

Unraveling Information Excited and Corrupted by Noise:

With Applications to Single Molecule Biophysics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sivaraman Rajaganapathy

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Murti Salapaka

February 2023

Acknowledgments

First and foremost, I would like to express my sincerest gratitude to my advisor, Prof. Murti Salapaka for his constant support throughout my Ph.D. journey. He has and continues to provide advice, resources, training, and opportunities to pursue my professional and research goals. I am grateful to him for being a true mentor.

I would like to thank Prof. James Ervasti for the unique opportunity and for enabling work on characterizing muscle proteins. I am thankful to Dr. Jackie McCourt, Dr. Maria Ramirez, Preston McCourt for being great collaborators.

I am very grateful to Prof. James Melbourne, for his patient advice and unique insights on mathematical analysis, career, and life.

I would like to thank my thesis committee members Prof. Andrew Lampeski, Prof. Nicola Elia, Prof. James Ervasti, and Prof. Ju Sun for their encouragement, insightful comments, and questions.

I wish to thank all my old teachers and mentors who inspired me to pursue research.

I wish to thank Dr. Saurav Talukdar who was instrumental in my joining the lab and made my initial years in the lab very exciting. I thank Rachit Shrivastava for the countless discussions as well as his friendship. I am grateful to Dr. Sayan Ghosal for his guidance on the experimental techniques in the field of single molecule science. My time at the lab was made immensely more exciting, thanks to the friendship and company of my current and former labmates, Dr. Sourav Patel, Dr. Govind Saraswat, Vivek Khatana, Soham Chakraborty, Arnab Dey, Dr. Shreyas Bhaban, Mohammed Tuhin Rana, Mishfad Veedu, and Connor Edlund. Special thanks go to Cailong Hua, who has been a pleasure to work with on various projects, old and new.

I would like to thank Esther Steier for her companionship and support during this journey. I wish to express my gratitude to my grandparents, for they toiled so I didn't have to.

Finally, I thank my parents - without their unwavering love and support I could not be where I am today.

Dedication

To my family and friends

Abstract

Single molecule studies enabled by recent nano-interrogation tools, offer unique insights not available to bulk studies constrained by averaging effects, and have revealed mechanisms of many fatal diseases including Alzheimer’s and Duchenne Muscular Dystrophy. However, gleaning insights at the molecular scale poses several challenges due to the extremely small dimensions and forces involved, the inherent stochastic behavior of the molecules, and the limitations of the instrumentation. Tools and algorithms are thus crucial for extracting information from and for building models using the single molecule data.

In the first part of this thesis, we demonstrate one such self-learning change detection algorithm which can operate without any prior information about the locations or the magnitude statistics of the steps. We show how this algorithm with provable guarantees can be adapted for applications such as the single molecule force spectroscopy of muscle proteins. In the next part of the thesis, we present our findings on the nano-mechanical properties of the muscle protein utrophin, enabled by our tools that automate data analysis and provide rigour to the conclusions. In the final part of the thesis, we demonstrate another avenue for intelligent algorithms to push the boundaries of single molecule science: working with limited data and providing measures of confidence on results obtained from non-equilibrium single molecule experiments.

Contents

1	Introduction	1
2	An Iterative Algorithm for Change Detection	5
2.1	Introduction	5
2.2	The step detection problem	8
2.2.1	Model Assumptions for the Detection Problem	8
2.2.2	The Step Detection Algorithm	11
2.3	Convergence of the Step Detection Algorithm	14
2.3.1	Increments in the Joint Conditional Probability	14
2.3.2	Binning Produces the Histogram with the Best Fit	19
2.3.3	SDA Proof of Convergence	21
2.3.4	SDA Formulation with Sensor Dynamics	26
2.4	Results and Comparisons	31
2.4.1	Performance Metrics	31
2.4.2	Empirical Results and Discussion	32
2.4.3	Sample Estimates	38
2.5	Conclusions	39
3	Single Molecule Force Spectroscopy of Muscle Proteins	48
3.1	Single Molecule Force Spectroscopy	48
3.1.1	Protein Pulling Experiments using AFMs	50
3.2	Change Detection for Automatic Data Analysis	51
3.2.1	Effectiveness of Data Analysis: Simulation	53
3.2.2	Effectiveness of Data Analysis: Experiments	54
3.3	Muscular Dystrophy	56
3.4	Dystrophin and Utrophin: Structure and Function	56

3.5	Single Molecule Force Spectroscopy of Utrophin	58
3.5.1	Atomic Force Microscopy Protocol for Utrophin Characterization	58
3.5.2	Data Analysis and Contour Length Filters	59
3.5.3	Kolmogorov-Smirnov Analysis	60
3.6	Utrophin Characterization Results	61
3.7	Conclusion	69
3.8	Supporting Information	71
4	Quantifying Errors in the Jarzynski Estimator	84
4.1	Introduction	84
4.2	The Jarzynski Estimator	87
4.3	Confidence Bounds	89
4.3.1	Via Mean and Range	90
4.3.2	Via the Cumulative Distribution Function	92
4.3.3	Estimating the Range of Work	97
4.4	Methods	99
4.5	Results	104
4.6	Discussion	109
4.7	Supporting Information	111
4.7.1	Time Averaged Potential of a Time-Multiplexed Optical Trap	111
4.7.2	Free Energy Difference for a Particle in a Time-Multiplexed Optical Trap	113
4.7.3	Performance Figures	114
5	Conclusion and Future Scope	121
5.1	Conclusion	121
5.2	Future Scope	122
	Bibliography	124

List of Tables

3.2.1 Titin Parameters: Reported vs SDA estimates 55

List of Figures

1.0.1	Comparison of characteristics of the protein utrophin from bulk and single molecule force spectroscopy. (a) Shows the circular dichroism melt curves of bulk utrophin constructs in solution measured at 1°C intervals by Dr. Jackie McCourt at the Ervasti Lab. Average melting temperatures $\pm$ standard deviation for Utr R1-10, R11-22, R1-22, and R6-15 are 45.7 ± 0.26 °C, 50.5 ± 0.3 °C, 47.0 ± 0.49 °C, and 45.4 ± 1.6 °C, respectively. (b) Mode of the unfolding forces of obtained from single molecule force spectroscopy experiments. The unfolding forces range from 100 pN to 250 pN for the 'brittle' utrophin fragments, Utr R1-10, R6-15, whereas the unfolding range from 100 pN to 600 pN or more for the utrophin fragments Utr R11-22, R1-R22, and the full length utrophin. The span of the forces for the brittle fragments are 150 pN, a factor of 4 lesser compared to the span of 500 pN observed for the stiffening fragments.	4
2.2.1	Underlying generative model assumed for solving the step detection problem. Here, x_t are the estimates of steps that are hidden, y_t the noisy observations of x_t , and $\Theta = \{\theta_1, \dots, \theta_K\}$ are the time invariant parameters of the step size distribution.	11
2.2.2	An upper semi-continuous histogram $\overline{\mathcal{H}}_{\Theta}(u)$	12
2.3.1	Underlying generative model $\mathcal{M}_{\mathcal{G}}^s$ assumed for solving the step detection problem, with the addition of sensor dynamics represented by G . The sequence $\{z_t\}$ is obtained by the convolution of the sequence $\{x_t\}$ with G . Here, y_t are the noisy observations of z_t , and $\Theta = \{\theta_1, \dots, \theta_K\}$ are the time invariant parameters of the step size distribution. . . .	28

2.4.1	Comparative performance with increasing noise magnitude. (a) The SDA has the lowest NRMSE for noise standard deviations up to 0.5, follows the TVR in performance for standard deviations beyond 0.5. (b) The SDA closely follows the VS-HMM algorithm when the performance is measured using F1 scores. The TVR performance is degraded in terms of F1 scores due to a large number of erroneous step fits, despite having small NRMSE. The methods PELT, PELT_KER, BINSEG, and BOTUP show improved performance for higher noise magnitude, while over-fitting the data for smaller noise magnitudes.	34
2.4.2	Comparative performance of the SDA with increasing number of steps available in the signal. (a) All the algorithms benefit from having multiple steps in the signal. The SDA has the lowest NRMSE amongst the algorithms tested. (b) The SDA closely follows the VS-HMM method when the performance is measured with F1 scores. The poor F1 scores of TVR are due to multiple erroneous fits.	35
2.4.3	Comparative performance of the SDA with increasing number samples per step in the signal. (a) A larger number of samples for any given step positively benefits each algorithm. The SDA has the lowest NRMSE. (b) The VS-HMM has ~15% better F1 scores compared to the SDA when the number of samples are small. For more than 200 samples per step, the SDA and VS-HMM have equivalent performance, with F1 scores close to 1.	36
2.4.4	Comparative performance of the SDA with sensor dynamics. A first order transfer function is assumed for the sensor, with unity gain and the time constant increased from 0.1 to 2s. (a) The SDA has a significantly lower NRMSE as it takes into account the sensor dynamics. (b) As expected the SDA has a larger F1 score compared to the other algorithms which do not account for sensor dynamics. F1 scores deteriorates beyond sensor time constants of 0.25s as sensors with large time constants (with respect to step duration) effectively destroy stepping information.	37

2.4.5	Histograms fit by the SDA at the end of each iteration for a sample step signal, with 20 steps of magnitude 1, noise standard deviation of 0.5. Empirically we observe that with each iteration, the histograms generated converge to the true stepping distribution.	38
3.1.1	Schematic of a protein pulling experiment. A polypeptide chain protein is stretched between a substrate and a cantilever tip of an AFM.	50
3.1.2	Measurements in a pulling experiment. The deflection in the cantilever as well as the distance between the tip and substrate are measured, allowing the force versus extension characteristics to be determined.	51
3.2.1	Estimates of the contour length with the reformulated SDA terminated at 20 iterations. The estimates in green were obtained with the persistence length assumed constant, which leads to the expected error in the contour length magnitudes. The estimates recomputed locally to improve accuracy are shown in blue. This is done after estimating the corresponding persistence lengths.	53
3.2.2	Persistence length estimates compared with the actual values when applying the SDA on simulated data.	54
3.2.3	Estimates of the force versus extension curve compared with the true and measured curves from the simulated system.	54
3.2.4	Sample force versus extension curve from pulling experiments on titin. The force estimates from the WLC models fit after each unfolding event are overlaid on the measurements.	55
3.4.1	Schematic of (a) dystrophin and (b) utrophin.	57

3.6.1	AFM extension characteristics of utrophin terminal constructs. a, Schematic of constructs analyzed by AFM. ABD - actin binding domain; DgBD - dystroglycan binding domain; SynBD - syntrophin binding domain; DBBD - dystrobrevin binding domain; NT - N-terminus; CR - cysteine rich domain; CT - C-terminus; ovals - spectrin-like repeats; diamonds - unstructured “hinge” regions. b-d, Force vs extension representative trace curves for Utr NT-R10 (b), Utr R11-CT (c), and full-length utrophin (d). Approach – force on the cantilever as it approaches the substrate; Retract – force on the cantilever as the molecule is extended; Unfolding events – force minima corresponding to domain unfolding. WLC – worm like chain model fit to the force-extension behavior.	65
3.6.2	Unfolding force distributions for utrophin terminal constructs reveal markedly different mechanical behaviors. a-c, Plots of the probability distribution of unfolding force vs the unfolding force for Utr NT-R10 (a), Utr-R11-CT (b), and full-length utrophin (c) collected from 3 batches of 300-500 successful pulls per batch. Here, ‘N’ represents the unfolding event count. For example, the distribution corresponding to N = 4 represents the distribution of the 4th unfolding event. . . .	66
3.6.3	Kolmogorov-Smirnov (KS) test metrics. KS test metrics which compare the distribution of unfolding forces for different unfolding event counts. The color-bars (heat maps) represent the KS metric with a 0 value (green) indicating similarity and value of 1 representing maximum dissimilarity (red). Each of the horizontal and vertical axes represents unfolding event counts from the corresponding molecule. The sub-figures (a), (b), and (c) compare the distributions within the same molecule (self-comparison). The sub-figures (d), (e), and (f) with horizontal and vertical axes from different molecules demonstrate cross-comparisons.	67

3.6.4	Unique mechanical behavior of utrophin halves is maintained upon deletion of terminal domains. a, Schematic of utrophin constructs lacking terminal domains and a 10-repeat construct spanning both N- and C-terminal halves. b-e, Plots of probability distributions vs unfolding force of each unfolding event (N) for Utr R1-10 (b), Utr R11-22 (c), Utr R1-22 (d), and Utr R6-15 (e).	68
3.6.5	Summary of utrophin mechanical behavior. a, Comparison of modes of the unfolding force distributions vs unfolding event count shows brittle vs stiffening behavior of utrophin constructs. Peak (or mode) of the distribution of unfolding forces expressed as a function of the number of domains unfolded (unfolding event count). Data are combined over 3 batches of 300-500 successful pulling experiments per batch for each protein construct and more detailed statistics are provided in Supplemental Figure S11. b, KS metrics for the distributions of unfolding force corresponding to different unfolding event counts from each of the utrophin constructs. The KS metrics are computed by using the force distribution of the first unfolding event of Utr NT-R10 as the reference distribution.	69
3.8.1	AFM statistics of titin I27O (an AFM reference protein with 8 repeats of the Ig 27 domain of human titin). a, Histogram plot of unfolding forces with the mode at 212 pN when pulled at 1 μ m/s with a cantilever spring constant of 6.67 pN/nm. This matches closely with the reported value of 225 pN [1]. b, Histogram plot of contour length increments indicates that the most probable value is 26.4 nm, while 28.4 ± 0.3 nm was the corresponding reported value [1]. c, The peak persistence length was measured to be 304 pm compared to the reported value of 390 ± 70 pm [1]. d, Box plots of the unfolding forces classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. The unfolding forces for titin I27O range from 200 pN to 300 pN (a 50% increase) for unfolding counts of 1 to 8, respectively.	71

3.8.2	AFM statistics of spectrin. a, Unfolding force histogram when pulled at $0.25\ \mu\text{m/s}$ with a cantilever of spring constant $6.88\ \text{pN/nm}$ with a most likely peak value of $40\ \text{pN}$ compared to $28\ \text{pN}$ reported in the literature under similar conditions [2]. b, Contour length increment histogram plot with a peak value of $26.6\ \text{nm}$, consistent with a value of $31.7\ \text{nm}$ from the earlier report [2]. c, Persistence length histogram plot with a peak measured $89\ \text{pm}$. d, Box plots of the unfolding forces classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. The unfolding forces vary from $26\ \text{pN}$ for the first unfolding event to $40\ \text{pN}$ for the sixth (a 54% increase), indicating that the molecule exhibits a mildly stiffening spring behavior.	72
3.8.3	Additional force vs extension curves with WLC fits. Candidate force vs extension curves of titin (a), spectrin (b), Utr NT-R10 (c), Utr R11-CT (d), and full-length utrophin (e). Approach - force on the cantilever as it approaches the substrate; retract - force on the cantilever as the molecule is extended; WLC fit - fit of the worm like chain model, which relates the force exerted on the molecule to its extension; unfolding event - force minima corresponding to domain unfolding.	73
3.8.4	Unfolding force histograms for the seven utrophin constructs. The unfolding forces are reported for a pulling speed of $1\ \mu\text{m/s}$ with cantilever spring constants between $5.04 - 9.06\ \text{pN/nm}$	74
3.8.5	Force extension characteristics of multiple protein pulls. a-c, Example force traces of Utr NT-R10 at high concentrations ($500\ \text{nM}$) with unfolding events exceeding the number of available domains for a single molecule ($11\ \text{domains}$ for Utr NT-R10).	75

3.8.6	Box plots of contour lengths as a function of unfolding count. Box plots of the contour length classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. These figures show contour length increasing linearly with unfolding events, where unfolding events are represented as ‘Ni’, $i = 1, 2, 3 \dots$ and ‘d’ denotes the number of pulling experiments that resulted in ‘Ni’ events. A nearly linear increase in the contour length is expected as the unfolding events occur.	76
3.8.7	The most probable contour length increments range between 23.6 to 41.1 nm. In previous reports, spectrin repeats of 106 amino acids had an average contour length increment of 31.7 nm [2] which gives a predicted contour length increment of 32.6 nm for utrophin repeats of 109 average amino acids.	77
3.8.8	Persistence length histograms for the seven utrophin constructs. The most probable persistence lengths were found to be between 23.4 to 38.2 pm for the utrophin constructs.	78
3.8.9	Heat maps of Kolmogorov-Smirnov (KS) statistics. KS test metrics which compare the distribution of unfolding forces for different unfolding event counts. The color-bars (heat maps) represent the KS metric with a 0 value (green) indicating similarity and value of 1 representing maximum dissimilarity (red).	79
3.8.10	Heat maps of KS test metrics which compare the distribution of unfolding forces for different unfolding event counts.	80

3.8.11	Box plots of unfolding forces as a function of unfolding count. Utr NT-R10, Utr R1-R10, and Utr R6-15 (a, d, and g) can be seen to exhibit brittle behavior, wherein the unfolding force for the 10th unfolding event is comparable to that of the 1st unfolding event. Utr R11-CT and UTR R11-R22 show a stiffening spring behavior, wherein the unfolding forces have a 3 to 5-fold increase for the 10th unfolding event compared to the 1st (b and e). Full-length utrophin and Utr R1-22 show a strongly stiffening spring behavior with 4 to 6-fold increases for the 10th unfolding events compared to their 1st (c and f). The relative frequencies of each unfolding event are represented in Fig. 3.8.13.	81
3.8.12	Peak to peak distance. Distribution of the elongation of the utrophin constructs in between successive unfolding events. Each construct exhibits an asymmetric distribution with single peak (mode). . . .	82
3.8.13	Unfolding events per trace. The probabilities of observing a particular number of events (vertical axes) consistently reduce with the increasing unfolding event counts (horizontal axes). The value Γ inset in the figures represents the probability of observing force vs extension curves where the total number of unfolding events exceeds the expected number of folded domains for the given protein construct. The highest value of Γ observed is 2.8% (a).	83
4.4.1	Schematic of the Spring Mass system	101
4.4.2	Time-Multiplexed Potentials created by optical tweezers. (A) A micro-sphere trapped by a laser beam experiencing a restoring force f towards the focus. (B) The parabolic potential created by the laser trap of width ω . (C) The resultant of a laser beam rapidly multiplexed between two different foci - L_l and L_r and (D) the resultant time averaged potential. Changing the duty ratio $\mathcal{D}$ from 0.5 to 0.7 results in a reshaping of the potential and W work is done on the micro-sphere.	102

