorporate the load-dependence through the transition rates. By using two separate states in the FK model to correspond to the single composite state of the SVB model, the isomerization can be accomplished by making the forward and reverse rates between these two states sufficiently large relative to the outgoing rates. Consider the following portion of a Markov jump process X_t:

$$\cdots 1 \underset{v_2'}{\overset{u_1}{\rightleftharpoons}} 2_- \underset{k_r}{\overset{k_f}{\rightleftharpoons}} 2_+ \underset{v_3}{\overset{u_2'}{\rightleftharpoons}} 3 \cdots . \tag{1.25}$$

The equilibrium constant for the composite state is equal to the ratio of the forward and reverse rates $K = \frac{k_f}{k_r}$. We represent the forward and backward rates as

$$k_f = \frac{1}{\eta}\tilde{k}_f, \quad k_r = \frac{1}{\eta}\tilde{k}_r, \tag{1.26}$$

where $\eta \ll 1$ represents a scaling parameter with the remaining transition rates u_i and v_i unchanged. Then the limiting behavior as $\eta \to 0$, corresponding to making the isomerization rates much larger than the other rates while holding the equilibrium fixed, is equivalent, in terms of transition rates, to the simpler model (Schnitzer et al., 2000)

$$\cdots 1 \underset{v_2}{\overset{u_1}{\rightleftharpoons}} 2 \underset{v_3}{\overset{u_2}{\rightleftharpoons}} 3 \cdots . \tag{1.27}$$

with transition rates

$$u_2 = \frac{u_2' K}{K+1}, \quad v_2 = \frac{v_2'}{K+1}. \tag{1.28}$$

Because of the success of the SVB model and the FK model, as well as the computation tools available for Markov processes, tightly-coupled models seem well-suited in describing the motion of kinesin. With a motivation for considering Markov processes, we next turn to the implementation of filtering algorithms.

1.4 Hidden Markov Models

Hidden Markov models (HMMs) (Rabiner, 1989; Elliott et al., 1997) attempt to describe a random sequence of observations that are generated from an unobservable Markov process. That is, we have a hidden Markov state process, $\{X_k; k \geq 1\}$, and an observation process, $\{Y_k; k \geq 1\}$. The state space for the Markov process, X_k, will consist of a finite number of states, which we can label $1, 2, \ldots, N$, and the dynamics are characterized by a transition matrix A and an initial distribution π_1. To be consistent with the notation of standard Markov processes, we let the i^{th} row of A contain the probability distribution of X_{k+1} given that $X_k = i$,

$$P[X_{k+1} = j | X_k = i] = a(i, j). \tag{1.29}$$

If π_k represents the distribution of X_k as a row vector, then the distribution π_{k+1} of X_{k+1} is given by the product of the transition matrix A and the initial distribution π_k, according to

$$\pi_{k+1} = \pi_k A. \tag{1.30}$$

Successive iterations are given by repeated multiplication by the transition matrix A. The observation process, Y_k, will also be a stochastic process, except that the distribution of Y_k depends only on the state X_k. Consequently, the sequence $\{Y_k\}$ will be a conditionally independent sequence given the sequence $\{X_k\}$. In the simplest case, this distribution will depend only on the value of X_k and not on the time index k. Typically, the distribution of Y_k given X_k will be absolutely continuous with respect to some common reference measure, ν. If the observation space is finite, then ν can be chosen to be counting measure on the state space. If the distribution is a continuous distribution, the reference measure might be Lebesgue measure. Given that $X_k = i$, we denote the density of Y_k with respect to ν as $b_i(y)$. We will let B denote the collection of all of the conditional densities. Consequently, the complete model is characterized by the triple (π_1, A, B). Typically, the transition matrix A

and the conditional densities B will be functions of some collection of parameters, collectively denoted θ.

Following the example of the classic tutorial by Rabiner (Rabiner, 1989), we introduce three fundamental problems for HMMs. There are three typical questions that might be asked given the observation sequence, $\vec{y} = (y_1, \ldots, y_K)$:

1. Given a model (π_1, A, B) (or, equivalently, the parameter θ), what is the likelihood of observing this sequence? Equivalently, what is the likelihood of the model given the observations?

2. Given a model and the observations, what is the best sequence of states, $\vec{x} = \{x_1, \ldots, x_K\}$ for the hidden state process X_k?

3. Given a parametrization for the model, what is the best choice of parameters θ to describe the observations?

We will discuss the classic solutions to each of these problems, as well as adaptations of the techniques to accommodate more general hidden Markov models. The ideas presented are essentially the same as the classic references (Baum and Petrie, 1966; Baum et al., 1970; Rabiner, 1989), but the presentation is designed to make later developments in the present work more natural.

1.4.1 Computing Likelihood: Forward-Backward Algorithm

We begin by considering the first problem, computing the likelihood of the observations. First, let $x_1, \ldots, x_K$ represent an arbitrary, fixed sequence of states for the state process X_k. The likelihood of the model for the complete system given the state sequence $\vec{x}$ and the observation sequence $\vec{y}$ is given by the product

$$L(\theta; \vec{x}, \vec{y}) = \pi_1(x_1) b_{x_1}(y_1) a(x_1, x_2) b_{x_2}(y_2) \cdots a(x_{K-1}, x_K) b_{x_K}(y_K). \tag{1.31}$$

However, since the actual sequence for the state process is unknown, we must sum over all possible sequences $\vec{x}$ that the state process might go through. Of course, enumerating the list of all such sequences and then computing the individual likelihoods is a particularly inefficient method, requiring the computation of the likelihood for each of the N^K different sequences. Instead, it is more effective to observe that if we define the matrix-valued function $B(y) : \mathbb{R} \to GL(\mathbb{R}, N)$ as the diagonal matrix with diagonal entries given by

$$B_{i,i}(y) = b_i(y), \tag{1.32}$$

then the sum over paths can be interpreted as the matrix product

$$L(\theta; \vec{y}) = \sum_{\vec{x}} L(\theta; \vec{x}, \vec{y}) \tag{1.33}$$

$$= \pi_1 B(y_1) \cdot AB(y_2) \cdots AB(y_K)\vec{1}, \tag{1.34}$$

where $\vec{1}$ is an N-dimensional vector containing all ones. The full-system likelihood $L(\theta; \vec{x}, \vec{y})$ represents the probability density with model parameter θ for the state process X_k and the observation process Y_k, $1 \leq k \leq K$, relative to the reference measure $(\chi \times \nu)^K$, where χ represents counting measure. Then the hidden Markov model likelihood $L(\theta; \vec{y})$ is the marginal density with respect to ν^K, with the summation over all states representing integration over the counting measure.

The forward algorithm (Baum et al., 1970) simply implements this matrix formulation, although it is typically presented as an inductive algorithm. Of interest are the forward variables $\alpha_k(i)$. Here is the standard description:

1. Initialization:

$$\alpha_1(i) = \pi_1(i) b_i(y_1), \quad 1 \leq i \leq N, \tag{1.35}$$

2. Induction:

$$\alpha_{k+1}(j) = \sum_{i=1}^{N} \alpha_k(i) a(i,j) b_j(y_{k+1}), \quad \begin{matrix} 1 \leq k \leq K-1, \\ 1 \leq j \leq N, \end{matrix} \tag{1.36}$$

3. Termination:

$$L(\theta; \vec{y}) = \sum_{i=1}^{N} \alpha_K(i). \tag{1.37}$$

The matrix interpretation of this result is that at each time k, we inductively define the row vector α_k as

1. Initialization: $\alpha_1 = \pi_1 \cdot B(y_1)$,

2. Forward Induction: $\alpha_{k+1} = \alpha_k \cdot AB(y_{k+1})$,

and the likelihood is given as the inner product of α_K with the vector of ones, $L(\theta; \vec{y}) = \alpha_K \cdot \vec{1}$. Each vector α_k actually represents the unnormalized conditional probability distribution for the state at time k, given the observations through time k under the model specified by θ. That is,

$$P_\theta[X_k = i|\vec{y}_k] = \frac{1}{\alpha_k \cdot \vec{1}} \alpha_k(i), \tag{1.38}$$

gives the conditional probability of the state at time k, given the forward partial observation sequence $\vec{y}_k$, where $\vec{y}_k$ represents the partial observation sequence $Y_1 = y_1, \ldots, Y_k = y_k$. Alternatively, the quantity $\alpha_k(i)$ can be interpreted directly as the likelihood of the partial sequence,

$$\alpha_k(i) = L(\theta; X_k = i, \vec{y}_k). \tag{1.39}$$

Whereas computation of the likelihood by direct enumeration of all paths for the state process grows exponentially in the length of the data, the forward algorithm requires only $O(KN^2)$ multiplications, with the complexity of the calculation growing linearly with the length of the data.

While the forward computation corresponds to computing a particular length of the left-hand side of the matrix calculation in Equation 1.34, a similar backward computation corresponds to the remaining right-hand side. Rather than present the standard inductive formulation, we simply state the algorithm in terms of the matrix

calculation. We define a sequence of column vectors β_k, the backward variables, according to the inductive relationship:

1. Initialize: $\beta_K = \vec{1}$,

2. Backward Induction: $\beta_k = AB(y_k) \cdot \beta_{k+1}$.

Consequently, we obtain for each k that the product of the forward and backward vectors will yield the hidden Markov model likelihood, $\alpha_k \cdot \beta_k = L(\theta; \vec{y})$. The interpretation of the backward variable $\beta_k(i)$ is as the conditional likelihood of the tail observations given $X_k = i$,

$$\beta_k(i) = L(\theta; \vec{y} \backslash \vec{y}_k | X_k = i), \tag{1.40}$$

where $\vec{y} \backslash \vec{y}_k$ represents $Y_{k+1} = y_{k+1}, \ldots, Y_K = y_K$.

In practice, subsequent iterations of the forward and backward relations tend to cause the terms of the vectors to geometrically decay, since a typical term in the product is less than one. This would lead to numerical underflow. Consequently, a rescaling of these vectors must occur, at least occasionally. One especially useful approach is to rescale the forward variables at each iteration, $\tilde{\alpha}_k = c_k \alpha_k$, with $c_k = 1/(\alpha_k \cdot \vec{1})$ so that $\tilde{\alpha}_k$ becomes the normalized conditional probability distribution for the state at time k given the first k observations. The backward variables can also be rescaled using the same scaling factors according to the relation $\tilde{\beta}_k = c_{k+1} \beta_k$ for $k < K$. The induction steps are modified to use the rescaled variables:

$$\alpha_{k+1} = \tilde{\alpha}_k \cdot AB(y_{k+1}), \tag{1.41}$$

$$\beta_k = AB(y_{k+1}) \cdot \tilde{\beta}_{k+1}. \tag{1.42}$$

Because all of the induction steps are simply linear transformations, the effect of the rescaling can be factored out to obtain a relationship between the scaled forward and

backward variables with the original, unscaled variables, denoted by a prime ($'$):

$$\tilde{\alpha}_k = \prod_{l=1}^{k} c_l \alpha'_k, \tag{1.43}$$

$$\tilde{\beta}_k = \prod_{l=k+1}^{K} c_l \beta'_k. \tag{1.44}$$

Therefore, we also obtain another method of calculating the likelihood as

$$L(\theta; \vec{y}) = \frac{1}{\prod_{k=1}^{K} c_k}. \tag{1.45}$$

In most cases this is exponentially small, so that the numerical calculation still leads to underflow for the likelihood. However, having written the likelihood as a product, we now also have a natural method to compute the log-likelihood as a sum of logarithms,

$$\log L(\theta; \vec{y}) = -\sum_{k=1}^{K} \log c_k. \tag{1.46}$$

By considering the term-wise product of α_k and β_k, we also obtain the following useful equalities, which we will use later in the context of the EM algorithm,

$$\gamma_{\theta,k}(i) \equiv P_\theta[X_k = i|\vec{y}] = \frac{\alpha_k(i)\beta_k(i)}{L(\theta; \vec{y})} = \tilde{\alpha}_k(i)\tilde{\beta}_k(i), \tag{1.47}$$

for the condition distribution of the state process, as well as the conditional probability of particular jumps at each time k,

$$\begin{aligned} \widehat{J}_{\theta,k}(i,j) \equiv E_\theta[J_k(i,j)|\vec{y}] &= P_\theta[X_k = i, X_{k+1} = j|\vec{y}] \\ &= \frac{\alpha_k(i)a(i,j)b_j(y_{k+1})\beta_{k+1}(j)}{L(\theta; \vec{y})} \qquad (1.48) \\ &= c_{k+1}\tilde{\alpha}_k(i)a(i,j)b_j(y_{k+1})\tilde{\beta}_{k+1}(j), \qquad (1.49) \end{aligned}$$

with $J_k(i,j) = I_i(X_k)I_j(X_{k+1})$ representing the occurrence of a jump at time k from state i to state j.

1.4.2 Best Sequence: Viterbi Algorithm

The second common question relating to a hidden Markov model is to determine the sequence $\vec{x} = (x_1, \ldots, x_K)$ for the hidden state process that are the best states given the model θ and the observations $\vec{y}$. To answer the question, one must explain what is meant by the best sequence. One common solution, the maximum *a posteriori* (MAP) sequence, is to determine for each time index k, the state x_k that is most likely given the observations. That is, we choose x_k so that

$$P[X_k = x_k|\vec{y}] = \max_i P[X_k = i|\vec{y}]. \tag{1.50}$$

Using the scaled forward and backward variables, $\tilde{\alpha}_k$ and $\tilde{\beta}_k$, and the formula in Equation 1.47, we can compute all of the individual posterior probabilities of each state for each index k. The disadvantage of this approach is that there is no relation between different indexes. In fact, the sequence which results may not even be a valid sequence, in that it may have zero likelihood. Another solution, the maximum likelihood (ML) sequence, is to ask for the sequence, out of all possible sequences, for which the full likelihood $L(\theta; \vec{x}, \vec{y})$ is maximized.

The ML sequence can be determined through a dynamic algorithm known as the Viterbi algorithm (Viterbi, 1967). The fundamental premise of the algorithm is that the partial sequences of an optimal sequence are also optimal. The algorithm can be explained as follows. For each time index k and state value i, we define the maximum likelihood of partial sequences with final state i,

$$\delta_k(i) = \max_{\vec{x}_k : x_k = i} L(\theta; \vec{x}_k, \vec{y}_k). \tag{1.51}$$

Because the likelihood factors, we also obtain the following inductive relationship:

$$\delta_{k+1}(j) = \max_i \delta_k(i) a(i,j) b_j(y_{k+1}). \tag{1.52}$$

The Viterbi algorithm recursively generates the partial sequence maximum likelihoods, and stores for each final state of the subsequence the previous state that led to the maximum likelihood path.

1. Initialization:

$$\delta_1(i) = \pi_1(i) b_i(y_1), \tag{1.53}$$

2. Recursion:

$$\delta_{k+1}(j) = \max_i \delta_k(i) a(i,j) b_j(y_{k+1}), \qquad \begin{array}{c} 1 \le k \le K-1, \\ 1 \le j \le N, \end{array} \tag{1.54}$$

$$\psi_k(j) = \arg\max_i \delta_k(i) a(i,j) b_j(y_{k+1}), \qquad \begin{array}{c} 1 \le k \le K-1, \\ 1 \le j \le N, \end{array} \tag{1.55}$$

3. Termination:

$$L^* = \max_i \delta_K(i), \tag{1.56}$$

$$x_K^* = \arg\max_i \delta_K(i), \tag{1.57}$$

4. Backtracking:

$$x_k^* = \psi_k(x_{k+1}^*), \qquad k = K-1, K-2, \ldots, 1. \tag{1.58}$$

The resulting sequence $\vec{x}^* = (x_1^*, \ldots, x_K^*)$ will then satisfy the maximum likelihood condition

$$L(\theta; \vec{x}^*, \vec{y}) = L^* = \max_{\vec{x}} L(\theta; \vec{x}, \vec{y}). \tag{1.59}$$

In terms of computation, we note that it is more practical to actually replace the maximum likelihood $\delta_k(i)$ with the maximum log-likelihood, analogous to the scaling that occurred in the forward-backward algorithm.

1.4.3 Model Estimation: Expectation-Maximization (EM) Algorithm

The final challenge in hidden Markov models is to find the best model within a particular parametrization for the given observation sequence. The most common choice for defining the best model is to find the parameter for which the likelihood

of the corresponding model is greatest. That is, we define the maximum likelihood estimator of the parametrization θ,

$$\widehat{\theta} = \arg\max_{\theta} L(\theta; \vec{y}). \tag{1.60}$$

In statistics, maximum likelihood estimators (MLEs) are particularly useful for many models because of their beneficial properties of consistency and asymptotic normality. If we write $\widehat{\theta}_n$ as the MLE given data of length n and suppose that θ_0 is the true parameter, then consistency means that the MLE converges to the true value $\widehat{\theta}_n \to \theta_0$, almost surely. Asymptotic normality refers to a central limit theorem-like behavior, so that natural error bounds can be placed around the estimates. Consistency and asymptotic normality of MLEs for HMMs in the case of a finite observation space was given by Baum and Petrie (1966), consistency for more general HMMs was proved by Leroux (1992), and asymptotic normality has been demonstrated for a variety of conditions (Bickel et al., 1998; Jensen and Petersen, 1999; Douc and Matias, 2000). Of course, these theorems require that certain assumptions on the models hold, such as appropriate continuity and differentiability conditions of the likelihood, ergodicity of the Markov state process, and identifiability of the model (up to labeling of the states).

There are two main challenges in finding MLEs for hidden Markov models. The first of these is the difficulty of maximizing the likelihood $L(\theta; \vec{y})$, given by Equation 1.33. Although maximization of the likelihood can be directly performed (Fredkin and Rice, 1992; Qin et al., 2000), the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) provides an alternative technique which generates a sequence of parameter estimates that are guaranteed to increase the likelihood. However, both direct maximization and the EM algorithm can potentially converge to local maxima which are not the true maximum. No known solution exists to this problem, and the general strategy is to try a number of different starting points for the maximization algorithms and select the best of the ending points.

The Baum-Welch algorithm (Baum et al., 1970) is a special case of the more general EM algorithm that applies for standard hidden Markov models. The algorithm consists of two fundamental steps: the expectation step (E-step) and the maximization step (M-step). Consider the simple example of an HMM consisting of a state process governed by the transition matrix $A = (a(i,j))$ where the observation process takes values in a finite collection $o_1, \ldots, o_M$. For each state i, the conditional probability that $Y_k = o_m$ given $X_k = i$ will be parametrized by the vector b_i:

$$P[Y_k = o_m | X_k = i] = b_i(m). \tag{1.61}$$

Then given an initial model parametrized by

$$\theta = \{\pi_1(i), a(i,j), b_i(m) : 1 \leq i \leq N, 1 \leq j \leq N, 1 \leq m \leq M\}, \tag{1.62}$$

the Baum-Welch algorithm gives a new model θ' using the following steps, recalling the formulas for $\gamma_{\theta,k}$ and $\widehat{J}_{\theta,k}$ given in Equations 1.47 and 1.49, respectively:

E: Use the Forward-Backward algorithm with model θ to compute the scaled forward and backward variables $\tilde{\alpha}_k$ and $\tilde{\beta}_k$.

M: Compute new model parameters using the following formulas:

$$\pi_1'(i) = \gamma_{\theta,1}(i), \tag{1.63}$$

$$\widehat{O}_{\theta,k}(i,m) = \gamma_{\theta,k}(i) I_m(y_k), \tag{1.64}$$

$$a'(i,j) = \frac{\sum_{k=1}^{K-1} \widehat{J}_{\theta,k}(i,j)}{\sum_{k=1}^{K-1} \gamma_{\theta,k}(i)}, \tag{1.65}$$

$$b_i'(m) = \frac{\sum_{k=1}^{K} \widehat{O}_{\theta,k}(i,m)}{\sum_{k=1}^{K} \gamma_{\theta,k}(i)}. \tag{1.66}$$

If the observation process is not from a finite set, but instead Y_k has a Gaussian distribution with mean μ_i and variance σ_i^2 given that the state $X_k = i$. Then the model is parametrized by $\theta = \{\pi_1(i), a(i,j), \mu_i, \sigma_i^2 : 1 \leq i \leq N, 1 \leq j \leq N\}$. In this case, Equation 1.66, the update of b_i, of the M-step is replaced with the corresponding M$'$-step updates for μ_i and σ_i^2,

M′:

$$\mu_i' = \frac{\sum_{k=1}^{K} y_k \gamma_{\theta,k}(i)}{\sum_{k=1}^{K} \gamma_{\theta,k}(i)}, \tag{1.67}$$

$$\sigma_i'^2 = \frac{\sum_{k=1}^{K} y_k^2 \gamma_{\theta,k}(i)}{\sum_{k=1}^{K} \gamma_{\theta,k}(i)} - \mu_i'^2. \tag{1.68}$$

In order to justify the use of the Baum-Welch algorithm, as well as the more general EM algorithm, we still need to show that the likelihood is nondecreasing, namely that $L(\theta';\vec{y}) \geq L(\theta;\vec{y})$. This is a consequence of an inequality that was originally proved in the context of the Baum-Welch algorithm (Baum and Petrie, 1966; Baum et al., 1970), but which was then generalized by Dempster, Laird and Rubin as part of the general EM algorithm (Dempster et al., 1977), which is suited for finding the MLE of a broad class of incomplete data problems, with HMMs becoming a special case. Introducing the auxiliary function $Q(\theta,\theta')$,

$$Q(\theta,\theta') = \sum_{\vec{x}} \frac{L(\theta;\vec{x},\vec{y})}{L(\theta;\vec{y})} \log L(\theta';\vec{x},\vec{y}), \tag{1.69}$$

we have the following inequality:

Theorem 1.4.1. *If $Q(\theta,\theta') \geq Q(\theta,\theta)$ then $L(\theta';\vec{y}) \geq L(\theta;\vec{y})$.*

Proof: We follow the standard proof (Dempster et al., 1977). This is simply an application of Jensen's inequality, applied to the concave function $\log t$, and the conditional probability $f_{X|Y,\Theta}(\vec{x}|\vec{y},\theta) = L(\theta;\vec{x},\vec{y})/L(\theta;\vec{y})$. We write

$$\begin{aligned} Q(\theta,\theta') - Q(\theta,\theta) &= \sum_{\vec{x}} f_{X|Y,\Theta}(\vec{x}|\vec{y},\theta) \log \frac{L(\theta';\vec{x},\vec{y})}{L(\theta;\vec{x},\vec{y})} \\ &= \sum_{\vec{x}} f_{X|Y,\Theta}(\vec{x}|\vec{y},\theta) \log \frac{L(\theta';\vec{x},\vec{y})/L(\theta';\vec{y})}{L(\theta;\vec{x},\vec{y})/L(\theta;\vec{y})} \\ &\quad + \sum_{\vec{x}} f_{X|Y,\Theta}(\vec{x}|\vec{y},\theta) \log \frac{L(\theta';\vec{y})}{L(\theta;\vec{y})} \\ &= S(\theta,\theta'|\vec{y}) + \log \frac{L(\theta';\vec{y})}{L(\theta;\vec{y})}, \end{aligned} \tag{1.70}$$

where we have introduced the conditional relative entropy $S(\theta, \theta'|\vec{y})$ for the model θ' relative to θ given the observations,

$$S(\theta, \theta'|\vec{y}) = \sum_{\vec{x}} f_{X|Y,\Theta}(\vec{x}|\vec{y}, \theta) \log \frac{f_{X|Y,\Theta}(\vec{x}|\vec{y}, \theta')}{f_{X|Y,\Theta}(\vec{x}|\vec{y}, \theta)}. \tag{1.71}$$

Note that the quantity $-S$ is also known as the conditional Kullback-Leibler information (Schervish, 1995). By Jensen's inequality, $S(\theta, \theta'|\vec{y}) \leq 0$ (or, equivalently, conditional Kullback-Leiber information is non-negative), resulting in

$$\begin{aligned} \log L(\theta'; \vec{y}) - \log L(\theta; \vec{y}) &= Q(\theta, \theta') - Q(\theta, \theta) - S(\theta, \theta'|\vec{y}) \\ &\geq Q(\theta, \theta') - Q(\theta, \theta). \end{aligned} \tag{1.72}$$

This completes the proof. ■

We finally show that the EM algorithm is equivalent to the Baum-Welch algorithm, so that the Baum-Welch algorithm has non-decreasing likelihood. If we consider the likelihood of the full model as given in Equation 1.31, then the log-likelihood can be written

$$\begin{aligned} \log L(\theta'; \vec{x}, \vec{y}) &= \log \pi'_1(x_1) + \sum_{k=1}^{K-1} \log a'(x_k, x_{k+1}) + \sum_{k=1}^{K} \log b'_{x_k}(y_k) \\ &= \sum_i I_i(x_1) \log \pi'_1(i) + \sum_{k=1}^{K-1} \sum_{(i,j)} I_i(x_k) I_j(x_{k+1}) \log a'(i, j) \\ &\quad + \sum_{k=1}^{K} \sum_i I_i(x_k) \log b'_i(y_k) \\ &= \sum_i I_i(x_1) \log \pi'_1(i) + \sum_{(i,j)} \log a'(i, j) \left(\sum_{k=1}^{K-1} I_i(x_k) I_j(x_{k+1}) \right) \\ &\quad + \sum_i \sum_{k=1}^{K} I_i(x_k) \log b'_i(y_k). \end{aligned} \tag{1.73}$$

Computing the conditional expectation of the log-likelihood can then be written as

$$\begin{aligned} Q(\theta,\theta') = &\sum_i \log \pi_1'(i) P_\theta[X_1 = i|\vec{Y} = \vec{y}] \\ &+ \sum_{(i,j)} \log a'(i,j) E_\theta[\sum_{k=1}^{K-1} I_i(X_k) I_j(X_{k+1})|\vec{Y} = \vec{y}] \\ &+ \sum_i E_\theta[\sum_{k=1}^{K} I_i(X_k) \log b_i'(Y_k)|\vec{Y} = \vec{y}]. \end{aligned} \tag{1.74}$$

Notice that the parameters π_1' and $a'(i,j)$ are factored out of the conditional expectation, so that we can maximize Q over these parameters directly. However, these parameters have constraints because they must represent probability distributions:

$$\sum_j \pi_1'(j) = 1, \tag{1.75}$$

$$\sum_j a(i,j) = 1, \qquad \text{for } 1 \le i \le N. \tag{1.76}$$

Performing a constrained optimization on these parameters, we obtain precisely the Baum-Welch formulas given in Equations 1.63 and 1.65.

The formulas for the observation process parameters b_i' were presented for the two situations of either a finite number of observation values or a Gaussian density. Both cases center on the expectation term

$$Q_b(\theta,\theta') = \sum_i E_\theta[\sum_{k=1}^{K} I_i(X_k) \log b_i'(Y_k)|\vec{Y} = \vec{y}].$$

For the case of a finite observation space, we rewrite Q_b as

$$\begin{aligned} Q_b(\theta,\theta') &= \sum_i \sum_m E_\theta[\sum_{k=1}^{K} I_i(X_k) I_m(Y_k) \log b_i'(m)|\vec{Y} = \vec{y}] \\ &= \sum_i \sum_m \log b_i'(m) E_\theta[\sum_{k=1}^{K} I_i(X_k) I_m(Y_k)|\vec{Y} = \vec{y}]. \end{aligned} \tag{1.77}$$

Again, we have a constraint on the parameter because $\sum_m b_i'(m) = 1$ for each i. Maximizing Q_b subject to this constraint leads to the Baum-Welch update Equation

1.66. For the case of a Gaussian density with mean μ_i' and variance $\sigma_i'^2$, we have

$$b_i'(y) = \frac{1}{\sqrt{2\pi}\sigma_i'} \exp(-\frac{(y-\mu_i')^2}{2\sigma_i'^2}), \tag{1.78}$$

so that $Q_b(\theta, \theta')$ is rewritten as

$$\begin{aligned} Q_b(\theta,\theta') &= \sum_i E_\theta[\sum_{k=1}^K I_i(X_k)\left(-\frac{1}{2}\log(2\pi\sigma_i'^2) - \frac{(Y_k-\mu_i')^2}{2\sigma_i'^2}\right)|\vec{Y}=\vec{y}] \\ &= \sum_i \left[-\frac{1}{2}\log(2\pi\sigma_i'^2)E_\theta[\sum_{k=1}^K I_i(X_k)|\vec{Y}=\vec{y}]\right. \\ &\quad \left. -\frac{1}{2\sigma_i'^2}E_\theta[\sum_{k=1}^K I_i(X_k)(Y_k^2 - 2\mu_i'Y_k + \mu_i'^2)|\vec{Y}=\vec{y}]\right]. \end{aligned} \tag{1.79}$$

The maximum occurs when each of the summands is maximized individually. Reminiscent of the calculations to perform least squares fitting for linear regression, we find the critical point (μ_i', σ_i') of the summand for each i, which corresponds exactly with the Baum-Welch update formulas of Equation 1.67 and 1.68. Note that the formula for μ_i' is independent of σ_i', so that the maximum must occur on the line defined by Equation 1.67, where σ_i' is allowed to vary. Applying the one-dimensional second derivative test on σ_i' verifies that when σ_i' satisfies Equation 1.68, $Q_b(\theta, \theta')$ is globally maximized.

We note that $Q(\theta, \theta')$ is actually the conditional expectation of the log-likelihood for the full model under θ', but calculated under the model θ,

$$Q(\theta,\theta') = E_\theta[\log L(\theta'; \vec{X}, \vec{Y})|\vec{Y}=\vec{y}]. \tag{1.80}$$

The generalization of the Baum-Welch algorithm, which is just a particular implementation of the EM algorithm, is to compute the conditional expectation $Q(\theta, \theta')$ for fixed θ, which gives the E-step its name. Then, the M-step involves maximizing over θ'. Because $\max_{\theta'} Q(\theta, \theta') \geq Q(\theta, \theta)$, this maximization will guarantee that the likelihood cannot decrease. The implementation of the EM algorithm for an HMM where the transition probabilities and observation densities depend on arbitrary parameters

is a direct extension of the Baum-Welch calculations, although the maximization may be more complex, particularly for parametrized transition probabilities (Michalek and Timmer, 1999).

1.4.4 Discussion of Hidden Markov Models

One of the primary obstacles to standard HMMs comes from the nature of the dependence of the observations on the hidden state process. In particular, the distribution of an observation Y_k depends only on a single state value X_k, conditionally independent of all other observations. If the distribution of Y_k depends on the recent history of the state process, $\{X_k, X_{k-1}, \ldots, X_{k-p}\}$, through a deterministic function $f(x_0, \ldots, x_p)$ in addition to an additive white noise sequence ε_k,

$$Y_k = f(X_k, X_{k-1}, \ldots, X_{k-p}) + \varepsilon_k, \tag{1.81}$$

then a corresponding meta-state Markov process can be used. By defining

$$\widetilde{X}_k = (X_k, X_{k-1}, \ldots, X_{k-p}), \tag{1.82}$$

the meta-state process $\widetilde{X}_k$ is also a Markov process, but with an enlarged state space of size N^{p+1}. Transitions involve shifting all coordinates to the right, and inserting a new coordinate in the first position according to the original transition probabilities. An especially important application of meta-states is when observations are modeled as a deterministic function of the hidden state with the addition of an autoregressive noise (Venkataramanan et al., 1998b,a; Qin et al., 2000). For noise modeled as a simple AR(1) process, it is possible to treat the problem without introducing meta-states (Qin et al., 2000), thus avoiding the general penalty of an exponentially growing state space.

Additional benefits arise by considering the likelihood in a more general context. The more general definition of likelihood arises in the context of a reference measure. The law governing the state and observation processes of the HMM must be absolutely

continuous with respect to this reference measure for every parameter. The Radon-Nikodym derivative of the law P_θ with respect to the reference measure ν, $dP_\theta/d\nu$, represents the likelihood of the model. It is beneficial for the reference measure to also represent a probability measure, P_0. Typically, this probability measure represents the law of a simpler model. Such a generalization allows for implementations of continuous time filtering (Dembo and Zeitouni, 1986; Zeitouni and Dembo, 1988).

In the context of discretized HMMs, a convenient reference measure corresponds to a trivial state and observation process, such as allowing both X_k and Y_k to be independent and identically distributed sequences, independent of each other (Elliott et al., 1997). That is, the observation Y_k is trivially dependent only on X_k by being independent of the entire state sequence. If π_0 represents the distribution of each state X_k under P_0, and b_0 represents the density of each observation Y_k and $\theta = (\pi_1, A, B)$ represents the parametrized model, then the Radon-Nikodym derivative of P_θ with respect to P_0 subject to the knowledge of $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$, which we denote by $\mathcal{F}_k$ (the σ-algebra generated by these random variables), is given by

$$\left.\frac{dP_\theta}{dP_0}\right|_{\mathcal{F}_k} = \frac{\pi_1(X_1)b_{X_1}(Y_1)a(X_1,X_2)b_{X_2}(Y_2)\cdots a(X_{K-1},X_K)b_{X_K}(Y_K)}{\pi_0(X_1)b_0(Y_1)\pi_0(X_2)b_0(Y_2)\cdots\pi_0(X_K)b_0(Y_K)}, \tag{1.83}$$

which has the form of a likelihood ratio. Under this formulation, the likelihood of partial information corresponds to a conditional expectation under P_0 of dP_θ/dP_0 conditioned on the available information. To be mathematically precise, if $\mathcal{G}$ represents the partial information ($\mathcal{G}$ is a sub σ-algebra of $\mathcal{F}_k$), then the likelihood given $\mathcal{G}$ is

$$\left.\frac{dP_\theta}{dP_0}\right|_{\mathcal{G}} = E_0[\frac{dP_\theta}{dP_0}|\mathcal{G}], \tag{1.84}$$

which corresponds in the earlier formulation to summing over the unknown states.

