UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Stochastic Modeling and Analysis of DNA Polymerase Kinetics Based on Observed Dwell Times

Permalink

https://escholarship.org/uc/item/5p47w4jt

Author

Labaria, George Reyes

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

STOCHASTIC MODELING AND ANALYSIS OF DNA POLYMERASE KINETICS BASED ON OBSERVED DWELL TIMES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS AND STATISTICS

by

George Reyes Labaria

June 2019

The Dissertation of George Reyes Labaria
is approved:

Hongyun Wang, Chair

Mark Akeson

Qi Gong

Daniele Venturi

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Copyright © by

George Reyes Labaria

2019

Table of Contents

List of Figures

List of Tables

Abstract

Stochastic Modeling and Analysis of DNA Polymerase Kinetics Based on Observed Dwell Times

by

George Reyes Labaria

DNA Polymerases (DNAPs) are enzymes that make DNA molecules by assembling nucleotides and are responsible for copying the genome in all cells. Fidelity in genome replication is essential for genome integrity. Replication errors could lead to mutations which lead to diseases, including cancer. DNAPs selectively bind a deoxyribonucleoside triphosphate (dNTP) that is complementary to the template nucleotide of the DNA they are copying. After the covalent incorporation of a complementary nucleotide into the newly synthesized DNA strand, the DNAP moves onto the next template nucleotide in the translocation step driven by thermal fluctuations, allowing for a new round of binding. The binding and incorporation of a nucleotide, along with the translocation step, consist of a full nucleotide addition cycle. Nanopore experiments allow us to observe the DNAP translocation along the template strand with single-nucleotide spatial precision and millisecond temporal resolution. We develop mathematical formulations and methods to infer the kinetic details of the nucleotide addition cycle from dwell time data obtained from the nanopore experiments. We fully characterize the uncertainty in the inferred kinetic details, and show that the uncertainty can be controlled in experimental design. We show that a dimensionless quantity based on the randomness parameter provides a lower and upper bounds on the number of biochemical states in the polymerization (pol) step of a replication cycle. Understanding the kinetic details of the nucleotide addition cycle is essential to

elucidating the mechanisms which regulate fidelity. The inference methods we developed can be applied to other single molecule experiments in which dwell time samples are observable. More importantly, the analysis results and methods for designing optimal experimental conditions will motivate more meaningful and informative single molecule measurements.

To my parents, who have loved and supported me every step of the way

Acknowledgments

I would like to thank my adviser Professor Hongyun Wang for mentoring me throughout my graduate career. Without his thoughtful guidance, unyielding patience, and support, this dissertation would not be possible. Thank you Hongyun for never losing faith in me and guiding me throughout my graduate studies. I would also like to thank Professor Daniele Venturi, Professor Mark Akeson, and Professor Qi Gong for serving on my qualifying examination and dissertation committees. Their constructive comments, suggestions, unique insights, and support played an invaluable role in the shaping of this work.

I would like to thank my fellow graduate students and friends Sara Nasab, Richard Shaffer, Matthew Simms, Catherine Brennan, Michael Lavell, and Yuanran Zhu for making research in a windowless office so much more bearable. Thank you all for your encouragement and unwavering support throughout the years.

I would also like to thank Caroline Harman who always reminded me to never give up, and that there is light at the end of the tunnel. Her caring support and encouragement always gave me an extra boost.

Last but certainly not least, I would like the thank my parents for their unconditional love and support. They supported me every step of the way, working tirelessly and selflessly to support me through my undergraduate and graduate studies. Their love, support, and encouragement were sometimes all that kept me going. Mom and Dad, I love you.

Chapter 1

Introduction

DNA Polymerases (DNAP) are enzymes that create DNA molecules by assembling nucleotides and are responsible for copying the genome in all cells. Fidelity of the replication is essential for avoiding mutations which can lead to disease, including cancer. DNAPs selectively bind a deoxyribonucleoside triphosphate (dNTP) that is complementary to the template nucleotide. This selection must be made over non-complementary dNTPs and over ribonucleoside triphosphates (rNTPs). In addition to discriminating against non-complementary dNTPs and rNTPs in nucleotide binding, fidelity is also achieved by 3' to 5' exonucleolytic editing of non-complementary dNTPs that have escaped the initial discrimination and were incorporated into the DNA. These coordinated activities give DNAPs an error rate of about one mistake per 10^8-10^9 base pairs, mostly in bacteria [63]. Error rates can be as high as one mistake per 10^2-10^3 base pairs in error-prone polymerase genes in humans [34]. After incorporation of a complementary dNTP, DNAPs must translocate a distance of one nucleotide to reset the DNAP active site for the next nucleotide addition cycle. Errors in translocation can cause frameshift mutations and deletion errors. Understanding the kinetic steps of nucleotide addition cycle–translocation, exonucleolytic editing, dNTP binding, dNTP incorporation, and the kinetic structure of the polymerization process are essential to elucidating the mechanisms which regulate fidelity.

Cells have multiple pathways that can correct replication errors
[23], [37], [40], [46], [47], [53], [54], [64], [66], and [69]. The mechanisms that discriminate against incorrect nucleobases and sugars for incoming nucleotide substrates has been studied for numerous DNAPs [7], [33], [39], and [75]. The structure of the DNAP domain is highly conserved and resembles a partially closed right hand. In this right hand analogy, the DNAP domain has palm, fingers, and thumb subdomains [6], [22], [25], [35], [30]. The palm subdomain contains residues that are required for catalysis. The thumb subdomain positions the primer-template in the active site. In DNAP-DNA complexes, containing complementary dNTP to the templating nucleotide, the fingers subdomain moves relaive to its position in complexes without a dNTP. Here, the fingers close and move toward the DNAP active site (the palm) to obtain a tight steric fit for the nascent base pair. Correct nucleotide substrates promote pre-chemistry conformational changes that are necessary to achieve optimal alignment in the DNAP active site, while incorrect nucleotides do this less efficiently [75]. The kinetics of nucleotide discrimination based on ensemble pre-steady state essays have been carried out in the past [33] [38]. Although studies such as these have provided lots of information on the kinetics of discrimination, many kinetics aspects of discrimination have yet to be determined.

The DNAP is a molecular motor that moves from the 3' end to the 5' end of the template strand of the DNA. The DNAP converts chemical energy into mechanical work and motion [1], [4], [41], [43], [44]. Chemical energy derived from the polymerization of the primer strand is used to drive the DNAP in a unidirectional manner, generating force against mechanical barriers and hydrodynamic drag [76].

The DNAP can be thought of as a small machine operating in a thermal bath, subject to fluctuations in conformation and chemical state. These microscopic fluctuations are not observable in the ensemble averages of bulk experiments, but some can be directly observed in single molecule experiments. This physical picture of the DNAP corresponds to a random walk of a particle guided by the free-energy surface of the system in which the mechanical motion and chemical reaction are coupled [4], [12], [41] (figure 1.1). The diffusion fluxes that result from this random walk give the the chemical reaction rates and mechanical velocities of the DNAP [12], [41].

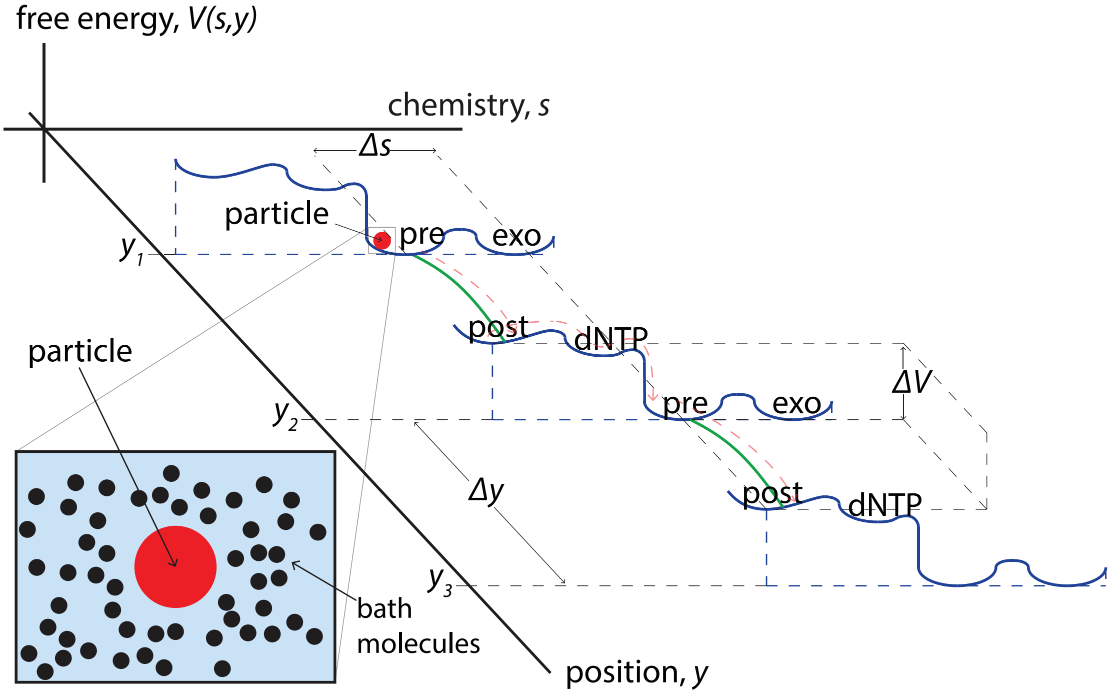

Figure 1.1: DNAP can be modeled as a particle undergoing a random walk on a free-energy surface. Collision by the bath molecules make the particle undergo Brownian motion that is statistically biased by the free-energy surface.

In this view, molecules in the surrounding bath collide with the DNAP (a larger particle), causing it to undergo Brownian motion; the resulting motion is biased by the free-energy surface. The particle spends a significant amount of time in equilibrium in the potential wells of the surface, occasionally fluctuating out, driven by the thermal noise, via low-energy barriers. If the particle starts in the potential well labeled “pre” in figure 1.1, the particle will likely fluctuate along the low-energy bridge (denoted by a green contour) to the “post” well. Since this fluctuation is parallel to the position coordinate, this manifests a change in position of the DNAP motor with no change in chemical occupancy. These fluctuations between the pre and post states may occur many times before the particle fluctuates to the “dNTP” well. The fluctuation to dNTP depends on the availability of dNTP molecules in bulk solution. It involves the change of chemical occupancy of the catalytic site and is parallel to the chemistry coordinate; it does not directly induce any translocation between the DNAP motor and DNA substrate. After dNTP binding, the system will incorporate the nucleotide, experiencing a large drop in free energy of magnitude ΔV. Then the system will fluctuate to the pre-translocation state of the next cycle. After nucleotide incorporation, in the absence of any catalyst, the particle will not fluctuate back to the previous cycle due the large free energy barrier.

This describes a Brownian ratchet. For a Brownian ratchet, chemical reaction does not provide a direct active force for the motion. Instead of directly driving the motor, chemistry selects the forward fluctuations and prevents the backwards fluctuations. Over a long time, the particle moves forward in position, producing useful work from the random thermal noise of the bath in the presence of the chemical gradient (see for example, [12]). The reasoning behind the way the free-energy surface was drawn in figure 1.1 and the claim that the DNAP is a Brownian ratchet motor will become clear in the later chapters.

Due to the small size of the DNAP, the effect of inertia is negligible. Each attempt of the particle at crossing over an energy barrier ends with either arriving at a new local minimum or returning to the current minimum, and the time spent in the actual crossing of the energy barrier is a small fraction of the total waiting time. It is believed that the potential well of each chemical state is a collection of small energy ripples, where the magnitude of the ripples are smaller than the available energy in the bath ($< k_B T$) where k_B is the Boltzmann constant, and T is the temperature [29] (see figure 1.2). The result is that the particle will fluctuate rapidly inside a well before eventually escaping to another potential well. Since the energy barriers of the well are much larger than $k_B T$, the fluctuations to other potential wells are on a much slower timescale. The effect of this separation of timescales is that the faster fluctuations inside the well will average out any differences in residence time of the particle within the well due to the exact position of the particle inside the well [29]. Approximately, as a result, the system looses all memory of the previous states; the previous attempts at crossing over an energy barrier; and the time elapsed since arriving at the present state. The future evolution of the system is hence solely governed by the present state and the

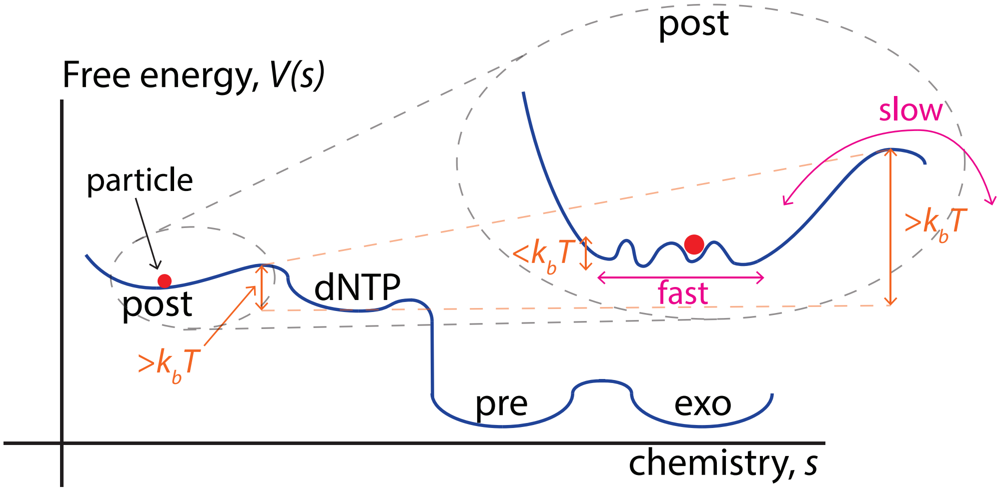

Figure 1.2: Low energy ripples ($< k_B T$) lead to fast fluctuations of the particle inside the well. Since the energy barriers around the well are much higher ($> k_B T$), the timescale of fluctuating between wells is much slower. This separation of timescales lead to "memorylessness" in the system.

stochastic thermal fluctuations from the environment. Thus we can characterize the system as a time-homogeneous, continuous-time Markov process [41], [45]. This provides a well-defined framework for studying the DNAP.

If the energy barriers surrounding the potential wells are not significantly higher, so that there is no longer adequate separation of timescales, the description of molecular motor dynamics using a memoryless particle undergoing discrete jumps between wells is no longer valid [44]. In this situation, the fluctuations between wells are not significantly slower than the relaxation time of a particle undergoing fluctuations within a well. The result here is that the residence time within the well may depend on the previous state. It is believed that the internal fluctuations within a well are on the nanosecond or faster timescale [29]. Hence Markov description of particle dynamics on the millisecond or larger timescale will be unlikely affected by these faster fluctuations. As we will briefly describe later in this section, and more thoroughly covered in any aforementioned references, the nanopore experiments we will be discussing here have time-resolution at the sub-millisecond ($>> ns$) temporal resolution. We thus do not concern ourselves with these complications. However, as future experimental techniques increase in resolution in both space and time, more complicated models for particle dynamics will have to be considered [77].

The DNAP moves in single nucleotide increments, and the movement from one nucleotide to the next is known as DNAP translocation or translocation step [18], [48]. The translocation and subsequent incorporation of a complementary nucleotide is known as the nucleotide addition cycle. Despite the importance of the translocation step, dNTP binding and incorporation, and exonucleolytic editing in understanding the mechanisms that regulate replication fidelity, their kinetics are not well understood. Previous work has been done to determine the kinetic structure and estimate the kinetic rates from dwell time data of non-synthesizing DNAP-DNA complexes [18], [19], [49], [50], [51].

Extracting kinetic parameters of the DNAP-DNA complex using ensemble methods is very difficult. Single-molecule experiments designed to study this also require a challenging combination of high spatial and temporal resolution since the translocation step involves a spatial displacement of only 0.3nm [3]. The translocation step of the bacteriophage ϕ-29 DNAP can be directly observed at the single-molecule level using an α-hemolysin nanopore with single nucleotide spatial and sub-millisecond temporal resolution [18]. In the past couple decades, nanopore experiments have become an important tool to study DNA and DNAPs at the single-molecule level [2], [5], [21], [26], and [48].

We use the bacteriophage ϕ-29 DNAP as a model system to study the translocation step and its kinetics. The ϕ-29 DNAP catalyzes highly processive DNA replication without the need for accessory proteins [9]. This provides a robust and high throughput experimental assay for studying the kinetics of the translocation step in the framework of rigorous mathematical models. The ϕ-29 DNAP is in the B family of DNAPs, which includes DNAPs δ and ϵ. Among members of the B family, the structures and mechanisms which contribute to replication fidelity are highly conserved [6], [31], [32], [71], [74]. Mutations in the human polymerase genes for pol δ and pol ϵ are linked to colon and endometrial cancers [17], [28], [61]. Thus to understand these cancers, the mechanisms which regulate fidelity in the ϕ-29 DNAP must be understood in detail.

The nanopore experiments allow us to observe individual DNAP-DNA complexes at specific known positions along the DNA template and control replication of DNA molecules [15], [18], [19], [48], [49], [50], and [51]. In each experiment, thousands of DNAP-DNA complexes can be examined individually.

In the nanopore experiments, DNAP-DNA complexes diffuse in bulk and captured atop an α-hemolysin nanopore which is embedded in a lipid membrane that separates two chambers. The nanopore is wide enough to accommodate only a single-strand DNA. A voltage is applied across the membrane and the ionic current trace, carried by potassium and chloride ions, is measured (figure 1.3). In this setup, DNAP-DNA complexes from bulk are driven towards the nanopore by the electric field. The nanopore is only wide enough to accommodate a single strand of DNA, and so the driven DNAP-DNA complex perches atop the nanopore with the single-strand of DNA suspended through the nanopore lumen (figure 1.4). More thorough introductions to the experimental setup can be found in [18], [19], [49], [50], and [51].

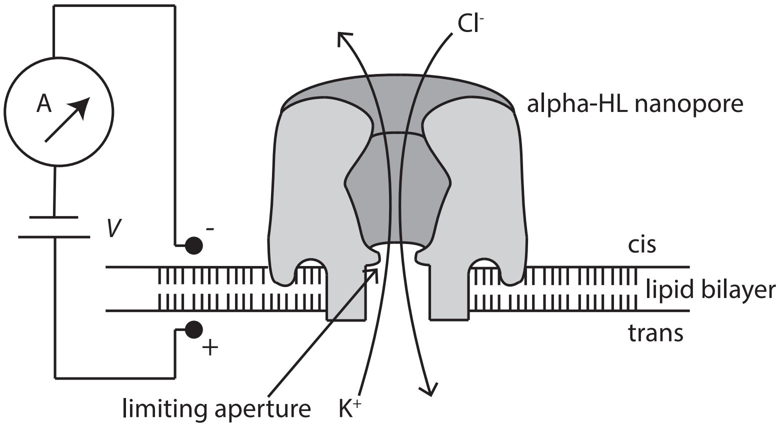

Figure 1.3: Schematic diagram of the nanopore experiment.

The ability to detect DNA displacement is achieved by a reporter group in the template strand, formed by five consecutive abasic residues. The abasic reporter group is thinner than the surrounding nucleobases, so the reporter group region allows for more ion flow through the limiting aperture of the nanopore. This in turn results in an increase in measured current amplitude as the reporter group nears the limiting aperture, and a decrease in amplitude as the reporter group moves further away from the aperture. The abasic reporter group thus reports on the direction and distance of the DNA substrate relative to the DNAP and nanopore during reactions [15] [18], [48] (figure 1.6).

Experimental data suggests that the ionic current trace will undergo iterative transitions across the translocation step when the DNAP-DNA complexes are not allowed to undergo synthesis (figure 1.4) [18]. The equilibrium across the DNAP translocation step is dependent on applied force, dNTP concentration, and by the DNA sequences close to the DNAP active site [18]. The experimental observations support a model in which the DNAP-DNA complex fluctuates between these two states and is driven by Brownian thermal motion [18]. The dNTP binding only occurs after transition to the post-translocation state, and the DNAP-DNA complex is rectified to the lower-amplitude, post-translocation state after dNTP binding [18]. The presence of dNTP shifts the translocation step equilibrium towards the lower amplitude, post-translocation state [50]. The pre-translocation state is also a branch point in which transfer of the primer strand to the exonuclease active site can occur [51]. When exonuclease activity is blocked, transition to the exonuclease is succeeded by a subsequent transition back to the pre-translocation state. In this setting, the DNAP-DNA complex will undergo stochastic transitions among the pre-translocation, post-translocation, exonuclease, and dNTP-bound states (figure 1.5). When using the DNA substrate described in [50], the upper and lower amplitudes are centered at about 32pA and 26pA, respectively when a voltage of 180mV is applied.

When the DNAP-DNA complex is allowed to undergo synthesis and complementary dNTP are provided in the cis chamber, then the ionic current will fluctuate in discrete amplitude levels. During synthesis, as the captured DNAP-DNA complex sits atop the nanopore, the template strand is drawn through the

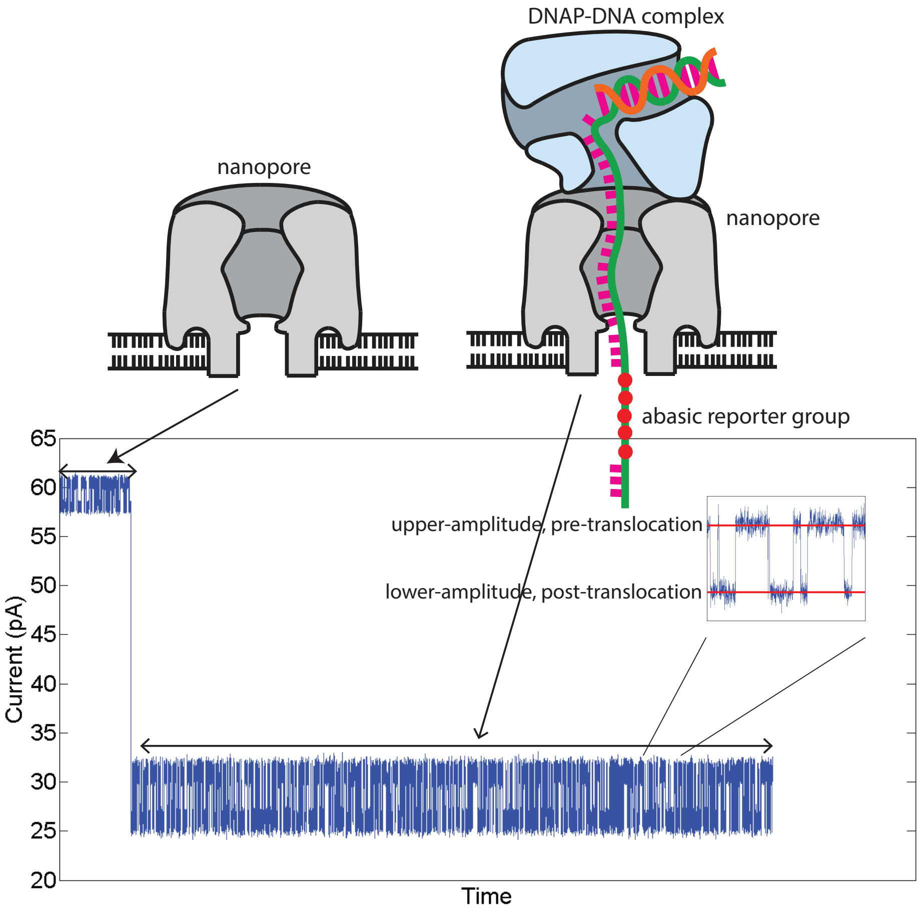

Figure 1.4: A schematic diagram of the membrane with embedded nanopore. The current amplitude level drops when a DNAP-DNA complex is captured atop the pore. When the captured DNAP-DNA complex is not allowed to undergo synthesis, the ionic current fluctuates between two distinct amplitude levels which corresponds to the upper-amplitude, pre-translocation and lower-amplitude, post-translocation states.

limiting aperture of the nanopore. The abasic reporter group is thus drawn closer to the limiting aperture and the measured ionic current level increases. When the reporter group is centered in the limiting aperture, the measured current traces reaches its maximum. Finally, after the reporter group passes through the limiting aperture, the current rapidly decreases (figure 1.6).

A blocking oligomer achieves (1) the protection of DNA in bulk phase from ϕ-29

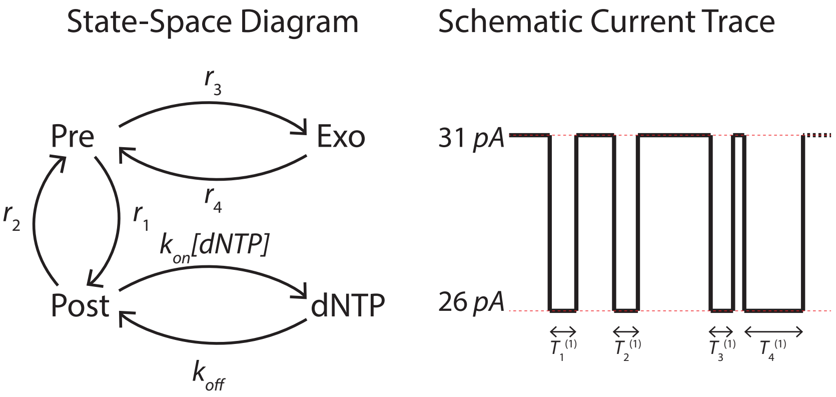

Figure 1.5: The relevant states and a representative current trace for one nucleotide addition cycle. Here, incorporation of a complementary nucleotide is blocked, preventing the DNAP-DNA complex from transitioning to the next nucleotide addition cycle.

DNAP-catalyzed replication and exonucleolysis; (2) capture-depended initiation of synthesis [15],[60]. Without the blocking oligomer, the suspended DNAP-DNA complexes in bulk will have already undergone synthesis before capture atop the nanopore. The blocking oligomer is attached to the template strand of the primer-template substrate immediately adjacent to the primer terminus and features a string of complementary residues capped by a tail of several abasic residues and a three-carbon spacer at the end. Upon capture of the DNAP-DNA complex atop the nanopore, the blocking oligomer is unzipped. This is facilitated by the non-complementary tail and the force induced by the voltage [15], [60].

In this setting, the abasic amplitude peak will be traversed twice [15] (see figure 1.7). The first traversal occurs as the blocking oligomer is unzipped upon capture and the template abasic reporter group is moved through the nanopore lumen into the trans chamber by the force induced by the voltage. Upon unzipping of the blocking oligomer, the DNAP encounters an exposed primer terminus and synthesis occurs. This draws the template strand with the abasic residues upwards. The abasic residues again move through the limiting aperture of the

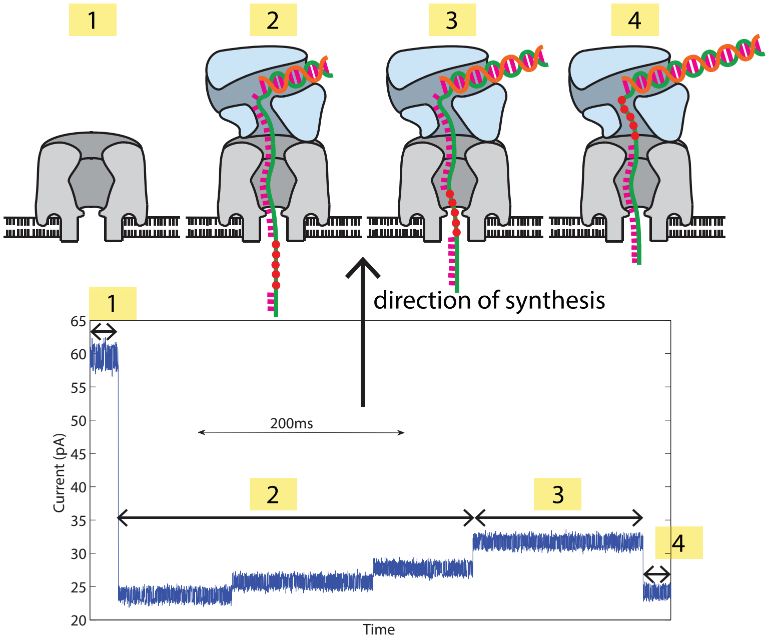

Figure 1.6: A schematic diagram of a DNAP-DNA complex undergoing synthesis captured atop a nanopore. With the absence of a DNAP-DNA complex captured atop the pore, the ionic current is at its highest (1). Upon capture, the ionic current is partially blocked, resulting in a large decrease in ionic current (2); As the complex begins synthesis, the reporter group is drawn towards the limiting aperture of the nanopore, manifested by an increase in current. When the reporter group is centered in the limiting aperture, the measured current trace reaches its maximum (3). After the reporter group passes through the limiting aperture of the nanopore, the current rapidly decreases (4).

nanopore, providing a second current peak qualitatively similar to figure 1.6.

Using the DNA substrate in [15], there are 25 nucleotide addition cycles catalyzed by the ϕ-29 DNAP. The amplitude levels corresponding to the 25 cycles were determined in a series of mapping experiments [15]. There are a subset of cycles that yield distinct current amplitude levels. Focusing on the 17th-19th nucleotide addition cycles, the current amplitudes for these cycles are about 31pA, 26pA, and 23.5pA, respectively at 180mV. Such a current amplitude trace will be qualitatively similar to the one in figure 1.8.

As we will see, the dwell times contain a lot of information about the ki-

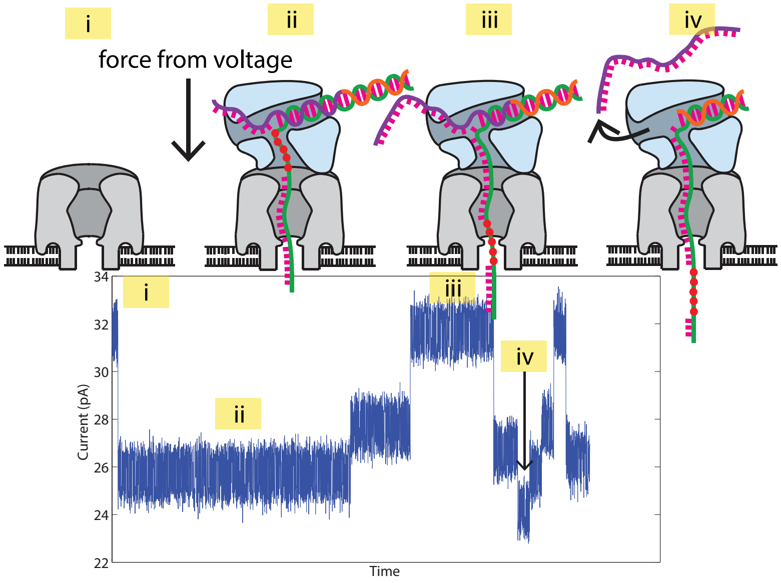

Figure 1.7: Schematic diagram of a DNAP-DNA complex with blocking oligomer. With the absence of the complex captured atop the nanopore, the ionic current is at its highest (i); upon capture, force induced by the voltage pulls the template strang downwards and unzips the blocking oligomer leading to successive increases in current (ii); when the abasic reporter group is directly aligned with the limiting aperture, the current trace is at its relative highest (iii); the current drops rapidly as the reporter group passes the limiting aperture and the blocking oligomer is ejected when it is fully unzipped from the template strand (iv); when the blocking oligomer is ejected, a primer-template terminus is exposed an the DNAP begins synthesis, drawing the template strand upwards against the force induced by the voltage. The current trace from here forward in time is qualitatively similar to that in labels (2)-(4) in figure 1.6.

netic structure of the nucleotide addition cycle. This dissertation focuses on what can be inferred from the dwell time data obtained during nanopore experiments. In chapter 2, we start with analyzing dwell time data from nanopore experiments in which the DNAP-DNA complexes cannot proceed to the chemical step of phosphodiester bond formation. In this situation, the DNAP-DNA complexes stochastically transition between four biochemical states: pre-translocation, post-translocation, exonuclease, and dNTP-bound states [18], [19], [49], [50], and [51]. We will review the previous work done in this regime and present a new method

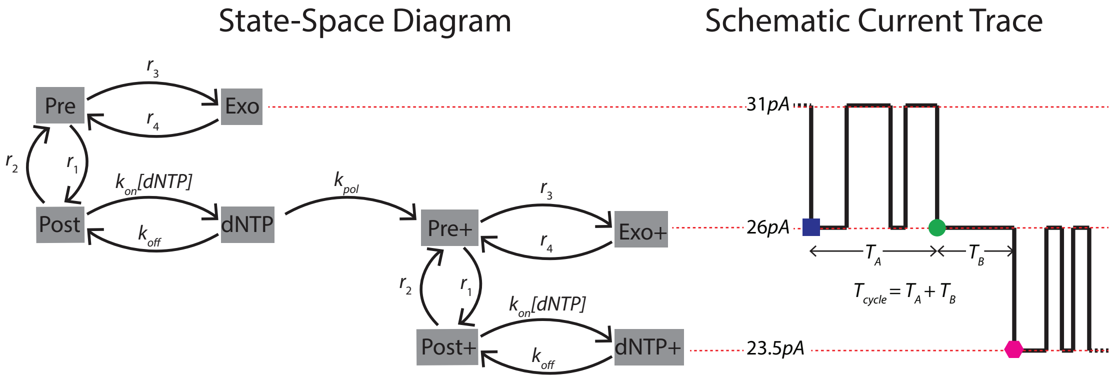

Figure 1.8: A state-space diagram for two nucleotide addition cycles in DNA replication. When the DNAP-DNA complex is allowed to undergo synthesis and a complementary dNTP is provided in the cis chamber, the DNAP-DNA complex can transition to the next nucleotide addition cycle–indicated by the "+" symbol after the state names. This is manifested as a change in the upper and lower amplitudes as the reporter group gets closer or further away from the nanopore lumen.

for inferring the kinetic rates based on maximum-likelihood estimation. We will show that the dNTP concentration, which can be controlled in the nanopore experiments, plays an important role in regulating the statistical uncertainty of the inferred dNTP binding and disassociation rates from dwell time data. Care must therefore be taken when choosing the dNTP concentration in the experiments. We characterize the inference uncertainty in the inferred dNTP binding and disassociation rates, and show how optimal experimental conditions can be determined. The methodology for choosing optimal experimental conditions will be extended to include constraints on the experimental time. We end this chapter with a characterization of the effects of multiplicative noise in the observed dwell times on the inferred kinetic rates.

In chapter 3, we will extend the results in chapter 2 by considering synthesizing DNAP-DNA complexes which can proceed through the polymerization process and incorporate a complmentary dNTP. In this context, the polymerization process is modeled as a single rate-limiting step from the dNTP-bound state to the pre-translocation state of the next nucleotide addition cycle. To the best of our knowledge, a stochastic model for DNAP-DNA complexes going through multiple nucleotide addition cycles based on observed dwell times has not been examined in the literature. We show that in regards to the relevant dwell times, synthesizing DNAP-DNA complexes can be mathematically mapped to an equivalent non-synthesizing complex with modified backwards translocation, dNTP binding, and dNTP disassociation rates. Therefore any inference methods and analysis based on dwell times for non-synthesizing complexes can be applied to synthesizing complexes.

Finally, in chapter 4, a general polymerization (pol) process is examined. In chapter 3, the pol process is modeled as a single rate-limiting step. The kinetic details of the polymerization process for DNAP-DNA complexes is largely unknown, but it consists of least binding, chemistry, and pyrophosphate release. Determining the number of effective kinetic states in the pol process is essential to discovering any fidelity regulating mechanisms in the dNTP incorporation step. We develop methods to infer the kinetic details of the pol process from dwell time data by examining a quantity based on the randomness parameter of the dwell times. In certain idealized situations, this quantity can determine the number of steps in the pol process exactly. In the more general case, we present a conjecture that puts a bounds on the number of steps on the polymerization process.

Chapter 2

Dynamics of dNTP Binding in Non-Synthesizing DNAP-DNA Complexes

2.1 Introduction

In this chapter, we determine the dNTP binding and disassociation rates using dwell time data for non-synthesizing DNAP-DNA complexes that cannot proceed to the chemical step of phosphodiester bond formation. In this setting, the DNAP-DNA complex stochastically transitions between the pre-translocation, post-translocation, exonuclease, and dNTP-bound states (figure 2.1). We derive the probability density function (PDF) underlying the dwell time data and determine the maximum-likelihood estimates of the binding and disassociation rates by use of the expectation-maximization (EM) algorithm. Previous work has been done to estimate these rates by use of a autocorrelation function of the entire current amplitude measured from nanopore experiments [50]. We will show that our

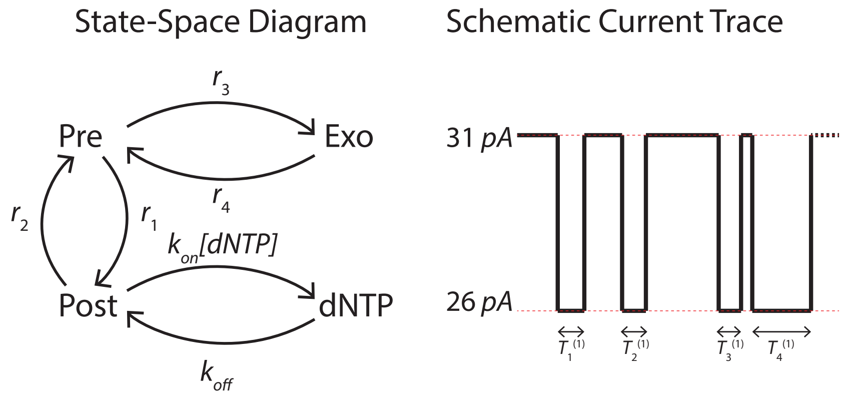

Figure 2.1: The relevant states and a representative current trace for one nucleotide addition cycle. Here, incorporation of a complementary nucleotide is blocked, preventing the DNAP-DNA complex from transitioning to the next nucleotide addition cycle.

method is robust against measurement noise and that the framework is general enough to be applied to other Markovian phenomena in which dwell time data is available.

Recall that in the synthesizing case, the ionic current trace covers more than one nucleotide addition cycle if complementary dNTP are provided in the cis chamber (figure 2.2). We define various dwell times of interest.

- T_A: the time from the first arrival to the post-translocation state of the current nucleotide addition cycle, to the last arrival to the post-translocation state of the current nucleotide addition cycle; this is shown graphically as the blue square to the green circle in figure 2.2.

- T_B: the time from the last arrival to the post-translocation state of the current nucleotide addition cycle to the first arrival to the post-translocation state of the next nucleotide addition cycle; this is shown graphically as the green circle to the magenta hexagon in figure 2.2.

- $T^{(1)}$: the lower-amplitude dwell times within the T_A dwell time segment

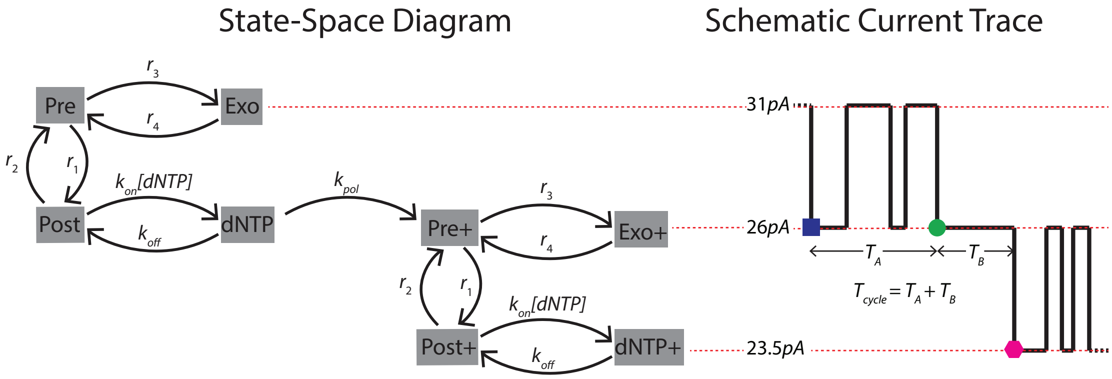

Figure 2.2: A state-space diagram for two nucleotide addition cycles in DNA replication. When the DNAP-DNA complex is allowed to undergo synthesis and a complementary dNTP is provided in the cis chamber, the DNAP-DNA complex can transition to the next nucleotide addition cycle–indicated by the "+" symbol after the state names. This is manifested as a change in the upper and lower amplitudes as the reporter group gets closer or further away from the nanopore lumen.

(figure 2.2). In any observation of T_A, there are likely to be many samples of $T^{(1)}$ and we label them as $T_1^{(1)}, T_2^{(1)}, T_3^{(1)}, \ldots$, etc (figure 2.1).

The transition rates $r_1, r_2, r_3, r_4, k_{on}, k_{off}$, and k_{pol} shown in figures 2.1 and 2.2 are defined as follows. Each transition rate is written next to an arrow originating from state i and ending at state j. That transition rate is the rate of which the DNAP-DNA complex transitions from state i to state j. For example, r_1 is the rate of which the DNAP-DNA complex transitions from the pre-translocation state to the post-translocation state. Mathematically, we can write

$$r_1 = \lim_{\Delta t \to 0^+} \frac{Pr\left(X\left(t + \Delta t\right) = \text{Post} \mid X\left(t\right) = \text{Pre}\right)}{\Delta t},$$

where $X\left(t\right)$ denotes the state of the Markov chain at time t. The other transition rates are defined in a similar manner.

In this chapter, we are interested in the case in which the DNAP-DNA complex cannot undergo synthesis; a mutation is engineered into the DNAP which prohibits dNTP incorporation, blocking the incorporation step of the polymerization process. Hence $k_{pol} = 0$ in figure 2.2. The DNAP-DNA complex will thus undergo stochastic transitions among the pre-translocation, post-translocation, exonuclease, and dNTP-bound states without ever proceeding to the next nucleotide addition cycle (figure 2.1). A mutation is also engineered into the exonuclease so that cleaving of the dNTP cannot occur. We are interested in inferring the transition rates k_{on} and k_{off} from the $T^{(1)}$ data. In this situation, observing $T^{(1)}$ is not in competition with T_B, since the DNAP-DNA complex will not go through the irreversible polymerization process. This provides a simplified situation in which to examine the information content of the $T^{(1)}$ data.

In [50], the transition rates k_{on} and k_{off} were inferred by from the measured current trace data by use of the autocorrelation function of the measured current trace. Here, we take a different approach. We consider only the lower-amplitudes of the current trace, the $T^{(1)}$ data, and derive its probability density function (PDF). Considering only the lower-amplitude data allows us to isolate the kinetic rates r_2, k_{on}, and k_{off}. We will show that the PDF of $T^{(1)}$ is a proper mixture of exponential modes and thus fits naturally into an expectation-maximization framework for finding the MLE estimates of the mixture parameters. The inferred mixture parameters are then mapped to the kinetic rates k_{on} and k_{off}. We will show that this method provides satisfactory results when tested against simulated data and is robust even when the observed $T^{(1)}$ data is subject to high measurement noise. The techniques used here to derive the PDF of $T^{(1)}$ and the subsequent setup of the EM framework can be used as a guide for inferring parameters from other Markov models using escape-time data.

We characterize the inference uncertainty by considering the total relative error which we define to be the sum of the relative errors of k_{on} and k_{off}. Using the observed Fisher information matrix, we can compute the inference uncertainty of the mixture parameters without the need for full Monte Carlo simulations. The inference uncertainty of the mixture parameters will then be propagated to the kinetic rates by a first-order Taylor expansion. To simplify our analysis, we will introduce scaling laws and show that the total relative error is a function of a scaled version of $[dNTP]$ and a scaled version of k_{off} only. Using this fact, we can build a table of which the total relative error of any k_{on} and k_{off} can be calculated from a priori. We mention that this table can be extended to compute the inference uncertainty for the dNTP binding, disassociation, and incorporation rates for synthesizing DNAP-DNA complexes. This will be covered in chapter 3.

We also discuss experimental design in finding the optimal $[dNTP]$. As will become evident in the Monte Carlo simulations and from the total relative error as a function of the scaled $[dNTP]$ and scaled k_{off}, there is a well defined minimum total relative error in the $[dNTP]$-direction for each k_{off}. As we will see, this optimal $[dNTP]$ may lead to long experimental run-times, so we extend the optimization problem to a constrained optimization problem in which the mean-field approximation to the experimental run-time is used for the constraint. We show numerically that using the mean-field approximation is justified.

Finally, we characterize the effect of multiplicative noise in the measured dwell time samples. We show that under multiplicative noise of the form $\exp(\sigma\zeta)$ where σ is the standard deviation and $\zeta \sim N(0,1)$, the effect on the inferred kinetic rates can be characterized exactly.

2.2 Mathematical Formulations

To derive the PDF of $T^{(1)}$, consider the general escape problem with state-space shown in figure 2.3. Note that for notational convenience, we have recycled the use of the rates r_1, r_2 and r_3; they are not related to the rates with the same name in the nucleotide addition cycle–they are any general transition rate. Let T be the time to escape to state 2 when the Markov process starts at state 1 at time $t = 0$. Let $X(t)$ be the state of the Markov process at time t. Define $p_j(t) = Pr(X(t) = j)$. We then have the initial conditions $p_1(0) = 1$ and $p_2(0) = p_3(0) = 0$.

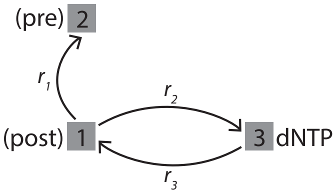

Figure 2.3: State-space diagram of a general escape problem of two transient states and one absorbing branch.

Proposition 1. *The PDF of T is of the form $\alpha\lambda_1 e^{-\lambda_1 t} + (1-\alpha)\lambda_2 e^{\lambda_2 t}$ with $0 < \alpha < 1$, $\lambda_1, \lambda_2 > 0$ and $\lambda_1 \neq \lambda_2$.*

Proof. By Kolmogorov's backwards equation, we have the following system of ODEs,

$$\frac{d}{dt}\begin{pmatrix} p_1 \\ p_3 \end{pmatrix} = \begin{pmatrix} -(r_1+r_2) & r_3 \\ r_2 & -r_3 \end{pmatrix}\begin{pmatrix} p_1 \\ p_3 \end{pmatrix}.$$

The characteristic polynomial is given by

$$f(\lambda) = \lambda^2 - (r_1 + r_2 + r_3)\lambda + r_1 r_3. \tag{2.1}$$

Hence solving $f(\lambda) = 0$ gives us the eigenvalues

$$\lambda_{1,2} = \frac{(r_1 + r_2 + r_3) \pm \sqrt{(r_1 + r_2 + r_3)^2 - 4r_1 r_3}}{2} \tag{2.2}$$

Now from the arithmetic mean-geometric mean (AM-GM) inequality, $(r_1 + r_3)/2 \geq$

$\sqrt{r_1 r_3}$ with equality if and only if $r_1 = r_3$. Thus we have that $(r_1 + r_3)^2 \geq 4r_1r_3$ and so $(r_1 + r_2 + r_3)^2 > 4r_1r_3$ since $r_1, r_2, r_3 > 0$. Also, by Descartes' Rule of Signs, both roots of the quadratic equation 2.1 are positive. Hence $p_j(t)$ is of the form of

$$p_j(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t}$$

Thus the total probability of the states 1 and 2 is $p_1(t) + p_3(t)$, which is of the form

$$p_1(t) + p_3(3) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t}$$

The PDF of the dwell time is given by

$$\begin{aligned} \rho(t) &= -\frac{d}{dt}\left(p_1(t) + p_3(t)\right) \\ &= c_1\lambda_1 e^{-\lambda_1 t} + c_2\lambda_2 e^{-\lambda_2 t} \end{aligned}$$

Now since $\lambda_1, \lambda_2 > 0$, we have that $1 = \int_0^\infty \rho\left(t\right) dt = c_1 + c_2$. Thus $c_1 = \alpha$ and $c_2 = 1 - \alpha$ for some $\alpha \in \mathbb{R}$. Hence we have

$$\rho(t) = \alpha\lambda_1 e^{-\lambda_1 t} + \left(1 - \alpha\right)\lambda_2 e^{-\lambda_2 t}$$

To determine α, we compare the value of $\rho\left(0\right)$ given by the expression above and the value based on the initial value problem,

$$\begin{aligned}\alpha\lambda_1 + (1-\alpha)\lambda_2 = \rho(0) = & \\ &= -\frac{d}{dt}\left(p_1(t)+p_3(t)\right)\Big|_{t=0} \\ &= -\frac{dp_1}{dt}(0) + \frac{dp_3}{dt}(0) \\ &= (r_1+r_2)\,p_1(0) - r_3p_3(0) - r_2p_1(0) + r_2p_3(0) \\ &= r_1p_1(0) \\ &= r_1\end{aligned}$$

The last equality is true since $p_1(0) = 1$ and $p_3(0) = 0$. Thus we have

$$\alpha\lambda_1 + (1-\alpha)\lambda_2 = r_1$$

Solving for α gives us

$$\alpha = \frac{\lambda_2 - r_1}{\lambda_2 - \lambda_1} \tag{2.3}$$

Now we show that $0 < \alpha < 1$. Without loss of generality, sort the two eigenvalues as $\lambda_1 < \lambda_2$, so that

$$\begin{aligned}\lambda_1 &= \frac{r_1+r_2+r_3-\sqrt{(r_1+r_2+r_3)^2 - 4r_1r_3}}{2}, \\ \lambda_2 &= \frac{r_1+r_2+r_3+\sqrt{(r_1+r_2+r_3)^2 - 4r_1r_3}}{2}.\end{aligned}$$

Note that

$$\begin{aligned}0 < \alpha < 1 &\Leftrightarrow 0 < \lambda_2 - r_1 < \lambda_2 - \lambda_1, \\ &\Leftrightarrow \lambda_1 < r_1 < \lambda_2.\end{aligned}$$

Hence it suffices to show that $\lambda_1 < r_1 < \lambda_2$.

Consider the quadratic equation given in equation 2.1. Notice that f satisfies $f(\lambda) > 0$ for $\lambda < \lambda_1$ and $\lambda > \lambda_2$. Also $f(\lambda) < 0$ for $\lambda_1 < \lambda < \lambda_2$. So we only need to show that $f(r_1) < 0$. Indeed, $f(r_1) = -r_1 r_2 < 0$. Thus we can conclude that $0 < \alpha < 1$.

□

2.3 Inference Method

From proposition 1, the PDF of $T^{(1)}$, $f_{T^{(1)}}$ is a mixture of two exponentials,

$$f_{T^{(1)}}(t) = \alpha \lambda_1 e^{-\lambda_1 t} + (1-\alpha)\lambda_2 e^{-\lambda_2 t}. \tag{2.4}$$

This gives us the mapping $(r_2, k_{on}, k_{off}) \mapsto (\alpha, \lambda_1, \lambda_2)$ with

$$\lambda_{1,2} = \frac{r_2 + k_{on}[dNTP] + k_{off} \pm \sqrt{(r_2 + k_{on}[dNTP] + k_{off})^2 - 4 r_2 k_{off}}}{2}. \tag{2.5}$$

The mixture weight α is given by $\alpha = (\lambda_2 - r_2)/(\lambda_2 - \lambda_1)$. The quantities $\theta := (\alpha, \lambda_1, \lambda_2)$ are referred to as the mixture parameters. We order the eigenvalues, $\lambda_1 < \lambda_2$. If r_2 and $[dNTP]$ are known, the mapping above is easily invertible; in fact, we can write

$$k_{off} = \frac{\lambda_1 \lambda_2}{r_2} \tag{2.6}$$

$$k_{on} = \frac{(1-\alpha)\lambda_1 + \alpha\lambda_2 - k_{off}}{[dNTP]}. \tag{2.7}$$

We will refer to the mapping in equations 2.6-2.7 as $K(\theta) = (k_{on}, k_{off})$.

The transition rate r_2 can be inferred from the $T^{(1)}$ data when $[dNTP] =$

0 [49]. When $[dNTP] = 0$, the DNAP-DNA complex transitions between the pre-translocation and post-translocation state. The dwell time of the lower-amplitude (which consist of only the post-translocation state) is a single exponential with rate r_2. Hence r_2 can be obtained from the $T^{(1)}$ data by using the fact that

$$\frac{1}{r_2} = \left\langle T^{(1)}\Big|_{[dNTP]=0} \right\rangle .$$

Also in the nanopore experiments, the $[dNTP]$ can be controlled accurately, and hence its value is assumed to be known. Hence for $[dNTP] > 0$, the mapping K defined above can be carried out in practice.

Since the distribution of $T^{(1)}$ is a proper mixture distribution, we can estimate the mixture parameters $\theta = (\alpha, \lambda_1, \lambda_2)$ using the expectation-maximization (EM) algorithm [13]. The mappings from equations 2.6 and 2.7 can then be used to obtain estimates for k_{on} and k_{off}.

Let $\theta = (\alpha, \lambda_1, \lambda_2)$. We denote $\theta^{(k)} = \left(\alpha^{(k)}, \lambda_1^{(k)}, \lambda_2^{(k)}\right)$ to be the k-th term in the EM sequence. Suppose that we observe $T_1^{(1)}, T_2^{(1)}, \ldots, T_n^{(1)} \overset{iid}{\sim} f_{T^{(1)}}$ where $f_{T^{(1)}}$ is the PDF of $T^{(1)}$ and iid means independently, identically distributed. Let $Z_1, \ldots, Z_n$ be the latent variable (hidden) that controls which exponential mode in $T_i^{(1)}$ is switched on in generating $T_i^{(1)}$,

$$T_i^{(1)} \mid \{Z_i = 1\} \sim \exp\left(t_i \mid \lambda_1\right),$$
$$T_i^{(1)} \mid \{Z_i = 0\} \sim \exp\left(t_i \mid \lambda_2\right),$$

where $\exp\left(t \mid \lambda\right)$ denotes the exponential distribution with rate λ. Note that $Z_i \sim \text{Bernoulli}\left(\alpha\right)$, so $Pr\left(Z_i = 1\right) = \alpha$, where $\text{Bernoulli}\left(p\right)$ is the Bernoulli distribution with probability of success p. The joint PDF of $\left(T_i^{(1)}, Z_i\right)$ is given by

$$\begin{aligned} f_{T_i^{(1)},Z_i}\left(t_i, z_i\right) &= f_{T_i^{(1)} \mid Z_i}\left(t_i \mid z_i\right) f_{Z_i}\left(z_i\right) \\ &= \left(\alpha\lambda_1 e^{-\lambda_1 t_i}\right)^{z_i} \left(\left(1-\alpha\right)\lambda_2 e^{-\lambda_2 t_i}\right)^{1-z_i}. \end{aligned}$$

The distribution of $Z_i \mid \left\{T_i^{(1)}, \theta\right\}$ is given by

$$\begin{aligned} Z_i \mid \left\{T_i^{(1)}, \theta\right\} &= \frac{Pr\left(Z_i, T_i^{(1)}\right)}{Pr\left(T_i^{(1)}\right)} \\ &\sim Bernoulli\left(\frac{\alpha\lambda_1 e^{-\lambda_1 t_i}}{\alpha\lambda_1 e^{-\lambda_1 t} + \left(1-\alpha\right)\lambda_2 e^{-\lambda_2 t_i}}\right), \end{aligned}$$

The complete data log-likelihood is given by

$$\begin{aligned} L\left(\theta \mid \left\{T^{(1)}, Z\right\}\right) &= \log f\left(\left\{T^{(1)}, Z\right\} \mid \theta\right) \\ &= \sum_{i=1}^{n}\left[z_i\left(\log\alpha + \log\lambda_1 - \lambda_1 t_i\right) + \left(1-z_i\right)\left(\log\left(1-\alpha\right) + \log\lambda_2 - \lambda_2 t_i\right)\right]. \end{aligned}$$

Note that Z_i is not in the data set. To use the above formulation to infer θ, we need to eliminate the hidden unknown Z_i. We accomplish this by taking the average based on the available value of θ from the previous iteration. Suppose that we have completed k iterations and $\theta^{(k)}$ is the most recent update on θ. The conditional expectation $\left\langle Z_i \mid T_i^{(1)}, \theta^{(k)}\right\rangle$ is given by

$$\beta_i^{(k)} := \left\langle Z_i \mid \left\{T_i^{(1)}, \theta^{(k)}\right\}\right\rangle = \frac{\alpha^{(k)}\lambda_1^{(k)} e^{-\lambda_1^{(k)} t_i}}{\alpha^{(k)}\lambda_1^{(k)} e^{-\lambda_1^{(k)} t_i} + \left(1-\alpha^{(k)}\right)\lambda_2^{(k)} e^{-\lambda_2^{(k)} t_i}}. \tag{2.8}$$

After taking the average, the result is a function of θ only, which we can then maximize to find the new updated approximation of θ. Hence

$$\begin{aligned}Q\left(\theta \mid \theta^{(k)}\right) &:= \left\langle L\left(\theta \mid \left\{T^{(1)}, Z\right\}\right)\right\rangle_{Z \mid \left\{T^{(1)}, \theta^{(k)}\right\}} \\ &= \left(\log \alpha + \log \lambda_1\right) \sum_{i=1}^{n} \beta_i^{(k)} - \lambda_1 \sum_{i=1}^{n} \beta_i^{(k)} t_i \\ &\qquad + \left(\log\left(1-\alpha\right) + \log \lambda_2\right) \sum_{i=1}^{n} \left(1-\beta_i^{(k)}\right) - \lambda_2 \sum_{i=1}^{n} \left(1-\beta_i^{(k)}\right) t_i.\end{aligned}$$

Hence the EM sequence is given by

$$\theta^{(k+1)} = \text{argmax}_{\theta} Q\left(\theta \mid \theta^{(k)}\right).$$

We can explicitly find the stationary points of Q:

$$\alpha^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \beta_i^{(k)}, \tag{2.9}$$

$$\lambda_1^{(k+1)} = \frac{\sum_{i=1}^{n} \beta_i^{(k)}}{\sum_{i=1}^{n} \beta_i^{(k)} t_i}, \tag{2.10}$$

$$\lambda_2^{(k+1)} = \frac{n - \sum_{i=1}^{n} \beta_i^{(k)}}{\sum_{i=1}^{n} \left(1-\beta_i^{(k)}\right) t_i}, \tag{2.11}$$

where $\beta_i^{(k)}$ are given in terms of $\left(\alpha^{(k)}, \lambda_1^{(k)}, \lambda_2^{(k)}\right)$ in equation 2.8. The analytical expressions in equations 2.9-2.10 provide an accurate and efficient way of calculating $\theta^{(k+1)}$ from $\theta^{(k)}$ and the dwell time data. Hence for each k, there is only one stationary point $\theta^{(k)} = \left(\alpha^{(k)}, \lambda_1^{(k)}, \lambda_2^{(k)}\right)^T$. For each k, taking second partial derivatives we find that

$$\begin{aligned}\frac{\partial^2 Q}{\partial \alpha^2} &= -\frac{1}{\alpha^2}\sum_{i=1}^{n}\beta_i^{(k)} - \frac{1}{\left(1-\alpha^{(k)}\right)^2}\sum_{i=1}^{n}\left(1-\beta_i^{(k)}\right),\\ \frac{\partial^2 Q}{\partial \lambda_1^2} &= -\frac{1}{\lambda_1^2}\sum_{i=1}^{n}\beta_i^{(k)},\\ \frac{\partial^2 Q}{\partial \lambda_2^2} &= -\frac{1}{\lambda_2^2}\sum_{i=1}^{n}\left(1-\beta_i^{(k)}\right),\end{aligned}$$

with all the mixed partial derivatives equal to 0. The Hessian matrix of Q is thus a diagonal matrix with entries $\operatorname{diag}\left(Q_{\alpha\alpha}, Q_{\lambda_1\lambda_2}, Q_{\lambda_2\lambda_2}\right)$. Note that for each k and i, $0 < \beta_i^{(k)} < 1$ so the diagonal elements of Q are negative, and hence Q is negative definite. This implies that the stationary point of Q given in equations 2.9-2.10 is a global maximum. It can then be shown that the EM sequence converges to the MLE of $(\alpha, \lambda_1, \lambda_2)$ for any initial guess $\theta^{(0)}$ [72].

2.4 Inference on Simulated Samples of Dwell Times

In this section, we conduct some numerical simulations to determine the validity of using the MLE method to infer k_{on} and k_{off} from $T^{(1)}$ data.

The following numerical simulation was done as follows. The random variable $T^{(1)}$ was sampled 10,000 times and this data was used to obtain MLE estimates for k_{on} and k_{off} This was then repeated 10,000 times to obtain a distribution for the MLE estimates.

The MLE estimates were centered with respect to their true values and then normalized by their true values; we use the notation

$$\operatorname{err}(k) := \frac{k^{MLE} - k^{true}}{k^{true}},$$

to denote the centered and normalized error of MLE estimates. The standard deviation off err (k) is also recorded, and we denote this by std $(\text{err}\,(k))$. In the simulations, we set the true values of k_{on} and k_{off} to be $k_{on} = 200$ and $k_{off} = 100$. For r_2, we set this rate to $r_2 = 100$. The rate r_2 can be determined from $T^{(1)}$ data when $[dNTP] = 0$ [49]. Thus for simplicity, we assume that this rate is known. The dNTP concentration can be controlled accurately in the nanopore experiments, so its value is also assumed to be known. The results from this numerical simulation are displayed in figures 2.4-2.5. They show that the inference accuracy is good and consistent over a wide range of dNTP concentrations. The accuracy is slightly better around $[dNTP] = 2$, as indicated by the smallest std(err) when compare to other dNTP concentrations. Since we use $k_{on} = 200$ and $k_{off} = 100, [dNTP] = 2$ corresponds to a slightly high concentration of $[dNTP]/K_d = 4$, where K_d is the dissociation constant defined to be $K_d = k_{off}/k_{on}$. For both transition rates, the bias is small.

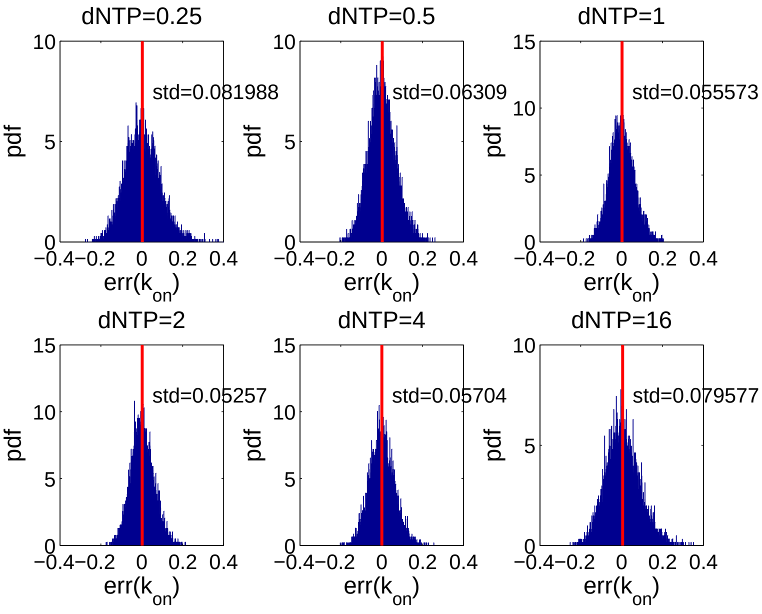

Figure 2.4: MLE results for k_{on} with no noise in $T^{(1)}$ observations.

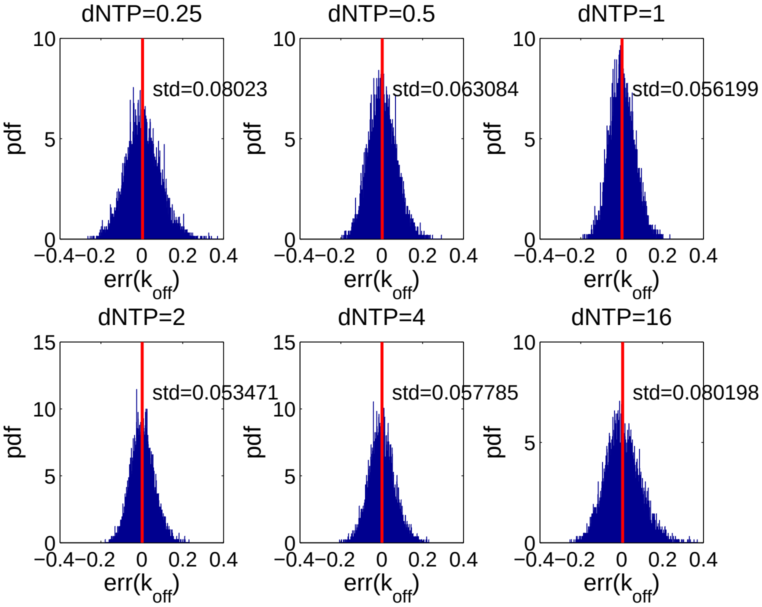

Figure 2.5: MLE results for k_{off} with no noise in $T^{(1)}$ observations.

2.5 Dependence of Inference Uncertainty Model Parameters

From the proceeding section, it is clear that the relative errors for k_{on} and k_{off} are dependent on the dNTP concentration. Even with no noise in the $T^{(1)}$ observations, the relative errors for k_{on} and k_{off} ranges from about 5%-8%, for the relatively small amount of dNTP concentrations that we tested. The error can of course, be a lot worse if a highly suboptimal $[dNTP]$ is chosen; for example, at $[dNTP] = 0.07$, the relative errors for k_{on} and k_{off} are about 16%-18%. The error in estimating k_{on} and k_{off} originate from the inference of the mixture parameters $\theta = (\alpha, \lambda_1, \lambda_2)$ in equation 2.4 by the EM algorithm. The inference uncertainty in θ is then propagated to k_{on} and k_{off} by the inverse mappings given in equations 2.6 and 2.6.

In order to control the error by tuning the dNTP concentration, we need to know the inference uncertainty of k_{on} and k_{off} a priori. We characterize the inference uncertainty of k_{on} and k_{off} by considering the total relative error. We define the total relative error as the sum of the relative errors of k_{on} and k_{off}. In the parameter regimes where the total relative error is small, the variance of the MLE estimates is a good approximation to the root-mean-squared due to the small inference bias in these regimes. Hence, we can write the total relative error as the sum of the standard deviations of $\text{err}\left(k_{on}\right)+\text{err}\left(k_{off}\right)$. Practically speaking, it is the small inference error regimes that are more useful for experimental design, so we only concern ourselves with approximating the total relative error in the small inference error regimes accurately. We will first show a way to obtain the total relative error without the need for full scale Monte Carlo simulation. Then we will show that the total relative error is a function of a scaled version of $[dNTP]$ and a scaled version of k_{off}.

2.5.1 Calculating of the Total Relative Error of the Kinetic Rates

Let $T_1^{(1)}, T_2^{(1)}, \ldots, T_n^{(1)}$ be a random sample from $f_{T^{(1)}}$, where $f_{T^{(1)}}$ is the PDF of the lower-amplitude segment given in equation 2.4. Let

$$L\left(\theta \mid t\right)=\sum_{i=1}^{n}\log\left(\alpha e^{-\lambda_1 t_i}+\left(1-\alpha\right)e^{-\lambda_2 t_i}\right) \tag{2.12}$$

be the log-likelihood function of the $T^{(1)}$ data, and let $\theta^{\text{MLE}}=\left(\alpha^{\text{MLE}},\lambda_1^{\text{MLE}},\lambda_2^{\text{MLE}}\right)$ be the MLE estimates of θ. The observed Fisher information matrix, H, defined by

$$H\left(\theta^{\mathrm{MLE}} \mid t\right) = -\begin{pmatrix} L_{\alpha,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\alpha,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\alpha,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \\ L_{\lambda_1,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_1,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_1,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \\ L_{\lambda_2,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_2,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_2,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \end{pmatrix}, \tag{2.13}$$

where we denote $L_{x,y}$ to mean $L_{x,y} = \partial^2 L / \left(\partial x \partial y\right)$. Note that equation 2.13 is the negative Hessian of L evaluated at the MLE estimates. It has been demonstrated that the inverse of H gives an approximation to the asymptotic covariance matrix of the MLE estimates of θ as the number of samples of $T^{(1)}$, $n \to \infty$ [24]. Hence for large n, $\mathrm{Cov}\left(\theta\right) \approx H^{-1}$.

We can propagate the inference uncertainty of the mixture parameters θ to k_{on} and k_{off} by a first-order Taylor expansion. Recall that we have the mapping $\theta \mapsto \left(k_{on}\left(\theta\right), k_{off}\left(\theta\right)\right)^T$ according to equations 2.6 and 2.7. Let $K\left(\theta\right) = \left(k_{on}\left(\theta\right), k_{off}\left(\theta\right)\right)^T$ be this mapping.