4.5.1	A summary of the performance and characteristics of the bounds on the synthetic data. (a) A detailed view of the performance results. The box plot shows (left) showcases the performance of the Combined Bounds at 95% confidence level on work distributions that are truncated normals, with $\langle w_d \rangle / \Delta F = 0.5$ as the number of samples are increased. Each box plot represents one case of the synthetic data validation, wherein each case comprises of a collection of 250 computations of the bounds. An example case ($N = 864$) is shown in more detail on the right. The summarized view of the bounds from the synthetic data study, with varying distributions, varying $\langle w_d \rangle$, and increasing N , is shown in (b). The performance curves have been obtained using 250 computations for each case and interpolating the modes of the data obtained in each case.	107
4.5.2	Comparison of our combined bounds with the GRB method on synthetic data with truncated normal distributions of work. The performance curves have been obtained using 250 computations for each case and interpolating the modes of the data obtained in each case.	108
4.5.3	Distributions of work and the corresponding bounds obtained from the simulated Spring Mass system.	108
4.5.4	(A) Distribution of the work done on a microsphere by changing the shape of the double well optical potential, from an initial duty ratio of 50% to a final duty ratio of 30%. (B) Confidence bounds and the true ΔF applied to the experimental work measurements.	109
4.7.1	Success rates for the bounds on the synthesized data with a known range.	114
4.7.2	Success rates for the GRB bounds on the synthesized data with a known range.	114
4.7.3	Success rates for the bounds on the Spring Mass data.	115
4.7.4	Particle in a Time-Multiplexed Potential: Simulated Results	115
4.7.5	Success rates for the bounds on the simulated data from the optical tweezers system.	116
4.7.6	Success rates for the bounds on the experimental data from the optical tweezers.	116

4.7.7	Performance and characteristics of the bounds on the synthesized data with range estimated using the range of measurements method. . .	117
4.7.8	Success rates for the bounds on the synthesized data with range estimated using the min-max method.	118
4.7.9	Performance and characteristics of the bounds on the synthesized data with range estimated using the empirical Chebyshev inequality. . .	119
4.7.10	Success rates for the bounds on the synthesized data with range estimated using the empirical Chebyshev inequality.	120

Chapter 1

Introduction

The development of instruments such as atomic force microscopes (AFM) and optical tweezers have enabled the study of biological systems at the single molecule scale [3, 4, 5, 6]. These instruments have enabled the resolution of measuring forces in the femto to multiple pico Newton range with spatial resolution at the nanometer scale. Mechanical forces play a critical role in many biological processes, such as transcription, translation, cell locomotion, protein folding and unfolding, and information transfer inside cells [7, 8]. Thus, single molecule experiments enabled by these instruments have the power to unravel the mechanisms of incurable diseases and finding potential cures. For example, characterizing the nano-mechanical properties of shock-absorbing muscle proteins dystrophin and utrophin could lead to an effective therapy for the fatal Duchenne Muscular Dystrophy, a disease that affects 1 in 4000 births [9, 10]. Coordination studies on molecular motor proteins have revealed mutations responsible for pathologies such as several cancers, neurodegenerative diseases, and deafness [11, 12, 13].

Studying biological systems one molecule at a time provides unique insights into their function and behavior, a view not afforded to conventional bulk studies obstructed by averaging effects. This difference is demonstrated in Fig. 1.0.1. In Fig. 1.0.1(a), the muscle protein utrophin and its fragments are denatured using heat and observed in bulk. The graph shows the fraction of the protein's three dimensional structure lost, as its temperature is increased in a solution. We observe that the curves are similar, have a single inflection point (also known as the 'melting' point) and the melting temperatures range from 45.4 °C to 50.5 °C for the different fragments of utrophin. Fig. 1.0.1(b) shows the modes of the forces required for the folded

domains of utrophin to unfold. In this case, a tensile force applied to single molecules of the protein denature the three dimensional structure of the protein instead of heat. In contrast to the thermal melt curves, we see marked difference in the behavior of the different fragments of utrophin. This observation could not have been predicted with the thermal study alone.

Even with advancements in simulations, experimental single molecule studies are irreplaceable, especially for large molecules [14, 15, 16]. Further, the number of single molecule investigations of biological processes is growing exponentially with a reported doubling time of 2.2 years [17, 18]. However, single molecule science has multiple challenges. The behavior of the proteins at the single molecule level is stochastic. For example, the folded domains of proteins unfold randomly, and the probability distribution of unfolding is usually unknown [1, 19]. This complicates their study, since the experiments must be repeated multiple times to capture the probabilistic behavior. In addition, the measurements are corrupted by unavoidable thermal and sensor noise, wherein the magnitude of noise is comparable to that of the signal of interest. Due to abrupt changes in the states of bio-molecules, the measured signals consist of sudden jumps. As the frequency domain contents of these abruptly changing signals and the noise overlap, conventional linear filters fail to be effective. These issues are not limited to the study of muscle proteins but are inherent to investigations of biological systems at the nano-scale in general.

This dissertation aims to develop tools that further single molecule biophysics. The outline of this thesis are summarized as follows:

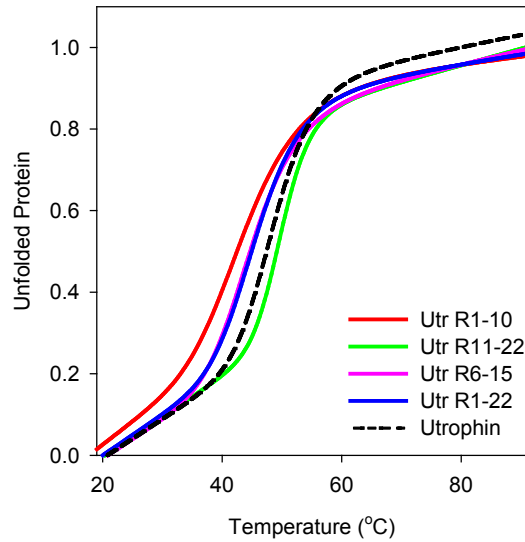
Chapter 2: Multiple domains, including single molecule biophysics, involve systems with abruptly changing states, that result in signals with signatures that are corrupted by noise and sensor dynamics. In many applications, prior information on the magnitudes and timing distributions are unknown, compounding the difficulty of filtering such signals. Several non-linear filtering techniques are available with varying efficacies, however only a few offer theoretical guarantees. Recent research has led to a heuristic, iterative, non-linear filtering technique that is found to be effective over a range of applications. In this chapter, we present this iterative algorithm that learns the distribution of the event magnitude and provides an estimate of when an event has occurred. Here, every iteration involves two stages: the first on providing an estimate of the true signal that uses a prior on the distribution of event magnitudes, followed by the second stage where a distribution of the event magnitudes is estimated

from an estimate of the true signal. It is shown that with every iterate, the performance of the algorithm only improves and convergence of the posterior probability is established. Our comparative tests show that our algorithm provides significant performance improvements, especially for high bandwidth applications where sensor dynamics cannot be neglected. In addition, we provide a python-based implementation of the algorithm and a comprehensive comparison with existing methods in practice. This work has been published in [20].

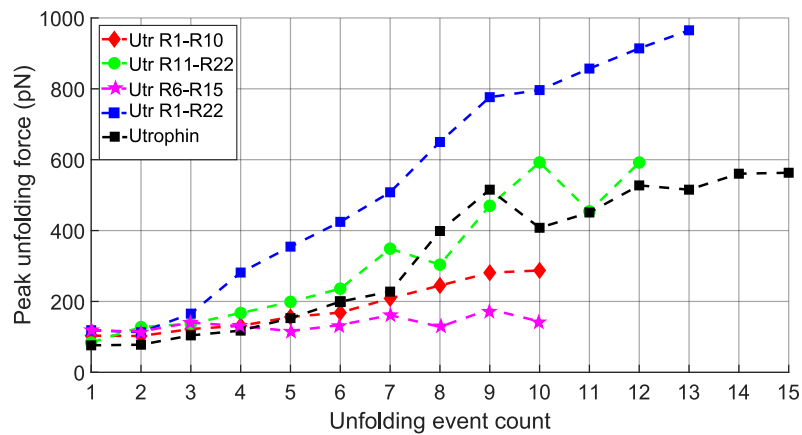
Chapter 3: Duchenne Muscular Dystrophy (DMD) is an incurable, muscle-wasting, genetic disease occurring in 1/4000 male births. Those affected have a life span of fewer than 28 years. Mutation of a shock-absorbing muscle protein dystrophin has been identified as the cause. Utrophin is being actively investigated as a possible replacement for dystrophin. We first expand the application of the change detection algorithm presented in Chapter 2 to estimate the parameters of proteins using data obtained from single molecule force spectroscopy experiments. This estimation technique was presented in [21]. We then present the findings on the nano-mechanical properties of utrophin enabled by automated data analysis and by algorithms to refine the data.

Chapter 4: Free energy is an important parameter in a wide range of domains including the study of systems at the nano-scale. The Jarzynski equality computes the equilibrium free energy difference between two states using the measurements of non-equilibrium work needed to move the system between the states. The Jarzynski equality uses the mean of the exponentiated work and thus holds true only in the limit of infinite data size. There are only a few studies estimating errors when employing the Jarzynski equality with a finite number of measurements; further, these estimates remain limited in scope. We provide non-restrictive and rigorous confidence bounds on the free energy difference computable from samples of measured work without the need for any prior information on the work distribution. These bounds are applicable for a large variety of experimental conditions that take a system far from the equilibrium. The analytical claims are corroborated using simulations and experiments utilizing a toolbox developed in Python.

Chapter 5: We provide a summary of the contributions of this thesis with a review of some directions that can expand the scope of this work.



(a)



(b)

Fig. 1.0.1: Comparison of characteristics of the protein utrophin from bulk and single molecule force spectroscopy. (a) Shows the circular dichroism melt curves of bulk utrophin constructs in solution measured at 1°C intervals by Dr. Jackie McCourt at the Ervasti Lab. Average melting temperatures $\pm$ standard deviation for Utr R1-10, R11-22, R1-22, and R6-15 are $45.7 \pm 0.26^\circ\text{C}$, $50.5 \pm 0.3^\circ\text{C}$, $47.0 \pm 0.49^\circ\text{C}$, and $45.4 \pm 1.6^\circ\text{C}$, respectively. (b) Mode of the unfolding forces of obtained from single molecule force spectroscopy experiments. The unfolding forces range from 100 pN to 250 pN for the 'brittle' utrophin fragments, Utr R1-10, R6-15, whereas the unfolding range from 100 pN to 600 pN or more for the utrophin fragments Utr R11-22, R1-22, and the full length utrophin. The span of the forces for the brittle fragments are 150 pN, a factor of 4 lesser compared to the span of 500 pN observed for the stiffening fragments.

Chapter 2

An Iterative Algorithm for Change Detection

2.1 Introduction

Systems with abrupt changes in their states and parameters arise in a number of diverse domains. The study of such systems thus involves the detection and analysis of signals with abrupt jumps or 'steps'. Examples include detecting DNA copy number variations in bio-informatics [22, 23, 24, 25], failure detection in statistics [26, 27], detecting edges in images (which are essentially a step changes in 2 dimensions) [28, 29], and detecting sudden changes in climate records [30].

Many biological systems at the nano-scale are driven by discontinuous processes [31, 32, 33, 18]. The study of processes at the molecular scale is crucial for understanding the failure mechanisms of many diseases. Here, high resolution instruments such as the atomic force microscope (AFM), optical, and magnetic tweezers that enable the study of biological systems at the single molecule scale are employed. Often the study of biology at the molecular scale depends on discerning the presence and size of abrupt events underlying the process. For instance, the malfunction of a shock absorbing muscle protein dystrophin leads to muscular dystrophy [34]. Muscular dystrophy is a fatal muscle wasting genetic disease, occurring in 1 in 4000 male births [35]. The AFM allows us to study the behavior of single molecules of dystrophin and similar muscle proteins in the nanometer scale, under tensile forces; in the order of piconewtons. Under tension, folded domains of the proteins unfold rapidly, leading to step changes in measured parameters. To illustrate further with another exam-

ple, molecular motor proteins that transport cargo within cells are known to take discrete steps on molecular tracks. It is now known that defects in such motor proteins are linked with serious diseases such as multiple sclerosis, Huntington’s disease, Alzheimer’s disease, and breast cancer [36, 11, 13]. The fact that the motor protein, kinesin, takes discrete 8-nm steps was first determined by analyzing stepping signals from an optical tweezer that tracked an artificial cargo carried by the protein [31]. Questions such as, how mutations affect step sizes, how multiple motors carry a cargo, whether motors coordinate, what is the impact of the cellular environment on motor performance, are of active research interest [37]. Answers to these questions need algorithms that can detect steps with high spatial as well as temporal resolution.

Step signals of interest are rarely noise free. Detecting steps in noisy measurements is a non-trivial problem. Standard linear filters typically work by eliminating noise by removing the high frequency content of the signal. However, the noise and the step signal both have significant power in overlapping ranges in the frequency domain. Consequently linear filters are not well suited for filtering step signals, and often result in the smoothing of the steps or the elimination of short duration steps. The detection problem is further compounded if little or no prior information is available about the underlying generative process such as the distribution of amplitudes and the likely timing of the steps. This is certainly the case when studying new bio-molecules at the single molecule scale. Furthermore, in single molecule investigations, the instruments are operated near the limits of their capabilities, which leads to low signal to noise ratios (SNR) in single molecule data, where the noise magnitudes could be as large as the signals of interest. Similarly, when investigating rapidly changing steps, the data is further corrupted due to the slow instrument dynamics. Subsequently, a number of nonlinear step detection techniques have arisen across multiple disciplines.

A comparative survey of the methods for detecting steps originating from multiple disciplines is presented in [38]. The survey summarizes that most algorithms operate by minimizing an objective function composed of four components: (i) pairwise distances between the measurements and the estimates, (ii) pairwise distances between the estimates, (iii) pairwise distances between the measurements, and (iv) positional distances. While the first component is used for reducing the distance between the estimates and the ground truth, components (ii), (iii), and (iv) are used frequently for regularization. Indeed total variation regularization (TVR) [39], originally proposed for denoising images, in a single dimension, reduces to minimizing the square of the

L_2 distance between the measurements and estimates, along with the L_1 norm of pairwise estimate distances or 'total variation' for regularization.

A selection of step detection tools with a focus on molecular biology applications described in [40, 41, 42, 43] are compared in [33]. The tools in the study were selected from many as they were particularly suited for signals with variable steps sizes and made no assumptions about the generating process. Of these, the χ^2 based tool described in [43] was found to have the best overall performance, and superior temporal resolution. A common feature among most of these algorithms is the presence hyper-parameters that need to be manually tuned for optimal performance. For instance the regularization trade-off weights in [39, 43] and the filter window sizes in [40, 42] are parameters that need tuning. Tuning these hyper-parameters is not trivial and often requires insight in to the algorithms' mechanics. In contrast, the tools in [44, 45, 46] do not require any hyper-parameter tuning. Variable step size hidden Markov model (VS-HMM), described in [45, 46] uses the Baum-Welch algorithm to train a hidden Markov model (HMM) from observed data. This trained HMM is then used by the Viterbi algorithm to generate an estimate of the signal after the HMM parameters are estimated. The VS-HMM has been shown to be robust under low SNR. A more recent survey of offline change detection algorithms is available in [47]. The authors classify the algorithms based on the cost function, the penalty function, and the search strategy used in the algorithms.

The algorithm in [44] is an iterative, dynamic programming based technique that is demonstrated to estimate the stepping probability under low SNRs. In addition, this algorithm can be readily modified to account for instrument dynamics. Its effectiveness has allowed it to be used in a number of biophysics applications such as the detection of wear in molecular tracks [48], characterizing the unfolding rates of the N2B domain of cardiac protein titin [49], the detection of fluorophore events in the study of nanodot arrays [50], and HP1 chromatin binding [51].

Inspired by the algorithm in [44], we develop the step detection algorithm (SDA). The SDA is an iterative algorithm for change detection, where each iteration consists of two stages: a stage where the location and magnitude of steps are determined and the second stage where statistics of the event size is determined from the output of the first stage. The statistics of the second stage is used by the first stage as input in addition to the data. The scheme is entirely data-driven with no prior assumption on event size or timing of events.

The chapter provides theoretical guarantees on convergence with nonrestrictive assumptions by developing a rigorous probabilistic framework through which the operational steps of histogram binning and regularized maximization, can be realized as coordinate ascent with an objective function given by the measurement conditioned posterior probability. Not only does this framework admit a natural proof of the algorithm’s self improvement property, it enables deeper theoretical understandings. We prove the convergence of the measurement conditioned posterior probability and extend the algorithm to accommodate sensor dynamics, where self improvement of estimates is also shown to hold. Finally, we comprehensively compare the empirical performance of our implementation with a selection of methods that includes the TVR, VS-HMM, χ^2 , Linearly Penalized Segmentation (PELT), Gaussian Kernel (KER), Binary-Segmentation (BINSEG), and Bottom-up (BOTUP) techniques described in [39, 45, 43, 52, 53, 54, 55]. In the preliminary conference version [21], we presented a theoretical framework for the SDA and sketched out how this can be used to show improvements in estimates with each iteration.

This paper is organized as follows. In Section 2.2, we formally introduce the step detection problem, the step detection algorithm, and assumptions. In Section 2.3, we show the iterations of the SDA improve and leading to a convergence of the posterior probability. The comparative empirical performance of the SDA is presented in Section 2.4 and discussed in Section 2.5.

2.2 The step detection problem

2.2.1 Model Assumptions for the Detection Problem

We will denote stochastic variables by **bold** characters and their realizations by normal characters. A complete sequence will have T elements and will be represented by capital letters, such as $X = \{x_1, x_2, \dots, x_T\}$. A partial sequence will be marked with the starting and ending indices in the sub and super scripts respectively, where $\{x_r, x_2, \dots, x_s\}$ is denoted by X_r^s . The step detection problem statement follows.