A number of recursive techniques have been developed by taking advantage of the reference measure approach through the use of martingale techniques (Elliott, 1994; Elliott et al., 1997). The general idea behind these techniques is to replace the forward-backward estimation procedure with a single forward, recursive sweep. In

order to complete the EM steps, however, the sums involved in the M-step (such as $\sum_{k=1}^{K} \gamma_{\theta,k}(i)$ and $\sum_{k=1}^{K-1} \widehat{J}_{\theta,k}(i,j)$) are replaced with conditional expectations of the actual sums, such as for the case of a finite observation space:

$$\sum_{k=1}^{K} \gamma_{\theta,k}(i) = E_\theta[\sum_{k=1}^{K} I_i(X_k)|\vec{y}], \tag{1.85}$$

$$\sum_{k=1}^{K-1} \widehat{J}_{\theta,k}(i,j) = E_\theta[\sum_{k=1}^{K-1} J_k(i,j)|\vec{y}], \tag{1.86}$$

$$\sum_{k=1}^{K} \widehat{O}_{\theta,k}(i,m) = E_\theta[\sum_{k=1}^{K} O_k(i,m)|\vec{y}], \tag{1.87}$$

where $O_k(i,m) = I_i(X_k)I_m(Y_k)$ indicates that at time k the state is i and the observation m. By working in the reference measure P_0 and converting results to the true measure P_θ by an application of a version of Bayes Theorem, these sums can be directly computed recursively by adding the effect of one observation at a time. There are two primary advantages to this approach. Perhaps the most significant of these is that it allows for on-line estimation. The forward-backward algorithm requires the complete set of data in order to do the backward iteration. Adding a new data point would require updating all of the backward variables β_k. The recursive approach only involves updating each of the required sums as well as the forward variable α_k, allowing parameters to be updated as data arrives (Ford and Moore, 1998). The second advantage is a reduced requirement for memory. The forward-backward calculation requires essentially $2KN$ memory locations to account for the forward and backward variables, where K is the data-length and N is the size of the state space. The recursive techniques essentially require N memory locations for each quantity which must be updated. As an example, in the case of the standard Baum-Welch update with M different possible observation values, we need to update the forward variables α, the N^2 different jump counts $\sum_{k=1}^{K-1} \widehat{J}(i,j)$, for $(i,j) \in N^2$, and the NM different observation counts $\sum_{k=1}^{K} \widehat{O}(i,m)$. The sum of occupation times can be recovered from the jumps. This results in a memory requirement of $N(N^2 + NM + 1)$, but which is

independent of the length of the data. The primary disadvantage is that each quantity being updated has roughly the same computational load as a single step in the update of the forward variables, although these calculations are essentially decoupled so that parallelizing might compensate for some of the increased burden.

Finally, because of the asymptotic properties of MLEs, confidence intervals and hypothesis tests can be constructed for the model parameters. For example, the negative of the Hessian of the likelihood, which measures the curvature, evaluated at the MLE $\widehat{\theta}$ gives the inverse of the covariance matrix of the estimator $\widehat{\theta}$ (Qin et al., 2000). The diagonal terms give the variance, while off-diagonal terms provide information about the correlation. The asymptotic normality of the MLE allows for the computation of confidence intervals using standard statistical techniques using the normal distribution. In addition, it formally justifies the use of hypothesis tests for nested models (Giudici et al., 2000). Suppose that $\widehat{\theta}_1$ is the MLE over all models under a particular parametrization, and let $\widehat{\theta}_0$ be the MLE for a restricted class of models within this parametrization. A likelihood ratio test determines whether the inequality $\frac{L(\widehat{\theta}_1;\vec{y})}{L(\widehat{\theta}_0;\vec{y})} \geq 1$ is significant enough to reject the null hypothesis that the restricted class represents the model. Asymptotic normality converts this question into a χ^2 test, using the asymptotic result that

$$2(\log L(\widehat{\theta}_1;\vec{y}) - \log L(\widehat{\theta}_0)) \sim \chi^2_k, \tag{1.88}$$

where the degrees of freedom k specifies the difference in the number of free parameters between the two hypotheses. In addition to requiring the two models to be nested, this technique further requires that the parameters are in the interior of the parameter space and that the models are identifiable, although there are situations where a more general parametrization can provide a suitable nesting (Wagner and Timmer, 2001).

Chapter 2

A Mathematical and Statistical Model

The extensive measurements of single-molecule assays of kinesin contain information about the protein's mechanical cycle. While past analyses of these data have successfully extracted some information about average behavior of the cycle, the highly detailed recordings suggest that more precise analysis might yield even more information. One approach to tracting the fine-scale information is to use model-based filtering. Hidden Markov models implement many of the features of the kinesin-bead assay and, therefore, promise to provide a useful implementation of a model-based filter.

Two major aspects distinguish the hidden Markov model that will be used for analyzing kinesin experiments from standard hidden Markov models. First, sequential measurements of the bead position will be correlated because of the viscous drag through the fluid and the elastic behavior of the bead relative to the kinesin and the trap. Fortunately, the correlation can be represented as a simple autoregressive model, which is easily incorporated into hidden Markov model theory without needing to expand the state space (Qin et al., 2000). Second, because the microtubule track creates a periodic lattice on which kinesin steps directionally, the hidden Markov model structure must be adapted to incorporate a lattice structure. Strictly speaking, the model will not be recurrent, as kinesin only rarely makes backward steps. The key to this adaptation lies in the chemical equivalence of all lattice sites resulting from the use of a force clamp. Consequently, information about a single mechanochemical state can be gathered from brief visits to different, but equivalent lattice sites.

This chapter focuses on the mathematical details required to implement a hidden Markov model filtering procedure to extract information from the experimental

data. In particular, we first discuss a physical model in continuous time for a kinesin molecule and an attached bead. Next, we discuss how to discretize this continuous time model to establish an appropriate discrete time model. Third, using this discrete model, we extend the framework of hidden Markov models to accommodate the non-recurrent, periodic structure of the kinesin-microtubule interaction. This involves explicitly defining the likelihood of the model, demonstrating how to compute the conditional relative log-likelihood function $Q(\theta, \theta')$, and implementing the EM algorithm by maximizing the function Q relative to θ'.

2.1 Continuous Time Model for Kinesin

The use of Markov processes to model motor proteins has been well established (Huxley and Simmons, 1971; Leibler and Huse, 1991; Schnitzer et al., 2000; Fisher and Kolomeisky, 2001). Traditionally, each state in the Markov process corresponds to a particular stage of the enzymatic cycle. Transitions between these stages are modeled as occurring after exponentially distributed times, characterized by corresponding transition rates. Typical chemical events for kinesin attached to a microtubule include the binding of ATP, hydrolysis of ATP, and the subsequent release of the products ADP and P_i. Furthermore, kinesin has two heads, each with an ATP-binding site and capable of catalyzing hydrolysis. However, whether both heads perform hydrolysis and how they might interact remains unclear (Hua et al., 2002). During the process of this uncertain chemical cycle, kinesin also undergoes a net displacement of approximately 8.2 nm for each ATP hydrolyzed. It is generally believed that different chemical states correspond to different physical conformations of the protein structure. The sequence of conformations corresponding to the chemical cycle induce the motion, but the conformations themselves remain unclear. Consequently, we model kinesin using a Markov process with a small, finite number ($N \leq 4$) of mechanochemical states. We will not assign an *a priori* physical interpretation to

the modeled states, which will be characterized only in terms of their parameters. Further, we remark that we model kinesin as a whole and not the individual heads.

We now state the mathematical model for kinesin. We assume that kinesin can be modeled as proceeding through N mechanochemical states for each overall step along the microtubule, which we will simply label with the numbers $\{1, 2, \ldots, N\}$. We will denote this state as a function of time by $C(t)$, where $t \in [0, t_{\max}]$ identifies the time during the experiment. We also introduce $X(t)$, which will take on integer values, to represent the lattice site at time t, with $X(0) = 0$ and positive values corresponding to the plus direction of the microtubule. We will let the joint stochastic process $Z(t) = (X(t), C(t))$ represent the complete state of kinesin at time t. The state space for $Z(t)$ is then $\mathcal{Z} = \mathbb{Z} \times \mathbb{Z}_N$.

We model the stochastic process $\{Z(t); t \geq 0\}$ as a Markov jump process. To characterize the dynamics of this process, we must specify, for each state $z \in \mathcal{Z}$, the instantaneous rate of a transition $\lambda_z > 0$ and the jump distribution μ_z. The rate λ_z is the rate of an exponential random variable which represents the dwell time for the process $Z(t)$ to remain in the state z. In other words, suppose $Z(t_0) = z$ and define the random variable $T_z = \inf\{t > 0 : Z(t + t_0) \neq z\}$, the amount of time until $Z(t)$ leaves z. Then T_z is exponentially distributed with a mean time $E_\theta[T_z] = 1/\lambda_z$. The jump distribution μ_z gives the distribution of the process at the time of the jump,

$$P_\theta[Z(t_0 + T_z) = z_f | Z(t_0) = z] = \mu_z(z_f). \tag{2.1}$$

The probability distribution for the process $Z(t)$ is determined from these transition rates according to the initial value problem involving the linear system of differential equations, known as the Kolmogorov forward equation, and an initial condition,

$$\frac{d}{dt} P_\theta[Z(t) = z_f] = \sum_{z \neq z_f} \mu_z(z_f) \lambda_z P_\theta[Z(t) = z] - \lambda_{z_f} P_\theta[Z(t) = z_f], \tag{2.2}$$

$$P_\theta[Z(0) = z] = \pi^0_\theta(z). \tag{2.3}$$

The specifics of our model place various restrictions on these modeling parameters.

Because of the periodicity of the model, the transition rates λ_z depend only on the state of the model, and not the site x: $\lambda_{(x,c)} = \lambda_c$. Similarly, the jump distribution can only depend on the initial and final states, c and c', respectively, and the difference in lattice sites:

$$\mu_{(x,c)}(x', c') = \mu_{(0,c)}(x' - x, c') = \mu_c(\Delta x, c'). \tag{2.4}$$

The structure, or topology, of the process is determined by identifying which transitions are allowed. The simplest model, which the present work adopts, imposes the restriction that transitions only occur between neighboring states. That is, if we define a distance between two states $z = (x, c)$ and $z' = (x', c')$ as

$$d(z; z') = |N(x - x') + (c - c')|,$$

then the jump distribution will vanish, $\mu_z(z') = 0$, for states farther apart than a unit length, $d(z; z') > 1$. That is, for each state c, there is only the probability of going forward $p_c = \mu_c(0, c + 1)$ and the probability of going backward $q_c = \mu_c(0, c - 1) = 1 - p_c$. We have introduced the extended states $c = 0$ and $c = N + 1$ according to the equivalence relations $(x, 0) \equiv (x-1, N)$ and $(x, N+1) \equiv (x+1, 1)$ to simplify notation. An equivalent representation for this model is to introduce forward and backward transition rates, u_c and v_c, respectively, according to the invertible relationship

$$\begin{array}{l} u_c = p_c \lambda_c \\ v_c = q_c \lambda_c \end{array} \qquad \Longleftrightarrow \qquad \begin{array}{c} \lambda_c = u_c + v_c \\ p_c = u_c/\lambda_c \\ q_c = v_c/\lambda_c, \end{array} \tag{2.5}$$

which is illustrated by the standard kinetic scheme,

$$\cdots (x-1, N) \underset{v_1}{\overset{u_N}{\rightleftharpoons}} (x, 1) \underset{v_2}{\overset{u_1}{\rightleftharpoons}} (x, 2) \underset{v_3}{\overset{u_2}{\rightleftharpoons}} \cdots \underset{v_N}{\overset{u_{N-1}}{\rightleftharpoons}} (x, N) \underset{v_1}{\overset{u_N}{\rightleftharpoons}} (x+1, 1) \cdots . \tag{2.6}$$

In addition to specifying the topology of the Markov process $Z(t)$, we must also indicate how the process depends on the experimental conditions, particularly the ATP concentration and the external load being exerted by the optical force clamp. In order to motivate and understand these modeling constraints, it is helpful to discuss

briefly some thermodynamic ideas for enzymatic cycles. The Markov process $C(t)$ can also be considered as representing an enzymatic cycle, where $X(t)$ represents the net number of forward cycles that have occured. Thermodynamics impose constraints on the transition rates, in that all transitions must be reversible such that under equilibrium conditions, detailed balance must be satisfied.

In order to consider properly the thermodynamic constraint of detailed balance, we briefly discuss the energetics of the enzymatic cycle (Hill, 1989). The different mechanochemical states of the model, $c = 1, \ldots, N$, will each have a corresponding energy level, E_c. The presence of an external force creates stresses within the protein, so that these energy levels may depend on the force F. In addition, we let ℓ_c represent the size of the conformational shift for the corresponding transition, with $\sum_c \ell_c = d$. By moving a distance, ℓ_c, against a force, F, the protein expends the energy $F\ell_c$ in work. Thus, the energy during a transition $c \to c+1$ subject to an isotonic force F will decrease by the amount

$$\Delta E_c(F) = E_c - E_{c+1} - F\ell_c. \tag{2.7}$$

If the transition involves binding ATP, then we must also subtract the chemical potential of ATP, μ_{ATP}. Similarly, the release of ADP or phosphate would require the addition of the chemical potential of ADP or P_i, given by μ_{ADP} or μ_{P_i}, respectively. The transition rates governing this transition, u_c and v_{c+1}, will depend on the force as well. The detailed balance condition requires that the rates satisfy

$$\frac{u_c(F)}{v_{c+1}(F)} = e^{\Delta E_c(F)/k_B T}. \tag{2.8}$$

By completing a full cycle, we obtain the requirement that

$$\frac{\prod_c u_c(F)}{\prod_c v_c(F)} = e^{(X_{\mathrm{ATP}} - Fd)/k_B T}, \tag{2.9}$$

where X_{ATP} here represents the thermodynamic force, $X_{\mathrm{ATP}} = \mu_{\mathrm{ATP}} - \mu_{\mathrm{ADP}} - \mu_{\mathrm{P}_i}$, arising from the hydrolysis of ATP. The thermodynamic force does not depend on

the motor protein at all; it depends only on the concentrations of the ATP, ADP, and phosphate relative to the equilibrium constant. Thus, any thermodynamically consistent model must have the relation

$$\frac{\prod_c u_c(F)}{\prod_c v_c(F)} = \frac{\prod_c u_c(0)}{\prod_c v_c(0)} e^{-Fd/k_B T}, \tag{2.10}$$

independent of the chemical concentrations.

The model of Fisher and Kolomeisky (Fisher and Kolomeisky, 2001) implements a Markov model for kinesin that satisfies this constraint in a reasonably straight-forward manner. As this model matched the global behavior of the VSB kinesin experiments, we adopt the FK parametrization for load dependence using (recall Equations 1.22 and 1.23)

$$u_c = u_c^0 e^{-\vartheta_c^+ Fd/k_B T}, \tag{2.11}$$

$$v_c = v_c^0 e^{+\vartheta_c^- Fd/k_B T}. \tag{2.12}$$

The rate suppression factors, ϑ^+ and ϑ^-, incorporate the sizes of the substeps for the transitions as well as the load-dependence of the energy levels of the macromolecule. That is, recalling the original detailed balance condition for transitions in Equation 2.8 with loads 0 and F, we must have

$$\vartheta_c^+ + \vartheta_{c+1}^- = \frac{\ell_c}{d} + \frac{E_c(0) - E_{c+1}(0)}{Fd} - \frac{E_c(F) - E_{c+1}(F)}{Fd}. \tag{2.13}$$

Thus, if the relative energy levels do not depend on the applied force, then the sum $\vartheta_c^+ + \vartheta_{c+1}^-$ corresponds to the fractional size of the substep for the transition.

Because kinesin must bind and subsequently hydrolyze an ATP molecule to proceed through the mechanical cycle, we will arbitrarily designate the first internal state $C = 1$ as the state before kinesin binds ATP and $C = 2$ as the state after kinesin has bound ATP. Thus, this first transition rate u_1 will be a second-order rate which depends on the concentration of ATP, so that we will model it as

$$u_1^0 = k_0[\text{ATP}]. \tag{2.14}$$

Since passing through the chemical cycle in reverse would correspond to formation of ATP by combining ADP and a phosphate, a reverse rate should depend on these concentrations as well. In principle, if one of the products, say ADP, were released in the transition $c \to c+1$, then the reverse rate v_{c+1} would have an analogous second-order relation depending on [ADP]. However, these concentrations are not measured, and experimentally are controlled to be very dilute by an ATP-regenerating system. Furthermore, we have no ability in the present work to identify which transitions might correspond to product release. To account for these problems, Fisher and Kolomeisky also introduce [ATP]-dependence in the reverse rate, leaving state $C = 1$ to return to $C = N$, by using a phenomenological form (Fisher and Kolomeisky, 2001),

$$v_1^0 = k_0' \frac{[\mathrm{ATP}]}{\sqrt{1 + [\mathrm{ATP}]/c_0}}. \tag{2.15}$$

The nonlinear dependence was primary introduced in that model to account for an [ATP]-dependence in the stall force, by reducing the probability of reverse cycles for high ATP concentrations.

2.2 Continuous Time Model for the Bead

Having specified a model for kinesin, we now turn to establishing a model for the attached bead. Forces exerted on the bead come from three sources. First, the kinesin exerts a force pulling the bead forward. Second, the optical trap exerts a force in the opposite direction. Finally, the bead interacts with the surrounding fluid. This interaction creates a viscous drag that opposes motion through the fluid in combination with a diffusive scattering because of the random collisions with thermal fluid particles. One way to incorporate these physical aspects is to model the position of the bead as a forced Ornstein-Uhlenbeck process.

The use of a force clamp in the experimental design makes a number of simplifications possible. In particular, the elasticity of the stalk of kinesin is rather complicated.

However, the force clamp maintains a relatively constant stretch in the stalk so that the force exerted by the stalk due to deviations from the equilibrium position can be approximated with a linear response, or in other words as a simple spring. The force of the optical trap is also well-approximated as being proportional to displacement. Thus, the coupled system of the force-clamp and the kinesin stalk can be modeled as a single spring, characterized by an effective spring stiffness k_{trap} and an equilibrium position.

The equilibrium position of the bead, $Y^0(t)$, will depend on the position of the kinesin on the microtubule, $Y^0(t) = f(Z(t))$, because the bead is connected to the kinesin through the stalk and anchored at the neck. By assuming that each state c in the kinesin cycle corresponds to a particular physical conformation of the protein molecule, transitions between two states will correspond to a physical shift in the position of the anchor site which, in turn, translates into an equivalent shift in the equilibrium position of the bead. After a complete cycle finishes, the new equilibrium position will have moved by a total distance d, the size of a lattice increment. Thus, we let $\epsilon(c) = \epsilon_c$ represent the cumulative fraction of a step that has occurred between state 1 and state c. Defining κ as the equilibrium position of the bead for the initial lattice site $X = 0$ and state $C = 1$, the equilibrium position can be modeled as

$$Y^0(t) = f(Z(t)) = \kappa + d\big(X(t) + \epsilon(C(t))\big). \tag{2.16}$$

The fluid interaction leads to two contributions to the model. First, the fluid exerts a force proportional to and in the opposite direction of the velocity of the bead, with a viscous drag coefficient of γ. Second, the fluid imparts thermal fluctuations to the bead. We will model these fluctuations as an additional white noise force, $\xi(t)$. The scale of this force depends on the viscosity as well as the thermal energy of the system, $k_B T$, according to the Einstein relation. Together, these forces combine to form the Langevin differential equation

$$\gamma \frac{dY}{dt}(t) = -k_{\text{trap}}\big(Y(t) - f(Z(t))\big) + \sqrt{2\gamma k_B T}\xi(t). \tag{2.17}$$

Mathematically, we rewrite this as a stochastic differential equation subject to Ito calculus with a Wiener process (Brownian motion) $W(t)$,

$$dY(t) = -\alpha\big(Y(t) - f(Z(t))\big)\,dt + \eta\,dW(t), \tag{2.18}$$

where we have introduced the rescaled parameters

$$\alpha = \frac{k_{\text{trap}}}{\gamma}, \tag{2.19}$$

$$\eta^2 = \frac{2k_BT}{\gamma}. \tag{2.20}$$

We can directly calculate a strong solution to the stochastic differential equation governing $Y(t)$. Using the method of variation of parameters, we find that

$$Y(t) = Y(0)e^{-\alpha t} + \alpha\int_0^t e^{-\alpha(t-s)}f(Z(s))ds + \eta\int_0^t e^{-\alpha(t-s)}dW(s). \tag{2.21}$$

If we label $\tau_0 = 0$ and τ_i as the time of the i^{th} transition, then for s in the interval $[\tau_i, \tau_{i+1})$, $Z(s) = Z(\tau_i)$ and we have $f(Z(s)) = f(Z(\tau_i))$. Let $N(t)$ represent the number of transitions that occur in the interval $[0, t]$. Define Δf_i as the size of the change in kinesin position at time τ_i,

$$\Delta f_i = f(Z(\tau_i)) - f(Z(\tau_{i-1})), \tag{2.22}$$

so that we can write $f(Z(s))$ as a weighted sum of indicator functions I_A,

$$f(Z(s)) = f(Z(0)) + \sum_{i=1}^{N(t)} \Delta f_i I_{[\tau_i,t]}(s). \tag{2.23}$$

Thus, we can rewrite the solution for $Y(t)$ as

$$Y(t) = f(Z(t)) - \sum_{i=1}^{N(t)} e^{-\alpha(t-\tau_i)}\Delta f_i + U(t), \tag{2.24}$$

where

$$U(t) = \big(Y(0) - f(Z(0))\big)e^{-\alpha t} + \eta\int_0^t e^{-\alpha(t-s)}dW(s). \tag{2.25}$$

Note that $U(t)$ is an Ornstein-Uhlenbeck process satisfying the stochastic differential equation (van Kampen, 1981)

$$dU(t) = -\alpha U(t)dt + \eta\, dW(t), \tag{2.26}$$

$$U(0) = Y(0) - f(Z(0)), \tag{2.27}$$

so that $U(t)$ is a Gaussian random variable with mean and variance,

$$E[U(t)] = U(0)e^{-\alpha t}, \tag{2.28}$$

$$\mathrm{Var}[U(t)] = \frac{\eta^2}{2\alpha}(1 - e^{-2\alpha t}). \tag{2.29}$$

If $f(Z(t))$ were constant in time, then $Y(t)$ would also be an Ornstein-Uhlenbeck process. Including the sudden changes in position Δf_i from transitions by kinesin, we introduce an exponentially decaying memory effect of when these jumps occur.

2.3 Discrete Time Approximate Models

Although the natural physical processes associated with the kinesin and bead system occur continuously in time, the experimental constraints for observations dictate that measurements occur at discrete times separated by a constant time interval. We let Δt represent the sampling interval, and t_0 be the time of the first observation. Then the subsequent sampling times differ from t_0 by integer multiples of Δt. In particular, the k^{th} observation takes place at time $t_k = t_0 + k\Delta t$. So rather than having access to the position of the bead at all times during the experiment, we are limited to the sampled positions, $\{Y_k = Y(t_k); k = 0, \ldots, K\}$, where $K+1$ is the number of recorded observations. The states of Z at the corresponding times, $\{Z_k = Z(t_k); k = 0, \ldots, K\}$, remain hidden. Thus, we consider the models for Y and Z in discrete time.

Creating the discrete time hidden process Z_k is straightforward. Because $Z(t)$ is a Markov jump process, the discretized sequence $\{Z_k\}$ is a Markov chain which is specified by its transition probability matrix $A = (a(z; z'))$,

$$P_\theta[Z_{k+1} = z' | Z_k = z] = a(z; z'), \tag{2.30}$$

and the initial distribution π_θ^0. If we consider the system of differential equations given by Equation 2.2, then for each fixed z, the matrix entries $a(z;z') = \pi_z(z', \Delta t)$ will be the solution to the initial value problem

$$\frac{d}{dt}\pi_z(z',t) = \sum_{z^* \neq z'} \mu_{z^*}(z')\lambda_{z^*}\pi_z(z^*,t) - \lambda_{z'}\pi_z(z',t), \qquad z' \in \mathcal{Z}, \tag{2.31}$$

$$\pi_z(z',0) = \delta_{z;z'}. \tag{2.32}$$

Introducing the generator matrix $Q = (q(z,z'))$, where

$$q(z;z') = \begin{cases} \lambda_z\mu_z(z') & \text{if } z \neq z' \\ -\lambda_z & \text{if } z = z', \end{cases} \tag{2.33}$$

the system of equations can be succinctly stated as a matrix problem. Let $\pi_z(t)$ be the row vector-valued function of time that contains the transition probabilities $\pi_z(z',t)$. Then Equation 2.31 can be rewritten as

$$\frac{d}{dt}\pi_z(t) = \pi_z(t)Q. \tag{2.34}$$

A well-known result of the theory of systems of linear differential equations is that the transition matrix will be the matrix exponential $A = \exp(\Delta t Q)$. In the simple case where transitions occur only between nearest neighbors, as given by Equation 2.5, Q can be represented as a tri-diagonal matrix.

Although the actual transition matrix A is infinite, there are a number of simplifications that make a finite representation appropriate. Because of the periodicity induced by the lattice structure, the rows of A will be shifted copies of each other

$$a(x,c;x',c') = a(0,c;x'-x,c') = \pi_c(\Delta x = x'-x, c') \tag{2.35}$$

so that only N rows, π_c, $c = 1,\ldots,N$ need to be calculated. Each of these rows is also infinite in each direction, which cannot be physically implemented. However, because of the topology imposed on the transitions, the probability of hopping to sites in time Δt becomes exponentially small except for sites sufficiently close to the

initial site. Thus, for implementation, we will replace the original, infinite model for the transition matrix with an approximate bounded jump model where transitions with $|\Delta x| > X_{\max}$ are prohibited.

We next consider Y_{k+1}, the observation of the bead at time t_{k+1}, as it relates to earlier observations and the state of the hidden process Z. Recalling the earlier formula, we have an exact representation

$$Y_{k+1} = f(Z_{k+1}) - \sum_{i=1}^{N_{k+1}} \Delta f_i e^{-\alpha(t_{k+1}-\tau_i)} + U_{k+1}, \tag{2.36}$$

where the sum

$$- \sum_{i=1}^{N_{k+1}} \Delta f_i e^{-\alpha(t_{k+1}-\tau_i)} \tag{2.37}$$

represents the exponentially decaying memory in the system of all past position increments. This memory term presents a serious complication. First of all, the sum includes information about events that occur at times in between sampling, including the possibility of leaving and returning to the state without being observed. Secondly, the sum includes the increments for each of the sampling intervals represented in the data, exponentially damped by the time elapsed from the event. These two issues have prevented our finding an implementation to account for the memory term. In principle, this memory term might be modeled as a third hidden process. But the increased complexity for our algorithms would make the computations too costly to be effective. Consequently, we model the bead as though the memory term were absent so that the observation process takes a particularly simple form,

$$Y_{k+1} = f(Z_{k+1}) + U_{k+1}. \tag{2.38}$$

When the Ornstein-Uhlenbeck process $U(t)$ is discretely sampled with a time interval Δt, the resulting process U_k is a simple autoregressive process (AR(1)) (van Kampen,

1981). That is, we can write

$$\rho = e^{-\alpha \Delta t}, \tag{2.39}$$

$$\sigma = \frac{\eta^2}{2\alpha}(1 - e^{-2\alpha \Delta t}), \tag{2.40}$$

$$U_{k+1} = \rho U_k + \sigma \varepsilon_{k+1}, \tag{2.41}$$

where ε_k is a Gaussian white noise—an independent, identically distributed sequence of standard normal random variables. Thus, an alternative representation for Y_{k+1}, in the absence of the memory term, would be

$$Y_{k+1} = f(Z_{k+1}) + \rho(Y_k - f(Z_k)) + \sigma \epsilon_{k+1}. \tag{2.42}$$

If the initial error term $U_0 = Y_0 - f(Z_0)$ has a normal distribution with mean zero and the stationary variance,

$$\sigma_0^2 = \frac{\sigma^2}{1 - \rho^2} = \frac{\eta^2}{2\alpha}, \tag{2.43}$$

then the sequence U_k of errors will be a stationary process.

Before proceeding with the discussion, we wish to discuss the role of the memory term in more detail. For each time at which the kinesin makes a transition, τ_i, the equilibrium position of the bead spontaneously changes by a displacement Δf_i. An equivalent (nonphysical) system would correspond to a bead that was instantaneously displaced by the opposite amount $-\Delta f_i$ while the equilibrium position remains fixed. The bead would continue to relax toward the equilibrium, while fluctuations continued to add random perturbations. For each step kinesin takes, the bead requires some time to relax in that direction. During this time, observations of the bead will include a bias towards the previous position. Next, suppose that kinesin takes a step and then returns. The bead will have already begun to relax to compensate for the original step when the kinesin returned to the original position. Then, at the subsequent sampling time, the state of kinesin would remain unchanged but there would be a bias in the bead position in the direction of this missed event. For states in rapid equilibrium,

caused by fast forward and backward transition rates, the effect of the memory term would be to average the two positions, as though the effective equilibrium position of the bead were somewhere between the two positions. The amount of information lost by neglecting this memory term is predominantly controlled by the relaxation time of the bead relative to the transition rates.

In addition to characterizing the models for kinesin and for the bead, we also must consider the amount of information available, which corresponds to the mathematical concept of a σ-field. A σ-field provides the mathematical description to determine for which sets one can compute a probability. When information increases as time progresses, the increasing family of σ-fields is known as a filtration. An ideal observer would have knowledge of the true position and mechanochemical state of the kinesin protein as well as the position of the bead. Thus, at time t_k, an ideal observer would have full information of all prior states, so that the full filtration, $\{\mathcal{F}_k\}$ is the increasing sequence of σ-fields created by completing

$$\mathcal{F}_k^0 = \sigma(\{Z_0, \ldots, Z_k, Y_0, \ldots, Y_k\}). \tag{2.44}$$

However, the experimentalist only has access to the position of the bead at these times. Thus, the experimentalist's filtration, $\{\mathcal{Y}_k\}$, corresponds to the σ-field at time k given by the completion of

$$\mathcal{Y}_k^0 = \sigma(\{Y_0, \ldots, Y_k\}). \tag{2.45}$$

The state process Z_k is $\mathcal{F}_k$-measurable, while Y_k is measurable with respect to both $\mathcal{Y}_k$ and $\mathcal{F}_k$. Any mathematical estimates based on the first k observed data points will also need to be $\mathcal{Y}_k$-measurable. Similarly, we define the hidden information filtration $\mathcal{H}_k$ as the complete filtration generated by $\sigma(Z_0, \ldots, Z_k)$. It will also be useful to define another family of σ-fields $\mathcal{M}_{i,j} = \sigma(\mathcal{H}_i, \mathcal{Y}_j)$ which corresponds to the information of the hidden process through step i and the observations through step j. Note that $\mathcal{M}_{k,k} = \mathcal{F}_k$.

2.4 Hidden Markov Model Likelihood

The class of models described above is characterized by a number of parameters. To describe the mechanochemical cycle of kinesin, we must specify a total of $2N$ forward and backward transition rates,

$$\vec{u} = (u_1, \ldots, u_N), \tag{2.46}$$

$$\vec{v} = (v_1, \ldots, v_N). \tag{2.47}$$

To describe the displacement of kinesin, and hence the equilibrium position of the bead, we require the lattice step-size d as well as the cumulative relative steps ϵ_c for each of the mechanochemical states $c = 2, \ldots, N$. By definition, we always have $\epsilon_1 = 0$. To complete the description of the equilibrium position, we also need the baseline position κ corresponding to the kinesin in state $Z = (0, 1)$. To parametrize the autoregressive behavior of the bead observations, we finally need the autocorrelation coefficient ρ and the noise scale σ. The stationary noise scale, which appears for the first bead observation, will be given by

$$\sigma_0^2 = \frac{\sigma^2}{1 - \rho^2}. \tag{2.48}$$

The collection of all such parameters will be summarized by the notation θ. When these quantities are further parametrized, such as to account for detailed balance, θ will implicitly include the understood parametrization. We have, at our disposal, extensive experimental data for the kinesin-bead system from the experiments of Visscher et al. (1999). We wish to determine the parameter θ which best describes the experimental data.

In fact, the experiment results in a number of independent recordings, each of which is recorded in a separate data file. Thus the full parametrization must include enough parameters to characterize all of these separate files. The files are naturally grouped according to the experimental conditions of ATP concentration and force

clamp load. Let $\mathcal{G}$ represent the set of all experimental groupings by [ATP] and load. For each such experimental condition $g \in \mathcal{G}$, we let $\mathcal{E}_g$ represent the listing of all experimental runs, corresponding to individual data files, in the group g. When transition rates are not parametrized, then each experimental condition will need a separate set of rate parameters. When transition rates are parametrized to deal with [ATP] but not the load, each common load will have a separate set of rate parameters. If the load is parametrized by detailed balance, but without parametrizing ATP concentration, then each ATP concentration level will have the base transition rates u_c^0, v_c^0 as well as the detailed balance parameters $\vartheta_c^+, \vartheta_c^-$. For rates that are completely parametrized to deal with both load and [ATP], there will only be one set of rates and detailed balance parameters for all experiments.

Parameters describing the observation process also will be categorized either at the level of an experimental condition or by individual files. Because the stiffness of the trap is changed in order to obtain different loads and because the stalk does not respond uniformly to different loads, we will have a different autocorrelation coefficient ρ and noise scale σ for each experimental grouping. In addition, each data file represents a different run of a kinesin-tethered bead along a microtubule. As such, each file needs to identify a separate offset parameter κ to allow for the different starting points. Furthermore, the recorded observations are actually a one-dimensional projection of a two-dimension trajectory. Experimentally, the microtubule was visually aligned with one of the observation axes, and the positions of the beads in this direction is recorded in the data files. Errors in alignment or non-straight microtubules will lead to an apparent shortening of the lattice size for the individual run. Thus, we also allow a different lattice step-size parameter d for each data file. Nevertheless, when computing transition probabilities, we will actually use $d' = 8.2$ nm when dealing with the detailed balance condition, which we do not consider to be a free parameter as it is confounded with the parameters $\vartheta_c^{\pm}$ and merely sets a length scale. The cumulative substep parameters ϵ_c will be allowed to vary by experimental condition.

Parameter	Description	Global	[ATP]‡, Load†	File
d	lattice step length			ABCD
κ	baseline offset			ABCD
ϵ_c	fractional substep		ABCD	
ρ	autocorrelation		ABCD	
σ	noise scale		ABCD	
$u_1^0/k_0, v_1^0/k_0',$ $u_2^0, v_2^0, \ldots, u_N^0, v_N^0$	transition rates	D	AB†C‡	
$\vartheta_1^+, \vartheta_1^-, \ldots, \vartheta_N^+, \vartheta_N^-$	det. bal. factors	D	C‡	

TABLE 2.1. Parameters are grouped according to how they are classified for each of four different possible implementations: (A) Each experimental condition is treated independently, (B) [ATP] is parametrized, (C) Load is parametrized with detailed balance, and (D) Both [ATP] and Load are parametrized. Superscripts ($\dagger, \ddagger$) indicate multiple experimental conditions grouped together by [ATP] or by Load, while the other condition is parametrized.