Consider the first-order Taylor expansion,

$$K\left(\theta\right) = K\left(\theta^{\mathrm{MLE}}\right) + J\left(\theta^{\mathrm{MLE}}\right)\left(\theta - \theta^{\mathrm{MLE}}\right) + o\left(\left\|\theta - \theta^{\mathrm{MLE}}\right\|\right),$$

where $J\left(\theta^{\mathrm{MLE}}\right)$ is the Jacobian of K evaluated at θ^{MLE}. Now

$$\begin{aligned} Cov\left(K\left(\theta\right)\right) &= Cov\left[K\left(\theta^{\mathrm{MLE}}\right) + J\left(\theta^{\mathrm{MLE}}\right)\left(\theta - \theta^{\mathrm{MLE}}\right) + o\left(\left\|\theta - \theta^{\mathrm{MLE}}\right\|\right)\right] \\ &= Cov\left(J\left(\theta^{\mathrm{MLE}}\right)\theta\right) \\ &= J\left(\theta^{\mathrm{MLE}}\right) Cov\left(\theta\right) J\left(\theta^{\mathrm{MLE}}\right)^T, \end{aligned}$$

where $Cov\left(\theta\right)$ is the covariance matrix of θ. Recall that $Cov\left(\theta\right)$ is approximated by H^{-1} where H is the observed information matrix (equation 2.13). The second equality follows since $K\left(\theta^{\text{MLE}}\right)$, θ^{MLE}, and $o\left(\left\|\theta-\theta^{\text{MLE}}\right\|\right)$ are constant vectors.

The result is that the diagonal entries of the covariance matrix $Cov\left(K\left(\theta\right)\right)$ are the asymptotic estimates of the variance of the MLE estimates of k_{on} and k_{off}. Using this, we can estimate the relative error of the MLE estimates of k_{on} and k_{off} a priori without the computational effort of full Monte Carlo simulations.

A useful metric which will guide our study of the inference uncertainty of k_{on} and k_{off} is the total relative error. We approximate the total relative error as $\text{std}\left(\text{err}\left(k_{on}\right)\right)+\text{std}\left(\text{err}\left(k_{off}\right)\right)$. In parameter regimes in which the total relative error is small, this is an adequate approximation. We can readily approximate the total relative error without the computational effort of Monte Carlo simulations by using the estimates for the relative error of the MLE estimates of k_{on} and k_{off} derived above in the following way,

$$\text{std}\left(\text{err}\left(k_{on}\right)\right)+\text{std}\left(\text{err}\left(k_{off}\right)\right)\approx\left\|\sqrt{\text{diag}\left(Cov\left(K\right)\right)}\odot\left(\frac{1}{k_{on}},\frac{1}{k_{off}}\right)^{T}\right\|_{1}, \quad (2.14)$$

where $\text{diag}\left(A\right)$ is the vector containing the diagonal entries of the matrix A, $A\odot B$ is the element-wise multiplication of the matrices A and B, and $\left\|\cdot\right\|_1$ denotes the Euclidean 1-norm. The square-root operator is taken to be applied element-wise on the entries of $\text{diag}\left(Cov\left(K\right)\right)$.

2.5.2 Characterizing the Total Relative Error

We will show that the total relative error in equation 2.14 is a function of k_{off} and $[dNTP]$ only.

Consider the following time-scaling result for a mixture of exponentials

Proposition 2. *Let $Z=\beta T^{(1)}$. Then the PDF of Z is given by $f_Z\left(z\right)=\frac{1}{\beta}f_{T^{(1)}}\left(\frac{z}{\beta}\right)$ where $f_{T^{(1)}}$ is the PDF of $T^{(1)}$.*

Proof. The PDF of Z is given by $\frac{d}{dz}Pr\left(z \leq Z\right) = \frac{d}{dz}Pr\left(\frac{z}{\beta} \leq T\right) = \frac{1}{\beta}f_{T^{(1)}}\left(\frac{z}{\beta}\right)$ by the chain rule and the fact that the derivative of the cumulative distribution function is the PDF. □

The consequence of Proposition 2 is that the scaled random variable $\beta T^{(1)}$ has PDF

$$\alpha\frac{\lambda_1}{\beta}e^{-\frac{\lambda_1}{\beta}t} + (1-\alpha)\frac{\lambda_2}{\beta}e^{-\frac{\lambda_2}{\beta}t}, \tag{2.15}$$

and hence is still a proper exponential mixture with mixture parameters $(\alpha, \lambda_1/\beta, \lambda_2/\beta)$. Note that since $\alpha = (\lambda_2 - r_2)/(\lambda_2 - \lambda_1)$, we have $r_2 = (1-\alpha)\lambda_2 + \alpha\lambda_1$. This and equations 2.6 and 2.7 gives us the mapping $(\alpha, \lambda_1/\beta, \lambda_2/\beta) \to (r_2/\beta, k_{on}/\beta, k_{off}/\beta)$. Hence $r_2 T_1^{(1)}$ gives us the scaling mapping $(r_2, k_{on}, k_{off}) \mapsto (1, k_{on}/r_2, k_{off}/r_2)$. An intuitive way to think about this is that the transition rates has units $[\text{time}]^{-1}$ and thus we can rescale time in such a way that $r_2 \mapsto 1$. Throughout the rest of this paper, we denote the scaled k_{off}/r_2 as $k := k_{off}/r_2$.

By the same reasoning, we can also scale $[dNTP]$ since $[dNTP]$ has units of concentration. Scaling $[dNTP]$ by k_{on}/r_2 gives us $S := k_{on}/r_2[dNTP]$. Hence after scaling, we have the following state-space diagram shown in figure 2.6, with scaled rates

$$\begin{aligned} r_2' &= r_2/r_2 = 1, \\ k_{on}' &= 1, \\ [dNTP]' &= S := k_{on}/r_2[dNTP], \\ k_{off}' &= k := k_{off}/r_2. \end{aligned}$$

Under the scaling, we are free to choose $k_{on}' = 1$.

From figure 2.6, we can conclude that the inference uncertainty of k_{on} and k_{off} is a function of S and k only. Hence in all of the following analysis, we can set

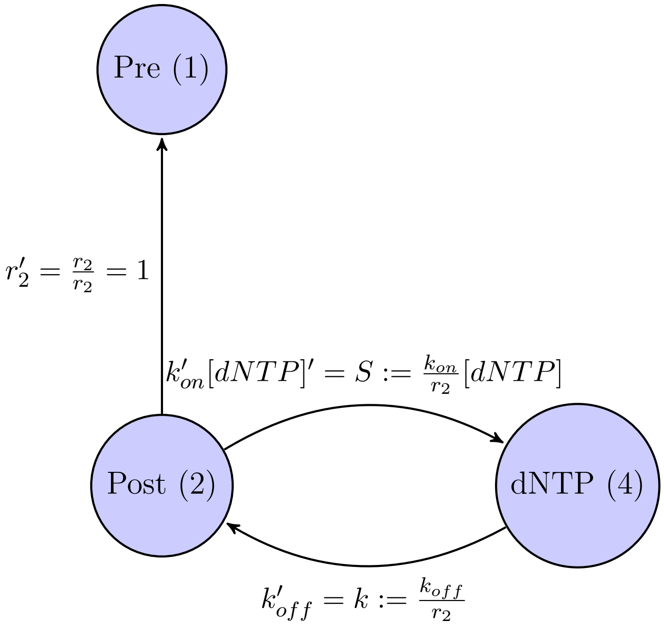
Pre (1)
$r_2' = \frac{r_2}{r_2} = 1$
$k_{on}'[dNTP]' = S := \frac{k_{on}}{r_2}[dNTP]$
Post (2)
dNTP (4)
$k_{off}' = k := \frac{k_{off}}{r_2}$

Figure 2.6: State-space diagram of the lower-amplitude segment of T_A after scaling.

$r_2 = k_{on} = 1$ and we can write equation 2.14 as

$$\operatorname{err}(S, k, n) = \left\| \sqrt{\operatorname{diag}\left(Cov\left(K\left(\theta\left(S, k, n\right)\right)\right)\right)} \odot \left(\frac{1}{k_{on}}, \frac{1}{k_{off}}\right)^T \right\|_1 \tag{2.16}$$

where K is the mapping $\theta \mapsto (k_{on}, k_{off})$, $\theta = (\alpha, \lambda_1, \lambda_2)$, and n is the number of $T^{(1)}$ samples. Note that the scaled dNTP concentration S, the scaled dNTP disassociation rate k, and the number of samples of $T^{(1)}$ n affect the mixture parameters θ (equation 2.5). We thus write $\theta\left(S, k, n\right)$ to emphasize θ's dependance on S, k, and n.

The total relative error function $\operatorname{err}\left(S, k\right)$ is very difficult to compute analytically. To numerically build the total relative error function, we discretize S and k over a range of values. Let $\mathcal{S}$ and $\mathcal{K}$ be the set of discrete points for S and k respectively. Enumerate the elements of $\mathcal{S} = \{S_1, \ldots, S_m\}$, where m is the number of S points used. At each $(S,k) \in \mathcal{S} \times \mathcal{K}$ point we sample $f_{T^{(1)}}$ $n_0 = 10,000$ times and estimate k_{on} and k_{off} using the EM method. The total relative error is then estimated by using equation 2.16. This is repeated 20 times for each $(S,k) \in \mathcal{S} \times \mathcal{K}$, giving us a cloud of total relative error data for each (S,k) point.

Let $\mathcal{E}_C\left(S,k\right)$ be the 20-point data cloud at the point (S,k). We then estimate the total relative error by fitting a quadratic polynomial in the S-direction using 41 points in the least squares sense in the following way.

Let $S_i \in \mathcal{S}$. Define the following subset of $\mathcal{S}$,

$$\mathcal{S}_i = \begin{cases} \{S_1, \ldots, S_{41}\} & \text{if } i < 21 \\ \{S_{m-40}, \ldots, S_m\} & \text{if } i > m - 20 \\ \{S_{i-20}, \ldots, S_{i+20}\} & \text{otherwise} \end{cases} \cdot$$

Here, $\mathcal{S}_i$ is selected to consist of 41 points entered around S_i with the index range shifted if necessary to be contained in $\mathcal{S}$. For each $k \in \mathcal{K}$, we do the following: for each $i = 1, \ldots, m$, a quadratic polynomial $P_{i,k}$ is fit to the set of points

$$\log\left(\mathcal{S}_i\right) \times \log\left(\bigcup_{S \in \mathcal{S}_i} \mathcal{E}_C\left(S,k\right)\right),$$

in the least squares sense where the logarithm function is understood to be taken over all the elements of the set; that is, $\log \mathcal{A} = \{\log a \ : \ a \in A\}$.

Since we are using $n_0 = 10000$ samples to build the numerical approximation to the total relative error along a grid of S and k points, define err_1 to be the function

$$\mathrm{err}_1\left(S,k\right) := \mathrm{err}\left(S,k,n\right)\Big|_{n=n_0}.$$

Here, err_1 is a function of only (S,k).

Then $\mathrm{err}_1(S,k) = P_{i,k}(S_i)$ is set to be the point-estimate of the total relative error for $k'_{on} = 1$ and k at (S_i, k). We use the log of the data for the local least squares fit since qualitatively the data is approximately quadratic on the log-scale.

After this procedure, a discrete grid of point-estimates for the total relative error of k_{on} and k_{off} using 10,000 $T^{(1)}$ samples is obtained:
$\mathcal{E} = \{\mathrm{err}_1(S,k) \;:\; (S,k) \in \mathcal{S} \times \mathcal{K}\}$. Using linear interpolation on $\mathcal{E}$, we can then compute err_1, for any S and k pair a priori. The resulting total relative error surface is shown in figure 2.7.

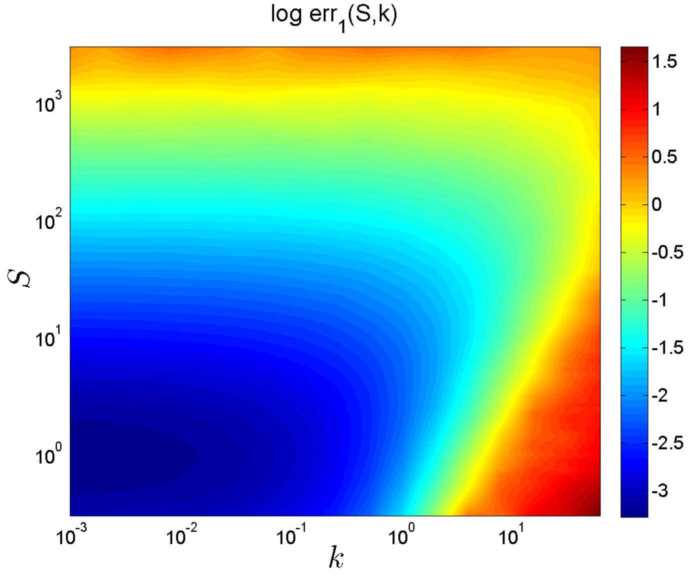

Figure 2.7: The total relative error surface $\mathrm{err}_1(S,k)$ by local quadratic polynomial least-squares.

The constructed total relative error function $\mathrm{err}_1(S,k)$ provides a good estimate to the total relative error of k_{on} and k_{off}. To show this, we re-sample the cloud of data $\mathcal{E}_C$ at each (S,k) point 1000 times to obtain the uncertainty of the total relative error estimate of k_{on} and k_{off} at each point. From figure 2.8, we see that the uncertainty of the total relative error of k_{on} and k_{off} is small and grows approximately proportional to the inference uncertainty of k_{on} and k_{off}. For each (S,k)-point, the covariance matrix of the MLE estimates of the mixture

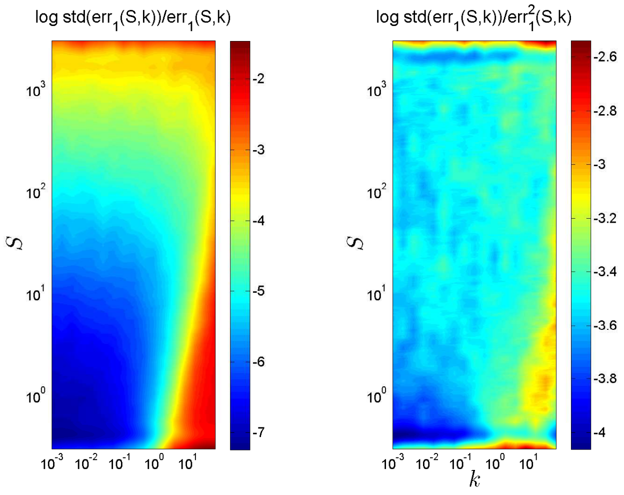

Figure 2.8: The left panel shows the uncertainty of the total relative error estimate of k_{on} and k_{off} as produced by boot-strap resampling of the cloud of 20 data points $\mathcal{E}_C$ at each (k, S) point. The right panel shows the top quantity divided by err_1.

parameters $(\alpha, \lambda_1, \lambda_2)$ is also saved. In doing so, we can easily extend this table to the $k_{pol} > 0$ case since conditioning on the escape to the pre-translocation state when $k_{pol} > 0$ forms an escape problem governing $T^{(1)}$ which is in the same form as the $k_{pol} = 0$ case. This will be discussed in chapter 3.

2.6 Optimum Experimental Condition

In this section, we examine the optimal experimental condition that when achieved, produces the least total relative error.

2.6.1 Finding the Optimal $[dNTP]$

From the scaling laws and the total relative error point estimates in $\mathcal{E}$, we can numerically obtain the $[dNTP]$ that yields the least total relative error for any k_{on} and k_{off}.

Let $[dNTP]^*$ denote the optimum dNTP concentration–optimum in the sense that it produces the least total relative error according to equation 2.16. From figure 2.6, we see that after scaling, the scaled optimum dNTP concentration S^* is a function of only k. Hence we can write,

$$S^* = F\left(k\right) \Leftrightarrow [dNTP]^* = \frac{r_2}{k_{on}} F\left(\frac{k_{off}}{r_2}\right). \tag{2.17}$$

Determining an expression for F analytically is very difficult, so we instead turn to a numerical approximation. For fixed k, the total relative error is locally quadratic in the log-scale around the minimum (figure 2.9).

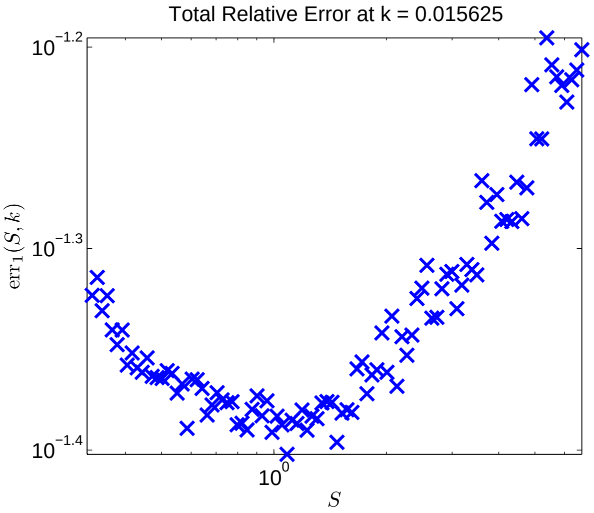

Figure 2.9: For fixed k, the total relative error is approximately quadratic near the minimum. This is a typical graph of $\mathrm{err}_1\left(S,k\right)$ with k fixed.

For fixed $k \in \mathcal{K}$, we approximate S^* by using the smoothing quadratic polynomial $P_{i,k}$ where i is any i such that $\mathrm{err}_1\left(S_i,k\right)$ is near the minimum for that fixed k. The minimum of the chosen $P_{i,k}$ is the approximated value for S^*. This can be extended for any arbitrary k by linear interpolation of the error grid $\mathcal{E}$. Figure 2.10 shows the approximation of the optimal S^* trajectory on the total relative error surface $\mathrm{err}_1\left(S,k\right)$. The trajectory $k \mapsto S^*$ provides a numerical approximation to F in equation 2.17.

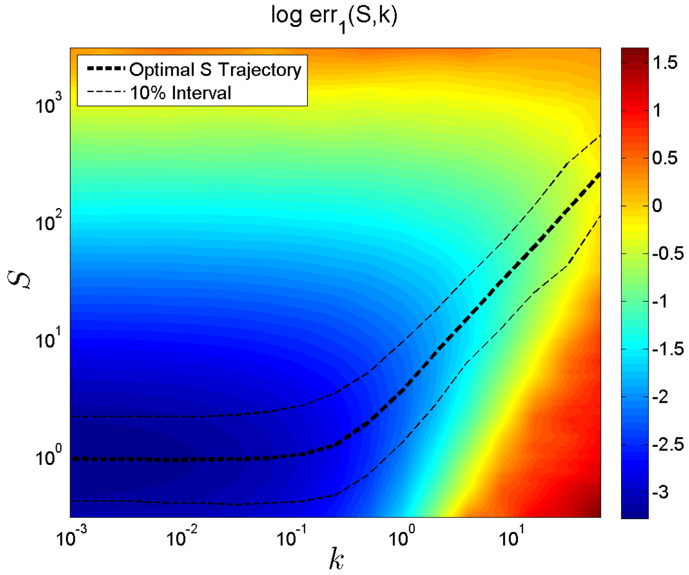

Figure 2.10: The total relative error surface $\mathrm{err}_1(S,k)$ with the estimated optimal S^* trajectory. For each k, the 10% error interval shown as black-dashed lines were obtained by finding the two S points such that $\mathrm{err}_1(S,k) = 1.1\mathrm{err}(S^*,k)$

For convenience, we also plot the total relative error along the optimal $[dNTP]$ trajectory (figure 2.11).

2.6.2 Behavior of the Minimum Total Relative Error

From figure 2.11, we see that the total relative error along the optimal S trajectory increases monotonically as k increases. It is also evident that there is no well defined least minimum total relative error as a function of k.

This can be intuitively explained as follows. When $k \to 0$, the post-translocation and dNTP-bound states become more "separated." That is, the dwell time of the dNTP-bound state increases as $k \to 0$. From figure 2.10, as $k \to 0$, $S^* \to 1$. In this region, when $S^* \to 1$, the probability of escape to the pre-translocation state and the probability dNTP binding approach each other. This means that the dwell time for the post-translocation state approaches the largest it can be.

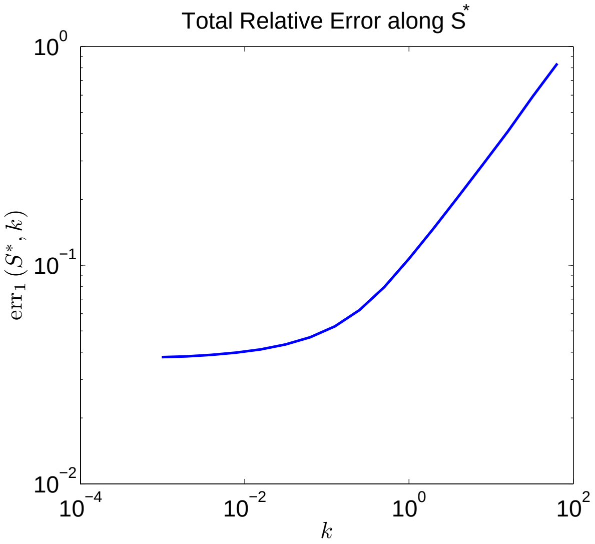

Figure 2.11: The total relative error along the optimal S^* trajectory.

Hence as $k \to 0$, the post-translocation and dNTP-bound states approach its largest separation and hence the total relative error approaches its infimum as $k \to 0$; that is, $\lim_{k\to 0} \mathrm{err}(S^*, k, n) = \inf_k \mathrm{err}(S^*, k, n)$ for fixed n.

The minimum total relative error approximately increases by an order of magnitude from $\inf_k \mathrm{err}(S^*, k, n)$ for $k > 50$. For k large, the optimal S^* is proportional to k. In this region, the total relative error is large since the post-translocation and dNTP-bound states are approximately in equilibrium. In this setting, the post-translocation and dNTP-bound states form a superstate, and resolution of the two exponential modes is very difficult.

2.6.3 Behavior of the Optimal $[dNTP]$

We now examine the behavior of the optimal $[dNTP]$ obtained from figure 2.10. Some immediate observations we can make are that as $k \to 0$, $S^* \to 1$, and for k larger than about 0.5, S^* increases proportional to k. To gain insight into this behavior, we investigate asymptotic cases for k and S.

Behavior of the Total Relative Error of the Mixture Parameters

Before diving into the asymptotic cases for k and S, we first take a digression to the behavior of the total relative error of the mixture parameters $(\alpha, \lambda_1, \lambda_2)$. The total relative error of the mixture parameters is defined to be the sum of the relative errors of α, λ_1, and λ_2. Recall the scaling law in proposition 2. We can scale the $T^{(1)}$ samples by λ_2, thereby obtaining the equivalent mixture parameters $(\alpha, \lambda_1/\lambda_2, 1)$. Hence the total relative error of the mixture parameters is a function of α and λ_1/λ_2 only.

Figure 2.12 shows a contour plot of the total relative error of the mixture parameters, where the total relative error was calculated from the covariance matrices obtained from the observed Fisher information matrix based on $n_0 = 10000$ samples of $T^{(1)}$. At each $(\alpha, \lambda_1/\lambda_2)$-point, this was repeated 40 times. The resulting total relative error data was put through the same quadratic polynomial smoothing algorithm in the α-direction described in section 2.5.2.

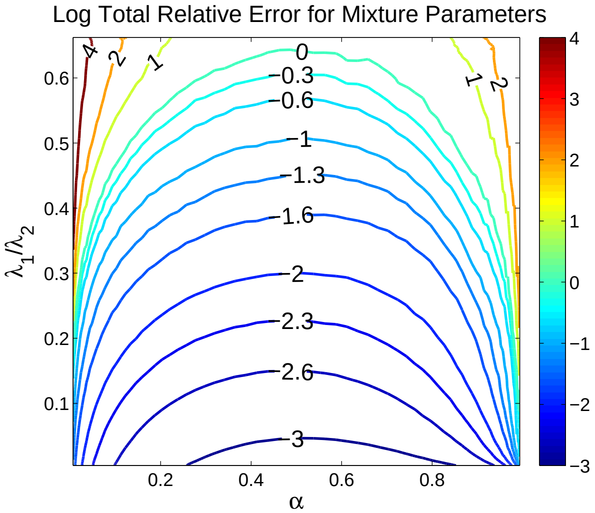

Figure 2.12: The total relative error of the mixture parameters $(\alpha, \lambda_1, \lambda_2)$ after quadratic polynomial smoothing. The total relative error increases along the boundary.

The error changes rapidly along the boundaries, so when generating the total relative error of the mixture parameters, we increases the resolution along the boundaries of $\alpha \approx 0$ and $\alpha \approx 1$. Although the upper limit of λ_1/λ_2 is 1, we stopped the simulation at $\lambda_1/\lambda_2 \approx 2/3$ since the total relative error was already very high in this region. To increase the resolution along the aforementioned boundaries, we generate a linear grid $\beta_\alpha \times \beta_{\lambda_1/\lambda_2}$ where β_α and $\beta_{\lambda_1/\lambda_2}$ consist of equally spaced points centered around 0. The following nonlinear mapping was applied,

$$\frac{\lambda_1}{\lambda_2} = \frac{\frac{2}{3}e^{\beta_{\lambda_1/\lambda_2}}}{1+e^{\beta_{\lambda_1/\lambda_2}}}, \tag{2.18}$$

$$\alpha = \frac{e^{\beta_\alpha}}{1+e^{\beta_\alpha}}, \tag{2.19}$$

to generate the $(\alpha, \lambda_1/\lambda_2)$-grid. This generates a non-uniform grid with more points concentrated along the boundaries $\alpha \approx 0, 1$ and $\lambda_1/\lambda_2 \approx 0, 2/3$.

The error plot in figure 2.12 confirm our intuition that the inference uncertainty for the mixture parameters increase as α gets close to 0 or 1 and as λ_1 and λ_2 approach each other. Recall the expression for the $T^{(1)}$ PDF, $\alpha\lambda_1 \exp(-\lambda_1 t) + (1-\alpha)\lambda_2 \exp(-\lambda_2 t)$. When $\alpha \approx 0$, the population of the faster exponential mode $(1-\alpha)\lambda_2 \exp(-\lambda_2 t)$ is substantially larger than the slower exponential mode $\alpha\lambda_1 \exp(-\lambda_1 t)$. In this situation, the inference uncertainty for λ_1 will be large. The opposite is true if $\alpha = 1$. In this situation, since the population of the lower exponential mode is substantially larger than the population of the faster exponential mode, the inference uncertainty for λ_2 will be large. A closer inspection of figure 2.12 reveals that the surface is not symmetric about $\alpha = 0.5$. In fact, the inference uncertainty for λ_2 is generally larger than the inference uncertainty for λ_1 for an equivalent distance of α from 1 and 0; i.e., let

α_1 and α_2 be values of α such that $0 < |1-\alpha_2| = |\alpha_1| << 1$, then the inference uncertainty for λ_2 is greater when $\alpha = \alpha_2$ than the inference uncertainty for λ_1 when $\alpha = \alpha_1$. This can be more easily seen when we plot the total relative error on the linear grid $\beta_\alpha \times \beta_{\lambda_1/\lambda_2}$ (figure 2.13). The reason for this is when $\alpha \approx 1$, not only is the population for the faster exponential mode substantially smaller than the population for the slower exponential mode, but the faster exponential mode decays faster than the slower exponential mode, further increasing the inference uncertainty for λ_2. Finally when $\lambda_1 \approx \lambda_2$, the two exponential modes are nearly indistinguishable and hence the inference uncertainty for α is increased.

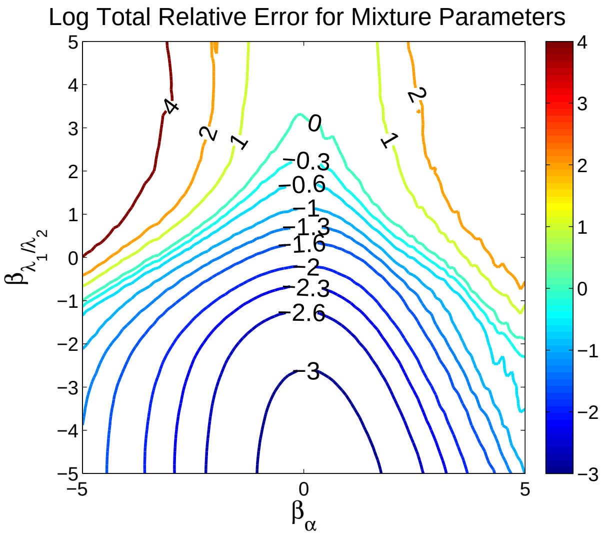

Figure 2.13: The total relative error of the mixture parameters $(\alpha, \lambda_1, \lambda_2)$ after quadratic polynomial smoothing plotted on the linear grid $\beta_\alpha \times \beta_{\lambda_1/\lambda_2}$.

Behavior of the Optimal $[dNTP]$: Asymptotic Studies of S and k

We know exmaine the behavior of the optimal $[dNTP]$. To gain some insight, we investigate asymptotic cases for k and S. In all of the following asymptotic analysis, let $0 < \epsilon << 1$ be a small parameter, and $a, b = O(1)$. Each of the following asymptotic cases are shown schematically in figure 2.14, with C1 referring to case 1, C2 referring to case 2, etc.

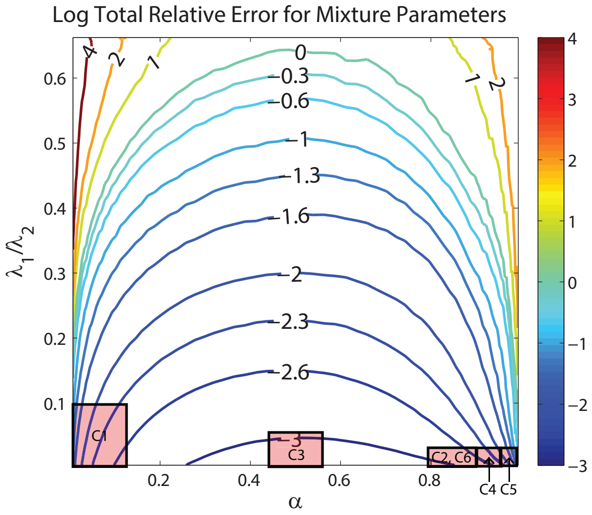

Figure 2.14: The total relative error of the mixture parameters $(\alpha, \lambda_1, \lambda_2)$ after quadratic polynomial smoothing plotted on the linear grid with schematic locations of the asymptotic regions for S and k.

- Case 1: $k = a\epsilon$ and $S = b\epsilon$.

$$\lambda_1 = a\epsilon + O\left(\epsilon^2\right),$$

$$\lambda_2 = 1 + b\epsilon + O\left(\epsilon^2\right),$$

$$\alpha = a\epsilon + O\left(\epsilon^2\right).$$

In this case, both k and S are small. When k is small, if the complex transitions to the dNTP-bound state, the complex will remain in that state for a long time. When S is small, then the complex has a high probability of immediately escaping to the pre-translocation state from the post-translocation state; hence the dNTP-bound state is visited less. This results in one of the exponential modes being sampled less. Indeed from the above asymptotic expansions for the mixture parameters, even though the exponential rates are well separated, $\alpha \approx 0$, resulting in the slow exponential with rate λ_1 being sampled less. This leads to a higher total relative error.

- Case 2: $k = a\epsilon$ and $S = b/\epsilon$.

$$\begin{aligned}\lambda_1 &= O\left(\epsilon^2\right), \\ \lambda_2 &= \frac{b}{\epsilon} + 1 + a\epsilon + O\left(\epsilon^2\right), \\ \alpha &= 1 - \frac{1}{b}\epsilon + O\left(\epsilon^2\right).\end{aligned}$$

Like the previous case, k is small so that if the complex transitions to the dNTP-bound state, the complex will remain in that state for a long time. When S is large, the complex has a high probability of immediately binding a dNTP. This results in a small dwell time for the post-translocation state. At the same time, the dwell time in the dNTP-bound state is large since k is small. This results in one of the exponential modes being sampled less. Indeed, from the asymptotic expansions of the mixture parameters, $\alpha \approx 1$ and hence the fast exponential mode is sampled less, increasing the total relative error.

- Case 3: $k = a\epsilon$ and $S = 1 + b\epsilon$.

$$\begin{aligned}\lambda_1 &= \frac{a}{2}\epsilon + O\left(\epsilon^2\right), \\ \lambda_2 &= 2 + \left(\frac{a}{2} + b\right)\epsilon + O\left(\epsilon^2\right), \\ \alpha &= \frac{1}{2} + \frac{a+b}{4}\epsilon + O\left(\epsilon^2\right).\end{aligned}$$

According to figure 2.10, for small k, the optimal $S^* \approx 1$. At $S = 1$, the probability of escape to the pre-translocation state and the probability of escape to the dNTP-bound state are equal. In this case, when the complex binds to a dNTP, it will remain there for a long time. When the dNTP disassociates, the complex has an equal proability of escaping to the pre-translocation state or binding another dNTP. Here, the dwell time in the post-translocation state is the longest it can be in this kinetic region, resulting in the least total relative error for small k. This can be seen from the asymptotic expansions of the mixture parameters. Here, not only are the exponential rates well separated, the parameters α is close to 1/2, resulting in equal sampling of both exponential modes. This results in the lowest total relative error for the mixture parameters (figure 2.13).

- Case 4: $k = \frac{a}{\epsilon}$ and $S = b\epsilon$.

$$\begin{aligned}
\lambda_1 &= 1 + O\left(\epsilon^2\right), \\
\lambda_2 &= \frac{a}{\epsilon} + b\epsilon + O\left(\epsilon^2\right), \\
\alpha &= 1 + O\left(\epsilon^2\right).
\end{aligned}$$

When k is large and when the complex transitions to the dNTP-bound state, the complex will transition back to the post-translocation state very quickly so that the dwell time in the dNTP-bound state is very short. If S is small, the complex quickly escapes to the pre-translocation state without ever visiting the dNTP-bound state. Any (rare) visit to the dNTP-bound state is quicky transitioned back to the post-translocation state and then back to the pre-translocation state. Here, $\alpha \approx 1$ resulting in the fast exponential mode being poorly sampled resulting in a high total relative error for the mixture parameters.

- Case 5: $k = \frac{a}{\epsilon}$ and $S = \frac{b}{\epsilon^2}$.

  $$\begin{aligned}\lambda_1 &= \frac{a}{b}\epsilon + O\left(\epsilon^2\right),\\ \lambda_2 &= \frac{b}{\epsilon^2} + \frac{a}{\epsilon} + 1 - \frac{a}{b}\epsilon + O\left(\epsilon^2\right),\\ \alpha &= 1 - \frac{\epsilon^2}{b} + O\left(\epsilon^3\right).\end{aligned}$$

  Like before, if k is large, when the complex transitions to the dNTP-bound state, the complex will transition back to the post-translocation state very quickly so that the dwell time of the dNTP-bound state is very short. When S is much larger than k, the complex will bind a dNTP very quickly. Since k is also large, both of the dwell times of the post-translocation and dNTP-bound states are very small. Hence in this regime, the post-translocation and dNTP-bound states are in equilibrium and resolving the exponential modes in this case is very difficult as a result. Indeed from the above asymptotic expansions, $\alpha \approx 1$ and so the slow exponential mode is hard to resolve.

- Case 6: $k = \frac{a}{\epsilon}$ and $S = \frac{b}{\epsilon}$.

  $$\begin{aligned}\lambda_1 &= \frac{a}{a+b} + \left(\frac{a^2}{(a+b)^3} - \frac{a}{(a+b)^2}\right)\epsilon + O\left(\epsilon^2\right),\\ \lambda_2 &= \frac{a+b}{\epsilon} + 1 - \frac{a}{a+b} + \left(\frac{a}{(a+b)^2} - \frac{a^2}{(a+b)^3}\right)\epsilon + O\left(\epsilon^2\right),\\ \alpha &= 1 + \left(\frac{a}{(a+b)^2} - \frac{1}{a+b}\right)\epsilon + O\left(\epsilon^2\right).\end{aligned}$$

  In this case, k is large and S is proportional to k. Here, the balance between the length of the dwell times for the post-translocation and the dNTP-bound states are approximately equal. This means that the situations described when S is small or when S is much larger than k are mitigated, producing the least total relative error for large k regimes. We can also see this from the asymptotic expansion for α for this case, α is further from 1 than when S is small or much larger than k.

2.6.4 Finding the Optimal $[dNTP]$ Under Experimental Time Constraints

It is useful to investigate the behavior of the total relative error and optimal dNTP concentration when under experimental time constraints. The PDF of the lower-amplitude dwell time $T^{(1)}$ is a function of r_2, k_{on}, k_{off}, and $[dNTP]$ with $[dNTP]$ being the only tunable parameter that can be controlled in the experiments. The unconstrained optimal $[dNTP]$, while producing the least total relative error, can result in long run-times in nanopore experiments. The experimental run-time is a function of the number of samples of $T^{(1)}$ that we choose to collect, $[dNTP]$, and the kinetic rates r_2, k_{on}, and k_{off}. After applying the scaling laws in section 2.5.1, we can write the experimental waiting time as a function of S, k, and the number of $T^{(1)}$ samples. Let n be the number of $T^{(1)}$ samples. The mean-field approximation to the total lower-amplitude waiting time can be written as

$$n \left\langle T^{(1)} \left(S, k\right) \right\rangle .$$

We write the random variable $T^{(1)}$ as $T^{(1)}\left(S, k\right)$ to emphasis its dependence on the scaled dNTP concentration and the scaled k_{off}.

It is reasonable to assume that $\text{err}\left(S, k, n\right) \sim O\left(1/\sqrt{n}\right)$ as $n \to \infty$, since the standard error of a parameter scales as $O\left(1/\sqrt{n}\right)$ where n is the number of samples [13]. Indeed, figure 2.15 shows that $\text{err}\left(S, k, n\right) \sim O\left(1/\sqrt{n}\right)$ at $\left(S, k\right) = \left(10, 1\right)$. It is hence reasonable to conclude the following scaling law for the total

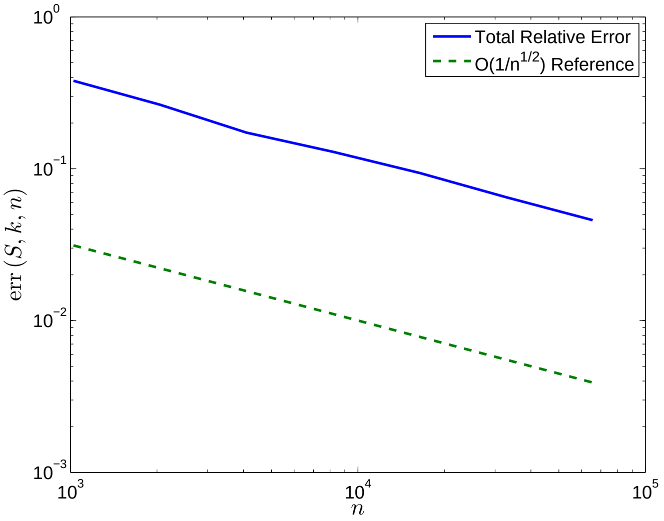

Figure 2.15: The total relative error at the point $(S, k) = (10, 1)$. The decaying of the error at $O\left(1/\sqrt{n}\right)$ is expected and numerically demonstrates the scaling law for the total relative error function (equation 2.20).

relative error function:

$$\mathrm{err}\,(S, k, n_1) \sim \mathrm{err}\,(S, k, n_2)\sqrt{\frac{n_2}{n_1}}. \tag{2.20}$$

Recall that the we numerically approximated total relative error function $\mathrm{err}_1\,(S, k)$ in equation 2.9 by constructing the discretized grid $\mathcal{E}$ using $n_0 = 10,000$ samples of $T^{(1)}$. Thus from equation 2.20, we have the approximation

$$\mathrm{err}\,(S, k, n) \approx \sqrt{\frac{n_0}{n}}\mathrm{err}_1\,(S, k)\,, \tag{2.21}$$

for large n.

Suppose that we do not want to wait more than $\tau_{\max}$ time for the total lower-amplitude run-time. We want to know the number of $T^{(1)}$ samples to collect and at what dNTP concentration to run the experiment that results in the least total relative error for the inference of k_{on} and k_{off}. The constrained optimization problem whose solution would result in those optimum n and S is thus given by

$$\text{minimize err}\left(S,k,n\right) \tag{2.22}$$

$$\text{subject to } n\left\langle T^{(1)}\left(S,k\right)\right\rangle = \tau_{\max}. \tag{2.23}$$

With k_{off} intrinsic to the system, the only tunable parameters are n and S.

We can recast the constrained optimization in equations 2.22-2.23 to an unconstrained optimization in S only in the following way. From the constraint in equation 2.23 we have that

$$n = \frac{\tau_{\max}}{\left\langle T^{(1)}\right\rangle}. \tag{2.24}$$

Thus from equation 2.21, the objective function to minimize which solves the optimization problem in equations 2.22-2.23 is given by

$$\text{err}_2\left(S,k\right) := \sqrt{\frac{n_0\left\langle T^{(1)}\right\rangle}{\tau_{\max}}}\text{err}_1\left(S,k\right). \tag{2.25}$$

Equation 2.25 can be used to create an error surface and find the optimum dNTP concentration which solves the constrained optimization problem given in equations 2.22-2.23 (see figure 2.16). The optimal number of $T^{(1)}$ samples to collect can then be calculated from equation 2.24 (see figure 2.17).

Generally the unconstrained optimal $[dNTP]$ will be higher than the constrained optimum $[dNTP]$ since at each k_{off}, higher dNTP concentrations increase the probability of the DNAP-DNA complex transitioning from the post-translocation to the dNTP-bound state and hence increases the $T^{(1)}$ dwell time. Thus when constraining the maximum experimental time, the constrained optimum dNTP concentration can be no larger than the unconstrained optimum dNTP concentration (figure 2.18). The total relative error at the constrained optimum $[dNTP]$ will therefore be higher than the total relative error at the un-

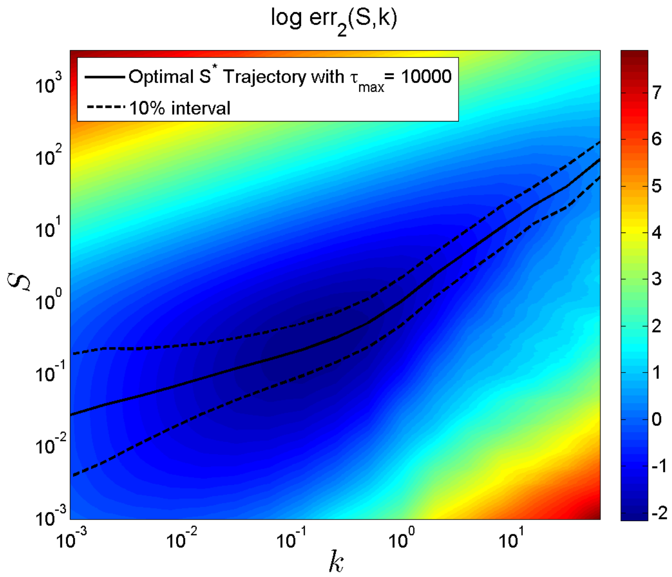

Figure 2.16: Constrained total relative error surface $\mathrm{err}_2(S,k)$ with optimal S^* concentration. The dashed lines are the 10% interval calculated in a similar manner as in figure 2.10.

constrained optimum $[dNTP]$ (figure 2.19). The minimum total relative error along the constrained $k \mapsto S^*$ trajectory occurs at a value of k not too small (figure 2.19), unlike in the unconstrained $k \mapsto S^*$ trajectory where the total relative error decreases with the decrease in k. This is because for very small values of k, the DNAP-DNA complex will take a longer time to transition from the dNTP-bound state to the post-translocation state, hence increasing the length of the $T^{(1)}$ segment. This results in low amounts of samples being used in order to maintain the total experimental time constraint in the constrained optimization problem in equations 2.22-2.23.

The scaled kinetic rate k and constrained optimal S^* that yields the least total relative error along the $k \mapsto S^*$ trajectory is the best possible system. Best possible in a sense that this system will yield the least total relative error at its optimal $[dNTP]$. Numerically finding the minimum of the total relative error at the constrained S^* in figure 2.19, we have $k_{\text{inf}} \approx 0.1096$ and correspondingly

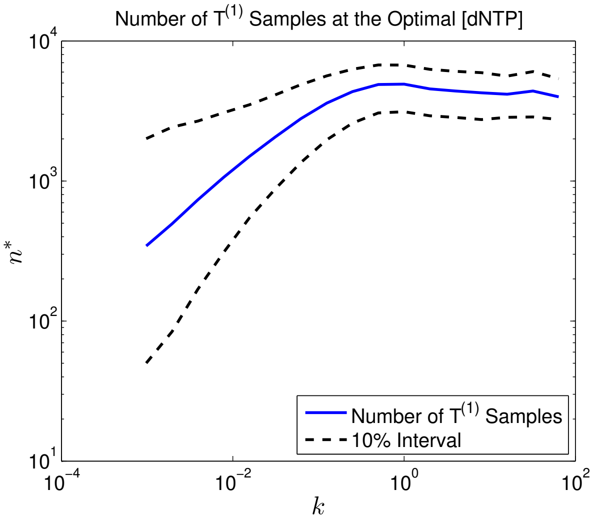

Figure 2.17: Number of $T^{(1)}$ samples at the optimal $[dNTP]$ with corresponding 10% interval as calculated from the optimal S^* trajectory found in figure 2.16.

$S^*_{\text{inf}} \approx 0.2281$ where k_{inf} and S^*_{inf} are the scaled k_{off} and $[dNTP]$ which produce the system that gives the least total relative error at its optimal $[dNTP]$.

Figures 2.20 and 2.21 show the unconstrained and constrained optimum S^* with intervals 10%, 25%, and 50%. These figures show how accurately the dNTP concentration must be in order to be within p% of the optimum scaled dNTP concentration S^* under the unconstrained and constrained conditions. Correspondingly, if the scaled dNTP concentration S is within p% of the optimum S^*, then the total relative error is no greater than $(1 + p\%/100)\,\text{err}_1\,(S, k)$ and $(1 + p\%/100)\,\text{err}_2\,(S, k)$ for the unconstrained case at 10,000 samples and the constrained case at $n\,(S, k)$ samples, respectively.

2.6.5 Finding the Optimal $[dNTP]$ Under Experimental Time Constraints with Overhead Cost

We can generalize the constrained optimization problem in equations 2.22-2.23 to include a fixed time-cost for collecting each $T^{(1)}$ sample. In the context of this

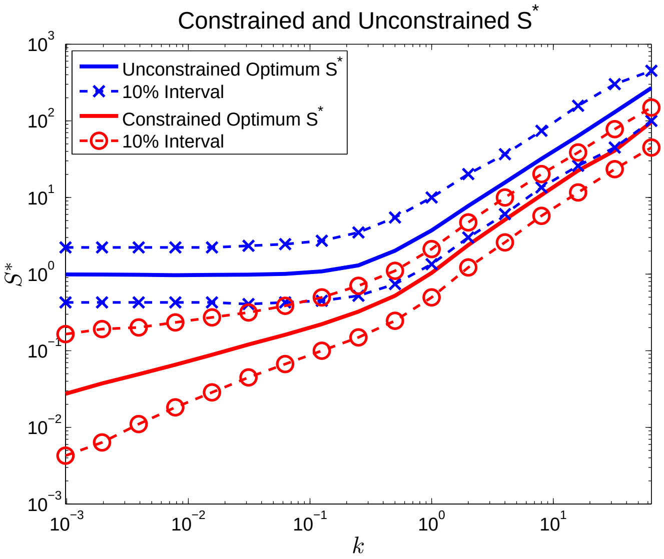

Figure 2.18: The optimum S^* for the unconstrained and constrained optimization problems along with their corresponding 10% intervals.

system, the fixed time-cost is the dwell time of the DNAP-DNA complex in the upper-amplitude state centered at 31pA, comprised of the pre-translocation and exonuclease states (see figure 1.5). Suppose that we have a fixed time-cost of t_0 for each $T^{(1)}$ sample. Note that the distribution of t_0 is a proper mixture of two exponential modes, determined completely by the transition rates r_1, r_3, and r_4 in a similar manner to Proposition 1 (figure 1.5). We include this fixed time cost into the maximum total time $\tau_{\max}$ allowed for the experiment. That is, we want to solve the constrained optimization problem,

$$\text{minimize } \operatorname{err}(S, k, n) \tag{2.26}$$

$$\text{subject to } n\left(\left\langle T^{(1)}(S, k)\right\rangle + t_0\right) = \tau_{\max}. \tag{2.27}$$

In a similar way for the constrained optimization problem in equations 2.22-2.23, we can recast this into an unconstrained optimization problem of S only.

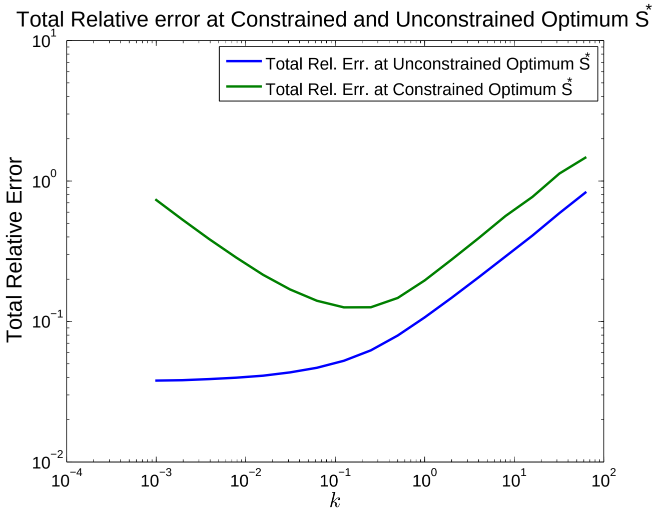

Figure 2.19: The total relative error at the constrained and unconstrained optimum S^*.

We thus want to minimize the objective function,

$$\mathrm{err}_3\left(S,k\right) := \sqrt{\frac{n_0\left(\langle T^{(1)}\rangle + t_0\right)}{\tau_{\max}}}\mathrm{err}_1\left(S,k\right). \tag{2.28}$$

Figure 2.22 shows the constrained optimum S^* for various values of t_0 along with the unconstrained optimum S^*. As seen in figure 2.22, the constrained optimum scaled dNTP approaches the unconstrained optimum dNTP as $t_0 \to \infty$. To see this, consider equation 2.28. We can re-write equation 2.28 as

$$\mathrm{err}_3\left(S,k\right) = \sqrt{t_0}\sqrt{\frac{n_0\left(\frac{\langle T^{(1)}\rangle}{t_0} + 1\right)}{\tau_{\max}}}\mathrm{err}_1\left(S,k\right).$$

Hence as $t_0 \to \infty$, we have the following asymptotic result,

$$\mathrm{err}_3\left(S,k\right) \sim \sqrt{\frac{t_0}{\tau_{\max}}n_0}\mathrm{err}_1\left(S,k\right) + O\left(\frac{1}{t_0}\right).$$

Thus we see that for any k, $\mathrm{argmin}_S\mathrm{err}_3\left(S,k\right) \to \mathrm{argmin}_S\mathrm{err}_1\left(S,k\right)$ as $t_0 \to \infty$.

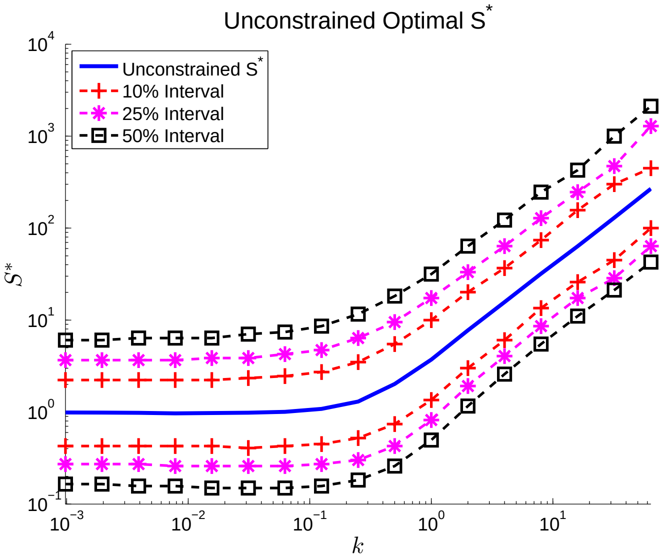

Figure 2.20: The optimum S^* for the unconstrained optimization problem along with the 10%, 25%, and 50% intervals.

The total relative error at the optimum S in this asymptotic regime will therefore scale as $O\left(\sqrt{t_0}\right)$.

In the constrained optimization problems, we are looking to minimize $\mathrm{err}\left(S, k, n\right)$ under the constraint $n\left(\left\langle T^{(1)}\right\rangle + t_0\right) = \tau_{\max}$. This constraint is the mean-field approximation to the experimental run time. However, under real experimental settings, the samples of $T^{(1)}$ would be collected one at a time until the constraint is met. That is, k samples of $T^{(1)}$ would be collected where k is the largest such that $T_1^{(1)} + \cdots T_k^{(1)} + t_0 \leq \tau_{\max}$. This is different than the mean-field approach taken in the optimization problems 2.22-2.23 and 2.26-2.27 in which the mean of $T^{(1)}$ is used instead of the observed total time of the $T^{(1)}$ samples.

The mean field approaches greatly simplifies the calculation of the solution to the constrained optimization problems by replacing the individual behavior of a large number of random variables $(T_1^{(1)}, \ldots, T_k^{(1)})$ with the ensemble average. We demonstrate the validity of the mean field approach by numerical simulation. For the constrained optimization case with $t_0 = 0$, we set $S = 0.1$, $k = 0.3$, and

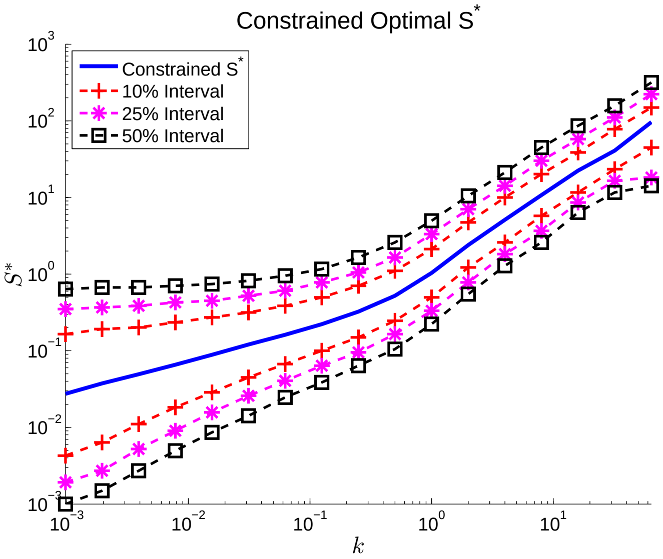

Figure 2.21: The optimum S^* for the constrained optimization problem along with the 10%, 25%, and 50% intervals.

$\tau_{\max} = \left\langle T^{(1)} \right\rangle n_{mf}$, where $n_{mf} = 10000$. The total relative error is then estimated using the mean field approach and using the constraint $\sum_i T_i^{(1)} \leq \tau_{\max}$ using 2000 data sets. As seen in figure 2.23, both approaches are in agreement and is well approximated by $\text{err}_1(S, k)$.

2.7 Inference on Simulated Samples of Dwell Times with Detection Uncertainty

Let σ denote the standard deviation of the measurement noise. In this section, we repeat the numerical experiments in section 2.4 with multiplicative noise in the observed $T^{(1)}$ samples; i.e., the observed value of $T^{(1)}$ is given by

$$T_{obs}^{(1)} := T^{(1)} e^{\sigma\zeta},$$

where $\zeta \sim N(0, 1)$.

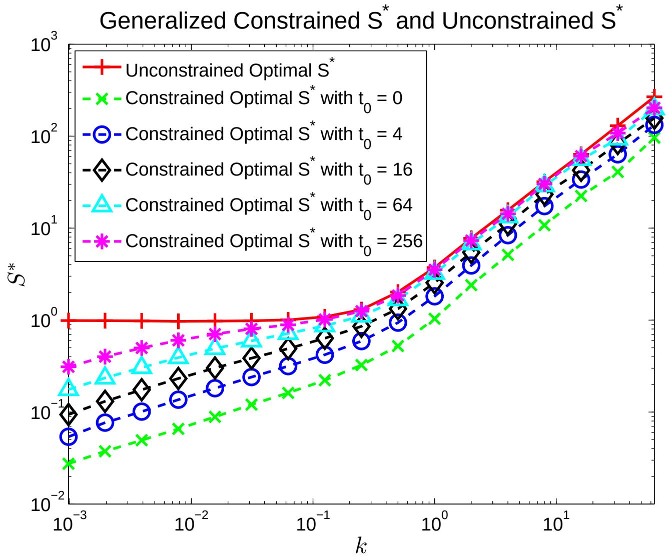

Figure 2.22: Constrained optimum S^* for various values of t_0 along with the unconstrained optimum S^*.

Let $\sigma = 0.01$. In this situation, the best MLE estimates are found around the concentration $[dNTP] = 2$. Like the ideal, no noise case, the bias is small throughout the [dNTP] ranges (figures 2.24 and 2.25).

Finally, the numerical experiment is repeated with a full magnitude increase in noise magnitude; i.e., $\sigma = 0.1$. For k_{off}, the best MLE estimates were provided at $[dNTP] = 2$ as before, but for k_{on}, the best MLE estimates were provided at $[dNTP] = 0.5$ (figures 2.26 and 2.27).

To get a sense of how $\mathrm{err}\,(k_{on})$ and $\mathrm{err}\,(k_{off})$ behave over a greater range of noise magnitudes, we plot $\mathrm{std}\,(\mathrm{err}\,(k_{on}))$ and $\mathrm{std}\,(\mathrm{err}\,(k_{off}))$ as a function of σ for each of the representative [dNTP] points (figure 2.28). From this figure we can see that measurement noise in $T^{(1)}$ samples affect the estimation of k_{off} more so than k_{on}. For k_{on}, the dNTP concentration of $[dNTP] = 0.5$ provides the estimates with the smallest relative error; even for a relatively high noise magnitude, the relative error for k_{on} is still small.

It is also valuable to see how the bias of the MLE estimates changes with

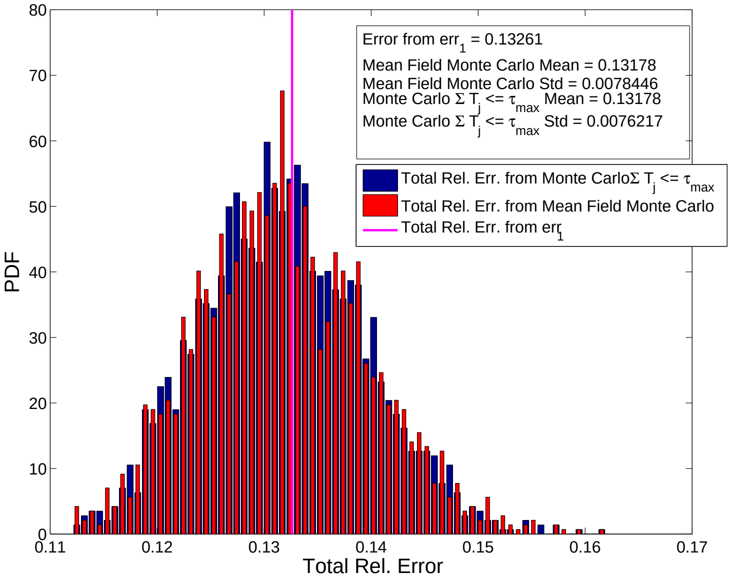

Figure 2.23: Comparison of the constrained total relative error err_2 obtained from the mean field approach and from enforcing the total time in the observed $T^{(1)}$ samples. Here, $S = 0.1$, $k = 0.3$, and $\tau_{\max} = \left\langle T^{(1)} \right\rangle n_{mf}$, where $n_{mf} = 10000$. The distributions of the mean field approach and approach $\sum_i T_i^{(1)} \leq \tau_{\max}$ are in agreement. The mean of the mean-field approach can be calculated from err_1.

respect to the noise magnitude. The inference bias is the mean of the MLE estimates, and we denote this by mean (err (k)) (figure 2.29). Here we see that, even for high noise magnitudes, the bias remains relatively small. A key conclusion from this is that even with relatively high noise, repeating the MLE method on many data sets of $T^{(1)}$ and then averaging out the MLE estimates of k_{on} and k_{off} across those data sets will yield satisfactory approximations of the transition rates.

We also plot the root-mean-squared (RMS) error as a function of σ. The RMS error contains information about both std(err(k)) and mean(err(k)). We denote the RMS of the MLE estimates as rms (err (k)) (figure 2.30).

It is also worthwhile to examine the response of standard deviation, inference

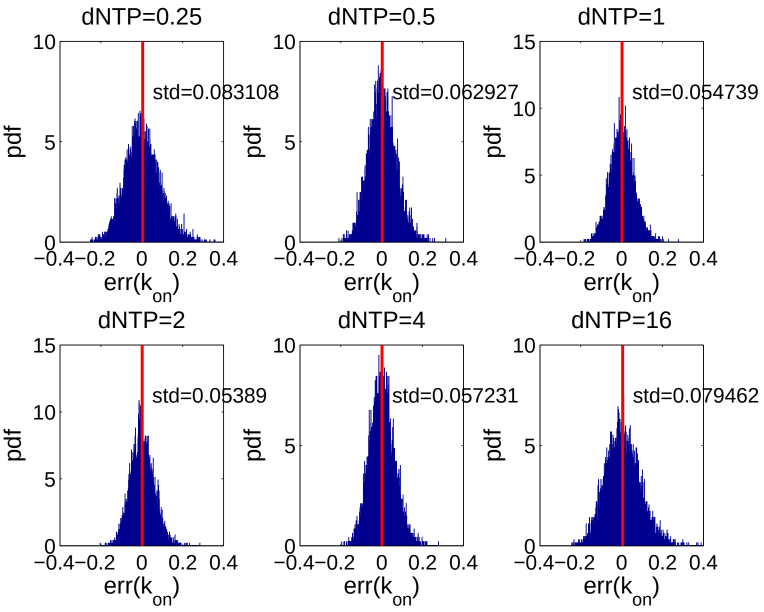

Figure 2.24: MLE results for k_{on} with multiplicative noise magnitude of $\sigma = 0.01$ in $T^{(1)}$ observations.

bias, and RMS of the MLE estimates of k_{on} and k_{off} as the number of $T^{(1)}$ samples n of $T^{(1)}$ change. Note that the number of data-sets consisting of n samples of $T^{(1)}$ remains fixed at 10,000. For this simulation, we hold the dNTP concentration fixed at $[dNTP] = 0.5$. According to the results in figures 2.28-2.30, $[dNTP] = 0.5$ provides the most robust concentration of the concentrations examined for estimating k_{on} and k_{off} (figures 2.31- 2.33). From the graphs, we can see that around 5000 samples of $T^{(1)}$ is sufficient to get a meaningful estimate of the transition rates k_{on} and k_{off}.

2.8 Characterizing the Effect of Measurement Noise

In this section, we characterize the effect of measurement noise on the observed $T^{(1)}$ samples. Suppose the true $T^{(1)}$ samples are perturbed by multiplicative noise

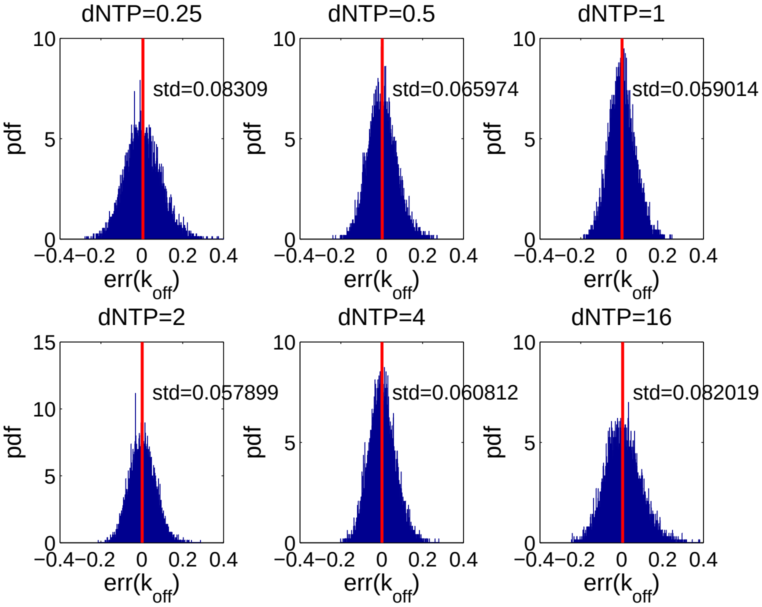

Figure 2.25: MLE results for k_{off} with multiplicative noise magnitude of $\sigma = 0.01$ in $T^{(1)}$ observations.

of the form $e^{\sigma\zeta}$ where $\zeta \sim N(0,1)$. That is, we observe the $T^{(1)}$ samples to be

$$T^{(1)}_{\text{obs}} := T^{(1)} e^{\sigma\zeta}. \tag{2.29}$$

In this section, we denote $k_{on}^{\text{MLE}}(\sigma)$ and $k_{off}^{\text{MLE}}(\sigma)$ to be the maximum-likelihood estimate of k_{on} and k_{off} respectively from the perturbed $T^{(1)}_{\text{obs}}$ data. To see how the MLE of k_{on} and k_{off} is effected by noise, we investigate the first-two moments and standard deviation of the quantities

$$z_{k_{on}} := k_{on}^{\text{MLE}}(\sigma) - k_{on}^{\text{MLE}}(0)$$
$$z_{k_{off}} := k_{off}^{\text{MLE}}(\sigma) - k_{off}^{\text{MLE}}(0).$$

For the following simulations, we use 10,000 data sets with $r_2 = k_{on} = [dNTP] = 1$ and $k_{off} = 0.25$ with varying the number of $T^{(1)}$ samples n, as

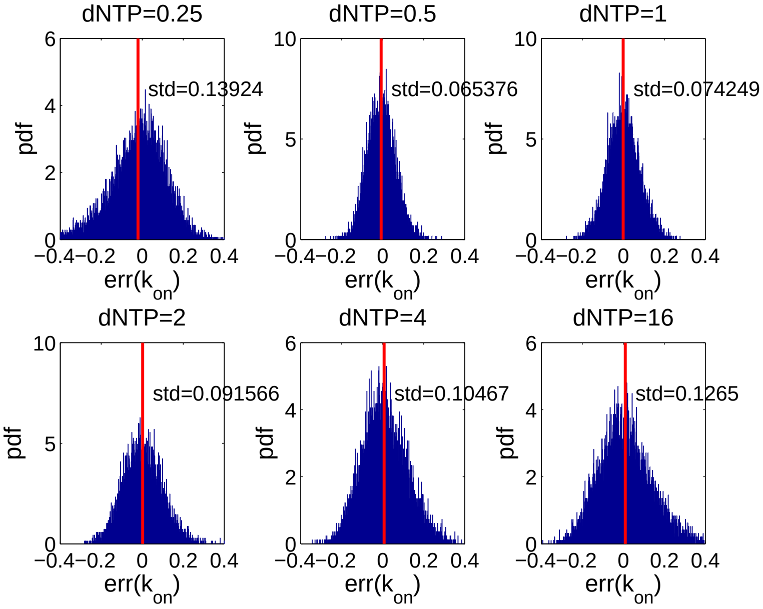

Figure 2.26: MLE results for k_{on} with multiplicative noise magnitude $\sigma = 0.1$ in $T^{(1)}$ observations.

well as the measurement noise standard deviation σ.

Figure 2.34 shows the squared-mean of $z_{k_{on}}$ and $z_{k_{off}}$. Here, we see that $\left\langle z_{k_{on}} \right\rangle, \left\langle z_{k_{off}} \right\rangle = O\left(\sigma^2\right)$.

Figure 2.35 and figure 2.36 shows the second-moment of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of σ and as a function of n, respectively. From these results, we see that as $\sigma \to 0$ and $n \to \infty$, $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle = O\left(\sigma^2/n\right)$. For large σ, we have that $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle = O\left(\sigma^4\right)$. We can write this more compactly as $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle = O\left(\sigma^2/n\right) + O\left(\sigma^4\right)$.

Figures 2.37 and 2.38 shows the variance of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of σ and n, respectively. From these results, we see that $var\left(z_{k_{on}}\right)$ and $var\left(z_{k_{off}}\right)$ behave as $O\left(\sigma^2/n\right)$.

The results of these simulations show that we have strong numerical evidence for the following claims:

1. $\left\langle z_{k_{on}} \right\rangle, \left\langle z_{k_{off}} \right\rangle = O\left(\sigma^2\right)$,

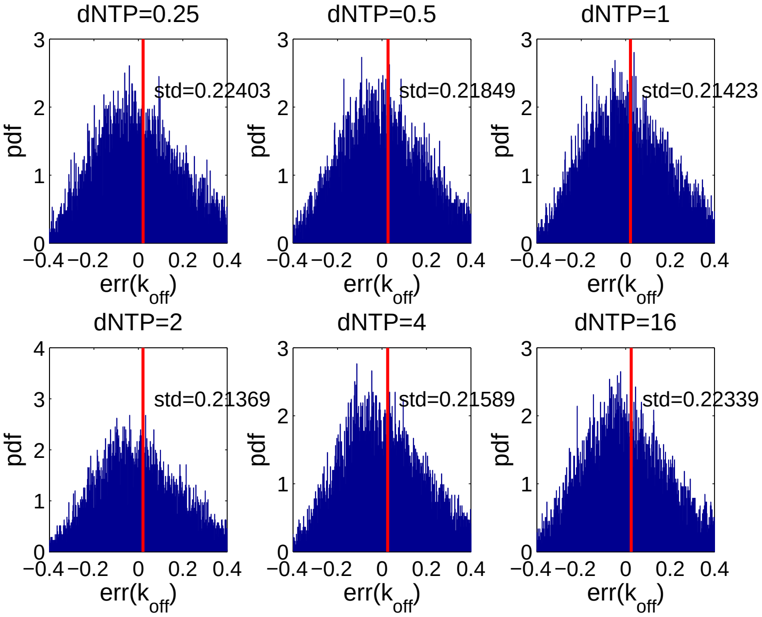

Figure 2.27: MLE results for k_{off} with multiplicative noise magnitude $\sigma = 0.1$ in $T^{(1)}$ observations.