Let $X^\dagger = \{x_1^\dagger, x_2^\dagger, \dots, x_T^\dagger\}$ be the true stepping signal generated by a system with its dynamics given by:

$$\mathbf{x}_t^\dagger = \mathbf{x}_{t-1}^\dagger + \mathbf{u}_t^\dagger, \quad (2.1)$$

with $\mathbf{x}_0^\dagger = 0$, where $\mathbf{u}_t^\dagger$ is a stochastic variable with an unknown distribution $p(\mathbf{u}^\dagger)$,

which is independent of time t . Let $Y = \{y_1, y_2, \dots, y_T\}$ be the observations of $X^\dagger$ corrupted by zero mean, Gaussian noise with,

$$\mathbf{y}_t = \mathbf{x}_t^\dagger + \boldsymbol{\eta}_t, \quad (2.2)$$

where $\{\boldsymbol{\eta}_t\}$ are independent and identically distributed with each $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma^2)$. The step detection problem is to estimate $X^\dagger$ and the density $p(\mathbf{u}^\dagger)$ given the observations Y . We remark that nothing is assumed about the *steps* $\mathbf{u}_t$, except that it admits a probability distribution function which is time-invariant (thus each $\mathbf{u}_t$, for all $t = 1, \dots, T$ are independent and identically distributed). This very generality is needed for many applications and is also the main reason for the challenges posed.

We approach the problem of estimating the stepping signal by assuming that *the probability distribution function $p(u)$ (assumed to be time-invariant) lies in a set of distributions, where each element in the set is determined by a parameter $\Theta \in \mathbb{R}^K$* . We denote an element in the class as $p(u|\Theta)$. As the distribution $p(u)$ is time-invariant (i.e. each $\mathbf{u}_t$, **for all** $t = 1, \dots, T$ being independent and identically distributed), for every Θ , $p(u|\Theta)$ will be assumed to be time-invariant, whereby

$$p(U = \{u_1 \dots u_T\} | \Theta) = \prod_{t=1}^T p(u_t | \Theta).$$

We are assuming there is an underlying true distribution $p(\Theta)$ for the parameter $\Theta \in \mathbb{R}^K$. A generative model $\mathcal{M}_G$ for the stepping dynamics as shown in Fig. 2.2.1. These are the model assumptions, where Fig. 2.2.1 shows the *Bayesian Network*, with vertices and edges, that describes the inter-dependencies of various stochastic variables for the generative model that gives rise to the steps. Here, as alluded to earlier, $\Theta \in \mathbb{R}^K$ is a random vector for $t \in [0, 1, \dots, T]$, Fig. 2.2.1 characterizes the dependencies arising from the structural equations: $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{u}_t$ where $\mathbf{u}_t \sim p(\mathbf{u}|\Theta)$. Thus, Θ and $\mathbf{x}_t$ determine $\mathbf{x}_{t+1}$. Now, $\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\eta}_t$ and thus $\mathbf{x}_t$ determines $\mathbf{y}_t$. The directed arrows in Fig. 2.2.1 go from the right hand side of the equations above to the left hand side variable. In graphical model, with vertices $\mathcal{V}$ and directed edges $\mathcal{E}$, if $(v_1 \rightarrow v_2)$ exists as an edge, then v_1 is a parent of v_2 and v_2 is a child of v_1 . If two vertices, v_i and v_j have a common child v_k , then v_i and v_j are *spouses*. If there is a chain $v_i \rightarrow v_{i+1} \rightarrow \dots \rightarrow v_{i+n} \rightarrow v_j$ then i is an ancestor of v_j and v_j is a descendant of v_i .

We now present the concept of a *Markov Blanket* which will prove to be useful for delineating conditional probabilities.

Definition 2.1. (Markov Blanket [56]) : In a Bayesian Network (such as the one shown in Fig. 2.2.1), the Markov Blanket of a node $\mathcal{A}$ is the smallest set of nodes $\mathcal{S}$ in the graph $\mathcal{G}$ such that $\mathcal{A}$ is independent of $\mathcal{G} - \mathcal{S}$ given $\mathcal{S}$. In a Bayesian network, the Markov blanket of $\mathcal{A}$ constitutes the parents, children, and children's parents of $\mathcal{A}$.

The following theorem will be used to infer a restricted set of inter-dependencies throughout the chapter.

Theorem 2.1. [56] The following holds in a Bayesian Network:

1. Suppose a node v_i has a Markov Blanket M_i . Then $p(v_i|\mathcal{V}\setminus v_i) = p(v_i|M_i)$.
2. Suppose $\mathcal{P}_i$ denotes the parents of a node v_i and $\tilde{\mathcal{D}}_i$ denotes all nodes that are not a descendant of v_i . Then $p(v_i|\{\mathcal{P}_i, \tilde{\mathcal{D}}_i\}) = p(v_i|\mathcal{P}_i)$.

Using Bayes rule, we have $p(X, \Theta|Y) = \frac{p(Y|X, \Theta)p(X|\Theta)p(\Theta)}{p(Y)}$. We note that X forms the Markov blanket for Y , thus $p(Y|X, \Theta) = p(Y|X)$. Therefore

$$p(X, \Theta|Y) = \alpha p(Y|X)p(X|\Theta)p(\Theta),$$

where $\alpha = \frac{1}{p(Y)}$. We now expand $p(Y|X)$ and $p(X|\Theta)$ using Bayes rule.

$$p(X, \Theta|Y) = \alpha \prod_{t=1}^T p(y_t|Y_1^{t-1}, X)p(x_t|X_1^{t-1}, \Theta)p(\Theta).$$

As x_t forms the Markov blanket for y_t , we have $p(y_t|Y_1^{t-1}, X) = p(y_t|x_t)$. Also $p(x_t|X_1^{t-1}, \Theta) = p(x_t|x_{t-1}, \Theta)$ where we have used the fact from Theorem 2.1 that x_t is independent of all non-descendants (X_1^{t-2}) given its parents (x_{t-1}, Θ). Thus we get

$$p(X, \Theta|Y) = \alpha \prod_{t=1}^T p(y_t|x_t)p(x_t|x_{t-1}, \Theta)p(\Theta).$$

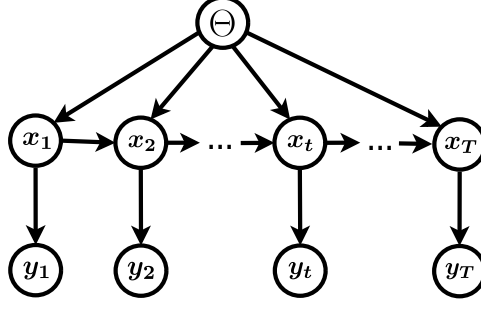


Fig. 2.2.1: Underlying generative model assumed for solving the step detection problem. Here, x_t are the estimates of steps that are hidden, y_t the noisy observations of x_t , and $\Theta = \{\theta_1, \dots, \theta_K\}$ are the time invariant parameters of the step size distribution.

2.2.2 The Step Detection Algorithm

In this section, we now describe the Step Detection Algorithm's (SDA) operational steps in a formal setting. We start by defining a histogram $\mathcal{H}$ over a variable u .

Definition 2.2. (Histogram $\mathcal{H}$) : Consider a bounded variable $u \in [u_{min}, u_{max}]$. For some finite K , select boundary set $\mathcal{B} = \{b_0, b_1, \dots, b_K\}$ such that

$$b_k - b_{k-1} = \frac{|u_{max} - u_{min}|}{K} \text{ for all } k,$$

with $b_0 = u_{min}$, $b_K = u_{max}$. With $r_k := (b_{k-1}, b_k)$, which are the open sets between the boundaries in $\mathcal{B}$, we form the bin ranges (also known as bins) $\bar{\mathcal{R}} = \{\bar{r}_1, \bar{r}_2, \dots, \bar{r}_K\}$ as

$$\bar{r}_k := r_k \cup \{b_{k-1}\}^{\bar{\delta}_{k-1}} \cup \{b_k\}^{\delta_k} \text{ for all } k,$$

where $\delta_k \in \{0, 1\}$ and $\bar{\delta}_k = 1 - \delta_k$ for all k . Note that whether a boundary element b_k belongs to bin range $\bar{r}_k$ is decided by the variable δ_k . Here, $\mathcal{D} = \{\delta_0, \delta_1, \dots, \delta_K\}$ with $\delta_0 = 0$, and $\delta_K = 1$ are the boundary decision parameters, and

$$\{b\}^\delta := \begin{cases} \text{null set} & \delta = 0 \\ \{b\} & \delta = 1. \end{cases}$$

For a fixed $\mathcal{B}$, with boundary decision parameters $\mathcal{D}$, and bin heights $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, $\theta_k \in \mathbb{R}$, the histogram $\mathcal{H}$ over a bounded variable u is defined as

$$\mathcal{H}_{\Theta, \mathcal{D}}(u) = \sum_{k=1}^K \theta_k \mathbb{I}(u \in \bar{r}_k).$$

We now define a upper semi-continuous histogram.

Definition 2.3. (Upper semi-continuous histogram $\bar{\mathcal{H}}$, see Fig. 2.2.2): Given a height set Θ , the corresponding histogram $\mathcal{H}_{\Theta, \mathcal{D}}(u) = \sum_{k=1}^K \theta_k \mathbb{I}(u \in \bar{r}_k)$ is upper semi-continuous if

$$\delta_k = \begin{cases} 1 & \theta_k \geq \theta_{k+1} \\ 0 & \theta_k < \theta_{k+1} \end{cases} \text{ for all } k \neq 0, K,$$

with $\delta_0 = 0$, and $\delta_K = 1$. We denote upper semi-continuous histograms as:

$$\bar{\mathcal{H}}_{\Theta}(u) = \sum_{k=1}^K \theta_k \mathbb{I}(u \in \bar{r}_k).$$

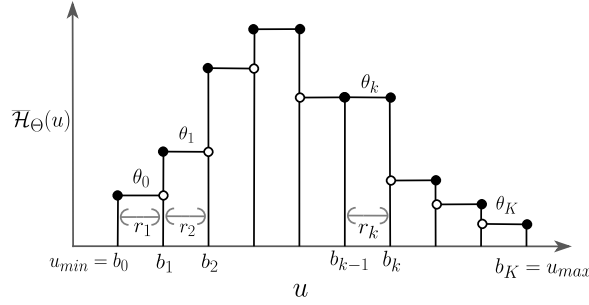


Fig. 2.2.2: An upper semi-continuous histogram $\bar{\mathcal{H}}_{\Theta}(u)$.

Note that a upper semi-continuous histogram $\bar{\mathcal{H}}$, defined over $u \in [u_{min}, u_{max}]$, is parameterized by the bin heights $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, and that definition 2.3 agrees with the usual, more general, definition of upper semi-continuity. For simplicity from this point on, ‘histogram’ will be used to mean the upper semi-continuous histogram $\bar{\mathcal{H}}$ unless otherwise stated. The SDA approximates the underlying distribution $\gamma(\cdot)$ of $\mathbf{u}$ by a histogram $\bar{\mathcal{H}}$ as described in Definition 2.3.

The SDA is composed of two primary stages, namely the minimization stage (M-Stage) followed by the evaluation stage (E-Stage) that at the n^{th} iteration generates an estimate $[\hat{X}^{(n)}, \Theta^{(n)}]$. Let $n \in \{1, 2, \dots, N\}$ represent the iteration count. A description of the two stages are as follows:

M-Stage: For a given $\Theta^{(n-1)}$, minimize the objective function:

$$J(X, \Theta^{(n-1)}) := \sum_{t=1}^T (y_t - x_t)^2 + W_t(X, \Theta^{(n-1)}),$$

with respect to X . Let

$$\hat{X}^{(n)} := \arg \min_X J(X, \Theta^{(n-1)}). \quad (2.3)$$

Here, $W_t(X, \Theta^{(n-1)})$ is a regularization term defined as:

$$W_t(X, \Theta^{(n-1)}) = \begin{cases} 9\sigma^2 \bar{\delta}(x_t - x_{t-1}) & n = 1 \\ -2\sigma^2 \log [p(u_t | \Theta^{(n-1)})] & n > 1 \end{cases} \quad (2.4)$$

with

$$\bar{\delta}(u) = \begin{cases} 0 & u = 0 \\ 1 & \text{otherwise.} \end{cases}$$

Note that, $p(u_t | \Theta^{(n-1)})$ is the probability density function of taking a step $\mathbf{u}_t$ evaluated for the step size u_t , with $u_t = x_t - x_{t-1}$ in iteration n , given $\Theta^{(n-1)}$, the previous estimate of parameters of the step size distribution. For the M-Stage, $\Theta^{(n-1)}$ is assumed to be known.

E-Stage: Consider a given $\hat{X}^{(n)}$. From the given $\hat{X}^{(n)}$, let $u_t^{(n)} := \hat{x}_{t+1}^{(n)} - \hat{x}_t^{(n)}$.

Evaluate $\Theta^{(n)} = \{\theta_1^{(n)}, \theta_2^{(n)}, \dots, \theta_K^{(n)}\}$ and select $\mathcal{D}^{(n)} = \{\delta_0^{(n)}, \delta_1^{(n)}, \dots, \delta_K^{(n)}\}$ with $\delta_0^{(n)} = 0$, and $\delta_K^{(n)} = 1$ using

$$\theta_k^{(n)} = \frac{s_k^{(n)}}{T} \text{ for all } k = 1, \dots, T, \quad (2.5)$$

such that $\Theta^{(n)}$ and $\mathcal{D}^{(n)}$ form an upper semi-continuous histogram.

Here $s_k^{(n)}$ is the number of steps, $u_t^{(n)}$, in the fit $\hat{X}^{(n)}$ contained within the bin $\bar{r}_k$. Thus $s_k^{(n)} = |\{t : u_t^{(n)} \in \bar{r}_k\}|_1$ where $|\Gamma|_1$ is the cardinality of the set Γ . Here T is the total number of observations of steps $u_t^{(n)}$ in the estimate $\hat{X}^{(n)}$, thus $T = \sum_{k=1}^K s_k^{(n)}$

(note that in the computation of T and s_k , $\{t : u_t = 0\}$ are also counted as being steps of zero amplitude).

An alternative to fixing the total number of iterations to N is to evaluate the M and E-Stages $|\hat{X}^n - \hat{X}^{n-1}| \leq \varepsilon_X$ and $|\Theta^n - \Theta^{n-1}| \leq \varepsilon_\Theta$ for some given tolerances ε_X and ε_Θ . Assuming N is known, the SDA is summarized in Algorithm 1.

Algorithm 1 Step Detection Algorithm

Input: Time series measurements $Y = \{y_1, y_2, \dots, y_T\}$.

Output: Step fit $\hat{X}^{(N)} = \{\hat{x}_1^{(N)}, \hat{x}_2^{(N)}, \dots, \hat{x}_T^{(N)}\}$ and Histogram bin parameters $\Theta^{(N)} = \{\theta_1^{(N)}, \theta_2^{(N)}, \dots, \theta_K^{(N)}\}$.

- 1: Set initial penalty using (2.4)
 - 2: Compute step fit $\hat{X}^{(1)}$ using (2.3)
 - 3: **for all** $n \in \{2, 3, \dots, N\}$ **do**
 - 4: **E-Stage:** Compute $\Theta^{(n)}$ using (2.5)
 - 5: **M-Stage:** Compute step fit $\hat{X}^{(n)}$ using (2.3)
 - 6: **end for**
-

2.3 Convergence of the Step Detection Algorithm

In this section, we analyze the convergence properties of the Step Detection Algorithm.

2.3.1 Increments in the Joint Conditional Probability

We now show that each iteration consisting of the M-Stage and E-Stage of the SDA improves upon the previous estimates whose joint conditional probability density $p(\hat{X}, \Theta | Y)$ evaluated at $\hat{X}^{(n)}, \Theta^{(n)}$ improves with each iterate. The SDA uses the knowledge gained from the previous estimates in the M-Stage, in the form of a previously estimated distribution $\overline{\mathcal{H}}_{\Theta^{(n-1)}}(u)$ of the step sizes.