This organization of the parameters is summarized in Table 2.1.

Preparing for the calculation of the likelihood of the parameter θ, we introduce an appropriate reference probability measure P_0 relative to which all of the parametrized probability measures P_θ will be absolutely continuous. Under P_0, the sequence of observations $Y_0, Y_1, \ldots$ will be a sequence of independent and identically distributed standard normal random variables. Also, the state sequence $Z_0, Z_1, \ldots$ will be a sequence of independent and identically distributed sequence of random variables taking values in the state space $\mathcal{Z}$ with the distribution

$$P_0[Z_l = (x, c)] = \frac{2^{-x}}{3N}, \tag{2.49}$$

which is chosen based on its property that it assigns positive probability to every possible state. To account for each of the experiments, identified by $g \in \mathcal{G}$ and $e \in \mathcal{E}_g$, we will assume that there is an independent and identically distributed pair of sequences, $(Z_l^{g,e}, Y_l^{g,e})$ for each experiment to be considered, and we similarly define the corresponding specific filtrations $\mathcal{F}_k^{g,e}$ and $\mathcal{Y}_k^{g,e}$.

Given a parameter θ, an experimental condition $g \in \mathcal{G}$, and a particular experi-

ment $e \in \mathcal{E}_g$, we introduce the $\mathcal{F}_k^{g,e}$-measurable random variable $\Lambda_k^{g,e}(\theta)$ which depends on $Z_0^{g,e}, \ldots, Z_k^{g,e}, Y_0^{g,e}, \ldots, Y_k^{g,e}$ according to the relation

$$\Lambda_k^{g,e}(\theta) = \frac{\pi_\theta^0(Z_0^{g,e})\phi_{\sigma_0}(U_0^{g,e})}{\pi_0(Z_0^{g,e})\phi_1(Y_0^{g,e})} \prod_{l=1}^{k} \frac{a(Z_{l-1}^{g,e}; Z_l^{g,e})\phi_\sigma(U_l^{g,e} - \rho U_{l-1}^{g,e})}{\pi_0(Z_l^{g,e})\phi_1(Y_l^{g,e})}, \tag{2.50}$$

where we write the abbreviation $U_l^{g,e}$ to represent the autoregressive residual

$$U_l^{g,e} = Y_l^{g,e} - f(Z_l^{g,e}), \tag{2.51}$$

and ϕ_σ is the density for a zero-mean Gaussian random variable with variance σ^2,

$$\phi_\sigma(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\xi^2/2\sigma^2}. \tag{2.52}$$

The transition matrix elements $a(z; z')$, the bead equilibrium position $f(z)$, and the autoregressive parameters ρ and σ also depend on the experiment (g, e) as discussed above, but we will suppress this dependency to reduce notational clutter. Noting that each experiment has only a finite duration, we let $K_{g,e}$ be the index of the last observation of experiment (g, e). Finally, using the independence of different experiments, we define the reference likelihood ratio $\Lambda(\theta)$ as the product

$$\Lambda(\theta) = \prod_{g \in \mathcal{G}} \prod_{e \in \mathcal{E}_g} \Lambda_{K_{g,e}}^{g,e}(\theta). \tag{2.53}$$

If the parameter under consideration is obvious from the context, then we suppress the dependence on θ.

Before continuing, we state the Conditional Bayes' Theorem (Elliott et al., 1997), which specifies how to compute conditional expectations under a change in measure.

Lemma 2.4.1 (Conditional Bayes' Theorem)**.** *Let H be a random variable, let $\mathcal{F}'$ be a sub-σ-field of the underlying σ-field $\mathcal{F}$, and let dP_θ/dP_0 be the Radon-Nikodym derivative of P_θ with respect to P_0. Then*

$$E_\theta[H|\mathcal{F}'] = \frac{E_0[H \frac{dP_\theta}{dP_0}|\mathcal{F}']}{E_0[\frac{dP_\theta}{dP_0}|\mathcal{F}']}. \tag{2.54}$$

We are now ready to justify the existence of the random variable $\Lambda(\theta)$.

Theorem 2.4.1. *$\Lambda(\theta)$ represents the Radon-Nikodym derivative of the parametrized model probability P_θ with respect to the reference measure P_0.*

That is, $\Lambda(\theta)$ provides the transformation giving $Z^{g,e}$ and $Y^{g,e}$ the distributions which govern the proposed models.

Proof: Define $\widetilde{P}_\theta$ as the measure which does have Radon-Nikodym derivative given by $\Lambda(\theta)$, and we consider the distribution under $\widetilde{P}_\theta$ of the processes $Z^{g,e}$ and $Y^{g,e}$. Because of the independence of the experiments, it is sufficient to show that under $\widetilde{P}_\theta$, $Z^{g,e}$ is a Markov process with transition matrix A and initial distribution π_θ^0 and that $Y^{g,e} - f(Z^{g,e})$ is a stationary, simple autoregressive sequence with parameters ρ and σ for an arbitrary experiment (g, e). We now fix the experiment under consideration. We write $\widetilde{E}_\theta$ to represent expectation with respect to the measure $\widetilde{P}_\theta$.

First of all, under P_0, the sequence $\Lambda_k^{g,e}$ is a martingale with respect to the filtration $\mathcal{F}_k^{g,e}$. To see this, we consider the conditional expectation,

$$\begin{aligned} E_0[\Lambda_{k+1}^{g,e}|\mathcal{F}_k^{g,e}] &= E_0[\Lambda_k^{g,e}\frac{a(Z_k;Z_{k+1})\phi_\sigma(U_{k+1}-\rho U_k)}{\pi_0(Z_{k+1})\phi_1(Y_{k+1})}|\mathcal{F}_k^{g,e}] \\ &= \Lambda_k^{g,e}E_0[\frac{a(Z_k;Z_{k+1})\phi_\sigma(U_{k+1}-\rho U_k)}{\pi_0(Z_{k+1})\phi_1(Y_{k+1})}|\mathcal{F}_k^{g,e}], \end{aligned} \tag{2.55}$$

which shows that we must only demonstrate that this final conditional expectation has a value of 1. To compute this, we use the fact that under P_0, all of the sequences are independent. Thus, we find

$$\begin{aligned} &E_0[\frac{a(Z_k;Z_{k+1})\phi_\sigma(U_{k+1}-\rho U_k)}{\pi_0(Z_{k+1})\phi_1(Y_{k+1})}|\mathcal{F}_k^{g,e}] \\ &\qquad = \sum_{z'}\pi_0(z')\int_{\mathbb{R}}dy'\,\phi_1(y')\frac{a(Z_k;z')\phi_\sigma(y'-f(z')-\rho U_k)}{\pi_0(z')\phi_1(y')} \\ &\qquad = \sum_{z'}a(Z_k;z')\int_{\mathbb{R}}dy'\,\phi_\sigma(y'-f(z')-\rho U_k). \end{aligned} \tag{2.56}$$

Because ϕ_σ is a probability density, for each fixed z', the integral over y' will be 1. Then because A is a transition probability matrix, the remaining sum over z' will

also be 1. Consequently, $\Lambda_k^{g,e}$ is a martingale. Furthermore, we compute in a similar manner

$$\begin{aligned} E_0[\Lambda_0^{g,e}] &= \sum_z \pi_0(z) \int_{\mathbb{R}} dy\, \phi_1(y) \frac{\pi_\theta^0(z)\phi_{\sigma_0}(y - f(z))}{\pi_0(z)\phi_1(y)} \\ &= \sum_z \pi_\theta^0(z) \int_{\mathbb{R}} \phi_{\sigma_0}(y - f(z)) \\ &= 1. \end{aligned} \tag{2.57}$$

Consequently, because the density $\Lambda_k^{g,e}$ is non-negative, $\widetilde{P}_\theta$ is actually a probability measure.

Next, we compute the distributions of the stochastic processes under $\widetilde{P}_\theta$. First, we consider the hidden state process Z_k. The probability that $Z_k = z$ given the information contained in $\mathcal{F}_{k-1}$ can be computed using the conditional Bayes' rule, along with the martingale property of $\Lambda_k^{g,e}$:

$$\begin{aligned} \widetilde{P}_\theta[Z_k = z|\mathcal{F}_{k-1}] = \widetilde{E}_\theta[I_z(Z_k)|\mathcal{F}_{k-1}] &= \frac{E_0[I_z(Z_k)\Lambda_k^{g,e}|\mathcal{F}_{k-1}]}{E_0[\Lambda_k^{g,e}|\mathcal{F}_{k-1}]} \\ &= \frac{\Lambda_{k-1}^{g,e}}{\Lambda_{k-1}^{g,e}} E_0[I_z(Z_k) \frac{a(Z_{k-1}; Z_k)\phi_\sigma(U_k - \rho U_{k-1})}{\pi_0(Z_k)\phi_1(Y_k)}|\mathcal{F}_{k-1}] \\ &= \sum_{z'} \pi_0(z') \int_{\mathbb{R}} dy\, \phi_1(y)\, I_z(z') \frac{a(Z_{k-1}; z')\phi_\sigma(y - f(z') - \rho U_{k-1})}{\pi_0(z')\phi_1(y)} \\ &= a(Z_{k-1}; z) \int_{\mathbb{R}} dy\, \phi_\sigma(y - f(z) - \rho U_{k-1}) = a(Z_{k-1}; z), \end{aligned} \tag{2.58}$$

which depends only on Z_{k-1}. Thus, the sequence Z_k is a Markov process with transition probability matrix A under the law $\widetilde{P}_\theta$. The initial distribution is easily seen to be π_θ^0, according to the calculation

$$\begin{aligned} \widetilde{P}_\theta[Z_0 = z] &= E_0[I_z(Z_0)\Lambda_0^{g,e}] \\ &= \sum_{z'} \pi_0(z') \int_{\mathbb{R}} dy' \phi_1(y') I_z(z') \frac{\pi_\theta^0(z')\phi_{\sigma_0}(y' - f(z'))}{\pi_0(z')\phi_1(y')} \\ &= \pi_\theta^0(z) \int_{\mathbb{R}} dy' \phi_{\sigma_0}(y' - f(z)) \\ &= \pi_\theta^0(z). \end{aligned} \tag{2.59}$$

We next turn to the observation process Y_k. We must show that under $\widetilde{P}_\theta$, the distribution of Y_{k+1} given the the information of the state process $\mathcal{M}_{k+1,k}$ is a normal distribution with mean $f(Z_{k+1}) + \rho(Y_k - f(Z_k))$ and variance σ^2, and that the distribution of Y_0 given Z_0 is a normal distribution with mean $f(Z_0)$ and variance $\sigma_0^2 = \frac{\sigma^2}{1-\rho^2}$. Let Γ be an arbitrary Borel set in $\mathbb{R}$. Starting with the initial distribution, we find

$$\begin{aligned}
\widetilde{P}_\theta[Y_0 \in \Gamma | Z_0] &= \widetilde{E}_\theta[I_\Gamma(Y_0)|Z_0] \\
&= \frac{E_0[I_\Gamma(Y_0)\Lambda_0^{g,e}|Z_0]}{E_0[\Lambda_0^{g,e}|Z_0]} \\
&= \frac{E_0[I_\Gamma(Y_0)\dfrac{\pi_\theta^0(Z_0)\phi_{\sigma_0}(Y_0 - f(Z_0))}{\pi_0(Z_0)\phi_1(Y_0)}|Z_0]}{E_0[\dfrac{\pi_\theta^0(Z_0)\phi_{\sigma_0}(Y_0 - f(Z_0))}{\pi_0(Z_0)\phi_1(Y_0)}|Z_0]} \\
&= \frac{E_0[I_\Gamma(Y_0)\dfrac{\phi_{\sigma_0}(Y_0 - f(Z_0))}{\phi_1(Y_0)}|Z_0]}{E_0[\dfrac{\phi_{\sigma_0}(Y_0 - f(Z_0))}{\phi_1(Y_0)}|Z_0]},
\end{aligned} \tag{2.60}$$

where the factors that depended only on Z_0 come out of the conditional expectation and cancel. Now, explicitly writing the expectation in P_0 as an integral, we obtain

$$E_0[I_\Gamma(Y_0)\frac{\phi_{\sigma_0}(Y_0 - f(Z_0))}{\phi_1(Y_0)}|Z_0] = \int_\Gamma dy\, \phi_{\sigma_0}(y - f(Z_0)), \tag{2.61}$$

which is precisely the probability for a normal random variable with mean $f(Z_0)$ and variance σ_0^2 to be in the Borel set Γ. The denominator of Equation 2.60 is 1, as seen by setting $\Gamma = \mathbb{R}$. Hence, the initial distribution under $\widetilde{P}_\theta$ corresponds to the initial distribution under P_θ.

We finally consider the distribution of Y_{k+1} given the information $\mathcal{M}_{k+1,k}$. Again,

let Γ represent an arbitrary Borel set in $\mathbb{R}$. Then, we find

$$\begin{aligned}
\widetilde{P}_\theta[Y_{k+1} \in \Gamma | \mathcal{M}_{k+1,k}] &= \widetilde{E}_\theta[I_\Gamma(Y_{k+1}) | \mathcal{M}_{k+1,k}] \\
&= \frac{E_0[I_\Gamma(Y_{k+1}) \Lambda^{g,e}_{k+1} | \mathcal{M}_{k+1,k}]}{E_0[\Lambda^{g,e}_{k+1} | \mathcal{M}_{k+1,k}]} \\
&= \frac{\Lambda^{g,e}_k \dfrac{a(Z_k; Z_{k+1})}{\pi_0(Z_{k+1})} E_0[I_\Gamma(Y_{k+1}) \dfrac{\phi_\sigma(Y_{k+1} - f(Z_{k+1}) - \rho U_k)}{\phi_1(Y_{k+1})} | \mathcal{M}_{k+1,k}]}{\Lambda^{g,e}_k \dfrac{a(Z_k; Z_{k+1})}{\pi_0(Z_{k+1})} E_0[\dfrac{\phi_\sigma(Y_{k+1} - f(Z_{k+1}) - \rho U_k)}{\phi_1(Y_{k+1})} | \mathcal{M}_{k+1,k}]} \\
&= \frac{E_0[I_\Gamma(Y_{k+1}) \dfrac{\phi_\sigma(Y_{k+1} - f(Z_{k+1}) - \rho U_k)}{\phi_1(Y_{k+1})} | \mathcal{M}_{k+1,k}]}{E_0[\dfrac{\phi_\sigma(Y_{k+1} - f(Z_{k+1}) - \rho U_k)}{\phi_1(Y_{k+1})} | \mathcal{M}_{k+1,k}]}.
\end{aligned} \tag{2.62}$$

Again, it is sufficient to compute an expression for the numerator, as the denominator immediately follows with $\Gamma = \mathbb{R}$. Writing the numerator as an integral, we obtain

$$E_0[I_\Gamma(Y_{k+1}) \frac{\phi_\sigma(Y_{k+1} - f(Z_{k+1} - \rho U_k))}{\phi_1(Y_{k+1})} | \mathcal{M}_{k+1,k}] = \int_\Gamma dy \, \phi_\sigma(y - f(Z_{k+1}) - \rho U_k), \tag{2.63}$$

which then demonstrates that, under $\widetilde{P}_\theta$, Y_{k+1} has a normal distribution with mean $f(Z_{k+1}) + \rho U_k$.

Consequently, the sequences Z_k and Y_k under $\widetilde{P}_\theta$ have the same distribution as the sequences defined for P_θ. ■

The classical definition of the likelihood corresponds to the numerator of the likelihood ratio $\Lambda(\theta)$. That is, suppose that for experiment (g, e), the ideal observer knew that $Z^{g,e}_0 = z_0, \ldots, Z^{g,e}_k = z_k$ as well as $Y^{g,e}_0 = y_0, \ldots, Y^{g,e}_k = y_k$. Then the likelihood $L^{g,e}(\theta; \vec{z}, \vec{y})$ for this experimental information with a model θ is given by the numerator of $\Lambda^{g,e}_k$,

$$L^{g,e}(\theta; \vec{z}, \vec{y}) = \pi^0_\theta(z_0) \phi_{\sigma_0}(u_0) \prod_{l=1}^{k} a(z_{l-1}; z_l) \phi_\sigma(u_l - \rho u_{l-1}), \tag{2.64}$$

with $u_l = y_l - f(z_l)$. The overall likelihood is given by the product of the likelihoods

of the individual experiments,

$$L(\theta; \vec{z}, \vec{y}) = \prod_{g \in \mathcal{G}} \prod_{e \in \mathcal{E}_g} L^{g,e}(\theta; \vec{z}, \vec{y}), \tag{2.65}$$

where we include $\vec{z}$ and $\vec{y}$ in the notation only to indicate explicitly the dependence of the likelihood on the full information, a slight abuse of notation.

We now turn to the hidden Markov model likelihood. That is, we now consider the experimentalist, who does not have access to the states of the hidden process Z_k. In the classical approach, when information is missing, a marginal density is computed by integrating over the missing information. In our case, this corresponds to summing over the possible states of $Z_0, \ldots, Z_k$,

$$\begin{aligned} L^{g,e}(\theta; \vec{y}) &= \sum_{z_0} \cdots \sum_{z_k} L^{g,e}(\theta; \vec{z}, \vec{y}) \\ &= \sum_{z_0} \cdots \sum_{z_k} \pi^0_\theta(z_0) \phi_{\sigma_0}(u_0) \prod_{l=1}^{k} a(z_{l-1}; z_l) \phi_\sigma(u_l - \rho u_{l-1}). \end{aligned} \tag{2.66}$$

The reference measure approach follows a similar idea, except that rather than computing a marginal density, one computes the conditional expectation of the Radon-Nikodym derivative. That is, $dP_\theta/dP_0 = \Lambda(\theta)$ represents the likelihood for the complete information of the experiments. The likelihood for a partial observation $\mathcal{Y}^{g,e}_k$, corresponding to knowledge of the bead position $Y^{g,e}_0, \ldots, Y^{g,e}_k$ in experiment (g, e) is represented as the restriction of dP_θ/dP_0 to the $\mathcal{Y}^{g,e}_k$-measurable sets, which is written $\left.\frac{dP_\theta}{dP_0}\right|_{\mathcal{Y}^{g,e}_k}$. This is computed as the conditional expectation under P_0 with respect to $\mathcal{Y}^{g,e}_k$:

$$\left.\frac{dP_\theta}{dP_0}\right|_{\mathcal{Y}^{g,e}_k} = E_0[\Lambda^{g,e}_k | \mathcal{Y}^{g,e}_k]. \tag{2.67}$$

To show that these two approaches are consistent with each other, we explicitly compute this conditional expectation by partitioning the underlying probability space

according to the hidden states (again suppressing dependence of Z_k and Y_k on (g, e)):

$$\begin{aligned} E_0[\Lambda_k^{g,e}|\mathcal{Y}_k^{g,e}] &= \sum_{z_0} \cdots \sum_{z_k} E_0[I_{z_0}(Z_0) \cdots I_{z_k}(Z_k)\Lambda_k^{g,e}|\mathcal{Y}_k^{g,e}] \\ &= \sum_{z_0} \cdots \sum_{z_k} E_0[I_{z_0}(Z_0) \cdots I_{z_k}(Z_k)|\mathcal{Y}_k^{g,e}] \frac{\pi_\theta^0(z_0)\phi_{\sigma_0}(Y_0 - z_0)}{\pi_0(z_0)\phi_1(Y_0)} \\ &\quad \times \prod_{l=1}^{k} \frac{a(z_{l-1}; z_l)\phi_\sigma(Y_l - f(z_l) - \rho(Y_{l-1} - f(z_{l-1})))}{\pi_0(z_l)\phi_1(Y_l)}. \end{aligned} \tag{2.68}$$

However, under P_0, the sequence Z is independent of the sequence Y, so that the final conditional expectation is just the probability under P_0 of the selected sequence,

$$E_0[I_{z_0}(Z_0) \cdots I_{z_k}(Z_k)|\mathcal{Y}_k^{g,e}] = P_0[Z_0 = z_0, \ldots, Z_k = z_k] = \prod_{l=0}^{k} \pi_0(z_l). \tag{2.69}$$

Consequently simplifying the conditional likelihood ratio, we obtain

$$\begin{aligned} E_0[\Lambda_k^{g,e}|\mathcal{Y}_k^{g,e}] &= \frac{1}{\prod_{l=0}^{k} \phi_1(Y_l)} \sum_{z_0} \cdots \sum_{z_k} \pi_\theta^0(z_0)\phi_{\sigma_0}(Y_0 - z_0) \\ &\quad \times \prod_{l=1}^{k} a(z_{l-1}; z_l)\phi_\sigma(Y_l - f(z_l) - \rho(Y_{l-1} - f(z_{l-1}))). \end{aligned} \tag{2.70}$$

Again, notice that the numerator of this sum is exactly the same as the unnormalized hidden Markov model likelihood. The denominator will always be independent of the parameter θ. Thus, we will make use of both expressions of the likelihood depending on the context.

We remark that for practical computation, only the standard likelihoods, $L(\theta; \vec{z}, \vec{y})$ and $L(\theta; \vec{y})$, are used. The use of a reference probability measure, which leads to the extra terms appearing in the likelihood ratio, Λ, serves the theoretical purpose of putting all of the models generated by different parameters on a common foundation. Additionally, theoretical calculations with Λ have the advantage of being already normalized, and so we use this approach for the development of the theory. For the implementations of the algorithms, however, we simply use the standard likelihoods.

We next introduce an alternative way to write the unnormalized hidden Markov model likelihood $L^{g,e}(\theta;\vec{y})$ for a particular experiment (g,e). Corresponding to the initial distribution for Y_0, we define a matrix-valued function $B_0^{g,e} : \mathbb{R} \to GL(\mathcal{Z})$ which generates a diagonal matrix for each value y,

$$b_0^{g,e}(y)(z;z') = \begin{cases} 0, & \text{if } z \neq z', \\ \phi_{\sigma_0}\big(y - f(z)\big), & \text{if } z = z', \end{cases} \tag{2.71}$$

where the dependence on the experiment (g,e) is implicit in f, ρ and σ. Similarly, we define a matrix-valued function $B^{g,e} : \mathbb{R}^2 \to GL(\mathcal{Z})$ to deal with transitions, defined as

$$b^{g,e}(y,y')(z;z') = \phi_\sigma\big(y' - f(z') - \rho(y - f(z))\big). \tag{2.72}$$

With these definitions, we can rewrite $L^{g,e}(\theta;\vec{y})$ as a matrix equation,

$$L^{g,e}(\theta;\vec{y}) = \pi_\theta^0 B_0^{g,e}(y_0) \cdot (A^{g,e} \star B^{g,e}(y_1,y_0)) \cdots (A^{g,e} \star B^{g,e}(y_k,y_{k-1})) \cdot \vec{1}, \tag{2.73}$$

where the operator $\star$ represents term by term multiplication of the matrices,

$$(A \star B)(z;z') = a(z;z') \cdot b(z;z'). \tag{2.74}$$

The product with the vector of all ones, $\vec{1}$, corresponds to the summation over the final state z_k. An intuitive interpretation of this formulation is that the matrices $B^{g,e}$ act as posterior weights on the original transition matrices A. That is, in the absence of any information about the observations, the distribution of Z_k would be given by the matrix product

$$\pi_\theta^k = \pi_\theta^0 \cdot A^{k+1}. \tag{2.75}$$

The observations provide additional information about which the transitions actually occured. This information is incorporated by weighting each transition according to how compatible the observations are with the transition, as given by the likelihood factor $b(y,y')(z;z')$.

2.5 Parameter Estimation

Recall that the hidden Markov model maximum likelihood estimator $\hat{\theta}$ is defined as the estimator for the parameter θ which maximizes the likelihood relative to the observed data,

$$\hat{\theta} = \arg\max_{\theta} L(\theta; \vec{Y}) = \arg\max_{\theta} E_0[\frac{dP_\theta}{dP_0}|\mathcal{Y}], \tag{2.76}$$

where $\mathcal{Y}$ represents the σ-field containing all of the observations over all experiments. Again using the independence of separate experiments, we obtain

$$\hat{\theta} = \arg\max_{\theta} \prod_{g\in\mathcal{G}} \prod_{e\in\mathcal{E}} E_0[\Lambda^{g,e}(\theta)|\mathcal{Y}^{g,e}_{K^{g,e}}]. \tag{2.77}$$

Because of the high computational cost in calculating the likelihood, scanning through possible parameters to find a maximal choice is an ineffective procedure. Some efforts have been made to use gradient information about the likelihood to create better direct maximization procedures (Qin et al., 2000), but the dimension of the current parameter space makes this difficult. The Expectation-Maximization (EM) algorithm provides an iterative scheme to create a sequence of parameter estimates. Each successive estimate has a likelihood at least as great as the previous estimate, so that the likelihoods converge to a local maximum. We demonstrate how to implement this classic procedure for the autoregressive model for the kinesin-bead system.

As discussed in the introduction, the EM algorithm focuses on maximizing an auxiliary function $Q(\theta, \theta')$ which is defined as

$$Q(\theta, \theta') = E_\theta[\log \frac{dP_{\theta'}}{dP_0}|\mathcal{Y}] = E_\theta[\log \Lambda(\theta')|\mathcal{Y}]. \tag{2.78}$$

Following the terminology of *elliott97:hmm*, we call the function $Q(\theta, \theta')$ the conditional pseudo-log-likelihood of θ' relative to θ. The essential feature of Q, as shown in Theorem 1.4.1, is that the difference in the log-likelihoods of θ and θ' is bounded below by

$$\log L(\theta'; \vec{y}) - \log L(\theta; \vec{y}) \geq Q(\theta, \theta') - Q(\theta, \theta). \tag{2.79}$$

Thus, maximizing $Q(\theta, \theta')$ over θ' to obtain $\hat{\theta}'$ guarantees that $L(\hat{\theta}'; \vec{y}) > L(\theta; \vec{y})$. The EM algorithm creates a sequence of parameter estimates $\{\theta_n\}$ in a recursive manner, starting with an initial estimate, θ_0. Given an estimate at the n^{th} iterate, θ_n, a single iteration of the EM algorithm consists of the following two steps:

E: Compute the conditional pseudo-log-likelihood $Q(\theta_n, \theta)$ as a function of θ,

M: Choose θ_{n+1} by maximizing $Q(\theta_n, \theta)$

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta_n, \theta). \tag{2.80}$$

When the likelihood is differentiable with respect to the parameter, fixed points of the EM algorithm correspond to critical points of the likelihood (Dempster et al., 1977). Unless the likelihood has only a single relative maximum, there is no guarantee that the EM algorithm will converge to the maximum likelihood estimate. However, the essentially intractable problem of directly maximizing the likelihood over the entire parameter space is converted to maximizing the function Q for a sequence of parameter estimates.

We next turn to computing the conditional pseudo-log-likelihood $Q(\theta_n, \theta)$ for a fixed parameter θ_n. By using the properties of logarithms, the products appearing in Λ become summations,

$$\begin{aligned} Q(\theta_n, \theta) &= E_{\theta_n}[\log \Lambda(\theta) | \mathcal{Y}] \\ &= E_{\theta_n}[\sum_{g \in \mathcal{G}} \sum_{e \in \mathcal{E}} \log \Lambda^{g,e}(\theta) | \mathcal{Y}] \\ &= \sum_{g \in \mathcal{G}} \sum_{e \in \mathcal{E}} E_{\theta_n}[\log \Lambda^{g,e}_{K_{g,e}}(\theta) | \mathcal{Y}^{g,e}_{K_{g,e}}]. \end{aligned} \tag{2.81}$$

So, in order to compute Q, we can compute the contributions from each individual experiment (g, e) separately. Thus, we consider the log-likelihood of a single experiment,

$\log \Lambda^{g,e}_{K_{g,e}}(\theta)$, which has the form

$$\begin{aligned}\log \Lambda^{g,e}_{K_{g,e}}(\theta) = \log \pi^0_\theta(Z_0) + \sum_{l=1}^{K_{g,e}} \log a(Z_{l-1}; Z_l) + \log \phi_{\sigma_0}(Y_0 - f(Z_0)) \\ + \sum_{l=1}^{K_{g,e}} \log \phi_\sigma(Y_l - f(Z_l) - \rho(Y_{l-1} - f(Z_{l-1}))) \\ - \sum_{l=0}^{K_{g,e}} \log \pi_0(Z_l) - \sum_{l=0}^{K_{g,e}} \log \phi_1(Y_l). \end{aligned} \tag{2.82}$$

The final two sums which come from the reference measure P_0 do not depend on the parameter θ. Consequently, we only need to compute the conditional expectation of

$$\begin{aligned} q^{g,e}(\theta) = \log \pi^0_\theta(Z_0) + \sum_{l=1}^{K_{g,e}} \log a(Z_{l-1}; Z_l) + \log \phi_{\sigma_0}\big(Y_0 - f(Z_0)\big) \\ + \sum_{l=1}^{K_{g,e}} \log \phi_\sigma\Big(Y_l - f(Z_l) - \rho\big(Y_{l-1} - f(Z_{l-1})\big)\Big). \end{aligned} \tag{2.83}$$

First, consider the term involving the jumps, $\sum_{l=1}^{K_{g,e}} \log a(Z_{l-1}; Z_l)$. Because transition probabilities only depend on the initial state, the final state, and the number of lattice sites changed, we recall the equivalence of Equation 2.35,

$$a(x, c; x', c') = a(0, c; x' - x, c') = \pi_c(\Delta x = x' - x, c'). \tag{2.84}$$

We define the jump function $J_l(x, c_i, c_f)$ to indicate that a jump occurred from initial state c_i at time l to a final state of c_f with a change of x sites at time $l + 1$,

$$J_l(x, c_i, c_f) = I_{c_i}(C_l) I_{c_f}(C_{l+1}) I_x(X_{l+1} - X_l). \tag{2.85}$$

Then we can rewrite the jump term of the log-likelihood as

$$\sum_{l=1}^{K} \log a(Z_{l-1}; Z_l) = \sum_x \sum_{c_i} \sum_{c_f} \log \pi_{c_i}(x, c_f) \sum_{l=0}^{K-1} J_l(x, c_i, c_f). \tag{2.86}$$

The conditional expectation of this can be written

$$E_{\theta_n}[\sum_{l=1}^{K} \log a(Z_{l-1}; Z_l) | \mathcal{Y}^{g,e}_{K_{g,e}}] = \sum_x \sum_{c_i} \sum_{c_f} \log \pi_{c_i}(x, c_f) E_{\theta_n}[\sum_{l=0}^{K-1} J_l(x, c_i, c_f) | \mathcal{Y}^{g,e}_{K_{g,e}}]. \tag{2.87}$$

Note that the required conditional expectation, $E_{\theta_n}[\sum J_l|\mathcal{Y}]$, depends only on the current parameter θ_n and can be computed once for all possible parameters θ.

Next we consider the terms involving the observations. Using the density ϕ_σ of a normal random variable given in Equation 2.52, the terms involving the sum of the logarithms of the density become sums of squares. In particular, recalling the explicit form of $f(Z)$ from Equation 2.16,

$$f(x,c) = \kappa + d\big(x + \epsilon(c)\big), \tag{2.88}$$

we obtain

$$\begin{aligned}
\log \phi_{\sigma_0}\big(Y_0 - f(Z_0)\big) + \sum_{l=1}^{K_{g,e}} \log \phi_\sigma\Big(Y_l - f(Z_l) - \rho\big(Y_{l-1} - f(Z_{l-1})\big)\Big) \\
= -\frac{K_{g,e}+1}{2}\log 2\pi - (K_{g,e}+1)\log\sigma + \frac{1}{2}\log(1-\rho^2) \\
- \frac{1-\rho^2}{2\sigma^2}\Big(Y_0 - d\big(X_0 + \epsilon(C_0)\big) - \kappa\Big)^2 \\
- \frac{1}{2\sigma^2}\sum_{l=1}^{K_{g,e}}\Big(Y_l - \rho Y_{l-1} - d\big(X_l + \epsilon(C_l)\big) \\
+ \rho d\big(X_{l-1} + \epsilon(C_{l-1})\big) - (1-\rho)\kappa\Big)^2.
\end{aligned} \tag{2.89}$$

Because we have $C \in \{1, \ldots, N\}$, it is convenient to use the representation

$$\epsilon(C) = \sum_{c=1}^{N} \epsilon_c I_c(C), \tag{2.90}$$

as well as to introduce a state-dependent absolute offset

$$\kappa_c = \epsilon_c d + \kappa. \tag{2.91}$$

This leads to an equivalent expression

$$\begin{aligned}
&\log \phi_{\sigma_0}\big(Y_0 - f(Z_0)\big) + \sum_{l=1}^{K_{g,e}} \log \phi_{\sigma}\big(Y_l - f(Z_l) - \rho(Y_{l-1} - f(Z_{l-1}))\big) \\
&\qquad = -\frac{K_{g,e}+1}{2}\log 2\pi - (K_{g,e}+1)\log\sigma + \frac{1}{2}\log(1-\rho^2) \\
&\qquad\quad - \frac{1-\rho^2}{2\sigma^2}\Big(Y_0 - dX_0 - \sum_c \kappa_c I_c(C_0)\Big)^2 \\
&\qquad\quad - \frac{1}{2\sigma^2}\sum_{l=1}^{K_{g,e}} \Big(Y_l - \rho Y_{l-1} - d\big(X_l - \rho X_{l-1}\big) - \sum_c \kappa_c\big(I_c(C_l) - \rho I_c(C_{l-1})\big)\Big)^2.
\end{aligned} \tag{2.92}$$

Motivated by these expressions, let U_l and V_l be two sequences and consider the sum

$$\begin{aligned}
&(1-\rho^2)(U_0V_0) + \sum_{l=1}^{k}(U_l - \rho U_{l-1})(V_l - \rho V_{l-1}) \\
&\qquad = (1-\rho^2)U_0V_0 + \sum_{l=1}^{k} U_lV_l - \rho\sum_{l=1}^{k}(U_lV_{l-1} + U_{l-1}V_l) + \rho^2\sum_{l=1}^{k} U_{l-1}V_{l-1} \\
&\qquad = (1+\rho^2)\sum_{l=0}^{k} U_lV_l - \rho^2(U_0V_0 + U_kV_k) - \rho\sum_{l=1}^{k}(U_lV_{l-1} + U_{l-1}V_l).
\end{aligned} \tag{2.93}$$

Defining the bilinear form

$$T_k^{\rho}(U,V) = (1+\rho^2)\sum_{l=0}^{k} U_lV_l - \rho^2(U_0V_0 + U_kV_k) - \rho\sum_{l=1}^{k}(U_lV_{l-1} + U_{l-1}V_l), \tag{2.94}$$

we can rewrite the observation terms of the log-likelihood function as

$$\begin{aligned}
&\log \phi_{\sigma_0}\big(Y_0 - f(Z_0)\big) + \sum_{l=1}^{K_{g,e}} \log \phi_{\sigma}\Big(Y_l - f(Z_l) - \rho\big(Y_{l-1} - f(Z_{l-1})\big)\Big) \\
&\qquad = -\frac{K_{g,e}+1}{2}\log 2\pi - (K_{g,e}+1)\log\sigma + \frac{1}{2}\log(1-\rho^2) \\
&\qquad\quad - \frac{1}{2\sigma^2}T^{\rho}_{K_{g,e}}\Big(Y - dX - \sum_c \epsilon_c d\, I_c(C) - \kappa, Y - dX - \sum_c \epsilon_c d\, I_c(C) - \kappa\Big).
\end{aligned} \tag{2.95}$$

Using the bilinearity of $T_\rho = T^\rho_{K_{g,e}}$, we obtain the formula

$$\begin{aligned} T^\rho_{K_{g,e}}\Big(Y - dX - \sum_c \epsilon_c d\, I_c(C) - \kappa, Y - dX - \sum_c \epsilon_c d\, I_c(C) - \kappa\Big) \\ = T_\rho(Y,Y) - 2d\, T_\rho(X,Y) - 2\sum_c \epsilon_c d\, T_\rho(Y, I_c(C)) - 2\kappa\, T_\rho(Y,1) \\ + d^2\, T_\rho(X,X) + 2\sum_c \epsilon_c d^2\, T_\rho(X, I_c(C)) + 2d\kappa\, T_\rho(X,1) \\ + \sum_c \sum_{c'} \epsilon_c \epsilon_{c'} d^2\, T_\rho(I_c(C), I_{c'}(C)) + 2\sum_c \epsilon_c d\kappa\, T_\rho(I_c(C),1) + \kappa^2\, T_\rho(1,1). \end{aligned} \tag{2.96}$$

Because the bilinear form T^ρ_k implicitly includes a dependence on ρ, it is more convenient to express T^ρ_k in terms of parameter free sums,

$$T^\rho_k(V,W) = (1+\rho^2)S_k(VW) - \rho^2 B_k(VW) - \rho R_k(V,W), \tag{2.97}$$

which are defined as

$$S_k(V) = \sum_{l=0}^{k} V_l, \tag{2.98}$$

$$B_k(V) = V_0 + V_k, \tag{2.99}$$

$$R_k(V,W) = \sum_{l=0}^{k} (V_l W_{l-1} + V_{l-1} W_l). \tag{2.100}$$

Therefore, just as for the term involving the jumps, the conditional expectation of the terms coming from the observations can be written in terms of conditional expectations of various sums of the sequences X_l, $I_c(C_l)$, and Y_l, without any dependence on the new parameters, so that the conditional expectation can be computed in a single pass.