2. $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle = O\left(\sigma^2/n\right) + O\left(\sigma^4\right)$, and

3. $var\left(z_{k_{on}}\right), var\left(z_{k_{off}}\right) = O\left(\sigma^2/n\right)$.

That is, both $z_{k_{on}}$ and $z_{k_{off}}$ are of the form $O\left(\sigma/\sqrt{n}\right)+O\left(\sigma^2\right)$, where the $O\left(\sigma/\sqrt{n}\right)$ term vanishes under the expectation.

To investigate the distribution of $z_{k_{on}}$ and $z_{k_{off}}$, we detail two different ways of collecting the noise-perturbed samples of $T^{(1)}$ and show that they are equivalent.

- Type 1: For each 10,000 data sets, n $T^{(1)}$ samples are generated and those n samples are perturbed by multiplicative noise with standard deviation σ as in equation 2.29.

- Type 2: n $T^{(1)}$ samples are generated, and those n samples are perturbed by 10,000 different realizations of multiplicative noise with standard deviation σ as in equation 2.29.

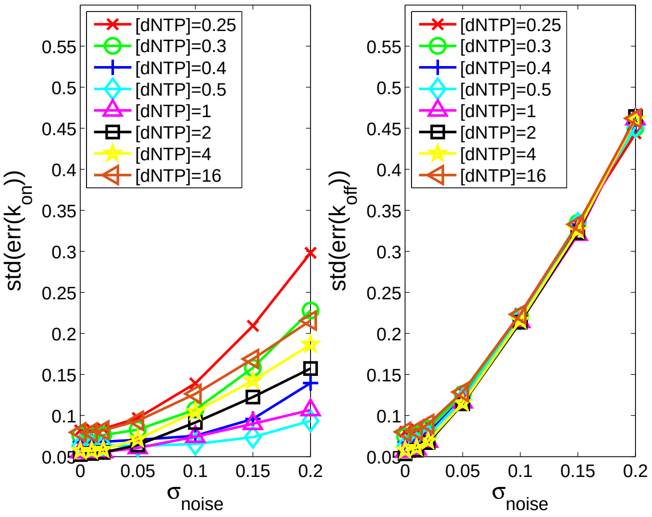

Figure 2.28: The quantities $\text{std}\left(\text{err}\left(k_{on}\right)\right)$ and $\text{std}\left(\text{err}\left(k_{off}\right)\right)$ as a function of σ for each dNTP concentration.

Figure 2.39 shows the distributions of $z_{k_{on}}$ and $z_{k_{off}}$ using the two different types of data sets, each with $n = 32000$ samples. Here, we see vary good agreement among the distributions of $z_{k_{on}}$ and $z_{k_{off}}$ using the two different types of data sets. Furthermore, we see that both $z_{k_{on}}$ and $z_{k_{off}}$ are normally distributed. From this observation and from claims 1-3 above, we can write $z_{k_{on}}$ and $z_{k_{off}}$ as

$$z_{k_{on}} = c_{2,k_{on}}\sigma^2 + c_{1,k_{on}}\frac{\sigma}{\sqrt{n}}\zeta, \tag{2.30}$$

$$z_{k_{off}} = c_{2,k_{off}}\sigma^2 + c_{1,k_{off}}\frac{\sigma}{\sqrt{n}}\zeta, \tag{2.31}$$

where $\zeta \sim N\left(0,1\right)$.

The consequence of this result is that the bias of $k_{on}^{\text{MLE}}\left(\sigma\right)$ and $k_{off}^{\text{MLE}}\left(\sigma\right)$ increases by a magnitude of $O\left(\sigma^2\right)$ relative to the bias of $k_{on}^{\text{MLE}}\left(0\right)$ and $k_{off}^{\text{MLE}}\left(0\right)$;

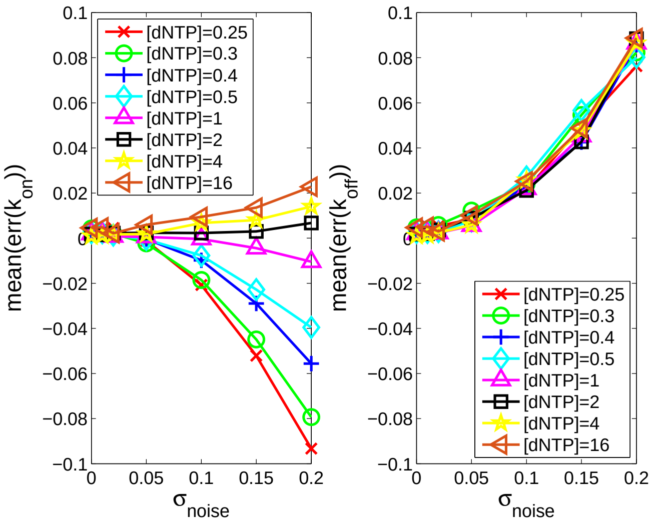

Figure 2.29: The quantities mean $(\text{err}\,(k_{on}))$ and mean $(\text{err}\,(k_{off}))$ as a function of σ for each dNTP concentration.

that is,

$$\left\langle k_{on}^{\text{MLE}}\left(\sigma\right)\right\rangle = \left\langle k_{on}^{\text{MLE}}\left(0\right)\right\rangle + O\left(\sigma^2\right),$$
$$\left\langle k_{off}^{\text{MLE}}\left(\sigma\right)\right\rangle = \left\langle k_{off}^{\text{MLE}}\left(0\right)\right\rangle + O\left(\sigma^2\right).$$

Also, the variance of the perturbed MLE estimates increase by $O\left(\sigma^2/n\right)$; that is,

$$var\left(k_{on}^{\text{MLE}}\left(\sigma\right)\right) = var\left(k_{on}^{\text{MLE}}\left(0\right)\right) + O\left(\frac{\sigma^2}{n}\right),$$
$$var\left(k_{off}^{\text{MLE}}\left(\sigma\right)\right) = var\left(k_{off}^{\text{MLE}}\left(0\right)\right) + O\left(\frac{\sigma^2}{n}\right),$$

since $k_{on}^{\text{MLE}}\left(0\right)$ and $k_{off}^{\text{MLE}}\left(0\right)$ are independent with the normal in 2.30-2.31.

We can numerically solve for the constants c_1 and c_2 for $z_{k_{on}}$ and $z_{k_{off}}$ in equations 2.30-2.31 by least-squares fitting. From equation 2.30, we have that $\left\langle z_{k_{on}}\right\rangle^2 = c_{2,k_{on}}^2\sigma^4$ and $var\left(z_{k_{on}}\right) = c_{1,k_{on}}\sigma^2/n$. The least-squares solution is given

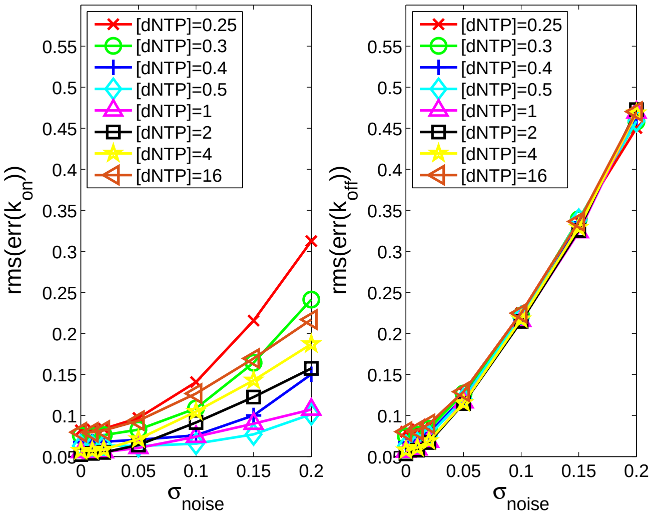

Figure 2.30: The quantities $\text{rms}\left(\text{err}\left(k_{on}\right)\right)$ and $\text{rms}\left(\text{err}\left(k_{off}\right)\right)$ as a function of σ for each dNTP concentration.

by

$$c_{1,k_{on}} = \sqrt{\frac{\sum_i var\left(z_{k_{on}}\right)\Big|_{\sigma=\sigma_i}}{\frac{\sigma_i^2}{n}}}, \tag{2.32}$$

$$c_{2,k_{on}} = \frac{\sum_i \left\langle z_{k_{on}} \right\rangle \Big|_{\sigma=\sigma_i}}{\sum_i \sigma_i^2}. \tag{2.33}$$

A least-squares solution for $c_{1,k_{off}}$ and $c_{k_{on}}$ can be derived in a similar manner.

To verify the validity of the least-squares fitting, we compare the mean and variance of $z_{k_{on}}$ and $z_{k_{off}}$ at $\sigma = 2^{-4}$ and $n = 32000$ with their predicted mean and variance as obtained through the least-squares fit above (table 2.1). Here, we obtained

- $c_{1,k_{on}} = 4.6580$
- $c_{1,k_{off}} = 1.2843$
- $c_{2,k_{on}} = -0.3467$

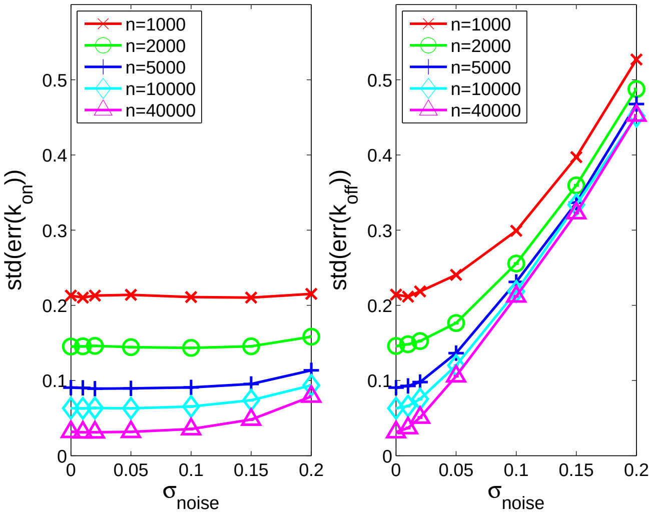

Figure 2.31: The quantities $\text{std}\left(\text{err}\left(k_{on}\right)\right)$ and $\text{std}\left(\text{err}\left(k_{off}\right)\right)$ as a function of σ for different sample sizes n.

- $c_{2,k_{off}} = -0.2219$

The distribution of $z_{k_{on}}$ and $z_{k_{off}}$ for these values of σ and n are shown in figure 2.39. Table 2.1 shows that there is good agreement between the predicted and observed mean and variances of $z_{k_{on}}$ and $z_{k_{off}}$.

	$z_{k_{on}}$	$z_{k_{off}}$
mean	-0.001102	-0.00089558
mean from fit	-0.0014	-0.00086661
var.	2.4636×10^{-6}	2.3314×10^{-7}
var. from fit	2.6486×10^{-6}	2.0136×10^{-7}

Table 2.1: Comparison between the observed mean and variance of $z_{k_{on}}$ and $z_{k_{off}}$ with their predicted means and variances obtained through the least-squares fit. Here, $\sigma = 2^{-4}$ and $n = 32000$. The results show good agreement between the observed and predicted mean and variances.

Throughout the number of samples examined in our numerical simulation ($n = 1000, 2000, 4000, 8000, 16000, 32000, 64000$), we observed little change in the least-squares solutions for c_1 and c_2 for $z_{k_{on}}$ and $z_{k_{off}}$. In fact, for these n's, the means of $c_{1,k_{on}}, c_{2,k_{on}}, c_{1,k_{off}}$, and $c_{2,k_{off}}$ are 4.6908, -0.3454, 1.2934, and -0.2213,

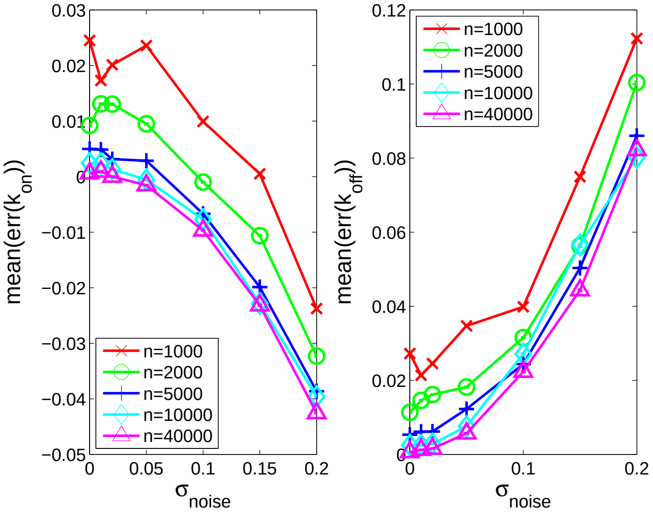

Figure 2.32: The quantities mean (err (k_{on})) and mean (err (k_{off})) as a function of σ for different sample sizes n.

respectively. The standard deviations are 0.0457, 0.0030, 0.0122, and 0.0009, respectively.

The importance of these results is that for any r_2, k_{on}, k_{off}, and $[dNTP]$, we can collect n unperturbed $T^{(1)}$ samples and perturb them m times to obtain m data sets. From this data, the coefficients c_1 and c_2 in equations 2.30-2.31 can be obtained by least-squares fitting (equations 2.32-2.33) and an accurate description of the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ can be obtained.

2.8.1 Analysis of the Single Exponential Mode

To gain some intuition as to why we can write the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ as in equations 2.30 and 2.30, we consider the simpler problem of inferring the rate from a single exponential mode: $T_1, \ldots, T_n \sim \exp(r)$, where the $T_1, \ldots T_n$ are independent and identically distributed (iid). The MLE of $1/r$ is then given by

$$\rho := \frac{1}{r} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

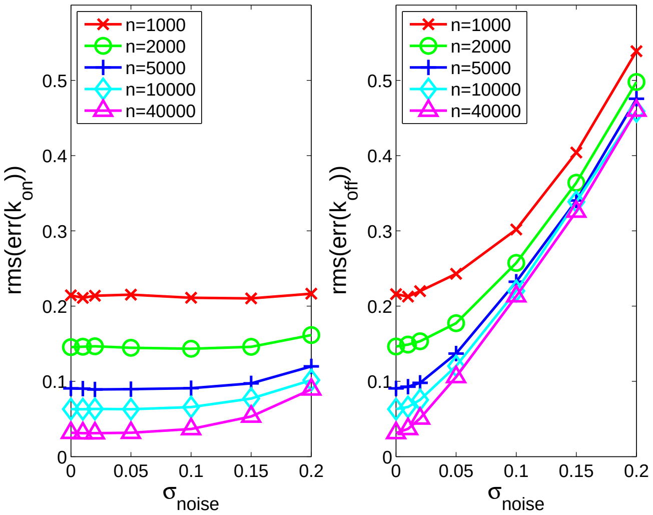

Figure 2.33: The quantities $\text{rms}\left(\text{err}\left(k_{on}\right)\right)$ and $\text{rms}\left(\text{err}\left(k_{off}\right)\right)$ as a function of σ_{noise} for different sample sizes n.

Consider the T_i samples perturbed by multiplicative noise, so that the observed T_i samples are of the form $T_i \exp\left(\sigma\zeta_i\right)$ where $\zeta_i \sim N\left(0,1\right)$, iid. Now the MLE of $1/r$ from the perturbed samples is given by

$$\rho\left(\sigma\right) := \frac{1}{r\left(\sigma\right)} = \frac{1}{n}\sum_{i=1}^{n} T_i e^{\sigma\zeta_i}.$$

We write $\rho\left(\sigma\right)$ to emphasize the dependence on σ. We can then write,

$$\rho\left(\sigma\right) := \frac{1}{r\left(\sigma\right)} = \frac{1}{n}\sum_{i=1}^{n} T_i + \frac{\sigma}{n}\sum_{i=1}^{n} T_i\zeta_i + \frac{\sigma^2}{2n}\sum_{i=1}^{n} T_i\zeta_i^2 + \cdots ,$$

after Taylor expansion. From here, we see that the first term is $\rho\left(0\right)$. From the Central Limit Theorem, the second term is approximately $\sigma N\left(0, 2/\left(nr^2\right)\right)$ for

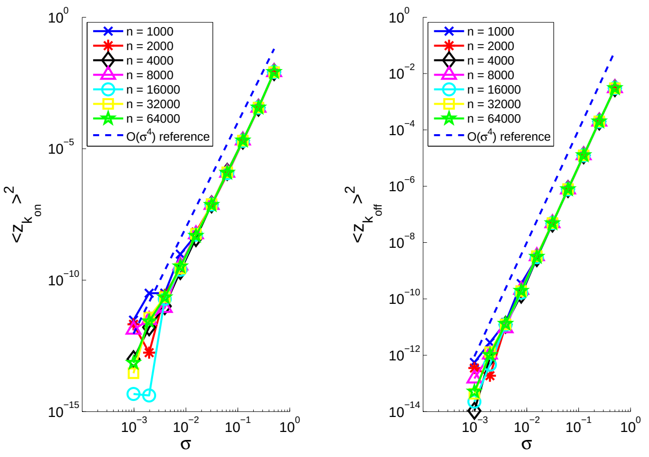

Figure 2.34: The squared-mean of the quantities $z_{k_{on}}$ and $z_{k_{off}}$. This shows that the squared-mean of $z_{k_{on}}$ and $z_{k_{off}}$ both follow $O\left(\sigma^4\right)$ as $\sigma \to 0$ and hence $\left\langle z_{k_{on}}\right\rangle, \left\langle z_{k_{off}}\right\rangle = O\left(\sigma^2\right)$.

large n. For the third term, we can write

$$\begin{aligned}\frac{\sigma^2}{2}\frac{1}{n}\sum_{i=1}^{n} T_i\zeta_i^2 &= \frac{\sigma^2}{2}\frac{1}{n}\sum_{i=1}^{n}\left(T_i\zeta_i^2 - \left\langle T_i\zeta_i^2\right\rangle + \left\langle T_i\zeta_i^2\right\rangle\right)\\ &= \frac{\sigma^2}{2}\frac{1}{n}\sum_{i=1}^{n}\left(T_i\zeta_i^2 - \left\langle T_i\zeta_i^2\right\rangle\right) + \frac{\sigma^2}{2r}.\end{aligned}$$

Notice that first term, $\frac{1}{n}\sum_{i=1}^{n}\left(T_i\zeta_i^2 - \left\langle T_i\zeta_i^2\right\rangle\right)$ is approximately normal $N\left(0, 8/\left(nr^2\right)\right)$ for large n by the Central Limit Theorem. Hence we have that

$$\frac{\sigma^2}{2}\frac{1}{n}\sum_{i=1}^{n} T_i\zeta_i^2 \text{ approximately } \frac{\sigma^2}{2}N\left(0, \frac{8}{nr^2}\right) + \frac{\sigma^2}{2r},$$

for large n. Notice that the stochastic contribution of $\sigma^2/2N\left(0, 8/\left(nr^2\right)\right)$ is small for small σ. Putting all of this together, we can formally write

$$\rho\left(\sigma\right) - \rho\left(0\right) = \frac{\sigma^2}{2r} + \frac{\sigma}{\sqrt{n}}N\left(0, \frac{2}{r^2}\right) + \cdots. \tag{2.34}$$

Notice that in equation 2.34, we have a bias of order $O\left(\sigma^2\right)$ and a variance

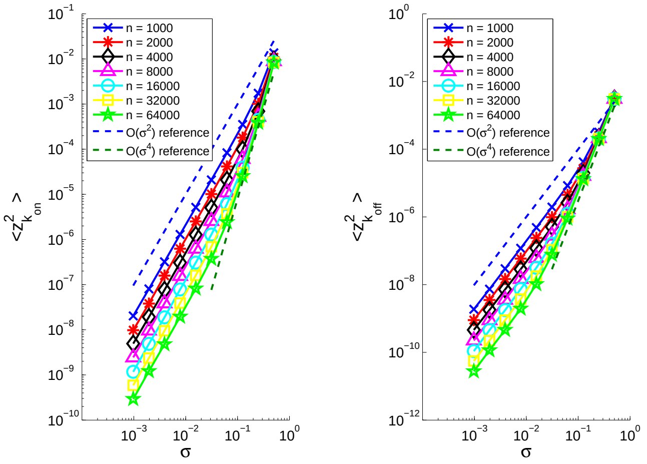

Figure 2.35: The second moments of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of σ. This shows that the second moments follow $O\left(\sigma^2\right) + O\left(\sigma^4\right)$ for fixed n.

of order $O\left(\sigma^2/n\right)$, similar to equations 2.30 and 2.31. The extension to the more general case of a proper exponential mixture is much harder, since in our context, k_{on} and k_{off} is intertwined in the mixture parameters $(\alpha, \lambda_1, \lambda_2)$. Nevertheless, this example of a single exponential mode gives us confidence in the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ given in equations 2.30 and 2.31.

2.9 Discussion and Concluding Remarks

The techniques used to derive the PDF of $T^{(1)}$ and the proceeding steps to derive the EM algorithm to infer the MLE of k_{on} and k_{off} can be applied to many phenomena which can be described as a Markov chain and for which dwell-time data can be gathered. The EM framework is dependent on the PDF of the observed data being in the form of a proper mixture distribution.

Using the EM algorithm to infer estimates for k_{on} and k_{off} has been shown to be robust under a wide range of noise magnitudes. We have found that the relative error of the inferred k_{on} and k_{off} rates are dependent on the dNTP concentration

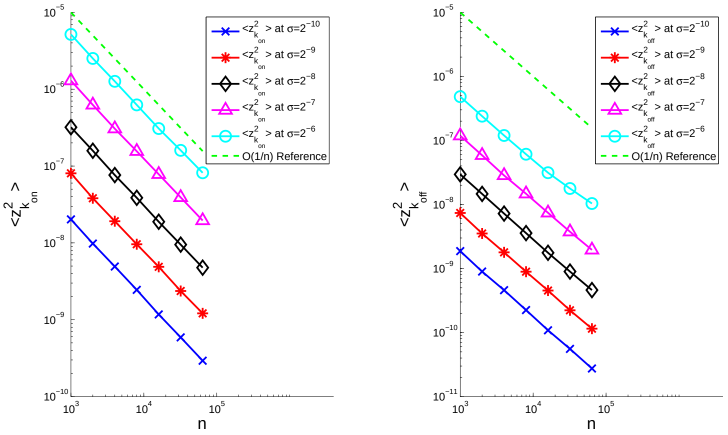

Figure 2.36: The second moments of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of n. This shows that the second moments follow $O(1/n)$ as $n \to \infty$ for fixed σ.

used. This is more evident in the inference of k_{on}, but it is only weakly dependent for the inference of k_{off}. For example, at $\sigma_{noise} = 0.2$, relative error for k_{on} ranges from about 9% to 30%, whereas for k_{off}, the relative error only ranges from about 44% to 46% throughout the dNTP concentrations examined for that noise magnitude. The relative error ranges tend to decrease for k_{on} as the noise magnitude decreases. For σ_{noise} greater than about 0.05, using $[dNTP] = 0.5$ provides the lowest relative error for k_{on} among the dNTP concentrations tested. Below 0.05, $[dNTP] = 2$ provides the lowest relative error among the dNTP concentrations examined. For both k_{on} and k_{off}, the inference bias remains under 10% throughout the dNTP concentrations tested and throughout the range of σ_{noise} examined.

We also examined the behavior of the relative error and inference bias for k_{on} and k_{off} when the number of samples of $T^{(1)}$ was varied. We found that observing around 5000 samples of $T^{(1)}$ is sufficient for the conclusions in the previous paragraph to be valid; recall that those conclusions were based on a data set in which 10000 samples of $T^{(1)}$ were observed.

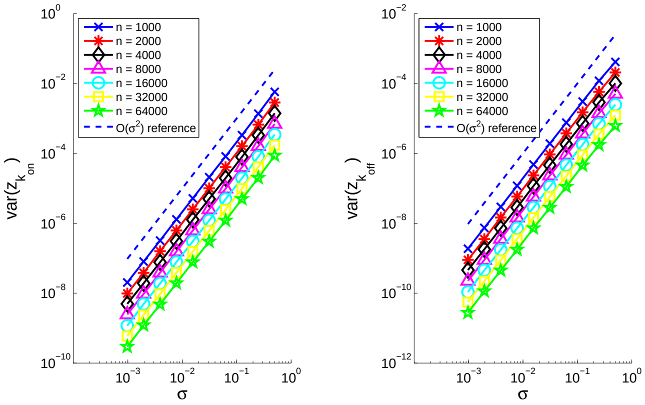

Figure 2.37: The standard deviation of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of σ. For fixed n, the variance of $z_{k_{on}}$ and $z_{k_{off}}$ behave as $O\left(\sigma^2\right)$.

The inference uncertainty of k_{on} and k_{off} was also investigated. We showed that the total relative error for inferring k_{on} and k_{off} is a function of the scaled $[dNTP]$ and scaled k_{off} only. We used the observed Fisher information matrix to obtain an asymptotic estimate for the covariance matrix for the MLE estimates and then we propagated that uncertainty to the k_{on} and k_{off} estimates through a first-order Taylor expansion. This and the scaling laws allowed us to build a numerical approximation to the total relative error for any k_{on} and k_{off}. From the scaling laws, we can also infer that the $[dNTP]$ that produces the least total relative error is a function of k only. The optimum $[dNTP]$ was also numerically estimated by the approximated total relative error function.

The total relative error table in (S, k) calculated in this paper for the numerical approximation to $\mathrm{err}_1\left(S, k\right)$ can also be applied to the synthesizing case in which the DNAP-DNA complex is allowed to incorporate a dNTP and proceed through the polymerization process. This extension can be made if the covariance matrix of the MLE estimates of the mixture parameters $(\alpha, \lambda_1, \lambda_2)$ are stored for each (S, k)-point in the table. It can be shown that for $k_{pol} > 0$, the escape problem governing $T^{(1)}$ can be re-written as an equivalent escape problem of the same

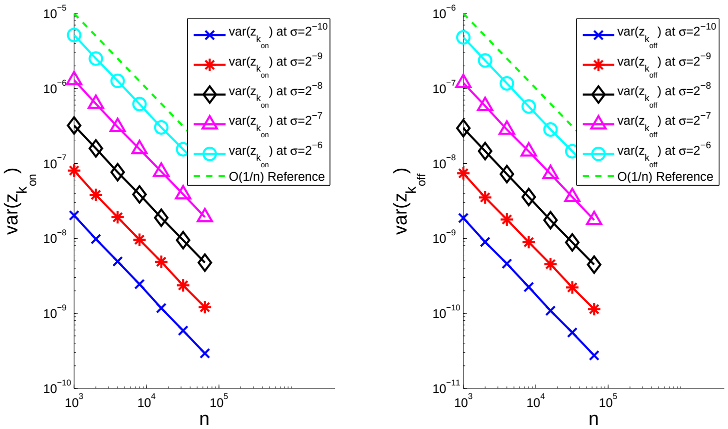

Figure 2.38: The standard deviation of $z_{k_{on}}$ and $z_{k_{off}}$ as a function of σ. For fixed σ, the variance of $z_{k_{on}}$ and $z_{k_{off}}$ behave as $O\left(1/n\right)$.

form as the escape problem governing $T^{(1)}$ for the $k_{pol} = 0$ case presented in this paper. Thus the saved covariance matrices for each (S, k)-point can be used in the synthesizing case.

We also looked into the constrained optimization problem in which the total experimental time was constrained using the mean-field approximation of the total experimental time and found the optimal number of $T^{(1)}$ samples to collect and the optimum $[dNTP]$ concentration to run the experiment which produces the least total relative error. The constrained optimization problem can be recast into an unconstrained optimization problem of $[dNTP]$ only. Using this technique, we were able to numerically estimate the $[dNTP]$ which produces the least relative error for each k. We showed that the use of the mean-field approximation to the total experimental run time was valid numerically.

The optimization problem was generalized to include the cost of obtaining each sample when considering the escape back to the lower-amplitude from the upper-amplitude. Again, we used the mean-field approximation to the total experimental time in this context. We also showed that as the cost of obtaining each sample

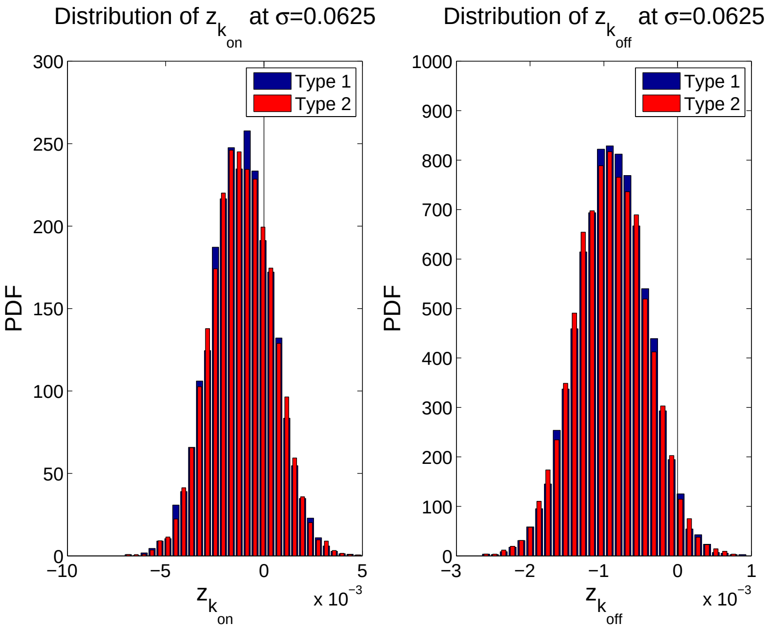

Figure 2.39: Distributions of $z_{k_{on}}$ and $z_{k_{off}}$ using two different types of data sets, each with $n = 32000$ samples. In the type 1 data sets, each data set is perturbed by a separate independent set of noises. In the type 2 data sets, all data sets are perturbed by the same set of noises. The distributions using the two types show good agreement.

increases, the optimal S approaches the optimal S in the unconstrained case.

The construction of the total relative error function and characterization of the optimal $[dNTP]$ thus provide a way to determine the experimental parameters which produce the least inference uncertainty when inferring the dNTP binding and disassociation rates. This a priori knowledge will allow researchers to make more accurate estimates for the dNTP binding and disassociation rates and further elucidate the dynamics of dNTP binding DNA-DNAP complexes.

Finally characterization of the MLE estimates from perturbed $T^{(1)}$ samples was also investigated. Using numerical simulations, we obtained strong numerical evidence to support the claims that the MLE estimates of k_{on} and k_{off} from the perturbed $T^{(1)}$ data differ from the MLE estimates of k_{on} and k_{off} from the unperturbed $T^{(1)}$ data by a Gaussian with deterministic mean of order $O\left(\sigma^2\right)$ and stochastic variance of order $O\left(\sigma^2/n\right)$, where σ is the standard deviation of the noise. Furthermore, the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ can be accurately described by least-squares fitting of the asymptotic coefficients to the squared-mean and variance of $z_{k_{on}}$ and $z_{k_{off}}$. The asymptotic coefficients are shown to have a weak dependence on n. This and numerical simulations examining the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ show that the distribution of $z_{k_{on}}$ and $z_{k_{off}}$ can be accurately obtained for any system in the following way: (1) generate n unperturbed $T^{(1)}$ samples and perturb them m times to create m data sets; and (2) determine the asymptotic coefficients of the squared-mean and variance of $z_{k_{on}}$ and $z_{k_{off}}$ by least-squares fitting.

Allowing the bound dNTP to proceed to the chemical step of phosphodiester bond formation is a natural extension to this paper. In this setting, after dNTP is bound (but not yet incorporated covalently), the DNA-DNAP complex can transition back to the post-translocation state or fully incorporate the bound dNTP and proceed to through the polymerization process and onto the next nucleotide addition cycle (figure 2.2). In this setting, the DNAP-DNA complex can escape to the next nucleotide addition cycle. Hence observing $T^{(1)}$ dwell times are in direct competition with the T_B dwell times. Developing statistical inference tools and optimal experimental design methodologies in this more general setting will allow for more robust control DNAP-DNA complexes allowed to undergo synthesis and illuminate the mechanisms which control replication fidelity.

Chapter 3

Extension to the Synthesizing Case: Inferring the Kinetic Rates of dNTP Binding and Incorporation

3.1 Introduction

In this chapter, we determine the dNTP binding, incorporation, and disassociation rates using dwell time data for synthesizing DNAP-DNA complexes. We derive the probability density function (PDF) underlying the dwell time data and determine the maximum-likelihood estimates (MLE) of the binding, incorporation, and disassociation rates. Previous work has been done to estimate the dNTP binding and disassociation rates in non-synthesizing DNAP-DNA complexes by use of a autocorrelation function of the entire current amplitude measured from nanopore experiments [50]. In the previous chapter, we also proposed a method of estimating the dNTP binding and disassociation rates in non-synthesizing complexes by deriving the underlying dwell time PDF and applying an expectation-maximization (EM) algorithm to obtain the MLE estimates. This chapter extends this to synthesizing case. Until now, inferring the binding, incorporation, and disassociation rates of synthesizing complexes have not yet been examined.

For an ionic current trace covering more than one nucleotide addition cycle, we define various dwell times (figure 3.1).

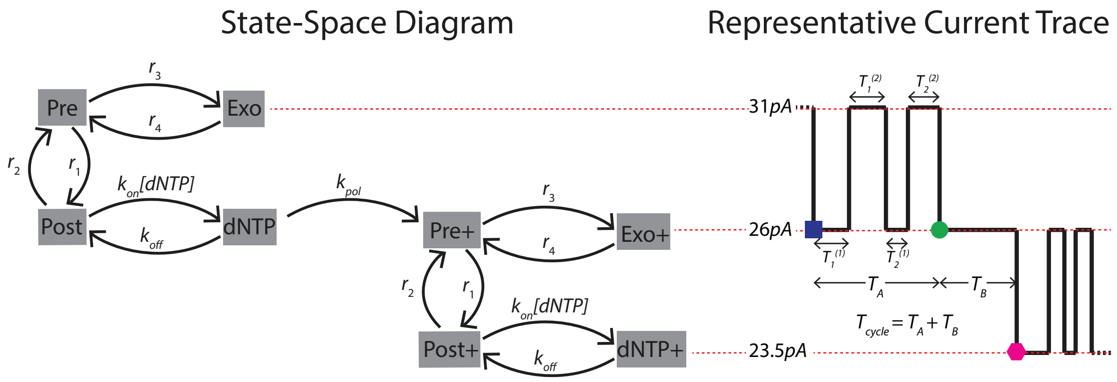

Figure 3.1: A state-space diagram for two nucleotide addition cycles in DNA replication. When the DNAP-DNA complex is allowed to undergo synthesis and a complementary dNTP is provided in the cis chamber, the DNAP-DNA complex can transition to the next nucleotide addition cycle–indicated by the "+" symbol after the state names. This is manifested as a change in the upper and lower amplitudes as the reporter group gets closer or further away from the nanopore lumen.

- T_A: the time from the first arrival to the post-translocation state of the current nucleotide addition cycle to the last arrival to the post-translocation state of the current nucleotide addition cycle; this is shown graphically as the blue square to the green circle in figure 3.1

- T_B: the time from the last arrival to the post-translocation state of the current nucleotide addition cycle to the first arrival to the post-translocation state of the next nucleotide addition cycle; this is shown graphically as the green circle to the magenta hexagon in figure 3.1

- $T^{(1)}$: the lower-amplitude dwell times within the T_A dwell time segment. In any observation of T_A, there are likely to be many samples of $T^{(1)}$ and we label them as $T_1^{(1)}, T_2^{(1)}, T_3^{(1)}, \ldots,$ etc (figure 3.1)

- $T^{(2)}$: the upper-amplitude dwell times within the T_A and T_B dwell time segments. Like $T^{(1)}$, there are likely to be many samples of $T^{(2)}$, so we label them as $T_1^{(2)}, T_2^{(2)}, \ldots,$ etc (figure 3.1). Note that the dwell time $T^{(2)}$ is not directly observable within the dwell time segment T_B. Within the T_B segment, this is denoted graphically as the left-opened cyan parenthesis to the right-opened cyan parenthesis in figure 3.5.

- T_{pol}: the time from the last arrival to the post-translocation state to the first arrival to the pre-translocation state in the next nucleotide addition cycle; this is the time that the DNAP-DNA complex completes the dNTP binding and incorporation steps. This is denoted by the green circle to the right-opened red parenthesis in figure 3.5.

In this chapter, we are interested in the case in which the DNAP-DNA complex is allowed to undergo synthesis. The DNAP-DNA complex will thus transition in discrete amplitude levels, each level corresponding to a nucleotide addition cycle. A mutation has been engineered into the exonuclease so that cleaving of the dNTP cannot occur, and hence the transition to the next nucleotide addition cycle is irreversible. We are interested in inferring the transition rates k_{on}, k_{off}, and k_{pol} from the $T^{(1)}$ and T_B data. In this situation, collecting $T^{(1)}$ samples is in competition with T_B in a sense that the probability of escaping to the next nucleotide addition cycle increases with the increase in dNTP concentration.

In [49], the transition rates k_{on} and k_{off} were inferred from the measured current trace data by use of the autocorrelation function of the measured current trace. In this situation, the DNAP-DNA complex was not allowed to undergo synthesis. In chapter 2, we re-examined this situation and inferred k_{on} and k_{off} by deriving the PDF of the lower-amplitude data, $T^{(1)}$. We then showed that the PDF is a proper exponential mixture in which the EM algorithm can be applied to determine the maximum-likelihood estimates (MLE) of the mixture parameters. The estimated mixture parameters can then be used to determine k_{on} and k_{off}.

Unlike in [49] and chapter 2, we are considering the case in which the DNAP-DNA complex is allowed to undergo synthesis and hence $k_{pol} > 0$. When the complex is allowed to undergo synthesis, the complex has a probability of incorporating the bound dNTP and proceeding to the next nucleotide addition cycle. The probability of incorporating the bound dNTP and proceeding to the next nucleotide addition cycle is determined by the transition rates k_{on}, k_{off}, k_{pol} and dNTP concentraion $[dNTP]$. The ability of the DNAP-DNA complex to escape to the pre-translocation state or to proceed to the next nucleotide addition cycle fundamentally changes the distribution of the lower-amplitude dwell time $T^{(1)}$. Derivation of the PDF of the $T^{(1)}$ in this setting must consider the possibility of the complex irreversibly escaping to the next nucleotide addition cycle. We will derive a new PDF for $T^{(1)}$ for this setting and show that it is still a proper exponential mixture. Thus the same methods used to determine the MLE estimates of the mixture weights from $T^{(1)}$ data via the EM algorithm as shown in chapter 2 can still be used.

In chapter 2, we characterized the inference uncertainty for k_{on} and k_{off} from $T^{(1)}$ data and show that the inference uncertainty of these kinetic rates can be controlled in experimental design. We also characterize the effect of noise on the inferred kinetic rates. We extend this to synthesizing DNAP-DNA complexes in this chapter. We first characterize the inference uncertainty of k_{on}, k_{off}, and k_{pol} and show that the inference uncertainty of these kinetic rates can also be controlled in experimental design. A simple extension of our noise study in the previous chapter characterizes the effect of noise on the inferred kinetic rates.

We will also examine the information content of T_B and show that the PDF of T_B is an improper mixture of four exponential modes. Here, we use the term "improper" to mean that one or more of the mixture weights are negative although the total sum of the weights still equal 1, and the overall mixture is still a PDF. The fact that some of the exponential weights are negative means that inference of the mixture parameters does not fit into the EM framework, and thus the MLE estimates from T_B have to be found by more naive approaches. Through numerical observation, there appears to be no advantage of using T_B over the $T^{(1)}$ dwell times.

3.2 Mathematical Formulations

In this section, we derive the PDFs of $T^{(1)}$ and T_B, as well as the mean cycle time.

Derivation of the PDF of $T^{(1)}$

Since the DNAP-DNA complex can transition to the next nucleotide addition cycle, the escape problem describing the dwell time $T^{(1)}$ is fundamentally different than what was derived in chapter 2. Consider figure 3.2 which shows the relevant states describing the $T^{(1)}$ dwell time. Throughout this section and the rest of the paper, the states may be referred to by their full name, abbreviated name, or number; for example, we will interchangeably refer to the post-translocation state of the current nucleotide addition cycle as "post" or 2. In this example, "post-translocation" is the full name of the state, "post" is the abbreviated name of the state, and 2 is the number assigned to that state. Any states in the next nucleotide addition cycle will have a "+" symbol appended to it; for example, "post+" or 2+ refers to the post-translocation state of the next nucleotide addition cycle (figures 3.2 and 3.4). The dwell time $T^{(1)}$ are the lower-amplitude dwell time segments of T_A, and so $T^{(1)}$ is the dwell time of the lower-amplitude conditioned on the event of escaping to the pre-translocation state when the complex starts in the post-translocation state. Throughout this section and the rest of this paper, let $X(t)$ denote the state of the Markov process at time t.

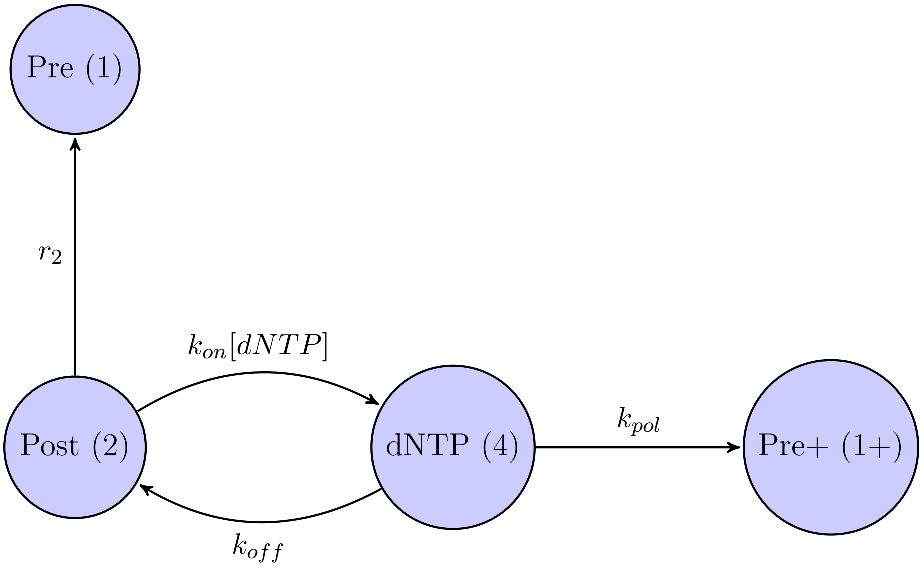

Figure 3.2: A state-space diagram of the relevant states for the escape problem pertaining to the T_1 and T_{pol} data.

It is helpful to define the following events:

$$\begin{aligned}
E_{pre}^{<t} &= \left\{X\left(t'\right) = pre, X\left(t\right) \neq pre, pre+ \ : \ 0 < r < t' < t\right\} \\
E_{pre}^{>t} &= \left\{X\left(t'\right) = pre, X\left(r\right) \neq pre, pre+ \ : \ 0 < r < t < r'\right\} \\
E_{pre}^{=t} &= \left\{X\left(t\right) = pre, X\left(r\right) \neq pre, pre+ \ : \ 0 < r < t\right\}, \\
E_{pre} &= \bigcup_{t>0} E_{pre}^{=t}.
\end{aligned}$$

Here, $E_{pre}^{<t}$ and $E_{pre}^{>t}$ are the events of the DNAP-DNA complex eventually escaping to the pre-translocation (pre) state of the current nucleotide addition cycle before and after time t, respectively. The event $E_{pre}^{=t}$ is the event of the complex escaping to the pre-translocation state of the current nucleotide addition cycle at exactly time t. Finally, E_{pre} is the event of the complex eventually escaping to the pre-translocation state of the current nucleotide addition cycle.

The following probabilities will be useful for our derivations:

$$p_{E_{pre}|2} = Pr\left(E_{pre} \mid X\left(0\right) = 2\right), \tag{3.1}$$

$$p_{E_{pre}|4} = Pr\left(E_{pre} \mid X\left(0\right) = 4\right). \tag{3.2}$$

Here $p_{E_{pre}|2}$ is the probability of escaping to the pre state provided that the DNAP-DNA complex starts at the post-translocation state (state 2), and $p_{E_{pre}|4}$ is the probability of escaping to the pre state provided that the DNAP-DNA complex starts at the dNTP-bound state (state 4). Note that only $p_{E_{pre}|2}$ is directly observable.

Consider the dwell time $\inf\left\{t \geq 0 \ : \ X\left(t\right) \neq \text{post, dNTP}\right\}$. This is the time-to-escape the lower-amplitude states: post-translocation and dNTP-bound. We can rigorously define the $T^{(1)}$ dwell time to be the following conditional random variable,

$$T^{(1)} = \inf\left\{t \geq 0 \ : \ X\left(t\right) \neq \text{post, dNTP}\right\} \ | \ \left\{E_{pre}, X\left(0\right) = post\right\}. \tag{3.3}$$

It is important to note that $T^{(1)}$ is a stochastic stopping time, and hence the Markov process before and after the dwell times can be thought of independent Markov processes by the strong Markov property (see for example [52] and [11]). In fact, it can be shown that any dwell time random variable is a stopping time; that is the random variable $\inf\left\{t \geq 0 \ : \ X\left(t\right) \notin \mathcal{U}\right\}$ where $\mathcal{U}$ denotes a subset of the state-space is a stopping time (see for example, page 119, example 7.2.2 of [59]). A consequence of this is that upon arrival of the complex at the post-translocation state, we can describe the escape problem underlying $T^{(1)}$ to be its own independent Markov process with state space {pre, post, dNTP, pre+} with the states {post, dNTP} transient and the states {pre,pre+} absorbing.

Let $X\left(t\right)$ be the state of the Markov process with state-space shown in figure 3.2. The infinitesimal generator Q of this Markov process is given by

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 \\ r_2 & -\left(r_2 + k_{on}[dNTP]\right) & k_{on}[dNTP] & 0 \\ 0 & k_{off} & -\left(k_{off} + k_{pol}\right) & k_{pol} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Now the transition matrix K of the embedded disrete-time Markov chain is given by

$$K = I + Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ r_2 & 1-(r_2 + k_{on}[dNTP]) & k_{on}[dNTP]\lambda & 0 \\ 0 & k_{off} & 1-(k_{off}+k_{pol}) & k_{pol} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We can write K in a canonical form

$$K = \begin{pmatrix} \mathcal{I} & \mathcal{O} \\ R & A \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ r_2 & 0 & 1-(r_2 + k_{on}[dNTP]) & k_{on}[dNTP] \\ 0 & k_{pol} & k_{off} & 1-(k_{off}+k_{pol}) \end{pmatrix}, \quad (3.4)$$

where R is the 2×2 matrix that gives the probability of transitioning from a transient state to an absorbing state; A is the 2×2 matrix that gives the probability of staying in a transient state; $\mathcal{I}$ is the identity matrix; and $\mathcal{O}$ is the zero matrix.

It can be shown that for a time-homogeneous, discrete-time Markov chain, the probability of being absorbed in absorbing state j from transient state i is given by $(\mathcal{I} - A)^{-1} R$ (theorem 3.3.7 page 52 of [42]). Computing this, we obtain

$$(\mathcal{I} - A)^{-1} R = \frac{1}{(k_{off}+k_{pol})\, r_2 + k_{pol} k_{on}[dNTP]} \begin{pmatrix} r_2\,(k_{off}+k_{pol}) & k_{on}[dNTP]k_{pol} \\ r_2 k_{off} & k_{pol}\,(r_2 + k_{on}[dNTP]) \end{pmatrix}. \quad (3.5)$$

Hence we can write,

$$p_{E_{pre}|2} = \frac{r_2\left(k_{off} + k_{pol}\right)}{\left(k_{off} + k_{pol}\right) r_2 + k_{pol} k_{on}[dNTP]}, \tag{3.6}$$

$$p_{E_{pre}|4} = \frac{r_2 k_{off}}{\left(k_{off} + k_{pol}\right) r_2 + k_{pol} k_{on}[dNTP]}. \tag{3.7}$$

For notational convenience, let I be the state-space of the Markov process; that is, $I = \{1, 2, 4, 1+\}$. Consider the conditional transition matrix, $P_{E_{pre}}(t) = \left(P_{i,j,E_{pre}}(t)\right)_{i,j \in I \times I}$, where

$$P_{i,j,E_{pre}}(t) = Pr\left(X(t) = j \mid X(0) = i, E_{pre}\right).$$

From here, we can obtain the conditional infinitesimal generator

$$Q_{E_{pre}} = \lim_{t \to 0^+} \frac{P_{E_{pre}}(t) - \mathcal{I}}{t},$$

for the Markov process governing the lower-amplitude escape problem conditioned on the escape to the pre-translocation state. We now derive the entries of the conditional transition matrix $P_{E_{pre}}(t)$.

$$\begin{aligned}
P_{2,4,E_{pre}}(t) &= Pr\left(X(t) = 4 \mid X(0) = 2, E_{pre}\right) \\
&= \frac{Pr\left(X(t) = 4, X(0) = 2, E_{pre}\right)}{Pr\left(E_{pre} \mid X(0)\right) Pr\left(X(0) = 2\right)} \\
&= \frac{Pr\left(E_{pre}, X(t) = 4 \mid X(0) = 2\right) Pr\left(X(0) = 2\right)}{Pr\left(E_{pre} \mid X(0) = 2\right) Pr\left(X(0) = 2\right)} \\
&= \frac{Pr\left(E_{pre}^{>t}, X(t) = 4 \mid X(0) = 2\right)}{Pr\left(E_{pre} \mid X(0) = 2\right)} \qquad \text{since } \{X(t) = 4\} \cap E_{pre} = E_{pre}^{>t} \\
&= \frac{Pr\left(E_{pre}^{>t} \mid X(t) = 4, X(0) = 2\right) Pr\left(X(t) = 4 \mid X(0) = 2\right)}{Pr\left(E_{pre} \mid X(0) = 2\right)} \\
&= \frac{Pr\left(E_{pre}^{>t} \mid X(t) = 4\right) Pr\left(X(t) = 4 \mid X(0) = 2\right)}{Pr\left(E_{pre} \mid X(0) = 2\right)} \\
&\qquad \text{from the Markov property} \\
&= \frac{\left(k_{on}[dNTP]t\right) p_{E_{pre}|4}}{p_{E_{pre}|2}} + o(t).
\end{aligned}$$

The entries $P_{4,2,E_{pre}}(t)$ and $P_{2,1,E_{pre}}(t)$ can be calculated in a similar manner.

$$\begin{aligned}
P_{4,2,E_{pre}}(t) &= Pr\left(X(t) = 2 \mid X(0) = 4, E_{pre}\right) \\
&= \frac{Pr\left(X(t) = 2, X(0) = 4, E_{pre}\right)}{Pr\left(E_{pre} \mid X(0) = 4\right) Pr\left(X(0) = 4\right)} \\
&= \frac{Pr\left(E_{pre}, X(t) = 2 \mid X(0) = 4\right) Pr\left(X(0) = 4\right)}{Pr\left(E_{pre} \mid X(0) = 4\right) Pr\left(X(0) = 4\right)} \\
&= \frac{Pr\left(E_{pre}, X(t) = 2 \mid X(0) = 4\right)}{Pr\left(E_{pre} \mid X(0) = 4\right)} \\
&= \frac{Pr\left(E_{pre}^{>t}, X(t) = 2 \mid X(0) = 4\right)}{p_2} \\
&= \frac{\left(k_{off}t\right) p_{E_{pre}|2}}{p_{E_{pre}|4}} + o(t).
\end{aligned}$$

$$\begin{aligned}
P_{2,1,E_{pre}}(t) &= Pr\left(X(t)=1 \mid X(0)=2, E_{pre}\right) \\
&= \frac{Pr\left(X(t)=1, X(0)=2, E_{pre}\right)}{Pr\left(E_{pre} \mid X(0)=2\right) Pr\left(X(0)=2\right)} \\
&= \frac{Pr\left(E_{pre}, X(t)=1 \mid X(0)=2\right) Pr\left(X(0)=2\right)}{Pr\left(E_{pre} \mid X(0)=2\right) Pr\left(X(0)=2\right)} \\
&= \frac{Pr\left(E_{pre}^{<t} \mid X(0)=2\right)}{Pr\left(E_{pre} \mid X(0)=2\right)} \\
&= \frac{r_2 t}{p_{E_{pre}|2}} + o(t).
\end{aligned}$$

Clearly, $P_{4,1+,E_{pre}} = 0$ since we are conditioning on E_{pre}. Also, $P_{2,1+,E_{pre}}(t) = o(t)$ and $P_{4,1,E_{pre}}(t) = o(t)$ since the probability of two or more transitions occuring in an interval $[0,t]$ is $o(t)$. And clearly, $P_{1,j,E_{pre}}(t) = 0$ for all $j \neq 1$ and $P_{1+,j,E_{pre}}(t) = 0$ for all $j \neq 1+$ since states 1 and 1+ are absorbing, and consequently $P_{1,1,E_{pre}}(t) = P_{1+,1+,E_{pre}}(t) = 1$. And finally, $P_{2,2,E_{pre}}(t) = 1 - \frac{r_2 t}{p_{E_{pre}|2}} - \frac{k_{on}[dNTP]p_{E_{pre}|4}t}{p_{E_{pre}|2}} + o(t)$, and $P_{4,4,E_{pre}}(t) = 1 - \frac{k_{off}p_{E_{pre}|2}t}{p_{E_{pre}|4}} + o(t)$. Hence the conditional infinitesimal generator matrix is given by

$$Q_{E_{pre}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{r_2}{p_{E_{pre}|2}} & -\frac{r_2 + k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} & \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} & 0 \\ 0 & \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}} & -\frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.8}$$

The state-space diagram for the Markov process generated by $Q_{E_{pre}}$ is shown in figure 3.3. The time to absorption of the DNAP-DNA complex starting in the post-translocation state of this Markov process is equivalent to the general escape problem shown in Proposition 1 of Chapter 2. Hence the PDF of $T^{(1)}$, $f_{T^{(1)}}(t)$, is

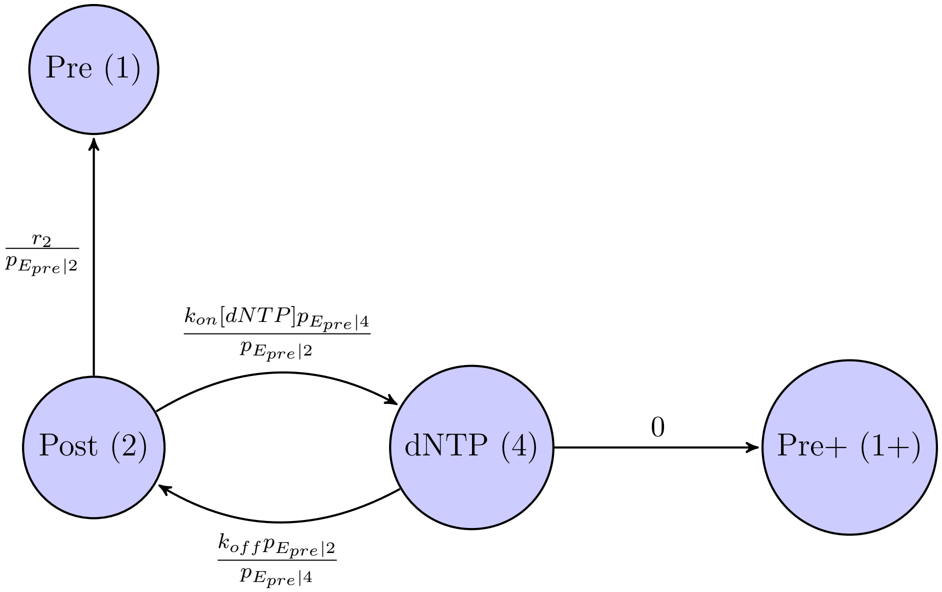
Pre (1)
$\frac{r_2}{p_{E_{pre}|2}}$
$\frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}}$
Post (2)
dNTP (4)
0
Pre+ (1+)
$\frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}}$

Figure 3.3: State-space diagram of the lower amplitude states conditioned on E_{pre}.

given by

$$f_{T^{(1)}} = \alpha\lambda_1 e^{-\lambda_1 t} + (1-\alpha)\,\lambda_2 e^{-\lambda_2 t}, \tag{3.9}$$

with

$$\lambda_{1,2} = \frac{B \pm \sqrt{B^2 - 4\frac{r_2 k_{off}}{p_{E_{pre}|4}}}}{2},$$

where

$$B = \frac{r_2}{p_{E_{pre}|2}} + \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} + \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}},$$

for notational compactness. We order the eigenvalues to be $\lambda_1 < \lambda_2$. Also, $\alpha = (\lambda_2 - r_2)\,/\,(\lambda_2 - \lambda_1)$ and $0 < \alpha < 1$. Hence $f_{T^{(1)}}$ is a proper exponential mixture.

Derivation of the PDF of T_B

Consider figure 3.4 which shows the relevant states for the escape problem describing T_B. Recall that T_B is the time for the DNAP-DNA complex to escape to the post-translocation state of the next nucleotide addition cycle (post+) when starting at the post-translocation state of the current nucleotide addition cycle (post). We can write $T_B = T_{pol} + T^{(2)}$ where T_{pol} is the time it takes the DNAP-DNA complex to complete the binding and incorporation segment of the nucleotide addition cycle and $T^{(2)}$ is the upper-amplitude segment of the next nucleotide addition cycle (figure 3.5). Unlike T_A, T_B, and $T^{(1)}$, the dwell times T_{pol} and $T^{(2)}$ are both unobservable. The latter is only unobservable when it is part of the T_B segment. Otherwise, $T^{(2)}$ data can be directly observed in between the

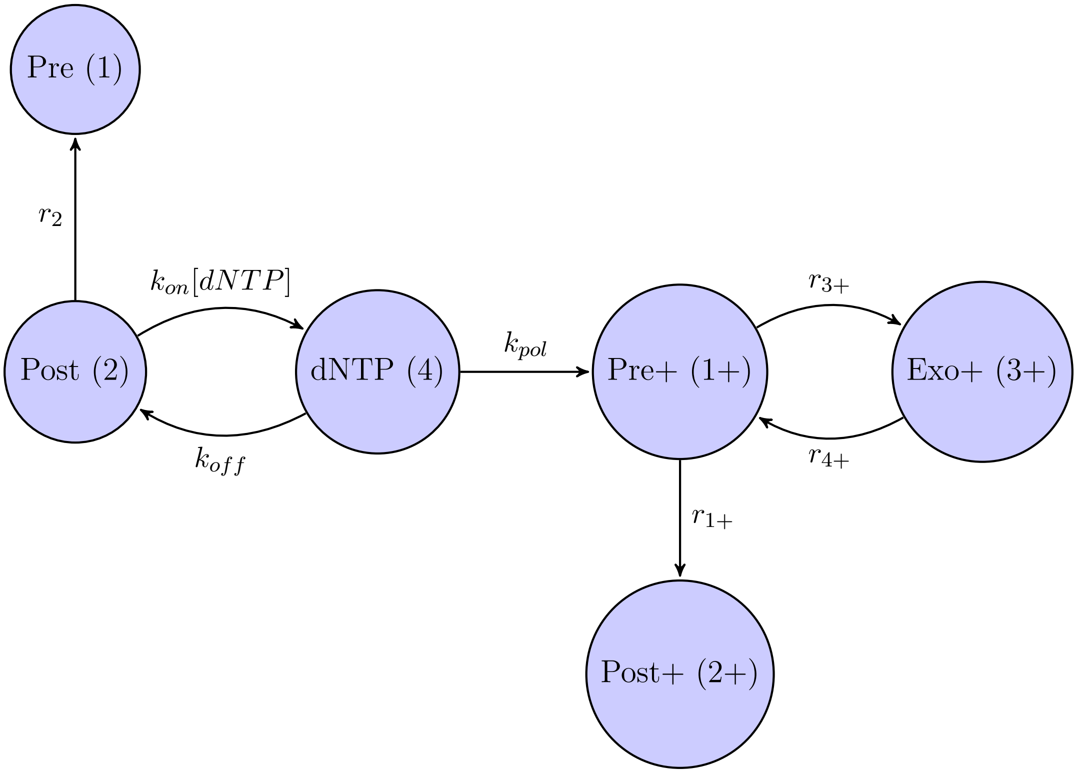

Figure 3.4: A state-space diagram of the relevant states for the escape problem pertaining to the T_B data.

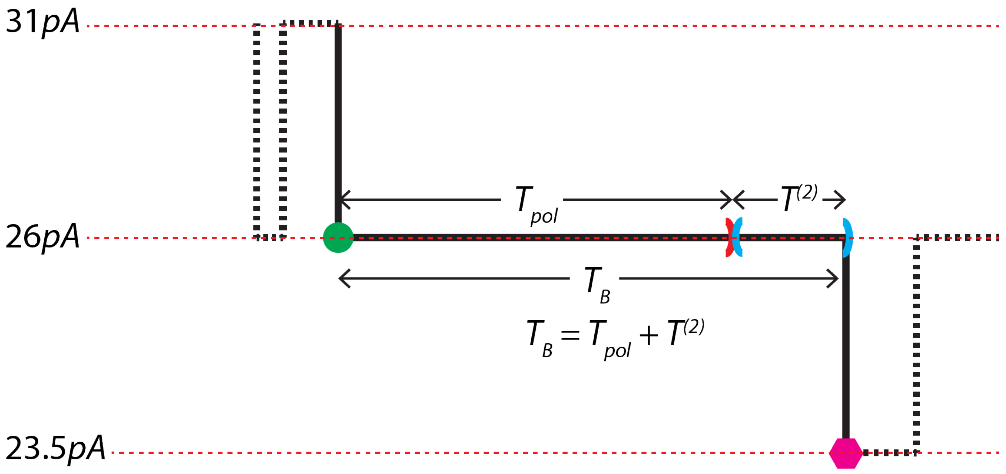

Figure 3.5: A schematic current trace depicting the T_B dwell time along with T_{pol} and $T^{(2)}$ dwell time segments which make up the T_B dwell time. The T_{pol} dwell time segment is graphically denoted from the green circle to the right-opened red parenthesis, and the $T^{(2)}$ dwell time segment is graphically denoted from the left-opened cyan parenthesis to the right-open cyan parenthesis. Both dwell times T_{pol} and $T^{(2)}$ are not directly observable since they do not manifest a change in current amplitude. The latter is only unobservable when part of the T_B dwell time segment.

DNAP-DNA transitions across the translocation in a similar manner in which $T^{(1)}$ data can be collected (figure 3.1). It is also important to note that T_{pol} and $T^{(2)}$ are stochastic stopping times, and hence the Markov process before and after the dwell times can be thought of independent Markov processes by the strong Markov property. An important consequence of this is that the random variables T_{pol} and $T^{(2)}$ are independent and hence T_B is a sum of two independent random variables. Using this fact, we can determine the PDF of T_B by determining the PDF of T_{pol} and $T^{(2)}$, and then using the fact that the PDF of a sum of two independent random variables in the convolution of their PDFs.

Derivation of the PDF of T_{pol}

To derive the PDF of T_{pol}, it is helpful to define the following events:

$$\begin{aligned}
E_{pol}^{<t} &= \{X(t') = pre+, X(t) \neq pre, pre+ \ : \ 0 < r < t' < t\} \\
E_{pol}^{>t} &= \{X(t') = pre+, X(r) \neq pre, pre+ \ : \ 0 < r < t < r'\} \\
E_{pol}^{=t} &= \{X(t) = pre+, X(r) \neq pre, pre+ \ : \ 0 < r < t\}, \\
E_{pol} &= \bigcup_{t>0} E_{pol}^{=t}.
\end{aligned}$$

Informally, the events $E_{pol}^{<t}$ and $E_{pol}^{>t}$ are the events of the DNAP-DNA complex escaping to the pre-translocation state of the next nucleotide addition cycle (pre+) before and after time t, respectively. The event $E_{pol}^{=t}$ is the event of the DNAP-DNA complex escaping to the pre+ state at exactly time t, and the event E_{pol} is the event of the complex eventually escaping to the pre+ state.

Define the following probabilities

$$p_{E_{pol}|2} = Pr\left(E_{pol} \mid X(0) = 2\right), \tag{3.10}$$

$$p_{E_{pol}|4} = Pr\left(E_{pol} \mid X(0) = 4\right). \tag{3.11}$$

Here $p_{E_{pol}|2}$ is the probability of escaping to the pre+ state provided that the DNAP-DNA complex starts at the post-translocation state (state 2), and $p_{E_{pol}|4}$ is the probability of escaping to the pre+ state provided that the DNAP-DNA complex starts at the dNTP-bound state (state 4).

Similar to the definition of $T^{(1)}$ in equation 3.3, we can define T_{pol} to be

$$T_{pol} = \inf\{t \geq 0 \ : \ X(t) \neq \text{Post, dNTP}\} \mid \{E_{pol}, X(0) = \text{Post}\}.$$

The state-space diagram relevant for this random variable is shown in figure 3.2.

For T_{pol} we are conditioning on the escape to the Pre+ state. Ximilar to our strategy for deriving the PDF of $T^{(1)}$, we will first derive the conditional transition matrix for the embedded discrete-time Markov process, conditioned on the escape to the Pre+ state. From equation 3.5, we have that

$$p_{E_{pol}|2} = \frac{k_{on}[dNTP]k_{pol}}{\left(k_{off}+k_{pol}\right)r_2 + k_{pol}k_{on}[dNTP]}, \tag{3.12}$$

$$p_{E_{pol}|4} = \frac{k_{pol}\left(r_2 + k_{on}[dNTP]\right)}{\left(k_{off}+k_{pol}\right)r_2 + k_{pol}k_{on}[dNTP]}. \tag{3.13}$$

We want to calculate the conditional transition matrix,

$$P_{E_{pol}}\left(t\right) = \left(P_{i,j,E_{pol}}\left(t\right)\right)_{i,j\in I},$$

where

$$P_{i,j,E_{pol}}\left(t\right) = Pr\left(X\left(t\right) = j \mid X\left(0\right) = i, E_{pol}\right).$$

$P_{2,4,E_{pol}}(t)$

$$\begin{aligned}
P_{2,4,E_{pol}}(t) &= Pr\left(X(t) = 4 \mid X(0) = 2, E_{pol}\right) \\
&= \frac{Pr\left(X(t) = 4, X(0) = 2, E_{pol}\right)}{Pr\left(X(0) = 2, E_{pol}\right)} \\
&= \frac{Pr\left(X(t) = 4, E_{pol} \mid X(0) = 2\right)}{p_{E_{pol}|2}} \\
&= \frac{Pr\left(E_{pol}^{>t}, X(t) = 4 \mid X(0) = 2\right)}{p_{E_{pol}|2}} \quad \text{since } \{X(t) = 4\} \cap E_{pol} = E_{pol}^{>t} \\
&= \frac{Pr\left(E_{pol}^{>t} \mid X(t) = 4, X(0) = 2\right) Pr\left(X(t) = 4 \mid X(0) = 2\right)}{p_{E_{pol}|2}} \\
&= \frac{Pr\left(E_{pol}^{>t} \mid X(t) = 4\right) Pr\left(X(t) = 4 \mid X(0) = 2\right)}{p_{E_{pol}|2}} \\
&= \frac{p_{E_{pol}|4} k_{on}[dNTP]t}{p_{E_{pol}|2}} + o(t).
\end{aligned}$$

$P_{4,2,E_{pol}}(t)$

$$\begin{aligned}
P_{4,2,E_{pol}}(t) &= Pr\left(X(t) = 2 \mid X(0) = 4, E_{pol}\right) \\
&= \frac{Pr\left(X(t) = 2, X(0) = 4, E_{pol}\right)}{Pr\left(E_{pol} \mid X(0) = 4\right) Pr\left(X(0) = 4\right)} \\
&= \frac{Pr\left(E_{pol}, X(t) = 2 \mid X(0) = 4\right)}{Pr\left(E_{pol} \mid X(0) = 4\right)} \\
&= \frac{Pr\left(E_{pol}^{>t}, X(t) = 2 \mid X(0) = 4\right)}{p_{E_{pol}|4}} \\
&= \frac{Pr\left(X_{pol}^{>t} \mid X(t) = 2\right) Pr\left(X(t) = 2 \mid X(0) = 4\right)}{p_{E_{pol}|4}} \\
&= \frac{p_{E_{pol}|2} k_{off} t}{p_{E_{pol}|4}} + o(t).
\end{aligned}$$

$P_{4,1+,E_{pol}}\left(t\right)$

$$\begin{aligned}
P_{4,1+,E_{pol}}\left(t\right) &= Pr\left(X\left(t\right)=1+\ \mid X\left(0\right)=4, E_{pol}\right)\\
&= \frac{Pr\left(E_{pol}, X\left(t\right)=1+\ \mid X\left(0\right)=4\right)}{Pr\left(E_{pol}\mid X\left(0\right)=4\right)}\\
&= \frac{Pr\left(E_{pol}^{<t}, X\left(t\right)=1+\ \mid X\left(0\right)=4\right)}{p_{E_{pol}|4}}\\
&= \frac{Pr\left(E_{pol}^{<t}\mid X\left(t\right)=1+\right)Pr\left(X\left(t\right)=1+\ \mid X\left(0\right)=4\right)}{p_{E_{pol}|4}}\\
&= \frac{k_{pol}t}{p_{E_{pol}|4}}+o\left(t\right).
\end{aligned}$$

Now, $P_{2,1,E_{pol}}\left(t\right) = P_{4,1,E_{pol}}\left(t\right) = 0$ since we are conditioning on E_{pol}. Also $P_{2,1+,E_{pol}}\left(t\right) = o\left(t\right)$ since the probability of transitioning twice in $[0,t]$ is $o\left(t\right)$. For the absorbing states, we have $P_{1,1,E_{pol}}\left(t\right) = P_{1+,1+,E_{pol}}\left(t\right) = 1$, and $P_{1,j,E_{pol}}\left(t\right) = P_{1+,j,E_{pol}}\left(t\right) = 0$ for $j = 1,\ldots,4$. Finally,

$$\begin{aligned}
P_{2,2,E_{pol}}\left(t\right) &= 1-\frac{p_{E_{pol}|4}k_{on}[dNTP]t}{p_{E_{pol}|2}}+o\left(t\right)\\
P_{4,4,E_{pol}}\left(t\right) &= 1-\frac{\left(p_{E_{pol}|2}k_{off}+k_{pol}\right)t}{p_{E_{pol}|4}}+o\left(t\right).
\end{aligned}$$

Hence the conditional infinitesimal generator matrix of the underlying continuous-time process is given by

$$
\begin{aligned}
Q_{E_{pol}} &= \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & -\frac{p_{E_{pol}|4} k_{on}[dNTP]}{p_{E_{pol}|2}} & \frac{p_{E_{pol}|4} k_{on}[dNTP]}{p_{E_{pol}|2}} & 0 \\
0 & \frac{p_{E_{pol}|2} k_{off}}{p_{E_{pol}|4}} & -\frac{p_{E_{pol}|2} k_{off} + k_{pol}}{p_{E_{pol}|4}} & \frac{k_{pol}}{p_{E_{pol}|4}} \\
0 & 0 & 0 & 0
\end{pmatrix} \\
&= \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & -\Lambda & \Lambda & 0 \\
0 & \frac{k_{off} k_{on}[dNTP]}{\Lambda} & -\frac{k_{off} k_{on}[dNTP]}{\Lambda} - \left(k_{pol} + \frac{k_{off} r_2}{\Lambda}\right) & k_{pol} + \frac{k_{off} r_2}{\Lambda} \\
0 & 0 & 0 & 0
\end{pmatrix}
\end{aligned}
$$

where $\Lambda = r_2 + k_{on}[dNTP]$ for notational compactness.