In the E-stage, the step size distribution parameters $\Theta^{(n)}$ are computed from $\hat{X}^{(n)}$ alone, without any direct use of previous estimates $\Theta^{(i)}$ and $\hat{X}^{(i)}$, for $i = 1, \dots, n-1$. In this manner, the estimate $\Theta^{(n)}$ is kept ‘unbiased’, wherein all $\Theta \in \mathbb{R}^K$ that form a valid histogram, are given equal preference for being selected as $\Theta^{(n)}$. On the contrary, utilizing information of the previous estimates in computing $\Theta^{(n)}$, would imply some histograms are favored over others. Due to the SDA’s unbiased computation of $\Theta^{(n)}$ at every stage, we can assume our step size histogram with unknown bin heights

$\Theta = \{\theta_1, \dots, \theta_K\}$, to be uniformly distributed over a measurable set of finite volume. For example, consider the case where the distribution Θ is represented by a histogram with K bins $\{\theta_1, \theta_2, \dots, \theta_K\}$. Then, for any integer K ,

$$\Omega = \{\Theta : \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\},$$

is a compact set of a K dimensional affine subspace $\mathbb{R}^K$, and thus can be endowed with a uniform probability measure. That is, we assume for any two histograms Θ' and Θ''

$$p(\Theta') = p(\Theta'') = \frac{1}{\mathcal{A}_K}.$$

Here, $\mathcal{A}_K$ is the area of the $K - 1$ dimensional hyper plane $\mathcal{S}^{K-1}(\lambda = 1)$, where

$$\mathcal{S}^{K-1}(\lambda) := \{\Theta \in \mathbb{R}^K \mid \sum_{k=1}^K \theta_k = \lambda, \theta_k \geq 0 \text{ for all } k\}.$$

The area $\mathcal{A}_K$ is given by (Proposition 2.3)

$$\mathcal{A}_K = \frac{\sqrt{K}}{(K-1)!}.$$

Lemma 2.1. *Maximizing the probability $p(X, \Theta|Y)$ with respect to X under the assumed Bayesian Network of Fig. 2.2.1 is equivalent to minimizing $J(X, \Theta)$ with respect to X , where*

$$J(X, \Theta) = \sum_{t=1}^T (y_t - x_t)^2 - 2\sigma^2 \log [p(\mathbf{u}_t | X_1^{t-1}, \Theta)].$$

That is, if $\hat{X}^ = \arg \max_X p(X, \Theta|Y)$ then $\hat{X}^* = \arg \max_X J(X, \Theta)$.*

Proof. We observe that by Bayes rule,

$$p(X, \Theta|Y) = \frac{p(Y|X, \Theta)p(X|\Theta)p(\Theta)}{p(Y)}.$$

Since Y is independent of Θ given X (X forms the Markov Blanket [56] for Y , see model in Fig. 2.2.1), $p(Y|X, \Theta) = p(Y|X)$, and since $p(\Theta)$ and $p(Y)$ do not depend

on X ,

$$\begin{aligned}\hat{X}^* &= \arg \max_X p(X, \Theta|Y) \\ &= \arg \max_X p(Y|X)p(X|\Theta).\end{aligned}$$

By repeated application of Bayes Rule, we get

$$\begin{aligned}\hat{X}^* &= \arg \max_X \prod_{t=1}^T p(y_t|Y_1^{t-1}, X)p(x_t|X_1^{t-1}, \Theta) \\ &= \arg \max_X \prod_{t=1}^T p(\mathbf{y}_t = y_t|Y_1^{t-1}, X)p(\mathbf{x}_t = x_t|X_1^{t-1}, \Theta) \\ &= \arg \max_X \prod_{t=1}^T p(\mathbf{y}_t = y_t|Y_1^{t-1}, X)p(\mathbf{x}_t = x_t|X_1^{t-1}, \Theta).\end{aligned}$$

Using $\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\eta}_t$, we observe that $\mathbf{y}_t \sim \mathcal{N}(x_t, \sigma^2)$. We have,

$$\begin{aligned}\hat{X}^* &= \arg \max_X \prod_{t=1}^T \frac{\exp\left\{-\frac{(y_t - x_t)^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} p(x_t|X_1^{t-1}, \Theta) \\ &= \arg \max_X \prod_{t=1}^T \exp\left\{-\frac{(y_t - x_t)^2}{2\sigma^2}\right\} p(x_t|X_1^{t-1}, \Theta) \\ &= \arg \min_X \sum_{t=1}^T \frac{(y_t - x_t)^2}{2\sigma^2} - \log p(\mathbf{u}_t = u_t|X_1^{t-1}, \Theta),\end{aligned}$$

with, $u_t = x_t - x_{t-1}$. Thus,

$$\begin{aligned}\hat{X}^* &= \arg \max_X p(X, \Theta|Y) \\ &= \arg \min_X \sum_{t=1}^T (y_t - x_t)^2 - 2\sigma^2 \log p(\mathbf{u}_t|X_1^{t-1}, \Theta).\end{aligned}$$

This completes the proof. □

Lemma 2.2. *For Θ uniformly distributed with $p(X, \Theta|Y)$ and $p(U|\Theta)$ governed by the Bayesian Network of Fig. 2.2.1,*

$$\Theta^* = \arg \max_{\Theta} p(X, \Theta|Y) = \arg \max_{\Theta} p(U|\Theta).$$

Proof. Using the definition of conditional probability, we note

$$\begin{aligned}
\Theta^* &= \arg \max_{\Theta} p(X, \Theta|Y) \\
&= \arg \max_{\Theta} p(\Theta|X, Y) \times p(X|Y) \\
&= \arg \max_{\Theta} p(\Theta|X, Y).
\end{aligned}$$

However, according to our model structure, Θ is independent of Y given X (X forms the Markov Blanket [56] for Θ), thus

$$\begin{aligned}
\Theta^* &= \arg \max_{\Theta} p(\Theta|X) \\
&= \arg \max_{\Theta} \frac{p(\Theta, X)}{p(X)} \\
&= \arg \max_{\Theta} p(\Theta, X) \\
&= \arg \max_{\Theta} p(X|\Theta) \times p(\Theta),
\end{aligned}$$

but $p(\Theta)$ is assumed to be uniform and thus does not depend on Θ . Therefore,

$$\Theta^* = \arg \max_{\Theta} p(X|\Theta).$$

We use Bayes rule to write

$$\begin{aligned}
p(X|\Theta) &= \prod_{t=1}^T p(\mathbf{x}_t = x_t | X_1^{t-1}, \Theta) \\
&= \prod_{t=1}^T p(\mathbf{u}_t = x_t - x_{t-1} | x_{t-1}, \Theta) \\
&= \prod_{t=1}^T p(\mathbf{u}_t = u_t | \Theta).
\end{aligned}$$

Thus

$$p(X|\Theta) = p(U|\Theta). \tag{2.1}$$

Therefore,

$$\arg \max_{\Theta} p(X, \Theta|Y) = \arg \max_{\Theta} p(U|\Theta).$$

□

Now we consolidate the results to establish that SDA improves the posterior probability with every iterate.

Theorem 2.2. *For Θ uniformly distributed, let the observations Y be corrupted by zero mean Gaussian noise with known variance σ^2 . That is, $\mathbf{y}_t = \hat{\mathbf{x}}_t + \boldsymbol{\eta}_t$, with $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma^2)$ for all t . Then, given an estimate $[\hat{X}^{(n-1)}, \Theta^{(n-1)}]$, the step detection algorithm's iteration produces an improved estimate $[\hat{X}^{(n)}, \Theta^{(n)}]$ whereby*

$$p(\hat{X}^{(n)}, \Theta^{(n)}|Y) \geq p(\hat{X}^{(n-1)}, \Theta^{(n-1)}|Y).$$

Proof. From Lemma 2.1, we note that the maximization of $p(X, \Theta|Y)$ with respect to X is equivalent to a minimization of $J(X, \Theta)$ with respect to X .

Now, in the **M-Stage** of SDA, we have $\Theta = \Theta^{(n-1)}$ and minimize $J(X, \Theta^{(n-1)})$ with respect to X to give $\hat{X}^{(n)}$. This is equivalent to maximizing $p(X, \Theta^{(n-1)}|Y)$ with respect to X . Therefore

$$p(X^{(n)}, \Theta^{(n-1)}|Y) \geq p(X, \Theta^{(n-1)}|Y) \text{ for all } X. \quad (2.2)$$

In the **E-Stage** of SDA, we compute the bin heights of the upper semi-continuous histogram as described in (2.5), which represents the step size distribution, using the estimate $\hat{X}^{(n)}$ available at stage n . In Lemma (2.3), we will show that this binning process is equivalent to maximizing $p(\hat{X}^{(n)}|\Theta)$ with respect to Θ to give $\Theta^{(n)}$. From Lemma 2.2, we observe that

$$\Theta^{(n)} = \arg \max_{\Theta} p(\hat{X}^{(n)}|\Theta) = \arg \max_{\Theta} p(\hat{X}^{(n)}, \Theta|Y).$$

Therefore,

$$p(\hat{X}^{(n)}, \Theta^{(n)}|Y) \geq p(\hat{X}^{(n)}, \Theta|Y) \quad \forall \Theta. \quad (2.3)$$

Consequently, from (2.2) and (2.3), we see that each iteration results in

$$\begin{aligned} p(\hat{X}^{(n)}, \Theta^{(n)}|Y) &\geq p(\hat{X}^{(n)}, \Theta^{(n-1)}|Y) \\ &\geq p(\hat{X}^{(n-1)}, \Theta^{(n-1)}|Y). \end{aligned}$$

This completes the proof. □

Remark: In the implementation, we compute a histogram of the step sizes U , with bin heights $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. In Lemma 2.3 we show that building such a histogram is equivalent to maximizing $p(U^{(n)}|\Theta)$ with respect to Θ , to give $\Theta^{(n)}$.

We briefly compare the similar features of the two algorithms as follows. From Lemma 2.1, Theorem 2.2, and 2.4, we observe that the SDA performs coordinate ascent on $p(X, \Theta|Y)$. With the assumption that $p(\Theta)$ is unbiased (i.e. a constant), and the observation that both $p(\Theta)$ and $p(Y)$ are independent of the values X and Θ , we have

$$\begin{aligned} \arg \max_{X \text{ or } \Theta} p(X, \Theta|Y) &= \arg \max_{X \text{ or } \Theta} p(X, \Theta|Y) p(Y) \\ &= \arg \max_{X \text{ or } \Theta} p(X, Y|\Theta) p(\Theta) \\ &= \arg \max_{X \text{ or } \Theta} p(X, Y|\Theta). \end{aligned}$$

Thus, the SDA performs a coordinate ascent on $p(X, Y|\Theta)$.

The EM algorithm maximizes the function $Q(f(X), \Theta)$, where $Q(f(X), \Theta) := \sum_{X \in \mathcal{X}} f(X) \left\{ \ln \frac{p(X, Y|\Theta)}{f(X)} \right\}$, and $f(X)$ is a probability density function of X . Specifically, the EM algorithm achieves this via coordinate ascent in two steps. In the E-step, the EM algorithm fixes $\Theta = \Theta^{n-1}$ and maximizes $Q(f(X), \Theta)$ along $f(X)$ by the assignment $f(X) = p(X|Y, \Theta^{n-1})$. In the M-step, the EM algorithm fixes $f(X) = p(X|Y, \Theta^{n-1})$, and maximizes $Q(f(X), \Theta)$ along Θ [57]. Thus, the SDA's coordinate maximization of $p(X^{n-1}, Y|\Theta)$ along Θ can be thought of as the coordinate maximization of $Q(f(X), \Theta)$ along Θ by the EM algorithm, under the substitution $f(X) = \delta(X - X^{n-1})$.

2.3.2 Binning Produces the Histogram with the Best Fit

In Lemma 2.3 we prove our earlier claim that building a histogram of the steps U in the estimate $\hat{X}$ in the E-Stage is equivalent to maximizing the probability $p(U|\Theta)$ with respect to the histogram bin heights Θ and thus provides the “best fit.”

Lemma 2.3. *Let $U = \{u_1, u_2, \dots, u_T\}$ be the observations of a stationary stochastic variable. Let $\overline{\mathcal{H}}$ be an upper semi-continuous histogram over $\mathbf{u}$ as described in Definition 2.3, approximating the distribution of U . Let the bin heights be $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. Let s_k be the number of observations u_t contained within the*

bin $\bar{r}_k$, that is $s_k(U, \Theta) = |\{t : u_t \in \bar{r}_k\}|$. Then, the local maxima of $\Theta \mapsto p(U|\Theta)$ satisfy $\theta_k = \frac{s_k(U, \Theta)}{T}$. Local maxima are global maxima iff they have minimum entropy among histograms satisfying $\theta_k = \frac{s_k(U, \Theta)}{T}$.

Proof. Let

$$\Theta^* := \arg \max_{\Theta} p(U|\Theta).$$

Here each $\mathbf{u}_t$ are independent and identically distributed. We then approximate the density $p(u_t|\Theta)$ by the histogram $\bar{\mathcal{H}}_{\Theta}(u_t)$, thus

$$\begin{aligned} \Theta^* &= \arg \max_{\Theta} \prod_{t=1}^T p(u_t|\Theta) = \arg \max_{\Theta} \prod_{t=1}^T \bar{\mathcal{H}}_{\Theta}(u_t) \\ &= \arg \max_{\Theta} \prod_{t=1}^T \sum_{k=1}^K \theta_k \mathbb{I}(u_t \in \bar{r}_k) = \arg \max_{\Theta} \prod_{k=1}^K \theta_k^{s_k} \\ &= \arg \max_{\Theta} \sum_{k=1}^K s_k \log [\theta_k]. \end{aligned}$$

Therefore

$$\Theta^* = \arg \max_{\Theta} p(U|\Theta) = \arg \max_{\Theta} \sum_{k=1}^K s_k \log [\theta_k]. \quad (2.4)$$

We now optimize $f(\Theta) = \sum_{k=1}^K s_k \log [\theta_k]$ with respect to bin heights Θ , subject to the equality constraint $g(\Theta) = \sum_{k=1}^K \theta_k - \frac{K}{u_{\max} - u_{\min}} = 0$. We form the Lagrangian to solve for this constrained optimum. We show in Lemma 2.7 that despite $s_k(U, \Theta)$ being a discontinuous function with respect to Θ , $f(\Theta)$ is continuous with respect to Θ . Further we show in Lemma 2.8, that the points at which $f(\Theta)$ is non-differentiable cannot be its maximizers and that at $\nabla_{\Theta} f(\Theta)$ exists near any maximizer Θ^* . With that, we form the Lagrangian:

$$\mathcal{Q}(\Theta, \lambda) := f(\Theta) - \lambda g(\Theta).$$

The necessary condition for optimality gives

$$\nabla \mathcal{Q}(\Theta^*, \lambda^*) = \bar{\mathbf{0}}.$$

This is equivalent to

$$\left[\frac{s_1}{\theta_1^*} - \lambda^*, \quad \dots, \quad \frac{s_K}{\theta_K^*} - \lambda^*, \quad \sum_{k=1}^K \theta_k^* - \frac{K}{u_{\max} - u_{\min}} \right] = \bar{\mathbf{0}}.$$

This implies that $T = \sum_{k=1}^K s_k = \sum_{k=1}^K \lambda^* \theta_k^* = \frac{K\lambda^*}{u_{\max} - u_{\min}}$. We then have $\theta_k^* = \frac{Ks_k}{T(u_{\max} - u_{\min})}$. We further note that

$$\nabla^2 \mathcal{Q}(\Theta^*, \lambda^*) = \begin{bmatrix} -\frac{s_1}{(\theta_1^*)^2} & 0 & 0 & 0 \\ 0 & -\frac{s_2}{(\theta_2^*)^2} & 0 & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \dots & -\frac{s_K}{(\theta_K^*)^2} & 0 \\ 0 & 0 & & 0 & 0 \end{bmatrix}.$$

Here we observe that $\nabla^2 \mathcal{Q}(\Theta^*, \lambda^*)$ is negative semi-definite, which shows that Θ^* is indeed a local maximizer of $p(U|\Theta)$, proving the first assertion. To prove the second assertion, observe that since local maximizers satisfy $s_k = \frac{\theta_k^* T(u_{\max} - u_{\min})}{K}$, any local maximizer satisfies

$$\begin{aligned} p(U|\Theta^*) &= \frac{T(u_{\max} - u_{\min})}{K} \sum_k \theta_k^* \log \theta_k^* \\ &= -\frac{T(u_{\max} - u_{\min})}{K} \psi(p_{\Theta^*}), \end{aligned}$$

where $\psi(p_{\Theta})$ denotes the differential entropy of p_{Θ} the probability density function corresponding to Θ . $\square$

2.3.3 SDA Proof of Convergence

From (2.3) and (2.3) we note that the M-Stage and the E-Stage are maximizing $p(X, \Theta|Y)$ with respect to X and Θ respectively. That is

$$X^{(n)} = \arg \max_X p(X, \Theta^{(n-1)}|Y)$$

$$\Theta^{(n)} = \arg \max_{\Theta} p(X^{(n)}, \Theta|Y).$$

The E and M stages can therefore be thought of as the stages of cyclic block coordinate descent (coordinate ascent, but we use 'descent' for convenience and uniformity with literature), with $p(X, \Theta|Y)$ as the objective function being maximized. We now find an expression for the objective function when the density $p(u_t|\Theta)$ is governed by a histogram $\overline{\mathcal{H}}_\Theta(u_t)$. As observed in Lemma 2.1, using Bayes rule, $p(X, \Theta|Y) = \frac{p(Y|X, \Theta)p(X|\Theta)p(\Theta)}{p(Y)}$. Here $p(Y)$ is not dependent on either X or Θ , and $p(\Theta)$ is assumed to be uniform. Further Y is independent of Θ given X in the model $\mathcal{M}_G$. Therefore

$$\arg \max_{X \text{ or } \Theta} p(X, \Theta|Y) = \arg \max_{X \text{ or } \Theta} p(Y|X)p(X|\Theta).$$

From (2.2),

$$p(Y|X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\sum_{t=1}^T -\frac{(y_t - x_t)^2}{2\sigma^2} \right).$$

As seen in (2.1), $p(X|\Theta) = p(U|\Theta)$. Now u_t are independent and identically distributed as per our model $\mathcal{M}_G$. So $p(U|\Theta) = \prod_{t=1}^T p(u_t|\Theta)$. Using the histogram $\overline{\mathcal{H}}_\Theta(u)$ for approximating the density $p(u|\Theta)$ as seen in Lemma 2.3 we get $p(X|\Theta) = \prod_{k=1}^K \theta_k^{s_k}$. Thus, $\zeta := \arg \max_{X \text{ or } \Theta} p(Y|X)p(X|\Theta)$ can be shown to be

$$\begin{aligned} \zeta &= \arg \max_{X \text{ or } \Theta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\sum_{t=1}^T -\frac{(y_t - x_t)^2}{2\sigma^2} \right) \prod_{k=1}^K \theta_k^{s_k} \\ &= \arg \max_{X \text{ or } \Theta} \sum_{t=1}^T -(y_t - x_t)^2 + \sum_{k=1}^K 2\sigma^2 s_k \log(\theta_k) \\ &= \arg \max_{X \text{ or } \Theta} \mathcal{J}(X, \Theta). \end{aligned}$$

Theorem 2.3. *For a finite length bounded observation sequence Y , the sequence*

$$\{p(\hat{X}^{(1)}, \Theta^{(1)}|Y), p(\hat{X}^{(2)}, \Theta^{(2)}|Y), \dots, p(\hat{X}^{(n)}, \Theta^{(n)}|Y), \dots\}$$

generated by the step detection algorithm converges. Further, the sequence $\{(\hat{X}^{(1)}, \Theta^{(1)}), (\hat{X}^{(2)}, \Theta^{(2)}), \dots, (\hat{X}^{(n)}, \Theta^{(n)}), \dots\}$, contains a convergent subsequence $\{(\hat{X}^{(v_1)}, \Theta^{(v_1)}), (\hat{X}^{(v_2)}, \Theta^{(v_2)}), \dots, (\hat{X}^{(v_m)}, \Theta^{(v_m)}), \dots\}$.

Proof. We observe from theorem (2.2) that $p(\hat{X}^{(n)}, \Theta^{(n)}|Y) \geq p(\hat{X}^{(n-1)}, \Theta^{(n-1)}|Y)$ for all n . Thus by monotone convergence, the sequence $\{p(\hat{X}^{(n)}, \Theta^{(n)}|Y)\}$ converges to a bounded limit.