Combining the terms, we can now completely describe the conditional pseudo-log-likelihood, $Q(\theta_n, \theta)$. For arbitrary random sequences V_l and W_l, we abbreviate the

conditional expectations

$$\widehat{S}_k(V) = E[S_k(V)|\mathcal{Y}_k], \tag{2.101}$$

$$\widehat{B}_k(V) = E[B_k(V)|\mathcal{Y}_k], \tag{2.102}$$

$$\widehat{R}_k(V,W) = E[R_k(V,W)|\mathcal{Y}_k], \tag{2.103}$$

$$\widehat{T}^{\rho}_k(V,W) = (1+\rho^2)\widehat{S}_k(VW) - \rho^2\widehat{B}_k(VW) - \rho\widehat{R}_k(V,W). \tag{2.104}$$

This finally allows us to rewrite the $Q^{g,e}$, ignoring terms that are independent of the parameters, as

$$\begin{aligned}
\widetilde{Q}^{g,e}(\theta_n,\theta) &= E_{\theta_n}[q^{g,e}(\theta)|\mathcal{Y}_k] \\
&= \sum_c \log \pi^0_\theta(0,c) E_{\theta_n}[I_c(C_0)|\mathcal{Y}^{g,e}_{K_{g,e}}] \\
&\quad + \sum_x \sum_{c_i} \sum_{c_f} \widehat{S}_{K_{g,e}-1}\big(J(x,c_i,c_f)\big) \log \pi_{c_i}(x,c_f) \\
&\quad - \frac{K_{g,e}+1}{2}\log 2\pi - (K_{g,e}+1)\log\sigma + \frac{1}{2}\log(1-\rho^2) \\
&\quad - \frac{1}{2\sigma^2}\Big[\widehat{T}^{\rho}_{K_{g,e}}(Y,Y) - 2d\,\widehat{T}^{\rho}_{K_{g,e}}(X,Y) + d^2\,\widehat{T}^{\rho}_{K_{g,e}}(X,X) \\
&\qquad - 2\kappa\,\widehat{T}^{\rho}_{K_{g,e}}(Y,1) + 2d\kappa\,\widehat{T}^{\rho}_{K_{g,e}}(X,1) + 2\sum_c \epsilon_c d\kappa\,\widehat{T}^{\rho}_{K_{g,e}}(I_c(C),1) \\
&\qquad - 2\sum_c \epsilon_c d\,\widehat{T}^{\rho}_{K_{g,e}}(Y,I_c(C)) + 2\sum_c \epsilon_c d^2\,\widehat{T}^{\rho}_{K_{g,e}}(X,I_c(C)) \\
&\qquad + \sum_c \sum_{c'} \epsilon_c\epsilon_{c'} d^2\,\widehat{T}^{\rho}_{K_{g,e}}(I_c(C),I_{c'}(C)) + \kappa^2\Big(1-\rho^2+(1-\rho)^2 K_{g,e}\Big)\Big].
\end{aligned} \tag{2.105}$$

Although we have written this in terms of $\widehat{T}$ to convey the formula, in reality, each such term is composed of the corresponding sum of $\widehat{S}$, $\widehat{B}$ and $\widehat{R}$ as given in Equation 2.104. Consequently, the expectation step consists of computing each of these conditional expectations of various sums. As stressed earlier, this calculation will be independent of the parameter θ which is maximized, as the parameters enter only as coefficients to these sums.

2.5.1 EM Step: Expectation

The conditional pseudo-log-likelihood, $Q(\theta_n, \theta)$, is evaluated in terms of a number of conditional expectations, each of which relies upon the application of the Conditional Bayes' Theorem, which we state in terms of the following lemma. Because the conditional expectations are computed for each experiment separately, we assume throughout this section that a particular experiment, (g, e), is being considered. Consequently, we will suppress these indexes except where needed for clarity.

Lemma 2.5.1. *Let $H : \mathcal{Z}^{K+1} \times \mathbb{R}^{K+1} \to \mathbb{R}$ be any measurable function. Write $\vec{Z} = Z_0, \ldots, Z_K$ and $\vec{Y} = Y_0, \ldots, Y_K$. Then*

$$E_{\theta_n}[H(\vec{Z}, \vec{Y})|\mathcal{Y}_K] = \sum_{z_0} \cdots \sum_{z_K} H(\vec{z}, \vec{Y}) \frac{L^{g,e}(\theta_n; \vec{z}, \vec{Y})}{L^{g,e}(\theta_n; \vec{Y})}. \tag{2.106}$$

Proof of Lemma: Since $\mathcal{Y}_k$ is a sub-σ-algebra of $\mathcal{F}_k$, the Conditional Bayes' Theorem leads to

$$E_{\theta_n}[H(\vec{Z}, \vec{Y})|\mathcal{Y}_K] = \frac{E_0[H(\vec{Z}, \vec{Y})\Lambda_K^{g,e}(\theta_n)|\mathcal{Y}_K]}{E_0[\Lambda_K^{g,e}(\theta_n)|\mathcal{Y}_K]}. \tag{2.107}$$

However, we also have

$$\begin{aligned}
&E_0[H(\vec{Z}, \vec{Y})\Lambda_K^{g,e}(\theta_n)|\mathcal{Y}_K] \\
&\qquad = E_0[\sum_{z_0} \cdots \sum_{z_k} H(\vec{z}, \vec{Y}) I_{z_0}(Z_0) \cdots I_{z_k}(Z_K) \Lambda_K^{g,e}(\theta_n)|\mathcal{Y}_K] \\
&\qquad = \sum_{z_0} \cdots \sum_{z_k} H(\vec{z}, \vec{Y}) \frac{L^{g,e}(\theta; \vec{z}, \vec{Y})}{\prod_{l=0}^{k} \pi_0(z_l)\phi_1(Y_l)} E_0[I_{z_0}(Z_0) \cdots I_{z_k}(Z_k)|\mathcal{Y}_k] \\
&\qquad = \sum_{z_0} \cdots \sum_{z_k} H(\vec{z}, \vec{Y}) \frac{L^{g,e}(\theta; \vec{z}, \vec{Y})}{\prod_{l=0}^{k} \phi_1(Y_l)}.
\end{aligned}$$

The denominator is exactly the same, but with $H = 1$,

$$E_0[\Lambda_\theta(k)|\mathcal{Y}_k] = \sum_{z_0} \cdots \sum_{z_k} \frac{L^{g,e}(\theta; \vec{z}, \vec{Y})}{\prod_{l=0}^{k} \phi_1(Y_l)} = \frac{L^{g,e}(\theta; \vec{Y})}{\prod_{l=0}^{k} \phi_1(Y_l)}. \tag{2.108}$$

Simplifying the ratio completes the proof. ■

Note that in principle, the formula suggests that the likelihood of each possible sequence of states $z_0, \ldots, z_k$ should be computed, which then serves as the unnormalized conditional density of the trajectory. The hidden Markov model likelihood $L^{g,e}(\theta_n; \vec{y})$, which normalizes the density, acts in a comparable role to a partition function in statistical mechanics. Of course, enumerating all possible hidden states is precisely what the classic forward algorithm for hidden Markov models helps to avoid, and this algorithm is directly generalized to the autoregressive model. We here establish an appropriate generalization of the forward algorithm to compute unnormalized conditional expectations. The classical forward algorithm actually only computes unnormalized conditional probabilities. The normalization factor is, as one should expect, the hidden Markov model likelihood.

Theorem 2.5.1. *Suppose that the measurable function $H : \mathcal{Z}^{K+1} \times \mathbb{R}^{K+1} \to \mathbb{R}$ can be written in the form*

$$H(\vec{z}, \vec{y}) = h_0(z_0, \vec{y}) \prod_{l=1}^{K} h_l(z_{l-1}, z_l, \vec{y}), \tag{2.109}$$

and define the diagonal matrix-valued function $H_0(\vec{y})$ with entries

$$H_0(\vec{y})(z; z') = \begin{cases} h_0(z, \vec{y}), & \text{if } z = z', \\ 0, & \text{if } z \neq z', \end{cases} \tag{2.110}$$

as well as the matrix-valued functions $H_l(\vec{y})$, for $l = 1, \ldots, K$, with entries

$$H_l(\vec{y})(z; z') = h_l(z, z', \vec{y}). \tag{2.111}$$

Let $\star$ denote term-by-term multiplication of matrices. Then the unnormalized conditional expectation of H can be written as a matrix multiplication,

$$E_0[H(\vec{Z}, \vec{Y}) \Lambda_K^{g,e}(\theta_n) | \mathcal{Y}_k] = \frac{1}{\prod_l \phi_1(y_l)} \pi^0_{\theta_n} \cdot (B_0 \star H_0(\vec{y})) \cdot (A \star B_1 \star H_1(\vec{y})) \cdots (A \star B_K \star H_K(\vec{y})) \cdot \vec{1}, \tag{2.112}$$

where $B_0 = B^0(y_0)$ *and* $B_l = B(y_{l-1}, y_l)$*. Consequently, the* P_{θ_n} *conditional expectation of* H *is equal to the ratio of matrix products,*

$$E_{\theta_n}[H(\vec{Z}, \vec{Y})|\mathcal{Y}_K] = \frac{\pi^0_{\theta_n} \cdot (B_0 \star H_0) \cdot \prod_{l=1}^K (A \star B_l \star H_l) \cdot \vec{1}}{\pi^0_{\theta_n} \cdot (B_0) \cdot \prod_{l=1}^K (A \star B_l) \cdot \vec{1}}. \tag{2.113}$$

Proof: First of all, note that $H = 1$ corresponds exactly to the case of computing the conditional likelihood $\frac{dP_{\theta_n}}{dP_0}\Big|_{\mathcal{Y}_K}$. The first result, Equation 2.112, is just a matter of commutativity, arranging the factors for H with the factors for $L^{g,e}(\theta_n; \vec{z}, \vec{y})$:

$$\begin{aligned}
&E_0[H(Z_0, \ldots, Z_k, Y_0, \ldots, Y_k)\Lambda_\theta(k)|\mathcal{Y}_k] \\
&\quad = \frac{1}{\prod_l \phi_1(y_l)} \sum_{\vec{z}} h_0(z_0, \vec{y}) \prod_{l=1}^K h_l(z_{l-1}, z_l, \vec{y}) \pi^0_{\theta_n}(z_0) \phi_{\sigma_0}(y_0 - f(z_0)) \\
&\qquad \times \prod_{l=1}^K a_\theta(z_{l-1}; z_l) \phi_\sigma\left(y_l - f(z_l) - \rho(y_{l-1} - f(z_{l-1}))\right) \\
&\quad = \frac{1}{\prod_l \phi_1(y_l)} \sum_{\vec{z}} \pi^0_{\theta_n}(z_0) \cdot \phi_{\sigma_0}(y_0 - f(z_0)) \cdot h_0(z_0, \vec{y}) \\
&\qquad \times \prod_{l=1}^K a_\theta(z_{l-1}; z_l) \cdot \phi_\sigma\left(y_l - f(z_l) - \rho(y_{l-1} - f(z_{l-1}))\right) \cdot h_l(z_{l-1}, z_l, \vec{y}) \\
&\quad = \frac{1}{\prod_l \phi_1(y_l)} \sum_{\vec{z}} \pi^0_{\theta_n}(z_0) \cdot B_0(y_0)(z_0) \cdot H_0(\vec{y})(z_0) \\
&\qquad \times \prod_{l=1}^K A(z_{l-1}; z_l) \cdot B(y_{l-1}, y_l)(z_{l-1}; z_l) \cdot H_l(\vec{y})(z_{l-1}; z_l) \\
&\quad = \frac{1}{\prod_l \phi_1(y_l)} \pi^0_{\theta_n} \cdot (B^0(y_0) \star H_1(\vec{y})) \cdot (A \star B(y_0, y_1) \star H_1(\vec{y})) \cdots \\
&\qquad\qquad \cdot (A \star B(y_{K-1}, y_K) \star H_K(\vec{y})) \cdot \vec{1}.
\end{aligned}$$

The final result follows immediately by applying Lemma 2.5.1. ■

The conditional pseudo-log-likelihood actually depends on sums of functions of consecutive hidden states rather than products. Each term in the sum can be represented in terms of a product; all but one of the factors are simply the constant function $h = 1$. Consequently, we can directly implement a forward-backward algorithm to compute this type of expectation, as demonstrated in the following corollary.

Corollary 2.5.1 (Forward-Backward Algorithm)**.** *Let the functions $h_0, \ldots, h_K$ be defined as for Theorem 2.5.1. Let the experiment (g, e) be fixed. Introduce the rescaled forward variables α_l according to the recursive procedure*

1. *Initialize: $\widetilde{\alpha}_0 = \pi^0_{\theta_n} \cdot B_0(Y_0)$,*
2. *Scale Factor: $c_l = \langle \widetilde{\alpha}_l, \vec{1} \rangle$, for $l = 0, \ldots, K$,*
3. *Normalize: $\alpha_l = \widetilde{\alpha}_l / c_l$, for $l = 0, \ldots, K$,*
4. *Recursion: $\widetilde{\alpha}_{l+1} = \alpha_l \cdot (A \star B(Y_l, Y_{l+1}))$, for $l = 0, \ldots, K-1$.*

Similarly, define the rescaled backward variables β_l according to

1. *Initialize: $\beta_K = \vec{1}$,*
2. *Recursion: $\widetilde{\beta}_{l-1} = (A \star B(Y_{l-1}, Y_l)) \cdot \beta_l$, for $l = K, \ldots, 1$,*
3. *Normalize: $\beta_{l-1} = \widetilde{\beta}_{l-1} / c_l$, for $l = K, \ldots, 1$.*

Then the P_{θ_n}-conditional expectation of h_0 is given by

$$E_{\theta_n}[h_0(Z_0, \vec{Y}) | \mathcal{Y}_K] = \alpha_0 \cdot H_0(\vec{Y}) \cdot \beta_0, \tag{2.114}$$

and for each $l = 1, \ldots, K$, the P_{θ_n}-conditional expectation of h_l is given by

$$E_{\theta_n}[h_l(Z_{l-1}, Z_l, \vec{Y}) | \mathcal{Y}_K] = \frac{1}{c_l} \alpha_{l-1} \cdot (A \star B_l \star H_l) \cdot \beta_l. \tag{2.115}$$

Proof: The essential observation is to see that α_l and β_l represent different sides of the matrix product defining the likelihood:

$$\alpha_l = \frac{1}{\prod_{i=0}^{l} c_i} \pi^0_{\theta_n} \cdot B_0 \cdot \prod_{i=1}^{l} (A \star B_i), \tag{2.116}$$

$$\beta_l = \frac{1}{\prod_{i=l+1}^{K} c_i} \prod_{i=l+1}^{K} (A \star B_i) \cdot \vec{1}. \tag{2.117}$$

Thus, as in the case of the classical HMM calculations, the vector product of corresponding forward and backward variables results in the likelihood divided by the product of scaling factors,

$$\alpha_l \cdot \beta_l = \frac{L^{g,e}(\theta_n; \vec{y})}{\prod_i c_i}. \tag{2.118}$$

Also, note that for every l, $\alpha_l \cdot \vec{1} = 1$. Since l is arbitrary and $\beta_K = \vec{1}$, we find, exactly analogous to classical HMM calculations,

$$\prod_{l=0}^{K} c_l = L^{g,e}(\theta_n; \vec{y}). \tag{2.119}$$

Multiplying each function h_l by K factors of 1, we apply Theorem 2.5.1 with $H_i = 1$ for $i \neq l$ to obtain for $l = 0$,

$$\begin{aligned} E_{\theta_n}[h_0(Z_0, \vec{Y})|\mathcal{Y}_K] &= \frac{\pi^0_{\theta_n}(B_0 \star H_0) \prod_{i=1}^{K}(A \star B_i \star 1) \cdot \vec{1}}{\pi^0_{\theta_n}(B_0) \prod_{i=1}^{K}(A \star B_i) \cdot \vec{1}} \\ &= \frac{\pi^0_{\theta_n} B_0}{c_0} \cdot H_0 \cdot \frac{\prod_{i=1}^{K}(A \star B_i) \cdot \vec{1}}{\prod_{i=1}^{K} c_i} \\ &= \alpha_0 \cdot H_0 \cdot \beta_0, \end{aligned} \tag{2.120}$$

and for $l = 1, \ldots, K$,

$$\begin{aligned} &E_{\theta_n}[h_l(Z_{l-1}, Z_l, \vec{Y})|\mathcal{Y}_K] \\ &\quad = \frac{\pi^0_{\theta_n}(B_0 \star 1) \prod_{i=1}^{l-1}(A \star B_i \star 1) \cdot (A \star B_l \star H_l) \cdot \prod_{i=l+1}^{K}(A \star B_i \star 1) \cdot \vec{1}}{\pi^0_{\theta_n}(B_0) \prod_{i=1}^{K}(A \star B_i) \cdot \vec{1}} \\ &\quad = \frac{\pi^0_{\theta_n} B_0 \cdot \prod_{i=1}^{l-1}(A \star B_i \star 1)}{\prod_{i=0}^{l-1} c_i} \cdot \frac{A \star B_l \star H_l}{c_l} \cdot \frac{\prod_{i=l+1}^{K}(A \star B_i) \cdot \vec{1}}{\prod_{i=l+1}^{K} c_i} \\ &\quad = \frac{1}{c_l} \alpha_{l-1} \cdot (A \star B_l \star H_l) \cdot \beta_l. \end{aligned} \tag{2.121}$$

■

In addition to obtaining a forward-backward procedure for computing expectations, we also obtain a recursive forward procedure for computing conditional expectations of sums.

Corollary 2.5.2. *Let the functions $h_0, \ldots, h_K$ be defined as for the Theorem 2.5.1. Define the partial sum*

$$S_k(h) = h_0(Z_0, \vec{Y}) + \sum_{l=1}^{k} h_l(Z_{l-1}, Z_l, \vec{Y}). \tag{2.122}$$

In addition to the forward variables α_k and the scaling factors c_k defined in the forward-backward algorithm, define the sequence of vectors $\gamma_k(S_k(h))$ according to the recursive definition

1. *Initialize:*

$$\gamma_0(S_0(h)) = \alpha_0 \cdot H_0, \tag{2.123}$$

2. *Recursion:*

$$\gamma_{k+1}(S_{k+1}(h)) = \frac{1}{c_{k+1}}\Big(\gamma_k(S_k(h)) \cdot (A \star B_{k+1}) + \alpha_k \cdot (A \star B_{k+1} \star H_{k+1})\Big). \tag{2.124}$$

Then the conditional expectation of the sum is given by the vector product

$$E_{\theta_n}[S_K(h)|\mathcal{Y}_K] = \gamma_K(S_K(h)) \cdot \vec{1}. \tag{2.125}$$

In the special case that each h_l is $\mathcal{F}_l$-measurable, so that it is a function of the observations only of $Y_0, \ldots, Y_l$ rather than the full sequence, then we also have for each k, the sequence of conditional expectations

$$E_{\theta_n}[S_k(h)|\mathcal{Y}_k] = \gamma_k(S_k(h)) \cdot \vec{1}. \tag{2.126}$$

Proof: Consider an individual term in the sum, h_l for $l \leq K$. Define $\gamma_k(h_l)$ according to the recursion

1. Initialize:

$$\gamma_0(h_l) = \begin{cases} \alpha_0 \cdot H_0, & \text{if } l = 0, \\ \alpha_0, & \text{if } l \neq 0, \end{cases} \tag{2.127}$$

2. Recursion:

$$\gamma_k(h_l) = \begin{cases} \gamma_{k-1}(h_l) \cdot \dfrac{(A \star B_k \star H_k)}{c_k}, & \text{if } k = l, \\ \gamma_{k-1}(h_l) \cdot \dfrac{(A \star B_k)}{c_k}, & \text{if } k \neq l. \end{cases} \tag{2.128}$$

Theorem 2.5.1 implies that

$$E_{\theta_n}[h_l(Z,Y)|\mathcal{Y}_K] = \gamma_K(h_l) \cdot \vec{1}. \tag{2.129}$$

Next, note that if $l > k$, then $\gamma_k(h_l) = \alpha_k$. Consequently, we also have the sequence of equalities

$$\gamma_k(S_k(h)) = \sum_{l=0}^{k} \gamma_k(h_l), \tag{2.130}$$

which is demonstrated by induction. The statement is clearly true for $k = 0$. So suppose that it is true for $k = k'$. By construction, we have

$$\begin{aligned} \gamma_{k'+1}(S_{k'+1}(h)) &= \frac{1}{c_{k'+1}} \gamma_{k'}(S_{k'}(h)) \cdot (A \star B_{k'+1}) + \frac{1}{c_{k'+1}} \alpha_{k'} \cdot (A \star B_{k'+1} \star H_{k'+1}) \\ &= \frac{1}{c_{k'+1}} \sum_{l=0}^{k'} \gamma_{k'}(h_l) \cdot (A \star B_{k'+1}) + \frac{1}{c_{k'+1}} \gamma_{k'}(h_{k'+1}) \cdot (A \star B_{k'+1} \star H_{k'+1}) \\ &= \sum_{l=0}^{k'} \gamma_{k'+1}(h_l) + \gamma_{k'+1}(h_{k'+1}), \end{aligned} \tag{2.131}$$

completing the induction step. Therefore, we have equality for $k = K$. By the linearity of conditional expectation, we have proved Equation 2.125. The final conclusion follows by realizing that if the functions h_l depend only on the first $l+1$ observations, then all of the previous results hold on the reduced data set $Y_0, \ldots, Y_k$ with $k < K$. ■

For the kinesin-bead HMM system, the functions h_l actually only depend on Y_{l-1} and Y_l, so that Equation 2.126 holds. We remark that this forward algorithm for computing expectation exactly recovers the recursive algorithms developed within the framework of a reference measure for standard hidden Markov models (Elliott, 1994; Elliott et al., 1997). The present derivation is much more transparent in its

relation with the forward-backward algorithms than the original works. As discussed in the introduction, the primary advantages of the forward algorithm are the ability to do on-line estimation without needing to recalculate the backward variables and a major reduction in memory use. However, each sum being calculated must be updated separately using Equation 2.124. The computational effort for each update will thus be comparable to updating the forward variable α_l. Therefore, this approach is much more expensive computationally than a single pass of the forward-backward algorithm.

We now recall the explicit quantities that need to be estimated to implement the EM algorithm. In particular, the function $Q(\theta_n, \theta)$, as expressed in Equation 2.105, depends on the conditional expectation of the sums of jumps $S_{K-1}(J(x, c_i, c_f))$ as well the various sums $S_K(UV)$, $B_K(UV)$ and $R_K(U, V)$ where U and V can each be any of the processes X, Y, or $I_c(C)$. Each of these computations can be calculated from the forward and backward variables. That is, each of these sums can be expressed in the form of the general sum

$$S(h) = h_0(Z_0, Y_0) + \sum_{l=1}^{K} h_l(Z_{l-1}, Z_l, Y_{l-1}, Y_l) \tag{2.132}$$

for appropriate choices of the functions h_l. Using the forward-backward algorithm, we find that the posterior initial distribution is given by

$$P_{\theta_n}[Z_0 = z|\mathcal{Y}_K] = \alpha_0(z) \cdot \beta_0(z), \tag{2.133}$$

and that the posterior probability of jumping from z to z' at time l is given by

$$\widehat{J}_l(z; z') = P_{\theta_n}[Z_{l-1} = z, Z_l = z'|\mathcal{Y}_K] \tag{2.134}$$

$$= \frac{1}{c_l}\alpha_{l-1}(z)(A \star B_l)(z; z')\beta_l(z'). \tag{2.135}$$

Therefore, the conditional expectation of each of the terms in the general sum can be

computed as

$$E_{\theta_n}[h_0(Z_0, Y_0)|\mathcal{Y}_K] = \sum_z h_0(z, Y_0)\, \alpha_0(z) \cdot \beta_0(z), \tag{2.136}$$

$$E_{\theta_n}[h_l(Z_{l-1}, Z_l, Y_{l-1}, Y_l)|\mathcal{Y}_K] = \sum_z \sum_{z'} h_l(z, z', Y_{l-1}, Y_l)\, \widehat{J}_l(z; z'). \tag{2.137}$$

For functions h_l that depend only on one of Z_{l-1} or Z_l, it is useful to note the identities

$$\sum_z \widehat{J}_l(z; z') = \alpha_l(z') \cdot \beta_l(z') = P_{\theta_n}[Z_l = z'|\mathcal{Y}_K], \tag{2.138}$$

$$\sum_{z'} \widehat{J}_l(z; z') = \alpha_{l-1}(z') \cdot \beta_{l-1}(z') = P_{\theta_n}[Z_{l-1} = z|\mathcal{Y}_K]. \tag{2.139}$$

In order to compute the required estimates of the number of relative jumps that go from state c_i to state c_f with a change of x lattice sites, we use the relation

$$\widehat{J}_l(x, c_i, c_f) = \sum_{x_i} \widehat{J}_l(x_i, c_i; x_i + x, c_f). \tag{2.140}$$

Thus, a single pass of the forward-backward algorithm computes all of the terms required for the conditional pseudo-log-likelihood, $Q(\theta_n, \theta)$, given the initial parameter θ_n.

2.5.2 EM Step: Maximization

The function $Q(\theta_n, \theta)$ is the sum of terms, each of which is the product of a function of the parameter θ times a conditional expectation of a statistic relative to the measure P_{θ_n}. The maximization of Q with respect to θ is complicated because of the relatively large number of individual data files. Table 2.1 summarizes the parameters and whether they are specified for each experiment. Each data file representing the experiment (g, e) introduces lattice parameters $d_{g,e}$ and $\kappa_{g,e}$ to describe the lattice step length and the initial offset, respectively. Additionally, each experimental condition g introduces additional parameters to describe the autocorrelation ρ^g and the noise σ^g. Further requiring parameters to describe the substep sizes and the transition

rates, the number of parameters becomes quite large. When Q is viewed as a function of all of these parameters at once, the number of parameters makes maximizing the function challenging.

We implement a method that maximizes a subset of the parameters at a time, creating a sequence of parameter estimates which increase the value of Q. We restate the functional form of Q showing explicitly how parameters are grouped by experimental conditions as

$$Q(\theta_n, \theta) = \sum_{g\in\mathcal{G}} \sum_{e\in\mathcal{E}_g} E_{\theta_n}[q^{g,e}(\theta)|\mathcal{Y}^{g,e}_{K_{g,e}}] \tag{2.141}$$

$$\begin{aligned} &= \sum_{g\in\mathcal{G}} \sum_{e\in\mathcal{E}_g} \Big\{ \sum_c \log \pi^0_\theta(0,c) E_{\theta_n}[I_c(C^{g,e}_0)|\mathcal{Y}^{g,e}_{K_{g,e}}] \\ &\qquad + \sum_x \sum_{c_i} \sum_{c_f} \widehat{S}^{g,e}_{K_{g,e}-1}(J(x,c_i,c_f)) \log \pi_{c_i}(x,c_f) \\ &\qquad - \frac{K_{g,e}+1}{2} \log 2\pi - (K_{g,e}+1)\log\sigma_g + \frac{1}{2}\log(1-{\rho_g}^2) \\ &\qquad - \frac{1}{2{\sigma_g}^2}\Big(\widehat{T}^{\rho}_{K_{g,e}}(Y^{g,e} - f_{g,e}(Z^{g,e}), Y^{g,e} - f_{g,e}(Z^{g,e}))\Big)\Big\}, \end{aligned} \tag{2.142}$$

where the function $f_{g,e}$ explicitly involves the parameters as

$$f_{g,e}(x,c) = d_{g,e} \cdot (x + \epsilon_g(c)) + \kappa_{g,e}. \tag{2.143}$$

The essence of our method is to group the parameters of θ into three updating clusters, $\theta = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$, where $\theta^{(1)} = \{d_{g,e}, \kappa_{g,e} : g \in \mathcal{G}, e \in \mathcal{E}_g\}$ consists of the parameters that vary for each experiment, $\theta^{(2)} = \{\rho_g, \sigma_g : g \in \mathcal{G}\}$ consists of parameters that vary over experimental conditions, and $\theta^{(3)} = \{\epsilon^g_c, u_c, v_c, \theta^+_c, \theta^-_c : c = 1, \ldots, N, g \in \mathcal{G}\}$ includes the remaining parameters to describe the size of substeps (which may vary by experiment group) and to characterize the rates. The procedure is to hold two parameter subgroups fixed while choosing parameters in the third subgroup to maximize Q in this restricted parameter space. That is, starting with parameter θ,

we define $\theta' = (\theta^{(1)'}, \theta^{(2)'}, \theta^{(3)'})$ according to

$$\theta^{(1)'} = \arg\max_{\theta^{(1)^*}} Q(\theta_n, (\theta^{(1)^*}, \theta^{(2)}, \theta^{(3)})), \tag{2.144}$$

$$\theta^{(2)'} = \arg\max_{\theta^{(2)^*}} Q(\theta_n, (\theta^{(1)'}, \theta^{(2)^*}, \theta^{(3)})), \tag{2.145}$$

$$\theta^{(3)'} = \arg\max_{\theta^{(1)^*}} Q(\theta_n, (\theta^{(1)'}, \theta^{(2)'}, \theta^{(3)^*})). \tag{2.146}$$

By this construction, we guarantee that $Q(\theta_n, \theta') \geq Q(\theta_n, \theta)$. This procedure is repeated until the parameter converges to a fixed point. The categorization of the parameter clusters makes the partial maximizations easy to implement.

First, consider the update for $\theta^{(1)}$, letting $\theta^{(2)}$ and $\theta^{(3)}$ remain fixed. The parameter pair $(d_{g,e}, \kappa_{g,e})$ only affects the contribution from a single experiment (g, e), $Q^{g,e}(\theta_n, \theta)$. Furthermore, within this function, a number of terms have no dependence on the lattice parameters. Neglecting the other fixed terms, for each experiment (g, e), we seek to choose $d'_{g,e}$ and $\kappa'_{g,e}$ to maximize

$$Q^{(1)}_{g,e}(d, \kappa) = -\frac{1}{2{\sigma_g}^2}\Big(\widehat{T}^{\rho}_{K_{g,e}}\big(Y^{g,e} - d(X^{g,e} + \epsilon(C^{g,e})) - \kappa, Y^{g,e} - d(X^{g,e} + \epsilon(C^{g,e})) - \kappa\big)\Big) \tag{2.147}$$

over d and κ. Using the bilinearity and symmetry of $\widehat{T}^{\rho}_{K_{g,e}}$, we can rewrite this as

$$\begin{aligned} Q^{(1)}_{g,e}(d, \kappa) = -\frac{1}{2{\sigma_g}^2}\Big(&\widehat{T}^{\rho_g}_{K_{g,e}}(Y^{g,e}, Y^{g,e}) - 2d\,\widehat{T}^{\rho_g}_{K_{g,e}}(X^{g,e} + \epsilon(C^{g,e}), Y^{g,e}) \\ &- 2\kappa\,\widehat{T}^{\rho_g}_{K_{g,e}}(Y^{g,e}, 1) + d^2\,\widehat{T}^{\rho_g}_{K_{g,e}}(X^{g,e} + \epsilon(C^{g,e}), X^{g,e} + \epsilon(C^{g,e})) \\ &+ 2d\kappa\,\widehat{T}^{\rho_g}_{K_{g,e}}(X^{g,e} + \epsilon(C^{g,e}), 1) + \kappa^2\big(1 - \rho_g^2 + (1 - \rho_g)^2 K_{g,e}\big)\Big). \end{aligned} \tag{2.148}$$

Setting the derivative with respect to (d, κ) equal to zero leads to a system of linear equations for d and κ,

$$\begin{pmatrix} \widehat{T}^{\rho}_{K}(X + \epsilon(C), X + \epsilon(C)) & \widehat{T}^{\rho}_{K}(X + \epsilon(C), 1) \\ \widehat{T}^{\rho}_{K}(X + \epsilon(C), 1) & (1 - \rho^2 + (1 - \rho)^2 K) \end{pmatrix} \begin{pmatrix} d \\ \kappa \end{pmatrix} = \begin{pmatrix} \widehat{T}^{\rho}_{K}(X + \epsilon(C), Y) \\ \widehat{T}^{\rho}_{K}(1, Y) \end{pmatrix}. \tag{2.149}$$

The solution to this system, $(d'_{g,e}, \kappa'_{g,e})$ is a local maximum of $Q^{(1)}_{g,e}$ because the matrix on the left is positive definite, arising from the bilinear operator T^{ρ}, thus giving an explicit method to calculate $\theta^{(1)\prime}$.