This gives rise to the Markov process with state-space diagram in figure 3.6. The escape problem shown in figure 3.6 is similar to the one underlying $T^{(1)}$. We

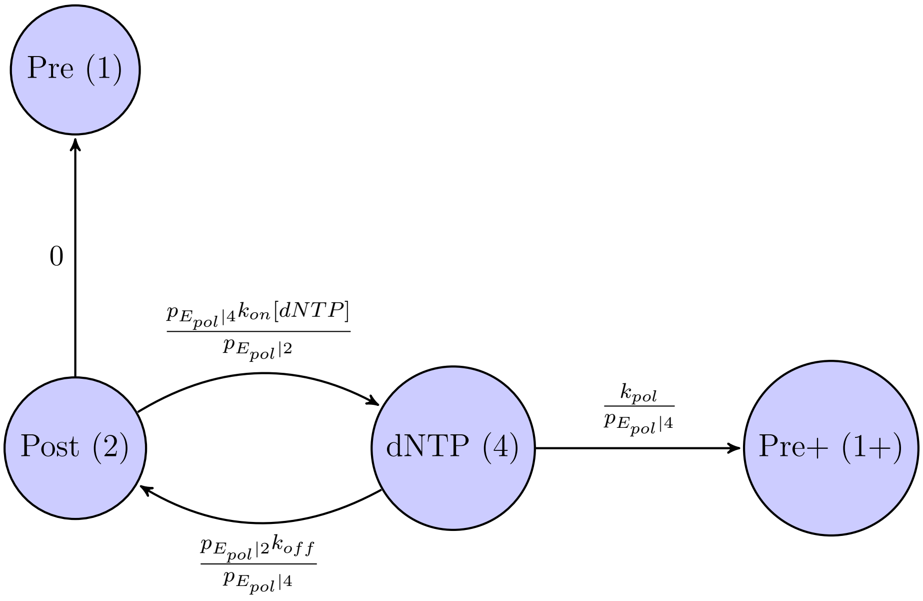
Pre (1)
0
$\frac{p_{E_{pol}|4} k_{on}[dNTP]}{p_{E_{pol}|2}}$
Post (2)
dNTP (4)
$\frac{k_{pol}}{p_{E_{pol}|4}}$
Pre+ (1+)
$\frac{p_{E_{pol}|2} k_{off}}{p_{E_{pol}|4}}$

Figure 3.6: A state-space diagram of the Markov process conditioned on the escape to the Pre+ state.

can derive the PDF of T_{pol} in a similar manner to $T^{(1)}$ as shown in Proposition 1 of Chapter 2. The key difference is that the escape to the absorbing state is sequential for T_{pol}, whereas for $T^{(1)}$, the absorbing state is a branch (compare figures 3.3 and 3.6).

Consider figure 3.7. Let T be the dwell time that it takes for the Markov process to escape to state 3, starting from state 1 in figure 3.7. We will derive the PDF of T below in a similar manner to Proposition 1 in Chapter 2.

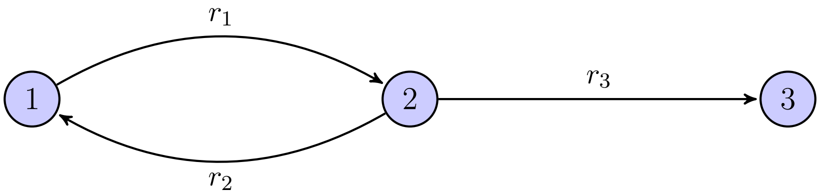

Figure 3.7: A state-space diagram of the general escape problem for a sequential chain with two transient states and one absorbing state.

Proposition 3. *The PDF of T is of the form $\alpha\lambda_1 e^{-\lambda_1 t} + (1-\alpha)\lambda_2 e^{-\lambda_2 t}$ with $\lambda_1, \lambda_2 > 0$ and $\alpha > 0$.*

Proof. By Kolmogorov's backwards equation, we have the following system of ODEs,

$$\frac{d}{dt}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} -r_1 & r_2 \\ r_1 & -(r_2+r_3) \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}.$$

The characteristic equation is given by

$$\lambda^2 - (r_1+r_2+r_3)\lambda + r_1 r_3 = 0 \tag{3.14}$$

This gives us the eigenvalues

$$\lambda_{1,2} = \frac{(r_1+r_2+r_3) \pm \sqrt{(r_1+r_2+r_3)^2 - 4r_1r_3}}{2} \tag{3.15}$$

Now from the arithmetic mean-geometric mean (AM-GM) inequality, $(r_1 + r_3)/2 \geq \sqrt{r_1 r_3}$ with equality if and only if $r_1 = r_3$. Thus we have that $(r_1 + r_3)^2 \geq 4r_1 r_3$ and so $(r_1 + r_2 + r_3)^2 > 4r_1 r_3$ since $r_1, r_2, r_3 > 0$. Also, by Descartes' Rule of Signs, both roots of the quadratic equation 3.14 are both positive. Hence $p_j(t)$ is of the form of

$$p_j(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t}$$

Thus the total probability of the states 1 and 2 is $p_1(t) + p_2(t)$, which is of the form

$$p_1(t) + p_2(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t}$$

The PDF of the dwell time is given by

$$\begin{aligned} \rho(t) &= -\frac{d}{dt}\left(p_1(t) + p_3(t)\right) \\ &= c_1 \lambda_1 e^{-\lambda_1 t} + c_2 \lambda_2 e^{-\lambda_2 t} \end{aligned}$$

Now since $\lambda_1, \lambda_2 > 0$, we have that $1 = \int_0^\infty \rho(t)\, dt = c_1 + c_2$. Thus $c_1 = \alpha$ and $c_2 = 1 - \alpha$ for some $\alpha \in \mathbb{R}$. Hence we have

$$\rho(t) = \alpha \lambda_1 e^{-\lambda_1 t} + (1 - \alpha)\, \lambda_2 e^{-\lambda_2 t}$$

To find α, consider

$$\begin{aligned}
\rho(0) &= \alpha\lambda_1 + (1-\alpha)\,\lambda_2 \\
&= -\left.\frac{d}{dt}\,(p_1(t)+p_2(t))\right|_{t=0} \\
&= -\frac{dp_1}{dt}(0) + \frac{dp_2}{dt}(0) \\
&= p_1\,(0)\,r_2 - p_2\,(0)\,r_2 + (r_2+r_3)\,p_2\,(0) - p_1\,(0)\,r_2 \\
&= 0
\end{aligned}$$

The last equality is true since $p_1(0) = 1$ and $p_2\,(0) = 0$ (the dwell time starts each time the complex arrives at the pre-translocation state from the post-translocation state in segment T_A). Thus we have

$$\alpha\lambda_1 + (1-\alpha)\,\lambda_2 = 0$$

Solving for α gives us

$$\alpha = \frac{\lambda_2}{\lambda_2 - \lambda_1}. \tag{3.16}$$

□

Hence we can writ the PDF of T_{pol}, $f_{T_{pol}}\,(t)$ in the form

$$f_{T_{pol}}\,(t) = \beta\eta_1 e^{-\eta_1 t} + (1-\beta)\,\eta_2 e^{-\eta_2 t},$$

with

$$\eta_{1,2} = \frac{B \pm \sqrt{B^2 - 4\frac{k_{on}[dNTP]k_{pol}}{p_{E_{pol}|2}}}}{2},$$

where

$$B = \frac{p_{E_{pol}|4}k_{on}[dNTP]}{p_{E_{pol}|2}} + \frac{p_{E_{pol}|2}k_{off}}{p_{E_{pol}|4}} + \frac{k_{pol}}{p_{E_{pol}|4}},$$

for notational compactness. The change in notation relative to what was used in Proposition 3 will be come clear after we derive the PDF of $T^{(2)}$ in the next subsection. Notice that we can write the PDF of T_{pol} as

$$f_{T_{pol}}(t) = \frac{\eta_1 \eta_2}{\eta_2 - \eta_1} e^{-\eta_1 t} - \frac{\eta_1 \eta_2}{\eta_2 - \eta_1} e^{-\eta_2 t},$$

so T_{pol} is a two-parameter hypoexponential distribution [10].

Derivation of the PDF of $T^{(2)}$

Recall that $T^{(2)}$ is the dwell time of the upper-amplitude. In this context, it is unobservable since it is part of the T_B dwell time segment and is not marked by a change in observed current amplitude (figure 3.5). Consider the state-space diagram shown in figure 3.8

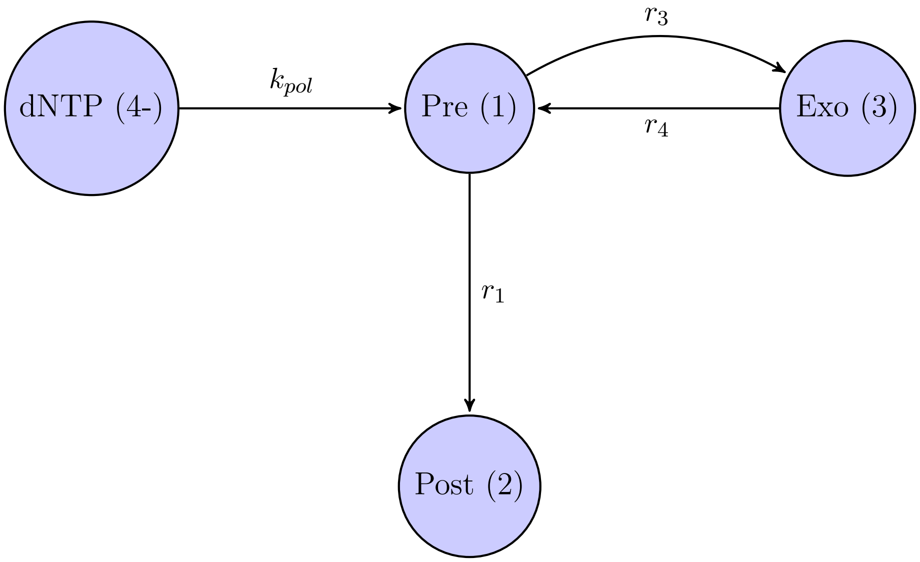

Figure 3.8: State-space diagram of the upper-amplitude and relevant states for the $T^{(2)}$ dwell time. The "-" symbol appended to the end of the dNTP-bound state means that the dNTP-bound state in this context is part of the previous nucleotide addition cycle.

We can define the dwell time $T^{(2)}$ as the following,

$$T^{(2)} = \inf \left\{ S\left(t\right) \neq \text{Pre, Exo} \right\} \mid \left\{ S\left(0\right) = \text{Pre} \right\}.$$

Note that once the complex enters the Pre state, the only way to escape to the lower-amplitude is transitioning to the Post state.

Notice that the state-space for the escape problem underlying $T^{(2)}$ is equivalent to the general escape problem shown in Proposition 1 in 2. Hence we can conclude that the PDF of $T^{(2)}$, $f_{T^{(2)}}$ is a proper exponential mixture in which we can write it in the form,

$$f_{T^{(2)}} = \gamma \mu_1 e^{-\mu_1 t} + \left(1 - \gamma\right) \mu_2 e^{-\mu_2 t}. \tag{3.17}$$

Derivation of the PDF of T_B

We are now ready to derive the PDF of T_B. Recall that $T_B = T_{pol} + T^{(2)}$ and that T_{pol} and $T^{(2)}$ are independent. Hence, the PDF of T_B, $f_{T_B}\left(t\right)$ is given by the convolution $f_{T_B}\left(t\right) = \left(f_{T_{pol}} * f_{T^{(2)}}\right)\left(t\right)$. By direct calculation, we find that

$$\begin{aligned} f_{T_B}\left(t\right) &= \int_0^t f_{T_{pol}}\left(t-\tau\right) f_{T^{(2)}}\left(\tau\right) d\tau \\ &= \left(\frac{\beta\gamma\eta_1\mu_1}{\eta_1 - \mu_1} + \frac{\left(1-\beta\right)\gamma\eta_2\mu_1}{\eta_2 - \mu_1}\right) e^{-\mu_1 t} \\ &\quad + \left(\frac{\beta\left(1-\gamma\right)\eta_1\mu_2}{\eta_1 - \mu_2} + \frac{\left(1-\beta\right)\left(1-\gamma\right)\eta_2\mu_2}{\eta_2 - \mu_2}\right) e^{-\mu_2 t} \\ &\quad - \left(\frac{\beta\gamma\eta_1\mu_1}{\eta_1 - \mu_1} + \frac{\beta\left(1-\gamma\right)\eta_1\mu_2}{\eta_1 - \mu_2}\right) e^{-\eta_1 t} \\ &\quad - \left(\frac{\left(1-\beta\right)\gamma\eta_2\mu_1}{\eta_2 - \mu_1} + \frac{\left(1-\beta\right)\left(1-\gamma\right)\eta_2\mu_2}{\eta_2 - \mu_2}\right) e^{-\eta_2 t}. \end{aligned}$$

Note that the coefficients of the exponential modes sum up to 1, but some of the coefficients are negative. Hence we can write the PDF of T_B as an improper mixture of four exponential modes,

$$f_{T_B}(t) = \Omega_1 e^{-\mu_1 t} + \Omega_2 e^{-\mu_2 t} + \Omega_3 e^{-\eta_1 t} + \Omega_4 e^{-\eta_2 t}, \tag{3.18}$$

where each Ω_i is their respective coefficient from above and $\Omega_1+\Omega_2+\Omega_3+\Omega_4 = 1$.

Derivation of the Mean Cycle Time

Let $T_{cycle} = T_A + T_B$ be the time it takes to complete a nucleotide addition cycle (figure 3.1). We will see that the mean cycle time plays a useful role in the inference of the transition rates k_{on}, k_{off}, and k_{pol}. The T_A dwell time segment consists of a random sum of $T^{(1)}$ and $T^{(2)}$ (figure 3.1),

$$T_A = \sum_{i=1}^{N_{T^{(1)}}} T_i^{(1)} + T_i^{(2)},$$

where $N_{T^{(1)}}$ is the number of $T^{(1)}$ samples per cycle. Hence we can write

$$T_{cycle} = \sum_{i=1}^{N_{T^{(1)}}} \left(T_i^{(1)} + T_i^{(2)}\right) + T_B.$$

From the law of total expectation, we can write

$$\begin{aligned}
\langle T_{cycle}\rangle &= \sum_{n=0}^{\infty} \langle T_{cycle} \mid N_{T^{(1)}} = n\rangle\, Pr\left(N_{T^{(1)}} = n\right) \\
&= \sum_{n=0}^{\infty} \left[n\left(\left\langle T^{(1)}\right\rangle + \langle T_{upper}\rangle\right) + \langle T_B\rangle\right] p_{E_{pre}|2}^{m}\left(1 - p_{E_{pre}|2}\right) \\
&= \left(\left\langle T^{(1)}\right\rangle + \left\langle T^{(2)}\right\rangle\right) \frac{p_{E_{pre}|2}}{1 - p_{E_{pre}|2}} + \langle T_B\rangle\,,
\end{aligned}$$

where the second equality comes from the fact that $N_{T^{(1)}} \sim$ geometric $\left(p_{E_{pol}|2}\right)$ where the geometric distribution has support $\mathbb{N} \cup \{0\}$ and $p_{E_{pol}|2} = 1 - p_{E_{pre}|2}$.

The calculation of $\left\langle T^{(2)}\right\rangle$ is straight-forward. Recall from equation 3.17 that the PDF of $T^{(2)}$ is a proper mixture of two exponential modes,

$$f_{T^{(2)}} = \gamma\mu_1 e^{-\mu_1 t} + (1-\gamma)\,\mu_2 e^{-\mu_2 t}.$$

Hence, by direct computation, $\left\langle T^{(2)}\right\rangle = \gamma/\mu_1 + (1-\gamma)\,\mu_2$.

Derivation of $\left\langle T^{(1)}\right\rangle$

To obtain the expression for $\left\langle T^{(1)}\right\rangle$, we could integrate $tf_{T^{(1)}}(t)$, but there is a cleaner way. Let $T_{post,j}$ and $T_{dNTP,j}$ be the dwell times of the post-translocation and the dNTP-bound states, respectively. Then,

$$\begin{aligned} T_{post,j} &\overset{\text{iid}}{\sim} \exp\left(r_2 + k_{on}[dNTP]\right), \\ T_{dNTP,j} &\overset{\text{iid}}{\sim} \exp\left(k_{off} + k_{pol}\right). \end{aligned}$$

Throughout the rest of this paper, let $Y(n)$ denote the state of the embedded discrete-time Markov chain of the continuous-time process $X(t)$. That is,

$$Y(n) = \begin{cases} X(\tau_n) & \text{if } \tau_n < \infty \\ \Diamond & \text{if } \tau_n = \infty \end{cases}$$

where $\tau_n = \inf\{t \geq \tau_{n-1} \ : \ X(t) \neq X(\tau_{n-1})\}$ and $\tau_0 = 0$ and $\Diamond$ is an arbitrary element not in the state-space of the Markov process $X(t)$. Note that the sequence $\{\tau_n\}_{n\in\mathbb{N}}$ is the sequence of transition times.

Define the conditional random variable

$$N_{dNTP} = \sum_{n>0} 1_{\{Y(n)=4\}} \,\bigg|\, E_{pre} \cap \{Y(0) = 2\},$$

where $1_{\{Y(n)\}} = 1$ is $Y(n) = 4$ and 0 otherwise. Here, N_{dNTP} is the number of arrivals at state 4, the dNTP-bound state given that the DNAP-DNA complex starts in the post-translocation state and eventually escapes to the pre-translocation state of the current nucleotide addition cycle. For notational convenience, let $h = Pr(N_{dNTP} = 0)$. We have that,

$$\begin{aligned} h &= Pr(N_{dNTP} = 0) \\ &= Pr(Y(1) = 1 \mid Y(0) = 2, E_{pre}) \\ &= \frac{Pr(Y(1) = 1, Y(0) = 2, E_{pre})}{Pr(E_{pre} \, | Y(0) = 2)\, Pr(Y(0) = 2)} \\ &= \frac{Pr(Y(1) = 1 \mid Y(0) = 2)\, Pr(Y(0) = 2)}{Pr(E_{pre} \mid X(0) = 2)\, P(Y(0) = 2)} \\ &= \frac{r_2}{(r_2 + k_{on}[dNTP])\, p_{E_{pre}|2}}. \end{aligned}$$

From the state-space diagram in figure 3.2, we see that $Pr(N_{dNTP} = n) = (1-h)^n h$ since the only possible way for there to be n arrivals at state 4, given E_{pre} and $Y(0) = X(0) = 2$ is for the DNAP-DNA complex to transition from state 2 to state 4 n times, and then escape to state 2. We can then write $T^{(1)}$ as the following sum of a random number of random variables,

$$T^{(1)} = \sum_{j=1}^{N_{dNTP}} (T_{post,j} + T_{dNTP,j}) + T_{post,(N_{dNTP}+1)}$$

The first moment can then be calculated by the law of total expectation, By the law of total expectation, we then have

$$\begin{aligned}\left\langle T^{(1)}\right\rangle &= \sum_{n=0}^{\infty}\left\langle T^{(1)} \mid N_{dNTP}=n\right\rangle Pr\left(N_{dNTP}=n\right)\\ &= \sum_{n=0}^{\infty}\left(\frac{n+1}{r_2+k_{on}[dNTP]}+\frac{n}{k_{off}+k_{on}}\right)(1-h)^n h\\ &= \left(\frac{1-h}{h}+1\right)\frac{1}{r_2+k_{on}[dNTP]}+\frac{1-h}{h}\frac{1}{k_{off}+k_{pol}}.\end{aligned}$$

Derivation of $\langle T_B \rangle$

We use the same strategy to determine the first moment of T_B. Define the conditional random variable,

$$N_{pol} = \sum_{n>0} 1_{\{Y(n)=4\}} \,\Bigg|\, E_{pol} \cap \{Y(0)=2\}.$$

Here, N_{pol} is the number of arrivals in the dNTP-bound state (state 4), given sucessful covalent incorporation of the nucleotide and that the DNAP-DNA complex starts with the initial state $X(0)=Y(0)=2$. We write the random variable T_{pol} as a random sum of a random number of random variables,

$$T_{pol} = \sum_{j=1}^{N_{pol}}\left(T_{post,j}+T_{dNTP,j}\right).$$

For notational convenience, let $q = Pr\left(N_{pol} = 1\right)$. We calculate,

$$\begin{aligned}
q = Pr\left(N_{pol} = 1\right) &= Pr\left(Y\left(2\right) = 1+, Y\left(1\right) = 4 \mid Y\left(0\right) = 2, E_{pol}\right) \\
&= \frac{Pr\left(Y\left(2\right) = 1+, Y\left(1\right) = 4, E_{pol} \mid Y\left(0\right) = 2\right) Pr\left(Y\left(0\right) = 2\right)}{Pr\left(E_{pol} \mid Y\left(0\right) = 2\right) Pr\left(Y\left(0\right) = 2\right)} \\
&= \frac{Pr\left(Y\left(2\right) = 1+, Y\left(1\right) = 4 \mid Y\left(0\right) = 2\right)}{p_{E_{pol}|2}} \\
&= \frac{Pr\left(Y\left(2\right) = 1+ \mid Y\left(1\right) = 4\right) Pr\left(Y\left(1\right) = 4 \mid Y\left(0\right) = 2\right)}{p_{E_{pol}|2}} \\
&= \frac{\frac{k_{pol}}{k_{off}+k_{pol}} \frac{k_{on}[dNTP]}{r_2+k_{on}[dNTP]}}{p_{E_{pol}|2}}.
\end{aligned}$$

Hence $Pr\left(N_{pol} = n\right) = 0$ if $n = 0$ and $Pr\left(N_{pol} = n\right) = \left(1 - q\right)^{n-1} q$ if $n > 0$, since for $N_{pol} = n$ to occur, the chain must have traveled from state 2 to 4, n times, and then escaped to state 1+.

Thus we can calculate by total expectation,

$$\begin{aligned}
\left\langle T_{pol}^{(1)} \right\rangle &= \sum_{n=0}^{\infty} \left\langle T_{pol}^{(1)} \mid N_{pol} = n \right\rangle Pr\left(N_{pol} = n\right) \\
&= \sum_{n=0}^{\infty} \left(\frac{n}{r_2 + k_{on}[dNTP]} + \frac{n}{k_{off} + k_{pol}} \right) Pr\left(N_{pol} = n\right) \\
&= \sum_{n=0}^{\infty} \left(\frac{n}{r_2 + k_{on}[dNTP]} + \frac{n}{k_{off} + k_{pol}} \right) \left(1 - q\right)^{n-1} q \\
&= \left(\frac{1}{r_2 + k_{on}[dNTP]} + \frac{1}{k_{off} + k_{on}} \right) \frac{1}{q}.
\end{aligned}$$

Hence the first moment of T_B is given by

$$\begin{aligned}
\left\langle T_B \right\rangle &= \left\langle T_{pol} \right\rangle + \left\langle T^{(2)} \right\rangle \\
&= \left(\frac{1}{r_2 + k_{on}[dNTP]} + \frac{1}{k_{off} + k_{pol}} \right) \frac{1}{q} + \frac{\gamma}{\mu_1} + \frac{1 - \gamma}{\mu_2}.
\end{aligned}$$

Putting this all together, we can write $\langle T_{cycle} \rangle$ as

$$\langle T_{cycle} \rangle = \frac{\left(r_1 r_4 + r_2 r_3 + r_2 r_4\right)\left(k_{off} + k_{pol}\right) + \left(r_3 k_{pol} + r_4 k_{pol} + r_1 r_4\right) k_{on}[dNTP]}{r_1 r_4 k_{pol} k_{on}[dNTP]}. \tag{3.19}$$

The first moment of T_{cycle} can be written in the Michaelis-Menten form,

$$\langle T_{cycle} \rangle = \frac{\frac{\left(r_1 r_4 + r_2 r_3 + r_2 r_4\right)\left(k_{off} + k_{pol}\right)}{\left(r_3 k_{pol} + r_4 k_{pol} + r_1 r_4\right) k_{on}} + [dNTP]}{\frac{r_1 r_4 k_{pol}}{r_3 k_{pol} + r_4 k_{pol} + r_1 r_4}[dNTP]} \tag{3.20}$$

$$= \frac{K_m + [dNTP]}{k_{cat}[dNTP]}, \tag{3.21}$$

where

$$K_m = \frac{\left(r_1 r_4 + r_2 r_3 + r_2 r_4\right)\left(k_{off} + k_{pol}\right)}{\left(r_3 k_{pol} + r_4 k_{pol} + r_1 r_4\right) k_{on}}, \tag{3.22}$$

and

$$k_{cat} = \frac{r_1 r_4 k_{pol}}{r_3 k_{pol} + r_4 k_{pol} + r_1 r_4}. \tag{3.23}$$

Here, K_m is the [dNTP] for which the reaction rate is half-maximum and k_{cat} is the maximum reaction rate.

3.3 Inferring the Transition Rates k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ and T_B Data and the Role of $\langle T_{cycle} \rangle$

In this section, we describe the strategies for inferring the transition rate parameters k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ and T_B dwell time data, and $\langle T_{cycle} \rangle$.

3.3.1 Inferring k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ Data

Recall that after conditioning on the escape to the pre-translocation state from the post-translocation state (figure 3.9)., the state-space of the Markov process describing the $T^{(1)}$ escape problem is of the same form as the escape problem for $T^{(1)}$ in chapter 2.

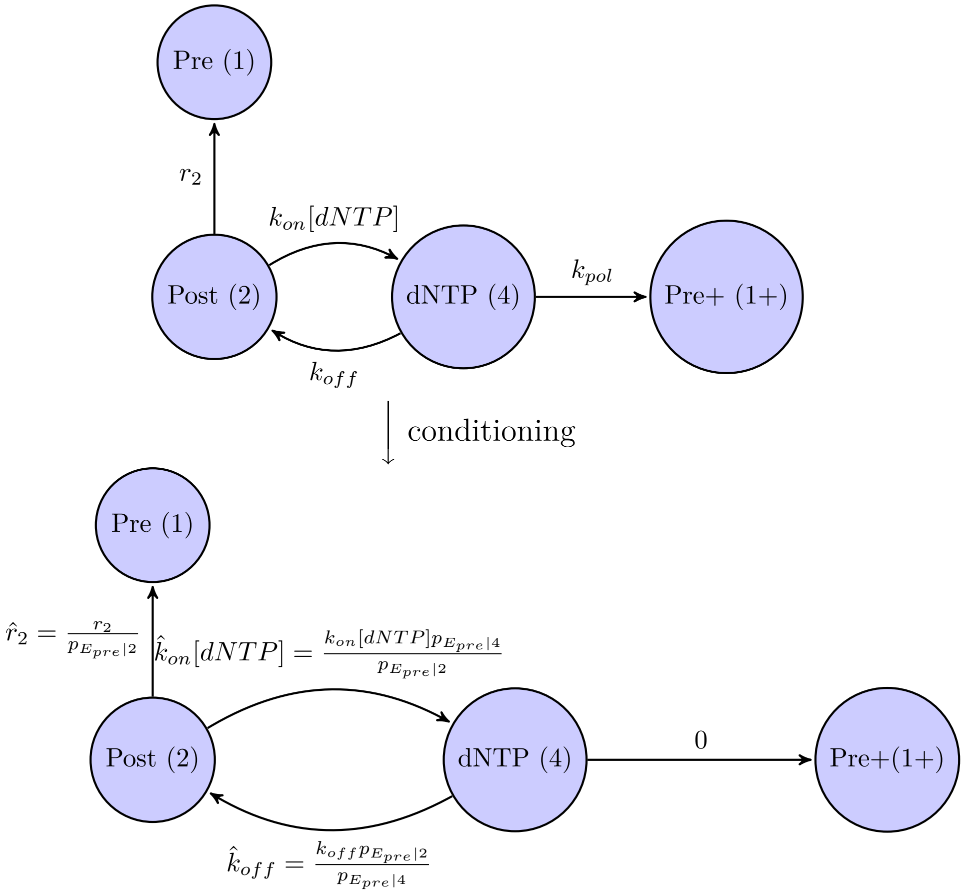
Pre (1)
r_2
$k_{on}[dNTP]$
Post (2)
dNTP (4)
k_{pol}
Pre+ (1+)
k_{off}
conditioning
Pre (1)
$\hat{r}_2 = \frac{r_2}{p_{E_{pre}|2}}$
$\hat{k}_{on}[dNTP] = \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}}$
Post (2)
dNTP (4)
0
Pre+(1+)
$\hat{k}_{off} = \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}}$

Figure 3.9: (top) State-space diagram of the lower-amplitude states describing the escape problem underlying $T^{(1)}$; (bottom) State-space diagram of the lower amplitude states after conditioning on the escape to the pre-translocation state. Recall that $p_{E_{pre}|2}$ and $p_{E_{pre}|4}$ are the probabilities of escape to the pre-translocation starting from states 2 and 4, respectively.

Recall that the PDF of $T^{(1)}$ is a mixture of two exponential modes,

$$f_{T^{(1)}} = \alpha\lambda_1 e^{-\lambda_1 t} + (1-\alpha)\,\lambda_2 e^{-\lambda_2 t},$$

with

$$\lambda_{1,2} = \frac{B \pm \sqrt{B^2 - 4\frac{r_2 k_{off}}{p_{E_{pre}|4}}}}{2},$$

where

$$B = \frac{r_2}{p_{E_{pre}|2}} + \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} + \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}},$$

for notational compactness. The mixture parameter α is given by $\alpha = (\lambda_2 - r_2)\,/\,(\lambda_2 - \lambda_1)$. The characteristic equation whose roots are $\lambda_{1,2}$ is given by

$$\lambda^2 - \left(\frac{r_2}{p_{E_{pre}|2}} + \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} + \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}}\right)\lambda + \frac{r_2 k_{off}}{p_{E_{pre}|4}} = 0.$$

Hence we have the following three nonlinear equations

$$\frac{r_2}{p_{E_{pre}|2}} + \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}} + \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}} = \lambda_1 + \lambda_2, \tag{3.24}$$

$$\frac{r_2 k_{off}}{p_{E_{pre}|4}} = \lambda_1\lambda_2, \tag{3.25}$$

$$\alpha = \frac{\lambda_2 - r_2}{\lambda_2 - \lambda_1}. \tag{3.26}$$

We solve the above system for k_{on}, k_{off}, and k_{pol} with the aide of a computer algebra solver if r_2 is known. Like in the $k_{pol} = 0$ case in chapter 2, r_2 is inferred separately when $[dNTP] = 0$. When $[dNTP] = 0$, the PDF of $T^{(1)}$ is a single exponential mode with mean $1/r_2$. Hence we can infer r_2 from the $T^{(1)}$ data when $[dNTP] = 0$ by computing,

$$r_2 = \frac{1}{\langle T^{(1)}\rangle}.$$

Hence for the purposes of inferring k_{on}, k_{off}, and k_{pol} (in this case $[dNTP] > 0$), we can consider r_2 to be known.

With the aide of a computer algebra solver, we solve 3.24-3.26 for k_{on}, k_{off}, and k_{pol}. Since the expressions are extraordinarily long and cumbersome, we write them down in the Appendix (equations A.1, A.2, and A.3). This solution gives us the mapping $(\alpha, \lambda_1, \lambda_2) \to (k_{on}, k_{off}, k_{pol})$. Hence we preserve the EM-framework thus providing a more efficient means of computing the estimates of k_{on}, k_{off}, and k_{pol}. Like in Chapter 2, we can estimate the mixture parameters α, λ_1, and λ_2 by using the EM algorithm since the distribution of $T^{(1)}$ is a proper exponential mixture [13]. Let $T_1^{(1)}, \ldots, T_n^{(1)}$ be a random sample of $f_{T^{(1)}}$. Let $\theta = (\alpha, \lambda_1, \lambda_2)$ be the mixture parameters and $\theta^{(k)} = \left(\alpha^{(k)}, \lambda_1^{(k)}, \lambda_2^{(k)}\right)$ the k-th term in the EM sequence. The analytical expression for $\theta^{(k)}$ is the same as in Chapter 2, which we rewrite here for convenience,

$$\alpha^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(k)}, \tag{3.27}$$

$$\lambda_1^{(k+1)} = \frac{\sum_{i=1}^{n} \xi_i^{(k)}}{\sum_{i=1}^{n} \xi_i^{(k)} t_i}, \tag{3.28}$$

$$\lambda_2^{(k+1)} = \frac{n - \sum_{i=1}^{n} \xi_i^{(k)}}{\sum_{i=1}^{n} \left(1 - \xi_i^{(k)}\right) t_i}, \tag{3.29}$$

where

$$\xi_i^{(k)} = \frac{\alpha^{(k)} \lambda_1^{(k)} e^{-\lambda_1^{(k)} t_i}}{\alpha^{(k)} \lambda_1^{(k)} e^{-\lambda_1^{(k)} t_i} + \left(1 - \alpha^{(k)}\right) \lambda_2^{(k)} e^{-\lambda_2^{(k)} t_i}}.$$

The analytical expressions in equations 3.27-3.29 are used to calculate $\theta^{(k+1)}$ from $\theta^{(k)}$ and the dwell time data. It can be shown that equations 3.27-3.29 will converge to the MLE of $(\alpha, \lambda_1, \lambda_2)$ for any initial guess $\theta^{(0)}$ [72] (see Chapter 2).

3.3.2 Inferring k_{on}, k_{off}, and k_{pol} from T_B Data

We also have the option of inferring the transition rates k_{on}, k_{off}, and k_{pol} from T_B data. However, unlike the dwell time $T^{(1)}$, the distribution of T_B is an improper mixture of four exponential modes (equation 3.18). Hence we lose the hierarchical structure of a proper mixture distribution and cannot use the EM algorithm. We instead have to use more straight-forward optimization techniques to find the MLE estimates of $(k_{on}, k_{off}, k_{off})$. Let $T_{B_1}, \ldots, T_{B_n}$ be a random sample of T_B. We can form the log-likelihood given by

$$L(\theta \mid t) = \sum_{i=1}^{n} \log\left(f_{T_B}(t_i \mid \theta)\right).$$

We maximize the log-likelihood numerically using Matlab's "fminsearch" function which employs the Nelder-Mead algorithm [57]. The value of $(k_{on}, k_{off}, k_{pol})$ which yields the maximum gives us the MLE estimates of the transition rates.

3.3.3 The role of $\langle T_{cycle} \rangle$ on the inference of k_{on}, k_{off}, and k_{pol}.

The average time of the nucleotide addition cycle plays a role in constraining the transition rates. Recall from equation 3.20 that $\langle T_{cycle} \rangle$ can be put in the Michaelis-Menten form. From the Michaelis-Menten parameter k_{cat} (equation 3.23), we can obtain an expression for k_{pol},

$$k_{pol} = \frac{r_1 r_4}{\frac{r_1 r_4}{k_{cat}} - (r_3 + r_4)}. \tag{3.30}$$

Provided that we know the transition rates r_1, r_3, and r_4, we can thus obtain the incorporation rate k_{pol} from saturating dNTP concentrations.

The Michaelis-Menten parameter K_m (equation 3.22) can be used to put a constraint on the binding and disassociation rates:

$$\begin{aligned} K_m \left(r_3 k_{pol} + r_4 k_{pol} + r_1 r_4 \right) k_{on} - \left(r_1 r_4 + r_2 r_3 + r_2 r_4 \right) k_{off} \\ - \left(r_1 r_4 + r_2 r_3 + r_2 r_4 \right) k_{pol} = 0. \end{aligned} \tag{3.31}$$

Provided that we can obtain k_{pol} from equation 3.30 or through another means, can accurately determine K_m, and we know the transition rates r_1, r_2, r_3, and r_4, we can constrain k_{on} and k_{off} to the line given in equation 3.31.

3.4 Numerical Simulations: Case of No Detection Uncertainty

In this section, we conduct some numerical simulations to determine the validity of using the MLE method to infer k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ and T_B data without noise in the $T^{(1)}$ and T_B observations. We also look at some numerical experiments to examine the reliability of inferring k_{pol} from the maximum reaction velocity.

3.4.1 Inferring k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ and T_B Data by the MLE Method

The following numerical simulation was done as follows. We generated 10,000 samples of the dwell times $T^{(1)}$ and T_B and we used these samples to obtain MLE estimates for k_{on}, k_{off}, and k_{pol}. This was then repeated 10,000 times to obtain a distribution for the MLE estimates.

The MLE estimates were centered with respect to their true values and then normalized by their true values. We use the notation

$$\mathrm{err}\,(k) = \frac{k^{MLE} - k^{true}}{k^{true}},$$

to denote the centered and normalized error of the MLE estimates. The standard deviation of $\mathrm{err}\,(k)$ is also recorded, and we denote this by $\mathrm{std}\,(\mathrm{err}\,(k))$. In the simulations, we set the true values of the transition rates to be:

- $r_1 = r_2 = 100$,
- $r_3 = r_4 = 1$,
- $k_{on} = 200$,
- $k_{off} = 100$,
- $k_{pol} = 50$.

The transition rate r_2 can be recovered from $T^{(1)}$ data when $[dNTP] = 0$. When $[dNTP] = 0$, $T^{(1)}$ is distributed by a single exponential mode and r_2 can be recovered by using an MLE similar to the method proposed in Chapter 2 or the autocorrelation method in [49]. The transition rates r_1, r_3, and r_4 can be inferred from upper-amplitude $T^{(2)}$ data in the exact same manner as in Chapter 2 or in [51]. The dNTP concentration can also be controlled accurately in the experiments. Thus for simplicity, we assume that r_1, r_2, r_3, r_4, and $[dNTP]$ are known and focus only on the inference of k_{on}, k_{off}, and k_{pol}.

For notational convenience, let $\theta = (k_{on}, k_{off}, k_{pol})$. We can infer θ from the $T^{(1)}$ and T_B dwell times separately. Using the $T^{(1)}$ data to infer θ, we see that higher dNTP concentrations lead to lower relative errors for these particular set of transition rates. The relative error decreases for k_{on}, k_{off}, and k_{pol} as $[dNTP]$ increases throughout the range of $[dNTP]$ tested; the minimum relative error for k_{on}, k_{off}, and k_{pol} is 2.5%, 2.8%, and 1.3%, respectively at $[dNTP] = 16$ (figures 3.10, 3.12, and 3.14).

Unlike $T^{(1)}$ the information content of T_B has no obvious pattern. The relative error for k_{on} increases as $[dNTP]$ increases when inferring from T_B data throughout the range of $[dNTP]$ tested, and the minimum relative error is about 10.7% at $[dNTP] = 0.25$ (figure 3.11). For inferring k_{off} from T_B data, the relative error is minimum at $[dNTP] = 1$ throughout the range of $[dNTP]$ tested. The relative error for k_{off} increases as $[dNTP]$ gets further from 1; at $[dNTP] = 1$, the relative error is about 5.2% (figure 3.13). Finally, when inferring k_{pol} from T_B data, the relative error decreases as $[dNTP]$ increases, obtaining a minimum of 1.6% at $[dNTP] = 16$, behaving similar to the relative error for k_{on}, k_{off}, and k_{pol} when inferring from $T^{(1)}$ data (figure 3.15). Since we set $k_{on} = 200$ and $k_{off} = 100$, $[dNTP] = 16$ corresponds to the high dNTP concentration of $[dNTP]/K_d = 32$ where $K_d = k_{off}/k_{on} = 0.5$.

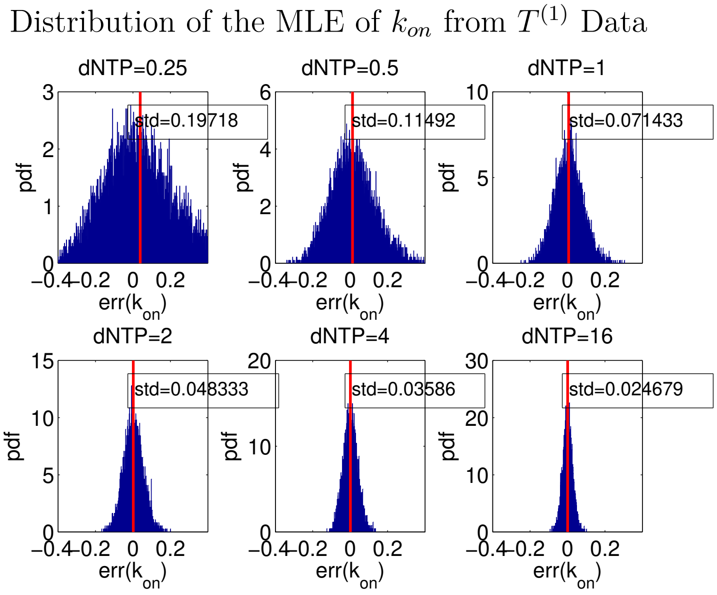

Figure 3.10: MLE results for k_{on} from $T^{(1)}$ data with no noise.

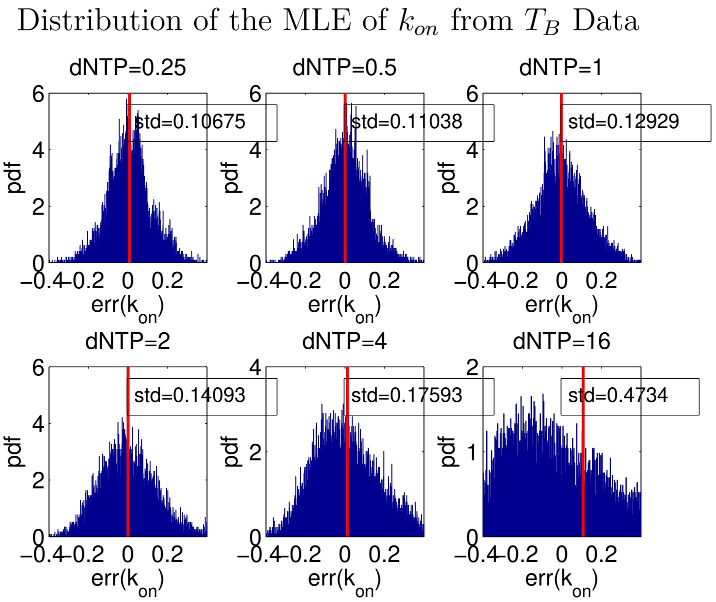

Figure 3.11: MLE results for k_{on} from T_B data with no noise.

3.4.2 Inferring k_{pol} from the Maximum Reaction Velocity

To test the validity of inferring k_{pol} from the maximum reaction velocity, we generated 10,000 samples of T_{cycle} of which to compute the mean cycle time $\langle T_{cycle} \rangle$. This was then repeated 1,000 times to obtain a set of estimates for k_{pol}. The results of $\text{std}\left(\text{err}\left(k_{pol}\right)\right)$, $\text{mean}\left(\text{err}\left(k_{pol}\right)\right)$, and $\text{rms}\left(\text{err}\left(k_{pol}\right)\right)$ are shown in figures 3.16-3.18. From figure 3.16, we see that the relative error actually increases from about 2% at $[dNTP] = 1$ to about 7% at $[dNTP] = 1024$. This is very unintuitive since we expect the error to decrease as the dNTP concentration saturates. However, as we can see from figure 3.17 the bias decreases from about 50% at $[dNTP] = 1$ to about 0.1% at $[dNTP] = 1024$. Examining the RMS error incorporates the bias and shows that the RMS error settles to about 7% at saturating dNTP concentrations.

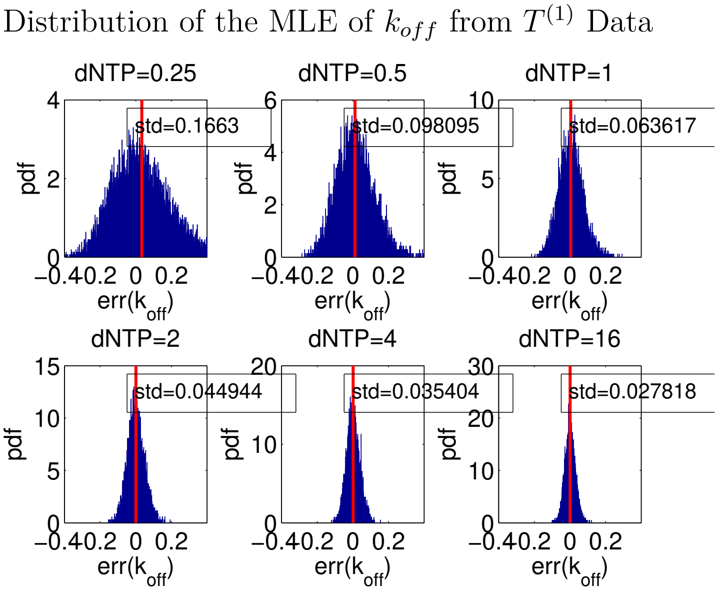

Figure 3.12: MLE results for k_{off} from $T^{(1)}$ data with no noise.

3.4.3 Inferring k_{on} and k_{off} from $T^{(1)}$ and T_B Data with Constraints

We also investigate the validity of inferring k_{on} and k_{off} from both the $T^{(1)}$ and T_B data while enforcing that k_{on} and k_{off} are constrained to the line given in equation 3.31. In this context, it makes sense to infer from both the $T^{(1)}$ and T_B data since (i) we cannot enforce the constraint in the EM method when using $T^{(1)}$ data, so we are forced to maximize the likelihood function of the $T^{(1)}$ data directly; and (ii) combining the likelihood functions for the $T^{(1)}$ and T_B data comes at no extra cost, as we can maximize the sum of the likelihood functions for $T^{(1)}$ and T_B. The results of $\text{std}\,(\text{err}\,(k))$, $\text{mean}\,(\text{err}\,(k))$, and $\text{rms}\,(\text{err}\,(k))$ are shown in figures 3.21-3.19. The increase in bias and relative error for k_{on} and k_{off} around $[dNTP] = 1$ is intriguing. This is likely due to the poor information content from T_B around $[dNTP] = 1$ creating large bias (figures 3.11 and 3.13). In this region, the likelihood function for $T^{(1)}$ was not able to compensate enough to produce

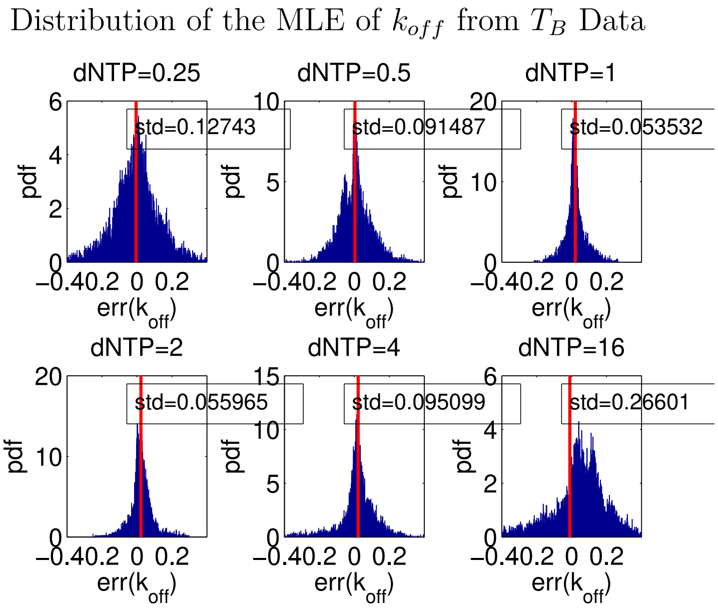

Figure 3.13: MLE results for k_{off} from T_B data with no noise.

comparable results to the surrounding regions of $[dNTP]$. This can be mitigated by only using the likelihood function for $T^{(1)}$, however careful thought must be carried out in regards to experimental waiting time to collect the desired number of samples of $T^{(1)}$ as we will see in the next section. Regardless, this numerical study shows that constraining k_{on} and k_{off} to the line obtained from the Michaelis-Menten parameter K_m produces very satisfactory inference results and improves upon the general unconstrained inference of k_{on}, k_{off}, and k_{pol} simultaneously (figures 3.10-3.13).

3.5 Dependence of Inference Uncertainty Model Parameters

Similar to the case of non synthesizing DNA-DNAP complexes, The dNTP concentration plays a role in the inference uncertainty of k_{on}, k_{off}, and k_{pol} (Chap-

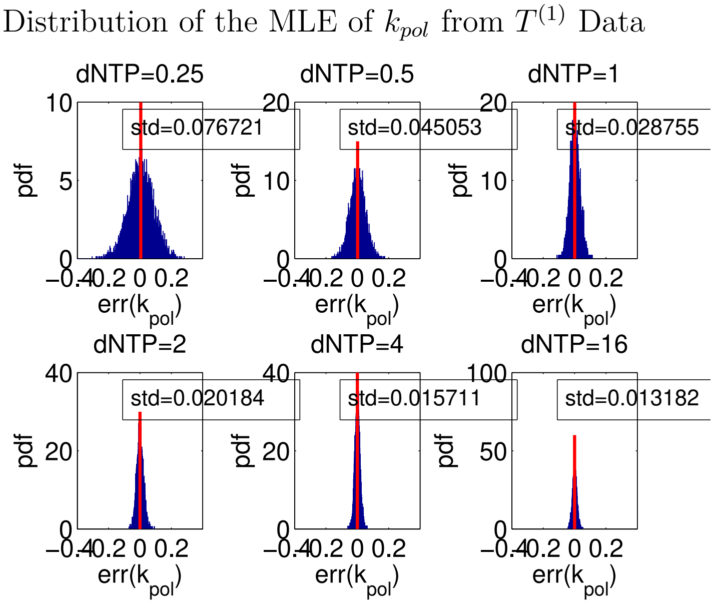

Figure 3.14: MLE results for k_{pol} from $T^{(1)}$ data with no noise.

ter 2). Tuning the dNTP concentration will allow us to control the total relative error of the inferred transition rates, and so the behavior of the total relative error as a function of the transition rates k_{on}, k_{off}, and k_{pol} and $[dNTP]$ has to be known a priori. We focus on inferring the transition rates from $T^{(1)}$ data only as numerical simulations from the previous section show that the information content of $T^{(1)}$ for the transition rates is superior to that of T_B for inferring the rates. We will do this by first deriving an estimate for the total relative error function of the inferred k_{on}, k_{off}, and k_{pol}. We will also show by conditioning and scaling laws that the total relative error is a function of scaled versions of k_{off} and $[dNTP]$ only.

Distribution of the MLE of k_{pol} from T_B Data

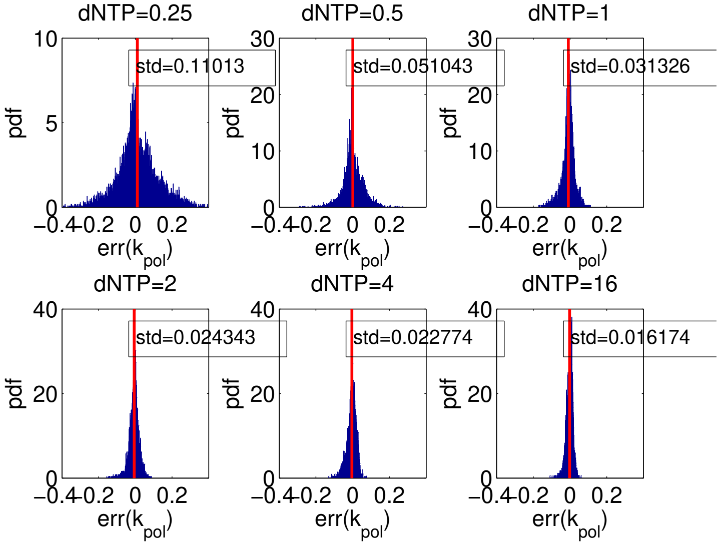

Figure 3.15: MLE results for k_{pol} from T_B data with no noise.

3.5.1 Derivation of Total Relative Error Function of k_{on}, k_{off}, and k_{pol}

The derivation of the total relative error is a straightforward generalization of the derivation presented in Chapter 2. We repeat it here for convenience. Let $T_1^{(1)}, \dots T_n^{(1)}$ be a random sample of $f_{T^{(1)}}$ where $f_{T^{(1)}}$ is the PDF of the lower-amplitude dwell time segment in T_A (equation 3.9).

Let

$$L(\theta \mid t) = \sum_{i=1}^{n} \log\left(\alpha e^{-\lambda_1 t_i} + (1-\alpha)\, e^{-\lambda_2 t_i}\right) \tag{3.32}$$

be the log-likelihood function of the $T^{(1)}$ data, and let $\theta^{\text{MLE}} = \left(\alpha^{\text{MLE}}, \lambda_1^{\text{MLE}}, \lambda_2^{\text{MLE}}\right)$ be the MLE estimates of θ. The observed Fisher information matrix, H, defined

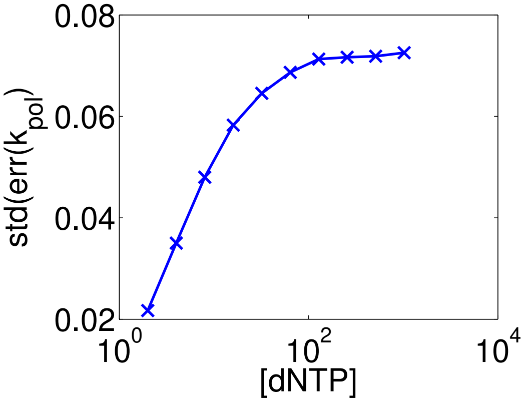

Figure 3.16: Plot of $\mathrm{std}\left(\mathrm{err}\left(k_{pol}\right)\right)$ vs $[dNTP]$ from the maximum reaction velocity

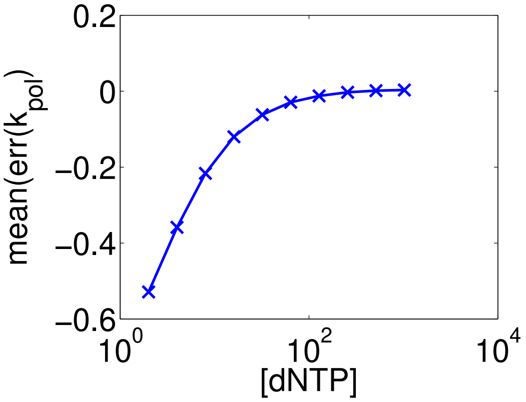

Figure 3.17: Plot of $\mathrm{mean}\left(\mathrm{err}\left(k_{pol}\right)\right)$ vs $[dNTP]$ from the maximum reaction velocity

by

$$H\left(\theta^{\mathrm{MLE}} \mid t\right) = -\begin{pmatrix} L_{\alpha,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\alpha,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\alpha,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \\ L_{\lambda_1,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_1,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_1,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \\ L_{\lambda_2,\alpha}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_2,\lambda_1}\left(\theta^{\mathrm{MLE}} \mid t\right) & L_{\lambda_2,\lambda_2}\left(\theta^{\mathrm{MLE}} \mid t\right) \end{pmatrix}, \tag{3.33}$$

where we denote $L_{x,y}$ to mean $L_{x,y} = \partial^2 L / \left(\partial x \partial y\right)$. Note that equation 3.33 is the negative Hessian of L evaluated at the MLE estimates. It has been demonstrated that the inverse of H gives an approximation to the asymptotic covariance matrix

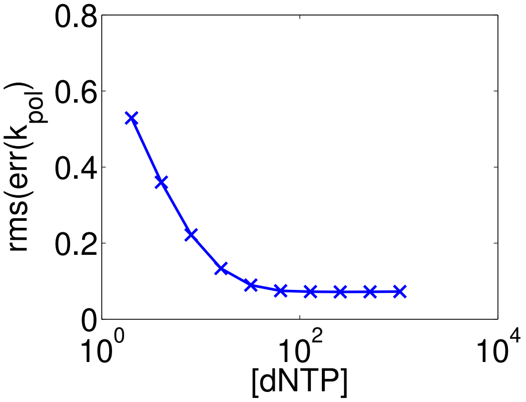

Figure 3.18: Plot of $\mathrm{rms}\left(\mathrm{err}\left(k_{pol}\right)\right)$ vs $[dNTP]$ from the maximum reaction velocity

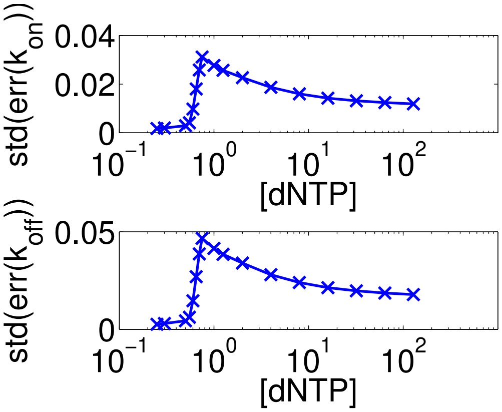

Figure 3.19: Plot of $\mathrm{std}\left(\mathrm{err}\left(k\right)\right)$ vs $[dNTP]$ the constrained maximization of the sum of the likelihood functions of $T^{(1)}$ and T_B.

of the MLE estimates of θ as the number of samples of $T^{(1)}$, $n \to \infty$ [24]. Hence, $\mathrm{Cov}\left(\theta\right) \approx H^{-1}$.

We can propagate the inference uncertainty of the mixture parameters θ to k_{on}, k_{off}, and k_{pol} by a first-order Taylor expansion. Recall that we have the mapping

$\theta \mapsto \left(k_{on}\left(\theta\right), k_{off}\left(\theta\right), k_{pol}\left(\theta\right)\right)^T$ according to equations A.1, A.2, and A.3 in the Appendix. Let $K\left(\theta\right) = \left(k_{on}\left(\theta\right), k_{off}\left(\theta\right), k_{pol}\left(\theta\right)\right)^T$ be this mapping.

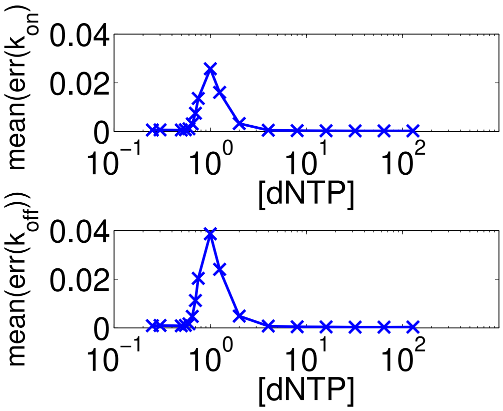

Figure 3.20: Plot of mean $(\mathrm{err}(k))$ vs $[dNTP]$ the constrained maximization of the sum of the likelihood functions of $T^{(1)}$ and T_B.

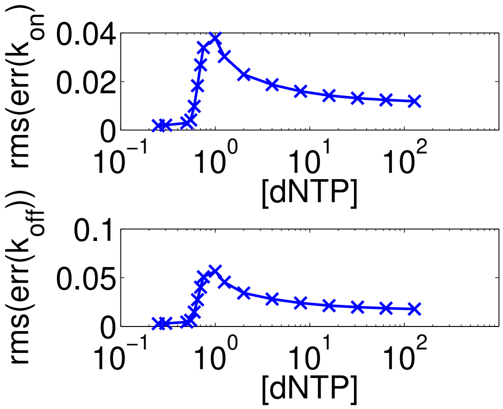

Figure 3.21: Plot of rms $(\mathrm{err}(k))$ vs $[dNTP]$ the constrained maximization of the sum of the likelihood functions of $T^{(1)}$ and T_B.

Consider the first-order Taylor expansion,

$$K(\theta) = K\left(\theta^{\mathrm{MLE}}\right) + J\left(\theta^{\mathrm{MLE}}\right)\left(\theta - \theta^{\mathrm{MLE}}\right) + o\left(\left\|\theta - \theta^{\mathrm{MLE}}\right\|\right),$$

where $J\left(\theta^{\text{MLE}}\right)$ is the Jacobian of K evaluated at θ^{MLE}. Now

$$\begin{aligned}
Cov\left(K\left(\theta\right)\right) &= Cov\left[K\left(\theta^{\text{MLE}}\right) + J\left(\theta^{\text{MLE}}\right)\left(\theta - \theta^{\text{MLE}}\right) + o\left(\left\|\theta - \theta^{\text{MLE}}\right\|\right)\right] \\
&= Cov\left(J\left(\theta^{\text{MLE}}\right)\theta\right) \\
&= J\left(\theta^{\text{MLE}}\right) Cov\left(\theta\right) J\left(\theta^{\text{MLE}}\right)^T,
\end{aligned}$$

where $Cov\left(\theta\right)$ is the covariance matrix of θ. Recall that $Cov\left(\theta\right)$ is approximated by H^{-1} where H is the observed information matrix (equation 3.33). The second equality follows since $K\left(\theta^{\text{MLE}}\right)$, θ^{MLE}, and $o\left(\left\|\theta - \theta^{\text{MLE}}\right\|\right)$ are constant vectors.

The result is that the diagonal entries of the covariance matrix $Cov\left(K\left(\theta\right)\right)$ are the asymptotic estimates of the variance of the MLE estimates of k_{on}, k_{off}, and k_{pol}. Using this, we can estimate the relative error of the MLE estimates of k_{on}, k_{off}, and k_{pol} a priori without the computational effort of full Monte Carlo simulations.

We will use the total relative error as a metric to study the inference uncertainty of k_{on}, k_{off}, and k_{pol}. The total relative error is the sum of the relative errors of k_{on}, k_{off}, and k_{pol}. In parameter regions in which the total relative error is small, the total relative error can be approximated by $\text{std}\left(\text{err}\left(k_{on}\right)\right) + \text{std}\left(\text{err}\left(k_{off}\right)\right) + \text{std}\left(\text{err}\left(k_{pol}\right)\right)$, since inference bias is small in these regions. We can approximate the total relative error without the computational efforts of a full Monte Carlo simulation for any k_{on}, k_{off}, and k_{pol} by using the estimates of the total relative error of their MLE estimates derived above in the following way,

$$\begin{aligned}
&\text{std}\left(\text{err}\left(k_{on}\right)\right) + \text{std}\left(\text{err}\left(k_{off}\right)\right) + \text{std}\left(\text{err}\left(k_{pol}\right)\right) \\
&\qquad \approx \left\|\sqrt{\text{diag}\left(Cov\left(K\right)\right)} \odot \left(\frac{1}{k_{on}}, \frac{1}{k_{off}}, \frac{1}{k_{pol}}\right)^T\right\|_1 := \text{Err}\left(k_{on}, k_{off}, k_{pol}, n\right) \quad (3.34)
\end{aligned}$$

where $\operatorname{diag}(A)$ is the vector containing the diagonal entries of the matrix A, $A \odot B$ is the element-wise multiplication of the matrices A and B, and $\|\cdot\|_1$ denotes the Euclidean 1-norm. The square-root operator is taken to be applied element-wise on the entries of $\operatorname{diag}(Cov(K))$. We denote the total relative error function as $\operatorname{Err}(k_{on}, k_{off}, k_{pol}, n)$, where n is the number of $T^{(1)}$ samples observed.

3.5.2 Behavior of the Total Relative Error Function

We apply the scaling law shown in Proposition 2 of chapter 2 to show that the total relative error function in equation 3.34 above is a function of scaled versions of k_{off}, k_{pol}, and $[dNTP]$ only.

Recall that after conditioning, we have effective translocation, dNTP binding, and dNTP disassociation rates

$$\begin{aligned}\hat{r}_2 &= \frac{r_2}{p_{E_{pre}|2}},\\ \hat{k}_{on} &= \frac{k_{on} p_{E_{pre}|4}}{p_{E_{pre}|2}},\\ \hat{k}_{off} &= \frac{k_{off} p_{E_{pre}|2}}{p_{E_{pre}|4}},\end{aligned}$$

(figure 3.9). Note that $p_{E_{pre}|2}$ and $p_{E_{pre}|4}$ are functions of $(r_2, k_{on}, k_{off}, k_{pol}, [dNTP])$ (equations 3.6-3.7. Using the scaling law in Proposition 2 of chapter 2, scaling the $T^{(1)}$ data by β gives the scaling map

$$\left(\hat{r}_2, \hat{k}_{on}, \hat{k}_{off}\right) \mapsto \left(\frac{\hat{r}_2}{\beta}, \frac{\hat{k}_{on}}{\beta}, \frac{\hat{k}_{off}}{\beta}\right).$$

Physically, this is analogous to scaling the unit of time such that $\hat{r}_2 \mapsto \hat{r}_2/\beta$. The scaling factor β is arbitrary, and the $T^{(1)}$ dwell time remains unchanged modulo a scaled factor. Since β is arbitrary, we can choose it to be any value. Choose

$\beta = \hat{r}_2$, so that

$$\left(\hat{r}_2, \hat{k}_{on}, \hat{k}_{off}\right) \mapsto \left(1, \frac{\hat{k}_{on}}{\hat{r}_2}, \frac{\hat{k}_{off}}{\hat{r}_2}\right).$$

Note that in the inference, we do not infer $\hat{r}_2$ and thus do not know its true value (we only know r_2). Nevertheless, in the theoretical treatment in showing that the relative error is a function of scaled k_{off} and scaled $[dNTP]$, we can choose β to be any value, namely we choose $\beta = \hat{r}_2$.

Denote ℓ as

$$\ell := \hat{k}_{off}/\hat{r}_2 = \frac{p^2_{E_{pre}|2}k}{p_{E_{pre}|4}}, \tag{3.35}$$

where $k := k_{off}/r_2$. Here k is the same scaled r_2 as in the non-synthesizing case in chapter 2.

We can also scale the units of concentration by $\hat{k}_{on}/\hat{r}_2$ so that we have the scaled $[dNTP]$,

$$U := \frac{\hat{k}_{on}}{\hat{r}_2}[dNTP] = p_{E_{pre}|4}S, \tag{3.36}$$

where $S := k_{on}/r_2[dNTP]$. Here S is the same scaled $[dNTP]$ as in the non-synthesizing case in chapter 2. The quantities ℓ and U can be viewed as the new effective binding and disassociation rates after scaling.

Let $k_p := k_{pol}/r_2$ be the scaled k_{pol}. From equations 3.6-3.7, it is easy to see that the absorption probabilities $p_{E_{pre}|2}$ and $p_{E_{pre}|4}$ can be written as

$$p_{E_{pre}|2} = \frac{k + k_p}{k + k_p + k_pS},$$
$$p_{E_{pre}|4} = \frac{k}{k + k_p + k_pS}.$$

That is, the probability to absorption to the pre-translocation state are functions of $S, k,$ and k_p only. Hence both U and ℓ are functions of $S, k,$ and k_p. We thus write them as $U\left(S, k, k_p\right)$ and $\ell\left(S, k, k_p\right)$ to emphasize this dependence.