Further $\{(\hat{X}, \Theta)\}$ are chosen from a compact set: Θ is chosen from a simplex of

finite and non-zero volume, and the candidates for $\hat{X}$ are selected from a compact set, indeed $x_t \in [x_0 + tu_{\min}, x_0 + tu_{\max}]$ for all $t = 1, \dots, T$. Note that $p(\hat{X}^{(n)}, \Theta^{(n)}|Y)$ is a probability density and hence non-negative for every n . For a finite number of bins K , with $u_{\max} \neq u_{\min}$, $\sup \left(\overline{\mathcal{H}}_{\Theta} \left(U^{(n)} \right) \right) = \frac{K}{u_{\max} - u_{\min}}$. Therefore,

$$0 \leq p(\hat{X}^{(n)}, \Theta^{(n)}|Y) \leq \frac{K}{u_{\max} - u_{\min}} \text{ for every } n.$$

This implies that $\left\{ \left(\hat{X}^{(n)}, \Theta^{(n)} \right) \right\}$ contain a convergent subsequence $\left\{ \left(\hat{X}^{(v_m)}, \Theta^{(v_m)} \right) \right\}$. $\square$

Lemma 2.4. *The solution space for the maximization, $X^* = \arg \max_X \mathcal{L}(X, \Theta)$ is restricted to*

$$x_t^* \in \{y_t\} \cup \{\mathcal{B} + x_{t-1}^*\} \text{ for all } t,$$

with $x_0^* = 0$.

Proof. Let $X = \arg \max_X \mathcal{L}(X, \Theta)$. Suppose there exist $x_t \neq y_t$ and $x_t \notin \{\mathcal{B} + x_{t-1}\}$ for some t . This implies that $x_t - x_{t-1} = u_t \in (b_{k-1}, b_k)$ for some k . Since $x_t \neq y_t$,

$$\mu := |y_t - x_t| > 0.$$

Here $\mathcal{L}(X, \Theta) = \sum_{t=1}^T -(y_t - x_t)^2 + 2\sigma^2 \sum_{k=1}^K s_k \log[\theta_k]$. Now choose $\tilde{x}_t \in (b_{k-1} + x_{t-1}, b_k + x_{t-1})$ and define $\widetilde{X} := \{\tilde{x}\} \cup \{X \setminus x_t\}$. Here s_k (and thus $\sum_{k=1}^K s_k \log[\theta_k]$) remains the same with X and $\widetilde{X}$.

Now we show that $\Delta = \mathcal{L}(X, \Theta) - \mathcal{L}(\widetilde{X}, \Theta) < 0$ in all of the following three mutually exclusive cases:

Case 1: $b_{k-1} + x_{t-1} \leq y_t \leq b_k + x_{t-1}$

We set $\tilde{x}_t = y_t$. Then

$$\begin{aligned} \Delta &= -(y_t - x_t)^2 + (y_t - \tilde{x}_t)^2 \\ &= -(y_t - x_t)^2. \end{aligned}$$

Therefore, $\Delta < 0$ since $x_t \neq y_t$. This implies $\mathcal{L}(X, \Theta) < \mathcal{L}(\widetilde{X}, \Theta)$, which is a contradiction that X is a maximizer of $\mathcal{L}(X, \Theta)$.

Case 2: $y_t \geq b_k + x_{t-1}$

We set $\tilde{x}_t = x_t + \varepsilon$ where $\varepsilon > 0$ such that $\tilde{x}_t - x_{t-1} \in (b_{k-1}, b_k)$. This implies that

$$x_t + \varepsilon < b_k + x_{t-1}$$

and since $y_t \geq b_k + x_{t-1}$, we get

$$\mu = y_t - x_t > \varepsilon.$$

Now,

$$\begin{aligned}\Delta &= -(y_t - x_t)^2 + (y_t - x_t - \varepsilon)^2 \\ &= -\mu^2 + (\mu - \varepsilon)^2 \\ &= \varepsilon(\varepsilon - 2\mu).\end{aligned}$$

Since $\mu > \varepsilon > 0$, $\Delta < 0$. Therefore $\mathcal{L}(X, \Theta) < \mathcal{L}(\tilde{X}, \Theta)$.

Case 3: $y_t \leq b_{k-1} + x_{t-1}$

We choose $\varepsilon > 0$ such that $\tilde{x}_t = x_t - \varepsilon$ and $\tilde{x}_t - x_{t-1} \in (b_{k-1}, b_k)$. Therefore

$$x_t - \varepsilon > b_{k-1} + x_{t-1} \geq y_t.$$

Thus,

$$\mu = x_t - y_t > \varepsilon.$$

Now,

$$\begin{aligned}\Delta &= -(y_t - x_t)^2 + (y_t - x_t + \varepsilon)^2 \\ &= -\mu^2 + (-\mu + \varepsilon)^2 \\ &= \varepsilon(\varepsilon - 2\mu).\end{aligned}$$

Since $\mu > \varepsilon > 0$, we observe that $\Delta < 0$ and hence $\mathcal{L}(X, \Theta) < \mathcal{L}(\tilde{X}, \Theta)$. This is a contradiction that X is a maximizer of $\mathcal{L}(X, \Theta)$. This completes the proof. □

Lemma 2.5. *For every $X = \{x_1, x_2, \dots, x_T\}$ and histogram $\mathcal{H}_{\Theta, \mathcal{D}}(U)$, the following inequality holds:*

$$\mathcal{J}(X, \overline{\mathcal{H}}_{\Theta}(U)) \geq \mathcal{J}(X, \mathcal{H}_{\Theta, \mathcal{D}}(U)).$$

Here,

$$\mathcal{J}(X, \mathcal{H}_{\Theta, \mathcal{D}}(U)) := \sum_{t=1}^T -(y_t - x_t)^2 + 2\sigma^2 \sum_{k=1}^K \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t),$$

and $\overline{\mathcal{H}}_{\Theta}(U)$ is the upper semi-continuous version of the histogram $\mathcal{H}_{\Theta, \mathcal{D}}(U)$.

Proof. We observe that

$$\begin{aligned} \Delta_{\mathcal{J}} &:= \mathcal{J}(X, \overline{\mathcal{H}}_{\Theta}(U)) - \mathcal{J}(X, \mathcal{H}_{\Theta, \mathcal{D}}(U)) \\ &= \sum_{t=1}^T \left[-(y_t - x_t)^2 + 2\sigma^2 \sum_{k=1}^K \log \overline{\mathcal{H}}_{\Theta}(u_t) \right] \\ &\quad - \sum_{t=1}^T \left[-(y_t - x_t)^2 + 2\sigma^2 \sum_{k=1}^K \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \right] \\ &= 2\sigma^2 \sum_{t=1}^T \sum_{k=1}^K \left(\log \overline{\mathcal{H}}_{\Theta}(u_t) - \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \right). \end{aligned}$$

Let $u_t \notin \mathcal{B}$ for all $t = \alpha_1, \alpha_2, \dots, \alpha_A$ and $u_t \in \mathcal{B}$ for all $t = \beta_1, \beta_2, \dots, \beta_B$. The descriptions of the histogram in Definition 2.2 and the upper semi-continuous histogram in Definition 2.3 imply that

$$\overline{\mathcal{H}}_{\Theta}(u_t) = \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \text{ for all } t \in \{\alpha_1, \alpha_2, \dots, \alpha_A\}.$$

Further for all $b_k \in \mathcal{B}$

$$\overline{\mathcal{H}}_{\Theta}(b_k) = \max\{\theta_k, \theta_{k+1}\},$$

while

$$\mathcal{H}_{\Theta, \mathcal{D}}(b_k) = \begin{cases} \theta_k & \delta_k = 1 \\ \theta_{k+1} & \delta_k = 0, \end{cases}$$

where $\delta_k \in \mathcal{D}$, the boundary selection variable. Therefore, for any choice of $\mathcal{D}$

$$\overline{\mathcal{H}}_{\Theta}(u_t) \geq \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \text{ for all } t \in \{\beta_1, \beta_2, \dots, \beta_B\}.$$

This gives us

$$\begin{aligned}
\Delta_{\mathcal{J}} &= 2\sigma^2 \sum_{t=\alpha_1}^{\alpha_A} \sum_{k=1}^K \left(\log \overline{\mathcal{H}}_{\Theta}(u_t) - \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \right) \\
&\quad + 2\sigma^2 \sum_{t=\beta_1}^{\beta_B} \sum_{k=1}^K \left(\log \overline{\mathcal{H}}_{\Theta}(u_t) - \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \right) \\
&= 2\sigma^2 \sum_{t=\beta_1}^{\beta_B} \sum_{k=1}^K \left(\log \overline{\mathcal{H}}_{\Theta}(u_t) - \log \mathcal{H}_{\Theta, \mathcal{D}}(u_t) \right) \\
&\geq 0.
\end{aligned}$$

□

Note: $\mathcal{J}(X, \overline{\mathcal{H}}_{\Theta}(U)) = \mathcal{L}(X, \Theta) = \sum_{t=1}^T -(y_t - x_t)^2 + 2\sigma^2 \sum_{k=1}^K s_k \log[\theta_k]$.

2.3.4 SDA Formulation with Sensor Dynamics

Measurements of stepping signals are often done with sensors whose dynamics have time constants which are of the same time scale as the stepping rates, and thus cannot be ignored. In such settings, accounting for the sensor dynamics improves the estimates. Now, we briefly describe this reformulation in the case of the SDA and show that the proof of improvement holds under the reformulation.

Note that $G(\mathcal{Z})$ represents the causal sensor transfer function in the Z-domain. Let $g(\tau)$ be the discrete time impulse response of the transfer function $G(\mathcal{Z})$. Let z_t be the output of the sensor at time t . Then $z_t = \sum_{l=-\infty}^{l=\infty} g(l) x(t-l)$, where we have further identified $x_t \equiv x(t)$. The reformulated SDA is summarized in Algorithm 2 and is given as follows:

M-stage: Only the M-Stage is modified, wherein the objective function $J_G(X, \Theta^{(n-1)})$ is minimized with respect to X to give $\hat{X}^{(n)}$, that is

$$\hat{X}^{(n)} := \arg \min_X J_G(X, \Theta^{(n-1)}), \quad (2.5)$$

with

$$J_G(X, \Theta^{(n-1)}) := \sum_{t=1}^T (y_t - z_t)^2 + W_t(X, \Theta^{(n-1)}),$$

where

$$z_t = \sum_{l=-\infty}^{l=\infty} g(l) x(t-l),$$

and

$$W_t(X, \Theta^{(n-1)}) = \begin{cases} 9\sigma^2 \bar{\delta}(x_t - x_{t-1}) & n = 1 \\ -2\sigma^2 \log [p(u_t | \Theta^{(n-1)})] & n > 1 \end{cases}. \quad (2.6)$$

E-Stage: This step remains unchanged from the standard SDA, and is described here for convenience. From the $\hat{X}^{(n)}$ computed in the modified M-Stage, compute $u_t^{(n)} = \hat{x}_{t+1}^{(n)} - \hat{x}_t^{(n)}$. Evaluate $\Theta^{(n)} = \{\theta_1^{(n)}, \theta_2^{(n)}, \dots, \theta_K^{(n)}\}$ and select $\mathcal{D}^{(n)} = \{\delta_0^{(n)}, \delta_1^{(n)}, \dots, \delta_K^{(n)}\}$ with $\delta_0^{(n)} = 0$, and $\delta_K^{(n)} = 1$ using

$$\theta_k^{(n)} = \frac{s_k^{(n)}}{T} \text{ for all } k = 1, \dots, T, \quad (2.7)$$

such that $\Theta^{(n)}$ and $\mathcal{D}^{(n)}$ form an upper semi-continuous histogram.

Algorithm 2 Step Detection Algorithm with Sensor Dynamics

Input: Time series measurements $Y = \{y_1, y_2, \dots, y_T\}$ and sensor dynamic model $G(z)$.

Output: Step fit $\hat{X}^{(N)} = \{\hat{x}_1^{(N)}, \hat{x}_2^{(N)}, \dots, \hat{x}_T^{(N)}\}$ and Histogram bin parameters $\Theta^{(N)} = \{\theta_1^{(N)}, \theta_2^{(N)}, \dots, \theta_K^{(N)}\}$.

- 1: Set initial penalty using (2.6)
 - 2: Compute step fit $\hat{X}^{(1)}$ using (2.5)
 - 3: **for all** $n \in \{2, 3, \dots, N\}$ **do**
 - 4: **E-Stage:** Compute $\Theta^{(n)}$ using (2.7)
 - 5: **M-Stage:** Compute step fit $\hat{X}^{(n)}$ using (2.5)
 - 6: **end for**
-

The dependence of J_G on Z has been omitted to emphasize the fact that given an X , Z is completely determined. The penalty $W_t(X, \Theta^{(n-1)})$ retains the definition given in (2.4). The generative model assumed is as given by

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u}_t, \quad (2.8)$$

where $\mathbf{x}_0 = 0$. Here the steps $\mathbf{u}_t$ are independent and identically distributed for every $t = 1, \dots, T$ and generated by a distribution $p(\mathbf{u}_t | \Theta)$, which is parameterized by $\Theta \in \mathbb{R}^K$. Here, $\Theta \in \mathbb{R}^K$ is an independent and identically distributed random

vector for $t = 1, \dots, T$.

$$\mathbf{z}_t = \sum_{l=-\infty}^{l=\infty} g(l) x(t-l), \quad (2.9)$$

$$\mathbf{y}_t = \mathbf{z}_t + \boldsymbol{\eta}_t, \quad (2.10)$$

where $\{\boldsymbol{\eta}_t\}$ are independent and identically distributed and each $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma^2)$.

This modified generative model $\mathcal{M}_{\mathcal{G}}^s$ which now includes the sensor dynamics is represented by the Bayesian Network shown in Fig. 2.3.1.

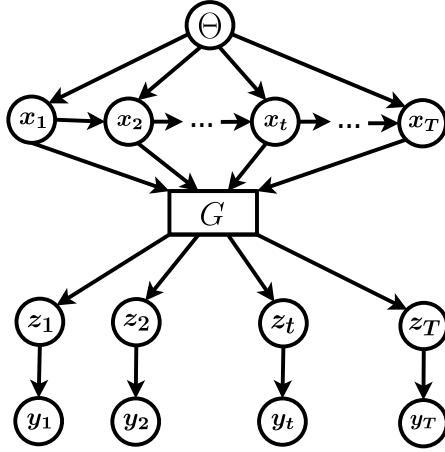


Fig. 2.3.1: Underlying generative model $\mathcal{M}_{\mathcal{G}}^s$ assumed for solving the step detection problem, with the addition of sensor dynamics represented by G . The sequence $\{z_t\}$ is obtained by the convolution of the sequence $\{x_t\}$ with G . Here, y_t are the noisy observations of z_t , and $\Theta = \{\theta_1, \dots, \theta_K\}$ are the time invariant parameters of the step size distribution.

Lemma 2.6. *Maximizing the probability $p(Z, X, \Theta|Y)$ with respect to X in the Bayesian Network as shown in Fig. 2.3.1 is equivalent to minimizing $J_G(X, \Theta)$ with respect to X where $Z = G(X)$ for some fixed, deterministic causal function G .*

Proof. We observe that using Bayes, we can rewrite $p(Z, X, \Theta|Y)$ as $\alpha p(Y|Z, X, \Theta) p(Z|X, \Theta) p(X|\Theta) p(\Theta)$ with $\alpha := p(Y)$, a constant. In the

Bayesian Network shown in Fig. 2.3.1, Z forms the Markov blanket for Y , thus $p(Y|Z, X, \Theta) = p(Y|Z)$. Thus,

$$p(Z, X, \Theta|Y) = \alpha p(Y|Z) p(Z|X, \Theta) p(X|\Theta) p(\Theta).$$

We also note that X forms the Markov blanket for Z and thus $p(Z|X, \Theta) = p(Z|X)$. Furthermore, given a sequence X , the sequence Z is completely determined and is unique. This is because the interconnecting function G is assumed to be a deterministic quantity. Thus $p(Z|X) = \delta(X)$, where $\delta(X)$ is the Dirac-delta function, and a deterministic quantity. Consequently,

$$p(Z, X, \Theta|Y) = \alpha \delta(X) p(Y|Z) p(X|\Theta) p(\Theta).$$

Now, applying Bayes rule, we can expand the terms of $p(Z, X, \Theta|Y)$ and write it as:

$$\alpha \delta(X) \prod_{t=1}^T p(y_t|Y_1^{t-1}, Z) p(x_t|X_1^{t-1}, \Theta) p(\Theta).$$

In the Bayesian Network (see Fig. 2.3.1), z_t forms the Markov blanket for y_t implying $p(y_t|Y_1^{t-1}, Z) = p(y_t|z_t)$. In addition, $p(x_t|X_1^{t-1}, \Theta) = p(x_t|x_{t-1}, \Theta)$ since x_t is independent of all non-descendants X_1^{t-2} given its parents (x_{t-1}, Θ) [56]. So we can write

$$p(Z, X, \Theta|Y) = \alpha \delta(X) \prod_{t=1}^T p(y_t|z_t) p(x_t|x_{t-1}, \Theta) p(\Theta).$$

From (2.10),

$$p(Y|Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\sum_{t=1}^T -\frac{(y_t - z_t)^2}{2\sigma^2}\right).$$

Since $p(Y)$ and $p(\Theta)$ are independent of X ,

$$\begin{aligned} \hat{X}^* &= \arg \max_X p(Z, X, \Theta|Y) \\ &= \arg \max_X p(Y|Z) p(X|\Theta) \\ &= \arg \max_X \prod_{t=1}^T p(y_t|Y_1^{t-1}, Z) p(x_t|X_1^{t-1}, \Theta) \\ &= \arg \min_X \left\{ \sum_{t=1}^T \frac{(y_t - z_t)^2}{2\sigma^2} - \log [p(x_t|X_1^{t-1}, \Theta)] \right\}. \end{aligned}$$

This shows the desired result. $\square$

Note: Lemma 2.6 holds for any deterministic function G when $Z = G(X)$. Therefore, it also holds for deterministic sensor dynamic transfer functions wherein $z_t = \sum_{l=-\infty}^{\infty} g(l) x(t-l)$. In Theorem 2.4, we show that the improvements in estimates hold true for the modified SDA which accommodates sensor dynamics.