Next we consider updating the experimental condition parameters ρ_g and σ_g which are included in $\theta^{(2)}$. Whereas in the case of $\theta^{(1)}$ we only needed to consider a single experiment (g, e), this time, we must consider all experiments that share the common experimental condition g. For a fixed $g \in \mathcal{G}$, by again dropping any terms that do not depend on these parameters and holding all other parameters fixed, we find that ρ' and σ' must maximize the function

$$Q^{(2)}_g(\rho, \sigma) = -\Big(\sum_{e\in\mathcal{E}_g}(K_{g,e}+1)\Big)\log\sigma + \frac{\#\mathcal{E}_g}{2}\log(1-\rho^2) \\ - \frac{1}{2\sigma^2}\Big((1-\rho^2)\sum_{e\in\mathcal{E}_g}\widehat{S}^{g,e} - \rho^2\sum_{e\in\mathcal{E}_g}\widehat{B}^{g,e} - \rho\sum_{e\in\mathcal{E}_g}\widehat{R}^{g,e}\Big), \quad (2.150)$$

where we have abbreviated the terms which are independent of ρ and σ

$$\widehat{S}^{g,e} = \widehat{S}^{g,e}_{K^{g,e}}\big((Y - d'_{g,e}(X+\epsilon(C)) - \kappa'_{g,e})^2\big), \quad (2.151)$$

$$\widehat{B}^{g,e} = \widehat{B}^{g,e}_{K^{g,e}}\big((Y - d'_{g,e}(X+\epsilon(C)) - \kappa'_{g,e})^2\big), \quad (2.152)$$

$$\widehat{R}^{g,e} = \widehat{R}^{g,e}_{K^{g,e}}\big(Y - d'_{g,e}(X+\epsilon(C)) - \kappa'_{g,e}, Y - d'_{g,e}(X+\epsilon(C)) - \kappa'_{g,e}\big), \quad (2.153)$$

and $\#\mathcal{E}_g$ is the number of experiments in the group g. For any choice of ρ, the value of σ which maximizes $Q^{(2)}_g$ will be

$$\sigma^2 = \Big(\sum_{e\in\mathcal{E}_g}(K_{g,e}+1)\Big)^{-1}\Big((1-\rho^2)\sum_{e\in\mathcal{E}_g}\widehat{S}^{g,e} - \rho^2\sum_{e\in\mathcal{E}_g}\widehat{B}^{g,e} - \rho\sum_{e\in\mathcal{E}_g}\widehat{R}^{g,e}\Big). \quad (2.154)$$

In principle, since σ^2 has been expressed as a function of ρ, the maximization has been reduced to a one-dimensional maximization, which can be shown to be a root of a cubic polynomial. Alternatively, it is equally simple to implement a conjugate gradient maximization routine in two dimensions, especially since we will need such an algorithm in arbitrary dimensions for the final maximization step. The gradient

of $Q_g^{(2)}$ is computed to be

$$\frac{\partial Q_g^{(2)}}{\partial \rho} = -\frac{\rho(\#\mathcal{E}_g)}{1-\rho^2} - \frac{1}{2\sigma^2}\Big(-2\rho \sum_{e\in\mathcal{E}_g} \widehat{S}^{g,e} - 2\rho \sum_{e\in\mathcal{E}_g} \widehat{B}^{g,e} - \sum_{e\in\mathcal{E}_g} \widehat{R}^{g,e}\Big), \tag{2.155}$$

$$\begin{aligned}\frac{\partial Q_g^{(2)}}{\partial \sigma^2} = &-\frac{1}{2\sigma^2}\Big(\sum_{e\in\mathcal{E}_g}(K_{g,e}+1)\Big) \\ &+ \frac{1}{2\sigma^4}\Big((1-\rho^2)\sum_{e\in\mathcal{E}_g}\widehat{S}^{g,e} - \rho^2 \sum_{e\in\mathcal{E}_g}\widehat{B}^{g,e} - \rho\sum_{e\in\mathcal{E}_g}\widehat{R}^{g,e}\Big).\end{aligned} \tag{2.156}$$

The third update to consider is for $\theta^{(3)}$, which contains parameters that affect every experiment such as rate parameters and relative substep positions. This involves maximizing the function $Q^{(3)}(\theta^{(3)})$ which is defined as

$$\begin{aligned}Q^{(3)}(\theta^{(3)}) = &\sum_{g\in\mathcal{G}}\sum_{x,c_i,c_f}\Big(\sum_{e\in\mathcal{E}_g}\widehat{J}^{g,e}_{K_{g,e}}(x,c_i,c_f)\Big)\log a_g(0,c_i;x,c_f) \\ &+ \sum_{g\in\mathcal{G}}\Big\{\sum_c \epsilon_g(c)\frac{1}{{\sigma'_g}^2}\Big(\sum_{e\in\mathcal{E}_g} d'_{g,e}\widehat{T}^{g,e}_{K_{g,e}}\big(I_c(C), Y - d'_{g,e}\,X - \kappa_{g,e}\big)\Big) \\ &\quad - \sum_c\sum_{c'}\frac{1}{2}\epsilon_g(c)\epsilon_g(c')\frac{1}{{\sigma'_g}^2}\Big(\sum_{e\in\mathcal{E}_g} {d'}^2_{g,e}\widehat{T}^{g,e}_{K_{g,e}}\big(I_c(C), I_{c'}(C)\big)\Big)\Big\}.\end{aligned} \tag{2.157}$$

If the detailed balance parameters $\vartheta^{\pm}$ are explicitly coupled to the size of the substep positions ϵ_c, then the two components of this function must be maximized together. However, when the positions of the substeps are decoupled from the transition rates, then the two components of $Q^{(3)}$ can be maximized independently each other. This is especially convenient, because the second component corresponds to minimizing a quadratic form. That is, for each experiment (g,e), the $(N-1)\times(N-1)$ matrix $T_C^{g,e}$ composed with entries

$$T_C^{g,e}(c-1,c'-1) = {d'}^2_{g,e}\widehat{T}^{g,e}_{K_{g,e}}\big(I_c(C), I_{c'}(C)\big), \quad c,c' = 2,\ldots,N, \tag{2.158}$$

is a positive definite matrix, and consequently, so are the matrices

$$T_C^g = \sum_{e \in \mathcal{E}_g} T_C^{g,e}, \tag{2.159}$$

$$T_C = \sum_{g \in \mathcal{G}} \frac{1}{{\sigma'}_g^2} T_C^g. \tag{2.160}$$

Also, introduce the vectors $\vec{\zeta}_{g,e} \in \mathbb{R}^{N-1}$ with entries

$$\zeta_{g,e}(c-1) = {d'}_{g,e} \widehat{T}^{g,e}_{K_{g,e}} \big(I_c(C), Y - {d'}_{g,e} X - {\kappa'}_{g,e} \big), \quad c = 2, \ldots, N, \tag{2.161}$$

as well as the corresponding sums

$$\vec{\zeta}_g = \sum_{e \in \mathcal{E}_g} \vec{\zeta}_{g,e}, \tag{2.162}$$

$$\vec{\zeta} = \sum_{g \in \mathcal{G}} \frac{1}{{\sigma'}_g^2} \vec{\zeta}_g. \tag{2.163}$$

If we write $\vec{\epsilon}_g = (\epsilon_g(2), \ldots, \epsilon_g(N))$, then we can rewrite $Q^{(3)}$ as

$$\begin{aligned} Q^{(3)}(\theta^{(3)}) = &\sum_{g \in \mathcal{G}} \sum_{x, c_i, c_f} \Big(\sum_{e \in \mathcal{E}_g} \widehat{J}^{g,e}_{K_{g,e}}(x, c_i, c_f) \Big) \log a_g(0, c_i; x, c_f) \\ &+ \sum_{g \in \mathcal{G}} \frac{1}{{\sigma'}_g^2} \Big\{ \langle \vec{\zeta}_g, \vec{\epsilon}_g \rangle - \frac{1}{2} \langle \vec{\epsilon}_g, T_C^g \cdot \vec{\epsilon}_g \rangle \Big\}. \end{aligned} \tag{2.164}$$

With the quadratic form now apparent, when the substep positions are decoupled from the transition rates, the substeps can be directly determined by solving the linear system,

$$T_C^g \cdot \vec{\epsilon}_g = \vec{\zeta}_g, \quad g \in \mathcal{G}, \tag{2.165}$$

for the case that each group has separate substep positions. If the parameters are constrained so that substep positions are the same over all experiments, $\vec{\epsilon}_g = \vec{\epsilon}$, then the linear systems for each g is replaced with the single system

$$T_C \cdot \vec{\epsilon} = \vec{\zeta}. \tag{2.166}$$

The terms that include the logarithms of transition probabilities coming from the matrices A also need to be maximized. To accomplish this, we simply use a conjugate gradient maximization routine over the transition rate parameters, computing the gradient numerically. If the substep positions are coupled with the transition rates through a detailed balance relationship, then we add to this numerical gradient the corresponding analytic derivatives of the quadratic form(s). The repeated computation of the transition probabilities A is typically the most costly computational step in the maximization, particularly when the transition rates are parametrized to satisfy detailed balance.

With these three partial maximizations, we update a current parameter choice θ to θ' such that $Q(\theta_n, \theta') \geq Q(\theta_n, \theta)$. Each of the steps maximizes a small collection of parameters at a time, as opposed to attempting to maximize over all of the parameters at once. Furthermore, most of the maximization routines can be computed explicitly, although some of them must be performed numerically. The three maximizations are repeated in turn until the parameter converges to complete the maximization step. Consequently, the EM algorithm can be performed efficiently, with only a single pass for the conditional expectation required per iteration, with a maximization step that is also straight forward. The E- and M-steps are then repeated until the parameters converge with the desired tolerance

2.5.3 Viterbi Algorithm

To round out the discussion of the hidden Markov model analysis, we give the extension of the Viterbi algorithm which was presented in the introduction. This algorithm, given a model described by the parameter θ, identifies the path out of all possible paths for the hidden Markov process $\{Z_k\}$ which maximizes the likelihood $L(\theta; \vec{z}, \vec{y})$. Because the ratio $\dfrac{L(\theta; \vec{z}, \vec{y})}{L(\theta; \vec{y})}$ represents the conditional density of the path $\vec{z}$ given the observations $\vec{y}$, this maximum likelihood path represents the best path conditioned

on the observations.

The algorithm is generalized in the obvious manner. We also explicitly change the algorithm to store the log-likelihood of the paths rather than the likelihood, as the true likelihood is subject to underflow. For each time index $k = 0, \ldots, K$ and state value z, we compute the maximum log-likelihood of partial sequences ending with final state z, $\delta_k(z)$. This is defined by the inductive relationship

1. Initialization:

$$\delta_0(z) = \log \pi_\theta^0(z) + \log b_0(y_0)(z_0; z_0), \tag{2.167}$$

2. Recursion:

$$\delta_{k+1}(z') = \max_z \delta_k(z) + \log a(z; z') + \log b(y_k, y_{k+1})(z; z'). \tag{2.168}$$

In addition, at each time $k = 1, \ldots, K$ and for each final state z', we record the previous site $\psi_k(z') = z$ which leads to the corresponding maximum log-likelihood,

$$\delta_k(z') = \delta_{k-1}(z) + \log a(z; z') + \log b(y_{k-1}, y_k)(z; z'). \tag{2.169}$$

The final path $(z_0^*, \ldots, z_K^*)$ is constructed by first finding z_K^* which maximizes δ_K and backtracking using the path stored in ψ:

$$z_K^* = \arg\max_z \delta_K(z), \tag{2.170}$$

$$z_{k-1}^* = \psi_k(z_k^*), \qquad k = 1, \ldots, K. \tag{2.171}$$

2.5.4 Bayesian Methods

Before concluding this chapter, it will also be useful to briefly discuss a Bayesian approach to parameter values. The idea is that one or more of the parameters are treated as additional unobserved random variables. So, we augment the probability space that has reference measure P_0 to include a random variable, Θ, representing

a parameter taking values in a discrete set Ω, and which has some specified prior distribution $\nu(\theta)$,

$$P_0[\Theta = \theta] = \nu(\theta). \tag{2.172}$$

Each possible parameter value, $\theta \in \Omega$, corresponds to a specific choice of parameters. The likelihood ratio $\Lambda(\theta)$ that induced the change of measure to P_θ is now interpreted as a conditional expectation on the partition of possible values of Θ,

$$\Lambda(\theta) = \frac{dP_\theta}{dP_0} = E_0[\frac{dP_\Theta}{dP_0}|\Theta = \theta]. \tag{2.173}$$

The posterior probability that $\Theta = \theta$ given the observations can then be calculated using Bayes' rule,

$$P_0[\Theta = \theta|\mathcal{Y}] = \frac{E_0[I_\theta(\Theta)\frac{dP_\Theta}{dP_0}|\mathcal{Y}]}{E_0[\frac{dP_\Theta}{dP_0}|\mathcal{Y}]} = \frac{\nu(\theta)L(\theta;\vec{y})}{\sum_{\theta'}\nu(\theta')L(\theta';\vec{y})}. \tag{2.174}$$

Adaptations of the forward algorithm to compute posterior probabilities of parameters directly have been derived elsewhere for standard hidden Markov models (Elliott et al., 1997), and are generalized easily to the current situation. Instead of computing a single vector of forward variables α_l, one computes a forward variable $\alpha_l(\theta)$ for each possible $\theta \in \Omega$ as described in the Forward-Backward Algorithm 2.5.1, where the transition matrix A and the observation weight matrix B are constructed using the parameter θ. The initialization of each $\alpha_l(\theta)$ incorporates the prior distribution of Θ,

$$\widetilde{\alpha}_0(\theta) = \nu(\theta)\,\pi^0_\theta \cdot B_{\theta,0}, \tag{2.175}$$

with the recursive step remaining

$$\widetilde{\alpha}_{l+1}(\theta) = \alpha_l(\theta) \cdot (A_\theta \star B_{\theta,l+1}). \tag{2.176}$$

The only fundamental change in the algorithm is the scaling factor c_l, which renormalizes the forward variables $\alpha_l(\theta)$,

$$\alpha_l(\theta) = \frac{1}{c_l}\widetilde{\alpha}_l(\theta). \tag{2.177}$$

Because there is an entire family of forward variables, the scaling factor c_l will be the calculated using the sum over Ω,

$$c_l = \sum_{\theta \in \Omega} \alpha_l(\theta) \cdot \vec{1}. \tag{2.178}$$

Consequently, we no longer have the identity that the sum of coordinates of $\alpha_l(\theta)$ would be 1. Instead, the sum now computes the posterior probability,

$$P_0[\Theta = \theta | \mathcal{Y}_l] = \alpha_l(\theta) \cdot \vec{1}. \tag{2.179}$$

Similar modifications can be applied to the backward variables β_l to obtain $\beta_l(\theta)$, thereby allowing calculations of conditional expectations.

However, our interest is not so much in computing in a Bayesian framework. Rather, we wish to compare different parameter choices. For example, our main application will be to create a grid of possible lattice step-sizes d and/or base offset positions κ. One approach of determining the best parameter on the grid is to compute the likelihoods of each grid point and to choose the grid point with the greatest likelihood. The other approach would be to perform a Bayesian calculation, and choosing the maximum *a posteriori* parameter. In fact, the Bayes' formula 2.174 for computing the posterior probabilities in terms of likelihood ties the two approaches together. Let θ and χ be two possible parameters and consider the ratio of their posterior probabilities:

$$\frac{P_0[\Theta = \theta | \mathcal{Y}]}{P_0[\Theta = \chi | \mathcal{Y}]} = \frac{\nu(\theta)\, L(\theta; \vec{y})}{\nu(\chi)\, L(\chi; \vec{y})}. \tag{2.180}$$

Since computing the log-likelihood is more common, we re-express this with logarithms,

$$\log L(\theta; \vec{y}) - \log L(\chi; \vec{y}) = \log(\frac{P_0[\Theta = \theta | \mathcal{Y}]}{\nu(\theta)}) - \log(\frac{P_0[\Theta = \chi | \mathcal{Y}]}{\nu(\chi)}). \tag{2.181}$$

When the prior distribution is uniform on Ω, then maximum likelihood parameters correspond exactly with maximum *a posteriori* parameters.

Chapter 3
Simulation Results

Having established the theoretical foundation for hidden Markov model analysis of single-molecule assays of the protein kinesin in the previous chapter, we now proceed to analyze the effectiveness of the implementation through the use of a simulated data set based on the two-state Fisher-Kolomeisky model. We also briefly look at a simulated data set with the four-state Fisher-Kolomeisky model, which pushes the limits of the algorithm as currently implemented so that we were unable to complete the analysis. This chapter will discuss the generation of the simulated data, details of implementing the filtering algorithms, and a discussion of results on analyzing the simulated data sets.

3.1 Generating Data

Creating simulated hidden Markov model data consists in generating the Markov process corresponding to the hidden state, $Z_0, \ldots, Z_K$, followed by generating the observation process $Y_0, \ldots, Y_K$. The Markov process is determined through the transition rates, and the observation process is determined from the lattice parameters d and κ, the relative substep positions ϵ_c, and the noise parameters ρ and σ. For the context of this discussion, we assume that these parameters are fixed. Their actual values will be given during the discussion of the results.

Randomness will be introduced using a pseudo-random number generator, implemented in the GNU Scientific Library (Galassi et al., 2001) and described in Numerical Recipes (Press et al., 1988). We use Park and Miller's MINSTD generator with Bayes-Durham shuffle to generate a sequence of pseudo-random numbers distributed uniformly on the interval $[0, 1]$. Random variables with other distributions

were generated from the sequence of uniform variables. Normally distributed random variables can be generated using the Box-Müller transform (Press et al., 1988).

Random variables with an arbitrary cumulative distribution function $F_V(v)$ can also be generated as a function of uniform random variables. The cumulative distribution function is defined as

$$F_V(v) = P[V \leq v], \tag{3.1}$$

and has the properties that it is non-decreasing, takes values between zero and one, inclusive, and has the limits

$$\lim_{v \to -\infty} F_V(v) = 0, \tag{3.2}$$

$$\lim_{v \to +\infty} F_V(v) = 1. \tag{3.3}$$

Because uniform random variables take values in the interval $[0, 1]$, which corresponds to the range of cumulative distribution functions, we convert uniform random variables through the function $V(u)$ defined as

$$V(u) = \sup\{v \in \mathbb{R} : F_V(v) < u\}. \tag{3.4}$$

If U is a uniform random variable, then $V(U)$ will have the distribution function F_V, as desired (Williams, 1991). When the cumulative distribution function is invertible, this function corresponds to the inverse $V(u) = F_V^{-1}(u)$. For example, an exponentially distributed random variable T with rate parameter λ has a cumulative distribution function, F_T, which is given by

$$F_T(t) = P[T \leq t] = (1 - e^{-\lambda t}) I_{[0,\infty)}(t). \tag{3.5}$$

The inverse of this function maps a number $u \in (0, 1)$ to the value

$$F_T^{-1}(u) = -\frac{1}{\lambda} \log(1 - u). \tag{3.6}$$

Since $1-U$ has a uniform distribution when U does, an exponential random variable can be generated by taking the negative logarithm of U and dividing by the rate parameter,

$$T = -\frac{1}{\lambda}\log U. \tag{3.7}$$

As another example, consider a discrete random variable. In this case, the cumulative distribution function is a step function with jumps at each of the possible discrete values. The prescription, then, says to find the jump at which the function F_V starts below u and ends above or at u,

$$F_V(V-) < u \text{ and } F_V(V) \geq u. \tag{3.8}$$

There are two approaches to computing the discretized Markov process. The first approach is to simulate the discretized model itself. That is, given the transition rates u_c and v_c and the length of the time interval Δt between samples, one computes the transition probabilities $\pi_{c_i}(x, c_f)$, which depend only on the initial state c_i, the final state c_f and the change in the number of lattice sites, and which are defined as

$$\pi_{c_i}(x, c_f) = P_\theta[Z(t+\Delta t) = (x_i + x, c_f) | Z(t) = (x_i, c_i)], \quad x_i \in \mathbb{Z}. \tag{3.9}$$

Given the state of the hidden process $Z_k = (x_i, c_i)$ at the k^{th} sample, the subsequent state, Z_{k+1}, is randomly selected from the distribution π_{c_i}. The second approach is to simulate the original, continuous time Markov jump process, $Z(t)$, and then use the state of the continuous process $Z_k = Z(t_k)$ at the sampling times $t_k = k\Delta t$ Gillespie (1976). To simulate the jump process, we generate the sequence of visited states $\{z_n = (x_n, c_n) : n \geq 0\}$ as well as the amount of time between transitions $\{\tau_n : n \geq 0\}$. The random time τ_n has an exponential distribution with rate parameter λ_{c_n}, the total outgoing transition rate for the current state z_n. The subsequent state z_{n+1} will move one state forward, $z_{n+1} = (x_n, c_n + 1)$, with probability p_{c_n} or one state backward, $z_{n+1} = (x_n, c_n - 1)$ with probability $q_{c_n} = 1 - p_{c_n}$, which are defined in terms of the

forward and backward transition rates u_c and v_c according to (recall Equation 2.5)

$$\begin{aligned} \lambda_c &= u_c + v_c, \\ p_c &= u_c/\lambda_c, \\ q_c &= v_c/\lambda_c. \end{aligned} \tag{3.10}$$

Transition times T_n specify when the transitions occur,

$$T_n = \sum_{i=0}^{n-1} \tau_i, \tag{3.11}$$

with $T_0 = 0$. Then for each $t \in [T_n, T_{n+1})$, the jump process takes the value $Z(t) = z_n$.

Even though the hidden process $Z_k = Z(t_k)$ can be generated without requiring the transition probabilities $\pi_{c_i}(x, c_f)$ for a given sampling time Δt, these probabilities will be required later to implement the forward-backward algorithm. Consequently, we now discuss how to compute these probabilities. Recall that the transition probabilities are defined as the solutions of the N initial value problems of the same linear system of differential equations (recall Equation 2.2), where for each $c_i = 1, \ldots, N$,

$$\frac{d}{dt}\pi_{c_i}(t)(x,c) = u_{c-1}\pi_{c_i}(t)(x,c-1) + v_{c+1}\pi_{c_i}(t)(x,c+1) - \lambda_c \pi_{c_i}(t)(x,c), \tag{3.12}$$

$$\pi_{c_i}(0)(x,c) = I_0(x) I_{c_i}(c), \tag{3.13}$$

and integrating until time Δt,

$$\pi_{c_i}(x,c) = \pi_{c_i}(\Delta t)(x,c). \tag{3.14}$$

One method of computing transition probabilities is to directly implement a numerical solution to the initial value problem, such as with a Runge-Kutta algorithm. At least one challenge to this technique is that the system becomes stiff when transition rates have significantly different time scales, requiring a very short time step to retain stability.

Instead of finding a generic algorithm for stiff systems, we instead use an algorithm that takes advantage of the fact that the system describes transition probabilities,

known as the uniformization procedure (Stewart, 1994). This method is based on the use of a fundamental solution for the system of differential equations. If Q represents the infinite matrix with transition rates from a given state on the corresponding row,

$$Q(x_i, c_i; x_f, c_f) = \begin{cases} u_{c_i}, & \text{if } x_f = x_i, c_f = c_i + 1, \\ -\lambda_{c_i}, & \text{if } x_f = x_i, c_f = c_i, \\ v_{c_i}, & \text{if } x_f = x_i, c_f = c_i - 1, \\ 0, & \text{otherwise,} \end{cases} \tag{3.15}$$

where the cases of the boundaries $c_i = 1$ or $c_i = N$ should be obvious. Then, the system of differential equations of Equation 3.12 can be written using matrix notation,

$$\frac{d}{dt}\pi_{c_i}(t) = \pi_{c_i}(t) \cdot Q. \tag{3.16}$$

The fundamental solution to this equation, e^{tQ}, maps the initial condition to the solution of System 3.12 at time t,

$$\pi_{c_i}(t) = \pi_{c_i}(0)e^{tQ}. \tag{3.17}$$

When the state space is finite, the fundamental solution corresponds to the matrix exponential,

$$e^{tQ} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n. \tag{3.18}$$

In our case, we really only need finite dimensional approximations to the N vectors,

$$\pi_{c_i} = \pi_{c_i}(\Delta t) = \pi_{c_i}(0)e^{\Delta t Q}. \tag{3.19}$$

The uniformization procedure provides a numerically stable method to compute these vectors. Letting $\lambda = \max_c \lambda_c$, we define a new matrix P, which will commute with Q, according to

$$P = I + \frac{1}{\lambda} Q, \tag{3.20}$$

where I is the identity matrix. The uniformization procedure is based on the equality

$$e^{tQ} = e^{-\lambda t} e^{\lambda t P}, \tag{3.21}$$

which can be alternately expressed as the series

$$e^{tQ} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n. \tag{3.22}$$

Because the sum of elements in the rows of Q add to zero, the sum of the elements in the rows of P add to one, with all of the elements non-negative. Thus, P is a stochastic matrix. Recall that λ_c represents the rate at which the process $Z(t)$ will make a transition out of the current state $Z(t) = (x, c)$. Then $\lambda \geq \lambda_c$, which represents a faster rate, would lead to more frequent transitions. The stochastic matrix P corresponds to the jump distributions for each of the states with this higher transition rate, corrected to account for pseudo-transitions which leave the state unchanged. The uniformization equation (3.22) expresses e^{tQ} in terms of making a random number of transitions, N_λ, each occuring at the rate λ, and where each of the transitions are determined by the matrix P. The numerical implementation of the uniformization procedure is to compute a truncated series, followed by multiplying by $e^{-\lambda \Delta t}$,

$$\pi_{c_i}(\Delta t) = e^{-\lambda \Delta t}(\pi_{c_i}(0) \sum_{n=0}^{n_{\max}} \frac{(\lambda \Delta t)^n}{n!} P^n). \tag{3.23}$$

Since P is stochastic, one can determine $n_{\max}$ so that the approximate transition probabilities are uniformly within a predetermined ε of the true transition probabilities (Stewart, 1994) by choosing $n_{\max}$ to satisfy

$$e^{-\lambda \Delta t} \sum_{n=0}^{n_{\max}} \frac{(\lambda \Delta t)^n}{n!} \geq 1 - \varepsilon. \tag{3.24}$$

3.2 Two-State Data

The primary set of simulated data that we discuss is for a simple model with two states, $N = 2$. These data were generated using the first proposed method for simulating data, based on the discrete time transition probabilities and without simulating each of the transitions. The rate parameters are modeled with the two-state model

[ATP]	Approximate Load		
	1 pN	3.5pN	5.5 pN
10 μM	1.05 pN	3.54 pN	5.61 pN
100 μM	1.05 pN	3.59 pN	5.61 pN
2 mM	1.06 pN	3.64 pN	5.76 pN

TABLE 3.1. Load vs [ATP] conditions for simulated and experimental data.

of Fisher and Kolomeisky (Fisher and Kolomeisky, 2001), as that model gave good fits to the observed velocity profiles of the experimental data under consideration. The model uses the detailed balance parametrization of Equations 2.11 and 2.12 and [ATP] dependence given by Equations 2.14 and 2.15, with parameters

$$\begin{array}{cc} k_0 = 1.8\ \mu\mathrm{M}^{-1}s^{-1}, & k_0' = 2.8\times 10^{-4}\ \mu\mathrm{M}^{-1}s^{-1}, \quad c_0 = 16\ \mu\mathrm{M}, \\ u_2^0 = 108\, s^{-1}, & v_2^0 = 6.0\, s^{-1}, \end{array} \tag{3.25}$$

and load factors

$$\vartheta_1^+ = 0.135, \quad \vartheta_1^- = 0.75, \quad \vartheta_2^+ = 0.035, \quad \vartheta_2^- = 0.08. \tag{3.26}$$

The data were generated for nine different conditions chosen to match nine experimental conditions selected from the true kinesin assay data which we will use for filtering later. These conditions were selected to come from the three ATP concentrations of 10, 100 and 2000 μM, and the imposed opposing loads of approximately 1, 3.5, and 5.5 pN. The precise values of the load for each of the three concentrations are given in Table 3.1. For each experimental condition, parameters for the autocorrelation and noise scale were chosen to be reasonably close to preliminary results on the experimental data, and are shown in Table 3.2. We also generated different numbers of runs for the different ATP concentrations, as the lower concentrations had longer individual run lengths than the higher concentrations, also shown in Table 3.2. The positions of the substep corresponding to the second state of the model was chosen as $\epsilon_2 = 0.215$ for each of the different experimental conditions, which corresponds to the literal interpretation of detailed balance coming only from spatial transitions,

$$\epsilon_2 - \epsilon_1 = 0.215 = 0.135 + 0.08 = \vartheta_1^+ + \vartheta_2^-. \tag{3.27}$$

Load	1 pN			3.5 pN		
[ATP]	10 μM	100 μM	2 mM	10 μM	100 μM	2 mM
ρ	0.3	0.3	0.3	0.25	0.25	0.25
σ [nm]	7.0	6.8	7.0	4.5	4.5	4.5
# runs	8	16	24	8	16	24

Load	5.5 pN		
[ATP]	10 μM	100 μM	2 mM
ρ	0.18	0.18	0.25
σ [nm]	3.0	3.0	3.0
# runs	8	16	24

TABLE 3.2. Autocorrelation (ρ), noise scale (σ), and the number of simulated runs for each of the experimental conditions.

The remaining parameters to characterize the model provide the lattice step-size d and offset position κ for each of the generated runs. The step-sizes d were selected from a normal distribution with a mean of 8.1 nm and a standard deviation of 0.05 nm, to provide some variability in the projected lattice spacing yet remaining reasonably close to the theoretical lattice length of 8.2 nm. The offsets κ were selected from a normal distribution with a mean of -80 nm and a standard deviation of 4 nm.

3.2.1 Identifying Lattice Parameters

These data were then analyzed using the EM algorithm for the hidden Markov model described in Chapter 2. Estimates of the lattice parameters d and κ for individual runs often did not converge to the externally known true values. Upon more careful examination, this behavior was found to be a consequence of the likelihood function having many local maxima caused by the lattice structure of the problem. Holding all other parameters fixed, the likelihood function can be visualized as a surface parametrized by d and κ. Figure 3.1 demonstrates a typical surface of the log-likelihood with all other parameters fixed at the true values for a particular data set from the 1 pN load and 100 μM [ATP] group. If the initial parameter estimate starts in the wrong part of a valley, the algorithm will send the sequence of estimates toward the top of the

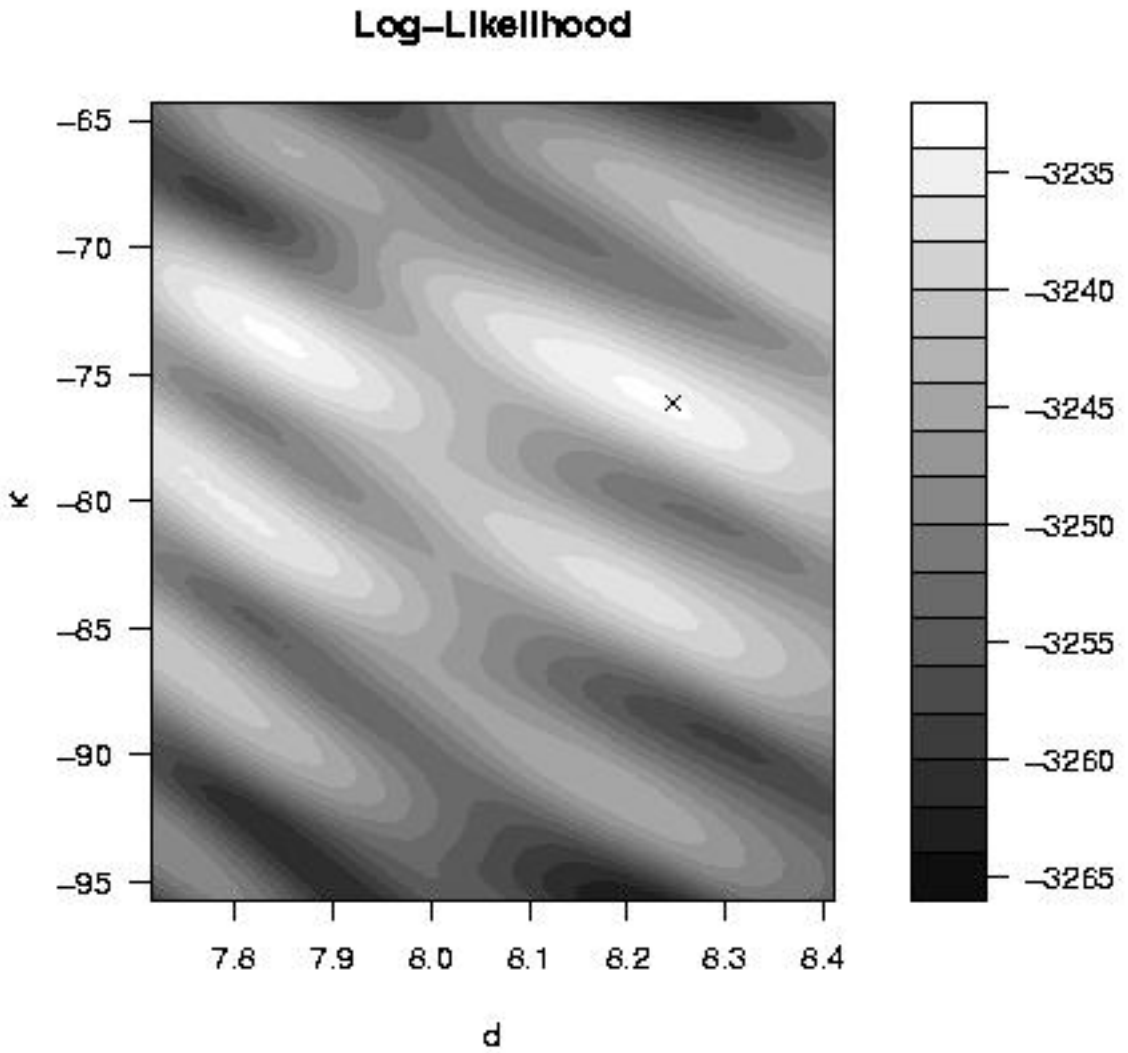


FIGURE 3.1. .
] A contour relief plot of the log-likelihood surface as a function of lattice step-size d and offset position κ for a simulated run selected from the experimental grouping of 1 pN load and 100 μM [ATP]. The true parameter position is marked with $\times$.

wrong peak.