Applying these scalings to the bottom state-space diagram in figure 3.9, we obtain the following state-space diagram in the bottom figure 3.22; here, figure 3.9 (top) is reproduced in figure 3.22 (top) for convenience. We see that the total relative error of k_{on}, k_{off}, and k_{pol} is thus a function of S, k, and k_p only.

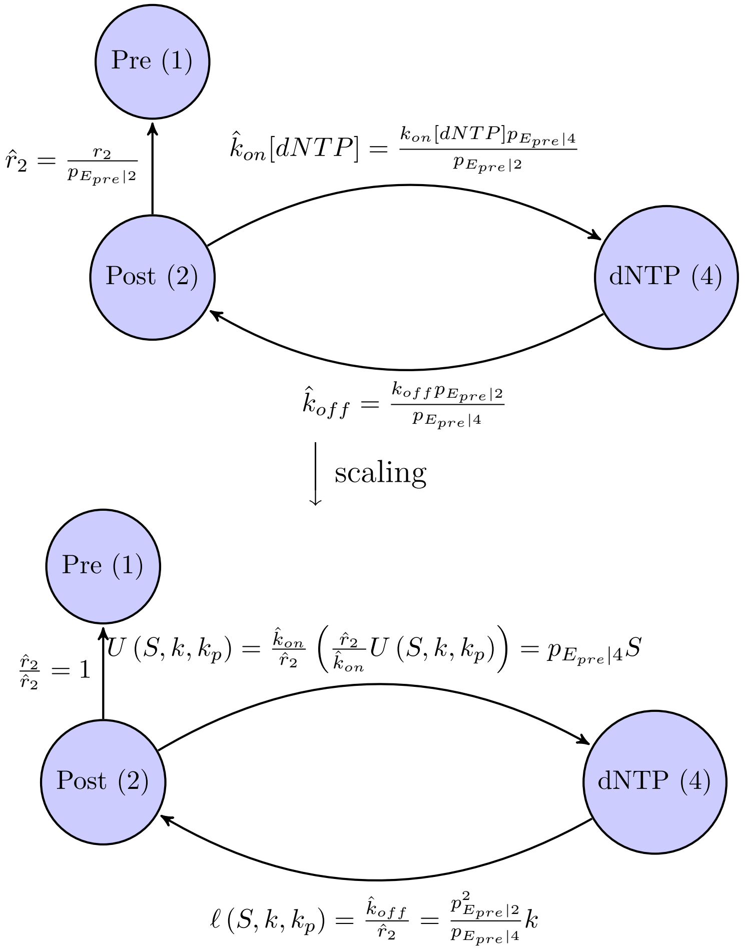
Pre (1)
$\hat{r}_2 = \frac{r_2}{p_{E_{pre}|2}}$
$\hat{k}_{on}[dNTP] = \frac{k_{on}[dNTP]p_{E_{pre}|4}}{p_{E_{pre}|2}}$
Post (2)
dNTP (4)
$\hat{k}_{off} = \frac{k_{off}p_{E_{pre}|2}}{p_{E_{pre}|4}}$
scaling
Pre (1)
$\frac{\hat{r}_2}{\hat{r}_2} = 1$
$U(S, k, k_p) = \frac{\hat{k}_{on}}{\hat{r}_2}\left(\frac{\hat{r}_2}{\hat{k}_{on}}U(S, k, k_p)\right) = p_{E_{pre}|4}S$
Post (2)
dNTP (4)
$\ell(S, k, k_p) = \frac{\hat{k}_{off}}{\hat{r}_2} = \frac{p^2_{E_{pre}|2}}{p_{E_{pre}|4}}k$

Figure 3.22: (top) State-space diagram of the lower-amplitude states after conditioning on E_{pre}; (bottom) State-space diagram of the lower amplitude states conditioned on E_{pre} after scaling on $\hat{r}_2$, $\hat{k}_{on}$, and $\hat{k}_{off}$. Here, we see that the total relative error of k_{on}, k_{off}, and k_{pol} is a function of S, k, and k_p only.

For clarity, figure 3.23 shows the state-space diagram governing the escape problem for $T^{(1)}$ before conditioning (top left), after conditioning (bottom left), and after scaling (bottom right).

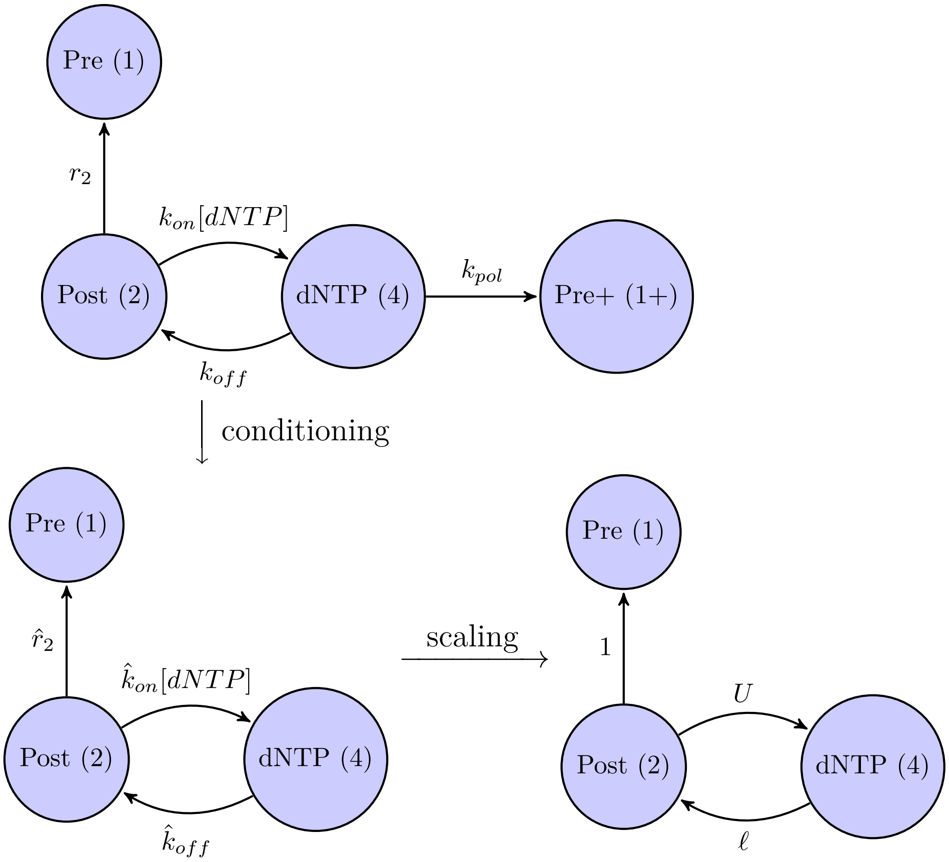

Figure 3.23: (top left) State-space diagram of the lower-amplitude states governing the escape problem for $T^{(1)}$; (bottom left) state space diagram after conditioning on the escape to the pre-translocation state; (bottom right) State-space diagram of the conditioned lower amplitude states after scaling by $\hat{r}_2$.

Hence let $\text{err}_{\text{pol}}\left(S,k,k_p,n\right) = \text{err}_{\text{pol}}\left(U\left(S,k,k_p\right),\ell\left(S,k,k_p\right),n\right)$ be the total relative error of k_{on}, k_{off}, and k_{pol} using n samples of $T^{(1)}$. We can write

$$\text{Err}\left(k_{on},k_{off},k_{pol},n\right) = \text{err}_{\text{pol}}\left(S,k,k_p,n\right).$$

The function $\text{err}_{\text{pol}}\left(S,k,k_p\right)$ is very difficult to compute analytically, so we turn to numerical approximation. To build the total relative error function numerically, we use a similar procedure as in approximating $\text{err}\left(S,k\right)$ shown in Chapter 2.

We set $r_2 = k_{on} = 1$. For each fixed k_p, we do the following. We vary S and

k over a range of values. Note that in this case, $S = [dNTP]$, $k = k_{off}$, and $k_p = k_{pol}$. Let $\mathcal{S}$ and $\mathcal{K}$ be the set of discrete points for S and k respectively. Enumerate the elements of $\mathcal{K} = \{k_1, \ldots, k_m\}$, where m is the number of k points used. At each $(S, k) \in \mathcal{S} \times \mathcal{K}$ point we sample $f_{T^{(1)}}$ $n_0 = 10,000$ times and estimate k_{on}, k_{off}, and k_{pol} using the EM method above. The total relative error is then estimated by using equation 3.34. This is repeated 20 times for each $(S, k) \in \mathcal{S} \times \mathcal{K}$, giving us a cloud of total relative error data for each (S, k) point.

Let $\mathcal{E}_C(S, k)$ be the 20-point data cloud at the point (S, k). We then estimate the total relative error by fitting a quadratic polynomial in the k-direction using 11 points in the least squares sense.

Let $k_i \in \mathcal{K}$. Define the following subset of $\mathcal{K}$,

$$\mathcal{K}_i = \begin{cases} \{k_1, \ldots, k_{11}\} & \text{if } i < 6 \\ \{k_{m-10}, \ldots, k_m\} & \text{if } i > m - 5 \\ \{k_{i-5}, \ldots, k_{i+5}\} & \text{otherwise} \end{cases} \cdot$$

For each $S \in \mathcal{S}$, we do the following: for each $i = 1, \ldots, m$, a quadratic polynomial $P_{i,S}$ is fit to the set of points

$$\log(\mathcal{K}_i) \times \log\left(\bigcup_{k \in \mathcal{K}_i} \mathcal{E}_C(S, k)\right),$$

in the least squares sense where the logarithm function is understood to be taken over all the elements of the set; that is $\log \mathcal{A} = \{\log a \ : \ a \in A\}$.

Since we are using $n_0 = 10000$ samples to build the numerical approximation to the total relative error along a grid of S and k points, define $\text{err}_{\text{pol},1}$ to be the function

$$\text{err}_{\text{pol},1}(S, k, k_p) := \text{err}_{\text{pol}}(S, k, k_p, n)\Big|_{n=n_0}.$$

Here, $\mathrm{err}_{\mathrm{pol},1}$ is a function of only (S, k, k_p).

Then $\mathrm{err}_{\mathrm{pol},1}(S, k, k_p) = P_{i,k}(S_i)$ is set to be the point-estimate of the total relative error for $k_{on} = 1$, k_{off}, and $k_{pol} = k_p$. We use the log of the data for the local least squares fit since qualitatively the data is approximately quadratic on the log-scale.

After this procedure, a discrete grid of point-estimates for the total relative error of k_{on}, k_{off}, and $k_{pol} = k_p$ using 10,000 $T^{(1)}$ samples is obtained: $\mathcal{E} = \{\mathrm{err}_{\mathrm{pol},1}(S, k, k_p) \ : \ (S, k) \in \mathcal{S} \times \mathcal{K}\}$. Using linear interpolation on $\mathcal{E}$, we can then compute $\mathrm{err}_{\mathrm{pol},1}$, for any S and k pair a priori. The resulting total relative error surface is shown in figure 3.24. This procedure is repeated for every k_p as desired.

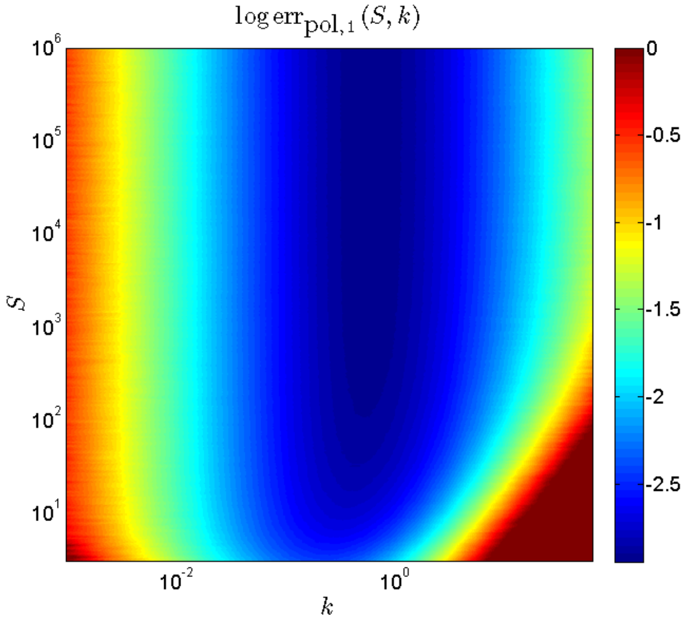

Figure 3.24: The total relative error surface $\mathrm{err}_{\mathrm{pol},1}(S, k)$ at $k_p = 0.5$ by local quadratic polynomial least-squares.

The black line in figure 3.25 is the trajectory $[dNTP] \mapsto (S, k)$ with $r_2 = 100, k_{on} = 200$, $k_{off} = 100$, and $k_{pol} = 50$ (so $k = 1$ and $k_p = 0.5$). We see that this trajectory shows good agreement with the full Monte Carlo simulations of std (err (k_{on})), std (err (k_{off})), and std (err (k_{pol})) in section 3.4.

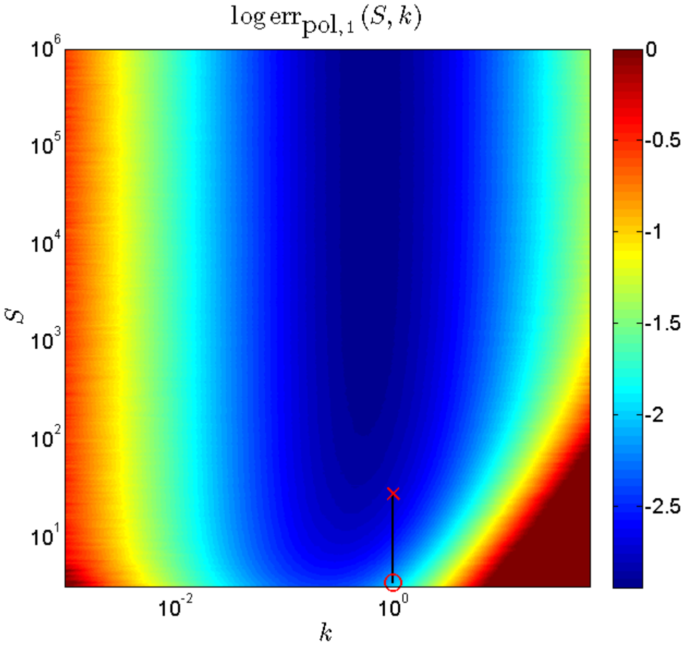

Figure 3.25: The total relative error surface $\mathrm{err}_{\mathrm{pol}}(U, \ell)$ at $k_p = 0.5$ by local quadratic polynomial least-squares. The black line is the trajectory $[dNTP] \mapsto (S, k)$ with $r_2 = 100, k_{on} = 200, k_{off} = 100$, and $k_{pol} = 50$. The red "O" and "X" denote the start and end of the trajectory at $[dNTP] = 2^{0.8}$ and $[dNTP] = 2^4$, respectively. We see that this trajectory shows good agreement with the full Monte Carlo simulations of *std* (*err* (k_{on})), *std* (*err* (k_{off})), and *std* (*err* (k_{pol}))

As evident by the expressions for U and ℓ in equations 3.35 and 3.36, both U and ℓ are dependent on k_p and hence the total relative error surface $\mathrm{err}_{\mathrm{pol,1}}(S, k, k_p)$ is dependent on k_p (figure 3.26). From figure 3.26, we see that at for high k_p, low values of k result in trajectories which lie entirely in high error regions. Thus we see that if k_p/k is large, then the total relative error may be high regardless of the $[dNTP]$ chosen.

The constructed total relative error $\mathrm{err}_{\mathrm{pol}}(S, k)$ provides a good estimate for the total relative error of k_{on}, k_{off}, and k_{pol}. To show this, we resample the cloud of data $\mathcal{E}_C$ at each (S, k) and repeat the quadratic polynomial smoothing procedure. This is repeated 1000 times and the resulting total relative error estimate at each

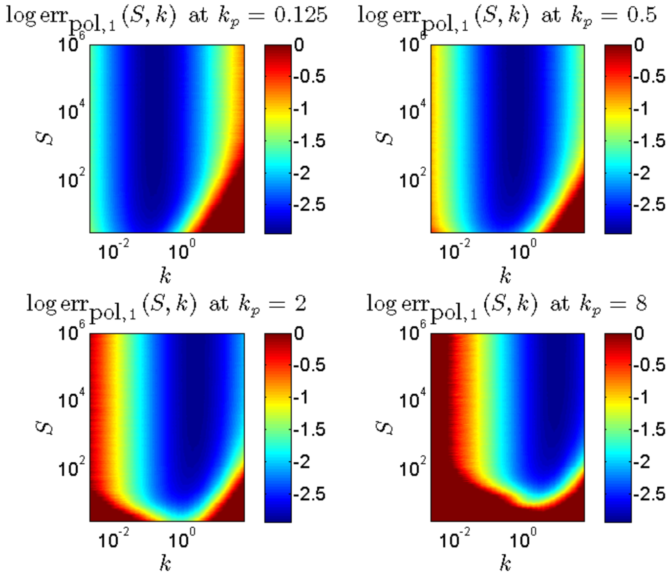

Figure 3.26: The total relative error surface $\mathrm{err}_{\mathrm{pol}}(S,k)$ at various values of k_p. This shows that the total relative error is dependent on the value of k_p.

(S,k) point is recorded so that a standard deviation at each (S,k) point can be obtained. Normalizing the standard deviation estimate by the total relative error gives the uncertainty of the total relative error estimate at each (S,k) point (figure 3.27). From the figure, we see that the relative error of the total relative error estimate is small and grows proportional to the inference uncertainty of k_{on}, k_{off}, and k_{pol}.

3.5.3 Comparison to the $k_{pol} = 0$ Case

After conditioning on the escape to the pre-translocation state, the state-space diagram of the Markov process describing the escape problem generating $T^{(1)}$ is seen in figure 3.22. This system can be formally viewed as a DNAP-DNA complex that cannot proceed to the pol-process and has effective dNTP binding and disassociation rates U and ℓ, respectively (equations 3.36 and 3.35). Viewed in this way, this modified system can be mapped to the $k_{pol} = 0$ case in Chapter 2.

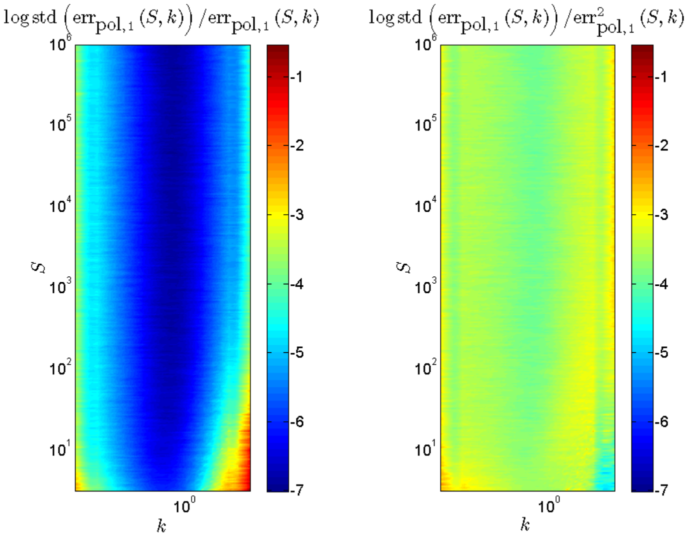

Figure 3.27: The left subplot shows the uncertainty of the total relative error estimate of k_{on}, k_{off}, and k_{pol} at $k_p = 0.5$ as produced by bootstrap resampling of the cloud of 20 data points $\mathcal{E}_C$ at each (S, k) point. The right subplot shows the left quantity divided by $\mathrm{err}_{\mathrm{pol}}$.

In fact, if in the $k_{pol} = 0$ case, the covariance matrix for MLE estimates of the mixture parameters was stored for each (S, k)-point, that same table can be used for the synthesizing $k_{pol} > 0$ cases.

Recall that U and ℓ are functions of k, k_p, and S. For fixed k_p, the effect on S and k on the effective binding and disassociation rates U and ℓ are not intuitive. In figure 3.28 we show the result of the mapping $(S, k) \mapsto (U(S, k, k_p), \ell(S, k, k_p))$ for fixed k_p.

The only parameter tunable in the nanopore experiments is $[dNTP]$. For fixed $r_2, k_{on}, k_{off}, k_{pol}$, both U and ℓ are functions of $[dNTP]$ only (since $k = k_{off}/r_2$ is fixed and $S = k_{on}/r_2[dNTP]$). We investigate the mapping $[dNTP] \mapsto (U, \ell)$ to see the behavior as a function of $[dNTP]$. Since we can formally view the system as a complex which cannot proceed to the pol-process with effective dNTP binding and disassociation rates U and ℓ, respectively, we plot a (U, ℓ) trajectory on the $\mathrm{err}_1(S, k)$ surface for the $k_{pol} = 0$ case (figure 3.29). Here we see that

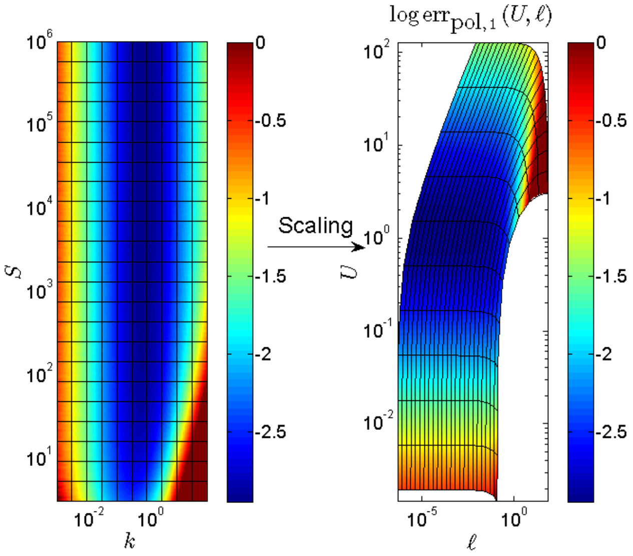

Figure 3.28: The total relative error surface $\mathrm{err}_{\mathrm{pol}}(S,k)$ at $k_p = 0.5$ when plotted on both the (S,k)-grid (left) and the (U,ℓ)-grid (right). The mapping from $(S,k) \mapsto (U(S,k,k_p), \ell(S,k,k_p))$ is not intuitive.

increasing $[dNTP]$ decreases the effective dNTP disassociation rate ℓ and increases the effective dNTP binding rate S. Note that in doing this, the total relative errors in $\mathrm{err}_{\mathrm{pol},1}$ and err_1 are slightly different. In $\mathrm{err}_{\mathrm{pol},1}$, the total relative error of k_{on}, k_{off}, and k_{pol} are recorded at each point, but in err_1, the total relative error of the effective binding and disassociation rates (i.e., U and ℓ, respectively) are recorded at each point.

From equations 3.35 and 3.36, we can readily see that at saturating $[dNTP]$, we see that $(U,\ell) \to (k_{off}/k_{pol}, 0) = (k/k_p, 0)$. As $[dNTP] \to 0$, we see that $(U,\ell) \to (0, (k_{off}+k_{pol})/r_2) = (0, k+k_p)$. To summarize,

$$\lim_{[dNTP]\to 0} (U,\ell) = (0, k+k_p), \tag{3.37}$$

$$\lim_{[dNTP]\to \infty} (U,\ell) = \left(\frac{k}{k_p}, 0\right). \tag{3.38}$$

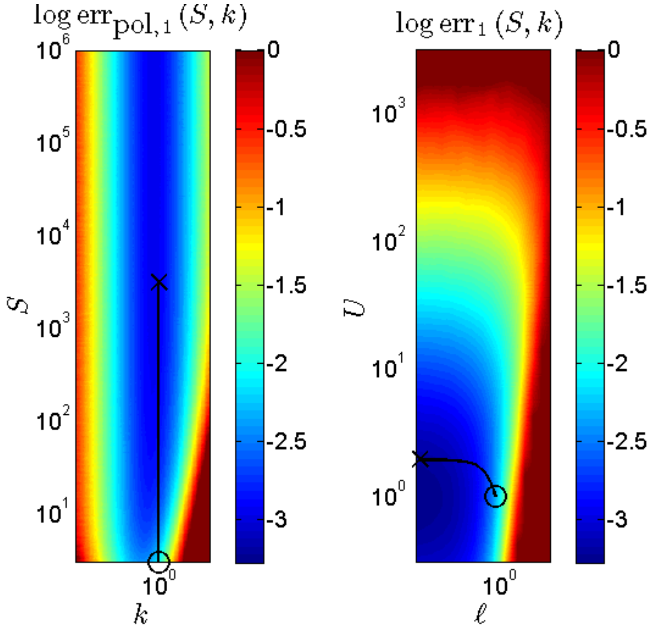

Figure 3.29: An example trajectory as $[dNTP]$ increases from $10^{0.5}$ to $10^{3.5}$ in the (S,k)-space for $k_p = 0.5$ and the (U,ℓ)-space in the equivalent non-synthesizing, $k_{pol} = 0$ system. The start and end of the trajectory are denoted by the "O" and "X" symbols, respectively.

Examining the derivatives of U and ℓ we have that

$$\begin{aligned}\frac{dU}{d[dNTP]} &= k_{on}k_{off}\frac{(k_{off}+k_{pol})\,r_2}{((k_{off}+k_{pol})\,r_2 + k_{pol}k_{on}[dNTP])^2}\\ \frac{d\ell}{d[dNTP]} &= -\left(k_{off}+k_{pol}\right)^2\left((k_{off}+k_{pol})\,r_2 + k_{pol}k_{on}[dNTP]\right)^{-2}k_{pol}k_{on}.\end{aligned}$$

Hence we see that

$$\begin{aligned}0 < \frac{dU}{d[dNTP]} &< \frac{k_{on}k_{off}}{2},\\ \frac{d\ell}{d[dNTP]} &< 0,\end{aligned}$$

for fixed r_2, k_{on}, k_{off}, and k_{pol}. The consequence of this is that for fixed transition rates, the trajectory $[dNTP] \mapsto (U,\ell)$ follows the line $\ell = (k_{off}+k_{pol})/r_2 = k+k_p$

for small $[dNTP]$ before bending left towards the line $U = k_{off}/k_{pol} = k/k_p$ for large $[dNTP]$. How quickly the trajectory bends left is determined by the derivatives of U and ℓ above.

From this analysis, we can see that as $[dNTP] \to \infty$ (and hence $S \to \infty$), the effective disassociation rate, ℓ, approaches 0; at the same time, the effective binding rate, U approaches the constant k/k_p. As $[dNTP] \to 0$ (and hence $S \to 0$), the effective disassociation rate approaches the constant $k + k_p$ and the effective binding rate approaches 0.

From here, we can intuitively see why saturating $[dNTP]$ gives the least total relative error but the total relative error does not decrease indefinitely, instead approaching a small constant. For large $[dNTP]$, the DNAP-DNA complex has a very high probability of binding a dNTP and traveling to the dNTP-bound state, determined by the effective binding rate, U. Once in the dNTP-bound state, the complex will remain there for a very long, but finite time, determined by the effective disassociation rate ℓ. Once the complex disassociates the bound dNTP, the complex binds another dNTP with high probability, remaining in the dNTP-bound state until eventual disassociation. This path does not continue indefinitely since there is a small probability of escaping to the pre-translocation state upon each visitation back to the post-translocation state. Eventually the complex escapes to the pre-translocation state after a very long time. In this process, both the post-translocation and dNTP-bound states must have been visited very many times and so the two exponential modes are easily discernible from the dwell-time data. This hence leads to a low total relative error for k_{on}, k_{off}, and k_{pol}. The total relative error does not decrease indefinitely with high $[dNTP]$ however since the effective binding rate is bounded.

Similarly, with low $[dNTP]$, the complex has a very low probability of binding a dNTP and hence immediately escapes to the pre-translocation state. The result is that the dNTP-bound state is rarely visited and hence the two exponential modes are not easily discernible.

3.6 Optimum Experimental Condition

In this section, we examine the optimal experimental condition that when achieved, produces the least total relative error.

3.6.1 Finding the optimal $[dNTP]$

Like in Chapter 2, from the scaling laws and the total relative error point estimates in $\mathcal{E}$, we can numerically obtain the $[dNTP]$ that yields the least total relative error for any k_{on}, k_{off}, and k_{pol}.

Notice that for fixed k and k_p, both U and ℓ are functions of S only. For fixed k and k_p we have the trajectory $S \mapsto \mathrm{err}_{\mathrm{pol},1}(S, k, k_p) = \mathrm{err}_{\mathrm{pol},1}(S)$. Let S^* be the optimum S in the sense that

$$S^* = \arg\min_S \mathrm{err}_{\mathrm{pol},1}(S).$$

For fixed k_p, S^* is entirely dependent on k. Hence we can write,

$$S^* = F(k) \Leftrightarrow [dNTP]^* = \frac{r_2}{k_{on}} F\left(\frac{k_{off}}{r_2}\right),$$

for fixed k_p and where F is some function of k.

Determining an expression for F is very difficult, so we turn to numerical approximation. For each $k \in \mathcal{K}$, we approximate F by finding the location of the minimum of $\mathrm{err}_{\mathrm{pol},1}(S)$. As it turns out, $\mathrm{err}_{pol,1}$ is minimum at saturating $[dNTP]$

for many values of k, and in this saturating $[dNTP]$ region, the total relative error approaches some small value (figure 3.30). This behavior is confirmed by the full Monte Carlo simulations for $k = 1$ and $k_p = 0.5$ (figures 3.10, 3.12, and 3.14).

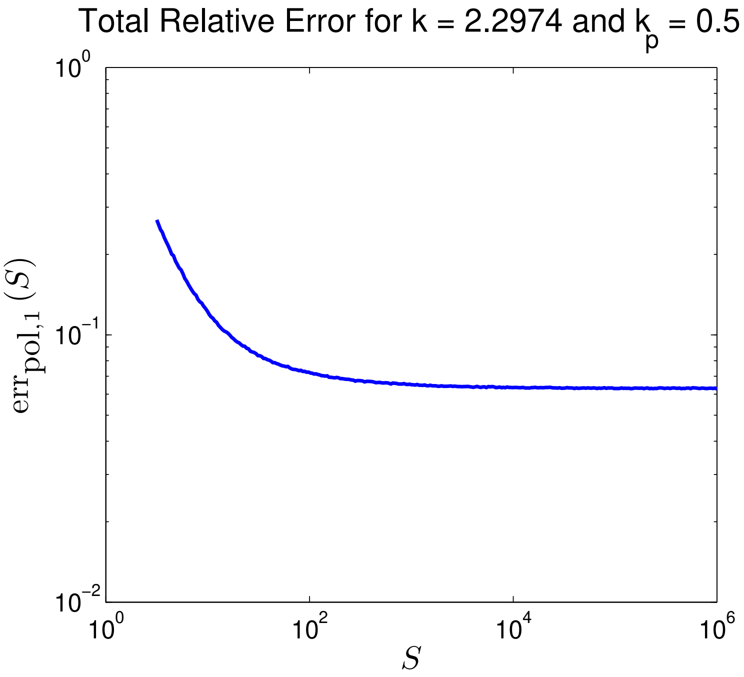

Figure 3.30: A typical plot of the total relative error along a trajectory $S \mapsto \mathrm{err}_{\mathrm{pol},1}(S)$. In agreement the plot of $\mathrm{err}_{\mathrm{pol},1}$ in figure 3.24, the minimum total relative error occurs at saturating $[dNTP]$ for most values of k.

The result of this is that the minimum total relative error is not numerically well defined in this region. Instead, it makes more sense to find the $p\%$ interval from the approximated minimum for each k. Any S within this interval is therefore guaranteed to give a total relative error within $p\%$ of the total relative error at S^*.

The $p\%$ interval is obtained as follows. For fixed k_p and for each $k \in \mathcal{K}$, $\min_S \mathrm{err}_{\mathrm{pol},1}(S)$ is obtained. Then the solution to the equation $\mathrm{err}_{\mathrm{pol},1}(S) = \left(\frac{p}{100} + 1\right) \mathrm{err}_{\mathrm{pol},1}(S^*)$ for each k gives the $p\%$ interval. After this procedure, we smooth the interval by a cubic smoothing spline [62]. The smoothing cubic spline is necessary since numerically finding the minimum total relative error produces noisy results due to the flat nature of the total relative error in saturating $[dNTP]$ regimes. Figure 3.31 shows a 5% interval for $k_p = 0.5$. The upper-bound of the interval does not exist the majority of k. For each k, any S within the bounds of the interval is guaranteed to produce a total relative error within 5% of the minimum total relative error.

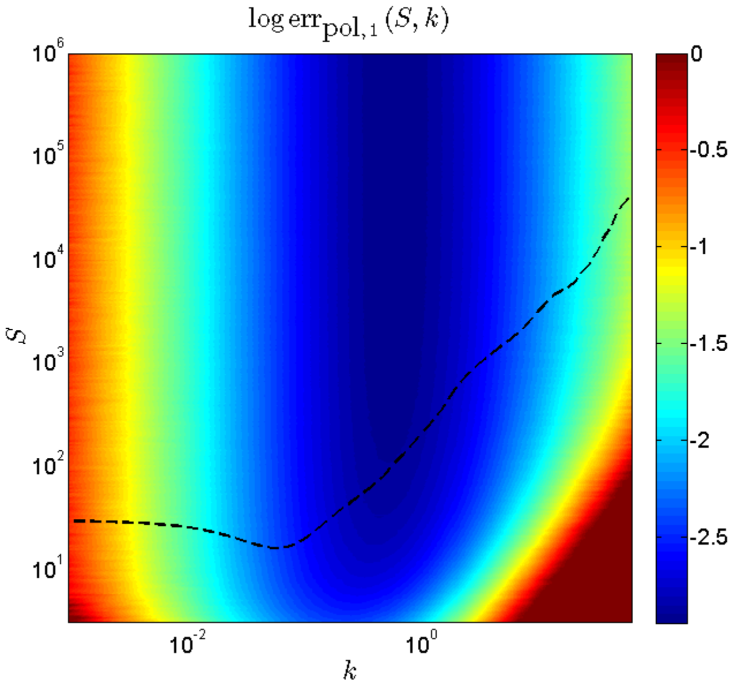

Figure 3.31: The total relative error $\mathrm{err}_{\mathrm{pol},1}$ for $k_p = 0.5$ with the 5% interval after applying a cubic smoothing polynomial to the lower-bound of the interval. The upper-bound of the interval does not exist the majority of k. For each k, any S within the bounds of the interval is guaranteed to produce a total relative error within 5% of the minimum total relative error.

Figure 3.32 shows the 5%, 10%, and 25% intervals plotted on $\mathrm{err}_{\mathrm{pol},1}$ for $k_p = 0.5$ along with the trajectory for $k = 1$ from $[dNTP] = 2^{0.8}$ to $[dNTP] = 2^4$. The different percentile intervals highlight the rate of decay of the total relative error for and show that at $[dNTP] = 2^4$, the total relative error obtained is within 25% of the minimum total relative error. Again, the upper-bound of the interval does

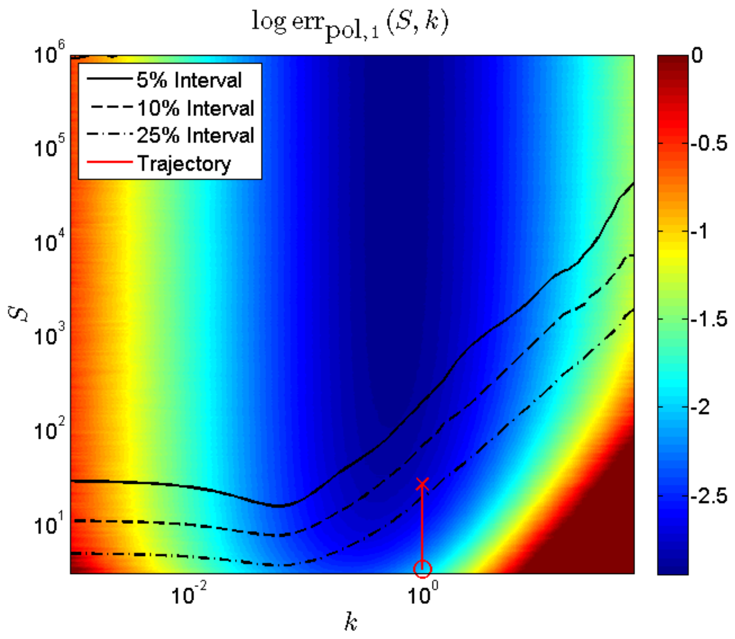

Figure 3.32: A plot of $\mathrm{err}_{\mathrm{pol},1}$ for $k_p = 0.5$ with 5%, 10%, and 25% intervals along with a trajectory for $k = 1$ from $[dNTP] = 2^{0.8}$ to $[dNTP] = 2^4$. The red "O" and "X" denote the start and end of the trajectory.

not exist for the majority of k, due to the monotonically decreasing total relative error as S increases for fixed k (figure 3.30).

3.6.2 Finding the Optimal $[dNTP]$ Under Experimental Time Constraints

The PDF of $T^{(1)}$ is a function of $r_2, k_{on}, k_{off}, k_{pol}$, and $[dNTP]$. The only tunable parameter in the nanopore experiments is $[dNTP]$. From the full Monte Carlo simulations in section 3.4 and the estimate of $\mathrm{err}_{\mathrm{pol}}$ above, we see that for most k_{off}/r_2, the total relative error decreases as $[dNTP]$ increases. However, it can be very expensive to collect a sufficient number of $T^{(1)}$ samples with high $[dNTP]$ concentrations. When $[dNTP]$ is high, there is a high probability of nucleotide binding. Thus upon nucleotide disassociation, the DNAP-DNA complex immediately encounters and binds another nucleotide. This immediate binding makes nucleotide incorporation through the pol-process highly probable, and hence the DNAP-DNA complex enters the next nucleotide addition cycle. This can be seen from the expression for $p_{E_{pol}|2}$ in equation 3.12. From this, we see that as $[dNTP] \to \infty$, $p_{E_{pol}|2} \to 1$. Hence for large $[dNTP]$, the number of $T^{(1)}$ samples will be very small.

In this section, we look into finding the optimal $[dNTP]$ that results in the least total relative error under a maximum experimental time-constraint $\tau_{\max}$. Here, $\tau_{\max}$ is the maximum time allowed for the experiment to run.

It is reasonable to assume that $\text{err}_{\text{pol}}(S, k, k_p, n) = O\left(1/\sqrt{n}\right)$ as $n \to \infty$ since the standard error of a parameter scales as $O\left(1/\sqrt{n}\right)$ as $n \to \infty$ [13]. Indeed, figure 3.33 shows $\text{err}_{\text{pol}}(S, k, k_p, n)$ at $S = 10^3$, $k = 1$, and $k_p = 0.5$ for n ranging from 10^4 to 10^5. Here, we see that err_{pol} scales as $O\left(1/\sqrt{n}\right)$ as $n \to \infty$ for fixed S, k, and k_p. Hence just like the non-synthesizing case in Chapter 2, it is reasonable

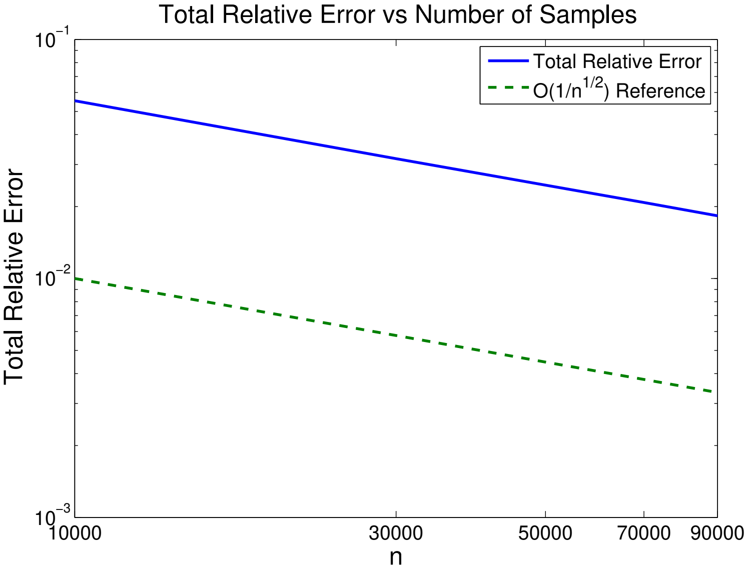

Figure 3.33: The total relative error function $\text{err}_{\text{pol}}(S, k, k_p, n)$ vs n for $S = 10^3$, $k = 1$, and $k_p = 0.5$. Here see that the total relative error scales as $O\left(1/\sqrt{n}\right)$ as $n \to \infty$.

to conclude the following scaling law for the total relative error function:

$$\mathrm{err}_{\mathrm{pol}}\left(S,k,k_p,n_1\right)\approx \mathrm{err}_{\mathrm{pol}}\left(S,k,k_p,n_2\right)\sqrt{\frac{n_2}{n_1}}, \tag{3.39}$$

for large n_1 and n_2.

Recall that we numerically approximated $\mathrm{err}_{\mathrm{pol},1}$ using $n_0 = 10000$ points. From equation 3.39, we thus have the approximation,

$$\mathrm{err}_{\mathrm{pol}}\left(S,k,k_p,n\right)\approx \sqrt{\frac{n_0}{n}}\mathrm{err}_{\mathrm{pol},1}\left(S,k,k_p\right), \tag{3.40}$$

for large n.

Let N be the number of $T^{(1)}$ samples obtained before the DNAP-DNA complex incorporates a dNTP and hence proceeds to the next nucleotide addition cycle. Notice that $N \sim \text{geometric}\left(p_{E_{pol}|2}\right)$, where $\text{geometric}\left(p\right)$ is the geometric distribution with parameter p with support $\mathbb{N}\cup\{0\}$. Hence the mean number of $T^{(1)}$ samples within a nucleotide addition cycle is given by

$$\langle N\rangle = \frac{1-p_{E_{pol}|2}}{p_{E_{pol}|2}}.$$

Let n be the number of $T^{(1)}$ samples that we want to observe from the nanopore experiments. The mean number of nucleotide addition cycles required to observe n samples is given by $n/\langle N\rangle$. Hence the mean amount of time required to observe n samples of $T^{(1)}$ is given by

$$\langle T_{cycle}\rangle \frac{n}{\langle N\rangle},$$

where $\langle T_{cycle}\rangle$ is the mean total cycle time (equation 3.19). We will use this quantity to constrain our optimization; that is, find the dNTP concentration which produces the minimum total relative error such that

$$\langle T_{cycle} \rangle \frac{n}{\langle N \rangle} = \tau_{\max}.$$

That is, we are interested in solving the constrained optimization problem,

$$\text{minimize } \mathrm{err}_{\mathrm{pol}} \left(S, k, k_p, n \right), \tag{3.41}$$

$$\text{subject to } \langle T_{cycle} \rangle \frac{n}{\langle N \rangle} = \tau_{\max}. \tag{3.42}$$

With k and k_p intrinsic to the DNAP-DNA complex, the only tunable parameters are n and S.

We recast the constrained optimization problem in equations 3.41-3.42 to an unconstrained optimization in S only. From the constraint in equation 3.42, we have

$$n = \tau_{\max} \frac{\langle N \rangle}{\langle T_{cycle} \rangle}. \tag{3.43}$$

Hence we have that

$$\mathrm{err}_{\mathrm{pol,2}} := \mathrm{err}_{\mathrm{pol}} \left(S, k, k_p, n \right) \approx \sqrt{\frac{n_0}{\tau_{\max}} \frac{\langle T_{cycle} \rangle}{\langle N \rangle}} \mathrm{err}_{\mathrm{pol,1}} \left(S, k, k_p \right). \tag{3.44}$$

From equation 3.44, we can find the optimal S which solves the optimization problem in equations 3.41-3.42.

Figure 3.34 shows the plot of $\mathrm{err}_{\mathrm{pol,2}}$ with $\tau_{\max} = 50000$ and $n_0 = 10000$, along with the optimal S trajectory as a function of k and its 5%, 10%, and 25% intervals. The constrained optimal trajectory and its 5%, 10%, and 25% intervals were found in the same way as the unconstrained case in which no constraints in the experimental time were used.

Unlike the unconstrained case, the optimal $[dNTP]$ has a well defined location

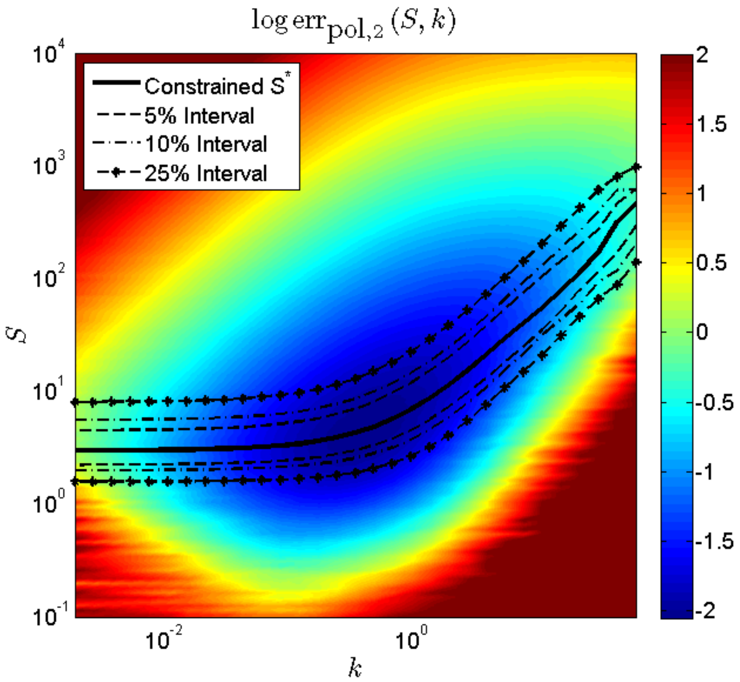

Figure 3.34: The constrained total relative error function $\mathrm{err}_{\mathrm{pol},2}(S,k,k_p)$ at $k_p = 0.5$ along with the optimal constrained S and its 5%, 10%, and 25% intervals with $\tau_{\max} = 50000$ and $n_0 = 10000$ for $k_p = 0.5$.

for the constrained optimization. This is because larger dNTP concentrations decrease the amount of $T^{(1)}$ samples that can be collected per nucleotide addition cycle.

3.6.3 Finding the Optimal $[dNTP]$ when Constraining the Number of Cycles

In the previous subsection, the total experimental time was constrained. In that situation, it is implicitly implied that the cost of collecting each cycle is small relative to to the cost of collecting each $T^{(1)}$ sample. In this subsection, we look at constraining the number of nucleotide addition cycles instead of the total time cost. This situation is useful in situations in which the cost of observing each cycle is high. For example, capturing a DNAP-DNA complex atop the nanopore at the beginning of each cycle takes a long time. Let $\eta_{\max}$ be the maximum number of cycles that is to be collected in the experiment. Then $\eta_{\max}\left\langle N\right\rangle$ is the average samples of $T^{(1)}$ collected with $\eta_{\max}$ number of cycles. The relevant optimization problem is thus

$$\text{minimize } \text{err}_{\text{pol}}\left(S,k,k_p,n\right), \tag{3.45}$$

$$\text{subject to } n=\eta_{\max}\left\langle N\right\rangle. \tag{3.46}$$

Like the previous constrained optimization problem, this is recast into an unconstrained problem by minimizing

$$\text{err}_{\text{pol,3}}\left(S,k,k_p\right):=\text{err}_{\text{pol}}\left(S,k,k_p,n\right)\approx\sqrt{\frac{n_0}{\eta_{\max}\left\langle N\right\rangle}}\text{err}_{\text{pol,1}}\left(S,k,k_p\right).$$

Figure 3.35 shows the constrained optimal S with 5%, 10%, and 25% intervals for $k_p=0.5$ with $\eta_{\max}=10000$. The graph is qualitatively similar to figure 3.34.

3.6.4 Validity of the Mean-Field Approaches

In this section, we numerically validate the mean-field approaches used in both of the constrained optimization problems.

In the constrained optimization problem, we minimized the total relative error $\text{err}_{\text{pol}}\left(S,k,k_p,n\right)$ under the constraint $\left\langle T_{cycle}\right\rangle n/\left\langle N\right\rangle=\tau_{\max}$. However, under real experimental settings, the samples of $T^{(1)}$ would be collected until the maximum time $\tau_{\max}$ has elapsed. Let $T_{cycle,j}$ be the j-th nucleotide addition cycle observed. Then

$$T_{cycle,j}=\sum_{i=1}^{N_j}\left(T_{j,i}^{(1)}+T_{j,i}^{(2)}\right)+T_{B,j},$$

where N_j is the number of $T^{(1)}$ segments in the j-th nucleotide addition cycle;

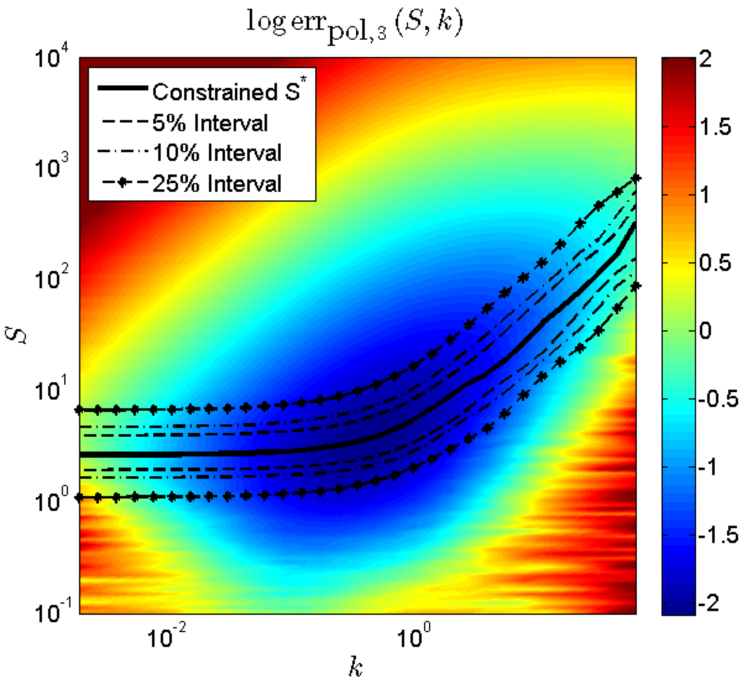

Figure 3.35: The error surface $\mathrm{err}_{\mathrm{pol},3}$ with $\eta_{\max} = 10000$ for $k_p = 0.5$ along with the optimal S trajectory with 5%, 10%, and 25% intervals.

that is, $N_j \sim \text{geometric}\left(p_{E_{pol}|2}\right)$ with support $\{0\} \cup \mathbb{N}$. Let M be the number of nucleotide addition cycles such that

$$\sum_{j=1}^{M} T_{cycle,j} \leq \tau_{\max}, \tag{3.47}$$

but

$$\sum_{j=1}^{M+1} T_{cycle,j} > \tau_{\max}. \tag{3.48}$$

The $T^{(1)}$ samples are collected until the inequality

$$\sum_{j=1}^{M} T_{cycle,j} + \sum_{i=1}^{n_{M+1}} \left(T^{(1)}_{M+1,i} + T^{(2)}_{M+1,i}\right) + T^{(1)}_{M+1,n_{M+1}+1} I \leq \tau_{\max} \tag{3.49}$$

is tight, where I is the indicator function

$$I = \begin{cases} 1 & \text{if } T^{(1)}_{M+1,n_{M+1}+1} \leq \tau_{\max} - \sum_{j=1}^{M} T_{cycle,j} + \sum_{i=1}^{n_{M+1}} \left(T^{(1)}_{M+1,i} + T^{(2)}_{M+1,i} \right) \\ 0 & \text{otherwise} \end{cases} .$$

Tight means that M is such that inequalities 3.47 and 3.48 hold and $n_{M+1} \leq N_{M+1}$ where n_{M+1} is such that

$$\sum_{j=1}^{M} T_{cycle,j} + \sum_{i=1}^{n_{M+1}} \left(T^{(1)}_{M+1,i} + T^{(2)}_{M+1,i} \right) \leq \tau_{\max},$$

but

$$\sum_{j=1}^{M} T_{cycle,j} + \sum_{i=1}^{n_{M+1}+1} \left(T^{(1)}_{M+1,i} + T^{(2)}_{M+1,i} \right) > \tau_{\max}.$$

The number of $T^{(1)}$ samples, n, observed constraining the experimental time to $\tau_{\max}$ is thus

$$n = \sum_{j=1}^{M} N_j + n_{M+1} + I.$$

This is different than the mean field approach taken when solving the optimization problem in equations 3.41 and 3.42.

The mean field approach greatly simplifies the calculation of the solution to the constrained optimization problem by replacing the behavior of the large number of random variables $(T_{cycle_1}, \ldots, T_{cycle_{M=m}})$ with the ensemble average. We demonstrate the validity of the mean field approach by numerical simulation. For the constrained optimization case, we set $S = 5$, $k = 0.5$, and $k_p = 0.5$ with $\tau_{\max} = \langle T_{cycle} \rangle n_{mf} / \langle N \rangle$ where $n_{mf} = 10000$. The total relative error is then estimated using the mean field approach and using the constraint in equation 3.49 using 2000 data sets. As seen in figure 3.36, both approaches are in agreement and is well approximated by $\mathrm{err}_{\mathrm{pol},2}(S, k, k_p)$.

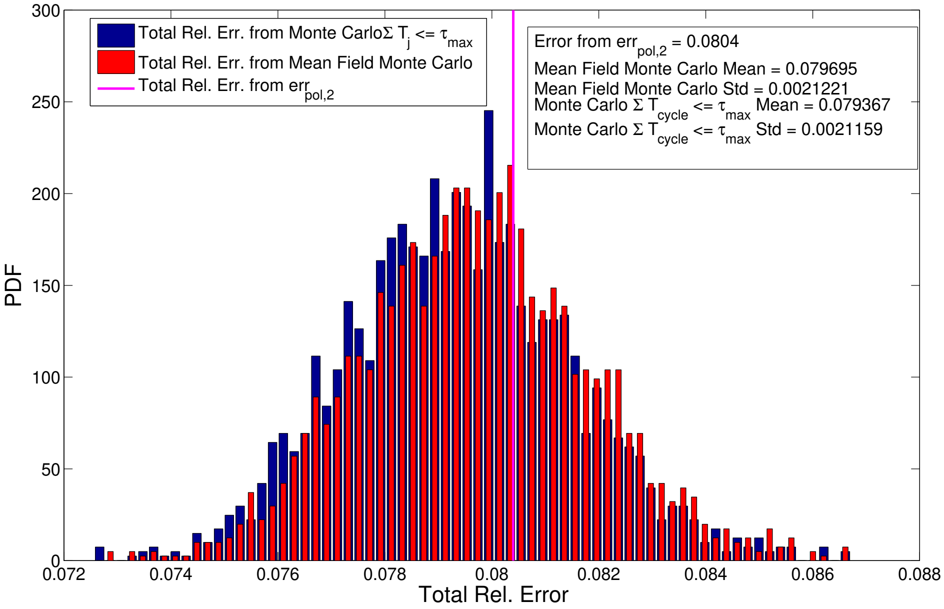

Figure 3.36: Comparison between the constrained total relative error $\mathrm{err}_{\mathrm{pol},2}$ obtained from the mean field approach and from using the total time from the observed $T^{(1)}$ samples using constraint 3.49. Here, we use $\sum_j T_{cycle,j}$ as an abbreviation for constraint 3.49. Also, $S = 5, k = 0.5, k_p = 0.5$, and $\tau_{\max} = \langle T_{cycle} \rangle n_{mf} / \langle N \rangle$, where $m_{mf} = 10000$. The distributions of the mean-field approach and the total run-time constrains are in agreement. The mean of the mean-field approach can be calculated from $\mathrm{err}_{\mathrm{pol},2}$.

We do the same comparison for the mean field approach taken for the optimization problem in equations 3.45-3.46. In this situation, we replaced the behavior of $N_1, \ldots, N_{\eta_{\max}}$ with the average $\langle N \rangle$. That is, the constraint in equation 3.46 is given by the mean number of $T^{(1)}$ samples after $\eta_{\max}$ nucleotide addition cycles, $n = \eta_{\max} \langle N \rangle$ instead of $n = N_1 + \cdots N_{\eta_{\max}}$. As before, this has the advantage of simplifying the optimization problem. The two approaches are equivalent as we see in figure 3.37.

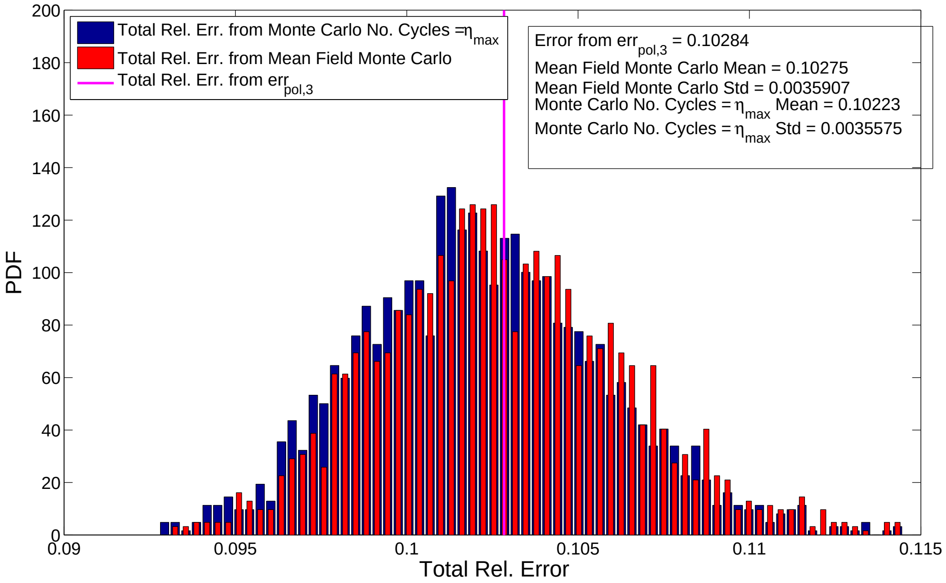

Figure 3.37: Comparison between the constrained total relative error $\mathrm{err}_{\mathrm{pol},3}$ obtained from the mean field approach and from using the total time from the observed $T^{(1)}$ samples using constraint 3.46. Also, $S = 5, k = 1, k_p = 0.5$, and $\eta_{\max} = m_{mf}/\langle N\rangle$, where $m_{mf} = 10000$. The distributions of the mean field approach and the total run-time constrains are in agreement. The mean of the mean-field can be calculated from $\mathrm{err}_{\mathrm{pol},3}$.

3.7 Numerical Simulations: Case with Detection Uncertainty

In this chapter, we repeat the numerical simulations done in chapter 3.4, adding multiplicative noise in the observed $T^{(1)}$ and T_B samples. Here, the observed $T^{(1)}$ and T_B samples are of the form

$$T^{(1)}_{\mathrm{obs}} = T^{(1)} e^{\sigma\zeta},$$

$$T_{B,\mathrm{obs}} = T_B e^{\sigma\zeta},$$

where $\zeta \sim N(0, 1)$.

The plots from the $T^{(1)}$ data with multiplicative noise are given in figures 3.38-3.40. The plots from the T_B data with multiplicative noise are given in figures 3.41-3.43

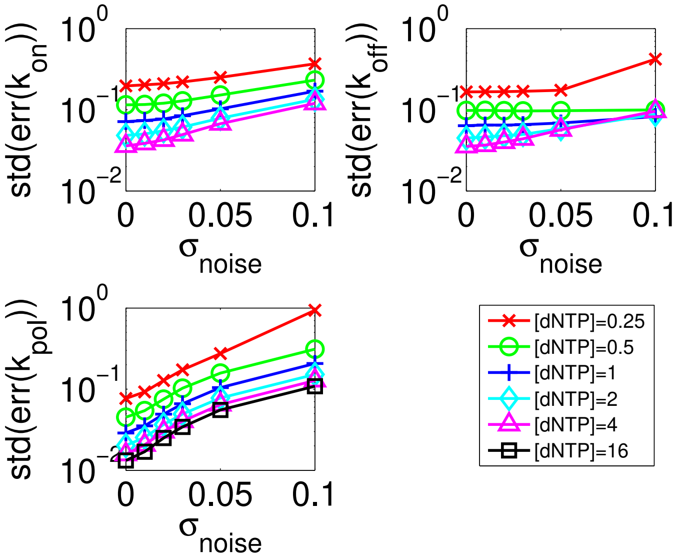

Figure 3.38: Plot of std (err (k)) vs σ from $T^{(1)}$ data.

For $T^{(1)}$, the higher dNTP concentrations produce the least amount of relative error throughout the noise magnitudes tested, and all of the concentrations except for $[dNTP] = 0.25$ have low bias. For T_B, the relative error for k_{on} is smallest throughout the noise magnitudes tested at $[dNTP] = 0.5$. For k_{off}, $[dNTP] = 1$ and $[dNTP] = 2$ produce the smallest relative error when using the T_B data. Like when inferring from $T^{(1)}$ data, for k_{pol}, the larger $[dNTP]$ concentrations produce the smallest relative error when using T_B data. The bias is larger when compared to the bias inference from the $T^{(1)}$ data for k_{on} and k_{off}. For k_{pol} the bias is small for $[dNTP] \geq 1$ when using T_B data. We see that throughout all the noise magnitudes tested, inferring from the $T^{(1)}$ samples produce lower relative errors for all of the transition rates than when inferring from the T_B data with respect to their optimum dNTP concentrations.

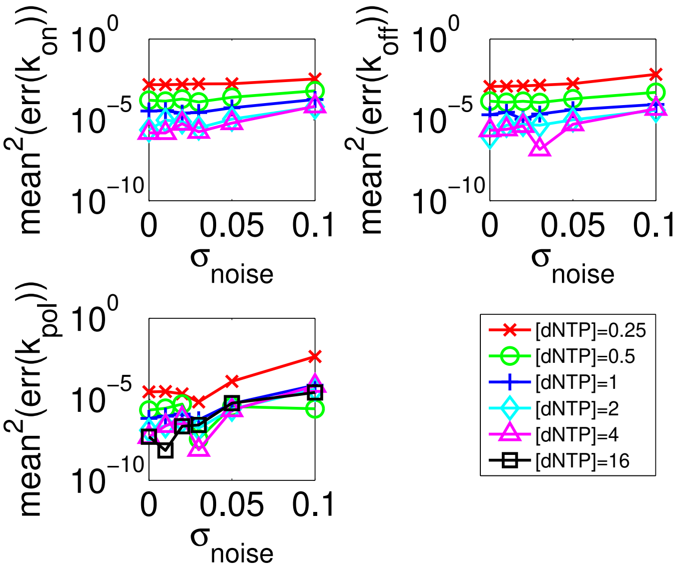

Figure 3.39: Plot of mean $(\text{err}(k))$ vs σ from $T^{(1)}$ data.

It is also worthwhile to investigate how the relative error changes as the number of samples of $T^{(1)}$ and T_B changes as a function of σ. We use $[dNTP] = 2$ and repeat the above numerical experiments but this time varying the number of samples of $T^{(1)}$ and T_B.

The plots from the $T^{(1)}$ data are given in figures 3.44-3.46. The plots from the T_B data are given in figures 3.47-3.49

Throughout the range of σ and the number of samples n tested, we see that using the $T^{(1)}$ samples at $[dNTP] = 2$ provides a lower relative error for each transition rate k_{on}, k_{off}, and k_{pol} when compared to inferring the transition rates using T_B data. We also see that when using $T^{(1)}$ data, collecting about 5000 samples is sufficient for obtaining reasonable estimates for the transition rates (the relative error is less than 20% for high noise magnitudes and less than 10$ for low noise magnitudes).

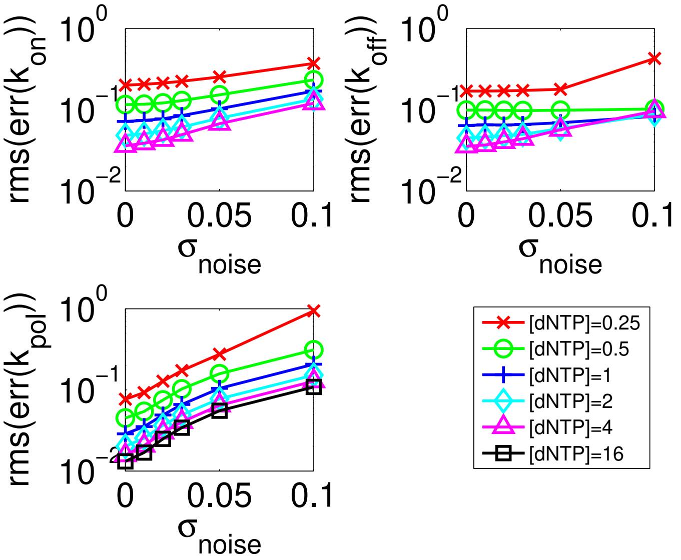

Figure 3.40: Plot of $\mathrm{rms}\left(\mathrm{err}\left(k\right)\right)$ vs σ from $T^{(1)}$ data.

3.8 Characterizing the Effect of Measurement Noise

In this section, we characterize the effect of measurement noise on the observed $T^{(1)}$ samples. Suppose that the true $T^{(1)}$ samples are perturbed by multiplicative noise of the form $e^{\sigma\zeta}$ where $\zeta \sim N\left(0,1\right)$. That is, we observe the $T^{(1)}$ samples to be

$$T^{(1)}_{\mathrm{obs}} := T^{(1)} e^{\sigma\zeta}.$$

In this section, we denote $k_{on}^{\mathrm{MLE}}\left(\sigma\right)$, $k_{off}^{\mathrm{MLE}}\left(\sigma\right)$, and $k_{pol}^{\mathrm{MLE}}\left(\sigma\right)$ to be the maximum-likelihood estimate of k_{on}, k_{off}, and k_{pol} respectively from the perturbed $T^{(1)}_{\mathrm{obs}}$ data

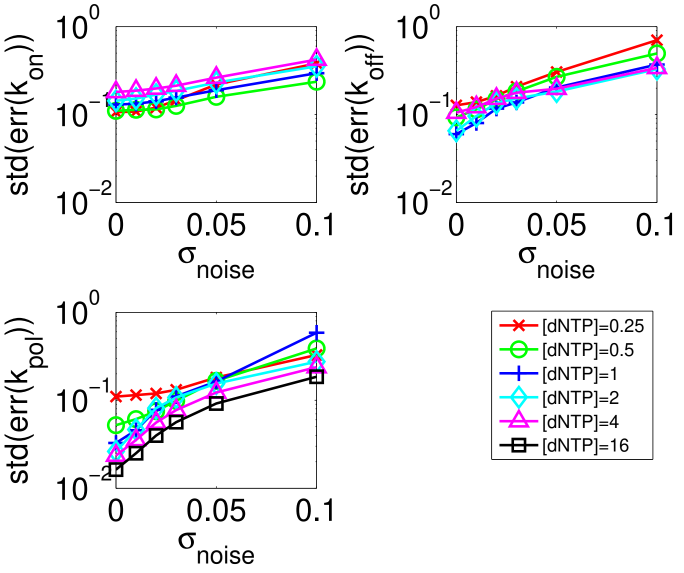

Figure 3.41: Plot of $\mathrm{std}\left(\mathrm{err}\left(k\right)\right)$ vs σ from T_B data.

with noise $e^{\sigma\zeta}$. We investigate the first-two moments and variance of the quantities

$$z_{k_{on}} := k_{on}^{\mathrm{MLE}}\left(\sigma\right) - k_{on}^{\mathrm{MLE}}\left(0\right),$$

$$z_{k_{off}} := k_{off}^{\mathrm{MLE}}\left(\sigma\right) - k_{off}^{\mathrm{MLE}}\left(0\right),$$

$$z_{k_{pol}} := k_{pol}^{\mathrm{MLE}}\left(\sigma\right) - k_{pol}^{\mathrm{MLE}}\left(0\right).$$

For the following simulations, we use 10,000 data sets with $r_2 = k_{on} = 1$, $k_{off} = 0.8$, $k_{pol} = 0.5$, and $[dNTP] = 10^3$. The number of $T^{(1)}$ samples, n, as well as the measurement noise σ is varied. Figure 3.50 shows the squared-mean of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ as a function of σ. From here, we see that $\left\langle z_{k_{on}}\right\rangle$, $\left\langle z_{k_{off}}\right\rangle$, and $\left\langle z_{k_{pol}}\right\rangle = O\left(\sigma^2\right)$.