Theorem 2.4. *Let Θ be uniformly distributed and observations Y corrupted by zero mean Gaussian noise with known variance σ^2 , for the Bayesian Network shown in Fig. 2.3.1. That is $\mathbf{y}_t = \mathbf{z}_t + \boldsymbol{\eta}_t$, with $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma^2)$ for all t . Here $Z = G(X)$ and G is a deterministic function. Then, given an estimate $[X^{(n-1)}, \Theta^{(n-1)}]$, the modified step detection algorithm's iteration produces an improved estimate $[X^{(n)}, \Theta^{(n)}]$ whereby*

$$p(X^{(n)}, \Theta^{(n)} | Y) \geq p(X^{(n-1)}, \Theta^{(n-1)} | Y).$$

Proof. From Lemma 2.6, we note that the M-Stage of the modified SDA is equivalent to the maximization of $p(Z, X, \Theta | Y)$ with respect to X . We note that by Bayes' theorem, $p(Z, X, \Theta | Y) = p(Z, |X, \Theta, Y) p(X, \Theta | Y)$. From the Bayesian Network of Fig. 2.3.1, observe that X forms the Markov Blanket for Z . Further, G is deterministic, thus $p(Z, |X, \Theta, Y) = p(Z | X) = \delta(X)$. Hence maximization of $p(Z, X, \Theta | Y)$ is equivalent to maximizing of $p(X, \Theta | Y)$ with respect to X .

In the M-Stage, we set $\Theta = \Theta^{(n-1)}$ and minimize $J_G(X, \Theta)$ with respect to X to give $X^{(n)}$. Therefore,

$$p(X^{(n)}, \Theta^{(n-1)} | Y) \geq p(X, \Theta^{(n-1)} | Y) \text{ for all } X.$$

The E-Stage of the modified SDA remains unchanged, wherein the estimate $X^{(n)}$ is binned to compute $\Theta^{(n)}$. As shown in Lemma 2.3, this step is equivalent to maximizing $p(X^{(n)}, \Theta | Y)$ with respect to Θ . Therefore, we have

$$p(X^{(n)}, \Theta^{(n)} | Y) \geq p(X^{(n)}, \Theta | Y) \text{ for all } \Theta.$$

Combining these observations gives us

$$p(X^{(n)}, \Theta^{(n)} | Y) \geq p(X^{(n-1)}, \Theta^{(n-1)} | Y),$$

proving the desired result for the modified SDA. $\square$

2.4 Results and Comparisons

We now compare the performance of the SDA with frequently employed step detection methods: Total Variation Regularization (TVR) [58], Hidden Markov Models (VS-HMM) [45, 46], the χ^2 based method developed by Kerssemakers (χ^2) [43], Linearly Penalized Segmentation (PELT) based search method described in [52], Gaussian Kernel (PELT_KER) [53], Binary-Segmentation (BINSEG)[54], and Bottom-up (BOTUP) [55]. For the last four methods, we use the consolidated python implementation given by [47].

2.4.1 Performance Metrics

The normalized root mean square error (NRMSE) is used for measuring the difference between the ground truth $X^\dagger$ and the estimates $\hat{X}$, where the NRMSE is defined by

$$e_{NRMSE} := \frac{1}{C} \sqrt{\frac{1}{T} \sum_{t=1}^T (x_t^\dagger - \hat{x}_t)^2},$$

with the normalization constant

$$C := \sqrt{\frac{1}{T} \sum_{t=1}^T (x_t^\dagger - \bar{x})^2}.$$

Here

$$\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t^\dagger.$$

The performance of the algorithms is also compared when the step detection algorithms are used as binary classifiers, wherein they predict the presence or absence of a step at any given time. Specifically, when the magnitude of an estimated step is predicted to be within $\pm 10\%$ of the magnitude of the true step, and the location of such a step is within ± 5 time steps of the true location, it is considered a true positive (TP). Any estimated steps of incorrect magnitude or location with respect to the previously described criterion are considered to be false positives (FP). When the algorithms fail to estimate a step within the location tolerance band of a true step, a false negative (FN) is assigned. True negatives constitute the remaining cases, i.e. no step estimates when there are no true steps. Since in the stepping signals the

number of true negatives far exceeds the number of true positives, we have used the F1 score [59] as a classification performance metric, which is defined as

$$F1 = 2 \frac{(\text{PPV})(\text{TPR})}{\text{PPV} + \text{TPR}},$$

where the positive predictive value is given by $\text{PPV} := \frac{\text{TP}}{\text{TP} + \text{FP}}$, and the true positive rate $\text{TPR} := \frac{\text{TP}}{\text{TP} + \text{FN}}$.

2.4.2 Empirical Results and Discussion

The algorithms are tested on 50 randomly generated stepping signals with the same randomization seeds for the signal and noise. The plots in Figures 2.4.1, 2.4.2, 2.4.3, and 2.4.4 show the mean of the NRMSE and F1 scores from the multiple runs. The error bars indicate 95% confidence intervals evaluated assuming the samples of the performance metric follow the Student's t distribution [60]. For each of the empirical observations, the default settings are magnitude of step sizes = 1, number of steps = 10, probability of a positive step = 0.5, average number of samples per step = 50, unless otherwise mentioned. For the SDA, the histogram bin widths are fixed to 0.05, with the stopping criteria being a change in histograms less than a tolerance value of 10^{-7} , or 100 iterations, whichever occurs earlier. The change in histograms at iteration n are computed using the Kolmogorov-Smirnov metric [60], defined by:

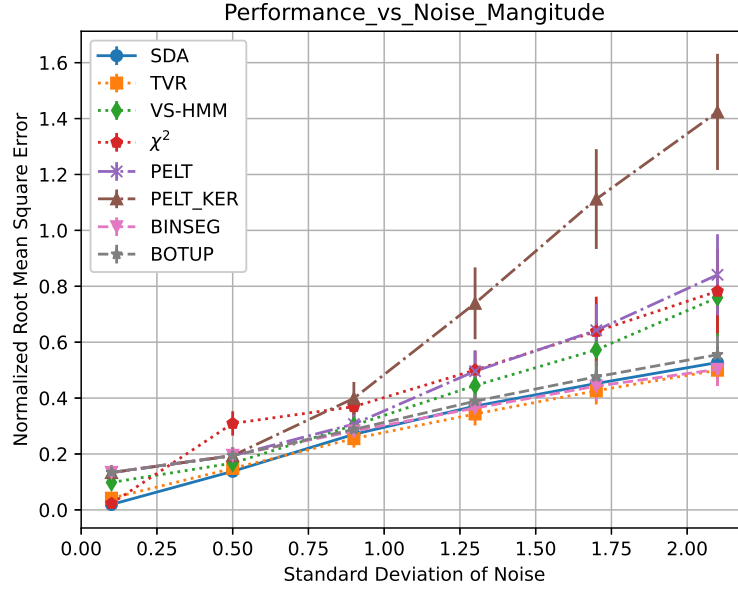
$$d_{KS} := \max_{k \in \{0, 1, \dots, K\}} \left| \sum_{i=0}^k \theta_k^n - \sum_{i=0}^k \theta_k^{n-1} \right|.$$

In Fig. 2.4.1, the performance of the algorithms is measured as the standard deviation of the noise is increased from 0.1 to 2.0. It can be seen that the deterioration of the SDA performance is graceful with increasing noise magnitude. The SDA balances both the NRMSE and F1 score performance measures, following closely the best algorithms with respect to each metric. Fig. 2.4.2 shows the empirical performance as the number of steps in a test signal are increased from 1 to 50. A larger number of steps in the signal improves the performance of all the algorithms.

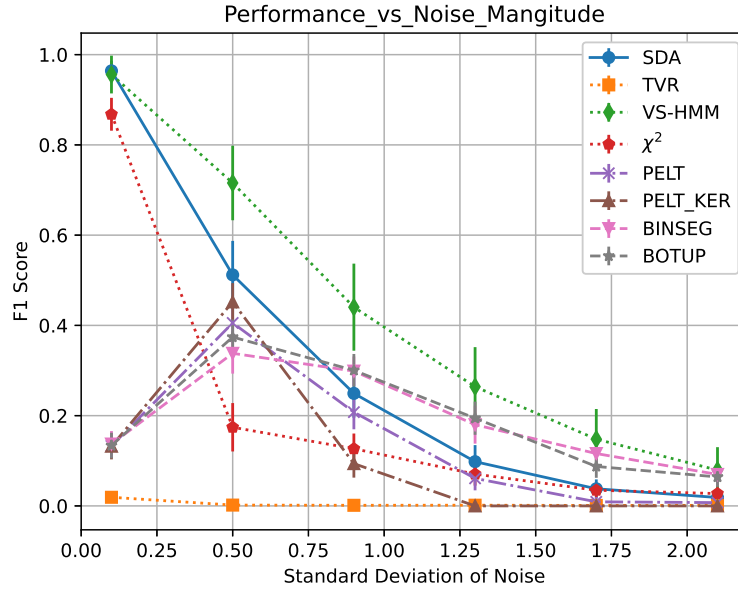
Fig. 2.4.3 compares performance of the algorithms as the average number of samples per step in raised from 10 to 1000. Note that in all the tests without sensor dynamics, the TVR along with the SDA provide the best NRMSE. Moreover the

VS-HMM together with the SDA provide the best F1 scores, in the absence of sensor dynamics.

In Fig. 2.4.4, the algorithms are tested on step signals filtered by first order sensor transfer function given by $G(s) = \frac{1}{\tau s + 1}$. The sensor time constant τ is increased from 0.1 to 2 seconds, wherein the assumption is that the sampling of the signal is done at 50 Hz. This implies that the average duration of a step is 1 second. It is seen that the SDA which can take into account the dynamics, has a 30-40% reduction in NRMSE compared to TVR and VS-HMM.

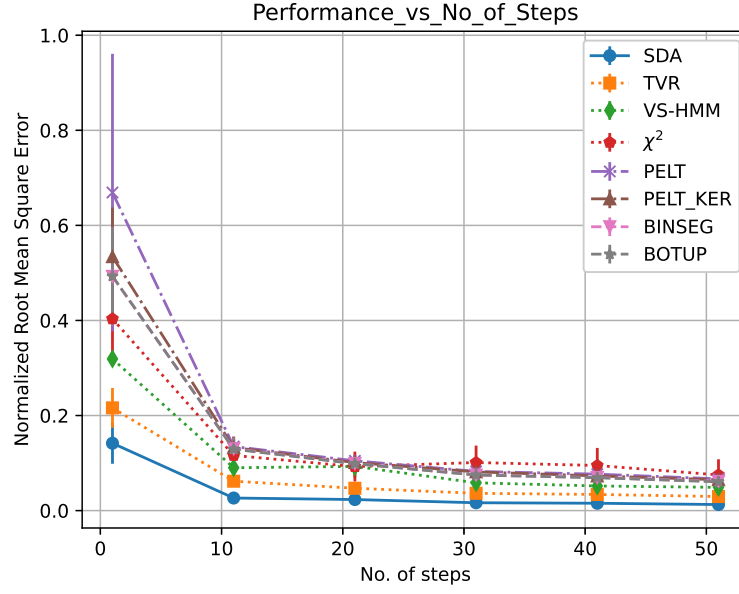


(a)

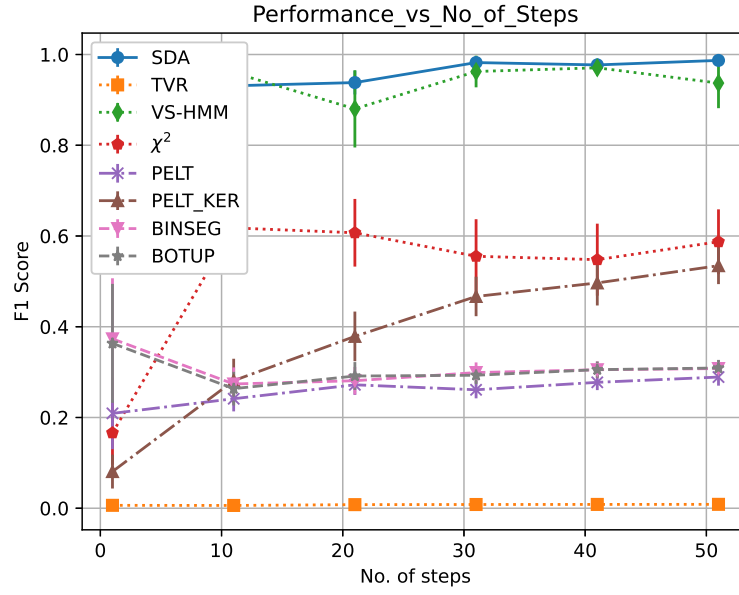


(b)

Fig. 2.4.1: Comparative performance with increasing noise magnitude. (a) The SDA has the lowest NRMSE for noise standard deviations up to 0.5, follows the TVR in performance for standard deviations beyond 0.5. (b) The SDA closely follows the VS-HMM algorithm when the performance is measured using F1 scores. The TVR performance is degraded in terms of F1 scores due to a large number of erroneous step fits, despite having small NRMSE. The methods PELT, PELT_KER, BINSEG, and BOTUP show improved performance for higher noise magnitude, while over-fitting the data for smaller noise magnitudes.

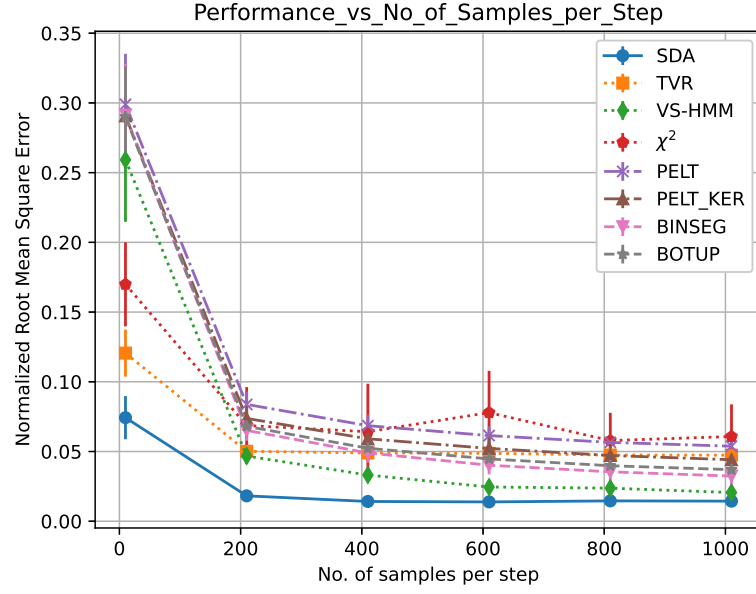


(a)

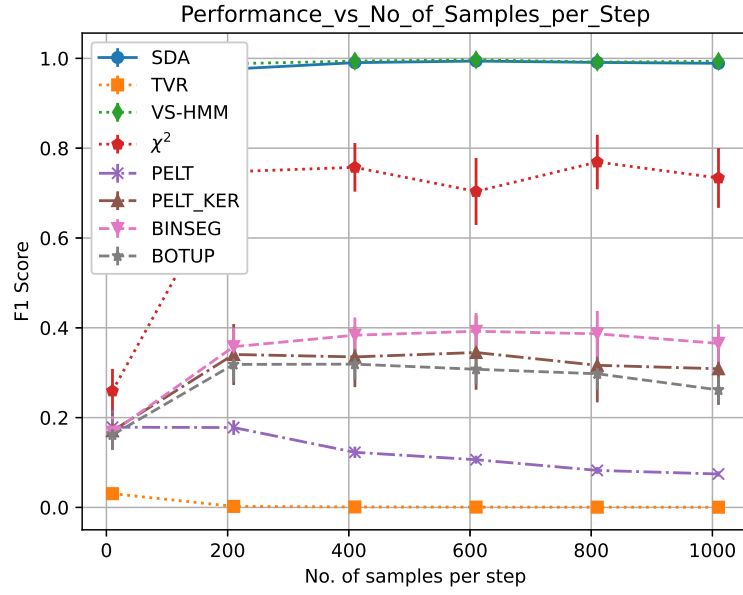


(b)

Fig. 2.4.2: Comparative performance of the SDA with increasing number of steps available in the signal. (a) All the algorithms benefit from having multiple steps in the signal. The SDA has the lowest NRMSE amongst the algorithms tested. (b) The SDA closely follows the VS-HMM method when the performance is measured with F1 scores. The poor F1 scores of TVR are due to multiple erroneous fits.

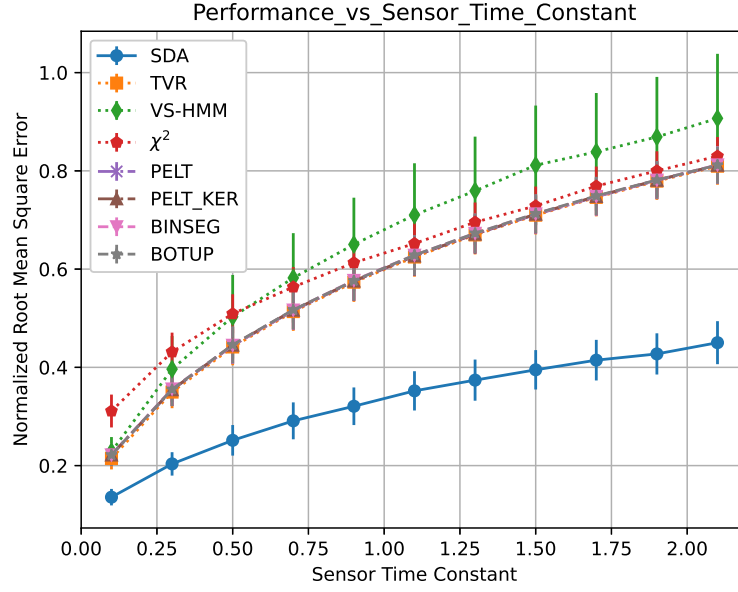


(a)

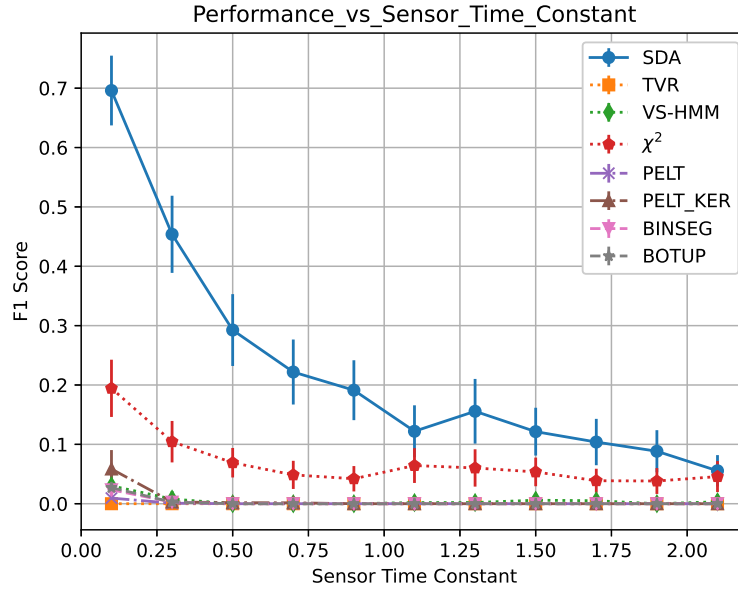


(b)

Fig. 2.4.3: Comparative performance of the SDA with increasing number samples per step in the signal. (a) A larger number of samples for any given step positively benefits each algorithm. The SDA has the lowest NRMSE. (b) The VS-HMM has ~15% better F1 scores compared to the SDA when the number of samples are small. For more than 200 samples per step, the SDA and VS-HMM have equivalent performance, with F1 scores close to 1.