The general features of this surface can be understood by considering how changing the step-size and offset position affect likelihood through the alignment of the true and modeled lattices. For simplicity in discussion, we consider a model that only has a single state. The log-likelihood can be regarded as a score for the extent of matching between the data and the model: the greater the match, the greater the score. First, consider the influence of the offset position κ for a fixed model lattice step-size d, which in general will not match the true lattice step-size. As κ varies, one might visualize the proposed lattice sliding relative to the true lattice. When sites on the proposed lattice align with the sites on the true lattice, the likelihood increases, whereas the likelihood decreases for sites out of alignment. The total likelihood accounts for the overall synchronization of sites on the true lattice relative to the model lattice. The likelihood as a function of κ, with d fixed, represents a vertical cross-section of the likelihood surface, such as in Figure 3.1. Changing the offset by the lattice step-size d does not change the location of lattice sites, but it does shift the labels on the model lattice by one position. This explains the near-periodicity of the likelihood in the κ-direction. The model assumes the initial condition $X_0 = 0$, so the first few observations of the bead will penalize the likelihood unless the model lattice has $X = 0$ near these observations. The highest peak in the log-likelihood as a function of κ corresponds to the best match for the initial lattice site.

Next consider the influence of the step-size d. When d is the correct lattice size, the two lattices can match exactly. As the model lattice size d increases, the alignment between all of the lattice sites degrade and the likelihood decreases. However, as d continues to increase, sites on the model lattice far from the origin will begin to align with the true lattice sites, even though they will not be a correct match. For a fixed offset position κ, this alignment continues to improve until the model lattice is off by one site relative to the true lattice over the sites that are visited by the run. A similar argument holds when d is decreased. As a rough guide, we expect peaks in

the likelihood to occur at step-sizes d when the span of the model lattice corresponds to the span of the true lattice with step-size d_0,

$$Md = M_0 d_0, \tag{3.28}$$

where M_0 is the number of lattice sites visited in the true lattice and M is the number of corresponding sites on the model lattice. If the offset position κ is allowed to vary when the lattice step-size d changes, then if d is increased, the alignment can be improved by causing κ to decrease, so that there is a smaller mismatch over a larger number of lattice sites. This accounts for the prominent directionality to the axes of the likelihood peaks.

Although this rough picture explains the prominent features of the likelihood surface, other factors influence the quality of fit as well. For instance, because of the inherent stochastic nature of the model, all sites on the lattice are not created equal. Different sites on the lattice will influence the likelihood with different weights, depending on how long the kinesin stays at the site. Also, the noise in the observations tends to smear the peaks out. Finally, the presence of intermediate steps due to the internal states leads to additional mismatch, particularly when different states have comparable lifetimes. Thus, the positions of the maxima in d-κ space is likely to change as the other parameters vary. The quantity and quality of data also influence the character of the likelihood surface. If either the number of samples at each lattice site increases or the scale of noise decreases, the separation in likelihood increases, both in terms of the difference in height between the peaks and the valleys, as well as the difference in height between neighboring peaks. This can be seen clearly by considering a second likelihood surface shown in Figure 3.2, coming from a slower experimental condition with [ATP]= 10 μM. If the number of sites visited increases, then the positions of the peaks move closer together in the d direction, in harmony with Equation 3.28.

Therefore, if the initial parameter estimates begin heading up the wrong peak, as

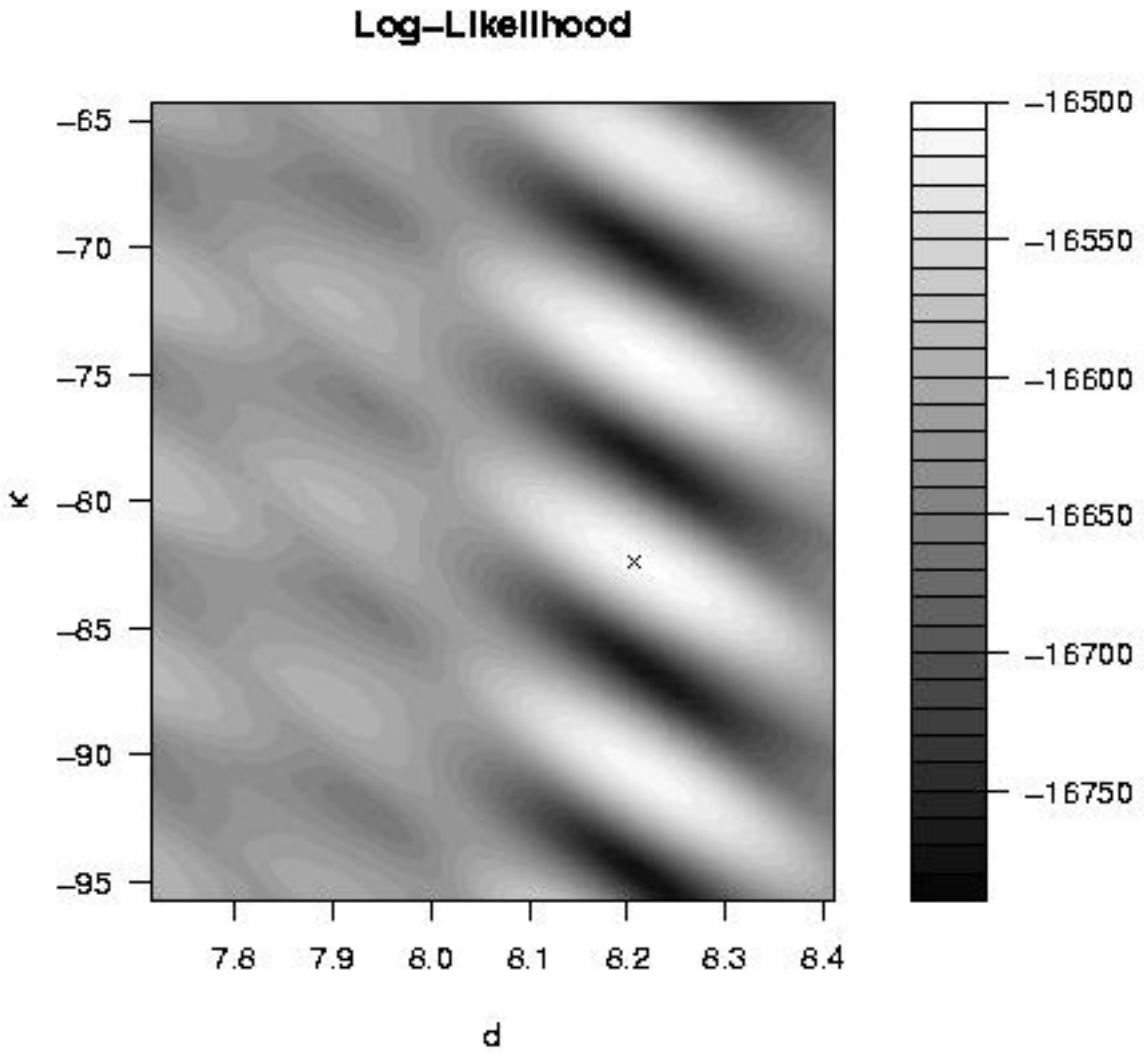


FIGURE 3.2. .

] A contour relief plot of the log-likelihood surface as a function of lattice step-size d and offset position κ for a simulated run selected from the experimental grouping of 1 pN load and 10 μM [ATP]. The true parameter position is marked with $\times$.

the EM algorithm progresses the lattice parameters will not converge to the correct position. One method we use to attempt to overcome this challenge is to evaluate the likelihood on a grid of points in d-κ space and then select that parameter from this collection which has the highest likelihood as the current parameter. Unfortunately, this is a costly procedure that grows proportionally to the number of points in the grid. Thus, a search for the maximum using a very refined grid is computationally expensive, particularly when dealing with all of the data runs. One possibility is to use a coarse, two-dimensional grid, so that one hopes at least to find the side of one of the higher peaks. If the EM algorithm has put the lattice parameters near the top of one of the peaks, one can choose to leave one of the lattice parameters fixed and see if the other parameter can be improved. Examining the log-likelihood profiles of Figures 3.1 and 3.2, we observe that the peaks occur on the lines of constant step-size d. But the peaks for approximately constant offset κ are not quite on horizontal lines. If the step-size d is held fixed, then a one-dimensional grid in the κ-direction should find the peaks corresponding to the offset positions $\kappa \pm d$. If the offset position κ is held fixed, then a one-dimensional grid in the d-direction runs the strong risk of missing the other peaks and only sampling the edges. To avoid this problem, one can instead search a narrow two-dimensional lattice.

3.2.2 Model Estimates

In addition to the lattice parameters d and κ, we also need the HMM algorithms to estimate effectively the transition rates and fractional positions of substeps. We will primarily discuss the results for the two-state model, both with uncoupled transition rates as well as with transition rates that obey detailed balance. With some exceptions, the estimated parameters were quite reasonable.

The theoretical substep of the generating model, $\epsilon_2 = 0.215$, was approximately recovered for nearly all of the nine experimental conditions. Figure 3.3 illustrates the

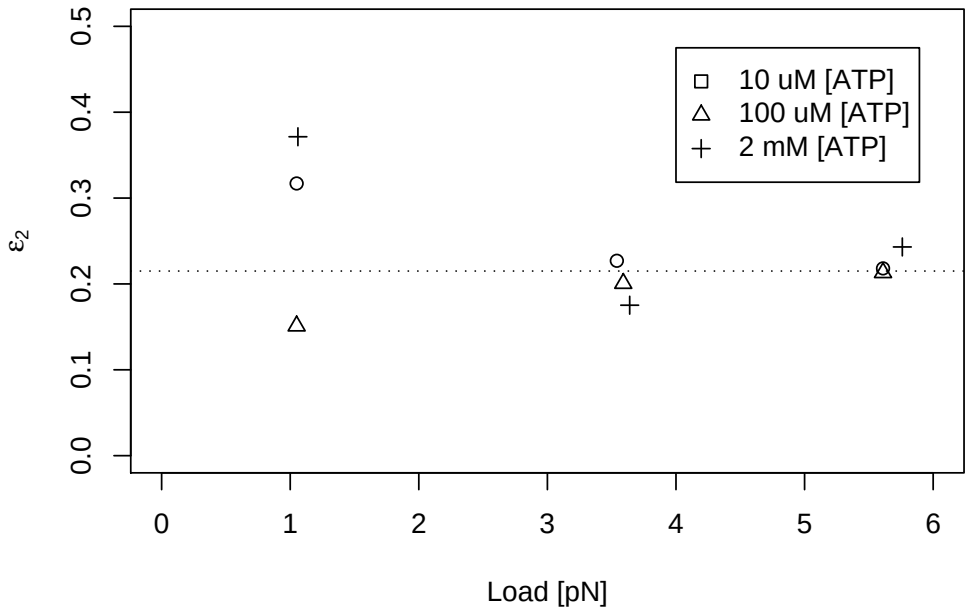


FIGURE 3.3. Estimated substep position for a two-state model with unconstrained transition rates. The true value is $\epsilon_2 = 0.215$.

estimated substep positions for each of the nine experimental conditions, clustered by load (since noise decreased as load increased), for the two-state model with unconstrained rate parameters. As expected, the low noise conditions do better than high noise. A similar illustration for the two-state model with rates described by detailed balance is shown in Figure 3.4, with the parameters very similar to the previous model.

Transition rates were estimated reasonably well, except for the experimental condition of 1 pN and 2 mM [ATP], which we discuss later. Figure 3.5 shows the estimated second order transition rate, k_0 for all of the experimental conditions. The scatter points represent point estimates coming from the unconstrained two-state model, while the curves represent the parametrized rate relative to the true underlying model and the estimated model with rates obeying detailed balance. Figure 3.6 shows the pseudo-second order transition rate k_0', and Figures 3.7 and 3.8 give the

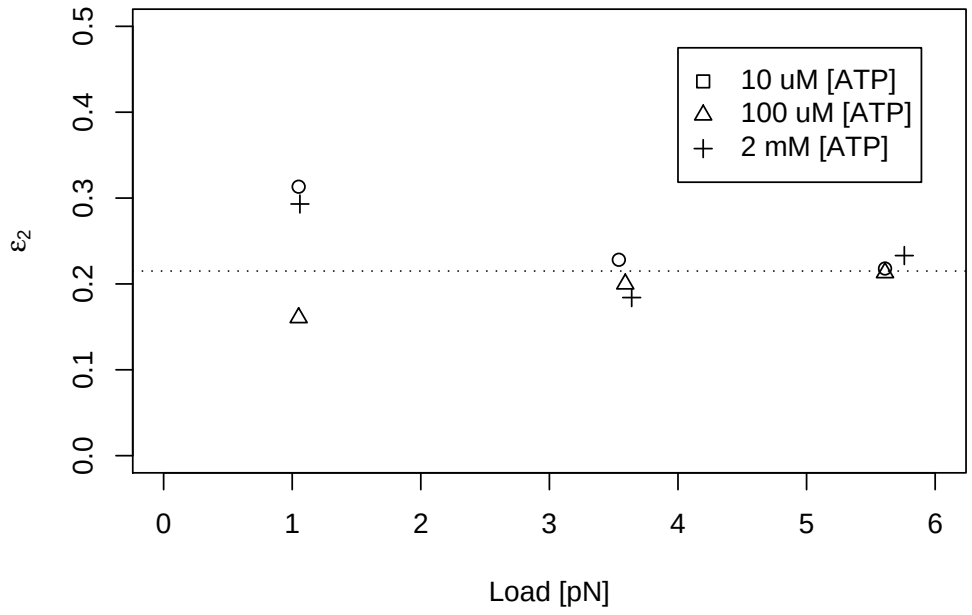


FIGURE 3.4. Estimated substep position for a two-state model with detailed balance. The true value is $\epsilon_2 = 0.215$.

forward and backward rates, u_2 and v_2, for the second state. Table 3.3 shows the estimated parameters for the detailed balance model with two states.

The experimental grouping of 1 pN and 2 mM [ATP] presented an interesting circumstance. Because the noise was relatively high and the time spent in a given state was relatively short, the one-state model did nearly as well as the two-state model, with a log-likelihood ratio of 8.377. If the corresponding χ^2 statistic were computed, $\chi^2 = 16.75$, with three degrees of freedom, then the P-value would be 0.00079. In a standard hypothesis test, this would be a significant result. As these models are not really nested, such a comparison is not technically justified, although it does suggest that the second state explains at least some aspect of the data. However, the improvement in likelihood is rather small compared to the improvements seen for other groups. Other experimental conditions often saw improvements of 100 or more, while the 3.5 pN, 2 mM [ATP] group had the second smallest change of approximately

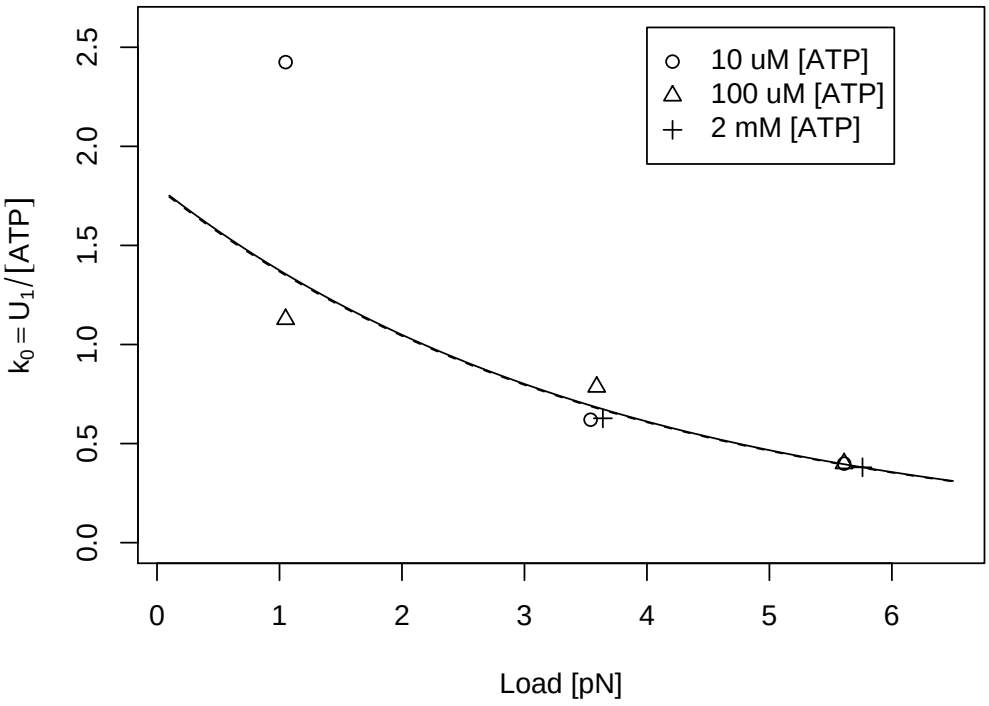


FIGURE 3.5. Estimated second-order transition rate k_0 for each experimental condition. The solid line represents the true underlying parametrization. The dashed line represents the parametrization given by the estimated model with detailed balance, but which is nearly exactly aligned with the true model. Scatter points represent unconstrained rate estimates for each of the separate conditions.

Parameter	True Value	Estimated Value
k_0	$1.8\ \mu\text{M}^{-1}s^{-1}$	$1.793\ \mu\text{M}^{-1}s^{-1}$
k_0'	$2.8 \times 10^{-4}\ \mu\text{M}^{-1}s^{-1}$	$1.4 \times 10^{-4}\ \mu\text{M}^{-1}s^{-1}$
u_2^0	$108\, s^{-1}$	$102.39\, s^{-1}$
v_2^0	$6.0\, s^{-1}$	$10.83\, s^{-1}$
ϑ_1^+	0.135	0.1353
ϑ_1^-	0.750	0.8114
ϑ_2^+	0.035	0.0298
ϑ_2^-	0.080	0.0235

TABLE 3.3. Maximum likelihood parameter estimates for the two-state detailed balance model.

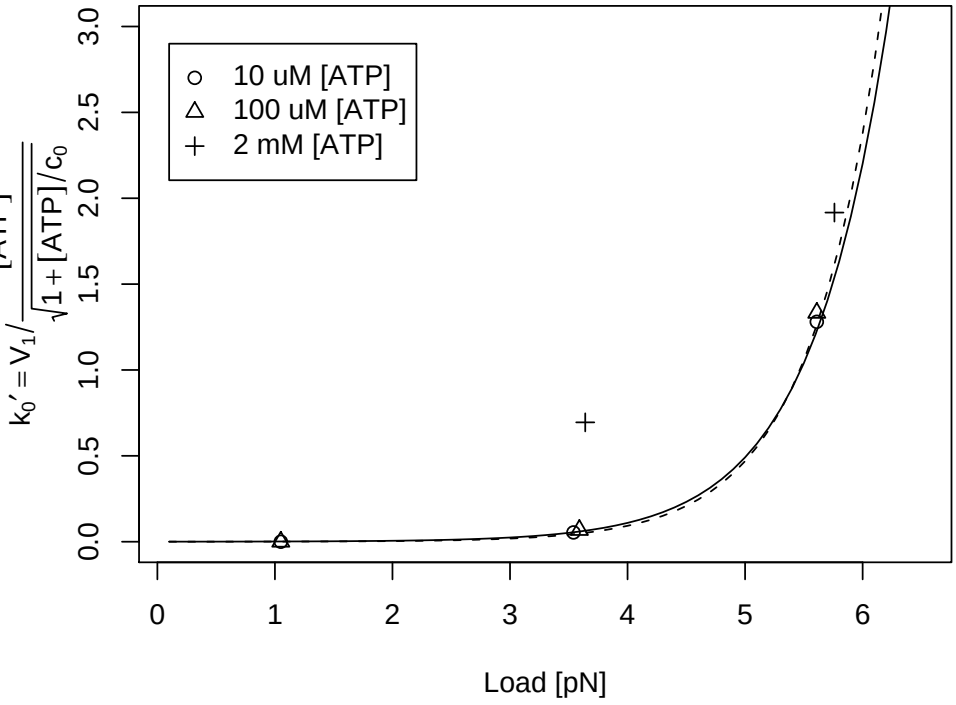


FIGURE 3.6. Estimated pseudo-second order transition rate k_0' for each experimental condition. The solid line represents the true underlying parametrization; the dashed line represents the parametrization given by the estimated model with detailed balance; the scatter points represent unconstrained rate estimates for each of the separate conditions.

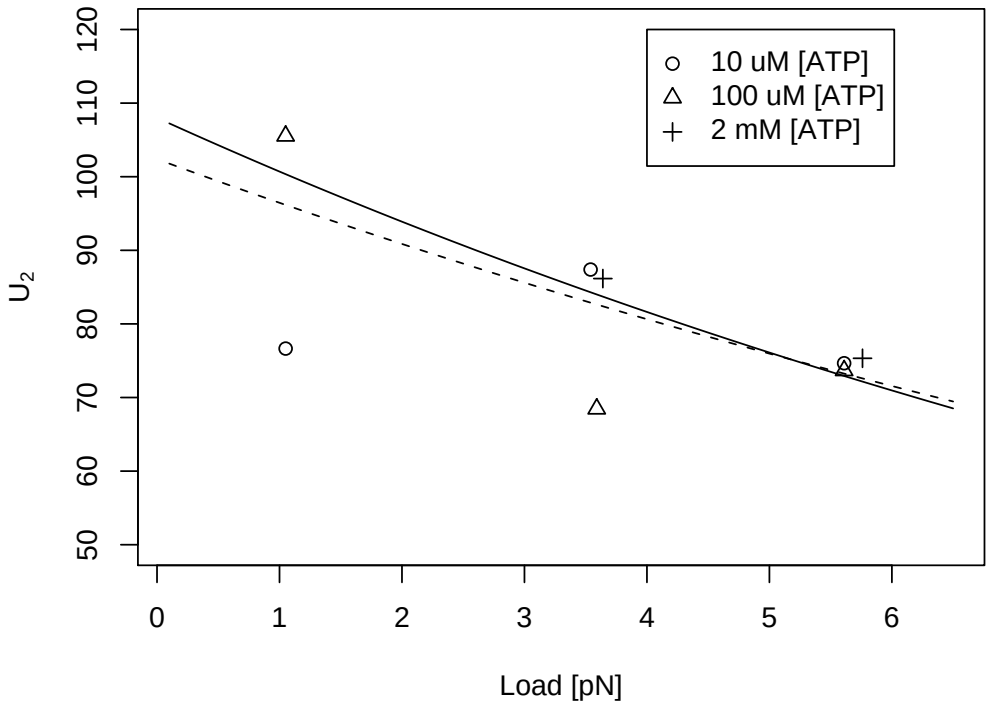


FIGURE 3.7. Estimated first-order transition rate u_2 for each experimental condition. The solid line represents the true underlying parametrization; the dashed line represents the parametrization given by the estimated model with detailed balance; the scatter points represent unconstrained rate estimates for each of the separate conditions.

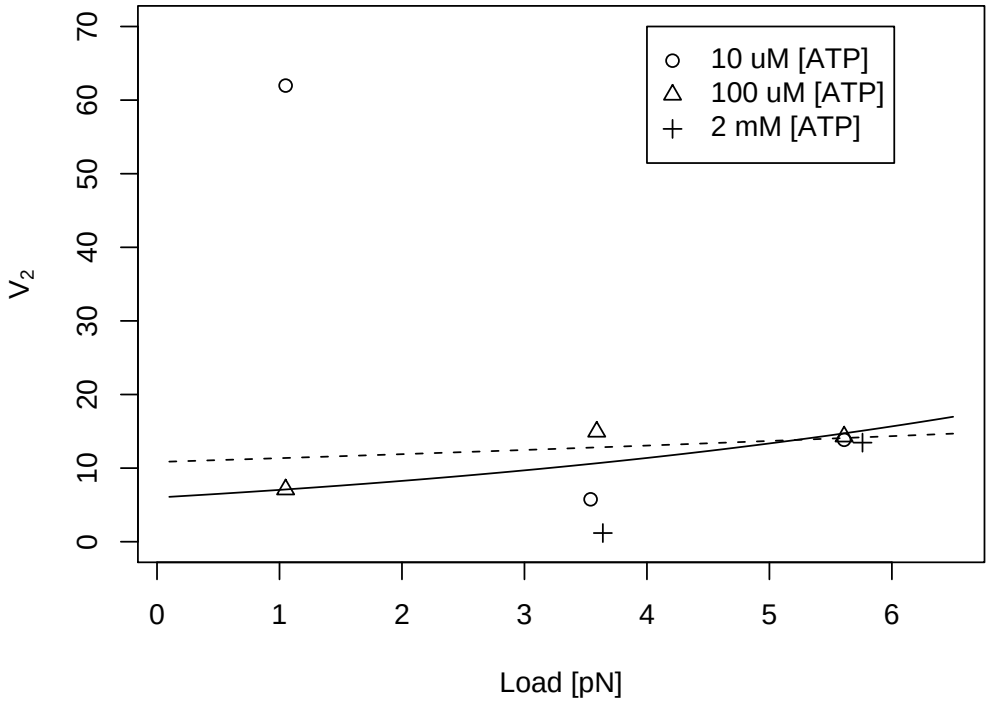


FIGURE 3.8. Estimated first-order transition rate v_2 for each experimental condition. The solid line represents the true underlying parametrization; the dashed line represents the parametrization given by the estimated model with detailed balance; the scatter points represent unconstrained rate estimates for each of the separate conditions.

18. The estimated transition rates for the problematic group took an extremely long time to even come close to stabilizing, requiring over 4500 iterations including attempts to manually guess where the parameters were headed. Most of the other groups required only a few hundred iterations, with the other 1 pN groups needing just over a thousand iterations. The likelihood was extremely insensitive to the faster of the two backward transition rates, which was found to vary over nearly three orders of magnitude while the log-likelihood changed by less than 3.0 (although the two forward rates also changed, though to a much lesser degree, to accommodate the average velocity). The final rates for this model, which do not appear on the earlier graphs because of these complications, are

$$\cdots \rightleftharpoons \; 1_l \; \underset{v_2 = 1 \times 10^{-11}}{\overset{u_1 = 139.9}{\rightleftharpoons}} \; 2_l \; \underset{v_1 = 1003.8}{\overset{u_2 = 2371.2}{\rightleftharpoons}} \; 1_{l+1} \; \rightleftharpoons \cdots . \tag{3.29}$$

These rapid forward and backward transitions are similar to a rapid equilibrium, and compatible with the likelihood's minimal improvement over the one-state model.

3.2.3 Model Selection

To consider the question of model selection, the EM algorithm was repeated on the two-state simulated data for several different models. To provide a reference point, the first model implemented for filtering was a one-state model. The transition rates remained unconstrained, so that each experimental condition was modeled independently of the data runs coming from other conditions. Subsequent models included a two-state model with unconstrained rate parameters, a two-state model with rate parameters constrained to the hypothesized [ATP]-dependence and detailed balance load-dependence as described in Chapter 2, and a three-state model with unconstrained rate parameters. The resulting log-likelihoods for fitting these models with the EM algorithm are summarized in Table 3.4, which we now discuss in more depth.

N	Rate Model	$\log L$
1	unconstr.	-1524061.9
2	unconstr.	-1515855.5
2	det. balance	-1515868.3
3	unconstr.	-1515852.3

TABLE 3.4. Log-likelihood results for the two-state simulated data sets.

Typically, models with greater complexity will be able to fit the data more closely. In fact, if a simpler model is a special case of a complex model, then the more complex model is guaranteed to do at least as well, with the worst case scenario being to simply use the simpler model. However, to describe the more complex model, one requires a larger number of descriptive parameters. Consequently, one needs a criterion to determine when the addition of complexity can be justified by a significant increase in the quality with which the model fits the data. When two models are nested, in which case one model is a sub-case of the other model but with a constraint on the parameters, then the difference in the log-likelihoods provides a natural measure for determining whether the more general model (with unconstrained parameters) provides a significantly better description of the data than the nested model. This is codified in the χ^2 test of the likelihood ratio. For models which are not nested, particularly when they have a different number of states, the χ^2 test does not technically apply. Assuming that the maximum likelihood estimator for the extended hidden Markov model is asymptotically normal, which has been proved for the standard hidden Markov model cases when the model is recurrent (Bickel et al., 1998; Jensen and Petersen, 1999; Douc and Matias, 2000) but has not yet been proved for the current case where the model is transient but periodic, the test statistic

$$\chi^2 = 2(\log L(\widehat{\theta}_1; \vec{y}) - \log L(\widehat{\theta}_0; \vec{y})) \tag{3.30}$$

will asymptotically have a χ^2 distribution with n degrees of freedom, where $\widehat{\theta}_0$ represents the maximum likelihood estimator for the nested model and $\widehat{\theta}_1$ represents the maximum likelihood estimator for the generalized model and n is the difference in

the number of independent parameters required for the two models.

First, consider the difference between the models with one and two states and unconstrained rate parameters. Adding a second state adds a forward and backward rate and a relative position for a substep for each of the nine experimental conditions, an increase of 27 parameters. Because models with different numbers of states are not nested models, using a χ^2-test to determine the significance of the likelihood ratio is not technically appropriate, although it may give a general sense of significance. The change in log-likelihood between the one- and two-state models, corresponding to the test statistic $\chi^2 = 2 \times 8206.4$, is several orders of magnitude greater than the reference value of $\chi^2_{27} = 63.16447$, which would be the 99.99 percent confidence cut-off for nested models. Most of this likelihood increase comes from the 5.5 pN data sets, although most of the individual groups improve the log-likelihood by at least 100. Consequently, there is significant evidence that the two state model is a better choice than the one state model. Next, we consider whether it is advantageous to add a third state to model the data. Introducing the new state requires 27 additional parameters to account for the three extra parameters for each experimental group. However, in this case, the increase in the log-likelihood is not nearly so dramatic, with the likelihood only improving by a few points, so that we would not reject the two state model for the three state model.

We remark that the addition of a third state and the corresponding parameters was computationally a serious challenge, as both the time for each iteration and the number of iterations required grew dramatically. As such, we were unable to try a variety of starting points to get a true optimal parameter set for this model. Nevertheless, the fact that the three-state models that were attempted did not significantly increase the likelihood suggests that other three-state models would similarly be moderately close in likelihood. In each instance that a second state was added to a one-state model, the likelihood increased immediately and dramatically. But when the third state was added, the likelihood only barely surpassed the original likelihood.

In addition, we note that lattice parameters have a significant impact in the likelihood values. At one point in the analysis, the apparent final likelihood of the three-state model was worse than the likelihood for two states. By comparing the likelihoods of individual files, we identified a single file that accounted for the entire difference. After we performed a more in-depth search for the lattice parameters for this file, the order of the two models reversed back to what one should expect.

We can also compare the two state model with rates parametrized by detailed balance and [ATP]-dependence to the two-state model with unconstrained rates discussed above. In this case, the detailed balance model represents a constrained model within the more general unconstrained model, so that a likelihood-ratio test using the χ^2 statistic will be appropriate. The null hypothesis is that detailed balance is satisfied, while the alternative is that it is not satisfied. The test statistic for this hypothesis test is $\chi^2 = 2 \times 12.8$. To model the detailed balance model, we have two rates and two load factors for each state, with a single constraint on the load factors that they add to 1. Thus, the detailed balance model has 7 independent parameters. The unconstrained model needs four rates for each of the nine experimental groups, leading to 36 independent parameters. Thus, the χ^2 test will have 29 degrees of freedom. The P-value of the hypothesis test is 0.647, so that we do not reject the null hypothesis that the constrained model with detailed balance and [ATP]-dependence is sufficient to describe the data. Of course, since the data were generated from this class of models, this is what one would hope results from the analysis.

3.2.4 Velocity and Randomness

Another interesting comparison between the different models is how well they estimate the velocity v and randomness r of the experimental data. Recall from Chapter 1 that these quantities characterize the mean and variance of the random time required to complete a full stepping cycle. That is, if T is the random time for a complete cycle

with mean τ and variance ν^2, then the expected mean velocity is $v = d/\tau$ and the randomness is the squared coefficient of variation $r = \nu^2/\tau^2$. The rate models completely characterize the Markov process as it moves through the intermediate states to complete a cycle, so that τ and ν^2 can be computed directly from the estimated rate parametrizations. There are a variety of ways to compute these moments based on the rate parameters (Schnitzer and Block, 1995; Fisher and Kolomeisky, 1999b). However, we found it easier to implement the calculations using two simple linear systems motivated by the Markov property.