Figures 3.51 and 3.52 show the second moment of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ as a function of σ and n, respectively. From these results, we see that as $\sigma \to 0$ and $n \to \infty$, we have that $\left\langle z_{k_{on}}^2\right\rangle, \left\langle z_{k_{off}}^2\right\rangle, \left\langle z_{k_{pol}}^2\right\rangle = O\left(\sigma^2/n\right)$. For large σ, we have that $\left\langle z_{k_{on}}^2\right\rangle, \left\langle z_{k_{off}}^2\right\rangle, \left\langle z_{k_{pol}}^2\right\rangle = O\left(\sigma^2/n\right) = O\left(\sigma^4\right)$. We can write this more

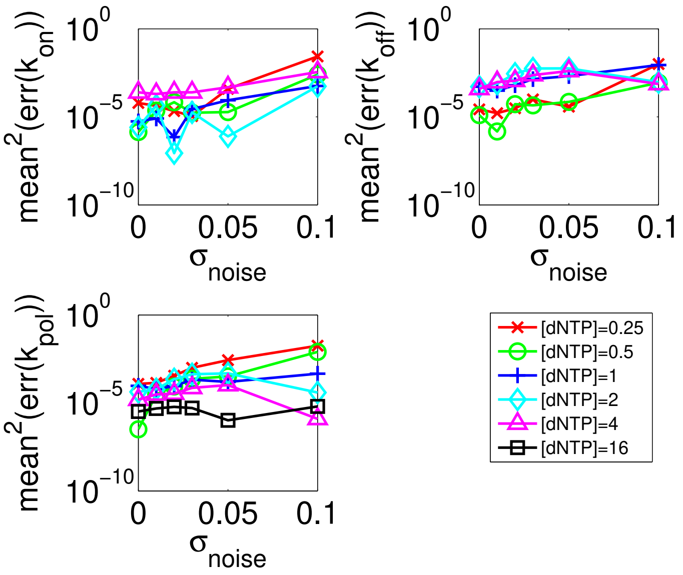

Figure 3.42: Plot of mean $(\text{err}(k))$ vs σ from T_B data.

compactly as $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle, \left\langle z_{k_{pol}}^2 \right\rangle = O\left(\sigma^2/n\right) + O\left(\sigma^4\right)$.

Figures 3.53 and 3.54 shows the variance of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ as a function of σ and n, respectively. From these results, we see that $var\left(k_{on}\right)$, $var\left(k_{off}\right)$, and $var\left(k_{pol}\right)$ behave as $O\left(\sigma^2/n\right)$.

The results in these numerical simulations show that we have strong numerical evidence for the following claims:

1. $\left\langle z_{k_{on}} \right\rangle, \left\langle z_{k_{off}} \right\rangle, \left\langle z_{k_{pol}} \right\rangle = O\left(\sigma^2\right)$,
2. $\left\langle z_{k_{on}}^2 \right\rangle, \left\langle z_{k_{off}}^2 \right\rangle, \left\langle z_{k_{pol}}^2 \right\rangle = O\left(\sigma^2/n\right) + O\left(\sigma^4\right)$, and
3. $var\left(z_{k_{on}}\right), var\left(z_{k_{off}}\right), var\left(z_{k_{pol}}\right) = O\left(\sigma^2/n\right)$.

From figure 3.55, we see that $z_{k_{on}}, z_{k_{off}}$, and $z_{k_{pol}}$ are normally distributed. There are primarily two ways of collecting the perturbed $T^{(1)}$ samples:

- Type 1: For each 10,000 data sets, n $T^{(1)}$ samples are generated and those n samples are perturbed by multiplicative noise with standard deviation σ.

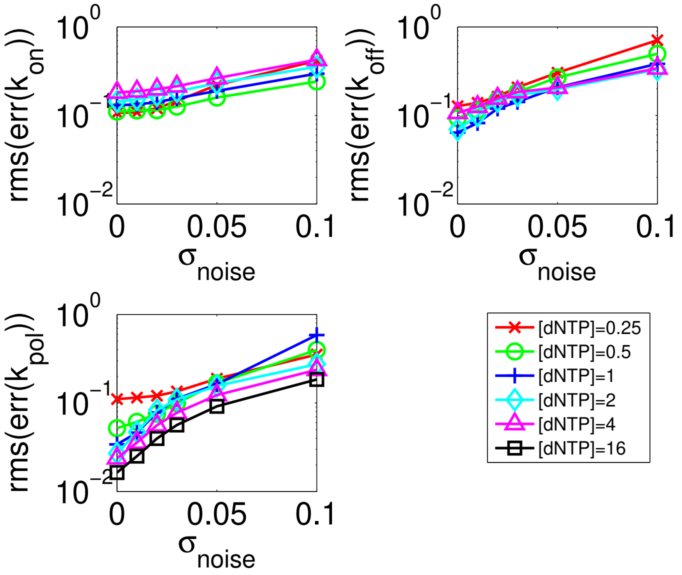

Figure 3.43: Plot of $\mathrm{rms}\left(\mathrm{err}\left(k\right)\right)$ vs σ from T_B data.

- n, $T^{(1)}$ samples are generated and those n samples are perturbed by 10,000 realizations of multiplicative noise with standard deviation σ.

As shown numerically in Chapter 2, both of these methods of perturbing $T^{(1)}$ are equivalent in distribution.

From claims 1-3 above and the observation that $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ are normally distributed, we can write

$$z_{k_{on}} = c_{2,k_{on}}\sigma^2 + c_{1,k_{on}}\frac{\sigma}{\sqrt{n}}\zeta, \tag{3.50}$$

$$z_{k_{off}} = c_{2,k_{off}}\sigma^2 + c_{1,k_{off}}\frac{\sigma}{\sqrt{n}}\zeta, \tag{3.51}$$

$$z_{k_{pol}} = c_{2,k_{pol}}\sigma^2 + c_{1,k_{pol}}\frac{\sigma}{\sqrt{n}}\zeta, \tag{3.52}$$

where $\zeta \sim N\left(0,1\right)$.

Similarly to the $k_{pol} = 0$ case in Chapter 2, the consequence of this result is that the bias of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ is deterministic and scales as $c_{2,k_{on}}\sigma^2$, $c_{2,k_{off}}\sigma^2$,

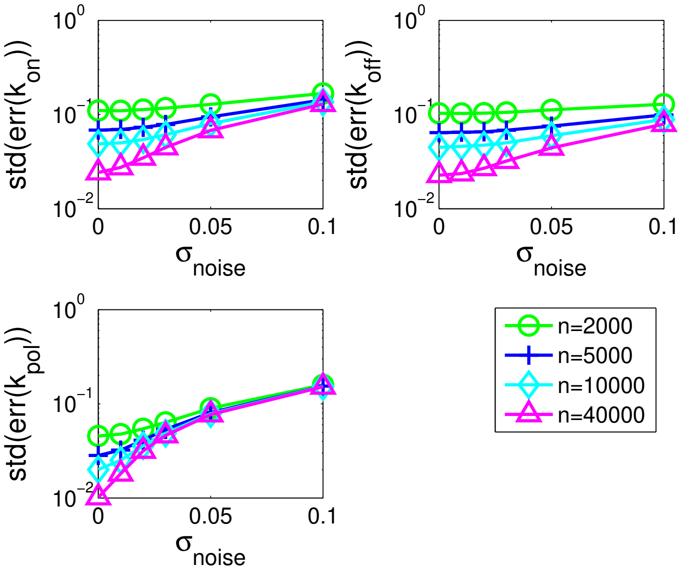

Figure 3.44: Plot of std (err (k)) vs σ from $T^{(1)}$ data for n samples of $T^{(1)}$.

and $c_{2,k_{pol}}\sigma^2$, respectively. The variance is stochastic and scales as $c_{1,k_{on}}\sigma^2/n$, $c_{1,k_{off}}\sigma^2/n$, and $c_{1,k_{pol}}\sigma^2/n$.

We can solve for the constants in equations 3.50-3.52 numerically. From equation 3.50, we have that $\langle z_{k_{on}}\rangle^2 = c_{2,k_{on}}^2\sigma^4$ and $var\left(z_{k_{on}}\right) = c_{1,k_{on}}\sigma^2/n$. The least-squares solution is given by

$$c_{1,k_{on}} = \sqrt{\frac{\sum_i var\left(z_{k_{on}}\right)\Big|_{\sigma=\sigma_i}}{\frac{\sigma_i^2}{n}}}, \tag{3.53}$$

$$c_{2,k_{on}} = \frac{\sum_i \langle z_{k_{on}}\rangle\Big|_{\sigma=\sigma_i}}{\sum_i \sigma_i^2}. \tag{3.54}$$

A least-squares solution for $c_{1,k_{off}}$, $c_{2,k_{off}}$, $c_{1,k_{pol}}$, and $c_{2,k_{pol}}$ can be derived in a similar manner.

To verify the validity of the least-squares fitting, we compare the mean and variance of $z_{k_{on}}$ and $z_{k_{off}}$ at $\sigma = 2^{-4}$ at $n = 32000$ with their predicted mean and variance as obtained through the least-squares fit above (table 3.1). Here, we

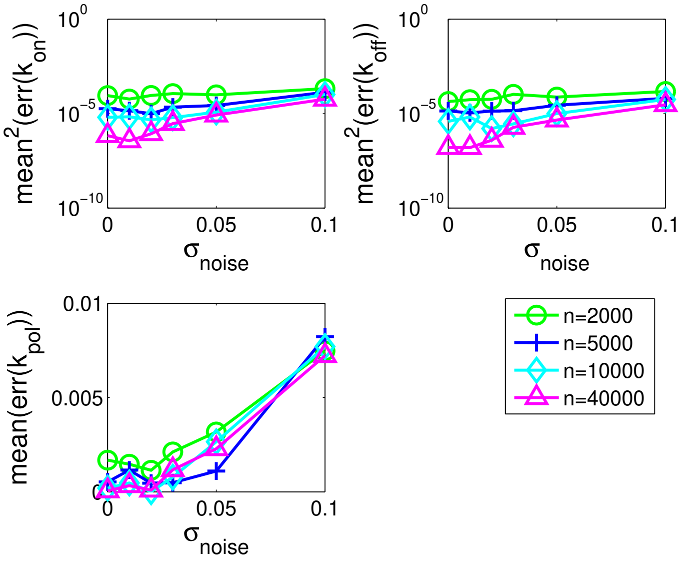

Figure 3.45: Plot of mean^2 (err (k)) vs σ from $T^{(1)}$ data for n samples of $T^{(1)}$.

obtained

- $c_{1,k_{on}} = 2.2652$
- $c_{1,k_{off}} = 1.5097$
- $c_{1,k_{pol}} = 0.87507$
- $c_{2,k_{on}} = -0.31581$
- $c_{2,k_{off}} = -0.34855$
- $c_{2,k_{pol}} = -0.23282$

Throughout the number of samples examined in our numerical simulations ($n = 1000, 2000, 4000, 8000, 16000, 32000, 64000$), we observe little change in the least-squares solutions for c_1 and c_2 for $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$. Table 3.2 shows the mean and standard deviation of the asymptotic coefficients c_1 and c_2 throughout the samples sizes.

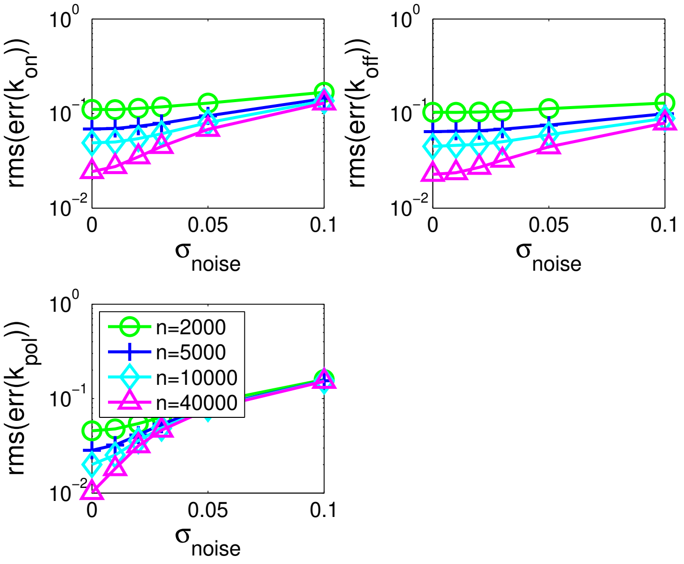

Figure 3.46: Plot of rms (err (k)) vs σ from $T^{(1)}$ data for n samples of $T^{(1)}$.

	$z_{k_{on}}$	$z_{k_{off}}$	$z_{k_{pol}}$
mean	-0.00036089	-0.00036063	-0.0002392
mean from fit	-0.00030841	-0.00034038	-0.00022737
var.	6.9199×10^{-7}	2.8203×10^{-7}	1.0156×10^{-7}
var. from fit	6.9127×10^{-8}	4.6073×10^{-8}	2.6705×10^{-8}

Table 3.1: Comparison between the observed mean and variance of $z_{k_{on}}$, $z_{k_{off}}$, $z_{k_{pol}}$ with their predicted means and variances obtained through the least-squares fit. Here, $\sigma = 2^{-4}$ and $n = 32000$. The results show good agreement between the observed and predicted mean and variances.

The importance of these results is that for any $r_2, k_{on}, k_{off}, k_{pol}$, and $[dNTP]$, we can collect n unperturbed $T^{(1)}$ samples and perturb them m times to obtain m data sets. From this data, the asymptotic coefficients c_1 and c_2 in equations 3.50-3.52 can be obtained by least-squares fitting, and an accurate description of the distribution of $z_{k_{on}}, z_{k_{off}}$, and $z_{k_{pol}}$ can be obtained.

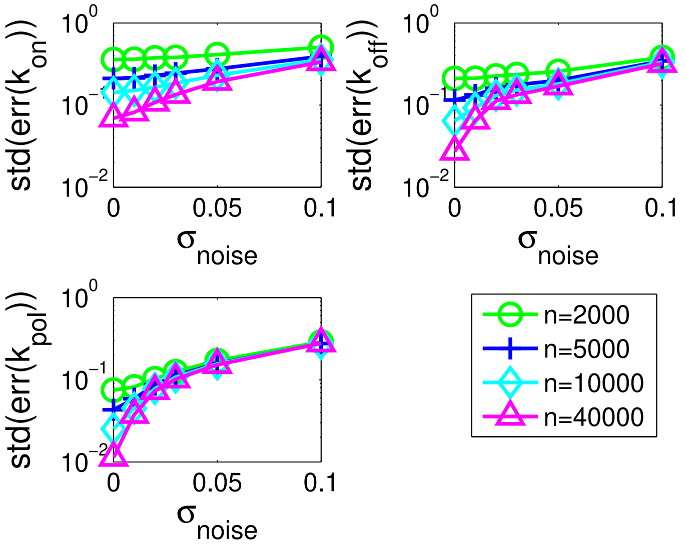

Figure 3.47: Plot of std (err (k)) vs σ from T_B data for n samples of $T^{(1)}$.

	$c_{1,k_{on}}$	$c_{1,k_{off}}$	$c_{1,k_{pol}}$	$c_{2,k_{on}}$	$c_{2,k_{off}}$	$c_{2,k_{pol}}$
mean	2.2523	1.5156	0.8779	-0.3149	-0.3481	-0.2323
std	0.0154	0.0075	0.0038	0.0012	8.0164×10^{-4}	8.1392×10^{-4}

Table 3.2: Mean and standard deviations of the asymptotic constants c_1 and c_2 for $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ throughout the different samples sizes $n = 1000, 2000, 4000, 8000, 16000, 32000, 64000$.

3.9 Further Improvements on Inference from $T^{(1)}$ Data

In subsection 3.3.1, we proposed a method for inferring k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ data by first using the EM algorithm to infer the mixture parameters α, λ_1, and λ_2 from the $T^{(1)}$ data and then mapping the inferred mixture parameters to k_{on}, k_{off}, and k_{pol} using equations A.1, A.2, and A.3. In this mapping, the probability of escape to the pre-translocation state, $p_{E_{pre}|2}$ was not used. In this section, we investigate the advantages of using the knowledge of $p_{E_{pre}|2}$. Intuitively, using more available information in the inference will reduce the total relative

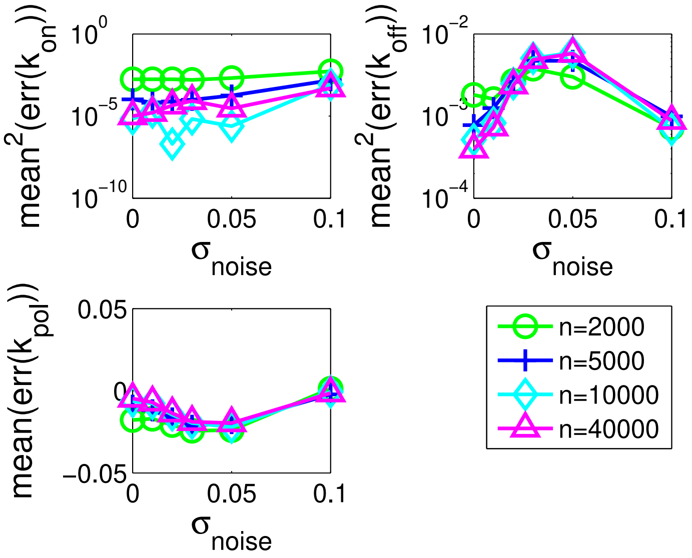

Figure 3.48: Plot of mean^2 (err (k)) vs σ from T_B data for n samples of $T^{(1)}$.

error.

The probability of escape back to the pre-translocation state, $p_{E_{pre}}$, can be calculated directly from observations in the nanopore experiment. We can approximate $p_{E_{pre}}$ from the data by,

$$p_{E_{pre}|2} \approx \frac{n_{T^{(1)}}}{n_{T^{(1)}} + n_{T_B}},$$

where $n_{T^{(1)}}$ and n_{T_B} are the total number of $T^{(1)}$ and T_B samples observed.

Recall figure 3.3 which shows the state-space diagram which pertains to $T^{(1)}$. Let

$$\hat{r}_2 := \frac{r_2}{p_{E_{pre}|2}},$$
$$\hat{k}_{on} := \frac{k_{on} p_{E_{pre}|4}}{p_{E_{pre}|2}},$$
$$\hat{k}_{off} := \frac{k_{off} p_{E_{pre}|2}}{p_{E_{pre}|4}}.$$

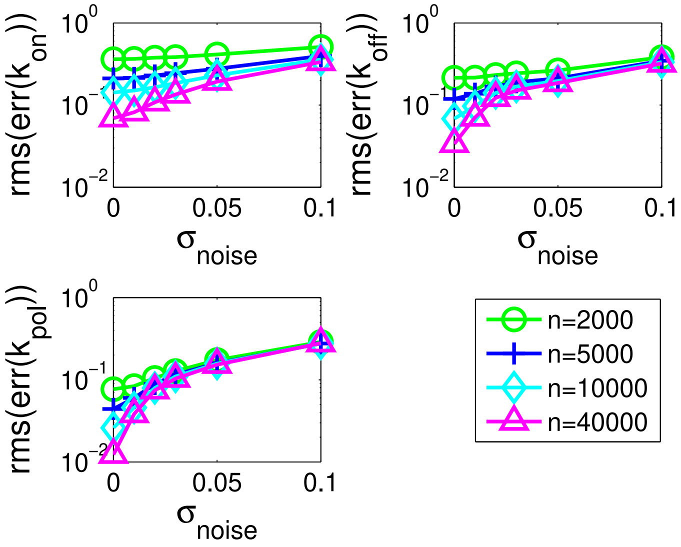

Figure 3.49: Plot of rms (err (k)) vs σ from T_B data for n samples of $T^{(1)}$.

We can calculate $p_{E_{pre}|2}$ as described above and r_2 can be calculated by setting $[dNTP] = 0$ and fitting a single exponential mode to the lower-amplitude data. Hence, we assume that $p_{E_{pre}|2}$ and r_2 is known; i.e., $\hat{r}_2$ is known.

Notice that the conditioned system is in the exact same form as the $k_{pol} = 0$ case in Chapter 2 with effective dNTP binding and disassociation rates $\hat{k}_{on}$ and $\hat{k}_{off}$, respectively. We can thus infer $\hat{k}_{on}$ and $\hat{k}_{off}$ in the exact same manner as k_{on} and k_{off} in the $k_{pol} = 0$ case in Chapter 2. In order to map $\hat{r}_2$, $\hat{k}_{on}$, and $\hat{k}_{off}$ to k_{on}, k_{off}, and k_{pol}, we solve the following system of equations

$$\begin{aligned}
p_{E_{pre}} &= \frac{r_2\left(k_{off} + k_{pol}\right)}{\left(k_{off} + k_{pol}\right) r_2 + k_{pol} k_{on} [dNTP]}, \\
\hat{k}_{on} &= \frac{k_{on} p_{E_{pre}|4}}{p_{E_{pre}|2}} = \frac{k_{on} k_{off}}{k_{off} + k_{pol}}, \\
\hat{k}_{off} &= \frac{k_{off} p_{E_{pre}|2}}{p_{E_{pre}|4}} = k_{off} + k_{pol}.
\end{aligned}$$

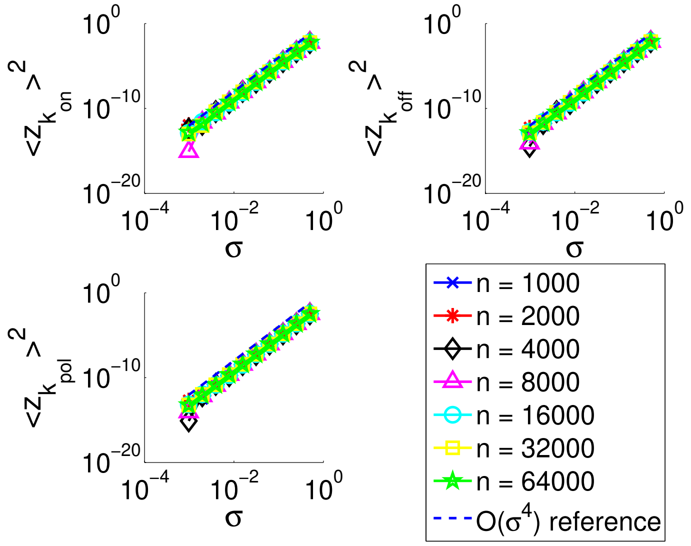

Figure 3.50: The squared-mean of the quantities $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$. This shows that the squared-mean of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ both follow $O\left(\sigma^4\right)$ as $\sigma \to 0$ and hence $\langle z_{k_{on}} \rangle$, $\left\langle z_{k_{off}} \right\rangle$, and $\left\langle z_{k_{pol}} \right\rangle = O\left(\sigma^2\right)$.

Indeed, the solution to this system of equations is given by

$$k_{pol} = \frac{\left(1 - p_{E_{pre}|2}\right) r_2 \hat{k}_{off}^2}{\left(1 - p_{E_{pre}|2}\right) r_2 \hat{k}_{off} + p_{E_{pre}|2} \hat{k}_{on} \hat{k}_{off} [dNTP]}, \tag{3.55}$$

$$k_{off} = \hat{k}_{off} - k_{pol}, \tag{3.56}$$

$$k_{on} = \frac{\hat{k}_{on} \hat{k}_{off}}{k_{off}}. \tag{3.57}$$

To recap, we first infer the mixture parameters α, λ_1, and λ_2 from $T^{(1)}$ data. The mixture parameters are then mapped to $\hat{k}_{on}$ and $\hat{k}_{off}$ in the exact same manner as in Chapter 2. Then the inferred k_{on}, k_{off}, and k_{pol} are given by equations 3.55-3.57. Equation 3.58 summarizes these series of mappings.

$$\left\{T_1^{(1)}, \ldots, T_n^{(1)}\right\} \xrightarrow{\text{MLE}} (\alpha, \lambda_1, \lambda_2) \xrightarrow{K} \left(\hat{k}_{on}, \hat{k}_{off}\right) \xrightarrow{G} (k_{on}, k_{off}, k_{pol}), \tag{3.58}$$

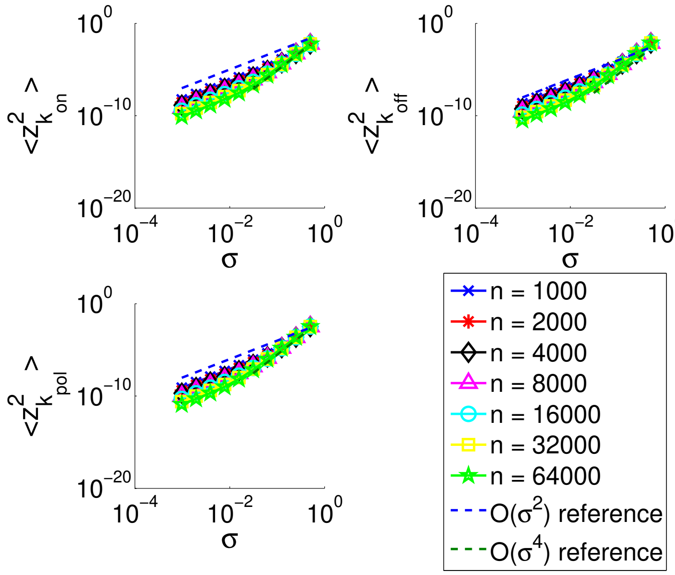

Figure 3.51: The second moment of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$. This shows that the second moments follow $O\left(\sigma^2\right) + O\left(\sigma^4\right)$ for fixed n.

where K and G are the mappings $(\alpha, \lambda_1, \lambda_2) \mapsto \left(\hat{k}_{on}, \hat{k}_{off}\right)$ and $\left(\hat{k}_{on}, \hat{k}_{off}\right) \mapsto (k_{on}, k_{off}, k_{pol})$ given in Chapter 2 and equations 3.55-3.57, respectively.

For notational convenience, let $\theta = (\alpha, \lambda_1, \lambda_2)$. Consider the first-order Taylor expansion,

$$G\left(K\left(\theta\right)\right) = G\left(K\left(\theta^{\text{MLE}}\right)\right) + J_{G\circ K}\left(\theta^{\text{MLE}}\right)\left(\theta - \theta^{\text{MLE}}\right) + o\left(\left\|\theta - \theta^{\text{MLE}}\right\|\right),$$

where $J_{G\circ K}\left(\theta^{\text{MLE}}\right)$ is the Jacobian of $G\circ K$ evaluated at θ^{MLE}. Now the covariance of the MLE estimates of k_{on}, k_{off}, and k_{pol} are given by,

$$Cov\left(K\left(\theta\right)\right) = J_{G\circ K}\left(\theta^{\text{MLE}}\right) Cov\left(\theta\right) J_{G\circ K}\left(\theta^{\text{MLE}}\right)^T,$$

where $Cov\left(\theta\right)$ is the covariance matrix of θ, which is approximated by H^{-1}, the inverse of the observed information matrix in equation 3.33. The total relative error can then be calculated using equation 3.34 in the same manner as before.

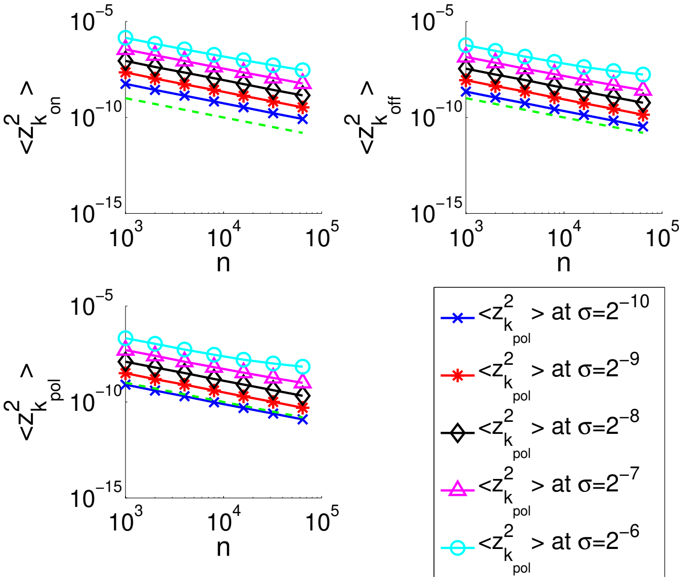

Figure 3.52: The second moment of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ a function of n. This shows that the second moments follow $O\left(1/n\right)$ for fixed σ.

Qualitatively, the total relative error when using knowledge of the escape probability $p_{E_{pre|2}}$ looks similar to when we did not use the escape probability (figure 3.56). However, closer inspection shows that this is not the case. Figure 3.57 shows the total relative error along the trajectory with $k = 1$, $k_p = 0.5$ and $[dNTP]$ ranging from 10^{-2} to 10^{3}. Using knowledge of $p_{E_{pre}}$ results in nearly a order of magnitude decrease in total relative error in small $[dNTP]$ regimes. However, this advantage is negligible in saturating $[dNTP]$ regimes where the total relative error is already at its lowest. Including knowledge of the escape probably will thus play a more imporant role in systems in which the optimal constrained $[dNTP]$ is small, otherwise the advantages of incorporating $p_{E_{pre|2}}$ are negligible.

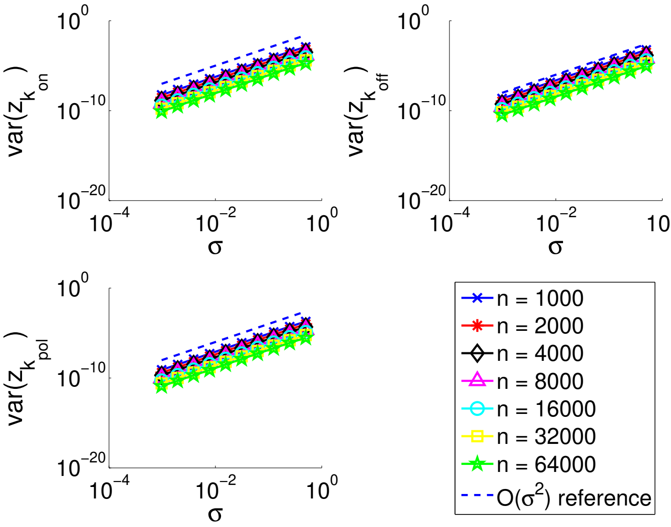

Figure 3.53: The variance of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ as a function of σ. For fixed n, the variance of $z_{k_{on}}, z_{k_{off}}$, and $z_{k_{pol}}$ behave as $O\left(\sigma^2\right)$.

3.10 Discussion and Concluding Remarks

The methods used here to derive the PDFs of $T^{(1)}$ and T_B can be applied to deriving dwell times of other Markovian phenomena in which dwell time data can be gathered. We showed that the PDF of $T^{(1)}$ is of the form of a proper mixture distribution and hence the mixture parameters can be accurately obtained from observed $T^{(1)}$ data through the EM algorithm.

We demonstrated through numerical simulations that the EM approach for inferring k_{on}, k_{off}, and k_{pol} from $T^{(1)}$ data is robust against measurement noise, and that the inference uncertainty decreases as $[dNTP]$ increases. For low measurement noise, the relative error is less than 20% and 17% respectively for k_{on} and k_{off} for low $[dNTP]$ and decreasing to under 5% for both k_{on} and k_{off} for high $[dNTP]$. Under low measurment noise, the inference uncertainty for k_{pol} is below 8% even for low $[dNTP]$. For high measurement noise, low $[dNTP]$ produces very high relative error. For $[dNTP] \geq 2$, the inference error is relatively

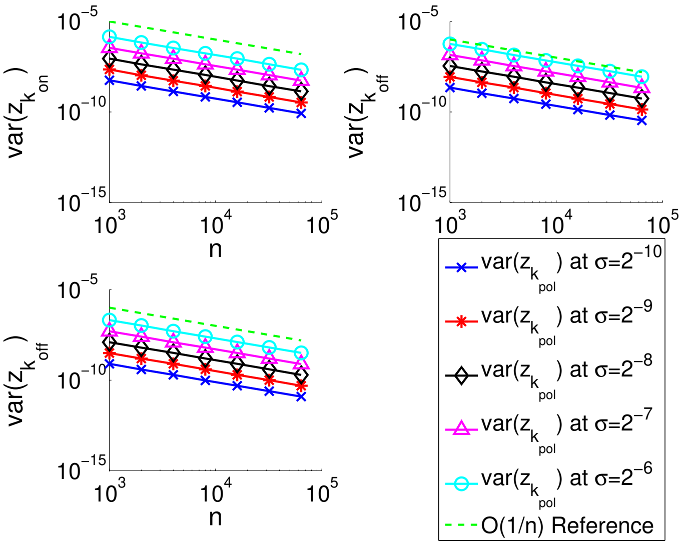

Figure 3.54: The variance of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ as a function of n. For fixed σ, the variance of $z_{k_{on}}, z_{k_{off}}$, and $z_{k_{pol}}$ behave as $O\left(1/n\right)$.

low: less than 14%, 9%, and 15% for k_{on}, k_{off}, and k_{pol}, respectively.

When using $T^{(1)}$ data, the measurement noise affects the inference of k_{on} and k_{off} the least, but its effects are relatively large for the inference of k_{pol}.

We also showed that the PDF of T_B is an improper mixture of four exponential modes. Because some of the mixture weights are negative, we lose the hierarchical structure of a proper mixture distribution and hence cannot use the EM method to infer k_{on}, k_{off}, and k_{pol}, unlike when using the $T^{(1)}$ data. We thus use the Nelder-Mead algorithm to maximize the log-likelihood function of the T_B data.

For low measurement noise, inferring k_{on} from T_B data produces relative errors of less than 18% when using $[dNTP] \leq 4$. The inference uncertainty of k_{on} grows to unacceptable levels for very high $[dNTP]$. The relative error for k_{off} is lowest when $[dNTP] = 1, 2$ (less than 14% for low noise), but the uncertainty becomes unacceptable for very high and low $[dNTP]$. The inference uncertainty for k_{pol} behaves similar to that of k_{pol} from the $T^{(1)}$ data, generally decreasing as $[dNTP]$ increases. For high measurement noise, $[dNTP] = 0.5$ produces the

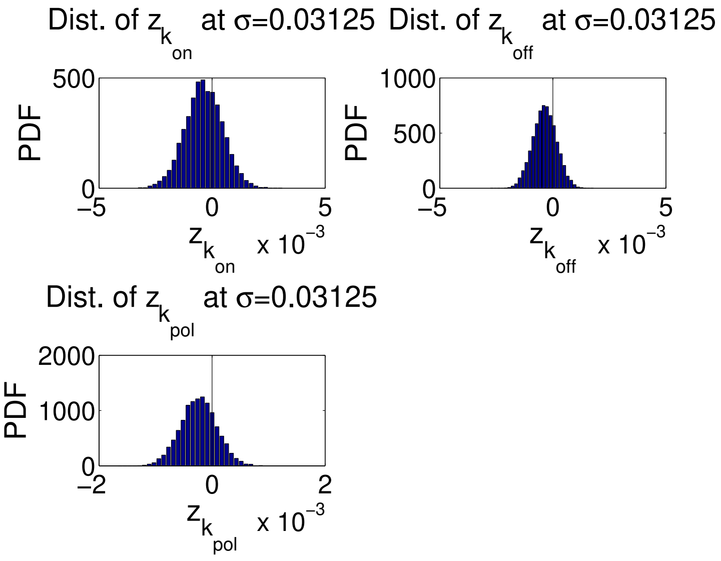

Figure 3.55: The distribution of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ at $\sigma = 2^{-4}$ with $n = 8000$. Here, all the quantities are normally distributed.

least relative error for k_{on}; $[dNTP] = 2$ produces the least relative error for k_{off}; and $[dNTP] = 16$ produces the least relative error for k_{pol}. However, for the high measurement noise, using the T_B data produces near unacceptable levels of uncertainty even at optimal $[dNTP]$.

We also examined how the relative errors behaved as a function of measurement noise when changing the number of samples of $T^{(1)}$ and T_B while keeping the $[dNTP]$ fixed. We found that around 5000 samples of $T^{(1)}$ are sufficient for obtaining reasonable estimates for k_{on}, k_{off}, and k_{pol}.

The maximum reaction velocity obtained from $\langle T_{cycle} \rangle$ provides an alternative method for calculating k_{pol}. Numerical simulations conclude that we can estimate k_{pol} to about 7% RMS error at saturating $[dNTP]$. This can provide an alternate means of inferring k_{pol}. The dNTP concentration which produces the half-maximum reaction velocity (parameter K_m) can be used to put a constraint on k_{on} and k_{off}. We found that the constrained maximization of the joint

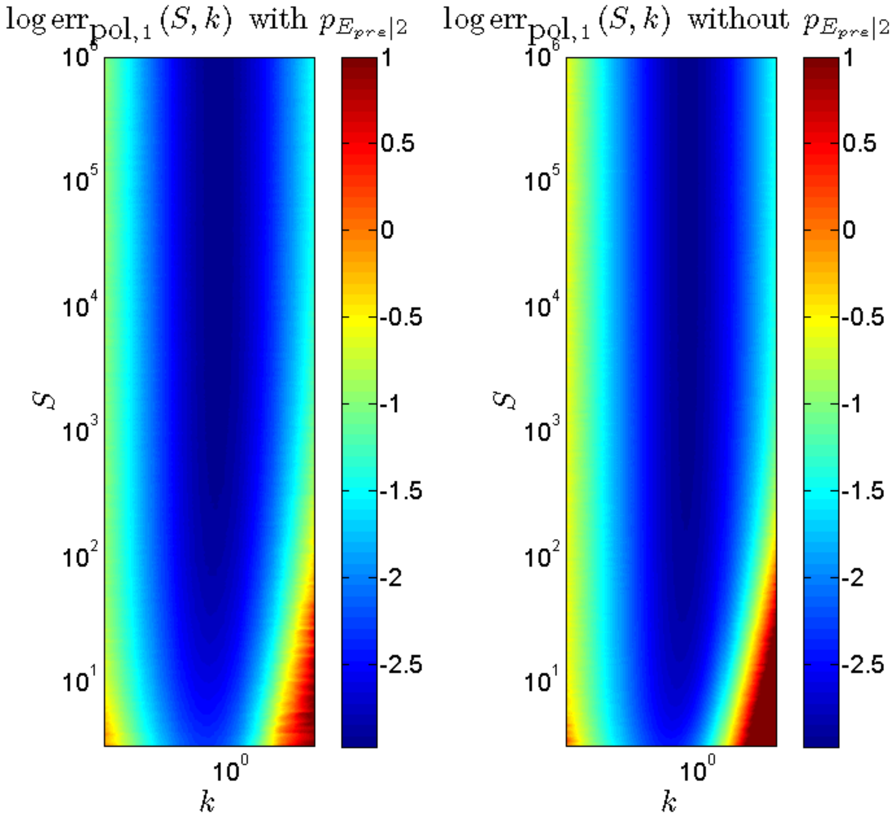

Figure 3.56: Comparison between the total relative errors calculated when knowledge of the escape probability $p_{E_{pre}|2}$ is used and when $p_{E_{pre}|2}$ is not used in mapping the mixture parameters to the kinetic rates k_{on}, k_{off}, and k_{pol}. The results are qualitatively similar.

likelihood function of $T^{(1)}$ and T_B formed by adding their respective likelihoods provides very satisfactory inference results and improves upon the uncertainty of inferring k_{on} and k_{off} when compared to the general unconstrained inference of k_{on}, k_{off}, and k_{pol} simultaneously.

We investigated the inference uncertainty of k_{on}, k_{off}, and k_{pol} estimated from the $T^{(1)}$ samples in terms of the total relative error. After conditioning on the escape to the pre-translocation state, the escape problem governing $T^{(1)}$ is of the same form as the $k_{pol} = 0$ case in Chapter 2 where r_2, k_{off}, and k_{pol} are scaled by the conditional probabilities $p_{E_{pre}|2}$ and $p_{E_{pre}|4}$. The scaled r_2 rate,

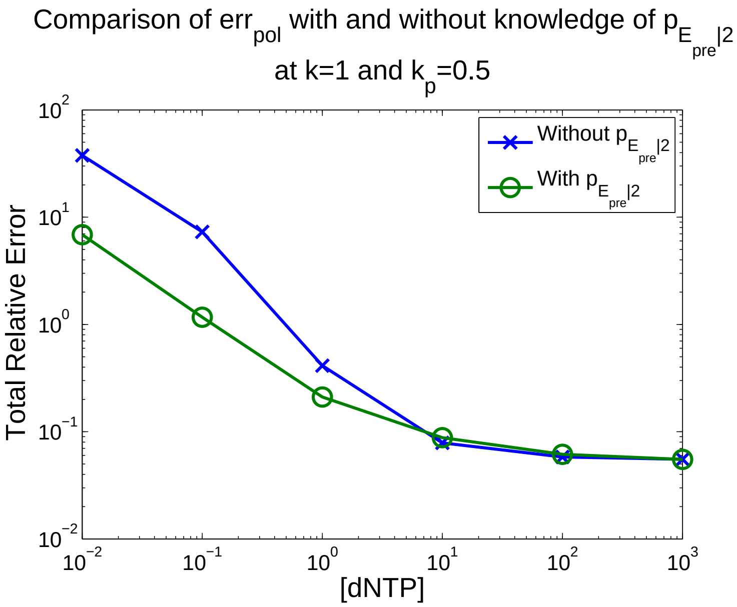

Figure 3.57: Comparison of the total relative error with and without knowledge of the escape probability $p_{E_{pre}|2}$ along the trajectory $k = 1$, $k_p = 0.5$ and $[dNTP]$ ranging from 10^{-2} to 10^{3}. Using knowledge of $p_{E_{pre}}$ results in nearly a order of magnitude decrease in total relative error in small $[dNTP]$ regimes. However, this advantage is negligible in saturating $[dNTP]$ regimes where the total relative error is already at its lowest.

$\hat{r}_2$ can be scaled to 1. For fixed $k_p = k_{pol}/r_2$, the total relative error of k_{on}, k_{off}, and k_{pol} can be shown to only depend on S and k–the scaled $[dNTP]$ and k_{off}, respectively. We used the observed Fisher information matrix to obtain an asymptotic estimate for the covariance matrix for the MLE estimates and then the uncertainty was propagated to the k_{on}, k_{off}, and k_{pol} estimates through a first-order Taylor expansion. This and the aforementioned scaling laws allowed us to build a numerical approximation to the total relative error for any k_{on}, k_{off}, and k_{pol}. The use of the observed Fisher information allowed us to build a database of the total relative error for any dNTP binding, disassociation, and incorporation rate without the use of full-scale Monte Carlo simulation.

There is no well defined optimum $[dNTP]$ which produces the least total relative error of k_{on}, k_{off}, and k_{pol} from the $T^{(1)}$ observations unlike the $k_{pol} = 0$

case in Chapter 2. The minimum total relative error occurs in saturating $[dNTP]$ regimes and the total relative error approaches a constant in this region. Although the total relative error is lowest in saturating $[dNTP]$ regions, collecting $T^{(1)}$ samples here is impractical due to the high probability of immediate escape into the pol-process.

We then examined how to obtain the optimal $[dNTP]$ under experimental time constraints; that is, the total run-time of the experiment is constrained using the mean-field approximation of the experimental run-time. This constrained optimization problem can be recast into an unconstrained optimization problem of $[dNTP]$ only and the $[dNTP]$ which produces the least total relative error in this recast problem is the optimal constrained $[dNTP]$. Under this setting, the optimal $[dNTP]$ occurs well below saturating $[dNTP]$ regions. We also examined the optimal $[dNTP]$ when constraining the number of cycles using the mean-field approximation as well. Like constraining the experimental time, the constrained optimization is recast into an unconstrained problem. The optimal $[dNTP]$ in this setting occurs well below saturating $[dNTP]$ regions. In both cases, the mean-field approximation is justified numerically.

The construction of the total relative error function and characterization of the optimal $[dNTP]$ thus provide a way to determine the experimental parameters which produce the least inference uncertainty when inferring dNTP binding, disassociation, and incorporation rates. This a priori knowledge will allow researchers to make more accurate estimates for the dNTP binding, disassociation, and incorporation rates and further elucidate the dynamics of dNTP binding and incorporation in DNAP-DNA complexes.

Characterization of the MLE estimates from perturbed $T^{(1)}$ samples with multiplicative noise was also investigated. Using numerical simulations, we obtained strong numerical evidence to support the claims that the MLE estimates of k_{on}, k_{off}, and k_{pol} from perturbed $T^{(1)}$ data differ from the MLE estimates of k_{on}, k_{off}, and k_{pol} from unperturbed $T^{(1)}$ data by a Gaussian with mean $O\left(\sigma^2\right)$ and variance $O\left(\sigma^2/n\right)$, where σ is the standard deviation of noise. Furthermore, the distribution of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ can be accurately described by least-squares fitting of the asymptotic coefficients to the squared-mean and variance of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$. The asymptotic coefficients are shown to have a weak dependence on n. This and numerical simulations examining the distribution of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ show that the distribution of $z_{k_{on}}$, $z_{k_{off}}$, $z_{k_{pol}}$ can be accurately obtained for any system in the following way: (1) generate n unperturbed $T^{(1)}$ samples and perturb them m times to create m data sets; and (2) determine the asymptotic coefficients of the squared-mean and variance of $z_{k_{on}}$, $z_{k_{off}}$, and $z_{k_{pol}}$ by least-squares fitting.

Finally, we looked at how further improvements can be made when inferring the kinetic rates from $T^{(1)}$ data by including the escape probability, $p_{E_{pre}|2}$. Gains in terms of reducing the total relative error by including this information in the mapping from the mixture parameters to the kinetic rates are negligible, but the reduction in total relative error when including the escape probability in the mapping is nearly an order of magnitude when $[dNTP]$ is small. This can be useful in systems in which the constrained optimal $[dNTP]$ occurs when $[dNTP]$ is small.

Chapter 4

Kinetic Structure of the DNAP Polymerization Process

4.1 Introduction

In this chapter, we focus on inferring the internal kinetic structure of the DNAP polymerization process. Previous work has been done to estimate the dNTP binding and disassociation rates in non-synthesizing DNAP-DNA complexes by use of a autocorrelation function of the entire current amplitude measured from nanopore experiments [50]. We also proposed a method of estimating the dNTP binding and disassociation rates in non-synthesizing complexes by deriving the underlying dwell time PDF and applying an expectation-maximization (EM) algorithm to obtain the MLE estimates (chapter 2). In the previous chapter, we extended this result to synthesizing DNAP-DNA complexes, inferring the kinetic rates of dNTP binding, disassociation, and incorporation (chapter 3).

However, in chapter 3, dNTP incorporation was modeled as a single kinetic step called k_{pol}. The actual number of steps of the polymerization process is unknown, but it contains at least activation, chemistry, and pyrophosphate release. In this chapter, we extend further the model for synthesizing DNAP-DNA complexes. We model the polymerization process as an arbitrary number of kinetic states and show how both the number of states and the kinetic rates of the polymerization process can be inferred from dwell time data.

For an ionic current trace covering more than one nucleotide addition cycle, we define various dwell times (figure 4.1):

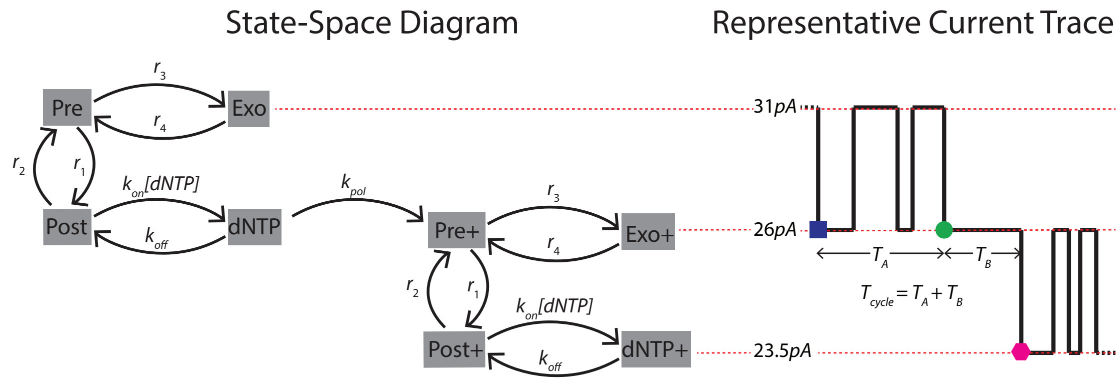

Figure 4.1: A state-space diagram for two nucleotide addition cycles in DNA replication. When the DNAP-DNA complex is allowed to undergo synthesis and a complementary dNTP is provided in the cis chamber, the DNAP-DNA complex can transition to the next nucleotide addition cycle–indicated by the "+" symbol after the state names. This is manifested as a change in the upper and lower amplitudes as the reporter group gets closer or further away from the nanopore lumen.

- T_A: the time from the first arrival to the post-translocation state of the current nucleotide addition cycle to the last arrival to the post-translocation state of the current nucleotide addition cycle; this is shown graphically as the blue square to the green circle in figure 4.1

- T_B: the time from the last arrival to the post-translocation state of the current nucleotide addition cycle to the first arrival to the post-translocation state of the next nucleotide addition cycle; this is shown graphically as the green circle to the magenta hexagon in figure 4.1

- $T^{(1)}$: the lower-amplitude dwell times within the T_A dwell time segment. In any observation of T_A, there are likely to be many samples of $T^{(1)}$ and we label them as $T_1^{(1)}, T_2^{(1)}, T_3^{(1)}, \ldots$, etc (figure 4.1)

- $T^{(2)}$: the upper-amplitude dwell times within the T_A and T_B dwell time segments. Like $T^{(1)}$, there are likely to be many samples of $T^{(2)}$, so we label them as $T_1^{(2)}, T_2^{(2)}, \ldots$, etc (figure 4.1). Note that the dwell time $T^{(2)}$ is not directly observable within the dwell time segment T_B. Within the T_B segment, this is denoted graphically as the left-opened cyan parenthesis to the right-opened cyan paranthesis in figure 4.3.

- T_{pol}: the time from the last arrival to the post-translocation state to the first arrival to the pre-translocation state in the next nucleotide addition cycle; this is the time that the DNAP-DNA complex completes the dNTP binding and incorporation steps. This is denoted by the green circle to the right-opened red parenthesis (figure 4.3).

The transition rates $r_1, r_2, r_3, r_4, k_{on}, k_{off}$, and k_{pol} shown in figures 4.2, and the rest of the state-space diagrams shown in this paper are defined as follows. Each transition rate is written next to an arrow originating from state i and ending at state j. That transition rate is the rate of which the DNAP-DNA complex transitions from state i to state j. For example, r_1 is the rate of which the DNAP-DNA complex transitions from the pre-translocation state to the post-translocation state. Mathematically, we can write

$$r_1 = \lim_{\Delta t \to 0^+} \frac{Pr\left(S\left(t+\Delta t\right) = \text{Post} \mid X\left(t\right) = \text{Pre}\right)}{\Delta t},$$

where $X(t)$ denotes the state of the Markov chain at time t. The other transition rates are defined in a similar manner.

In this paper, we are interested in the case in which the DNAP-DNA complex is allowed to undergo synthesis. The DNAP-DNA complex will thus transition in discrete amplitude levels, each level corresponding to a nucleotide addition cycle. A mutation has been engineered into the exonuclease so that cleaving of the dNTP cannot occur, and hence the transition to the next nucleotide addition cycle is irreversible. In the previous chapter, we inferred k_{on}, k_{off}, and k_{pol} from the lower-amplitude, $T^{(1)}$ data.

The dNTP incorporation rate k_{pol} modeled the polymerization (pol) process as a single rate-limiting step. The polymerization process is actually multiple internal kinetic steps which includes at least activation, chemistry, and pyrophosphate release. In this paper, we model the pol process as an arbitrary number of kinetic steps with the last step irreversible and introduce ways to infer the number of internal states and their kinetic rates in the pol process from dwell time data available from the nanopore experiments (figure 4.2).

The dwell time T_B incorporates information about the pol process. Recall that T_B is the time for the DNAP-DNA complex to escape to the post-translocation state of the next nucleotide addition cycle (post+) when starting at the post-translocation state of the current nucleotide addition cycle (post). We can write $T_B = T_{pol} + T^{(2)}$ where T_{pol} is the time it takes the DNAP-DNA complex to complete the binding and incorporation segment of the nucleotide addition cycle and includes the pol process; and $T^{(2)}$ is the upper-amplitude segment of the next nucleotide addition cycle (figure 4.3).

In this paper, we will develop methods to infer the number of states in the DNAP polymerization process from the dwell time data. The methods all make

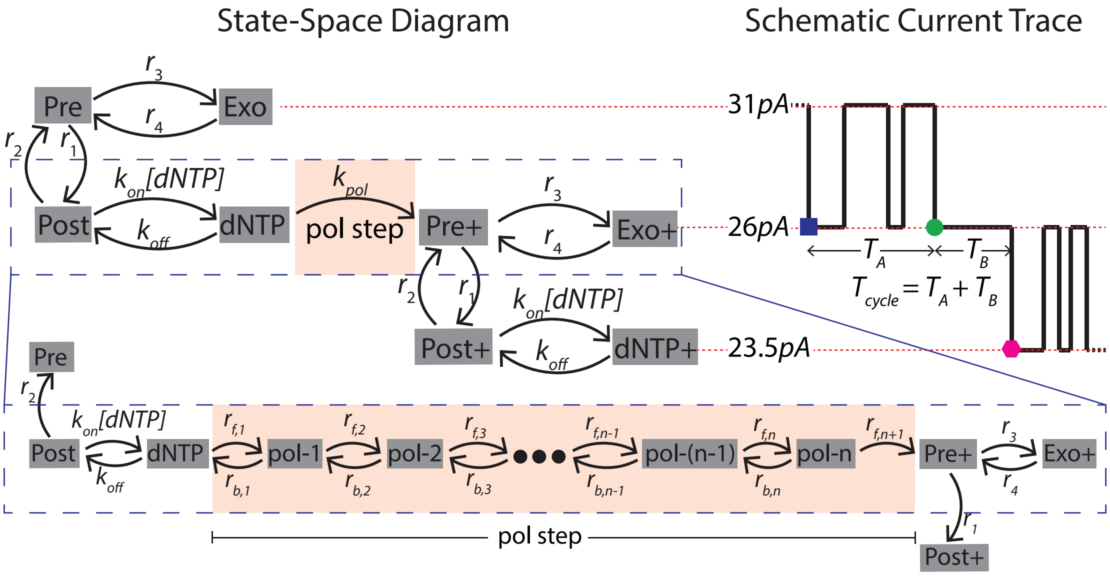

Figure 4.2: The state-space diagram of a DNAP-DNA complex allowed to undergo synthesis. The complex has been engineered to exonuclease activity cannot occur, and hence incorporation of a dNTP is irreversible. The pol process is modeled as an arbitrary number of internal kinetic steps with the last step being irreversible. The blue-box zoom emphasizes the T_B escape problem in which the pol process resides.

use of the randomness parameter of the dwell time data. We will show that under a restricted class of Markov processes, we can improve upon previous results in the literature that provide a bounds on the possible number of states in a continuous-time, discrete-state Markov process based on the randomness parameter of the observed dwell time data.

4.1.1 Introduction to the Randomness Parameter

In the context of molecular motors, the randomness parameter is defined to be $r = 2D/(vd)$ where D is the effective diffusion constant of the enzyme, v is the average rate of the enzyme, and d is the step size of the molecular motor. Due to fluctuations, if two identical motors are started at the same location at the same time, they will eventually separate with a squared distance that increases linearly

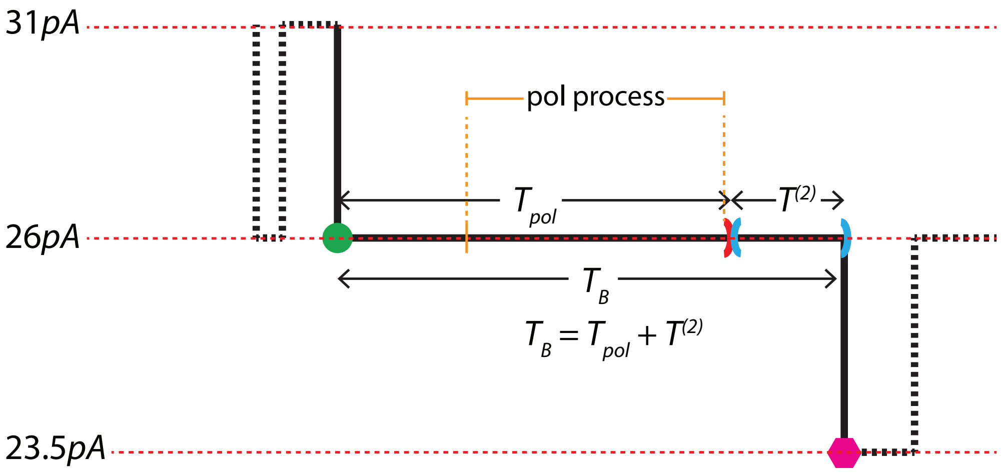

Figure 4.3: A representative current trace depicting the T_B dwell time along with T_{pol} and $T^{(2)}$ dwell time segments which make up the T_B dwell time. The T_{pol} dwell time segment is graphically denoted from the green circle to the right-opened red parenthesis, and the $T^{(2)}$ dwell time segment is graphically denoted from the left-opened cyan parenthesis to the right-open cyan parenthesis. The pol process occurs within the T_{pol} dwell time. Both dwell times T_{pol} and $T^{(2)}$ are not directly observable since they do not manifest a change in current amplitude. The latter is only unobservable when part of the T_B dwell time segment.

with time [56]. The quantity D is a measure of this diffusive behavior [67], [70]. In the limit that the motor takes a uniform step size and direction, the randomness parameter reduces to a function of only the first two moments of the dwell times

$$R = \frac{var\,(T)}{\langle T \rangle^2},$$

where T is the cycle time [67], [70]. In the broader context of stochastic modeling, this quantity is also known as the squared coefficient of variation.

The randomness parameter can be used to put a lower bound on the number of kinetic states in a system:

$$\frac{1}{R} \leq n, \tag{4.1}$$

where n is the number of states in the system. The inequality was first introduced a conjecture in the context of single-molecule experiments [67], [70]. It has been formally proven for any continuous-time, discrete-state Markov process by the use of martingales in the context of phase-type distributions [20].

The lower-bound inequality in equation 4.1 can be intuitively explained as follows. If we have a system with a single kinetic state, then the randomness parameter is 1 since the variance of the dwell time is the square of the mean for an exponential distribution. As more kinetic states are added, the mean of the dwell time increases more quickly than the variance. Thus the ratio of the squared mean to the variance increases.

An advantage of using the randomness parameter to study the kinetic structure of a system is that the quantity can be accurately measured when the moments of the cycle time are corrupted by noise [67], [70].

In certain situations however, the randomness parameter fails to follow the inequality in equation 4.1. This occurs when the motor step size is not uniform or when the kinetic pathway varies [14], [68], [73]. In this setting, correction terms must be added to the randomness parameter so that the inequality in equation 4.1 is valid. In our context, the DNAP-DNA complex translocation step size is assumed to be uniform, and we can observe the individual dwell time events from the nanopore experimental data. The dwell time moments can thus be calculated, and hence so can R.

4.1.2 Application to the Randomness Parameter to Molecular Motors

The application of the randomness parameter to molecular motors was first introduced in the context of single molecule experiments in [67], [70]. Here, a sequential enzymatic pathway of n irreversible reactions was considered (figure 4.4).

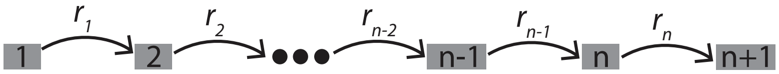

Figure 4.4: A sequential enzymatic pathway with irreversible transitions.

For this system the randomness parameter is given by

$$R = \frac{\sum_{i=1}^{n} r_i^{-2}}{\left(\sum_{i=1}^{n} r_i^{-1}\right)^2},$$

as calculated by Laplace Transforms in [67], [70]. Suppose that k of the r_i rates are comparable, and the $n-k$ other rates are much smaller. That is, $r_{i_1}, \ldots, r_{i_k} = O(\eta)$ and $r_{i_{k+1}}, \ldots, r_{i_n} = O(\epsilon)$ where $0 < \epsilon << \eta$ and $r_{i_j} \in \{1, \ldots, n\}$ with $r_{i_j} \neq r_{i_l}$ for $j \neq l$. In this case, $R \approx 1/k$. Hence $1/R$ gives an approximation to the number of rate-limiting steps in the reaction.

In [67], the sequential kinetic chain was generalized to include a reversible step between states 1 and 2 with the binding step first order in the substrate concentration (figure 4.5).

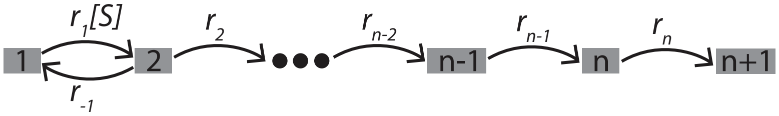

Figure 4.5: A sequential enzymatic pathway with irreversible transitions and the first transition reversible. The binding step first order in the substrate concentration, $[S]$.

The randomness parameter in this case is given by

$$R = \frac{\sum_{i=3}^{n} \frac{1}{r_i^2} + \frac{(r_1[S]+r_2+r_{-1})^2 - 2r_1r_2[S]}{(r_1r_2[S])^2}}{\left(\sum_{i=3}^{n} \frac{1}{r_i} + \frac{r_1[S]+r_2+r_{-1}}{r_1r_2[S]}\right)^2}$$

as calculated in [67]. It was shown in [67] that if all of forward rates $r_1, \ldots r_n$ are comparable, then $r \to 1/n$ as $[S] \to \infty$. This is because as $[S] \to \infty$, the unbinding rate r_{-1} is negligible compared to the binding rate. As $[S] \to 0$, $r \to 1$ since the substrate binding step becomes rate-limiting. Thus at saturating substrate concentrations, the number of states can be recovered for this system.

The randomness parameter has been studied in numerous papers over the years in the context of sequential systems [67], [70], [78] among others; or in the context of more complicated systems in which branches or parallel kinetic pathways are present [68], [55], [14] among others.

In this chapter, we extend the application of the randomness parameter to systems in which the escape problem governing the dwell time has two absorbing states. In particular, we apply this theory to the DNAP-DNA complex when synthesis is allowed (figure 4.2). We will develop methods using the randomness parameter to infer the number of kinetic states and kinetic rates of the polymerization process. In this situation, the DNAP-DNA complex has two choices after arrival at the post-translocation state: (i) the complex can transition back to the pre-translocation state, exiting the lower-amplitude; or (ii) the complex can bind and incorporate a complementary dNTP and proceed irreversibly through the entire pol-process, eventually arriving at the post-translocation state of the next nucleotide addition cycle. This application thus has two escape possibilities, or equivalently, two absorbing states. We will develop methods to infer the kinetic structure and details in this setting in which the individual kinetic states cannot be directly observed. Such is the case with the kinetic states in the pol-process; the states in the pol-process do not manifest a change in ionic current and hence cannot be directly observed.

4.2 Determining the Number of States and Transition Rates in a Birth-Death Process

Consider a birth-death process with two absorbing boundary states (figure 4.6). The birth-death process with two absorbing boundary states is of the same form as the escape problem describing the lower-amplitude up to and including covalent incorporation of the nucleotide. In this setting, the two absorbing states are the pre-translocation state and pre-translocation state of the next nucleotide addition cycle. This is apparent from figure 4.2; a Markov process describing the escape problem with transient states $\{\text{post}, \text{dNTP}, \text{pol-1}, \ldots, \text{pol-n}\}$ and absorbing states $\{\text{pre}, \text{pre+}\}$ is isomorphic to the birth-death process described in figure 4.6. Throughout the rest of this paper, let $X(0) = 1$ (or $X(0) = \text{post}$ if viewed in the setting of the DNAP-DNA complex), where $X(t)$ is the state of the Markov process at time t. Hence the birth-death process describes the escape from the lower-amplitude up to and including covalent incorporation of the nucleotide as mentioned before. In this section, we introduce methods to infer the number of states in the birth-death process and the forward and backwards (birth and death respectively) rates.

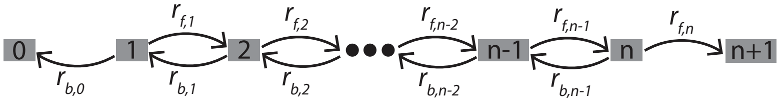

Figure 4.6: A birth death process with two absorbing boundary states.

4.2.1 Symmetric Birth-Death Process with Two Absorbing States

We first consider the simple case of a symmetric birth-death process. Suppose that the forward and backwards rates of the birth-death process in figure 4.6 are equal; that is, $r_{f,i} = r_{b_j} = r$ for all i and j. We look into methods for inferring the number of states n and the forward and backwards transition rate r from data.

Unconditional Escape Time

Let T be the time to absorption; that is the time that the process takes to transition into one of the absorbing states, 0 or $n+1$. For notational simplifity, let $h_k = \langle T \mid X(0) = k \rangle$. We can then write $h_k = r\Delta t h_{k-1} + r\Delta t h_{k+1} + (1 - 2r\Delta t)\, h_k + \Delta t + o(\Delta t)$. Dividing by Δt and passing the limit $\Delta t \to 0$, we obtain the nonhomogeneous difference equation

$$r h_{k-1} + r h_{k+1} - 2r h_K = -1,$$

with boundary conditions $h_0 = h_{n+1} = 0$. Solving this difference equation gives the solution

$$h_k = \frac{n+1}{2r} k - \frac{k^2}{2r}.$$

Hence, the quantity we are interested in is given by

$$h_1 = \frac{n+1}{2r} - \frac{1}{2r} = \frac{n}{2r}. \tag{4.2}$$

This gives the mean escape time when the birth-death process starts at state 1 at time $t = 0$.

To derive the second moment of T, we introduce some more notation. Let

$T_k = T \mid \{S(0) = k\}$ be the conditional random variable. Then we can write,

$$T_k = \Delta t + \begin{cases} T_{k-1} & \text{with probability } r\Delta t + o(\Delta t) \\ T_{k+1} & \text{with probability } r\Delta t + o(\Delta t) \\ T_k & \text{with probability } 1 - 2r\Delta t + o(\Delta t) \end{cases}$$

We can write

$$\begin{aligned} \left\langle T_k^2 \right\rangle = (\Delta t)^2 + 2\Delta t \left[r\Delta t \left\langle T_{k-1} \right\rangle + r\Delta t \left\langle T_{k+1} \right\rangle + (1 - 2r\Delta t) \left\langle T_k \right\rangle \right] \\ + r\Delta t \left\langle T_{k-1}^2 \right\rangle + r\Delta t \left\langle T_{k+1}^2 \right\rangle + (1 - 2r\Delta t) \left\langle T_k^2 \right\rangle + o(\Delta t). \end{aligned}$$

For ease of notation, let $u_k = \langle T_k^2 \rangle$ and recall that $h_k = \langle T_k \rangle$ and that

$$h_k = r\Delta t h_{k-1} + r\Delta t h_{k+1} + (1 - 2r\Delta t) h_k + \Delta t + o(\Delta t),$$

so that

$$h_k - \Delta t = r\Delta t h_{k-1} + r\Delta t h_{k+1} + (1 - 2r\Delta t) h_k + o(\Delta t).$$

Hence

$$u_k = 2\Delta t (h_k - \Delta t) + r\Delta t u_{k-1} + r\Delta t u_{k+1} + (1 - 2r\Delta t) u_k + o(\Delta t).$$

Dividing by Δt and passing the limit $\Delta t \to 0$, we obtain the inhomogeneous difference equation

$$\begin{aligned} u_{k-1} + u_{k+1} - 2u_k &= -\frac{2}{r} h_k \\ &= -\frac{2}{r} \left(\frac{n+1}{2r} k - \frac{k^2}{2r} \right). \end{aligned}$$

Solving this difference equation, we obtain

$$u_k = \left(\frac{n+1}{12r^2} + \frac{(n+1)^3}{12r^2} \right) k - \frac{k^2}{12r^2} - \frac{n+1}{6r^2} k^3 + \frac{k^4}{12r^2}.$$

Plugging in $k = 1$ gives us the desired quantity

$$u_1 = \frac{n\,(n+1)\,(n+2)}{12r^2}. \tag{4.3}$$

From the derivations of the first two moments above (equations 4.2 and 4.3), the randomness parameter of T is given by

$$R_T = \frac{n^2+2}{3n}. \tag{4.4}$$

We consider the quantity $3R_T$. Notice that for $n \geq 3$, we have $n < 3R_T \leq n + 2/3$ with equality to $n + 2/3$ if and only if $n = 3$. Hence we can infer the number of states in the symmetric birth-death process by computing $n = \lfloor 3R_T \rfloor$ when $n \geq 3$. However, if $n = 1$ or $n = 2$, we have $3R_T = 3$. The number of states cannot be recovered from just R_T in this case.

Advantages of Conditioning on the Location of Escape

In order to fully recover the number of states, we need more information. Consider the probability of escaping to the forward absorbing boundary; that is, let p_{f_k} be the probability of escaping to the $n+1$ state starting from state k. Like T, we can derive a linear, homogeneous difference equation in p_{f_k}.

We have the following probabilities

$$Pr\left(S\left(\Delta t\right)=k+1 \mid \left(0\right)=k\right)=r\Delta t+o\left(\Delta t\right)$$
$$Pr\left(S\left(\Delta t\right)=k-1 \mid \left(0\right)=k\right)=r\Delta t+o\left(\Delta t\right)$$
$$Pr\left(S\left(\Delta t\right)=k \mid \left(0\right)=k\right)=1-2r\Delta t+o\left(\Delta t\right)$$

Hence we can write $p_{f_k}=r\Delta t p_{f_{k-1}}+r\Delta t p_{f_{k+1}}+\left(1-2r\Delta t\right)p_{f_k}+o\left(\Delta t\right)$. Now dividing by Δt and r and passing the limit $\Delta t\to 0$, we obtain, the linear, homogeneous difference equation $p_{f_{k-1}}+p_{f_{k+1}}-2p_{f_k}=0$. Solving this difference equation with boundary conditions $p_{f_0}=0$ and $p_{f_{n+1}}=1$, we obtain $p_{f_k}=k/\left(n+1\right)$. Hence the desired quantity is $p_{f_1}=1/\left(n+1\right)$, which gives us the probability of absorption to $n+1$ given that $S\left(0\right)=1$. From here, we have the obvious corollary that $n=1/p_{f_1}-1$.