(a)



(b)

Fig. 2.4.4: Comparative performance of the SDA with sensor dynamics. A first order transfer function is assumed for the sensor, with unity gain and the time constant increased from 0.1 to 2s. (a) The SDA has a significantly lower NRMSE as it takes into account the sensor dynamics. (b) As expected the SDA has a larger F1 score compared to the other algorithms which do not account for sensor dynamics. F1 scores deteriorates beyond sensor time constants of 0.25s as sensors with large time constants (with respect to step duration) effectively destroy stepping information.

2.4.3 Sample Estimates

In Fig. 2.4.5, the SDA's step size estimates are shown on a sample step signal, for illustrating the estimation process. The sample step signal was generated with 20 randomly spaced unidirectional steps of size 1, and noise standard deviation of 0.5. Observe that the step size distributions converge to the true distribution as the iterations increase. At the end of iteration 0, there are more incorrectly identified step sizes, which reduces in the subsequent estimates.

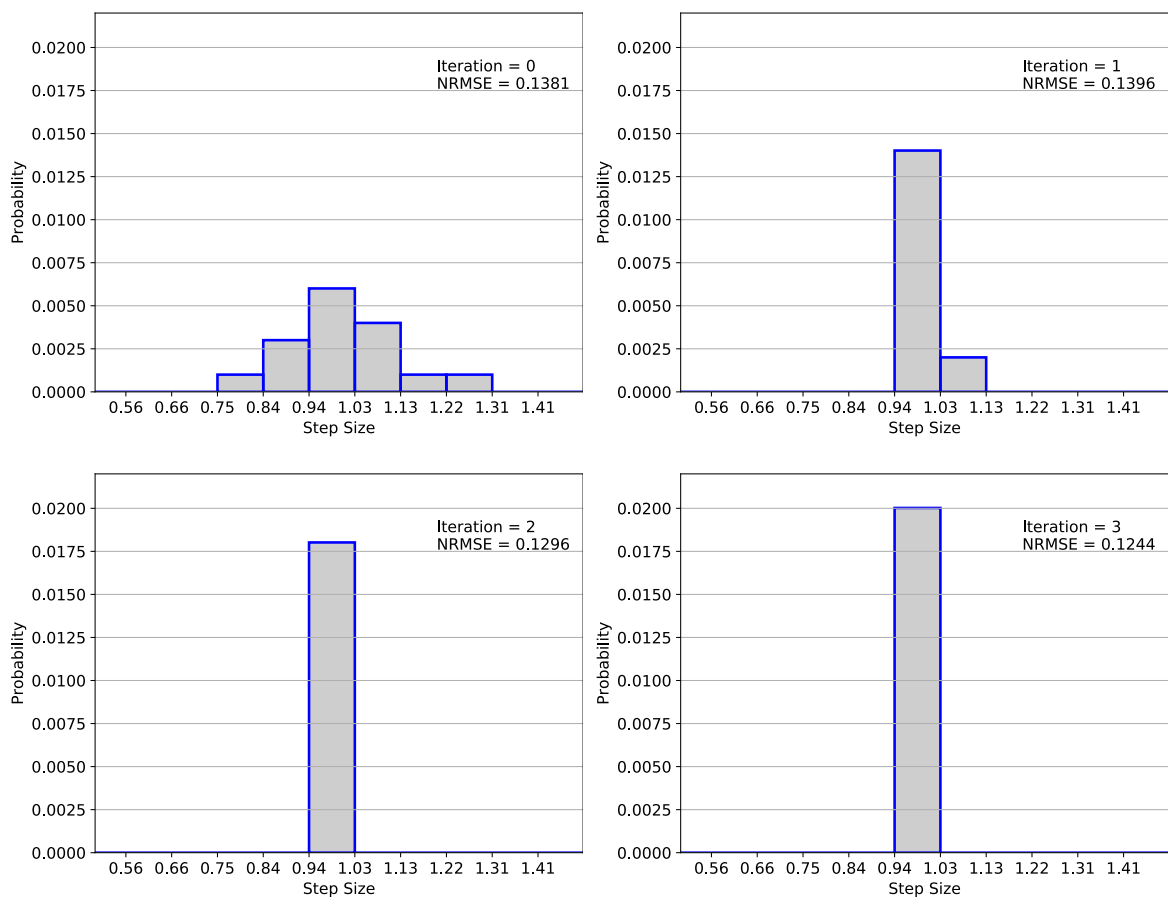


Fig. 2.4.5: Histograms fit by the SDA at the end of each iteration for a sample step signal, with 20 steps of magnitude 1, noise standard deviation of 0.5. Empirically we observe that with each iteration, the histograms generated converge to the true stepping distribution.

2.5 Conclusions

The need for detecting steps in noisy signals is frequently encountered in a number of fields. We have provided in this chapter a Step Detection algorithm, where we supply a theoretical foundation for analysis and show that each iteration of the algorithm necessarily improves the posterior probability of estimates given the observations. Further the iterations result in a converging sequence of the posterior probability. The chapter also provides a new technical toolbox for the analysis of iterations that uses histograms to approximate probability density functions. The improvements in estimates holds in the presence sensor dynamics, a situation often encountered, for example, when studying the physics of biological systems at the molecular scale. Comparative empirical analysis supports the theoretical description and the SDA's efficacy.

Supporting Information

Proposition 2.1. *Markov Blanket: In a Bayesian network $\mathcal{N}_{\mathcal{B}}$, the nodes Z are independent of every other node in the network, given the parents (P), children (C), and children's co-parents (S) of Z [56].*

Proof. Let X nodes in $\mathcal{N}_{\mathcal{B}}$, let their ordering be $\{x_1, x_2, \dots, x_N\}$ such that there is no descendant of x_i in $\{x_{i+1}, x_{i+2}, \dots, x_N\}$ for all i . Further, a node x_i is independent of all its non-descendants given its parents. Thus by Bayes rule

$$\begin{aligned} p(Z|X \setminus Z) &= \frac{p(X)}{\sum_{x_i \in Z} p(X)} \\ &= \frac{\prod_{x_i \in X} p(x_i|x_{i+1}, x_{i+2}, \dots, x_N)}{\sum_{x_i \in Z} \prod_{x_i \in X} p(x_i|x_{i+1}, x_{i+2}, \dots, x_N)} \\ &= \frac{\prod_{x_i \in X} p(x_i|\text{parents}(x_i))}{\sum_{x_i \in Z} \prod_{x_i \in X} p(x_i|\text{parents}(x_i))}. \end{aligned}$$

Now we split the product $\prod_{x_i \in X} p(x_i|\text{parents}(x_i))$ into those that involve $x_i \in Z \cup C$ and those that do not. That is we express $p(X) = \mathcal{U}\mathcal{V}$, where

$$\mathcal{U} := \prod_{x_i \in X \setminus (Z \cup C)} p(x_i|\text{parents}(x_i)),$$

and

$$\mathcal{V} := \prod_{x_i \in Z \cup C} p(x_i|\text{parents}(x_i)).$$

$$\begin{aligned} p(Z|X \setminus Z) &= \frac{p(X)}{\sum_{x_i \in Z} p(X)} \\ &= \frac{\mathcal{U}\mathcal{V}}{\sum_{x_i \in Z} \mathcal{U}\mathcal{V}} \\ &= \frac{\mathcal{U}\mathcal{V}}{\mathcal{U} \sum_{x_i \in Z} \mathcal{V}} \\ &= \frac{\prod_{x_i \in Z \cup C} p(x_i|\text{parents}(x_i))}{\sum_{x_i \in Z} \prod_{x_i \in Z \cup C} p(x_i|\text{parents}(x_i))}. \end{aligned}$$

By rearrangement, we observe $\prod_{x_i \in Z} p(x_i|\text{parents}(x_i)) = p(Z|P)$.

Similarly, $\prod_{x_i \in C} p(x_i|\text{parents}(x_i)) = p(C|Z, S)$. Note also that $p(C|Z, S) =$

$p(C|Z, S, P)$, and $p(Z|P) = p(Z|P, S)$. Thus,

$$\begin{aligned}
p(Z|X \setminus Z) &= \frac{p(Z|P)p(C|Z, S)}{\sum_{x_i \in Z} p(Z|P)p(C|Z, S)} \\
&= \frac{p(S, P)p(Z|S, P)p(C|Z, S, P)}{p(S, P) \sum_{x_i \in Z} p(Z|S, P)p(C|Z, S, P)} \\
&= \frac{p(C, Z, S, P)}{\sum_{x_i \in Z} p(C, Z, S, P)} \\
&= p(Z|C, S, P).
\end{aligned}$$

This completes the proof. $\square$

Proposition 2.2. *With $\mathcal{V}_K(\lambda) := \{X \in \mathbb{R}^K | \sum_{k=1}^K x_k = \lambda, x_k \geq 0 \text{ for all } k\}$ for $\lambda \in (0, \infty)$,*

$$|\mathcal{V}_K(\lambda)|_K = \frac{\lambda^K}{K!}$$

Proof. When $K = 1$, this is the statement that the length of the interval $[0, \lambda]$ is λ . For $K > 1$ we proceed by induction. Notice, from the parameterization

$$\mathcal{V}_K(\lambda) = \{(x, y) | x \in (0, \lambda), y \in \mathcal{V}_{K-1}(\lambda)\},$$

we have

$$|\mathcal{V}_K(\lambda)|_K = \int_0^\lambda |\mathcal{V}_{K-1}(\lambda)|_{K-1} dx.$$

By induction the right hand side is

$$\int_0^\lambda \frac{x^{K-1}}{(K-1)!} dx = \frac{\lambda^K}{K!}.$$

This is the required result. $\square$

Proposition 2.3. *The set $\mathcal{S}^{K-1}(\lambda) := \{\theta_K | \sum_{k=1}^K \theta_k = \lambda, \theta_k \geq 0 \text{ for all } k\}$ satisfies*

$$|\mathcal{S}^{K-1}(\lambda)|_{K-1} = \sqrt{K} \frac{\lambda^{K-1}}{(K-1)!}$$

Proof. Observe that for small h ,

$$\frac{h}{\sqrt{K}} |\mathcal{S}^{K-1}(\lambda)|_{K-1} = |\mathcal{V}_K(\lambda + h)|_K - |\mathcal{V}_K(\lambda)|_K + o(h).$$

Thus,

$$\begin{aligned} |\mathcal{S}^{K-1}(\lambda)|_{K-1} &= \sqrt{K} \lim_{h \rightarrow 0} \frac{(\lambda + h)^K - \lambda^K}{hK!} \\ &= \sqrt{K} \frac{\lambda^{K-1}}{(K-1)!}, \end{aligned}$$

which is the desired result. $\square$

Lemma 2.7. *For an upper semi-continuous histogram $\overline{\mathcal{H}}$ (see Definition 2.3), the function $f(\Theta) = \sum_{k=1}^K s_k \log[\theta_k]$ is continuous with respect to Θ , when $\theta_k \neq 0$ if $s_k \neq 0$.*

Proof. Here $s_k := |\{t : u_t \in \bar{r}_k\}|$. For any U and Θ , consider a perturbation $\epsilon_\Theta = [\epsilon_{\theta_1}, \epsilon_{\theta_2}, \dots, \epsilon_{\theta_K}]$ around Θ . Then for any $\delta > 0$, we will now show that there exists $|\epsilon_{\theta_k}| > 0$, for all k , such that

$$|\Delta f| = |f(\Theta) - f(\Theta + \epsilon_\Theta)| < \delta.$$

Let $\bar{r}'_k$ be the bins of $\overline{\mathcal{H}}$ with bin heights $\Theta + \epsilon_\Theta$. Let $s'_k := |\{t : u_t \in \bar{r}'_k\}|$. Let $s_k = s'_k$ for all $k = \alpha_1, \alpha_2, \dots, \alpha_A$. Let $s_k \neq s'_k$ for $k = \beta_1, \beta_2, \dots, \beta_B$. We note that

$$\sum_{k=1}^K s_k = \sum_{k=1}^K s'_k.$$

Further,

$$\sum_{k=\alpha_1}^{\alpha_A} s_k = \sum_{k=\alpha_1}^{\alpha_A} s'_k.$$

Therefore,

$$\sum_{k=\beta_1}^{\beta_B} s_k = \sum_{k=\beta_1}^{\beta_B} s'_k.$$

By Definition 2.3 of $\overline{\mathcal{H}}$, $s_k \neq s'_k$ only if there exists at least one u_t such that

$$u_t = b_{k-1} \text{ and } \theta_{k-1} = \theta_k \text{ or}$$

$$u_t = b_k \text{ and } \theta_k = \theta_{k+1}.$$

Let $\beta_{i_j}, \beta_{i_j+1}, \dots, \beta_{i_j+I_j}$ be the adjacent bins for which $\theta_{\beta_{i_j}} = \theta_{\beta_{i_j+1}} = \dots = \theta_{\beta_{i_j+I_j}}$. Let this be true for all J groups. Then,

$$\sum_{k=\beta_{i_j}}^{\beta_{I_j}} s_k = \sum_{k=\beta_{i_j}}^{\beta_{I_j}} s'_k = q_{I_j}, \text{ for all } j.$$

Therefore,

$$\begin{aligned} |\Delta f| &= \sum_{k=\alpha_1}^{\alpha_A} s_k \log \theta_k - \sum_{k=\alpha_1}^{\alpha_A} s'_k \log(\theta_k + \epsilon_{\theta_k}) \\ &\quad + \sum_{k=\beta_1}^{\beta_B} s_k \log \theta_k - \sum_{k=\beta_1}^{\beta_B} s'_k \log(\theta_k + \epsilon_{\theta_k}) \\ &= \sum_{k=\alpha_1}^{\alpha_A} s_k \log \left(\frac{\theta_k}{\theta_k + \epsilon_{\theta_k}} \right) \\ &\quad + \sum_{k=\beta_{i_1}}^{\beta_{I_1}} s_k \log \theta_{\beta_{i_1}} - \sum_{k=\beta_{i_1}}^{\beta_{I_1}} s'_k \log(\theta_{\beta_{i_1}} + \epsilon_{\theta_k}) \\ &\quad + \dots \\ &\quad + \sum_{k=\beta_{i_J}}^{\beta_{I_J}} s_k \log \theta_{\beta_{i_J}} - \sum_{k=\beta_{i_J}}^{\beta_{I_J}} s'_k \log(\theta_{\beta_{i_J}} + \epsilon_{\theta_k}) \\ &= \sum_{k=\alpha_1}^{\alpha_A} s_k \log \left(\frac{\theta_k}{\theta_k + \epsilon_{\theta_k}} \right) \\ &\quad + \sum_{k=\beta_{i_1}}^{\beta_{I_1}} s'_k \log \left(\frac{\theta_{\beta_{i_1}}}{\theta_{\beta_{i_1}} + \epsilon_{\theta_k}} \right) \\ &\quad + \dots \\ &\quad + \sum_{k=\beta_{i_J}}^{\beta_{I_J}} s'_k \log \left(\frac{\theta_{\beta_{i_J}}}{\theta_{\beta_{i_J}} + \epsilon_{\theta_k}} \right). \end{aligned}$$

Here we observe that

$$\lim_{\epsilon_k \rightarrow 0} \sum_{k=\alpha_1}^{\alpha_A} s_k \log \left(\frac{\theta_k}{\theta_k + \epsilon_{\theta_k}} \right) = 0, \quad (2.1)$$

and

$$\lim_{\epsilon_k \rightarrow 0} \sum_{k=\beta_{i_1}}^{\beta_{I_1}} s'_k \log \left(\frac{\theta_{\beta_{I_j}}}{\theta_{\beta_{I_j}} + \epsilon_{\theta_k}} \right) = 0 \text{ for all } j. \quad (2.2)$$

Therefore, from 2.1 and 2.2, for any $\delta > 0$, there exists $|\epsilon_{\theta_k}| > 0$, for all k , such that

$$|\Delta f| = |f(\Theta) - f(\Theta + \epsilon_{\Theta})| < \delta.$$

This proves the lemma. □

Lemma 2.8. *If $f(\Theta) = \sum_{k=1}^K s_k \log [\theta_k]$ is non-differentiable with respect to Θ , then Θ cannot be a maximizer of $f(\Theta)$.*

Proof. We first show that the function $f(\Theta)$ is non-differentiable with respect to Θ when there exists at least one $u_t \in b_k$ and $\theta_k = \theta_{k+1} = \theta$ for some k . Consider the directional derivative $\nabla_V f(\Theta)$ along $V = e_k - e_{k-1}$, where $e_k \in \mathbb{R}^{K \times 1}$ are the standard basis vectors defined as $e_k := [0 \ \dots \ 1 \ \dots \ 0]^T$. (Here we assume without loss of generality that $\theta_{k-1} \neq \theta$ and $\theta_{k+2} \neq \theta$. However, in the cases where $\theta_{k-p} = \theta_{k-p+1} = \dots = \theta_k = \theta_{k+1} = \dots = \theta_{k+1+q} = \theta$ for some $p, q \geq 1$, then we redefine our direction vector V to be $V := e_{k-p} + e_{k-p+1} + \dots + e_k - e_{k+1} - \dots - e_{k+1+q}$, to arrive at the same conclusions.) Then the directional derivative of $f(\Theta)$ along V is defined as

$$\nabla_V f(\Theta) := \lim_{h \rightarrow 0} \frac{f(\Theta + Vh) - f(\Theta)}{h}.$$