Suppose that the initial mechanochemical state is $C(0) = c$. The total cycle time can be decomposed into the time it takes to arrive in state $(1, 1)$ plus the time it takes to go from $(1, 1)$ to $(1, c)$. By the strong Markov property, these times are independent. In addition, because of the periodicity of the lattice, the time required to go from $(1, 1)$ to $(1, c)$ has the same distribution as the time to go from $(0, 1)$ to $(0, c)$. Consequently, the time to complete a forward cycle starting in an arbitrary state c has the same distribution as the time required to start in state $(0, 1)$ and end in state $(1, 1)$. In other words, the cycle time is independent of the initial state. We define the random time S as the time when we arrive at $(1, 1)$,

$$S = \inf\{t > 0 : Z(t) = (1, 1)\}. \tag{3.31}$$

Let $F_S^{(c)}$ be the conditional distribution of S given $C(0) = c$. The distribution of T, the time to complete a cycle, or equivalently to start at $(0, 1)$ and end at $(1, 1)$, will then be $F_S^{(1)}$. So, consider the distribution of S with an initial state $C(0) = c$. Let S_1 be the departure time,

$$S_1 = \inf\{t > 0 : Z(t) \neq (0, c)\}, \tag{3.32}$$

and let S_2 be the completion time

$$S_2 = \inf\{t > 0 : Z(S_1 + t) = (1, 1)\}. \tag{3.33}$$

Again applying the strong Markov property, S_1 and S_2 are independent. Furthermore, S_1 has an exponential distribution with rate parameter λ_c, and S_2 will have the distribution $F_S^{(c+1)}$ with probability p_c and the distribution $F_S^{(c-1)}$ with probability q_c. In the case of $c = N$, the distribution $F_S^{(N+1)}$ corresponds to the random variable always taking the value zero. The other case requiring special consideration is when $c = 1$. Taking a backward step puts the new state at $(-1, N)$. In order to complete the cycle, one must first return to $(0, 1)$ and then to $(1, 1)$. This random time will have a distribution that is equivalent to the sum of two independent random variables, one with distribution $F_S^{(N)}$ and the other with distribution $F_S^{(1)}$.

The linear system mentioned earlier establishes a linear relations for the mean and variance of the distributions $F_S^{(c)}$,

$$\tau_c = E[S|C(0) = c], \tag{3.34}$$

$$\nu_c^2 = E[(S - \tau_c)^2|C(0) = c]. \tag{3.35}$$

By linearity of conditional expectation, we obtain a linear relation for each c

$$\tau_c = E[S_1 + S_2|C(0) = c] = \frac{1}{\lambda_c} + p_c\tau_{c+1} + q_c\tau_{c-1}, \tag{3.36}$$

which can be rewritten as

$$-q_c\tau_{c-1} + \tau_c - p_c\tau_{c+1} = \frac{1}{\lambda_c}. \tag{3.37}$$

The boundary case of $c = 1$ leads to

$$-q_1(\tau_1 + \tau_N) + \tau_1 - p_1\tau_2 = \frac{1}{\lambda_1}. \tag{3.38}$$

By independence, we also obtain a linear relation for the variances,

$$\nu_c^2 = \mathrm{Var}[S_1 + S_2|C(0) = c] = \frac{1}{\lambda_c^2} + \mathrm{Var}[S_2|C(0) = c]. \tag{3.39}$$

To compute the second variance, we apply a simple case of the conditional variance formula to obtain

$$\mathrm{Var}[S_2|C(0) = c] = p_c\nu_{c+1}^2 + q_c\nu_{c-1}^2 + p_c\tau_{c+1}^2 + q_c\tau_{c-1}^2 - (p_c\tau_{c+1} + q_c\tau_{c-1})^2. \tag{3.40}$$

Combining the variances, we obtain a linear relation with the same form as for the means,

$$-q_c\nu_{c-1}^2 + \nu_c^2 - p_c\nu_{c+1}^2 = \frac{1}{\lambda_c^2} + p_c\tau_{c+1}^2 + q_c\tau_{c-1}^2 - (p_c\tau_{c+1} + q_c\tau_{c-1})^2, \tag{3.41}$$

with the special boundary case for $c = 1$ of

$$-q_1(\nu_1^2 + \nu_N^2) + \nu_1 - p_1\nu_2 = \frac{1}{\lambda_1^2} + p_1\tau_2^2 + q_1(\tau_1 + \tau_N)^2 - (p_1\tau_2 + q_1(\tau_1 + \tau_N))^2. \tag{3.42}$$

Solving for the mean times τ_c using Equations 3.37-3.38, these values can be substituted into the right hand side of Equations 3.41-3.42 to solve for the variances ν_c^2. Because the linear systems are the same, even these reasonably small matrices need be inverted only once.

Tables 3.5 and 3.6 summarize the estimates for the velocity and randomness for the various models describing the two-state data. In addition, the tables show the theoretical velocity and randomness for the original model from which the data were derived. They also display the velocity and randomness that were explicitly calculated based on the observed positions, using the methods given in chapter 1, using Equations 1.3 and 1.8. For each of the estimated values, the maximum likelihood parameters found through the EM procedure were used to compute the velocity and randomness. The velocity is well-matched by all of the models. However, the randomness is not well-described by the one-state model. The randomness for the true model takes values less than 1, but one-state models are constrained to have a randomness of at least 1. Also, although the simulation-based observed randomness varied from the theoretical values, based on a limited amount of data, the randomness based on the fitted models remained very consistent with the true values.

3.2.5 Residual Analysis

Another possible tool to help evaluate the appropriateness of a given model is residual analysis through the use of the Viterbi algorithm. Recall that the Viterbi algorithm

Experimental Group	2 state predict	simul. observed	1 state unconstr.	2 state unconstr.	2 state det. bal.	3 state unconstr.
1 pN, 10 μM	92.2	93.2	94.2	93.6	88.1	93.4
3.5 pN, 10 μM	46.4	45.7	44.2	44.2	44.9	44.2
5.5 pN, 10 μM	11.5	13.2	12.7	12.6	11.4	12.6
1 pN, 100 μM	435.6	436.3	435.6	434.5	441.6	433.1
3.5 pN, 100 μM	283.5	262.9	269.0	268.6	277.4	267.4
5.5 pN, 100 μM	103.3	102.3	102.7	103.1	102.0	102.9
1 pN, 2 mM	790.8	780.1	777.2	778.6	766.0	775.7
3.5 pN, 2 mM	635.6	617.4	605.4	607.0	630.4	604.2
5.5 pN, 2 mM	367.4	361.1	363.6	365.0	365.2	363.6

TABLE 3.5. Estimated velocities (nm/s) for each experimental group according to 1, 2, and 3 state models, as they compare to the predicted and observed velocities for simulated, two-state data.

Experimental Group	2 state predict	simul. observed	1 state unconstr.	2 state unconstr.	2 state det. bal.	3 state unconstr.
1 pN, 10 μM	0.814	0.691	1.000	0.827	0.823	0.826
3.5 pN, 10 μM	0.905	0.923	1.066	0.908	0.910	0.903
5.5 pN, 10 μM	3.019	3.342	10.007	2.895	3.028	2.893
1 pN, 100 μM	0.540	0.501	1.000	0.547	0.555	0.529
3.5 pN, 100 μM	0.590	0.609	1.056	0.625	0.598	0.616
5.5 pN, 100 μM	1.485	1.262	4.207	1.505	1.489	1.499
1 pN, 2 mM	0.931	0.690	1.000	0.924	0.933	0.903
3.5 pN, 2 mM	0.896	0.782	1.009	0.915	0.895	0.911
5.5 pN, 2 mM	1.089	0.953	1.276	1.088	1.090	1.086

TABLE 3.6. Randomness computed for maximum likelihood models with 1, 2, and 3 states, as they compare to the predicted and observed randomness for simulated, two-state data.

computes the maximum likelihood path for the hidden Markov process. Using the maximum likelihood parameters for each proposed model for the Viterbi algorithm, the estimated equilibrium position of the bead can be calculated. Subtracting the estimated equilibrium position from the observed bead position, we compute an estimated sequence for the residuals. If, instead of the Viterbi estimates, the true equilibriums position were used at each time step, then the residual sequence would behave as a simple autoregressive process. Thus, as a qualitative exploration of model, we considered how the behavior of the residuals arising from the estimated states relate to a simple autoregressive process.

Given the residual sequence, which we denote $\{\varepsilon_k\}$, we look at the autocorrelations of the residuals through the use of the autocorrelation function (ACF) and the partial autocorrelation function (PACF). Theoretically, if the residuals $\{\varepsilon_k\}$ were a simple autoregressive process with an autocorrelation coefficient ρ, then the ACF would be a geometrically decaying sequence,

$$\mathrm{ACF}(k) = \rho^k, \quad k = 0, 1, 2, \ldots, \tag{3.43}$$

while the PACF, which is defined for $k \geq 1$, will have

$$\mathrm{PACF}(k) = \begin{cases} \rho, & k = 1, \\ 0, & k > 1. \end{cases} \tag{3.44}$$

(See a text on time series analysis, such as Brockwell and Davis (1996)). We used the standard `ts` package of the R statistical programming language for computing and graphing the ACF and PACF (Ihaka and Gentleman, 1996). In addition to plotting the estimated autocorrelation and partial autocorrelation, this package also plots a cut-off band, where values inside the band would fail to be significantly different from zero (0) with 95 percent confidence. Thus, if one observes the PACF and sees more than one or two indexes $k > 1$ with values outside of this band, then the model is likely not sufficiently characterizing the data.

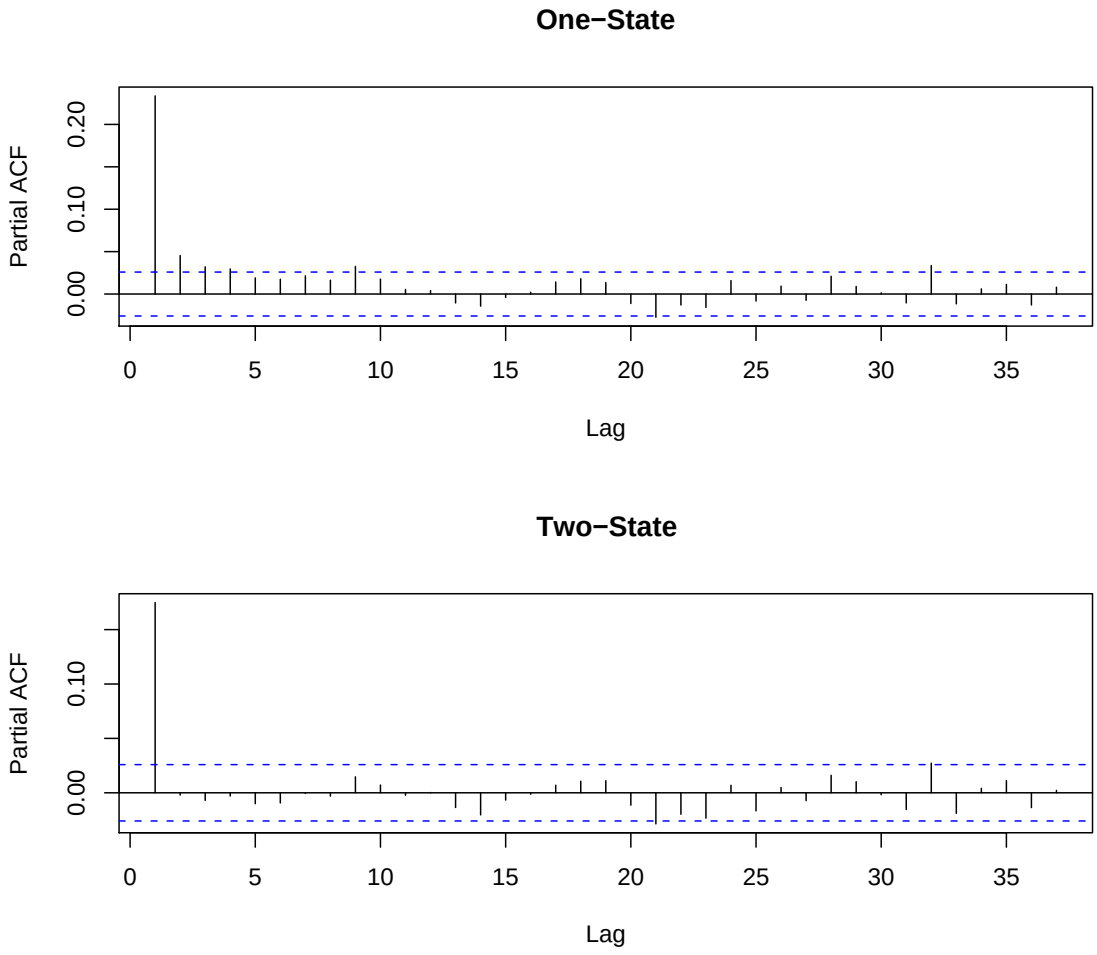


FIGURE 3.9. The partial autocorrelation function for the Viterbi residuals of a single data set in the 5.5 pN 100 μM group generated from a two-state model and estimated using one- and two-state models. The one-state model (left) shows additional structure in the residuals, while the two-state model (right) does not. The horizontal dashed lines are approximate 95 percent confidence bounds for zero autocorrelations.

Consider the effects of one- and two-state models on a particular data set from the simulated set with a force of 5.5 pN and an ATP concentration of 100 μM. Figure 3.9 illustrates the PACF generated from estimated Viterbi residuals for the MLE estimates described earlier. Note that for this data file, the one-state model shows a significant number of partial autocorrelations outside of the cut-off band. This suggests that the residuals coming from a single-state model still show significant structure beyond a simple autocorrelation. However, examining the PACF for the same data set but using the two-state model, nearly all of the partial autocorrelations now lie within the cut-off band. Thus, the addition of the second state eliminated the structure that the PACF was identifying.

By examining another simulated run, we see that the PACF does not always identify whether the model is sufficient. In this case, we consider a data set with a force of 3.5 pN and ATP concentration of 100 μM. The PACF for both the one- and two-state models are illustrated in Figure 3.10. In both cases, the Viterbi residuals appear to have simple autoregressive structure, with nearly all partial autocorrelations inside of the cut-off region. One of the possible factors that will impact the effectiveness of the PACF in identifying extra structure is the scale of the noise. In the 5.5 pN data set, the noise had a scale of only 3 nm, while the scale was 4.5 nm for the 3.5 pN data set. By reducing the scale of noise, observations will be more tightly clustered about the equilibrium positions of the underlying model. This extra structure may be interpreted as extra autocorrelation in the PACF. But for greater noise, the observations are only loosely clustered about the equilibrium positions, and the PACF may fail to interpret the remaining structure as autocorrelation. If we consider the change in log-likelihood for these individual data sets as the second state is added, then the strength of the structure will determine how much improvement is seen. The 3.5 pN data set increased the log-likelihood by 10.753, while the 5.5 pN data set increased by 110.915 although it was only about three times longer in length.

Using the Viterbi algorithm to compute the estimated equilibrium position of the

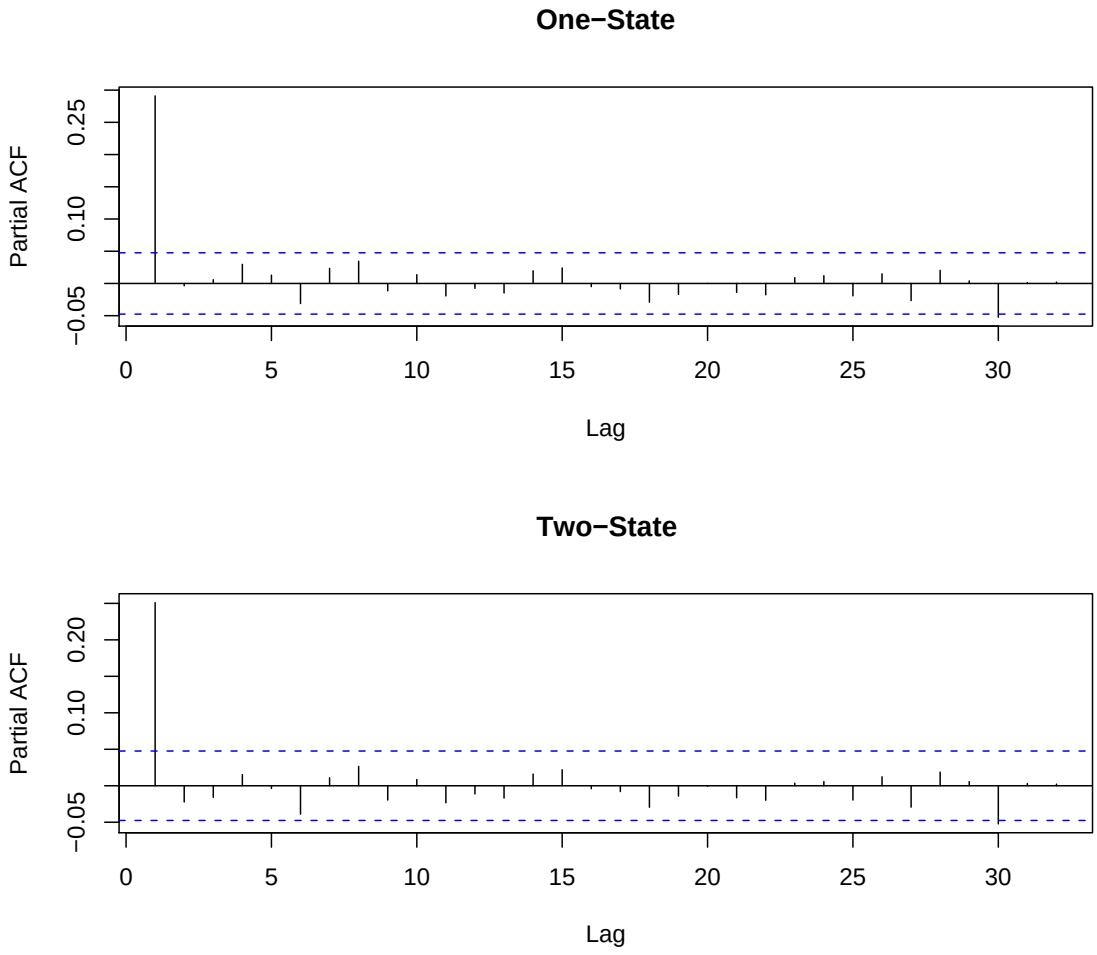


FIGURE 3.10. The partial autocorrelation function for the Viterbi residuals of a single data set in the 3.5 pN 100 μM group generated from a two-state model and estimated using one- and two-state models. The horizontal dashed lines are approximate 95 percent confidence bounds for zero autocorrelations. Neither of the two models show significant structure beyond the simple autocorrelation for the noise.

bead, and hence the residuals, does potentially introduce additional correlations in the measurements. As the algorithm selects the sequence that maximizes the likelihood rather than according to the properties of the residuals, the resulting residual sequence will include artifacts arising from misspecified states. Fast events, such as a short-lived forward and backward substep, will tend to be overlooked as the likelihood pays less penalty for a slightly larger discrepancy in a residual arising from an erroneous state assignment than it would pay for two relatively rare transitions. Nevertheless, the apparent success of this procedure for at least some of the data suggests that it might be a helpful tool to consider.

3.3 Four-State Data

In addition to the two-state data discussed above, we also attempted to analyze a collection of simulations based on the four-state model of Fisher and Kolomeisky (Fisher and Kolomeisky, 2001). For these simulations, we explicitly model the continuous time model, so that we have knowledge of all the transitions. This extra information allows us to include the memory term describing the relaxation of the bead position to the equilibrium position. Because the hidden Markov model does not account for this aspect of the true behavior, we used this data set to explore the impact of the existence of the memory term.

The most severe challenge to accomplish this analysis was the time required for the computers to perform the EM algorithm, particularly for the models with three and four states, where we managed only three to four iterations per hour. The maximization step, in particular, took quite a bit longer for the larger models due to the increase in number of rate parameters needing to be estimated and the corresponding cost to differentiate the transition probabilities. When it was clear that we would not be able to complete the same thorough model comparison as was performed for the two-state simulations, we directed the majority of our resources to completing the

two-state analysis and to do as much analysis of the real data sets as possible.

By starting with the true four-state model and parameters, we performed the EM algorithm for 200 iterations. While this is far from enough to complete any analysis—applying a three-state model to filter the two-state data required approximately 2000 iterations before parameters really settled down—it did reveal that the true observation parameters, particularly the position of intermediate substeps, were not stable. Figure 3.11 illustrates the evolution of these positions over these 200 iterations (solid lines). In both the 1 pN and the 3.5 pN groups, at least one of the intermediate states seems to be lost in the data. It should also be remarked that when the three offsets maintain a common relative separation while still moving on the graph, that this can also be interpreted as the first position ϵ_1 moving in the opposite direction relative to the other three, such as for the group with 5.5 pN force and 2 mM [ATP].

In order to determine how much the inclusion of the memory term controls this unfortunate behavior, we generated the identical data set except that the memory term was not included. In particular, we used the same seed for the pseudo-random number generator. We verified that the data were the same by comparing the hidden states in a number of different data files and observing that they were identical. We repeated the EM algorithm for another 200 iterations with the same initial conditions. The behavior was essentially identical to the results with the memory term, with the evolution of substep positions showing similar trends. The most prominent effect in several of the data sets was an overall offset between the substep positions with the memory present relative to the positions with the memory absent. The effect is most visible for the groups with force of 5.5 pN. The presence of the memory term causes the bead position not to reflect the true equilibrium position near transition events, mixing the information about neighboring states.

Because the overall behavior of the substep positions did not depend significantly on the presence of the memory effect, it appears that the scale of noise and autocorrelation has a much more important role in identifying the positions of the intermediate

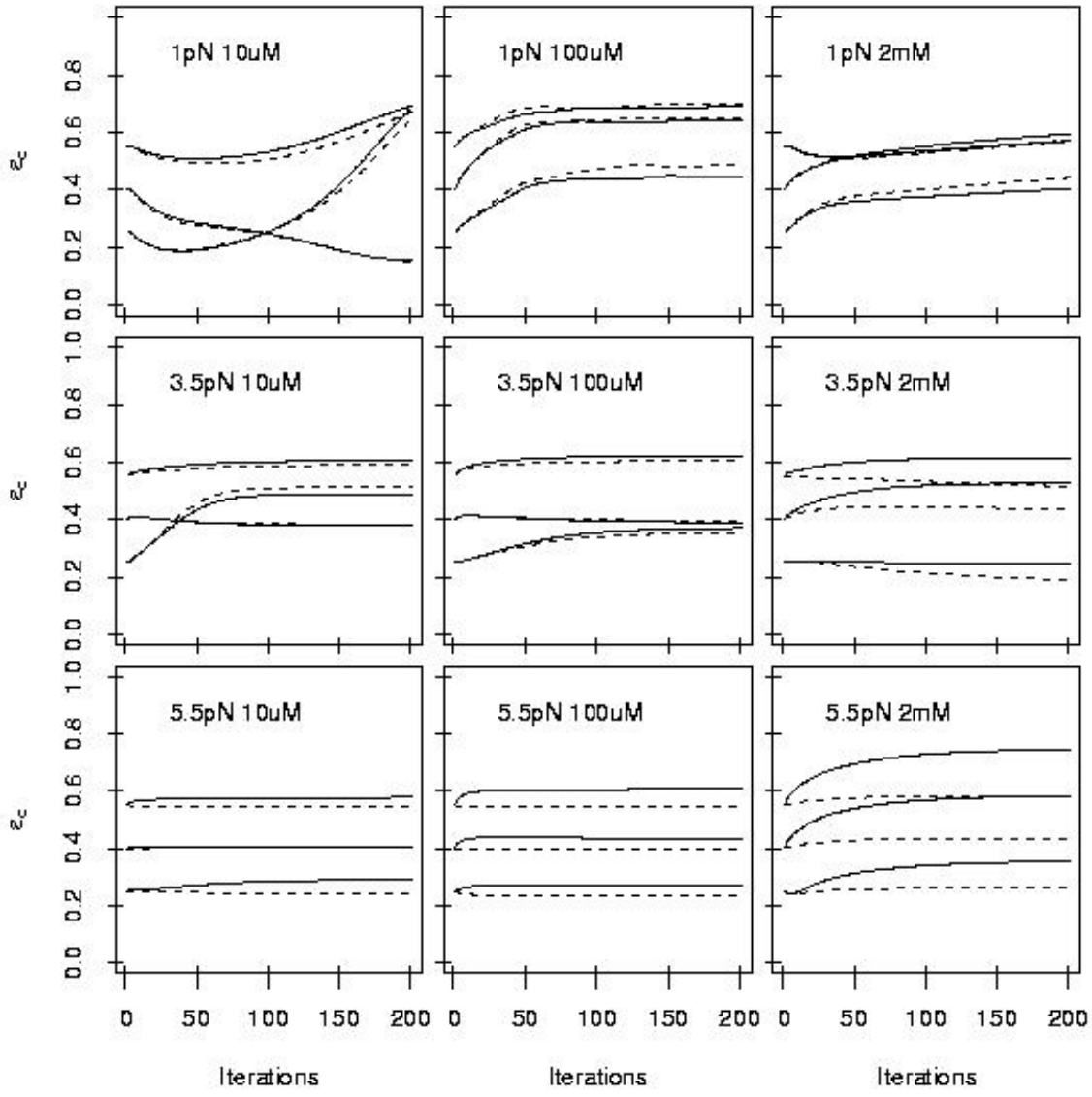


FIGURE 3.11. Evolution of the offsets ϵ_c from the true values, $\epsilon_2 = 0.25$, $\epsilon_3 = 0.40$, and $\epsilon_4 = 0.55$, for a four-state model for 200 iterations of the EM algorithm. Solid lines are for data which include the memory term; dashed lines are for data which have the memory term eliminated.

states. The data sets were generated with progressively smaller sizes of noise and autocorrelation, corresponding to the behavior of the original physical system where a stiffer trap exerts a higher load and creates smaller autocorrelation times. The 1 pN data sets were generated with a noise scale of approximately 7 nm, the 3.5 pN data sets with 4.5 nm, and the 5.5 pN data sets with 3 nm. Similarly the autocorrelation coefficients dropped as the load was increased, with values of 0.3, 0.25, and 0.18 for the respective force levels, corresponding to respective autocorrelation times of 0.416 ms, 0.361 ms, and 0.292 ms. The sampling time used for the data was 0.5 ms. In addition, the model caused the data with higher loads to spend more time in the underlying states, due to the suppression of transitions due to load-dependence. Thus, in all senses, the high-load experiments contained much more information about the hidden model. The confounding of the autocorrelation with the scale of noise makes it impossible from this experiment to determine whether the noise or autocorrelation plays the more important role. The experiments do, however, suggest that the data need to be sampled more frequently.

Chapter 4
Kinesin Data Results

We finally discuss the results of applying the hidden Markov model filtering procedures on the single-molecule kinesin data arising from the experiments of Visscher et al. (1999). The data sets we used come from the same nine experimental conditions selected for the simulated data, shown in Table 3.1, and the existence of the actual data motivated the choice of simulated conditions. Because the speed at which kinesin moves depends so strongly on the experimental conditions, the typical number of data points in each file varied widely over the different experimental conditions. With the original intent of performing a global fit to the detailed balance model, we want each experimental condition to provide a comparable weight to the overall likelihood. With the slow conditions of high force and low ATP concentrations having so many data points per file, we randomly selected data sets to exclude from these conditions so that the overall likelihood does not come overwhelmingly from these conditions.

In the process of performing the EM algorithm using the hidden Markov model methods developed for this project, we realized that the data of the experiment were not quite appropriate for the attempted analysis. Consequently, we do not believe that the particular model parameters obtained from this analysis have physical relevance. The results on the simulations initially suggest that a first problem is insufficient data sampling, particularly for models with more than two states. This chapter will discuss the results of our analysis on the kinesin with simple two-state models, and the additional issues that arise. In the conclusion, Chapter 5, we suggest possible improvements to the experimental design that will be better compatible with hidden Markov model filtering.

4.1 Step-Size Variation

Perhaps the most serious challenge in obtaining reliable parameter estimates for the kinesin data was excessive variability in the estimates of d, the lattice step-size. Recall that the hidden Markov model estimates of the summary statistics, on which all of the parameters are based, are computed using the observations weighted according to their distance from the assumed lattice positions. If the model lattice is reasonably close to the true lattice, then the EM algorithm will successfully converge to consistent estimates of the true model parameters. But if the lattice is not valid, then these summary statistics will represent mixtures of different states of the true model. Consequently, if the lattice is improperly specified, then all other parameter estimates become unreliable.

Physically, the lattice structure of the microtubule is known to have a step-size of 8.2 nm. Because the recorded observations come from projections of kinesin movement in two dimensions into a single dimension, effective lattice step-sizes should be slightly shorter than 8.2 nm. For example, an alignment error of 10 degrees would correspond to a reduced length of 98.5 percent of the original, 8.075 nm. Lattice step-sizes larger than 8.2 nm would not be expected, except perhaps for some slight error inherent in statistical estimation. Additional errors might also come from instrument calibration.

Our earlier results for simulated data indicated that the lattice structure naturally creates a multi-modal likelihood surface. Recall that changing the step-size enough to add or remove a site on the lattice corresponded approximately with the positions of the local maxima in the d-direction. The kinesin experiments typically involve runs with a span of approximately 300 nm. This corresponds to a range of at least 36 different lattice sites. Thus, an error in the step-size larger than 0.22 nm ($\frac{1}{36}$ of 8.2 nm) would lead to a mismatch in the lattices, and consequently would invalidate other parameter estimates.

When the EM algorithm was applied to the kinesin data, the estimated step-sizes

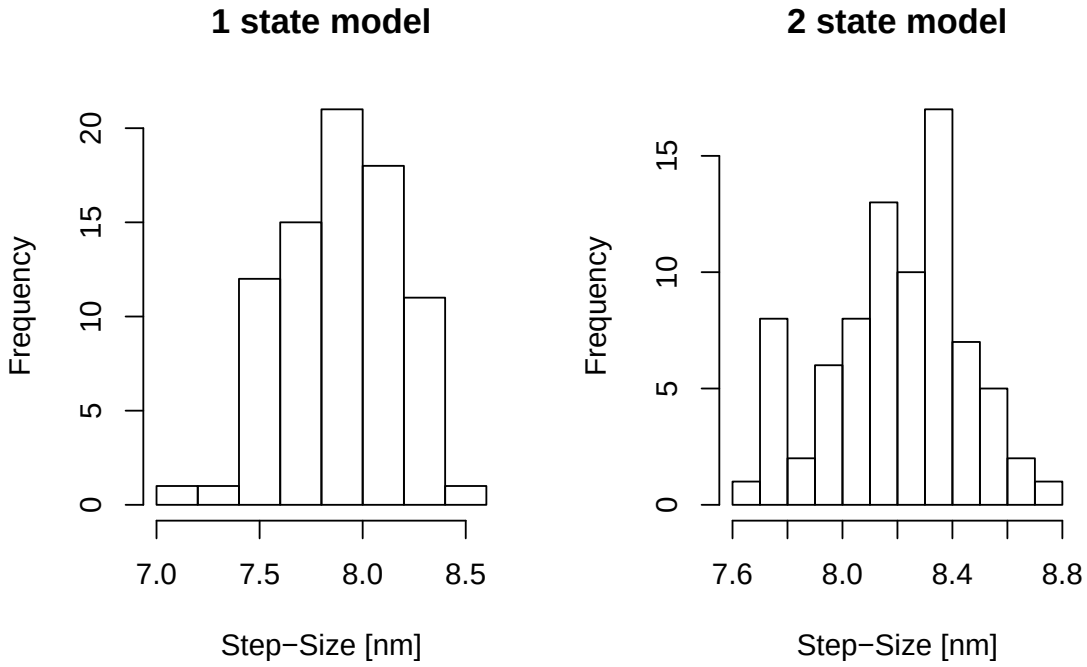


FIGURE 4.1. Histograms of the estimated step-size d for the 80 data sets comprising the experimental conditions of 3.5 pN and 100 μM [ATP] for both one- and two-state models.

had much larger variability than allowed for consistent parameter estimates. For example, consider the experimental condition with 100 μM concentration of ATP and 3.5 pN load. This collection had 80 different data files. The histogram of step-sizes for a one-state model, given in Figure 4.1, shows that the estimates are broadly centered about a mean of 7.89 nm and standard deviation of 0.277 nm. The values range from a low of 7.196 nm to a high of 8.404 nm. The second half of the figure gives the histogram of step-sizes for a two-state model. This set has a mean of 8.20 nm and a standard deviation of 0.246 nm, with a range from 7.69 to 8.80 nm. Similar results appeared for other experimental conditions (not shown). In both cases, the range of step-sizes is much too wide. Particularly, the estimated step-sizes greater than 8.2 nm cast serious doubt to their validity, as do the extremely low estimates.

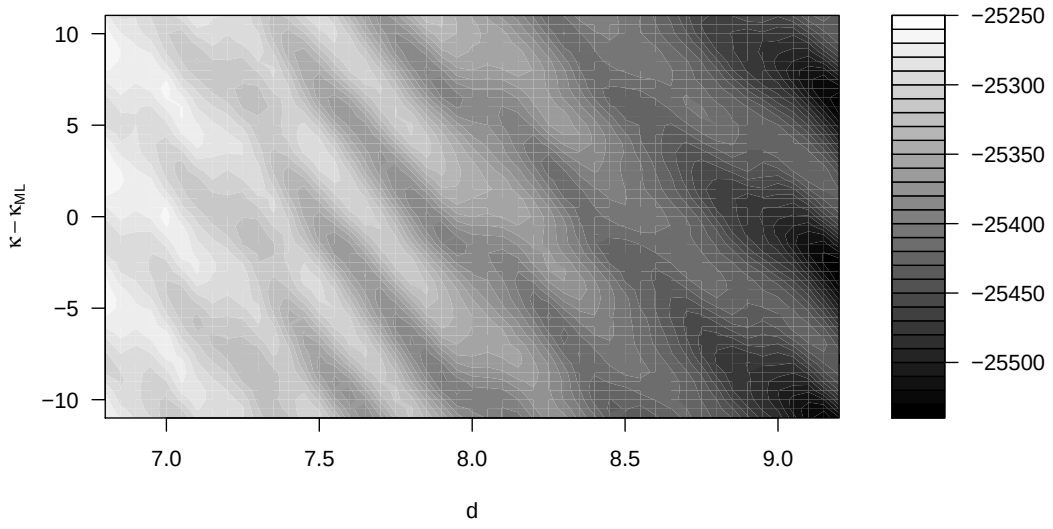
However, even a marginally low step-size, such as 7.8 nm, potentially creates a model lattice sufficiently mismatched to the true lattice to cause problems with parameter estimates.