We see that conditioning on the escape to either of the absorbing boundary states yields more information content in regards to the number of states then the unconditional escape time T. From the forward escape probability alone, the number of states n can be recovered for a symmetric birth-death process.

To infer the forward and backward transition rate r, recall that $\langle T \mid S\left(0\right)=1\rangle = n/\left(2r\right)$ from equation 4.2. Hence after inferring n from the probability of forward escape, p_{f_1}, we can infer r from $r=n/\left(2\langle T \mid S\left(0\right)=1\rangle\right)$.

A Look into the Conditional Escape Time

In this section, we look at the conditional escape time which demonstrates the advantages of conditioning on the escape direction. Although both the number of states n and the forward and backwards rates can already be determined by p_1 and T, the conditional escape time nevertheless offers an interesting theoretical study.

Let T_b and T_f be the time to absorption to the 0 state and $n+1$ states respectively. We will also refer to T_b and T_f as the backwards escape time and forward escape times respectively. The birth-death process on the transient states is fully described by the infinitesimal generator matrix pertaining to the transient states $\{1, \dots, n\}$. Here the infinitesimal generator matrix of the transient states is tri-diagonal with super and sub-diagonal entries r and main diagonal entries $-2r$; let Q be this matrix.

Let $p_k\left(t\right) = Pr\left(S\left(t\right) = k\right)$, where k is a transient state. The state probabilities p_k are given by the solution to the differential equation

$$\frac{d}{dt} = Qp\left(t\right),$$

where $p\left(t\right)$ is vector $p\left(t\right) = \left(p_1\left(t\right), \dots, p_n\left(t\right)\right)$. The solution to this differential equation is given by $p\left(t\right) = \exp\left(tQ\right)p\left(0\right)$. Here, $p\left(0\right) = \left(1, 0, \dots, 0\right)^T$. Let $Q = VDV^T$ be the eigenvalue decomposition of Q. Note that $V^{-1} = V^T$ since V is orthogonal. Hence we have

$$\begin{aligned} p_1\left(t\right) &= \sum_{k=1}^{n} v_{1,k}^2 e^{t\lambda_k}, \\ p_n\left(t\right) &= \sum_{k=1}^{n} v_{n,k} v_{1,k} e^{t\lambda_k}. \end{aligned}$$

The PDF of T_b and T_f is then given by

$$\begin{aligned} f_{T_b}\left(t\right) &= \frac{p_1\left(t\right)}{\int_0^\infty p_1\left(t\right) dt}, \\ f_{T_f}\left(t\right) &= \frac{p_n\left(t\right)}{\int_0^\infty p_n\left(t\right) dt}. \end{aligned}$$

Computing this, we obtain

$$f_{T_b}\left(t\right) = \sum_{k=1}^{n} \omega_1 v_{1,k}^2 e^{t\lambda_k} \tag{4.5}$$

$$f_{T_f}\left(t\right) = \sum_{k=1}^{n} \omega_n v_{n,k} v_{1,k} e^{t\lambda_k} \tag{4.6}$$

where

$$\omega_1 = -\left(\sum_{k=1}^{n} \frac{v_{1,k}^2}{\lambda_k}\right)^{-1},$$

$$\omega_n = -\left(\sum_{k=1}^{n} \frac{v_{n,k} v_{1,k}}{\lambda}_{k}\right)^{-1}.$$

The first-two moments of T_b and T_f can then be easily computed,

$$\langle T_b \rangle = \sum_{k=1}^{n} \omega_1 \frac{v_{1,k}^2}{\lambda_k^2}, \tag{4.7}$$

$$\langle T_f \rangle = \sum_{k=1}^{n} \omega_n \frac{v_{n,k} v_{1,k}}{\lambda_k^2}, \tag{4.8}$$

$$\left\langle T_b^2 \right\rangle = -\sum_{k=1}^{n} 2\omega_1 \frac{v_{1,k} v_{1,k}}{\lambda_k^3}, \tag{4.9}$$

$$\left\langle T_f^2 \right\rangle = -\sum_{k=1}^{n} 2\omega_n \frac{v_{n,k} v_{1,k}}{\lambda_k^3}. \tag{4.10}$$

Since Q is tri-diagonal and Toeplitz, we have closed-form expressions for the eigenvalues and eigenvectors [58]. The eigenvalues and eigenvectors are given by

$$\lambda_k = 2r\left(\cos\left(\frac{k\pi}{n+1}\right) - 1\right), \tag{4.11}$$

$$v_{j,k} = \sin\left(\frac{kj\pi}{n+1}\right), \tag{4.12}$$

for $k, j = 1, \ldots, n$ respectively.

We can then write the first-two moments of T_b and T_f as,

$$\langle T_b \rangle = -\frac{1}{2r}\frac{\sum_{k=1}^{n}\frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}}{\sum_{k=1}^{n}\frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}}, \tag{4.13}$$

$$\left\langle T_b^2 \right\rangle = \frac{1}{2r^2}\frac{\sum_{k=1}^{n}\frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^3}}{\sum_{k=1}^{n}\frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}}, \tag{4.14}$$

$$\langle T_f \rangle = -\frac{1}{2r}\frac{\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}}{\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}}, \tag{4.15}$$

$$\left\langle T_f^2 \right\rangle = \frac{1}{2r^2}\frac{\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^3}}{\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}}. \tag{4.16}$$

Amazingly, these finite trigonometric sums can be calculated analytically by use of the residue theorem or by the use of generating functions established by expansions of trigonometric polynomials in partial fractions [8], [16]. In Appendix B.1, we calculate these summations explicitly. It can then be shown that the first-two moments of T_b and T_f can be written as,

$$\langle T_b \rangle = \frac{1}{6r}\left(2n+1\right), \tag{4.17}$$

$$\left\langle T_b^2 \right\rangle = \frac{1}{r^2}\left(\frac{2n^3}{45}+\frac{8n^2}{45}+\frac{19n}{90}+\frac{1}{15}\right), \tag{4.18}$$

$$\langle T_f \rangle = \frac{1}{6r}\left(n^2+2n\right), \tag{4.19}$$

$$\left\langle T_f^2 \right\rangle = \frac{1}{8r^2}\left(\frac{14n^4}{45}+\frac{56n^3}{45}+\frac{74n^2}{45}+\frac{4n}{5}\right). \tag{4.20}$$

The randomness parameters of T_b and T_f are thus

$$R_{T_b} = \frac{4n^2 + 4n + 7}{5\left(2n+1\right)}, \tag{4.21}$$

$$R_{T_f} = \frac{2n^2 + 4n + 9}{5n\left(n+2\right)}. \tag{4.22}$$

Here we see that asymptotically, we have $R_{T_b} \sim O\left(n\right)$ and $R_{T_f} \sim O\left(1\right)$. This result implies that the ratio R_{T_b}/R_{T_f} can be used to infer the number of states n, for n large. Indeed this can be taken a step further and can be shown to hold for all values of n as the next theorem shows

Theorem 4. *Let $R = R_{T_b}/R_{T_f}$. Then for $n \in \mathbb{N}$, $n - \frac{2}{125} \leq R < n + \frac{1}{2}$ with equality to $n - \frac{2}{125}$ if any only if $n = 2$. Furthermore, R is asymptotic to $n + \frac{1}{2}$ as $n \to \infty$.*

Proof. From equations 4.21 and 4.22, we obtain

$$R = \frac{4n^4 + 12n^3 + 15n^2 + 14n}{4n^3 + 10n^2 + 22n + 9}.$$

By direct computation, $R = 1$ at $n = 1$ and $R = 248/125 = 2 - 2/125$ at $n = 2$. Also, we can write,

$$R = n + \frac{1}{2} - \frac{12n^2 + 6n + \frac{9}{2}}{4n^3 + 10n^2 + 22n + 9}. \tag{4.23}$$

Thus, clearly for $n \geq 3$, $n < R < n + \frac{1}{2}$. Combined with the above, we obtain the desired inequality. From equation 4.23, it is easy to see that R is asymptotic to $n + \frac{1}{2}$ as $n \to \infty$. □

The consequence of this theorem is that if $R = 1$, then $n = 1$, and if $R = 248/125$, then $n = 2$. For $n \geq 3$, $n = \lfloor R \rfloor$.

It turns out that we only need the first moments of T_b and T_f to determine the number of states n as the next theorem shows.

Theorem 5. *For all $n \in \mathbb{N}$, $n \leq 2\frac{\langle T_f \rangle}{\langle T_b \rangle} - 1 < n + \frac{1}{2}$ with equality to n if any only if $n = 1$. Furthermore, the quantity $2\frac{\langle T_f \rangle}{\langle T_b \rangle} - 1$ is asymptotic to $n + \frac{1}{2}$ as $n \to \infty$.*

Proof. From equations 4.17 and 4.19, we have

$$2\frac{\langle T_f \rangle}{\langle T_b \rangle} - 1 = \frac{2n^2 + 4n}{2n+1} = n + \frac{1}{2} - \frac{\frac{3}{2}}{2n+1}.$$

The result follows. □

The consequence here of course is that n can be recovered from the T_f and T_b data by

$$n = \left\lfloor 2\frac{\langle T_f \rangle}{\langle T_b \rangle} - 1 \right\rfloor,$$

for any n.

Theorems 4 and 5 above can be used to determine n from the T_b and T_f data. The transition rate r can then be recovered easily from the first moment of T_b. From equation 4.17, we see that $r = (2n+1) / (6 \langle T_b \rangle)$.

By now, we have seen the advantages of conditioning on the direction of escape. Without conditioning, the unconditional dwell time T does not have enough information content to infer n. After conditioning on the direction of escape, the number of states n can be inferred from p_{f_1} of from the moments of T_b and T_f. In either case, the forward and backward transition rate r can be recovered from the first moment of T_b or from the first moment of T.

4.2.2 Non-Symmetric Birth-Death Process with Two Absorbing States

In this section we generalize the results in the last section by considering that the backwards transition rates are all equal and the forwards transition rates are all equal but different than the backwards transition rates. That is, $r_{b,0} = \cdots = r_{b,n-1} = r_b$, $r_{f,1} = \cdots = r_{f,n} = r_f$, and $r_b \neq r_f$. Thus without loss of generality, we write $r_f = \beta r_b$ and $r_b = r$. Here, $\beta \neq 1$.

Like the symmetric case, we first look at the first-two moments of the unconditional escape time T, starting from state k. Let $h_k = \langle T \mid S(0) = k \rangle$. We can then write

$$h_k = r\Delta t h_{k-1} + r\beta\Delta t h_{k+1} + \left[1 - (1+\beta)\, r\Delta t\right] h_k + \Delta t + o(\Delta t)\,.$$

This leads to the inhomogeneous difference equation

$$rh_{k-1} + r\beta h_{k+1} - (1+\beta)\, rh_k = -1.$$

The solution to this equation is given by

$$h_k = \frac{n+1}{r(\beta-1)\left[1-\left(\frac{1}{\beta}\right)^{n+1}\right]} - \frac{n+1}{r(\beta-1)\left[1-\left(\frac{1}{\beta}\right)^{n+1}\right]}\left(\frac{1}{\beta}\right)^k - \frac{k}{r(\beta-1)},$$

for $\beta \neq 1$. Hence the desired quantity is given by

$$h_1 = \frac{1}{r(\beta-1)}\left[\frac{n+1}{1-\left(\frac{1}{\beta}\right)^{n+1}}\frac{\beta-1}{\beta} - 1\right].$$

The derivation of the second moment of T is similar to the symmetric case.

Again, we have that

$$T_k = \Delta t + \begin{cases} T_{k-1} & \text{with probability } r\Delta t + o\left(\Delta t\right) \\ T_{k+1} & \text{with probability } r\beta\Delta t + o\left(\Delta t\right) \\ T_k & \text{with probability } 1 - r\Delta t\left(\beta+1\right) + o\left(\Delta t\right) \end{cases}$$

We can then write

$$\begin{aligned} \left\langle T_k^2 \right\rangle = {} & \left(\Delta t\right)^2 + 2\Delta t\left[r\Delta t\left\langle T_{k-1}\right\rangle + r\beta\Delta t\left\langle T_{k+1}\right\rangle + \left(1 - r\Delta t\left(\beta+1\right)\right)\right] \\ & + r\Delta t\left\langle T_{k-1}^2\right\rangle + r\beta\Delta t\left\langle T_{k+1}^2\right\rangle + \left[1 - r\Delta t\left(\beta+1\right)\right]\left\langle T_k^2\right\rangle + o\left(\Delta t\right). \end{aligned}$$

Recall that $h_k = \langle T_k \rangle$ and that

$$h_k - \Delta t = r\Delta t h_{k-1} + r\beta\Delta t h_{k+1} + \left[1 - r\Delta t\left(\beta+1\right)\right]h_k + o\left(\Delta t\right).$$

Hence we can write

$$\left\langle T_k^2\right\rangle = 2\Delta t\left(h_k - \Delta t\right) + r\Delta t\left\langle T_{k-1}^2\right\rangle + r\beta\Delta t\left\langle T_{k+1}^2\right\rangle + \left[1 - r\Delta t\left(\beta+1\right)\right]\left\langle T_k^2\right\rangle + o\left(\Delta t\right)$$

Using the notation $u_k = \langle T_k^2 \rangle$, we have the inhomogeneous difference equation

$$u_{k-1} + \beta u_{k+1} - \left(\beta+1\right)u_k = -\frac{2}{r}h_k.$$

The solution is of the form

$$u_k = C_1 + C_2\left(\frac{1}{\beta}\right)^k + B_1 k + B_2 k\left(\frac{1}{\beta}\right)^k + B_3 k^2,$$

where

$$C_1 = \frac{B_1\left(n+1\right) + B_2\left(n+1\right)\left(\frac{1}{\beta}\right)^{n+1} + B_3\left(n+1\right)^2}{\left(\frac{1}{\beta}\right)^{n+1} - 1}$$

and $C_2 = -C_1$. Here, $B_1 = \left[A_1 - B_3\left(1+\beta\right)\right]/\left(\beta-1\right)$, $B_2 = A_2/\left(1-\beta\right)$, and $B_3 = A_3/\left(2\beta-2\right)$. The A_i coefficients are given by

$$\begin{aligned}
A_1 &= -\frac{2\left(n+1\right)}{r^2\left(\beta-1\right)\left[1-\left(\frac{1}{\beta}\right)^{n+1}\right]} \\
A_2 &= \frac{2\left(n+1\right)}{r^2\left(\beta-1\right)\left[1-\left(\frac{1}{\beta}\right)^{n+1}\right]} \\
A_3 &= \frac{2}{r^2\left(\beta-1\right)}.
\end{aligned}$$

Hence the desired quantity is given by

$$u_1 = C_1 +_2 \left(\frac{1}{\beta}\right) + B_1 + B_2\left(\frac{1}{\beta}\right) + B_3,$$

The randomness parameter is thus a function of both n and β, unlike the symmetric case. Here, we have no hope of recovering n or β unless we also look at the probability of forward (or backwards) escape.

The probability of forward escape can be derived in a similar manner as the symmetric case. Again, let p_{f_k} be the probability of forward escape with $S\left(0\right) = k$. Here we have the boundary conditions $p_{f_0} = 0$ and $p_{f_{n+1}} = 1$. We also have the following transition probabilities

$$\begin{aligned}
Pr\left(S\left(\Delta t\right) = k-1 \mid S\left(0\right) = k\right) &= r\Delta t + o\left(\Delta t\right) \\
Pr\left(S\left(\Delta t\right) = k+1 \mid S\left(0\right) = k\right) &= \beta r\Delta t + o\left(\Delta\right) \\
Pr\left(S\left(\Delta t\right) = k \mid S\left(0\right) = k\right) &= 1 - \left(1+\beta\right) r\Delta t + o\left(\Delta t\right)
\end{aligned}$$

Hence we can write,

$$p_{f_k} = r\Delta t p_{f_{k-1}} + \beta r \Delta t p_{f_{k+1}} + \left[1 - (1+\beta) r \Delta t\right] p_{f_k} + o\left(\Delta t\right).$$

Dividing by Δt and passing the limit $t \to 0$, we obtain the difference equation

$$r p_{f_{k-1}} + \beta r p_{f_{k+1}} - (1+\beta) r p_{f_k} = 0.$$

Solving this difference equation gives the result

$$p_{f_k} = \frac{1}{1-\left(\frac{1}{\beta}\right)^{n+1}} - \frac{1}{1-\left(\frac{1}{\beta}\right)^{n+1}} \left(\frac{1}{\beta}\right)^{k},$$

for all $\beta \neq 1$. Thus, the quantity we are interested in is given by

$$p_{f_1} = \frac{1}{1-\left(\frac{1}{\beta}\right)^{n+1}} - \frac{1}{1-\left(\frac{1}{\beta}\right)^{n+1}} \left(\frac{1}{\beta}\right). \tag{4.24}$$

The forward escape probability and the randomness parameter of T provide a mapping $(n,\beta) \mapsto (p_{f_1}, R_T)$. Unlike the symmetric case, it appears that analytically inverting this mapping is intractable. We thus develop a method to invert this mapping numerically by least-squares.

Determining n and β from p_{f_1} and R_T Numerically

Let F denote the mapping

$$F(n,\beta) = (p_{f_1}, R_T) := (f_1(n,\beta), f_2(n,\beta)) \tag{4.25}$$

Let F^{obs} denote the observed value of (n,β), so that $F^{\text{obs}} = \left(f_1^{\text{obs}}, f_2^{\text{obs}}\right)$. Define a grid of (n,β)-points and descritize in the β-direction. Let $\mathcal{N}$ and $\mathcal{B}$ be the set of points for n and discrete set of points for β, respectively. Enumerate the n and β points,

$$\mathcal{N} = \{n_1, \ldots, n_q\}\,,$$
$$\mathcal{B} = \{\beta_1, \ldots, \beta_m\}\,,$$

where q and m are the number of n and β-points respectively (figure 4.7).

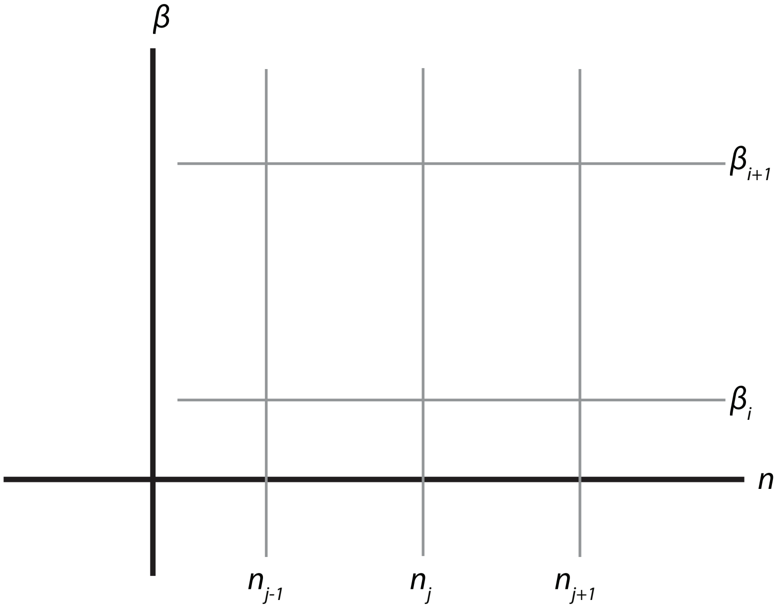

Figure 4.7: A schematic grid of the n and β points. By definition, n is already discrete. The discretization occurs in the β-direction.

For each line segment defined by (n_j, β_i) to (n_j, β_{i+1}) for $j = 1, \ldots, q$ and $i = 1, \ldots, m-1$, we solve the following least-squares problem. Let

$$s := \frac{\beta - \beta_i}{\beta_{i+1} - \beta_i},$$

for $\beta_i \leq \beta < \beta_{i+1}$. Define the functions

$$g_1(s,i) = (1-s) f_1(n,\beta_i) + s f_1(n,\beta_{i+1}),$$
$$g_2(s,i) = (1-s) f_2(n,\beta_i) + s f_2(n,\beta_{i+1}).$$

For each n_j, $j = 1,\ldots,q$ and $i = 1,\ldots,m$ solve the following

$$s_{n_j,i} = \arg\min_{s} \left(\left(g_1(s,i) - f_1^{\mathrm{obs}} \right)^2 + \left(g_2(s,i) - f_2^{\mathrm{obs}} \right)^2 \right). \tag{4.26}$$

Let $s_{n_j,i}$ be the solution to the above least-squares problem and let $m_{n_j,i}$ be the corresponding minimum. Discard $s_{n_j,i}$ and the corresponding $m_{n_j,i}$ if $s_{n_j,i} < 0$ or $s_{n_j,i} \geq 1$. Let

$$n^* = \arg\min_{n_j} m_{n_j,i}.$$

This is the estimate of n. The corresponding estimate of β is given by $\beta^* = \beta_i + (\beta_{i+1} - \beta_i) s_{n^*}$, where i is the corresponding index in which the least n^* was obtained.

We summarize the above in the following algorithm for finding the estimate of n and β from the p_{f_1} and R_T observations.

Algorithm 6.

```
input: $(p_{f_1}, R_T)$.
output: $(n^*, \beta^*)$.
begin:
define: $n = (n_1, \ldots, n_q)$.
define: $\beta = (\beta_1, \ldots, \beta_m)$.
define: snj, mnj array of size $q \times m$.
for $j = 1, \ldots, q$ do
```

```
        for $i = 1, \dots, m$ do
            $snj\left(j,i\right) = \arg\min_s \left(\left(g_1\left(s,i\right) - f_1^{obs}\right)^2 + \left(g_2\left(s,i\right) - f_2^{obs}\right)^2\right).$
            $mnj(j,i) = \left(g_1\left(snj(j,i),i\right) - f_1^{obs}\right)^2 + \left(g_2\left(snj(j,i),i\right) - f_2^{obs}\right)^2.$
            if $snj < 0$ or $snj \geq 1$ do
                    $snj(j,i) = -1.$
            endif
        endfor
endfor
$ind = where\left(\min\left(mnj\right)\right).$
$n^* = n\left(ind(1)\right).$
$\beta^* = \beta\left(ind\left(2\right)\right) + \left(\beta\left(ind\left(2\right)+1\right) - \beta\left(ind\left(2\right)\right)\right) snj\left(ind(1), ind(2)\right).$
end
```

Determining n and β from p_{f_1}, R_{T_b}, and R_{T_f}

We can modify the above algorithm to use the randomness parameter of conditional escape times T_b and T_f instead of the unconditional escape time T. The strategy of the algorithm is the same. Only a slight modification to the mapping F in equation 4.25 and the least-squares objective function in equation 4.26.

Let F be the mapping

$$F\left(n,\beta\right) = \left(p_{f_1}, R_{T_b}, R_{T_f}\right) := \left(f_1\left(n,\beta\right), f_2\left(n,\beta\right), f_3\left(n,\beta\right)\right).$$

Define the functions

$$g_k\left(s,i\right) = \left(1-s\right) f_k\left(n,\beta_i\right) + s f_k\left(n,\beta_{i+1}\right),$$

for $k = 1, 2, 3$. Hence the new least-squares objective function is given by

$$s_{n_j,i} = \arg\min_s \sum_{k=1}^{3} \left(g_k\left(s, i\right) - f_1^{\text{obs}}\right)^2 .$$

The following algorithm for estimating n and β from p_{f_1}, R_{T_b}, and R_{T_f} is a slight modification of algorithm 6.

Algorithm 7.

```
***input:*** $\left(p_{f_1}, R_{T_b}, R_{T_f}\right)$.
***output:*** $(n^*, \beta^*)$.
***begin:***
***define:*** $n = (n_1, \ldots, n_q)$.
***define:*** $\beta = (\beta_1, \ldots, \beta_m)$.
***define:*** $snj$, $mnj$ array of size $q \times m$.
***for*** $j = 1, \ldots, q$ ***do***
    ***for*** $i = 1, \ldots, m$ ***do***
        $snj\,(j, i) = \arg\min_s\ \sum_{k=1}^{3} \left(g_k\left(snj(j,i), i\right) - f_k^{obs}\right)^2$.
        $mnj(j, i) = \sum_{k=1}^{3} \left(g_k\left(snj(j,i), i\right) - f_k^{obs}\right)^2$.
        ***if*** $snj < 0$ ***or*** $snj \geq 1$ ***do***
                $snj(j, i) = -1$.
        ***endif***
    ***endfor***
***endfor***
$ind = where\,(\min\,(mnj))$.
$n^* = n\,(ind(1))$.
$\beta^* = \beta\,(ind\,(2)) + (\beta\,(ind\,(2) + 1) - \beta\,(ind\,(2)))\, snj\,(ind(1), ind(2))$.
***end***
```

4.2.3 Numerical Simulations

In this section, we conduct numerical simulations to see the performance of the two least-squares algorithms above; one utilizing (p_{f_1}, R_T) and the other utilizing $\left(p_{f_1}, R_{T_b}, R_{T_f}\right)$ to infer n and β.

We first start with $n_{\text{true}} = 5$ and $\beta_{\text{true}} = 1.1$ as the true values of n and β. Here, n and β are varied over a grid from 3 to 7 and 0.5 to 1.5 with 200 equally spaced points, respectively. We test the sensitivity of the algorithms by introducing multiplicative noise into the observations of p_{f_1}, R_T, R_{T_b}, and R_{T_f} of the form

$$p_{f_1}^{\text{obs}} = p_{f_1} e^{\sigma\zeta},$$

where $\zeta \sim N(0,1)$ and σ is the standard deviation. The perturbed samples of R_T^{obs}, $R_{T_b}^{\text{obs}}$, and $R_{T_f}^{\text{obs}}$ are defined in a similar manner. The numerical experiment is repeated 1,000 times to obtain a distribution of estimated n and β values for both least-squares codes. For each graph, we plot the quantity

$$\text{err}(\beta) = \frac{\beta^{\text{LS}} - \beta^{\text{true}}}{\beta^{\text{true}}}.$$

The quantity $\text{err}(n)$ is defined similarly. In the following figures, we plot $\text{err}(\beta)$ and $\text{err}(n)$ for $\sigma = 0, 2^{-6}, 2^{-4}$. We refer to the unconditional least-squares code as algorithm 6 and the conditional least-squares code as algorithm 7.

With $\sigma = 0$, both least-squares codes recover n and β exactly. The distribution of the inferred n and β is therefore a point-mass at 5 and 1.1 respectively.

Figures 4.8 and 4.9 show the results for $\sigma = 2^{-6}$ and $\sigma = 2^{-4}$, respectively. As we see from the figures, the unconditional least-squares code outperforms the conditional least-squares code when inferring β. Both codes are comparable when inferring n. To see why this is the case, consider figure 4.10 which shows

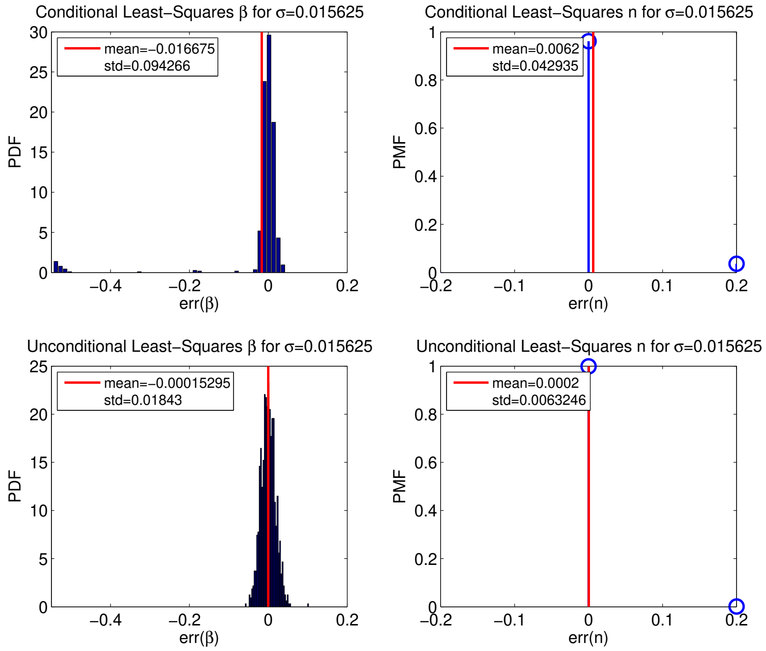

Figure 4.8: Comparison of the conditional and unconditional least-squares code for $n = 5$, $\beta = 1.1$, and $\sigma = 2^{-6}$.

p_{f_1}, R_T, R_{T_b}, and R_{T_f} as a function of β and n. From the plots of R_{T_b} and R_{T_f} as a function of β for fixed n, we see that R_{T_b} and R_{T_f} are not injective on the interval $[0.5, 1.5]$; hence are are two values of β that correspond to the same R_{T_b} and R_{T_f}. This explains why in figures 4.8 and 4.9, the distribution of $\operatorname{err}(\beta)$ has a small cluster of mass around -0.5. We note that the same result would likely occur for R_T if the search-grid for β was expanded and the true value of β was centered around the location of the maximum of R_T, $\beta \approx 0.6$.

From these simulations, a better strategy for inferring n and β from the dwell time data would be to run the least-squares code for (p_{f_1}, R_T) or $\left(p_{f_1}, R_{T_b}, R_{T_f}\right)$ to infer n. Then fixing n and re-running the least-squares code on just p_{f_1} to infer β. From figure 4.10, p_{f_1} is injective throughout the range of β for fixed n; hence the the inference of β from p_{f_1} is a numerically easier task.

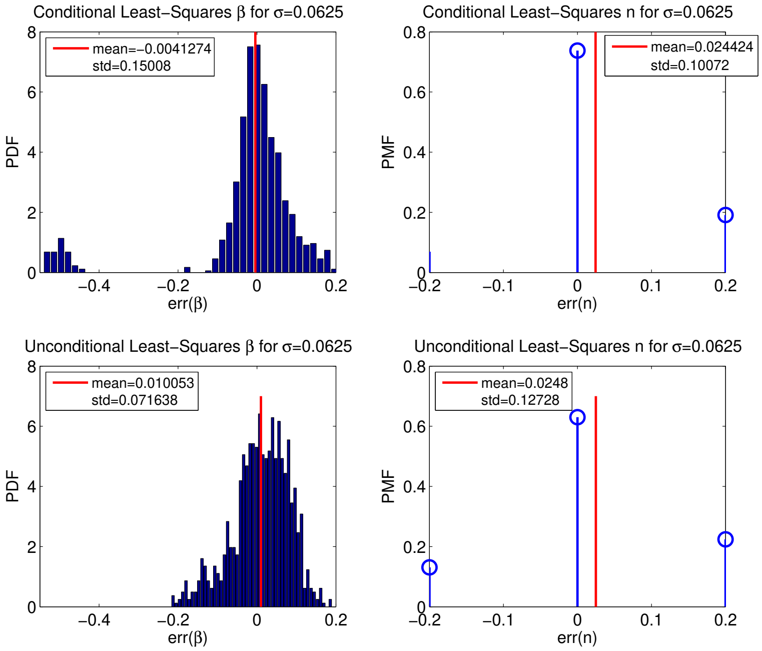

Figure 4.9: Comparison of the conditional and unconditional least-squares code for $n = 5$, $\beta = 1.1$, and $\sigma = 2^{-4}$.

We re-run these numerical simulations with the true $n = 30$. Examining figure 4.10, we expect the inference for n would be less accurate for large n since p_{f_1} and R_T are relatively flat in this region. Like the case for $n = 5$, for no multiplicative noise, both least-squares codes recover n and β exactly. Figures 4.11 and 4.12 show the distribution of the quantities $\text{err}\,(\beta)$ and $\text{err}\,(n)$ for $\sigma = 2^{-6}$ and $\sigma = 2^{-4}$, respectively. Here we see that the estimates for β in the conditional and unconditional least-squares codes are comparable. For estimating n, the conditional least-squares code outperforms the unconditional least-squares code. For $\sigma = 2^{-4}$, the conditional least-squares code has about a 7% reduction in relative error and a 2% reduction in bias over the unconditional least-squares code.

Figures 4.10, 4.11, and 4.12 highlight the advantages of conditioning–using R_{T_b} and R_{T_f} for the inference of n in particular. For large n, using R_{T_b} and R_{T_f} to

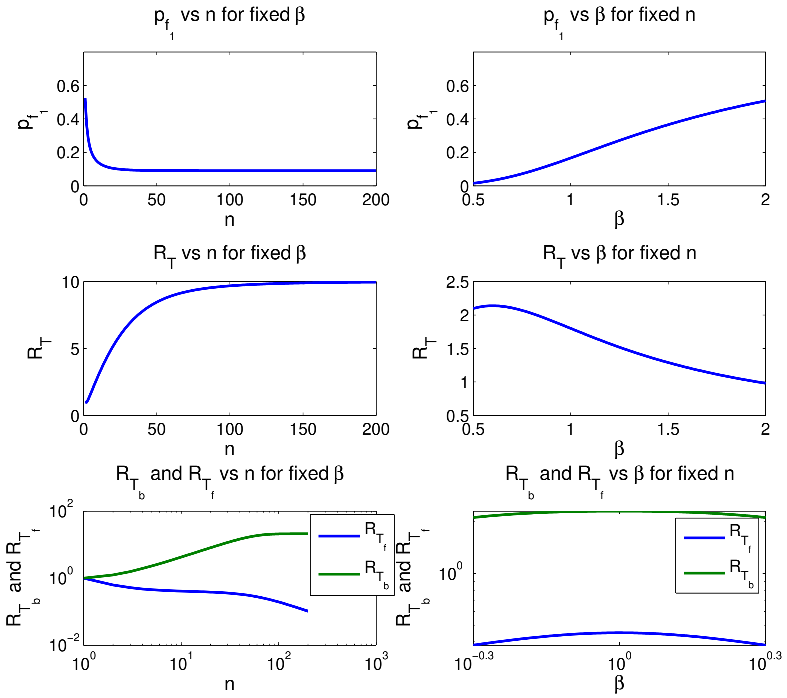

Figure 4.10: Plot of p_{f_1}, R_T, R_{T_b}, and R_{T_f} as a function of β for fixed $n = 5$ (left); and plot of p_{f_1}, R_T, R_{T_b}, and R_{T_f} as a function of n for fixed $\beta = 1.1$ (right).

infer n is advantages over using R_T because R_T is relatively flatter in this region.

4.3 Inferring the Kinetic Structure of the Polymerization Process

In this section, we extend our results from the previous section to infer the kinetic structure of the polymerization process in DNAP-DNA complexes. The most straightforward approach is to examine the randomness parameter of T_B. Recall that $T_B = T_{pol} + T^{(2)}$. The dwell times T_{pol} and $T^{(2)}$ are independent,

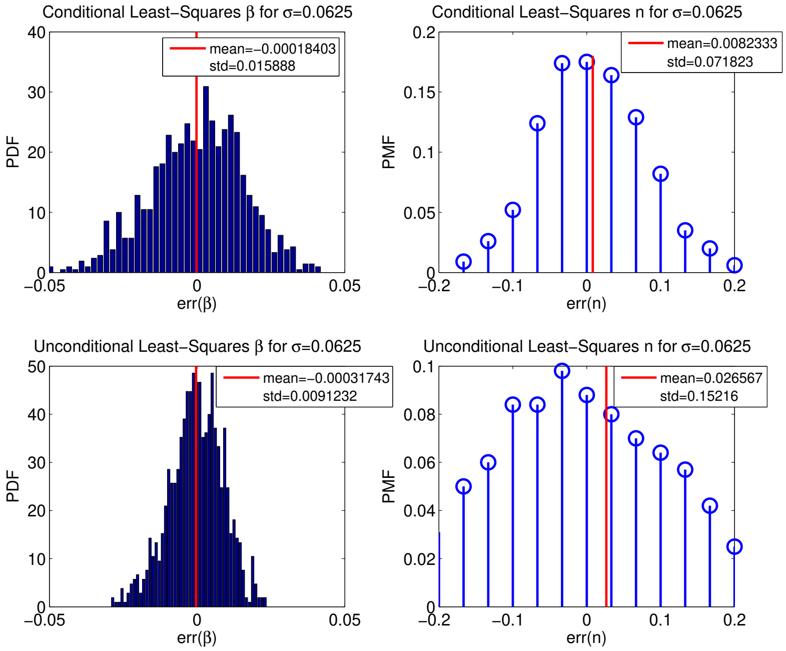

Figure 4.11: Comparison of the conditional and unconditional least-squares code for $n = 5$, $\beta = 1.1$, and $\sigma = 2^{-6}$.

hence we can write the randomness parameter of T_B as

$$R_{T_B} = \frac{R_{T_{pol}} \langle T_{pol} \rangle^2 + R_{T^{(2)}} \left\langle T^{(2)} \right\rangle^2}{\left(\langle T_{pol} \rangle + \langle T^{(2)} \rangle\right)^2}.$$

This shows that we can study the randomness parameters of T_{pol} and $T^{(2)}$ individually to gain insight into the randomness parameter of T_B.

4.3.1 Randomness Parameter of $T^{(2)}$

The randomness parameter of $T^{(2)}$ is worth special consideration, and we study its behavior here. Recall that $T^{(2)}$ is the time to escape the upper-amplitude state when starting at the pre-translocation state (figure 1.5). The state-space structure of this escape problem is more akin to the T_b escape times studied in the birth-

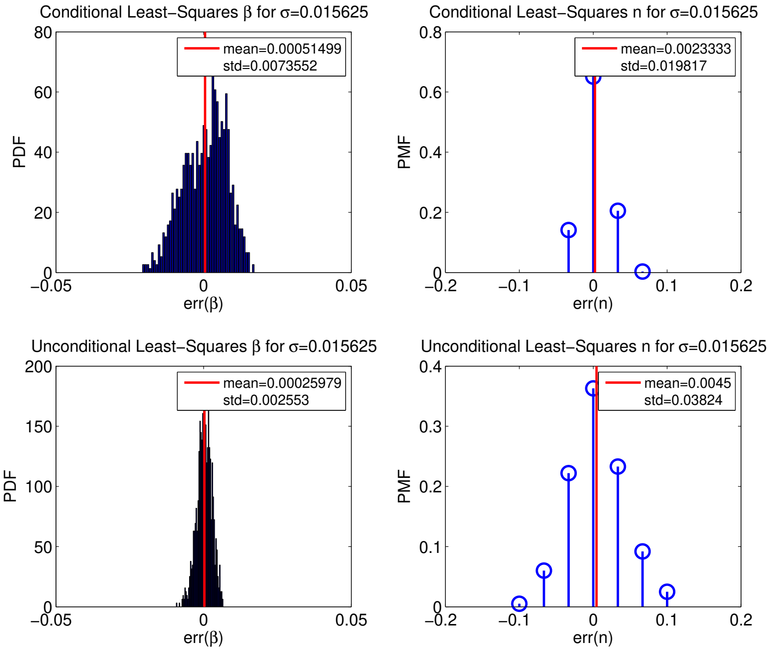

Figure 4.12: Comparison of the conditional and unconditional least-squares code for $n = 5$, $\beta = 1.1$, and $\sigma = 2^{-4}$.

death processes with absorbing boundary states above. As the reader may have noticed the randomness parameter of T_b can exceed 1; indeed, that is the case for the randomness parameter of $T^{(2)}$ (figure 4.13). Note that due to scaling, we can write the randomness parameter of $T^{(2)}$ as a function of r_3 and r_4 only. To see this, note that the transition rates have units $[time]^{-1}$. So we can scale $T^{(2)}$ such that $r_1 \mapsto 1$, $r_3 \mapsto r_3/r_1$, and $r_4 \mapsto r_4/r_1$ (see Proposition 2 in Chapter 2).

As seen in figure 4.13, the randomness parameter can exceed 1. This provides a sufficient but not necessary condition for determining the existence of branches in a Markov process. The contour plot of $R_{T^{(2)}}$ suggests that $R_{T^{(2)}} \to \infty$ as $r_3, r_4 \to 0$. Indeed this makes sense since the time-scale to equilibrium $1/(r_3 + r_4) \to \infty$ as $r_3, r_4 \to 0$; hence the variance of the escape time $T^{(2)}$ approaches infinity.

The overall shape of the contour lines in figure 4.13 is interesting, and is

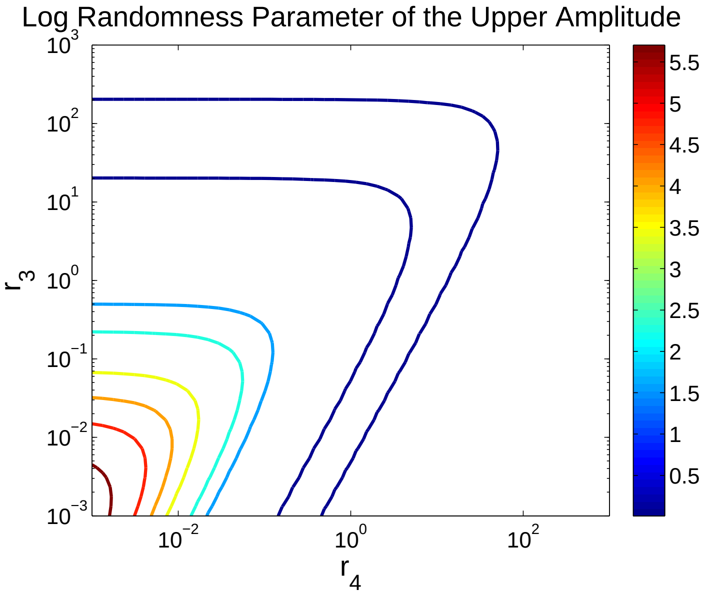

Figure 4.13: The randomness parameter of $T^{(2)}$, $R_{T^{(2)}}$, as a function of r_3 and r_4.

asymmetric. To examine the behavior of $R_{T^{(2)}}$, we investigate some asymptotic cases for r_3 and r_4. Let $0 < \epsilon << 1$, so that $1/\epsilon$ is large.

1. $r_1 = 1$ and $r_3 = r_4 = 1/\epsilon$:

   $R_{T^{(2)}} = 1 + \frac{r_1 \epsilon}{2} + (\epsilon^2)$.

2. $r_1 = r_3 = 1$ and $r_4 = \frac{2}{\epsilon} - 1$:

   $R_{T^{(2)}} = 1 + \frac{r_1 \epsilon^2}{2} + O\left(\epsilon^2\right)$.

3. $r_1 = r_4 = 1$ and $r_3 = \frac{2}{\epsilon} - 1$:

   $R_{T^{(2)}} = 1 + r_1 \epsilon + O\left(\epsilon^2\right)$.

4. $r_1 = 1$, $r_3 = \frac{1-\delta}{\epsilon}$, and $r_4 = \frac{1-\delta}{\epsilon}$, where $-1 < \delta < 1$ is fixed:

   $R_{T^{(2)}} = 1 - \frac{r_1(\delta - 1)\epsilon}{2} + O\left(\epsilon^2\right)$.

Cases (1)-(3) provide a sense of how close the pre-translocation and exonuclease states are to being one "superstate." Cases (2) and (3) show that $R_{T^{(1)}}$ approaches 1 faster in the r_4 direction than in the r_3 direction in agreement with figure 4.13. Case (4) examines the behavior around the "bend" of the contour lines.

4.3.2 Randomness Parameter of T_{pol}

In this section, we investigate the randomness parameter of T_{pol}. The dwell time T_{pol} is the time from the last arrival to the post-translocation state to the first arrival of the pre-translocation state of the next nucleotide addition cycle. Figure 4.14 shows the state-space diagram of the Markov process governing the escape problems for the polymerization process. Conditioning on starting at the post-translocation state and escaping to the pre-translocation state of the next nucleotide addition cycle generates T_{pol}. Note that the Markov process whose state-space diagram is shown in 4.14 is isomorphic to the birth-death process with two absorbing states introduced in section 4.2. Viewed in this way, T_{pol} is the same as T_f, the forward escape time in the birth-death process with two absorbing states.

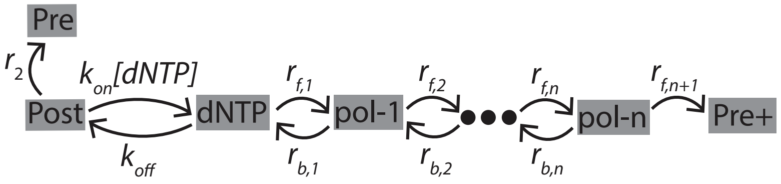

Figure 4.14: State-space diagram of the Markov process governing the escape problems for the polymerization process. Conditioned on starting at the post-translocation state and escaping to the pre-translocation state of the next nucleotide addition cycle generates T_{pol}.

For a Markov process with absorbing states, we can condition on the escape to any of the absorbing states and write down an equivalent Markov process describing this conditioning. Before we present how to write down a conditioned Markov process, we introduce some definitions.

Let $\mathcal{A}$ be the set of absorbing states, and $\mathcal{T}$ be the set of transient states. Let $a \in \mathcal{A}$ and define the following events:

$$\begin{aligned}
E_a^{=t} &= \left\{X\left(t\right) = a, X\left(r\right) \notin \mathcal{A} \ : \ 0 < r < t\right\}, \\
E_a^{<t} &= \left\{X\left(t'\right) = a, X\left(r\right) \notin \mathcal{A} \ : \ 0 < r < t' \leq t\right\}, \\
E_a^{>t} &= \left\{X\left(t'\right) = a, X\left(r\right) \notin \mathcal{A} \ : \ 0 < r < t < t'\right\}, \\
E_a &= \bigcup_{t>0} E_a^{=t}.
\end{aligned} \tag{4.27}$$

Informally,

- $E_a^{=t}$ is the event of arriving at a for the first time *at* time t before arriving in any other state in $\mathcal{A}$;

- $E_a^{<t}$ is the event of arriving at a for the first time *by* time t before arriving in any other state in $\mathcal{A}$;

- $E_a^{>t}$ is the event of arriving at a for the first time *after* time t before arriving in any other state in $\mathcal{A}$;

- E_a is the event of arriving at a for the first time before arriving in any other state in $\mathcal{A}$.

Now consider the following proposition.

Proposition 8. *Let $S\left(t\right)$ be a Markov process with state-space $\mathcal{I} = \mathcal{A} \cup \mathcal{T}$, where $\mathcal{A}$ denotes the set of absorbing states and $\mathcal{T}$ denotes the set of transient states. Let Q be the infinitesimal generator matrix characterizing the Markov process. Let E_a be the event of escaping to absorbing state $a \in \mathcal{A}$. Let $p_{E_a|k} = Pr\left(E_a \mid X\left(0\right) = k\right)$.*

Then the (i,j)-component of the conditional infinitesimal generator matrix Q_{E_a} is given by

$$Q_{i,j,E_a} = \begin{cases} 0 & \text{if } i \in \mathcal{A} \text{ and } j \in \mathcal{A} \cup \mathcal{T} \text{ or } i \in \mathcal{T} \text{ and } j \in \mathcal{A} \backslash \{a\} \\ \frac{p_{E_a|j} Q_{i,j}}{p_{E_a|i}} & \text{if } i,j \in \mathcal{T} \text{ and } i \neq j \\ -\sum_{\substack{j \in \mathcal{A} \cup \mathcal{T} \\ j \neq i}} Q_{i,j,E_a} & \text{if } i,j \in \mathcal{T} \text{ and } i = j \\ \frac{Q_{i,j}}{p_{E_a|i}} & \text{if } i \in \mathcal{T} \text{ and } j = a \end{cases}$$

Proof. Clearly if $i \in \mathcal{A}$, then the transition rate is 0. Also if i is transient and j is an absorbing state different than a, then the transition rate is 0 (since we are conditioning on the escape to a).

When $i, j \in \mathcal{T}$ and $i \neq j$, we have

$$\begin{aligned} & Pr\left(X\left(t\right) = j \mid X\left(0\right) = i, E_a\right) \\ & \quad = \frac{Pr\left(E_a, X\left(t\right) = j \mid X\left(0\right) = i\right) Pr\left(X\left(0\right) = i\right)}{Pr\left(E_a \mid X\left(0\right) = i\right) Pr\left(X\left(0\right) = i\right)} \\ & \quad = \frac{Pr\left(E_a^{>t}, X\left(t\right) = j \mid X\left(0\right) = i\right)}{Pr\left(E_a \mid X\left(0\right) = i\right)} \text{ since } \{X\left(t\right) = j\} \cap E_a = E_a^{>t} \\ & \quad = \frac{Pr\left(E_a^{>t} \mid X\left(t\right) = j, X\left(0\right) = i\right) Pr\left(X\left(t\right) = j \mid X\left(0\right) = i\right)}{p_{E_a|i}} \\ & \quad = \frac{Pr\left(E_a^{>t} \mid X\left(t\right) = j\right) Pr\left(X\left(t\right) = j \mid X\left(0\right) = i\right)}{p_{E_a|i}} \text{ by the Markov property} \\ & \quad = \frac{p_{E_a|j} Q_{i,j} t}{p_{E_a|i}} + o\left(t\right). \end{aligned}$$

Hence

$$Q_{i,j,E_a} = \lim_{t \to 0} \frac{Pr\left(X\left(t\right) = j \mid X\left(0\right) = i, E_a\right)}{t} = \frac{p_{E_a|j} Q_{i,j}}{p_{E_a|i}}.$$

The case $i, j \in \mathcal{T}$, $i \neq j$ follows since the probabilities must add up to 1. And finally, if $i \in \mathcal{T}$ and $j = a$, then

$$\begin{aligned}
&Pr\left(X\left(t\right) = j \mid X\left(0\right) = i, E_a\right) \\
&\qquad = \frac{Pr\left(E_a, X\left(t\right) = j \mid X\left(0\right) = i\right) Pr\left(X\left(0\right) = i\right)}{Pr\left(E_a \mid X\left(0\right) = i\right) Pr\left(X\left(0\right) = i\right)} \\
&\qquad = \frac{Pr\left(E_a, X\left(t\right) = j \mid X\left(0\right) = i\right)}{P_{E_a|i}} \\
&\qquad = \frac{Pr\left(E_a \mid X\left(t\right) = j, X\left(0\right) = i\right) Pr\left(X\left(t\right) = j \mid X\left(0\right) = i\right)}{p_{E_a|i}} \\
&\qquad = \frac{Pr\left(E_a \mid X\left(t\right) = j\right) Q_{i,j} t}{p_{E_a|i}} + o\left(t\right).
\end{aligned}$$

Hence,

$$Q_{i,j,E_a} = \lim_{t \to 0} \frac{Pr\left(X\left(t\right) = j \mid X\left(0\right) = i, E_a\right)}{t} = \frac{Q_{i,j}}{p_{E_a|i}}.$$

□

After conditioning on the escape to the pre-translocation state of the next nucleotide addition cycle, the state-space diagram of the Markov process governing the T_{pol} dwell time is given in figure 4.15. Here, we label the states $\{\text{pre}, \text{post}, \text{dNTP}, \text{pol-1}, \ldots, \text{pol-n}, \text{pre+}\}$ as $\{0, 1, 2, 3, \ldots, n+2, n+3\}$, respectively for notational convenience. We will refer to the states as either the name, shortened name, or number; i.e., "post-translocation", "post," or 2. Here, $\rho_{f,i} := Pr\left(E_{pre+} \mid X\left(0\right) = i\right)$. We can compute the absorption probabilities $\rho_{f,i}$ by con-

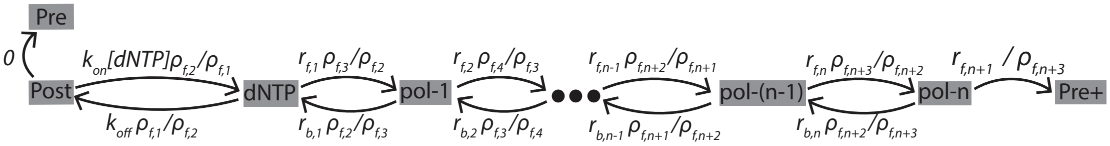
Pre
0
Post
$k_{on}[dNTP]\rho_{f,2}/\rho_{f,1}$
$k_{off}\rho_{f,1}/\rho_{f,2}$
dNTP
$r_{f,1}\rho_{f,3}/\rho_{f,2}$
$r_{b,1}\rho_{f,2}/\rho_{f,3}$
pol-1
$r_{f,2}\rho_{f,4}/\rho_{f,3}$
$r_{b,2}\rho_{f,3}/\rho_{f,4}$
$r_{f,n-1}\rho_{f,n+2}/\rho_{f,n+1}$
$r_{b,n-1}\rho_{f,n+1}/\rho_{f,n+2}$
pol-(n-1)
$r_{f,n}\rho_{f,n+3}/\rho_{f,n+2}$
$r_{b,n}\rho_{f,n+2}/\rho_{f,n+3}$
pol-n
$r_{f,n+1}\,/\rho_{f,n+3}$
Pre+

Figure 4.15: State-space diagram of the Markov process governing the escape problem for T_{pol} after conditioning. Here, we label the states $\{\text{pre}, \text{post}, \text{dNTP}, \text{pol-1}, \ldots, \text{pol-n}, \text{pre+}\}$ as $\{0, 1, 2, 3, \ldots, n+2, n+3\}$, respectively for notational convenience. Here, $\rho_{f,i} := Pr\left(E_{pre+} \mid X\left(0\right) = i\right)$.

sidering the subordinated discrete-time Markov chain with transition matrix $K = I + Q_{E_{pre+}}$ where I is the identity matrix and $Q_{E_{pre+}}$ is the infinitesimal generator matrix characterizing the Markov process whose state-space is given in figure 4.15. Let U be the submatrix of K corresponding to the probabilities of transitioning among states in $\mathcal{T}$, and let R be the submatrix of K corresponding to the probabilities of transitioning from transitioning from a transient state to an absorbing state. Then it can be shown that the absorption probabilities are given by the entries of the matrix $B = (I - U)^{-1} R$ (see [42]). The matrix $(I - U)^{-1}$ is known as the fundamental matrix of the absorbing Markov chain.

The absorption probabilities $\rho_{f,i}$ in figure 4.15 are complicated functions of all the transition rates and $[dNTP]$, so it is not clear analytically how the randomness parameter of T_{pol} behaves as a function of $[dNTP]$. Lets assume that the forward transition rates $r_{f,i}$ are all equal and that the forward rates are a constant multiple of the backwards transition rates; i.e., $r_{f,i} = \beta r_{b,i}$, where $\beta > 0$.

The PDF and hence the first-two moments for T_{pol} can in principle be written down, but it is nearly intractable, so we turn to numerical computation. We can largely follow the strategy, with modification, of computing the PDF and moments of T_f in section 4.2.1 for computing T_{pol}. Let $Q_{\mathcal{T}}$ be the infinitesimal generator corresponding to the transient states. Due to scaling, we can scale the units of time so that the backwards transition rates $r_{b,i} \mapsto 1$ and the forward transition rates $r_{f,i} = \beta r_{b,i} \mapsto \beta$. Thus without loss of generality, we can set $r_{f,i} = \beta$ and $r_{b,i} = 1$. The matrix $Q_{\mathcal{T}}$ will thus be tri-diagonal with the diagonal entries $(-\beta[dNTP] - 1, -\beta - 1, \ldots, -\beta - 1)$, super-diagonal entries $(\beta, \ldots, \beta)$, and sub-diagonal entries $(1, \ldots, 1)$. Here, the eigenvectors will not be orthogonal since $Q_{\mathcal{T}}$ is not symmetric.

We introduce a transformation to make $Q_{\mathcal{T}}$ symmetric. Let $q_i\left(t\right)$ be such that

$p_i(t) = \beta^{i/2} q_i(t)$, where $p'(t) = Q_{\mathcal{T}} p(t)$ are the state-probability equations and $p_i(t) = Pr(X(t) = i)$. Thus the transformed system becomes $q'(t) = \tilde{Q}_{\mathcal{T}} q(t)$ with the transformed infinitesimal generator matrix $\tilde{Q}_{\mathcal{T}}$ being tri-diagonal with main diagonal $(-\beta[dNTP] - 1, -\beta - 1, \ldots, -\beta - 1)$, super-diagonal $\left(\sqrt{\beta}, \ldots, \sqrt{\beta}\right)$, and sub-diagonal $\left(\sqrt{\beta}, \ldots, \sqrt{\beta}\right)$. The transformed matrix $\tilde{Q}_{\mathcal{T}}$ is thus symmetric and its eigenvectors are orthogonal. The PDF of T_{pol} and its first-two moments can then be computed according to equations 4.6, 4.8, and 4.10.

Figure 4.16 shows the reciprocal of $R_{T_{pol}}$ as a function of $[dNTP]$ for various numbers of transient states n. We see that over a large range of $[dNTP]$, the randomness parameter of T_{pol} does not reveal much about the number of states. Over the range of n examined, the randomness parameter of T_{pol} does not change significantly throughout the $[dNTP]$.

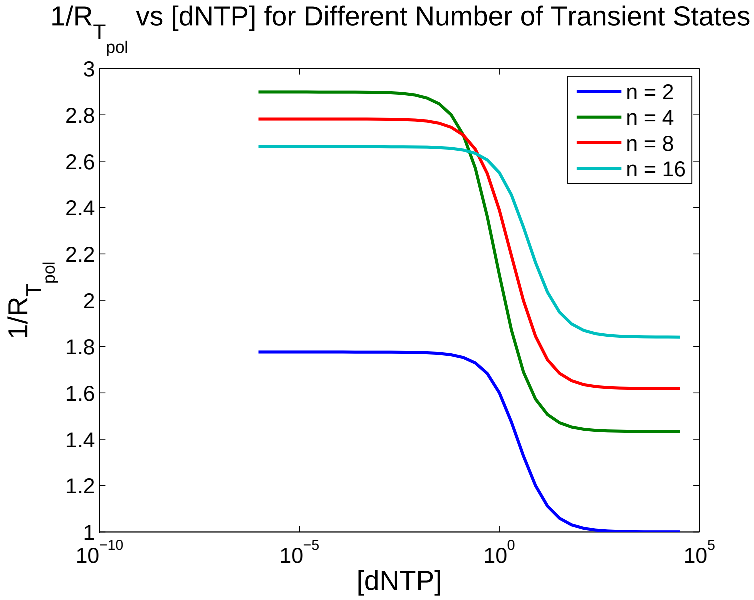

Figure 4.16: The reciprocal of the randomness parameter $1/R_{T_{pol}}$ as a function of $[dNTP]$ for various number of transient states, n. In this simulation, the forward rates, $r_{f,i} = \beta r_{b,i}$ where $\beta = 1.1$.

4.3.3 Randomness Parameter of T_B

In this section, we consider the randomness parameter of T_B. Since the T_B dwell time can be exactly observed, we can readily calculate R_{T_B} from the T_B data. In the last two sections, we examined the randomness parameters of T_{pol} and $T^{(2)}$ which both make-up components of T_B. Alone, these dwell times do not reveal much information about the number of kinetic steps in the pol-process. Here, we examine R_{T_B} to see if any information about the kinetic structure of the pol-process can be obtained. Figure 4.17 shows the reciprocal of R_{T_B} as a function of $[dNTP]$ for various numbers of transient states n in the T_{pol} segment of T_B.

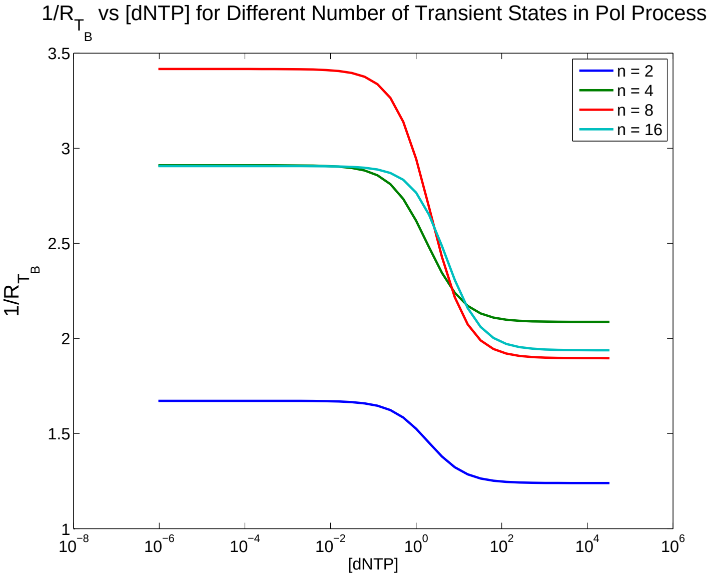

Figure 4.17: The reciprocal of the randomness parameter $1/R_{T_B}$ as a function of $[dNTP]$ for various number of transient states, n. In this simulation, the forward rates, $r_{f,i} = \beta r_{b,i}$ where $\beta = 1.1$ in the T_{pol} segment of T_B. We also set $r_1 = r_3 = r_4 = 1$ in the $T^{(2)}$ segment of $T^{(2)}$.

Here, we see that the randomness parameter of T_B also does not reveal much information about the number of states in the pol-process. For the range of transient states in the T_{pol} segment, n examined, the randomness parameter R_{T_B} changes very little throughout the range of $[dNTP]$ examined. We therefore need a more robust quantity that reveals in a tighter bound, the number of states in the pol-process.

4.3.4 Randomness Parameter of $T^{(1)}$

The randomness parameter of $T^{(1)}$ can be computed in the exact same way as the randomness parameter of T_{pol}. Here, instead we are conditioning on the escape to the pre-translocation state of the current nucleotide addition cycle, given that the Markov process starts in the post-translocation state at time $t = 0$. Figure 4.14 shows the state-space diagram of the Markov process governing $T^{(1)}$ before conditioning. As mentioned previously, the state-space is isomorphic to the one in the birth-death process with two absorbing states introduced in section 4.2. Viewed in this way, $T^{(1)}$ is the same as T_b in the birth-death process.

Using proposition 8, we obtain the conditional infinitesimal generator $Q_{E_{pre}}$ characterizing the Markov process governing $T^{(1)}$ (figure 4.18). The absorption probabilities $\rho_{b,i}$ are complicated functions of all the transition rates and $[dNTP]$.

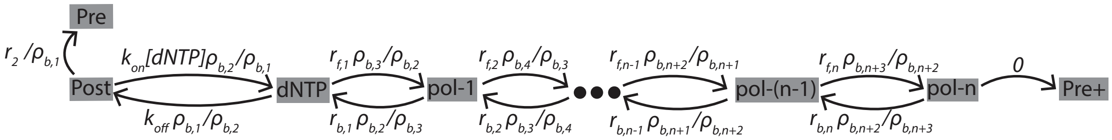

Figure 4.18: State-space diagram of the Markov process governing the escape problem for $T^{(1)}$ after conditioning. Here, we label the states $\{\text{pre}, \text{post}, \text{dNTP}, \text{pol-1}, \ldots, \text{pol-n}, \text{pre+}\}$ as $\{0, 1, 2, 3, \ldots, n+2, n+3\}$, respectively for notational convenience. Here, $\rho_{b,i} := Pr\left(E_{pre} \mid X\left(0\right) = i\right)$.

Like T_{pol}, the PDF and the first-two moments of $T^{(1)}$ are difficult to write down, so we turn to a numerical solution. Using the transformed infinitesimal generator matrix $Q_{\mathcal{T}}$, the PDF of $T^{(1)}$ and its first-two moments can be computed according to equations 4.5, 4.7, and 4.9. Figure 4.19 shows the randomness parameter of $T^{(1)}$ as a function of $[dNTP]$ for various number of transient states n. Unlike the

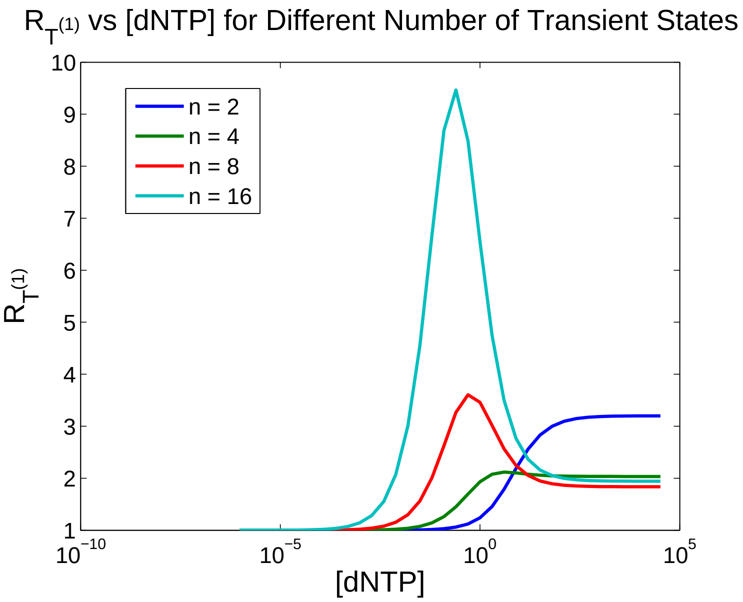

Figure 4.19: The randomness parameter $R_{T^{(1)}}$ as a function of $[dNTP]$ for various number of transient states, n. In this simulation, the forward rates, $r_{f,i} = \beta r_{b,i}$ where $\beta = 1.1$.

randomness parameters of T_{pol} and T_B, the randomness parameter of $T^{(1)}$ reveals quite a bit of information about the number of states in the pol-process. For example, it appears that $\max_{[dNTP]} R_{T^{(1)}}$ increases with n for fixed β.

4.3.5 Conjectures

In this section, we will detail some conjectures that we recently discovered about inferring the number of states in the pol-process. The conjectures are presented in a general setting of a birth-death process with state-space given in figure 4.20.

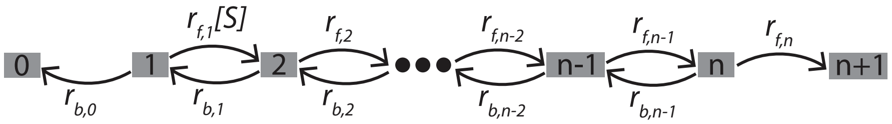

Figure 4.20: A birth-death process with two absorbing boundary states and the first forward transition rate dependent on the substrate concentration $[S]$.

In the previous section, we investigated the randomness parameter of $T^{(1)}$, which is akin to the T_b conditional dwell time introduced in section 4.2. We demonstrated numerically, for fixed β, the maximum of $R_{T^{(1)}}$ increases with n. For the symmetric birth-death process case in section 4.2.1, the ratio of randomness parameters R_{T_b}/R_{T_f} provided a suitable quantity of which the number of transient states n can be inferred. We investigate that quantity in this context.

Consider the quantity $R_{T^{(1)}}/R_{T_{pol}}$. This quantity is the same as R_{T_b}/R_{T_f} from section 4.2.1. Figure 4.21 shows the behavior of $R_{T^{(1)}}/R_{T_{pol}}$ as a function of $[dNTP]$ for various numbers of transient states in the pol-process. Here, we see

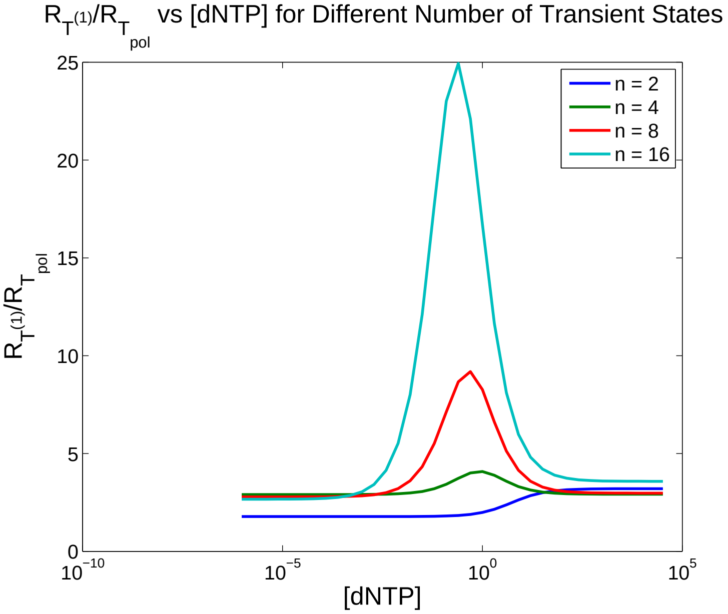

Figure 4.21: The ratio $R_{T^{(1)}}/R_{T_{pol}}$ as a function of $[dNTP]$ for various number of transient states, n. In this simulation, the forward rates, $r_{f,i} = \beta r_{b,i}$ where $\beta = 1.1$ and $r_{b,i} = 1$.

that the ratio $R_{T^{(1)}}/R_{T_{pol}}$ generally increases with n for fixed β. The previous section suggests that some quantity of R_{T_b} and R_{T_f} can be used to obtain the number of transient states n. Indeed, this motivates our first conjecture.

Conjecture 9. *Let $X(t)$ be a birth-death process with two absorbing boundary* *states (figure 4.20); here,* $[S] > 0$ *is a tunable parameter such as* $[dNTP]$*. Suppose that* $r_{f,i} = r_f e^{\epsilon_i}$ *and* $r_{b,i} = r_b e^{\eta_i}$*, where* $|\epsilon|, |\eta| << 1$*. Let* T_b *be the time to absorption to state 0 when* $X(0) = 1$*, and let* T_f *be the time to absorption to state* $n+1$ *when* $X(0) = 1$*; that is,*

$$T_b := \inf \{t \geq 0 \ : \ X(t) \neq 1, \ldots, n\} \ | \ \{E_0, X(0) = 1\},$$
$$T_f := \inf \{t \geq 0 \ : \ X(t) \neq 1, \ldots, n\} \ | \ \{E_{n+1}, X(0) = 1\},$$

where E_0 *and* E_{n+1} *are the escape events defined in 4.27. Then one of the following inequalities hold,*

$$\lim_{[S]\to 0^+} \frac{R_{T_b}}{R_{T_f}} \leq n \leq \lim_{[S]\to\infty} \frac{R_{T_b}}{R_{T_f}}, \ \ \textit{or}$$
$$\max\left\{\lim_{[S]\to 0^+} \frac{R_{T_b}}{R_{T_f}}, \lim_{[S]\to\infty} \frac{R_{T_b}}{R_{T_f}}\right\} \leq n \leq \max_{[S]} \frac{R_{T_b}}{R_{T_f}}.$$

Based on numerical evidence, the inequality appears to hold for r_b and r_f and for ϵ_i and η_i small (for example, see figure 4.22). Also, we note that the bounds are relatively tight for small n (n less than about 5), but very lose for n large. We can give a partial proof of this conjecture for the lower-bounds of the first inequality.

Proof. Partial proof of lower-bounds of conjecture 9: From [20], the reciprocal of the randomness parameter is always less than or equal to the number of states, so $1/R_{T_f} \leq n$, for all $[S] > 0$. Now as $[S] \to 0^+$, $R_{T_b} \to 1^+$ since $\operatorname{var}(T_b) \to \langle T_b \rangle^2$ as $[S] \to 0^+$. Hence $\lim_{[S]\to 0^+} R_{T_b}/R_{T_f} \leq n$. □

The practical implications of this conjecture is profound. If true, this implies that there exist a substrate concentration which provides the exact number of bio-

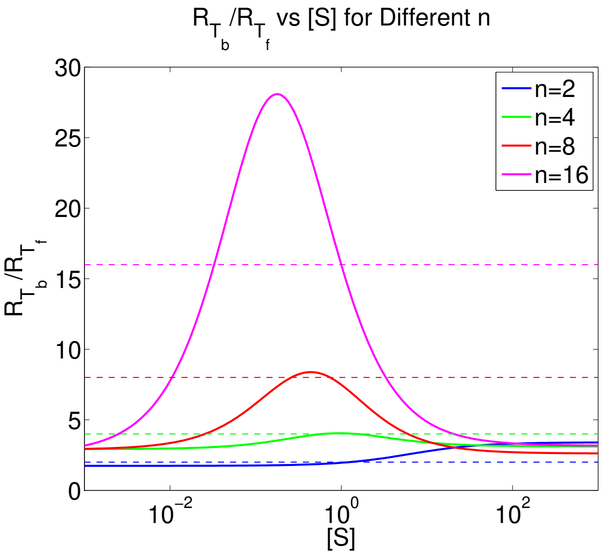

Figure 4.22: Some numerical evidence for conjecture on the bounds of n. Here, $r_{b_i} = 1.1e^{0.05\zeta}$, $r_{f_i} = e^{0.05\zeta}$, where $\zeta \sim N(0, 1)$.

chemical steps in the polymerization process. For n small, the conjecture provides a bounds which is fairly tight. However, the utility of this conjecture decreases rapidly as n increases. Nevertheless, this conjecture can improve upon the lower-bound $1/R_T$ on the number of states, where R_T is a randomness parameter of any dwell time T, proven for any continuous-time discrete-state Markov process [20].

The following conjecture is an obvious corollary to conjecture 9, and is a consequence of the intermediate value theorem.

Conjecture 10. *There exist* $[S]_0 > 0$ *such that*

$$n = \left.\frac{R_{T_b}}{R_{T_f}}\right|_{[S]=[S]_0},$$

where n *is the number of transient states in figure 4.20.*

As alluded to before, the consequence of this conjecture is that there exist at least one value of the substrate concentration $[S]$ that gives the number of states of the polymerization process. Of course, finding these values of $[S]$ may be very difficult. If a constructive proof of conjecture 9 can be made, it is possible that such a proof will illuminate a method to finding such a substrate concentration $[S]$.

4.4 Discussion and Concluding Remarks

The randomness parameter is a function of the first two-moments of the dwell time, and has been shown to be a robust quantity for inferring the number of states. In [20], it has been shown that the reciprocal of the randomness parameter provides a lower-bound on the number of states of any continuous-time, discrete-state Markov chain. This fact has been applied in the inference of the number of kinetic steps in biochemical processes in which the reaction consisted of sequential, irreversible steps ([67], [70], [78] among others). We extended these results to include two absorbing boundary states. Such an extension is motivated by the problem of inferring the number of kinetic states in the DNAP polymerization process from dwell time data. In this situation, the polymerization process is modeled as a birth-death process with two absorbing boundary states.

We first studied an symmetric birth-death process with two absorbing boundary states to give us insight into more general reaction models. The unconditional escape time T was not sufficient in inferring the number of states in a symmetric birth-death process. The number of states can be inferred by the forward escape probability p_{f_1}. This highlights the advantages of conditioning. After the number of states were inferred, the forward and backward kinetic rates can be obtained from R_T, the randomness parameter of T.

For theoretical study, we also looked at the conditional forward and backward escape times T_f and T_b. We derived closed-form analytical expressions for the first-two moments of T_f and T_b, and thus the randomness parameters of T_f and

T_b. We showed that the number of internal states n can be obtained by just the first moment of T_f and T_b. As a theoretical exercise, we also showed that n can be recovered from a ratio of the randomness parameters R_{T_b}/R_{T_f}. After the number of states are inferred, the forward and backward kinetic rate can be obtained from the first moment of T_b. Although redundant and perhaps not useful in practice, the study of the randomness parameter of T_f and T_b illuminated the possible advantages of conditioning on the direction of escape for inferring the number of internal states and the ratio of the forward and backward kinetic rates for birth-death processes with two absorbing states where $r_{f_i} = \beta r_{b,i-i}$ and $r_{b,i} = r_{b,j}$.

For non-symmetric birth-death processes with two absorbing boundary states, the inference of n and β (the ratio of the forward to backward kinetic rates) is difficult to do analytically. We developed two least-squares codes to infer n and β from dwell time data: (1) using p_{f_1} and R_T (algorithm 6); and (2) using p_{f_1}, R_{T_b}, and R_{T_f} (algorithm 7). These codes are referred to as the unconditional least-squares or conditional least-squares codes, respectively. For small values of n, both codes were comparable in inferring n. However, the unconditional least-squares code was more accurate in inferring β, about a 7% decrease in relative error over the conditional least-squares code. For large values of n, the conditional least-squares code was more accurate in inferring n, about a 7% decrease in relative error over the unconditional least-squares code.

We extended the results for non-symmetric birth-death processes to study the inference of the number of kinetic steps of the polymerization process. Naturally, the T_B segment is a good starting point since it fully contains the polymerization process. Since $T_B = T_{pol} + T^{(2)}$, we studied the randomness parameters of T_{pol} and $T^{(2)}$ individually to gain insight into the randomness parameter of T_B. The dwell time $T^{(2)}$ is the same as the T_b dwell time that we defined for the birth-death process with $n = 2$. We examined the behavior of the randomness parameter of $T^{(2)}$ as a function of r_3 and r_4. We showed that the randomness parameter of $T^{(2)}$ can exceed 1, due to branching of the state-space structure. The randomness parameter of T_{pol} was also then examined and was found to not change significantly throughout the range of $[dNTP]$ as n varies. Finally, the randomness parameter of R_{T_B} was studied and it was also found to not change significantly as n increases.

The ratio of randomness parameters R_{T_b}/R_{T_f} studied in the symmetric birth-death process context provided motivation for examining the quantity $R_{T^{(1)}}/R_{T_{pol}}$. Numerical simulations showed that $\max_{[dNTP]} R_{T^{(1)}}/R_{T_{pol}}$ increased as n increased, providing motivation for studying this quantity further. For a more general context, we look at a birth-death process of the form in figure 4.20. In this context, the quantity R_{T_b}/R_{T_f} was examined, akin to $R_{T^{(1)}}/R_{T_{pol}}$. We have strong numerical evidence for a bounds on n based on the quantity R_{T_b}/R_{T_f}, where n is the number of transient states in figure 4.20. We were able to provide a proof of the lower-bounds of one of the inequalities, however a proof or counterexample of the upper-bounds or lower bounds of the other inequality remains an open problem. An obvious corollary of this conjecture is that the number of states in the polymerization process can be obtained from the quantity $R_{T^{(1)}}/R_{T_{pol}}$ at some $[dNTP] > 0$, though a method of finding this dNTP concentration is not known. In any case, the dNTP concentration can be set $[dNTP] = 1$ and the least-squares codes utilizing (p_{f_1}, R_T) and $\left(p_{f_1}, R_{T_b}, R_{T_f}\right)$ from section 4.2.2 can be used to infer n and β in the polymerization process.

Chapter 5

Discussion and Concluding Remarks

In the past couple decades, nanopore experiments have become an important tool to study DNA and DNAPs at the single-molecule level
[2], [5], [21], [26], [48] [19], [18], [49], [50], and [51]. We used the ϕ-29 DNAP as a model system for studying the DNAP since the ϕ-29 can undergo processive replication without the need for any accessory proteins [9]. The nanopore experiments allow us to observe the DNAP translocation step at specified positions along the DNA template and control replication [48], [15], [19], [18], [49], [50], and [51].

The kinetic structure for non-synthesizing DNAP-DNA complexes has been determined in [19], [18], [49], [50], and [51]. In chapter 2, we looked at non-synthesizing complexes and did a complete theoretical study. In [50], the dNTP binding and disassociation rates were inferred by use of the autocorrelation function of the measured ionic current. We complemented this method by showing that the lower-amplitude dwell time, $T^{(1)}$ is a proper mixture of exponential modes. Mixture distributions naturally fit in the expectation-maximization framework for finding the maximum-likelihood estimation. We infer the dNTP binding and disassociation rates by using the EM method.

We extended the results in [50] and completely characterized the uncertainty of the inferred kinetic rates for dNTP binding and disassociation. We found that the inference uncertainty is dependent on scaled versions of k_{off} and $[dNTP]$ (k and S, respectively). Since k is intrinsic to the system, the only tunable experimental parameter that can influence the uncertainty is $[dNTP]$ and hence S. We found that an optimal concentration of dNTP exist that produces the least inference uncertainty for each k. Collecting sufficient amounts of $T^{(1)}$ samples at the optimal dNTP concentration may be difficult in practice. Larger dNTP concentrations decrease the probability of escape to the pre-translocation state, so long experimental run-times may be required to observe a sufficient amount of $T^{(1)}$ samples. To address this, we looked at the constrained optimization problem in which we constrain the experimental run-time to a maximum time; that is, we find the dNTP concentration which produces the least inference uncertainty subject to the constraint that the run-time is no larger than $\tau_{\max}$. The constrained optimization can be solved relatively easily by the mean-field approximation to the constraint. In doing so, for any maximum experimental time $\tau_{\max}$, we can find the optimal $[dNTP]$ that produces the least total relative error such that the experimental run-time is approximately no greater than $\tau_{\max}$.

We also characterized the effect of measurement noise on the observed $T^{(1)}$ samples. We found that for multiplicative noise of the form $\exp(\sigma\zeta)$ where $\zeta \sim N(0,1)$, the difference between the MLE estimates of with $\sigma = 0$ and $\sigma > 0$ is of the form $c_2\sigma^2 + c_1\sigma/\sqrt{n}\zeta$. This means that the introduction of noise increases the bias deterministically by $O(\sigma^2)$ and affects the variance stochastically by an order of $O(\sigma^2/n)$. The coefficients c_1 and c_2 were estimated numerically.

The model in chapter 2 was extended to synthesizing DNAP-DNA complexes in chapter 3. Here, the DNAP-DNA complex was allowed to proceed to the chemical step of phosphodiester bond formation and onto the next nucleotide addition cycle. The DNAP polymerization process is modeled as a single rate limiting step k_{pol}. In this context, we infer the dNTP binding, disassociation, and incorporation (k_{on}, k_{off}, and k_{pol}, respectively). In this context, we have two relevant dwell times: $T^{(1)}$ and T_B which are the time from the arrival to the post-translocation state to the arrival to the pre-translocation state of the current nucleotide addition cycle; and the time from the last arrival to the post-translocation state to the first arrival to the post-translocation state of the next nucleotide addition cycle respectively.

We examined the information content of both $T^{(1)}$ and T_B in regards to inferring k_{on}, k_{off}, and k_{pol}. Throughout the range of $[dNTP]$ examined, we found no advantage of using T_B for the inference; that is, the uncertainty when using $T^{(1)}$ for the inference was lower throughout the range of $[dNTP]$ examined. We derived an equivalent Markov process governing the escape problem for $T^{(1)}$ by conditioning on the arrival to the pre-translocation state when starting at the post-translocation state. In doing this, we showed that the PDF of $T^{(1)}$ is a proper mixture of two exponential modes, and hence the same inference strategy in the non-synthesizing $k_{pol} = 0$ case can be used. After applying scaling laws, we showed that the total relative error is a function of scaled versions of k_{off}, $[dNTP]$, and k_{pol} (k, S, and k_p, respectively); these k and S are the same as in the $k_{pol} = 0$ case. Hence for fixed scaled k_p, the total relative error is a function of the same scaled k and scaled S as in the $k_{pol} = 0$ case. For fixed scaled k_p, if the covariance matrices for each (S, k)-point were saved from the $k_{pol} = 0$ case, the same table can be used to calculate the total relative error of k_{on}, k_{off}, and k_{pol}.

Like the non-synthesizing $k_{pol} = 0$ case, the inference uncertainty was shown to be dependent on k and S for each fixed k_p. However, unlike the $k_{pol} = 0$ case, the total relative error monotonically decreased and approached a horizontal asymptote as $[dNTP] \to \infty$. The result is that there is no well defined optimal $[dNTP]$ which yields the least total relative error. This result is impractical, since at saturating dNTP concentrations, the probability of escape to the pre-translocation state starting from the post-translocation state is approximately 0. Hence experimental times are essentially infinite to obtain a sufficient amount of $T^{(1)}$ samples for inference.

We studied the optimal $[dNTP]$ under two different kinds of constraints: (1) constraining the experimental run-time to a maximum of $\tau_{\max}$; or (2) constraining the number of cycles to a maximum of $\eta_{\max}$. For both constraints, we used the mean-field approximation. Both constraints pull the optimal $[dNTP]$ to a finite region in which there is a clear minimum total relative error. This is because high $[dNTP]$ concentrations limit the number of $T^{(1)}$ samples that can be observed. Intuitively, the optimal $[dNTP]$ under any of these constraints will be a balance between decreasing the total relative error and observing a sufficient number of $T^{(1)}$ samples.

Like the non-synthesizing $k_{pol} = 0$ case, the effect of measurement noise on the observed $T^{(1)}$ samples is completely characterized. The difference between the MLE estimates of k_{on}, k_{off}, and k_{pol} with noise and without is of the form $c_2\sigma^2 + c_1\sigma/\sqrt{n}\zeta$; that is the mean of the difference is deterministic with order $O\left(\sigma^2\right)$ and the variance is stochastic of order $O\left(\sigma^2/n\right)$. The coefficients c_1 and c_2 can be estimated numerically from the data for k_{on}, k_{off}, and k_{pol}.

We looked at further improvements on the inference from $T^{(1)}$ data. After the EM algorithm is used to find the MLE estimates of the mixture parameters for the PDF of $T^{(1)}$, the estimated mixture parameters are mapped to the kinetic rates k_{on}, k_{off}, and k_{pol} directly without taking into account that the probability of escape to the pre-translocation state from the post-translocation state, $p_{E_{pre}|2}$ can be calculated directly from the ionic current observations and hence can be treated as known. The advantages of incorporating knowledge of $p_{E_{pre}|2}$ in the mappings from the mixture parameters to the kinetic rates are apparent for small $[dNTP]$–as much as nearly an order of magnitude decrease in total relative error. We showed that incorporating knowledge of $p_{E_{pre}|2}$ makes sense if the constrained optimal $[dNTP]$ is small, otherwise the advantages of knowing $p_{E_{pre}|2}$ is negligible.

Mentioned briefly in section 3.3.3, the mean cycle time $\langle T_{cycle} \rangle$ can be used to infer the dNTP incorporation rate k_{pol} at saturating $[dNTP]$ with about 8% relative error. After writing the mean cycle time in Michaelis-Menten form, the parameter K_M–the concentration of which half the reaction velocity if obtained–can be used to constrain k_{on} and k_{off} to a line, further improving inference on these kinetic rates.

In chapter 4, we extended the model further by modeling the polymerization process as an unknown number of kinetic steps with the last kinetic step being irreversible. In the context of our model for the DNAP-DNA complex, we define the polymerization process as the kinetic states after dNTP binding and before transition into the pre-translocation of the next nucleotide addition cycle. These states, along with the post-translocation state of the current cycle and dNTP-bound state is modeled as a birth-death process with two absorbing boundary states. Here, the transient states of the birth-death process are the post-translocation, dNTP-bound, and polymerization process states; and the absorbing boundary states are the pre-translocation states of the current and next nucleotide addition cycle (figure 4.20).

To infer the number of kinetic states in the polymerization process, we looked at the randomness parameter of the dwell times and the forward escape probability. The dwell times that we considered was the overall, unconditional escape time, T; the escape time to the forward absorbing state, T_f; and the escape time to the backward absorbing state, T_b. To gain insight into how these dwell times and the forward escape probability can be used to infer the kinetic structure of the polymerization process, we considered two simple cases for the birth-death process model: (1) symmetric birth-death rates and (2) non-symmetric birth-death rates. In the symmetric birth-death process, all the transition rates are equal. In the non-symmetric birth-death process, all of the forward rates are equal and a constant multiple of the backward rates; i.e., $r_{f,i} = \beta r_{b,i-1}$ for all i.

The dwell time T provided the least information about the number of states in the birth-death process. For the symmetric birth-death case, we found that the randomness parameter of T, R_T was not sufficient to determine the number of states when $n \leq 2$. This hinted at the advantages of conditioning on the direction of escape. For the symmetric birth-death case, we found that the forward escape probability, p_{f_1} alone was sufficient enough to determine the number of states in the process. After the number of states has been determined, the unconditional escape time T can then be used to determine the forward and backward rates of the process.

We also looked at the conditional escape time T_f and T_b. Like R_T and p_{f_1}, we were able to derive analytical closed-form expressions for R_{T_f} and R_{T_b}. We found that the quantity $R := R_{T_b}/R_{T_f}$ can be used to determine the number of states in the process. After the number of states has been determined, the first moment of T_b can be used to infer the forward and backward transition rates.

The same quantities: R_T, p_{f_1}, R_{T_f}, and R_{T_b} were also examined for the non-symmetric birth-death process case. In this case, we were able to derive analytical, closed-form expressions for R_T and p_{f_1} but not R. However unlike the symmetric case, inverting the mapping $(n, \beta) \mapsto (p_{f_1}, R_T)$ or $(n, \beta) \mapsto \left(p_{f_1}, R_{T_f}, R_{T_b}\right)$ appears to be analytically intractable. Instead, we developed numerical codes to invert the mappings which utilize least-squares.

With no numerical noise, both least-squares codes inferred β and n exactly from the data. With multiplicative noise introduced into the observed p_{f_1}, R_T, R_{T_f}, and R_{T_b} data, both least-squares codes behaved differently. For small values of n, both codes were comparable in inferring n. However, the unconditional least-squares code was more accurate in inferring β, about a 7% decrease in relative error over the conditional least-squares code. For large values of n, the conditional least-squares code was more accurate in inferring n, about a 7% decrease in relative error over the unconditional least-squares code.

We extended the birth-death process model to a more general case which resembles the DNAP polymerization process. The forward transition from state 1 to state 2 has been replaced with a first-order substrate dependent rate: $r_{f_1}[S]$. States 1 and 2 correspond to the post-translocation and dNTP-bound states in the DNAP-DNA state-space diagram, respectively. In this case, we assumed that the forward rates are comparable to each other, and the backward rates are comparable to each other; i.e., $r_{f,i} = r_f e^{\epsilon_i}$ and $r_{b,i} r_b e^{\eta_i}$ where ϵ_i and η_i are small in magnitude.

Motivated by the simpler birth-death process model, we looked at the quantity $R = R_{T_f}/R_{T_b}$ and found that it increases with n for fixed β. We found that there is strong numerical evidence which puts a lower and upper bound on the number of states in the pol-process for the more general case. The consequence of this conjecture (conjecture 9), if true, is that there exist a substrate concentration which yields the number of states n.

In conclusion, we developed a mathematical framework that: (1) utilizes the unique advantage of nanopore experiments in measuring translocation position with single nucleotide precision and millisecond time resolution; (2) allows MLE inference on kinetic rates based on observed dwell times; (3) formulates the inference uncertainty as a table via scaling law to facilitate efficient computation; and (4) allows the adaptive selecting of $[dNTP]$ in experiments to minimize inference uncertainty. The methodology and analysis we developed can be applied to any single molecule experiment in which dwell time data is available. Lastly, the results and methods for designing optimal experimental conditions presented in this dissertation will motivate more meaningful and informative single molecule measurements.

Appendix A

Appendix for Chapter 3

A.1 Analytical Solution for k_{on}, k_{off} and k_{pol}

In this Appendix section, we write down the solution k_{on}, k_{off}, k_{pol} from the system of equations 3.24-3.26. The solution was computed using a computer algebra solver.

A.1.1 Analytical Solution for k_{on}

For convenience, define the following

$$
\begin{aligned}
\Lambda &= \lambda_1^2 - 2\lambda_1\lambda_2 + \lambda_2^2 \\
A_1 &= \frac{\frac{\lambda_1^2}{2} + \lambda_1\lambda_2 - r_2\lambda_1 + \frac{\lambda_2^2}{2} - r_2\lambda_2}{[dNTP]\,(\lambda_1 + \lambda_2)} \\
A_2 &= 2\alpha\lambda_1^3 + 2\alpha\lambda_2^3 + \frac{\Lambda^{\frac{3}{2}}}{2} + \lambda_1\lambda_2^2 + \lambda_1^2\lambda_2 + \lambda_1^2 r_2 + \lambda_2^2 r_2 - \lambda_1^3 - \lambda_2^2 \\
A_3 &= -\frac{\left(\frac{\lambda_1^2}{2} + \lambda_1\lambda_2 - r_2\lambda_1 + \frac{\lambda_2^2}{2} - r_2\lambda_2\right)\left(2\lambda_1\lambda_2 + 2\alpha\lambda_1^2 + 2\alpha\lambda_2^2 - \lambda_1^2 - \lambda_2^2 - 4\alpha\lambda_1\lambda_2\right)}{\lambda_1 + \lambda_2} \\
A_4 &= -2\lambda_1\lambda_2 r_2 - 2\alpha\lambda_1\lambda_2^2 - 2\alpha\lambda_1^2\lambda_2 - 2\alpha\lambda_1^2 r_2 - 2\alpha\lambda_2^2 r_2 + 4\alpha\lambda_1\lambda_2 r_2.
\end{aligned}
$$

We have that

$$k_{on} = A_1 + \frac{A_2 + A_3 + A_4}{[dNTP]\left(2\lambda_1\lambda_2 + \lambda_1\sqrt{\Lambda} + \lambda_2\sqrt{\Lambda} + 2\alpha\lambda_1^2 + 2\alpha\lambda_2^2 - \lambda_1^2 - \lambda_2^2 - 4\alpha\lambda_1\lambda_2\right)} \tag{A.1}$$

A.1.2 Analytical Solution for k_{off}

For convenience define the following

$$\begin{aligned}
\Lambda &= \lambda_1^2 - 2\lambda_1\lambda_2 + \lambda_2^2 \\
B_1 &= \frac{\lambda_1^4\Lambda}{2} + \frac{\lambda_2^2\Lambda}{2} + \Lambda\Big(2\alpha\lambda_1^5 + 2\alpha\lambda_2^5 + 3\lambda_1\lambda_2^4 + 3\lambda_1^4\lambda_2 - \lambda_1^5 - \lambda_2^5 \\
&\qquad - 2\lambda_1^2\lambda_2^3 - 2\lambda_1^3\lambda_2^2 + 4\alpha\lambda_1^2\lambda_2^3 + 4\alpha\lambda_1^3\lambda_2^2 - 6\alpha\lambda_1\lambda_2^4 - 6\alpha\lambda_1^4\lambda_2\Big) \\
B_2 &= -\frac{\Lambda^3}{2} - 2\lambda_1\lambda_2^5 - 2\lambda_1^5\lambda_2 + \lambda_1^3\Lambda^{\frac{3}{2}} + \lambda_2^3\Lambda^{\frac{3}{2}} + 8\lambda_1^2\lambda_2^4 - 12\lambda_1^3\lambda_2^3 + 8\lambda_1^4\lambda_2^2 - 2\alpha\lambda_1^3\Lambda^{\frac{3}{2}} \\
&\qquad - 2\alpha\lambda_2^3\Lambda^{\frac{3}{2}} - \lambda_1^2\lambda_2^2\Lambda \\
B_3 &= -\lambda_1\lambda_2^2\Lambda^{\frac{3}{2}} - \lambda_1^2\lambda_2\Lambda^{\frac{3}{2}} - 32\alpha\lambda_1^2\lambda_2^4 + 48\alpha\lambda_1^3\lambda_2^3 - 32\alpha\lambda_1^4\lambda_2^2 - 8\alpha^2\lambda_1\lambda_2^5 - 8\alpha^2\lambda_1^5\lambda_2 \\
B_4 &= 32\alpha^2\lambda_1^2\lambda_2^4 - 48\alpha^2\lambda_1^3\lambda_2^3 + 32\alpha^2\lambda_1^4\lambda_2^2 + 8\alpha\lambda_1\lambda_2^5 + 8\alpha\lambda_1^5\lambda_2 \\
&\qquad + 2\alpha\lambda_1\lambda_2^2\Lambda^{\frac{3}{2}} + 2\alpha\lambda_1^2\lambda_2 + 2\alpha\lambda_1^2\lambda_2^2\Lambda^{\frac{3}{2}} \\
C_1 &= 2\lambda_1\lambda_2 + \lambda_1\sqrt{\Lambda} + \lambda_2\sqrt{\Lambda} + 2\alpha\lambda_1^2 + 2\alpha\lambda_2^2 - \lambda_1^2 - \lambda_2^2 - 4\alpha\lambda_1\lambda_2 \\
C_2 &= 4\alpha\lambda_1^3 + 4\alpha\lambda_2^3 + \Lambda^{\frac{3}{2}} + 2\lambda_1\lambda_2^2 + 2\lambda_1^2\lambda_2 + 2\lambda_1^2 r_2 + 2\lambda_2^2 r_2 + \lambda_1^2\sqrt{\Lambda} \\
C_3 &= \lambda_2^2\sqrt{\Lambda} - 2\lambda_1^3 - 2\lambda_2^3 - 4\lambda_1\lambda_2 r_2 + 2\lambda_1\lambda_2\sqrt{\Lambda} \\
&\qquad - 2\lambda_1 r_2\sqrt{\Lambda} - 2\lambda_2 r_2\sqrt{\Lambda} - 4\alpha\lambda_1\lambda_2^2 - 4\alpha\lambda_1^2\lambda_2 - 4\alpha\lambda_1^2 r_2 - 4\alpha\lambda_2^2 r_2 + 8\alpha\lambda_1\lambda_2 r_2.
\end{aligned}$$

The equation for k_{off} is given by

$$k_{off} = \frac{B_1 + B_2 + B_3 + B_4}{C_1\left(C_2 + C_3\right)}. \tag{A.2}$$

A.1.3 Analytical Solution for k_{pol}

For convenience, define the following

$$
\begin{aligned}
\Lambda &= \lambda_1^2 - 2\lambda_1\lambda_2 + \lambda_2^2 \\
D_1 &= r_2\Lambda^{\frac{3}{2}} - 2\lambda_1\lambda_2^3 - 2\lambda_1^3\lambda_2 + 4\lambda_1^2\lambda_2^2 + 2\lambda_1\lambda_2^2\sqrt{\Lambda} + 2\lambda_1^2\lambda_2\sqrt{\Lambda} \\
D_2 &= -\lambda_1^2 r_2\sqrt{\Lambda} - \lambda_2^2 r_2\sqrt{\Lambda} - 8\alpha\lambda_1^2\lambda_2^2 + 4\alpha\lambda_1\lambda_2^3\lambda + 4\alpha\lambda_1^3\lambda_2 - 2\lambda_1\lambda_2 r_2\sqrt{\Lambda} \\
F_1 &= 4\alpha\lambda_1^3 + 4\alpha\lambda_2^3 + \Lambda^{\frac{3}{2}} + 2\lambda_1\lambda_2^2 + 2\lambda_1^2\lambda_2 + 2\lambda_1^2 r_2 + 2\lambda_2^2 r_2 + \lambda_1^2\sqrt{\Lambda} \\
F_2 &= \lambda_2^2\sqrt{\Lambda} - 2\lambda_1^3 - 2\lambda_2^3 - 4\lambda_1\lambda_2 r_2 + 2\lambda_1\lambda_2\sqrt{\Lambda} - 2\lambda_1 r_2\sqrt{\Lambda} - 2\lambda_2 r_2\sqrt{\Lambda} \\
F_3 &= -4\alpha\lambda_1\lambda_2^2 - 4\alpha\lambda_1^2\lambda_2 - 4\alpha\lambda_1^2 r_2 - 4\alpha\lambda_2^2 r_2 + 8\alpha\lambda_1\lambda_2.
\end{aligned}
$$

k_{pol} is given by

$$k_{pol} = \frac{D_1 + D_2}{F_1 + F_2 + F_3} \tag{A.3}$$

Appendix B

Appendix for Chapter 4

B.1 Derivations of the first-two moments of T_b and T_f for a Symmetric Birth-death Process

In this section, we derive the analytical expressions for the first-two moments of T_b and T_f for a symmetric birth-death process. Recall that we can write the first-two moments of T_b and T_f as

$$\left\langle T_b \right\rangle = -\frac{1}{2r} \frac{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}}{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}},$$

$$\left\langle T_b^2 \right\rangle = \frac{1}{2r^2} \frac{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^3}}{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}},$$

$$\left\langle T_f \right\rangle = -\frac{1}{2r} \frac{\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}}{\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}},$$

$$\left\langle T_f^2 \right\rangle = \frac{1}{2r^2} \frac{\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^3}}{\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}}.$$

In this section, we evaluate these finite trigonometric summations analytically.

Lemma 11. $\left\langle T_b \right\rangle = \frac{1}{6r}\left(2n+1\right)$.

Proof.

$$\begin{aligned}
\langle T_b \rangle &= -\frac{1}{2r} \frac{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}}{\sum_{k=1}^{n} \frac{\sin^2\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1}\right)-1}} \\
&= -\frac{1}{2r} \frac{\sum_{k=1}^{n} \frac{\cos\left(\frac{k\pi}{n+1}\right)+1}{\cos\left(\frac{k\pi}{n+1}\right)-1}}{\sum_{k=1}^{n} \left(\cos\left(\frac{k\pi}{n+1}\right)+1\right)} \\
&= \frac{1}{2r} \frac{\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2(n+1)}\right)}{n + \frac{1}{2} + \frac{\sin\left(\frac{2n+1}{2}\frac{\pi}{n+1}\right)}{2\sin\left(\frac{\pi}{2(n+1)}\right)} - 1} \\
&= \frac{1}{2r} \frac{\frac{1}{6}\left(2n+1\right)\left(2n\right)}{n + \frac{1}{2} + \frac{\sin\left(\frac{2n+1}{2}\frac{\pi}{n+1}\right)}{2\sin\left(\frac{\pi}{2(n+1)}\right)} - 1} \\
&= \frac{1}{2r} \frac{\frac{1}{6}\left(2n+1\right)\left(2n\right)}{n + \frac{1}{2} + \frac{1}{2} - 1} \\
&= \frac{1}{6r}\left(2n+1\right),
\end{aligned}$$

where the 3rd and 4th equalities follow from the fact that

$$\sum_{k=0}^{n} \cos\left(kx\right) = \frac{1}{2} + \frac{\sin\left(\frac{2n+1}{2}x\right)}{2\sin\left(\frac{x}{2}\right)},$$

and

$$\sum_{k=1}^{n-1} \cot^2\left(\frac{k\pi}{n}\right) = \frac{\left(n-1\right)\left(n-2\right)}{3}. \tag{B.1}$$

Hence from equation B.1, we have

$$\sum_{k=1}^{\left\lfloor \frac{n-1}{2} \right\rfloor} \cot^2\left(\frac{k\pi}{n}\right) = \frac{\left(n-1\right)\left(n-2\right)}{6}.$$

□

Equation B.1 can be found in tables of series such as [27], [36], and [65].

Various proofs of this series evaluation exists, for example, in [8].

Lemma 12. $\langle T_b^2 \rangle = \frac{1}{r^2}\left(\frac{2n^3}{45} + \frac{8n^2}{45} + \frac{19n}{90} + \frac{1}{15}\right)$.

Proof. From Lemma 11, we see that the denominator of equation 4.14 is one. Thus following a similar strategy for the derivation of the exact polynomial expression for $\langle T_b \rangle$, we obtain

$$\begin{aligned}
\left\langle T_b^2 \right\rangle &= \frac{1}{2r^2}\sum_{k=1}^{n} \frac{\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right) - 1\right)^3} \\
&= -\frac{1}{2r^2}\sum_{k=1}^{n} \frac{\cos\left(\frac{k\pi}{n+1}\right) - 1}{\left(\cos\left(\frac{k\pi}{n+1}\right) - 1\right)} \\
&= -\frac{1}{2r^2}\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right) \frac{1}{\cos\left(\frac{k\pi}{n+1}\right) - 1} \\
&= \frac{1}{2r^2}\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right) \frac{\frac{1}{2}}{\frac{1-\cos\left(\frac{2k\pi}{2(n+1)}\right)}{2}} \\
&= \frac{1}{4r^2}\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right) \csc^2\left(\frac{k\pi}{2\left(n+1\right)}\right) \\
&= \frac{1}{4r^2}\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right) \left(1 + \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right)\right) \\
&= \frac{1}{4r^2}\left[\sum_{k=1}^{n} \cot^2\left(\frac{k\pi}{2\left(n+1\right)}\right) + \sum_{k=1}^{n} \cot^4\left(\frac{k\pi}{2\left(n+1\right)}\right)\right] \\
&= \frac{1}{4r^2}\left[\frac{1}{6}2n\left(2n+1\right) + \frac{8n^4}{45} + \frac{32n^3}{45} + \frac{8n^2}{45} - \frac{n}{45}\right].
\end{aligned}$$

The last equality comes from equation B.1 and from the fact that

$$\sum_{k=1}^{\left\lfloor \frac{n-1}{2} \right\rfloor} \cot^4\left(\frac{k\pi}{n}\right) = \frac{\left(n-1\right)\left(n-2\right)\left(n^2+3n-13\right)}{90},$$

as proven in [8]. Hence after elementary simplification, we get

$$\left\langle T_b^2 \right\rangle = \frac{1}{r^2}\left(\frac{2n^3}{45} + \frac{8n^2}{45} + \frac{19n}{90} + \frac{1}{15}\right).$$

□

Lemma 13. $\langle T_f \rangle = \frac{1}{6r}\left(n^2 + 2n\right)$.

Proof. We start by evaluating the denominator of equation 4.15. By using elementary trigonometric identities,

$$\begin{aligned}
\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\cos\left(\frac{k\pi}{n+1} - 1\right)} &= -\sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{2\left(n+1\right)}\right) \\
&= -\sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right)\left[\cot\left(\frac{k\pi}{n+1}\right) + \csc\left(\frac{k\pi}{n+1}\right)\right] \\
&= -\left[\sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right) + \sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}\right].
\end{aligned}$$

In the last equality above, we show that

$$\sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)} = \begin{cases} 1 & \text{if } n \text{ odd} \\ 0 & \text{if } n \text{ even} \end{cases} \tag{B.2}$$

and

$$\sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right) = \begin{cases} 0 & \text{if } n \text{ odd} \\ 1 & \text{if } n \text{ even} \end{cases} \tag{B.3}$$

To show equation B.2, notice that for k odd, $\sin\left(\frac{nk\pi}{n+1}\right) = \sin\left(\frac{k\pi}{n+1}\right)$. To see this, note that $\sin\left(x\right) = \sin\left(y\right)$ if and only if $x = \alpha\pi + \left(-1\right)^{\alpha} y$ for $\alpha \in \mathbb{Z}$. Now for k odd, $k = 2j+1$. Choose $x = \frac{nk\pi}{n+1} = \frac{2(2j+1)\pi}{n+1}$, $y = \frac{k\pi}{n+1} = \frac{(2j+1)\pi}{n+1}$, and $\alpha = 2j+1$.

Similarly, for k even, we see that $\sin\left(\frac{nk\pi}{n+1}\right) = -\sin\left(\frac{k\pi}{n+1}\right)$. To see this, note that $\sin\left(x\right) = -\sin\left(y\right)$ if any only if $x = \alpha\pi + \left(-1\right)^{\alpha+1} y$ for $\alpha \in \mathbb{Z}$. For $k = 2j$, choose $x = \frac{nk\pi}{n+1} = \frac{n2j\pi}{n+1}$, $y = \frac{k\pi}{n+1} = \frac{2j\pi}{n+1}$, and $\alpha = 2j$. Hence the result in equation B.2 follows.

To show equation B.3, we use Corollary 4.2 in [8]. As in [8], define

$$e_m\left(n,a\right)=\sum_{k=1}^{n-1}\sin\left(\frac{2\pi ak}{n}\right)\cot^m\left(\frac{\pi k}{n}\right),$$

for $m, n, a \in \mathbb{N}$ and $a < n$. For n even, set $a = n/2$, and $m = 1$. Then by Corollary 4.2 in [8], $e_1\left(n+1, n/2\right) = 1$. For the case n is odd, rewrite

$$\sum_{k=1}^{n}\sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right)=\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\cos\left(\frac{k\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}.$$

By the reasoning on establishing equation B.2 above, $\frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}=\left(-1\right)^{k+1}$. Hence

$$\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\cos\left(\frac{k\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}=\sum_{k=1}^{n}\left(-1\right)^{k+1}\cos\left(\frac{k\pi}{n+1}\right).$$

In this case, the above equation is 0 for n odd, which can be argued by symmetry. Hence the denominator of equation 4.15 is -1.

Now we consider the numerator of equation 4.15. By elementary trigonometric identities, we can write

$$
\begin{aligned}
&-\frac{1}{2r}\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^2}\\
&\qquad= -\frac{1}{2r}\sum_{k=1}^{n}\frac{1}{2}\sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{2\left(n+1\right)}\right)\csc^2\left(\frac{k\pi}{2\left(n+1\right)}\right)\\
&\qquad= -\frac{1}{2r}\frac{1}{2}\sum_{k=1}^{n}\left[\sin\left(\frac{nk\pi}{n+1}\right)\csc\left(\frac{k\pi}{n+1}\right)\right.\\
&\qquad\quad+\sin\left(\frac{nk\pi}{n+1}\right)\csc^3\left(\frac{k\pi}{n+1}\right)\\
&\qquad\quad+3\sin\left(\frac{nk\pi}{n+1}\right)\csc^2\left(\frac{k\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right)\\
&\qquad\quad+3\sin\left(\frac{nk\pi}{n+1}\right)\csc\left(\frac{k\pi}{n+1}\right)\cot^2\left(\frac{k\pi}{n+1}\right)\\
&\qquad\quad+\sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right)\\
&\qquad\quad\left.+\sin\left(\frac{nk\pi}{n+1}\right)\cot^3\left(\frac{k\pi}{n+1}\right)\right].
\end{aligned}
$$

For ease of notation, call s_i the i-th term in the summation of the last equality above. We see that

$$
s_1=\sum_{k=1}^{n}\sin\left(\frac{nk\pi}{n+1}\right)\csc\left(\frac{k\pi}{n+1}\right)=\sum_{k=1}^{n}\frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}=\begin{cases}1 & \text{if } n \text{ odd}\\ 0 & \text{if } n \text{ even}\end{cases} \tag{B.4}
$$

from equation B.2.

For s_2, we can write it as

$$
s_2=\sum_{k=1}^{n}\sin\left(\frac{nk\pi}{n+1}\right)\csc^3\left(\frac{k\pi}{n+1}\right)=-\sum_{k=1}^{n}\left(-1\right)^k\csc^2\left(\frac{k\pi}{n+1}\right),
$$

since $\frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)}=1$ if k is odd and -1 if k is even. Now from Corollary 3.2 in [8], we have the result that for n odd,

$$\frac{1}{n+1}\sum_{k=1}^{n}(-1)^k\csc^2\left(\frac{k\pi}{n+1}\right) = -2^3\sum_{\substack{j_0,j_1,j_2\geq 0\\ j_0+j_1+j_2=1}}(n+1)^{2j_0-1}\prod_{r=0}^{2}\left(2^{j_r-1}-1\right)\frac{B_{2j_r}}{(2j_r)!}, \tag{B.5}$$

where B_n is the n-th Bernoulli number. Computing equation B.5 above gives us

$$s_2 = \begin{cases} \frac{n^2+2n+3}{6} & \text{if } n \text{ odd} \\ 0 & \text{if } n \text{ even} \end{cases} \tag{B.6}$$

The even case for s_2 can be argued by symmetry.

The term s_3 can be computed in a similar fashion. The key is, writing it as an alternate series, whose value can be looked up in a table of finite trigonometric sums.

$$\begin{aligned} s_3 &= \sum_{k=1}^{n} 3\sin\left(\frac{nk\pi}{n+1}\right)\csc^2\left(\frac{k\pi}{n+1}\right)\cot\left(\frac{k\pi}{n+1}\right) \\ &= -3\sum_{k=1}^{n}(-1)^k\cos\left(\frac{k\pi}{n+1}\right)\csc^2\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

This alternating cosine and cosecant sum was computed in page 129 of [16] giving us,

$$s_3 = \begin{cases} 0 & \text{if } n \text{ odd} \\ \frac{n^2+2n}{2} & \text{if } n \text{ even} \end{cases} \tag{B.7}$$

where the odd case can be argued by symmetry.

For s_4, we can write

$$\begin{aligned} s_4 &= \sum_{k=1}^{n} 3\sin\left(\frac{nk\pi}{n+1}\right)\csc\left(\frac{k\pi}{n+1}\right)\cot^2\left(\frac{k\pi}{n+1}\right) \\ &= -3\sum_{k=1}^{n}(-1)^k\cot^2\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

Appealing to page 155 in [16], we obtain

$$s_4 = \begin{cases} \frac{n^2+2n-3}{2} & \text{if } n \text{ odd} \\ 0 & \text{if } n \text{ even} \end{cases} \tag{B.8}$$

The even case can be proven by symmetry.

The term s_5 was calculated already in equation B.3 above. Here,

$$s_5 = \begin{cases} 0 & \text{if } n \text{ odd} \\ 1 & \text{if } n \text{ even} \end{cases} \tag{B.9}$$

Finally, the s_6 term can be calculated by rewriting the summation as an alternating cosine and cotagent sum.

$$s_6 = \sum_{k=1}^{n}\sin\left(\frac{nk\pi}{n+1}\right)\cot^3\left(\frac{k\pi}{n+1}\right) = -\sum_{k=1}^{n}(-1)^k\cos\left(\frac{k\pi}{n+1}\right)\cot^2\left(\frac{k\pi}{n+1}\right).$$

This finite sum was evaluted in page 140 of [16]. Hence,

$$s_6 = \begin{cases} 0 & \text{if } n \text{ odd} \\ \frac{n^2+2n-6}{6} & \text{if } n \text{ even} \end{cases} \tag{B.10}$$

Combining the results from equations B.4, B.6, B.7, B.8, B.9, and B.10, we see that the numerator is given by

$$-\frac{1}{4r}\sum_{k=1}^{6} s_k = -\frac{n^2+2n}{6}. \tag{B.11}$$

Thus combining this with that fact that the denominator of equation 4.15 is -1, we obtain that $\langle T_f \rangle = \frac{n^2+2n}{6r}$ as desired. □

Lemma 14. $\left\langle T_f^2 \right\rangle = \frac{1}{8r^2}\left(\frac{14n^4}{45} + \frac{56n^3}{45} + \frac{45n^2}{45} + \frac{4n}{5}\right)$.

Proof. The numerator of $\left\langle T_f^2 \right\rangle$ is given by

$$\begin{aligned}\frac{1}{2r}\sum_{k=1}^{n}&\frac{\sin\left(\frac{nk\pi}{n+1}\right)\sin\left(\frac{k\pi}{n+1}\right)}{\left(\cos\left(\frac{k\pi}{n+1}\right)-1\right)^3}\\ &= -\frac{1}{8r}\sum_{k=1}^{n}\sin\left(\frac{nk\pi}{n+1}\right)\cot\left(\frac{k\pi}{2\,(n+1)}\right)\csc^4\left(\frac{k\pi}{2\,(n+1)}\right)\end{aligned} \tag{B.12}$$

For ease of writing, let $\theta_k = k\pi/\left(n+1\right)$. Equation B.12 can be written as

$$\begin{aligned}-\frac{1}{8r}\sum_{k=1}^{n}\Bigg\{&\sin\left(n\theta_k\right)\Big[\cot^5\left(\theta_k\right)+2\cot^3\left(\theta_k\right)+\cot\left(\theta_k\right)+\csc^5\left(\theta_k\right)+2\csc^3\left(\theta_k\right)\\ &+\csc\left(\theta_k\right)+5\cot^4\left(\theta_k\right)\csc\left(\theta_k\right)+10\cot^3\left(\theta_k\right)\csc^2\left(\theta_k\right)\\ &+10\cot^2\left(\theta_k\right)\csc^3\left(\theta_k\right)+6\cot^2\left(\theta_k\right)\csc\left(\theta_k\right)\\ &+5\cot\left(\theta_k\right)\csc^4\left(\theta_k\right)+6\cot\left(\theta_k\right)\csc^2\left(\theta_k\right)\Big]\Bigg\}\end{aligned}$$

There are twelve terms in the summand, which we call $s_1, \ldots, s_{12}$. These s_k's will be evaluated and the result will be calculated as $-\frac{1}{8r}\left(s_1 + \cdots + s_{12}\right)$.

For s_1, we can write it as an alternating sum of cosine and an even power of cotangent,

$$\begin{aligned} s_1 &= \sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right) \cot^5\left(\frac{k\pi}{n+1}\right) \\ &= -\sum_{k=1}^{n} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \cot^4\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

This has been evaluated in page 140 of [16],

$$s_1 = \begin{cases} 0 \quad \text{if } n \text{ odd} \\ \frac{7(n+1)^4 - 110(n+1)^2 + 463}{360} \quad \text{if } n \text{ even} \end{cases} \tag{B.13}$$

The odd case can be argued by symmetry.

Similar as s_1, for s_2, we can write it as an alternating sum of cosine and and even power of cotangent.

$$\begin{aligned} s_2 &= \sum_{k=1}^{n} 2\sin\left(\frac{nk\pi}{n+1}\right) \cot^3\left(\frac{k\pi}{n+1}\right) \\ &= -2\sum_{k=1}^{n} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \cot^2\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

This was evaluated in page 140 of [16]. Hence

$$s_2 = \begin{cases} 0 \quad \text{if } n \text{ odd} \\ \frac{(n+1)^2 - 7}{3} \quad \text{if } n \text{ even} \end{cases} \tag{B.14}$$

The quantity s_3 is exactly equation B.3.

The term s_4 can be written as an alternating sum of an even power of cosecant,

$$\begin{aligned} s_4 &= \sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right) \csc^5\left(\frac{k\pi}{n+1}\right) \\ &= -\sum_{k=1}^{n} (-1)^k \csc^4\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

This case was evaluated in page 149 of [16]. Hence

$$s_4 = \begin{cases} \frac{7(n+1)^4+40(n+1)^2+88}{360} & \text{if } n \text{ odd} \\ 0 \ \ \text{if } n \text{ even} \end{cases} \tag{B.15}$$

The even case can be argued by symmetry.

The quantity s_5 can be written as an alternating cosecant squared sum,

$$\begin{aligned} s_5 &= \sum_{k=1}^{n} 2\sin\left(\frac{nk\pi}{n+1}\right) \csc^3\left(\frac{k\pi}{n+1}\right) \\ &= -2\sum_{k=1}^{n} (-1)^k \csc^2\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

By page 149 of [16], this becomes

$$s_5 = \begin{cases} \frac{(n+1)^2+2}{3} & \text{if } n \text{ odd} \\ 0 \ \ \text{if } n \text{ even} \end{cases} \tag{B.16}$$

The quantity s_6 is simply an alternating summation,

$$s_6 = \sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right) \csc\left(\frac{k\pi}{n+1}\right) = -\sum_{k=1}^{n} (-1)^k.$$

This is clearly 1 is n is odd and 0 otherwise.

s_7 can be written as an alternating series of an even power of cotangent,

$$\begin{aligned} s_7 &= \sum_{k=1}^{n} 5 \sin\left(\frac{nk\pi}{n+1}\right) \csc\left(\frac{k\pi}{n+1}\right) \cot^4\left(\frac{k\pi}{n+1}\right) \\ &= -5 \sum_{k=1}^{n} (-1)^k \cot^4\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

Again, this summation was evaluated in page 156 of [16]. Hence

$$s_7 = \begin{cases} \frac{(n+3)(n-1)}{72}\left(7\left(n+1\right)^2 - 52\right) & \text{if } n \text{ odd} \\ 0 \;\; \text{if } n \text{ even} \end{cases} \tag{B.17}$$

The term s_8 can be split into two finite sums of alternating cosine-cotangent products,

$$\begin{aligned} s_8 &= \sum_{k=1}^{n} 10 \sin\left(\frac{nk\pi}{n+1}\right) \cot^3\left(\frac{k\pi}{n+1}\right) \csc^2\left(\frac{k\pi}{n+1}\right) \\ &= 10 \sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right) \cot^3\left(\frac{k\pi}{n+1}\right)\left(1 + \cot^2\left(\frac{k\pi}{n+1}\right)\right) \\ &= 10 \sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)} \cos\left(\frac{k\pi}{n+1}\right) \cot^2\left(\frac{k\pi}{n+1}\right) \\ &\quad + 10 \sum_{k=1}^{n} \frac{\sin\left(\frac{nk\pi}{n+1}\right)}{\sin\left(\frac{k\pi}{n+1}\right)} \cos\left(\frac{k\pi}{n+1}\right) \cot^4\left(\frac{k\pi}{n+1}\right) \\ &= -10 \sum_{k=1}^{n} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \cot^2\left(\frac{k\pi}{n+1}\right) \\ &\quad - 10 \sum_{k=1}^{n} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \cot^4\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

General alternating cosine-cotangent summations for even powers of cotagent are given in page 140 of [16]. Hence,

$$s_8 = \begin{cases} 0 \quad \text{if } n \text{ odd} \\ \frac{5}{3}\left((n+1)^2 - 7\right) + \frac{7(n+1)^4 - 110(n+1)^2 + 463}{36} \quad \text{if } n \text{ even} \end{cases} \tag{B.18}$$

The term s_9 can be written as two alternating sums of even powers of cotangents,

$$\begin{aligned} s_9 &= \sum_{k=1}^{n} 10 \sin\left(\frac{nk\pi}{n+1}\right) \cot^2\left(\frac{k\pi}{n+1}\right) \csc^3\left(\frac{k\pi}{n+1}\right) \\ &= -10 \sum_{k=1}^{n} (-1)^k \cot^2\left(\frac{k\pi}{n+1}\right)\left(1 + \cot^2\left(\frac{k\pi}{n+1}\right)\right) \\ &= -10 \sum_{k=1}^{n} (-1)^k \cot^2\left(\frac{k\pi}{n+1}\right) - 10 \sum_{k=1}^{n} (-1)^k \cot^4\left(\frac{k\pi}{n+1}\right). \end{aligned}$$

Again, this is evaluated in page 156 of [16].

$$s_9 = \begin{cases} \frac{5}{3}(n+3)(n-1) + \frac{(n+3)(n-1)\left(7(n+1)^2 - 52\right)}{36} \quad \text{if } n \text{ odd} \\ 0 \quad \text{if } n \text{ even} \end{cases} \tag{B.19}$$

s_{10} can be written as an alternating sum of cotangent, hence

$$\begin{aligned} s_{10} &= \sum_{k=1}^{n} 6 \sin\left(\frac{nk\pi}{n+1}\right) \cot^2\left(\frac{k\pi}{n+1}\right) \csc\left(\frac{k\pi}{n+1}\right) \\ &= -6 \sum_{k=1}^{n} (-1)^k \cot^2\left(\frac{k\pi}{n+1}\right) \end{aligned}$$

Hence,

$$s_{10} = \begin{cases} (n+3)(n-1) \quad \text{if } n \text{ odd} \\ 0 \quad \text{if } n \text{ even} \end{cases} \tag{B.20}$$

For s_{11}, we can write it as an alternating sum of cosine and an even power of cosecant.

$$\begin{aligned}s_{11} &= \sum_{k=1}^{n} 5 \sin\left(\frac{nk\pi}{n+1}\right) \cot\left(\frac{k\pi}{n+1}\right) \csc^4\left(\frac{k\pi}{n+1}\right) \\ &= -5 \sum_{k=1}^{m} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \csc^4\left(\frac{k\pi}{n+1}\right).\end{aligned}$$

This was evaluated in page 129 of [16]. Hence

$$s_{11} = \begin{cases} 0 \;\text{ if } n \text{ odd} \\ \frac{n(n+2)}{72}\left(7\left(n+1\right)^2 + 17\right) \;\text{ if } n \text{ even}\end{cases} \tag{B.21}$$

And finally, s_{12} can also be written as an alternating series of cosine and cosecant.

$$\begin{aligned}s_{12} &= \sum_{k=1}^{n} \sin\left(\frac{nk\pi}{n+1}\right) \cot\left(\frac{k\pi}{n+1}\right) \csc^2\left(\frac{k\pi}{n+1}\right) \\ &= -6 \sum_{k=1}^{n} (-1)^k \cos\left(\frac{k\pi}{n+1}\right) \csc^2\left(\frac{k\pi}{n+1}\right).\end{aligned}$$

Hence,

$$s_{12} = \begin{cases} 0 \;\text{ if } n \text{ odd} \\ n\left(n+2\right) \;\text{ if } n \text{ even}\end{cases} \tag{B.22}$$

Putting together the expressions for $s_1, \ldots, s_{12}$ together from equations B.13-B.22, we see that the numerator of 4.16 is given by

$$-\frac{1}{8r^2}\frac{2n\left(n+2\right)\left(7n^2+14n+9\right)}{45}.$$

Thus,

$$\left\langle T_f^2 \right\rangle = \frac{1}{8r^2}\left(\frac{14n^4}{45} + \frac{56n^3}{45} + \frac{74n^2}{45} + \frac{4n}{5}\right). \tag{B.23}$$

□

Bibliography

[1] Rachid Ait-Haddou and Walter Herzog. Brownian ratchet models of molecular motors. *Cell biochemistry and biophysics*, 38(2):191–213, 2003.

[2] Mark Akeson, Daniel Branton, John J Kasianowicz, Eric Brandin, and David W Deamer. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single rna molecules. *Biophysical journal*, 77(6):3227–3233, 1999.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4 edition, 2002.

[4] R Dean Astumian. Thermodynamics and kinetics of molecular motors. *Biophysical journal*, 98(11):2401–2409, 2010.

[5] Seico Benner, Roger JA Chen, Noah A Wilson, Robin Abu-Shumays, Nicholas Hurt, Kate R Lieberman, David W Deamer, William B Dunbar, and Mark Akeson. Sequence-specific detection of individual dna polymerase complexes in real time using a nanopore. *Nature nanotechnology*, 2(11):718, 2007.

[6] Andrea J Berman, Satwik Kamtekar, Jessica L Goodman, Jose M Lazaro, Miguel de Vega, Luis Blanco, Margarita Salas, and Thomas A Steitz. Structures of phi29 dna polymerase complexed with substrate: the mechanism of translocation in b-family polymerases. *The EMBO journal*, 26(14):3494–3505, 2007.

[7] Oya Bermek, Nigel DF Grindley, and Catherine M Joyce. Prechemistry nucleotide selection checkpoints in the reaction pathway of dna polymerase i and roles of glu710 and tyr766. *Biochemistry*, 52(36):6258–6274, 2013.

[8] Bruce C Berndt and Boon Pin Yeap. Explicit evaluations and reciprocity theorems for finite trigonometric sums. *Advances in Applied Mathematics*, 29(3):358–385, 2002.

[9] Luis Blanco, Antonio Bernad, José M Lázaro, Gil Martín, Cristina Garmendia, and Margarita Salas. Highly efficient dna synthesis by the phage phi 29 dna polymerase. symmetrical mode of dna replication. *Journal of Biological Chemistry*, 264(15):8935–8940, 1989.

[10] Gunter Bolch, Stefan Greiner, Hermann De Meer, and Kishor S Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2006.

[11] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

[12] Carlos Bustamante, David Keller, and George Oster. The physics of molecular motors. *Accounts of chemical research*, 34(6):412–420, 2001.

[13] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[14] Yann R Chemla, Jeffrey R Moffitt, and Carlos Bustamante. Exact solutions for kinetic models of macromolecular dynamics. *The Journal of Physical Chemistry B*, 112(19):6025–6044, 2008.

[15] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of dna in a nanopore at 5-å precision. *Nature biotechnology*, 30(4):344, 2012.

[16] Wenchang Chu and Alberto Marini. Partial fractions and trigonometric identities. *Advances in Applied Mathematics*, 23(2):115–175, 1999.

[17] David N Church, Sarah EW Briggs, Claire Palles, Enric Domingo, Stephen J Kearsey, Jonathon M Grimes, Maggie Gorman, Lynn Martin, Kimberley M Howarth, Shirley V Hodgson, et al. Dna polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. *Human molecular genetics*, 22(14):2820–2828, 2013.

[18] Joseph M Dahl, Ai H Mai, Gerald M Cherf, Nahid N Jetha, Daniel R Garalde, Andre Marziali, Mark Akeson, Hongyun Wang, and Kate R Lieberman. Direct observation of translocation in individual dna polymerase complexes. *Journal of Biological Chemistry*, pages jbc–M111, 2012.

[19] Joseph Michael Dahl. *Single Molecule Studies of DNA Polymerase Fidelity.* PhD thesis, UC Santa Cruz, 2016.

[20] Aldous David and Shepp Larry. The least variable phase type distribution is erlang. *Stochastic Models*, 3(3):467–473, 1987.

[21] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5):518, 2016.

[22] Sylvie Doublie, Stanley Tabor, Alexander M Long, Charles C Richardson, and Tom Ellenberger. Crystal structure of a bacteriophage t7 dna replication complex at 2.2 å resolution. *Nature*, 391(6664):251, 1998.

[23] PS Eder, RY Walder, and JA Walder. Substrate specificity of human rnase h1 and its role in excision repair of ribose residues misincorporated in dna. *Biochimie*, 75(1-2):123–126, 1993.

[24] Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483, 1978.

[25] Matthew C Franklin, Jimin Wang, and Thomas A Steitz. Structure of the replicating complex of a pol α family dna polymerase. *Cell*, 105(5):657–667, 2001.

[26] Daniel R Garalde, Christopher A Simon, Joseph M Dahl, Hongyun Wang, Mark Akeson, and Kate R Lieberman. Distinct complexes of dna polymerase i (klenow fragment) for base and sugar discrimination during nucleotide substrate selection. *Journal of Biological Chemistry*, pages jbc–M111, 2011.

[27] Eldon R Hansen. A table of series and products. *Prentice Hall Series in Automatic Computation, Englewood Cliffs: Prentice Hall, 1975*, 1975.

[28] Ellen Heitzer and Ian Tomlinson. Replicative dna polymerase mutations in cancer. *Current opinion in genetics & development*, 24:107–113, 2014.

[29] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964, 2007.

[30] Matthew Hogg, Pia Osterman, Göran O Bylund, Rais A Ganai, Else-Britt Lundström, A Elisabeth Sauer-Eriksson, and Erik Johansson. Structural basis for processive dna synthesis by yeast dna polymerase eta. *Nature structural & molecular biology*, 21(1):49, 2014.

[31] Erik Johansson and Nicholas Dixon. Replicative dna polymerases. *Cold Spring Harbor perspectives in biology*, 5(6):a012799, 2013.

[32] Erik Johansson and Stuart A MacNeill. The eukaryotic replicative dna polymerases take shape. *Trends in biochemical sciences*, 35(6):339–347, 2010.

[33] Kenneth A Johnson. The kinetic and chemical mechanism of high-fidelity dna polymerases. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(5):1041–1048, 2010.

[34] Robert E Johnson, M Todd Washington, Satya Prakash, and Louise Prakash. Fidelity of human dna polymerase η. *Journal of Biological Chemistry*, 275(11):7447–7450, 2000.

[35] Sean J Johnson, Jeffrey S Taylor, and Lorena S Beese. Processive dna synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proceedings of the National Academy of Sciences*, 100(7):3895–3900, 2003.

[36] Leonard Benjamin William Jolley. *Summation of series*. Courier Corporation, 2012.

[37] Rebecca M Jones and Eva Petermann. Replication fork dynamics and the dna damage response. *Biochemical Journal*, 443(1):13–26, 2012.

[38] Catherine M Joyce. Techniques used to study the dna polymerase reaction pathway. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(5):1032–1040, 2010.

[39] Catherine M Joyce, Olga Potapova, Angela M DeLucia, Xuanwei Huang, Vandana Purohit Basu, and Nigel DF Grindley. Fingers-closing and other rapid conformational changes in dna polymerase i (klenow fragment) and their role in nucleotide selectivity. *Biochemistry*, 47(23):6103–6116, 2008.

[40] Lawrence Kazak, Aurelio Reyes, and Ian J Holt. Minimizing the damage: repair pathways keep mitochondrial dna intact. *Nature reviews Molecular cell biology*, 13(10):659, 2012.

[41] David Keller and Carlos Bustamante. The mechanochemistry of molecular motors. *Biophysical Journal*, 78(2):541–556, 2000.

[42] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. D. Van Nostrand Company, Inc., 1960.

[43] Anatoly B Kolomeisky. Motor proteins and molecular motors: how to operate machines at the nanoscale. *Journal of Physics: Condensed Matter*, 25(46):463101, 2013.

[44] Anatoly B Kolomeisky and Michael E Fisher. Molecular motors: a theorist's perspective. *Annu. Rev. Phys. Chem.*, 58:675–695, 2007.

[45] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*, volume 31. Springer Science & Business Media, 2012.

[46] Thomas A Kunkel. Balancing eukaryotic replication asymmetry with replication fidelity. *Current opinion in chemical biology*, 15(5):620–626, 2011.

[47] Thomas A Kunkel and Dorothy A Erie. Dna mismatch repair. *Annu. Rev. Biochem.*, 74:681–710, 2005.

[48] Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive replication of single dna molecules in a nanopore catalyzed by phi29 dna polymerase. *Journal of the American Chemical Society*, 132(50):17961–17972, 2010.

[49] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Mark Akeson, and Hongyun Wang. Dynamics of the translocation step measured in individual dna polymerase complexes. *Journal of the American Chemical Society*, 134(45):18816–18823, 2012.

[50] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Ashley Cox, Mark Akeson, and Hongyun Wang. Kinetic mechanism of translocation and dntp binding in individual dna polymerase complexes. *Journal of the American Chemical Society*, 135(24):9149–9155, 2013.

[51] Kate R Lieberman, Joseph M Dahl, and Hongyun Wang. Kinetic mechanism at the branchpoint between the dna synthesis and editing pathways in individual dna polymerase complexes. *Journal of the American Chemical Society*, 136(19):7117–7131, 2014.

[52] Thomas Milton Liggett. *Continuous time Markov processes: an introduction*, volume 113. American Mathematical Soc., 2010.

[53] Robyn L Maher, Amy M Branagan, and Scott W Morrical. Coordination of dna replication and recombination activities in the maintenance of genome stability. *Journal of cellular biochemistry*, 112(10):2672–2682, 2011.

[54] Miriam R Menezes and Joann B Sweasy. Mouse models of dna polymerases. *Environmental and molecular mutagenesis*, 53(9):645–665, 2012.

[55] Jeffrey R Moffitt and Carlos Bustamante. Extracting signal from noise: kinetic mechanisms from a michaelis–menten-like expression for enzymatic fluctuations. *The FEBS journal*, 281(2):498–517, 2014.

[56] Jeffrey R Moffitt, Yann R Chemla, and Carlos Bustamante. Methods in statistical kinetics. In *Methods in enzymology*, volume 475, pages 221–257. Elsevier, 2010.

[57] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[58] Silvia Noschese, Lionello Pasquini, and Lothar Reichel. Tridiagonal toeplitz matrices: properties and novel applications. *Numerical linear algebra with applications*, 20(2):302–326, 2013.

[59] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

[60] Felix Olasagasti, Kate R Lieberman, Seico Benner, Gerald M Cherf, Joseph M Dahl, David W Deamer, and Mark Akeson. Replication of individual dna molecules under electronic control using a protein nanopore. *Nature nanotechnology*, 5(11):798, 2010.

[61] Claire Palles, Jean-Baptiste Cazier, Kimberley M Howarth, Enric Domingo, Angela M Jones, Peter Broderick, Zoe Kemp, Sarah L Spain, Estrella Guarino, Israel Salguero, et al. Germline mutations affecting the proofreading domains of pole and pold1 predispose to colorectal adenomas and carcinomas. *Nature genetics*, 45(2):136, 2013.

[62] DSG Pollock. Smoothing with cubic splines, 1993.

[63] Leslie Pray. Dna replication and causes of mutation. *Nature education*, 1(1):214, 2008.

[64] Marc J Prindle and Lawrence A Loeb. Dna polymerase delta in dna replication and genome maintenance. *Environmental and molecular mutagenesis*, 53(9):666–682, 2012.

[65] AP Prudnikov, Yu A Brychkov, and OI Marichev. Integrals and series, vol. 1. *Gordon and Breach Science Publishers*, 14:16, 1986.

[66] Bjorn Rydberg and John Game. Excision of misincorporated ribonucleotides in dna by rnase h (type 2) and fen-1 in cell-free extracts. *Proceedings of the National Academy of Sciences*, 99(26):16654–16659, 2002.

[67] Mark J Schnitzer and SM Block. Statistical kinetics of processive enzymes. In *Cold spring harbor symposia on quantitative biology*, volume 60, pages 793–802. Cold Spring Harbor Laboratory Press, 1995.

[68] Joshua W Shaevitz, Steven M Block, and Mark J Schnitzer. Statistical kinetics of macromolecular dynamics. *Biophysical journal*, 89(4):2277–2285, 2005.

[69] Justin L Sparks, Hyongi Chon, Susana M Cerritelli, Thomas A Kunkel, Erik Johansson, Robert J Crouch, and Peter M Burgers. Rnase h2-initiated ribonucleotide excision repair. *Molecular cell*, 47(6):980–986, 2012.

[70] Karel Svoboda, Partha P Mitra, and Steven M Block. Fluctuation analysis of motor protein movement and single enzyme kinetics. *Proceedings of the National Academy of Sciences*, 91(25):11782–11786, 1994.

[71] Michael K Swan, Robert E Johnson, Louise Prakash, Satya Prakash, and Aneel K Aggarwal. Structural basis of high-fidelity dna synthesis by yeast dna polymerase δ. *Nature structural & molecular biology*, 16(9):979, 2009.

[72] Florin Vaida. Parameter convergence for em and mm algorithms. *Statistica Sinica*, pages 831–840, 2005.

[73] Hongyun Wang. A new derivation of the randomness parameter. *Journal of Mathematical Physics*, 48(10):103301, 2007.

[74] Mina Wang, Shuangluo Xia, Gregor Blaha, Thomas A Steitz, William H Konigsberg, and Jimin Wang. Insights into base selectivity from the 1.8 å resolution structure of an rb69 dna polymerase ternary complex. *Biochemistry*, 50(4):581–590, 2010.

[75] Weina Wang, Eugene Y Wu, Homme W Hellinga, and Lorena S Beese. Structural factors that determine selectivity of a high fidelity dna polymerase for deoxy-, dideoxy-, and ribonucleotides. *Journal of Biological Chemistry*, 287(34):28215–28226, 2012.

[76] Gijs JL Wuite, Steven B Smith, Mark Young, David Keller, and Carlos Bustamante. Single-molecule studies of the effect of template tension on t7 dna polymerase activity. *Nature*, 404(6773):103, 2000.

[77] Jianhua Xing, Hongyun Wang, and George Oster. From continuum fokker-planck models to discrete kinetic models. *Biophysical journal*, 89(3):1551–1563, 2005.

[78] Yajun Zhou and Xiaowei Zhuang. Kinetic analysis of sequential multistep reactions. *The Journal of Physical Chemistry B*, 111(48):13600–13610, 2007.