4.2 Variability within Conditions

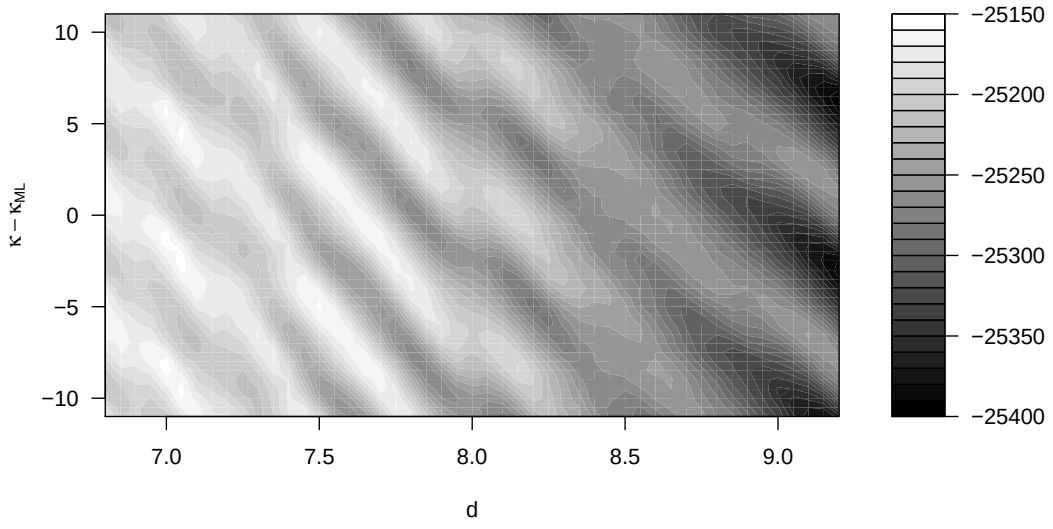
Concerned that the estimated step-sizes were simply a result of the multimodal nature of the likelihood surface, we also explored the likelihood surface restricted to the two dimensions of step-size d and starting offset κ. The expected behavior, established by the previous simulations, would be for a band of maximum peaks near $d = 8.1$ nm (because of the apparent shortening) with additional bands to either side that decrease in likelihood. Holding all other parameters fixed at the final estimated values, we varied the step-size and starting offsets for a few different data sets and examined their likelihood surfaces. If the problem was that the algorithm accidentally selected the wrong peak, in spite of our efforts to avoid this through occasionally using a coarse grid for the lattice parameters, then a finer grid should reveal an appropriate peak near the expected position.

Instead, the likelihood surfaces indicated that something else was occurring. The first contour of Figure 4.2 shows the likelihood surface for a data set which had a step-size estimate of 7.00 nm. Rather than having a likelihood centered around a value near 8.1 nm, the surface displays a strong preference for short step-sizes. Similarly, the first contour of Figure 4.3 shows the likelihood surface for a second data set which led to an estimated step-size of 9.13 nm. This surface shows a preference toward large step-sizes. Evidently, some information in the data causes step-size estimates for particular data sets to move toward extreme values.

One possible cause for the observed behavior is that the underlying Markov model varies between different observations within a given experimental condition. Each experimental condition includes many different kinesin runs, coming from several

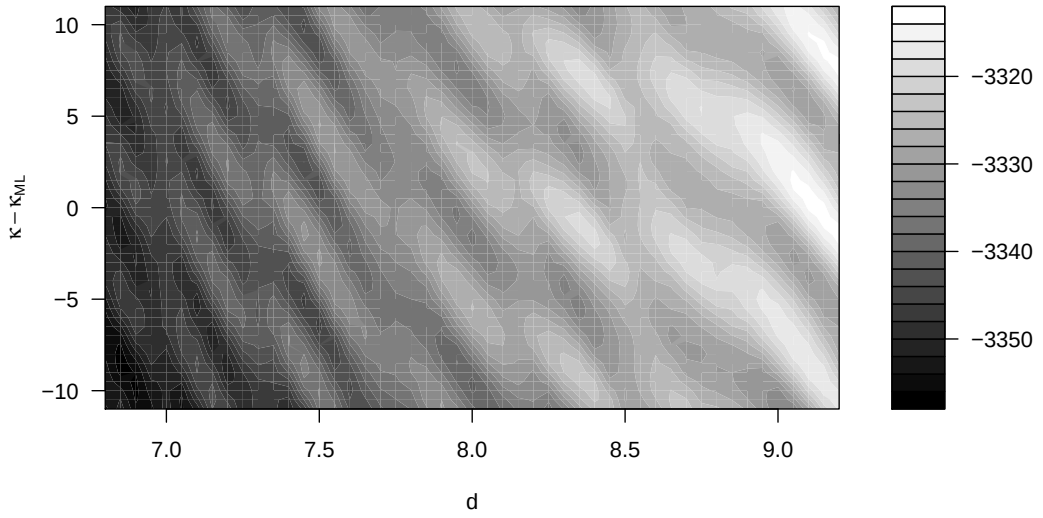


(a) Group Rates

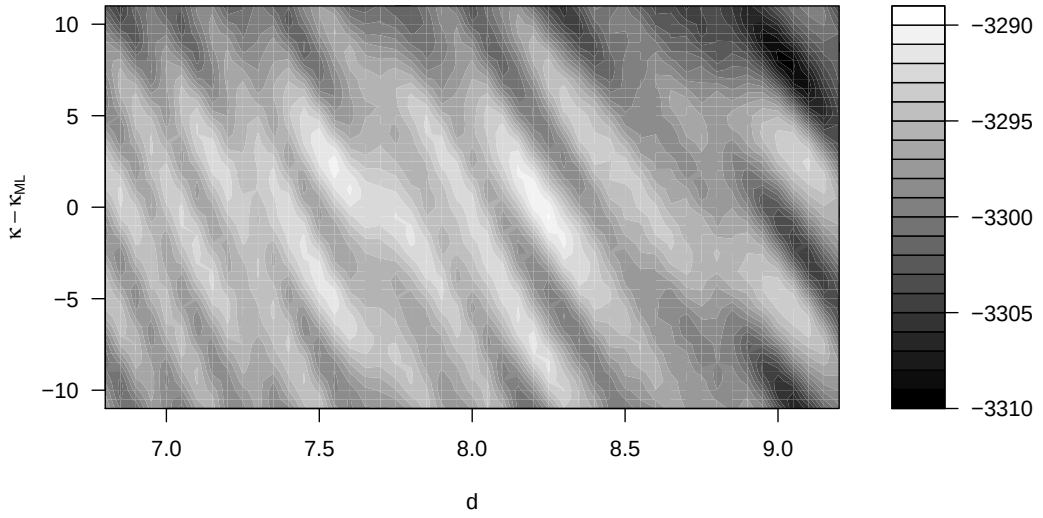


(b) Individual Rates

FIGURE 4.2. Log-likelihood profile for a data set in the 1 pN 10 μM group which estimated the step-size near 7 nm. The κ-axis is centered around the original MLE κ_{ML}. Subfigure (a) gives the surface when using transition rates derived from all of the data sets in the group, while subfigure (b) gives the surface when using transition rates optimized for the individual data set.



(a) Group Rates



(b) Individual Rates

FIGURE 4.3. Log-likelihood profile for a data set in the 1 pN 100 μM group which estimated the step-size near 9 nm. The κ-axis is centered around the original MLE κ_{ML}. Subfigure (a) gives the surface when using transition rates derived from all of the data sets in the group, while subfigure (b) gives the surface when using transition rates optimized for the individual data set.

different kinesin-bead systems. Additionally, these different runs may come from different flow-cell preparations or even from different days. The amount of variability in stepping behavior due to differences between individual proteins, as well as different flow-cells, remains unclear. If a particular run has stepping statistics different from the experimental group, then it may be beneficial to the likelihood for the model artificially to increase or decrease the number of estimated steps to remain compatible with the average transition rates.

In order to see if internal variability might influence the likelihood profile, we isolated the data sets being examined and repeated the EM algorithm on each file individually. To encourage the estimated model to be consistent with the physical step-size of 8.2 nm, we ran the algorithm with the step-size constrained at 8.1 nm for 200 iterations, allowing other parameters to accommodate this prior information. Then we relaxed the constraint and repeated the EM algorithm, allowing the step-size to converge to a local maximum. This provided each file with an independent estimate of rate parameters for the underlying Markov model. Finally, we recreated the likelihood surface over the original range of step-sizes to see what changed. These surfaces based on individualized parameters are shown in the second half of the previously mentioned figures, Figures 4.2 and 4.3.

The resulting surfaces showed some improvement, but did not necessarily position the overall maximum likelihood near the expected 8.1 nm. First consider the likelihood surfaces of Figure 4.2, which started with a bias toward small step-sizes. This data set shows some progress, although the bias is still strongly apparent. We note that the log-likelihood for this individual data set improved by about 100. With four parameters to describe the transition rates and three parameters to describe the autocorrelation ρ, noise scale σ, and intermediate substep position ϵ_2, splitting each data file out of an experimental group introduces seven new parameters. The 99.99 percent cut-off for a χ^2 test with 7 degrees of freedom is 29.878. Thus, statistical evidence indicates that the data set is significantly different from the parameters derived from

the full group. Next, consider the likelihood surfaces of Figure 4.3, where the bias was toward larger step-sizes. In this instance, the strongest band for the likelihood with individualized parameters now appears at approximately 8.2 nm. The overall likelihood increase for this data set is 23.18, with a χ^2 statistic value of $\chi^2 = 46.36$. Again, the data demonstrate a statistically significant improvement for separating the data set from the rest of the group. Finally, we remark that the amount that the likelihood changed for these two data sets should not be directly compared with each other, as the sets came from different experimental conditions.

4.3 Unexpected Noise Characteristics

Another issue that arose, at least in the high noise conditions of 1 pN load, was that the data included some unexplained noise characteristics. When analyzing the two-state model, the EM algorithm consistently selected an intermediate step position ϵ_2 which was a moderately large, negative number. The values selected in the maximum likelihood models for the 10 μM, 100 μM and 2 mM conditions are $\epsilon_2 = -0.596$, -0.567 and -0.194, respectively. Because a backward substep as part of a forward step seems physically problematic, we examined this behavior more carefully to identify the source.

Since the estimate for a given substep position will be based on a weighted average of those observations that most likely come from that particular state, we used the Viterbi algorithm to generate maximum likelihood paths of the kinesin state for the data sets in question. Figure 4.4 illustrates a sample data set from the 1 pN 100 μM group, paralleled by the estimated equilibrium bead position, offset to separate it from the original signal. Notice that the spikes in the equilibrium position correspond with corresponding extreme positions in the bead position. Examining the original data sets in the proximity of these spikes, one finds occasional sudden jumps for a single observation of 20-30 nm, after which the observations immediately return to

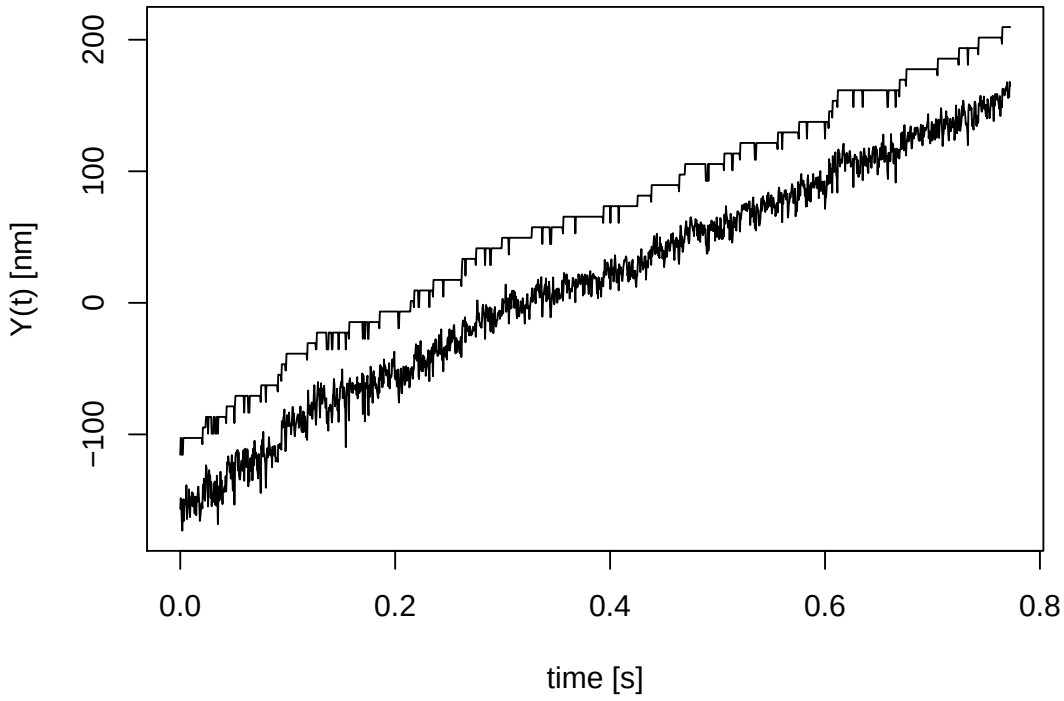


FIGURE 4.4. The estimated equilibrium position of the bead for a particular data set in the 1 pN 100 μM experimental group, offset from the original sequence of observations. The downward spikes are a modeling artifact arising from occasional position measurements that are unexplained by noise or standard transitions.

values comparable to other nearby observations.

Evidently, as the EM algorithm progresses, the model must minimize the penalty on the likelihood caused by spikes. One possibility is to treat the spike as though it were caused by a complete cycle in one direction which was immediately followed by a cycle in the opposite direction. The primary difficulty with this approach to dealing with the spikes is that it would necessarily raise all of the backward transitions, implying an increase in the rate of reverse cycles. But the majority of the data do not support frequent backward cycles, as backward steps are relatively rare. An alternative approach is for one of the states in a multi-state model to represent all of the extreme spikes in a particular direction. By making the total transition rate

sufficiently large, the time spent in the state will be kept short. This second method minimizes the likelihood penalty better than the first, as the cost for the cycle to proceed through the pike state without being observed is minimal. The most extreme spikes eventually cause the model to use one of the states to account for spikes. As the EM algorithm progresses, other observations which may be only moderately in the direction of the spike gradually contribute additional weight to the state. A side effect of this behavior is that true short-term backward steps might also be interpreted as coming from the spike state.

We are unable to determine the physical source of these spikes. The data sets with lower noise from the higher loads of 3.5 and 5.5 pN do not select for a state corresponding to spikes. In fact, they estimate an intermediate fractional substep ϵ_2 between 0.3 and 0.7. It may be that the spikes are present in the other data sets, but due to the smaller noise scale, the model attains a better likelihood by incorporating internal structure in the data other than the spikes. Alternatively, there may have been specific, yet unknown experimental cause of the spikes that was present only for some of the experiments. In any event, the unphysical nature of the fractional substep positions indicates that these data sets are inadequate to identify internal structure.

4.4 Additional Experimental Issues

In addition to the data characteristics already mentioned, a few other potential problems may also contribute to make the analysis ineffective—kinesin slippage and experimental drift—although it was unclear whether they occurred in the nine conditions considered.

Early during the project, we did perform one complete iteration of the EM algorithm on every data set available from the experiments of Visscher et al. (1999). Occasionally, extremely rare, large backward events would occur, causing a serious

penalty to the likelihood of the particular data set. By recording which data set and for which time index the event occurred, we examined the original data to observe that occasionally the data show a sudden drop in the bead position. Some of these events represented extremely strong spikes, where the bead position would subsequently return to the neighborhood of the original position. But other times, the kinesin simply made a sudden shift without recovering the loss. One possibility is that the kinesin performed a rare sequence of backward cycles. However, it seems more likely that the kinesin slipped. This would occur if the kinesin detached from the microtubule and then reattached at another site before the bead drifted away from the microtubule. The second possibility would necessarily be interpreted as backward cycles, as the present model provides no alternative pathways for backward motion. Consequently, the hidden Markov model estimation would artificially inflate the backward rates to compensate for any data with the presence of slippage.

Another issue that we observed in some experimental conditions was the possibility of drift, particularly for conditions with very slow velocity. When kinesin remained at particular lattice sites for time intervals on the scale of several seconds, the average position appeared to vary slowly in time. This drift arises, at least in part, from the slow motion of the experimental equipment, such as the microscope stage and tracking system. The nine experimental conditions selected for analysis had velocities fast enough that drift was not visibly apparent at individual lattice sites, although the possibility of drift can not be rejected. If drift were significant, then it would corrupt the estimates of the step-size d as well as the positions of the intermediate states. This, in turn, would affect the transition rates between these states.

Chapter 5
Conclusion

Hidden Markov model analysis of single-molecule assays provides a promising tool to extract information about the internal states of the chemical cycles. Although simulations indicate that the method should work for appropriate data, current experimental data for the motor protein kinesin are not quite adequate for the present implementation of the algorithm. The experiments under consideration did not provide enough data points sampling the underlying states relative to the size of noise and autocorrelation. In addition, the speed of the algorithm provided a major obstacle to pursuing more extensive analysis. Both of these issues likely need to be resolved in order to make this form of analysis helpful to understanding the mechanochemical cycle of proteins such as kinesin.

5.1 Experimental Suggestions

The results of analyzing the kinesin data suggest that future experiments must be designed for the needs of hidden Markov model analysis rather than simply applied *post hoc*. In particular, bead positions need to be recorded more frequently, so that the number of observations per underlying state can increase. Sampling rates should ideally be at least several factors larger than the fastest transition rates in the model. In addition, the autocorrelation time caused by the viscous drag of the bead through the fluid needs to be reduced so that the additional data points will be able to contribute significant new information. The possibility of variability of the underlying transition rates for different kinesin molecules suggests that experiments should be undertaken to increase the number of repeated runs for individual molecules, rather than relying on many different proteins within a particular experimental condition. In

addition, it would be helpful to extend the length of the runs, to increase the number of sites visited on the lattice. Some control or tracking of the experimental drift during the course of an experiment would help make any estimates more reliable. Preliminary analysis of bead records might also help identify irregular noise characteristics during the experiment.

5.2 Model Suggestions

In addition to needing more refined experiments, the experimental results indicated that the model might need some additional adjustments. The appearance of irregular noise suggests that if the experimental source can not be identified, then some adaptation to the model might be required. In addition, the appearance of slippage suggests that some effort might be needed to allow this type of behavior. Introducing these additional transitions would, unfortunately, increase the complexity of the model and transition rates. A simple alternative would be to use the existing hidden Markov models to identify the occurrence of these rare events, and then split the data set into different runs separated at the slip times. This has the disadvantage of reducing the number of data points available for each of the resulting sets, although this would be reduced with faster sampling rates. Finally, the current model entirely neglects the memory term which arises from the spatial transitions of kinesins to which the bead responds with a viscous delay. The model would benefit from the inclusion of the memory term, if it could be added with seriously increasing the complexity of the algorithms.

5.3 Implementation Issues

The final area that needs serious resolution deals with the computational effort to complete the EM algorithm. Two issues to address are the time to compute the derivatives of the transition probabilities and the number of iterations required to

obtain convergence. Rather than compute the derivatives numerically by repeatedly computing the transition probabilities, one might find a method to compute them directly. Some work has been done in this direction (Michalek and Timmer, 1999), but it is presently unclear how to implement this effectively for the number of states introduced with an underlying lattice. The second challenge, reducing the number of iterations required for convergence, might also be solved using a direct approach. The apparent challenge to the EM algorithm is that the likelihood surface is relatively flat in certain directions. Consequently, sequential estimates remain close together. To overcome this challenge, more information about the likelihood surface needs to be incorporated than the EM algorithm uses. Direct maximization of the likelihood surface can take advantage of both the gradient and curvature of the surface (Qin et al., 2000), and an extension of these ideas may be possible to the lattice model. The greatest obstacle to this approach is the large number of estimated parameters. However, it may be possible to maximize the likelihood in stages, working first with the lattice parameters and followed by the remaining parameters. In such an approach, care should be taken to minimize the number of times that the likelihood must be calculated.

References

Astumian, R. D. (1997). Thermodynamics and kinetics of a brownian motor. *Science*, 276:917–22.

Astumian, R. D. and Bier, M. (1994). Fluctuation driven ratchets: Molecular motors. *Phys. Rev. Lett.*, 72(11):1766–9.

Astumian, R. D. and Derényi, I. (1999). A chemically reversible Brownian motor: Application to kinesin and ncd. *Biophysical Journal*, 77:993–1002.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilitic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1554–63.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–71.

Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Stat.*, 26(4):1614–35.

Block, S. M., Goldstein, L. S., and Schnapp, B. J. (1990). Bead movement by single kinesin molecules studied with optical tweezers. *Nature*, 348(6299):348–52.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer-Verlag, New York.

Chung, S. H., Moore, J. B., Xia, L., Premkumar, L. S., and Gage, P. W. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *Philos. T. Roy. Soc. B*, 329:265–85.

Churchill, G. A. (1989). Stochastic-models for heterogeneous DNA-sequences. *B. Math. Biol.*, 51(1):79–94.

Coy, D. L., Wagenbach, M., and Howard, J. (1999). Kinesin takes one 8-nm step for each ATP that it hydrolyzes. *J. Biol. Chem.*, 274(6):3667–71.

Crevel, I. M.-T. C., Lockhart, A., and Cross, R. A. (1996). Weak and strong states of kinesin and ncd. *J. Mol. Biol.*, 257(1):66–76.

Dembo, A. and Zeitouni, O. (1986). Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stoch. Proc. Appl.*, 23:91–113.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38.

Douc, R. and Matias, C. (2000). Propriétés asymptotiques de l'estimateur de maximum de vraisemblance pour des modèles de markov cachés généraux. *C. R. Acad. Sci. I–Math*, 330(1):135–8.

Duke, T. and Leibler, S. (1996). Motor protein mechanics: a stochastic model with minimal mechanochemical coupling. *Biophys. J.*, 71:1235–47.

Elliott, R. J. (1994). Measure change estimates for hidden Markov models. *Syst. Control Lett.*, 23:149–57.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1997). *Hidden Markov Models: Estimation and Control*. Springer-Verlag, New York.

Feynman, R. P., Leighton, R. B., and Sands, M. (1963). *The Feynman Lectures on Physics*, volume 1, chapter 46. Addison-Wesley.

Fisher, M. E. and Kolomeisky, A. B. (1999a). The force exerted by a molecular motor. *P. Natl. Acad. Sci. USA*, 96:6597–602.

Fisher, M. E. and Kolomeisky, A. B. (1999b). Molecular motors and the forces they exert. *Physica A*, 274:241–66.

Fisher, M. E. and Kolomeisky, A. B. (2001). Simple mechanochemistry describes the dynamics of kinesin molecules. *P. Natl. Acad. Sci. USA*, 98(14):7748–53.

Ford, J. J. and Moore, J. B. (1998). On adaptive HMM state estimation. *IEEE T. Signal Proces.*, 46(2):475–86.

Fredkin, D. R. and Rice, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *P. Roy. Soc. Lond. B*, 249:125–32.

Friedman, D. S. and Vale, R. D. (1999). Single-molecule analysis of kinesin motility reveals regulation by the cargo-binding tail domain. *Nat. Cell Biol.*, 1(5):293–7.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., and Rossi, F. (2001). *GNU Scientific Library Reference Manual*. Network Theory Ltd.

Gelles, J., Schnapp, B. J., and Sheetz, M. P. (1988). Tracking kinesin-driven movements with nanometre-scale precision. *Nature*, 331(6155):450–3.

Gibbons, I. R. (1965). Chemical dissection of cilia. *Arch. Biol.*, 76:317–52.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Physics*, 22:403–34.

Giudici, P., Ryden, T., and Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics*, 56(3):742–7.

Hackney, D. D. (1988). Kinesin ATPase: Rate-limiting ADP release. *P. Natl. Acad. Sci. USA*, 85(17):6314–8.

Hancock, W. O. and Howard, J. (1998). Processivity of the motor protein kinesin requires two heads. *J. Cell Biol.*, 140(6):1395–405.

Hill, T. L. (1989). *Free Energy Transduction and Biochemical Cycle Kinetics*. Springer-Verlag, New York.

Hirokawa, N., Sato-Yoshitake, R., Kobayashi, N., Pfister, K. K., Bloom, G. S., and Brady, S. T. (1991). Kinesin associates with anterogradely transported membranous organelles in vivo. *J. Cell Biol.*, 114(2):295–302.

Howard, J., Hudspeth, A. J., and Vale, R. D. (1989). Movement of microtubules by single kinesin molecules. *Nature*, 342(6246):154–8.

Hua, W., Chung, J., and Gelles, J. (2002). Distinguishing inchworm and hand-over-hand processive kinesin movement by neck rotation measurements. *Science*, 295(5556):844–8.

Hua, W., Young, E. C., Fleming, M. L., and Gelles, J. (1997). Coupling of kinesin steps to ATP hydrolysis. *Nature*, 388(6640):390–3.

Hunt, A. J., Gittes, F., and Howard, J. (1994). The force exerted by a single kinesin molecule against a viscous load. *Biophys. J.*, 67(2):766–81.

Huxley, A. F. (1957). Muscle structure and theories of contraction. *Prog. Biophys. Biophys. Chem.*, 7:255–318.

Huxley, A. F. and Simmons, R. M. (1971). Proposed mechanism of force generation in striated muscle. *Nature*, 233:533–8.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, 5(3):299–314.

Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Stat.*, 27:514–35.

Jülicher, F., Ajdari, A., and Prost, J. (1997). Modeling molecular motors. *Rev. Mod. Phys.*, 69(4):1269–81.

Kolomeisky, A. B. and Fisher, M. E. (2000). Periodic sequential kinetic models with jumping, branching and deaths. *Physica A*, 279:1–20.

Kolomeisky, A. B. and Widom, B. (1998). A simplified "ratchet" model of molecular motors. *J. Stat. Phys.*, 93(3/4):633–45.

Kozielski, F., Sack, S., Marx, A., Thormahlen, M., Schonbrunn, E., Biou, V., Thompson, A., Mandelkow, E. M., and Mandelkow, E. (1997). The crystal structure of dimeric kinesin and implications for microtubule-dependent motility. *Cell*, 91(7):985–94.

Kuo, S. C., Gelles, J., Steuer, E., and Sheetz, M. P. (1991). A model for kinesin movement from nanometer-level movements of kinesin and cytoplasmic dynein and force measurements. *J. Cell Sci. Suppl.*, 14:135–8.

Kuo, S. C. and Sheetz, M. P. (1993). Force of single kinesin molecules measured with optical tweezers. *Science*, 260(5105):232–4.

Leibler, S. and Huse, D. A. (1991). A physical model for motor proteins. *C. R. Acad. Sci. III*, 313:27–35.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, 40:127–43.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Tech. J.*, 62(4):1035–74.

Magnasco, M. O. (1993). Forced thermal ratchets. *Phys. Rev. Lett.*, 71(10):1477–81.

Malik, F., Brillinger, D., and Vale, R. D. (1994). High-resolution tracking of microtubule motility driven by a single kinesin motor. *P. Natl. Acad. Sci. USA*, 91(10):4584–8.

Marx, A., Thormahlen, M., Muller, J., Sack, S., Mandelkow, E. M., and Mandelkow, E. (1998). Conformations of kinesin: solution vs. crystal structures and interactions with microtubules. *Eur. Biophys. J.*, 27(5):455–65.

Meyhofer, E. and Howard, J. (1995). The force generated by a single kinesin molecule against an elastic load. *P. Natl. Acad. Sci. USA*, 92(2):574–8.

Michalek, S. and Timmer, J. (1999). Estimating rate constants in hidden Markov models by the EM algorithm. *IEEE T. Signal Proces.*, 47(1):226–8.

Millonas, M. M. and Dykman, M. I. (1994). Transport and current reversal in stochastically driven ratchets. *Phys. Lett. A*, 185:65–9.

Muresan, V. (2000). One axon, many kinesins: What's the logic? *J. Neurocytol.*, 29(11-12):799–818.

Peskin, C. S., Ermentrout, G. B., and Oster, G. F. (1994). The correlation ratchet: a novel mechanism for generating directed motion by ATP hydrolysis. In *Symposium Cell Mechanics and Cellular Engineering*, pages 479–89. Springer-Verlag, New York.

Peskin, C. S. and Oster, G. (1995). Coordinated hydrolysis explains the mechanical behavior of kinesin. *Biophys. J.*, 68:202–11s.

Press, W. H., Flannery, B. P., Teukolosky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York.

Qian, H. (1997). A simple theory of motor protein kinetics and energetics. *Biophys. Chem.*, 67:263–7.

Qin, F., Auerbach, A., and Sachs, F. (2000). A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.*, 79:1915–27.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *P. IEEE*, 77(2):257–86.

Ray, S., Meyhofer, E., Milligan, R. A., and Howard, J. (1993). Kinesin follows the microtubule's protofilament axis. *J. Cell Biol.*, 121(5):1083–93.

Rice, S. et al. (1999). A structural change in the kinesin motor protein that drives motility. *Nature*, 402(6763):778–84.

Romberg, L. and Vale, R. D. (1993). Chemomechanical cycle of kinesin differs from that of myosin. *Nature*, 361(6403):168–70.

Rosenfeld, S. S., Jefferson, G. M., and King, P. H. (2001). ATP reorients the neck linker of kinesin in two sequential steps. *J. Biol. Chem.*, 276(43):40167–74.

Rosenfeld, S. S., Rener, B., Correia, J. J., Mayo, M. S., and Cheung, H. C. (1996). Equilibrium studies of kinesin-nucleotide intermediates. *J. Biol. Chem.*, 271(16):9473–82.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.

Schnapp, B. J. and Reese, T. S. (1989). Dynein is the motor for retrograde axonal transport of organelles. *P. Natl. Acad. Sci. USA*, 86(5):1548–52.

Schnitzer, M. J. and Block, S. M. (1995). Statistical kinetics of processive enzymes. *Cold Spring Harb Symp. Quant. Biol.*, 60:793–802.

Schnitzer, M. J. and Block, S. M. (1997). Kinesin hydrolyses one ATP per 8-nm step. *Nature*, 388(6640):386–90.

Schnitzer, M. J., Visscher, K., and Block, S. M. (2000). Force production by single kinesin motors. *Nat. Cell Biol.*, 2(10):718–23.

Smith, D. A., Steffen, W., Simmons, R. M., and Sleep, J. (2001). Hidden-Markov methods for the analysis of single-molecule actomyosin displacement data: The variance-hidden-Markov method. *Biophys. J.*, 81:2795–816.

Sosa, H., Peterman, E. J., Moerner, W. E., and Goldstein, L. S. (2001). ADP-induced rocking of the kinesin motor domain revealed by single-molecule fluorescence polarization microscopy. *Nat. Struct. Biol.*, 8(6):540–4.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton Univeristy Press.

Svoboda, K. and Block, S. M. (1994a). Biological applications of optical forces. *Annu. Rev. Biophys. Biom.*, 23:247–85.

Svoboda, K. and Block, S. M. (1994b). Force and velocity measured for single kinesin molecules. *Cell*, 77(5):773–84.

Svoboda, K., Mitra, P. P., and Block, S. M. (1994). Fluctuation analysis of motor protein movement and single enzyme kinetics. *P. Natl. Acad. Sci. USA*, 91:11782–6.

Svoboda, K., Schmidt, C. F., Schnapp, B. J., and Block, S. M. (1993). Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 365(6448):721–7.

Vale, R. D., Funatsu, T., Pierce, D. W., Romberg, L., Harada, Y., and Yanagida, T. (1996). Direct observation of single kinesin molecules moving along microtubules. *Nature*, 380(6573):451–3.

Vale, R. D. and Oosawa, F. (1990). Protein motors and Maxwell's demons: does mechanochemical transduction involve a thermal ratchet? *Adv. Biophy.*, 26:97–134.

Vale, R. D., Reese, T. S., and Sheetz, M. P. (1985a). Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell*, 42(1):39–50.

Vale, R. D., Schnapp, B. J., Mitchison, T., Steuer, E., Reese, T. S., and Sheetz, M. P. (1985b). Different axoplasmic proteins generate movement in opposite directions along microtubules in vitro. *Cell*, 43(3 Pt 2):623–32.

Vallee, R. B. and Bloom, G. S. (1991). Mechanisms of fast and slow axonal transport. *Annu. Rev. Neurosci.*, 14:59–92.

van Kampen, N. G. (1981). *Stochastic processes in physics and chemistry*. North-Holland, New York.

Venkataramanan, L., Kuc, R., and Sigworth, F. J. (1998a). Identification of hidden Markov models for ion channel currents–Part II: State-dependent excess noise. *IEEE T. Signal Proces.*, 46(7):1916–29.

Venkataramanan, L., Walsh, J. L., Kuc, R., and Sigworth, F. J. (1998b). Identification of hidden Markov models for ion channel currents–Part I: Colored background noise. *IEEE T. Signal Proces.*, 46(7):1901–15.

Visscher, K. and Block, S. M. (1998). Versatile optical traps with feedback control. *Methods Enzymol.*, 298:460–89.

Visscher, K., Schnitzer, M. J., and Block, S. M. (1999). Single kinesin molecules studied with a molecular force clamp. *Nature*, 400(6740):184–9.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory*, 13(2):260–9.

Wagner, M. and Timmer, J. (2001). Model selection in non-nested hidden Markov models for ion channel gating. *J. Theor. Biol.*, 208:439–50.

Walker, R. A., Salmon, E. D., and Endow, S. A. (1990). The Drosophila claret segregation protein is a minus-end directed motor molecule. *Nature*, 347(6295):780–2.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

Woehlke, G. and Schliwa, M. (2000). Directional motility of kinesin motor proteins. *Biochim. Biophys. Acta*, 1496(1):117–27.

Wriggers, W. and Schulten, K. (1998). Nucleotide-dependent movements of the kinesin motor domain predicted by simulated annealing. *Biophys. J.*, 75(2):646–61.

Xing, J., Wriggers, W., Jefferson, G. M., Stein, R., Cheung, H. C., and Rosenfeld, S. S. (2000). Kinesin has three nucleotide-dependent conformations. Implications for strain-dependent release. *J. Biol. Chem.*, 275(45):35413–23.

Young, E. C., Mahtani, H. K., and Gelles, J. (1998). One-headed kinesin derivatives move by a nonprocessive, low-duty ratio mechanism unlike that of two-headed kinesin. *Biochemistry*, 37(10):3467–79.

Zeitouni, O. and Dembo, A. (1988). Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE T. Inform. Theory*, 34(4):891–3.