Now the limit from above

$$\nabla_V f(\Theta)^+ := \lim_{h \rightarrow 0^+} \frac{f(\Theta + Vh) - f(\Theta)}{h}$$

when $\theta_k = \theta_{k+1} = \theta$ is given by

$$\begin{aligned}
\nabla_V f(\Theta)^+ &= \lim_{h \rightarrow 0^+} \frac{1}{h} s'_k \log [\theta_k + h] \\
&\quad + \frac{1}{h} s'_{k+1} \log [\theta_{k+1} - h] \\
&\quad - \frac{1}{h} \{s_k \log [\theta_k] + s_{k+1} \log [\theta_{k+1}]\} \\
&= \lim_{h \rightarrow 0^+} \frac{1}{h} \{s'_k \log [\theta + h] - s_k \log [\theta]\} \\
&\quad + \frac{1}{h} \{s'_{k+1} \log [\theta - h] - s_{k+1} \log [\theta]\}.
\end{aligned}$$

For $h \geq 0$, let $s'_k = s_k$ and $s'_{k+1} = s_{k+1}$. That is we have assumed that by default when $\theta_k = \theta_{k+1} = \theta$ the upper semi-continuous histogram places $u_t \in b_k$ in the bin k (consistent with Definition 2.3). With l'Hopital's rule, we then have

$$\begin{aligned}
\nabla_V f(\Theta)^+ &= \lim_{h \rightarrow 0^+} \frac{s_k}{h} \{\log [\theta + h] - \log [\theta]\} \\
&\quad + \frac{s_{k+1}}{h} \{\log [\theta - h] - \log [\theta]\} \\
&= \lim_{h \rightarrow 0^+} \frac{s_k}{\theta + h} - \frac{s_{k+1}}{\theta - h} \\
&= \frac{s_k - s_{k+1}}{\theta}.
\end{aligned} \tag{2.3}$$

Whereas, the limit from below

$$\nabla_V f(\Theta)^- := \lim_{h \rightarrow 0^-} \frac{f(\Theta + Vh) - f(\Theta)}{h}$$

when $\theta_k = \theta_{k+1} = \theta$ is given by

$$\begin{aligned}
\nabla_V f(\Theta)^- &= \lim_{h \rightarrow 0^-} \frac{1}{h} s'_k \log [\theta_k + h] \\
&\quad + \frac{1}{h} s'_{k+1} \log [\theta_{k+1} - h] \\
&\quad - \frac{1}{h} \{s_k \log [\theta_k] + s_{k+1} \log [\theta_{k+1}]\} \\
&= \lim_{h \rightarrow 0^-} \frac{1}{h} \{s'_k \log [\theta + h] - s_k \log [\theta]\} \\
&\quad + \frac{1}{h} \{s'_{k+1} \log [\theta - h] - s_{k+1} \log [\theta]\}.
\end{aligned}$$

Let $\varepsilon = -h$, then

$$\begin{aligned}\nabla_V f(\Theta)^- &= \lim_{\varepsilon \rightarrow 0^+} \frac{f(\Theta - V\varepsilon) - f(\Theta)}{-\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0^+} \frac{1}{-\varepsilon} \left\{ s'_k \log[\theta - \varepsilon] - s_k \log[\theta] \right\} \\ &\quad + \frac{1}{-\varepsilon} \left\{ s'_{k+1} \log[\theta + \varepsilon] - s_{k+1} \log[\theta] \right\}.\end{aligned}$$

For $h < 0$, $s'_k = s_k - C$ and $s'_{k+1} = s_{k+1} + C$, with $C \geq 1$. With this substitution and l'Hopital's rule, this implies,

$$\begin{aligned}\nabla_V f(\Theta)^- &= \lim_{\varepsilon \rightarrow 0^+} -\frac{s_k}{\varepsilon} \{\log[\theta - \varepsilon] - \log[\theta]\} \\ &\quad - \frac{s_{k+1}}{\varepsilon} \{\log[\theta + \varepsilon] - \log[\theta]\} \\ &\quad - \frac{C}{\varepsilon} \{-\log[\theta - \varepsilon] + \log[\theta + \varepsilon]\} \\ &= \lim_{\varepsilon \rightarrow 0^+} \frac{s_k}{\theta - \varepsilon} - \frac{s_{k+1}}{\theta + \varepsilon} \\ &\quad - \lim_{\varepsilon \rightarrow 0^+} C \left(\frac{1}{\theta - \varepsilon} + \frac{1}{\theta + \varepsilon} \right) \\ &= \frac{s_k - s_{k+1} - 2C}{\theta}.\end{aligned}\tag{2.4}$$

Thus, from 2.3 and 2.4, we observe that

$$\lim_{h \rightarrow 0^-} \frac{f(\Theta + Vh) - f(\Theta)}{h} \neq \lim_{h \rightarrow 0^+} \frac{f(\Theta + Vh) - f(\Theta)}{h}.$$

Thus, $\nabla_V f(\Theta)$ does not exist. Therefore $f(\Theta)$ is not differentiable when there exists at least one $u_t \in b_k$ and $\theta_k = \theta_{k+1} = \theta$ for some k .

We note that when $\theta_k \neq \theta_{k+1}$ for all k , $f(\Theta)$ is differentiable with respect to Θ regardless of the value of u_t for all t . In such as situation, s_k remains invariant for all k with small perturbations $\theta_k + \epsilon_k$, as long as for $\epsilon > \epsilon_k > 0$ and $2\epsilon < |\theta_k - \theta_{k+1}|$ for all k . Thus, within $B_\epsilon(\Theta) := \{\psi \in \mathbb{R}^K | \epsilon > |\psi_k - \theta_k|\}$, $f(\Theta)$ is linear combination of logarithm functions and hence differentiable. Now, to show that the non-differentiable points of $f(\Theta)$ cannot be its maximizers, consider the one sided directional derivative along the direction $-V$. That is

$$\nabla_{-V} f(\Theta)^+ := \lim_{h \rightarrow 0^+} \frac{f(\Theta - Vh) - f(\Theta)}{h}$$

$$\begin{aligned}
\nabla_{-V} f(\Theta)^+ &= \lim_{h \rightarrow 0^+} \frac{1}{h} s'_k \log [\theta_k - h] \\
&\quad + \frac{1}{h} s'_{k+1} \log [\theta_{k+1} + h] \\
&\quad - \frac{1}{h} \{s_k \log [\theta_k] + s_{k+1} \log [\theta_{k+1}]\} \\
&= \lim_{h \rightarrow 0^+} \frac{1}{h} \{s'_k \log [\theta_k - h] - s_k \log [\theta_k]\} \\
&\quad + \frac{1}{h} s'_{k+1} \log [\theta_{k+1} + h] \\
&\quad - \frac{1}{h} s_{k+1} \log [\theta_{k+1}].
\end{aligned}$$

When there exists at least one $u_t \in b_k$ and $\theta_k = \theta_{k+1} = \theta$, we make the substitution $s'_k = s_k - C$ and $s'_{k+1} = s_{k+1} + C$, and apply l'Hopital's rule to determine the limit. This gives

$$\begin{aligned}
\nabla_{-V} f(\Theta)^+ &= \lim_{h \rightarrow 0^+} \frac{s_k}{h} \{\log [\theta - h] - \log [\theta]\} \\
&\quad + \frac{s_{k+1}}{h} \{\log [\theta + h] - \log [\theta]\} \\
&\quad + \frac{C}{h} \{-\log [\theta - h] + \log [\theta + h]\} \\
&= \lim_{h \rightarrow 0^+} -\frac{s_k}{\theta - h} + \frac{s_{k+1}}{\theta + h} \\
&\quad + \lim_{h \rightarrow 0^+} C \left(\frac{1}{\theta - h} + \frac{1}{\theta + h} \right) \\
&= \frac{-s_k + s_{k+1} + 2C}{\theta}. \tag{2.5}
\end{aligned}$$

From 2.3 and 2.5, if

$$\lim_{h \rightarrow 0^+} \frac{f(\Theta + Vh) - f(\Theta)}{h} = \frac{s_k - s_{k+1}}{\theta} \leq 0,$$

then

$$\lim_{h \rightarrow 0^+} \frac{f(\Theta - Vh) - f(\Theta)}{h} = \frac{-s_k + s_{k+1} + 2C}{\theta} > 0.$$

This implies that $f(\Theta)$ increases with respect to Θ along the direction $-V$ [61]. Thus, Θ with $\theta_k = \theta_{k+1} = \theta$ when there exists at least one $u_t \in b_k$ cannot be a maximizer of $f(\Theta)$. This completes the proof. $\square$

Chapter 3

Single Molecule Force Spectroscopy of Muscle Proteins

In chapter 2, we provide an algorithm that can detect abrupt changes with minimal prior information needed. In this chapter, we showcase its modification for analysing data from single molecule force spectroscopy. We then present the nano-mechanical characterization of the muscle protein utrophin. We describe the details of the algorithms and tools we developed to conduct the characterization study.

3.1 Single Molecule Force Spectroscopy

Instruments such as atomic force microscopes (AFMs) and optical tweezers have enabled resolution of measuring forces in the femto to multiple pico Newton range with spatial resolution at the nanometer scale. These instruments have made it possible to perform protein folding and unfolding experiments [3, 4, 5, 6], where domains fold and unfold with force differentials in the pico-Newton and femto-Newton range. The length of the domains are in the tens to hundreds of nanometers. In force spectroscopy experiments using AFMs (see Fig. 3.1.1), a solution with protein molecules to be studied is deposited on a substrate, and a cantilever with a sharp tip is pressed against the substrate to enable a protein molecule to attach to the tip. Subsequent to pressing the cantilever against the substrate, the cantilever is retracted away from it. If there is a successful attachment the protein is extended and comes under tension. The domains of the protein unfold under the application of tension. The unfolding of a domain leads to a step change in the length of the

protein. How much time is needed for the domains to unfold, the relationship to the force applied, the changes in the structure under folding and unfolding events are questions that are studied with AFM based force spectroscopy. Most of the studies involve numerous iterations of experiments. Often, data is collected from a large number of experiments, not just to ensure a high confidence on statistical information, but also because experiment success rates are sometimes intentionally kept low. For example, in single molecule force spectroscopy, the protein solution is diluted enough to ensure that the the success rate of an attachment forming between the tip and protein is between two to ten percent [32, 62]. The low probabilities of attachment ensure that the probability of attaching to multiple protein molecules is low [32, 62]. The task of obtaining accurate yet precise statistics of events under large uncertainties and the confounding dynamics of the instrumentation is a daunting one. For example, in AFM experiments, the protein domains can unfold at rapid rates where the dynamics of the cantilever cannot be ignored. The timing of domain unfolding or folding and changes in protein structure have to be ascertained from the measured cantilever deflection data which is corrupted by effects of thermal noise (process noise), measurement noise and the dynamics of the cantilever probe. It is evident that the needs of fundamental biophysics research necessitate the development of tools for automatic detection and estimation of abrupt changes in measurements and parameters of models used to describe the experimental data. In addition, these tools must be capable of operating with minimal prior information on the systems being studied. Here we apply the change detection algorithm discussed in Chapter 2 for the analysis of data from atomic force microscope (AFM) based protein pulling experiments.

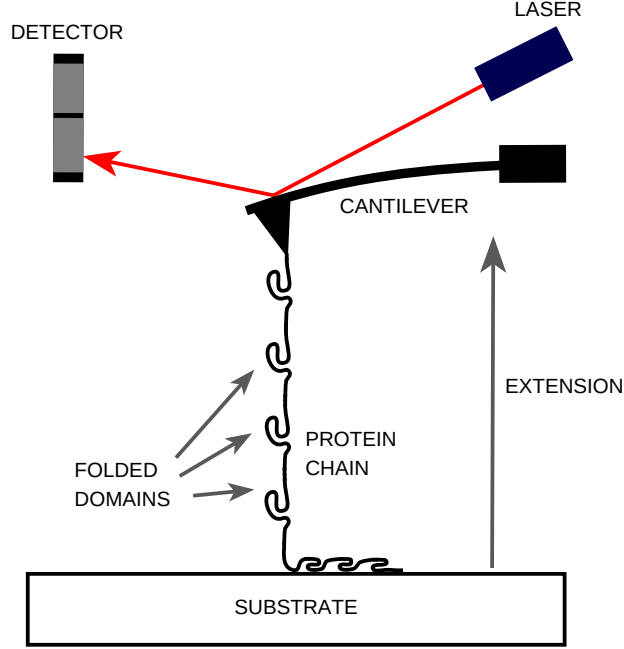


Fig. 3.1.1: Schematic of a protein pulling experiment. A polypeptide chain protein is stretched between a substrate and a cantilever tip of an AFM.

3.1.1 Protein Pulling Experiments using AFMs

In a protein pulling experiment, a single molecule of a poly-peptide chain protein is attached between a substrate and a cantilever tip of an AFM as shown in Fig. 3.1.2. The distance z between the cantilever tip and the substrate is measured and is controlled by a piezoelectric transducer. The deflection d in the cantilever is measured via a laser photo-diode sensor arrangement. Thus, the extension ($x = z - d$) can be calculated. The deflection in the cantilever is multiplied by its spring constant to convert it to the force f in the protein. During an experiment, the distance z is increased at a constant speed, so as to apply an increasing tensile force on the protein. It is known (see [2]) that force f is well characterized by the force-extension relation of the worm like chain model f^{WLC} , described by (3.1).

$$f^{WLC}(x, l, p) = \frac{k_B T}{p} \left[\frac{1}{4} \left(1 - \frac{x}{l} \right)^{-2} - \frac{1}{4} + \frac{x}{l} \right], \quad (3.1)$$

where k_B is the Boltzmann constant, and T is the temperature. Here, l , which represents the contour length and p , which represents the persistence of the protein are the free parameters whose estimates are of interest in the experimental study. Further,

during a pulling experiment, folded domains in the protein unfold rapidly, thereby abruptly changing both the contour and persistence lengths [32]. The data analysis then involves the detection of such abrupt changes in the contour and persistence lengths.

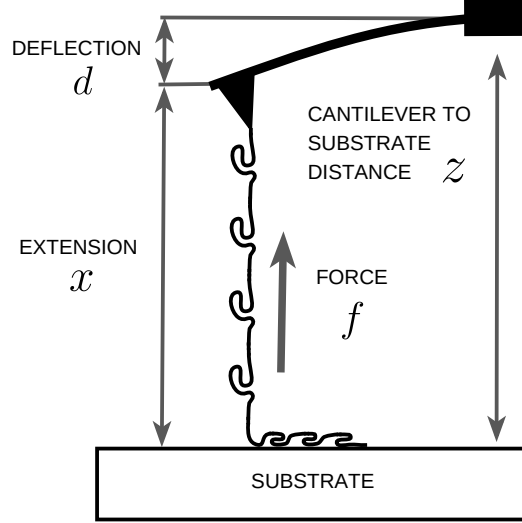


Fig. 3.1.2: Measurements in a pulling experiment. The deflection in the cantilever as well as the distance between the tip and substrate are measured, allowing the force versus extension characteristics to be determined.

3.2 Change Detection for Automatic Data Analysis

Our task now is to find the estimates $\hat{L}$ and $\hat{P}$ of $L = \{l_1, l_2, \dots, l_T\}$ and $P = \{p_1, p_2, \dots, p_T\}$ given the measurements of force $F = \{f_1, f_2, \dots, f_T\}$ and extension $X = \{x_1, x_2, \dots, x_T\}$. While experimental data shows that the noise in the measured force is stationary, the corresponding noise in the contour and persistence lengths when inverting the WLC function is non-stationary and dependent on the extension. Further, the percentage change in persistence length is typically smaller than the percentage change in the contour length, leading to a poorer SNR in the estimates of p compared to those of l . Due to these challenges, the change detection algorithm is reformulated as follows:

M-Stage: Minimize the objective function $J^{WLC}(L, \Theta^{(n-1)})$ with respect to L to

give $\hat{L}^{(n)}$, that is

$$\hat{L}^{(n)} = \arg \min_X J^{WLC}(L, \Theta^{(n-1)}),$$

where, $J^{WLC}(L, \Theta^{(n-1)})$ is defined as

$$\sum_{t=1}^T \left[f_t - f_t^{WLC}(x_t, l_t, p_o) \right]^2 + W_t(L, \Theta^{(n-1)}),$$

with

$$W_t(L, \Theta^{(n-1)}) = \begin{cases} 9\sigma^2\bar{\delta}(l_t - l_{t-1}) & n = 1 \\ -2\sigma^2 \log \left[p(u_t | L_1^{t-1}, \Theta^{(n-1)}) \right] & n > 1 \end{cases}.$$

E-Stage: Evaluate $\Theta^{(n)}$ using $\Theta^{(n)} = \arg \max_{\Theta} p(U^{(n)} | \Theta)$. Here σ^2 is the variance of noise in the force measurements F , n the iteration count, and p_o is an initial guess of the mean persistence length. The steps u_t are redefined as $u_t = l_t - l_{t-1}$. This reformulation allows us to determine the steps in L . Since the step locations in L are caused by unfolding events, the locations of the steps in P must be the same. At this stage, the regions of the force versus extension curves corresponding to the step changes in L and P are extracted. Local estimates for both are then recomputed using a least squares fit to improve the estimation accuracy.

3.2.1 Effectiveness of Data Analysis: Simulation

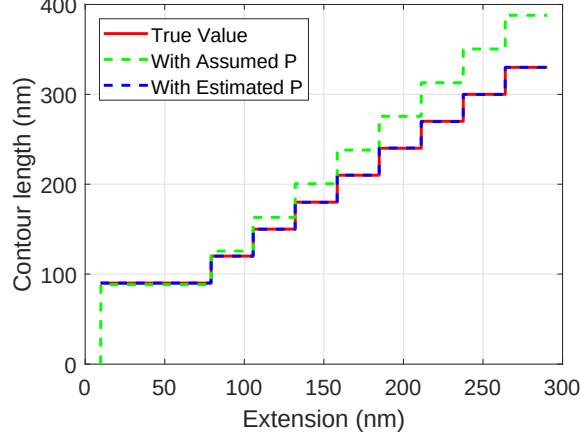


Fig. 3.2.1: Estimates of the contour length with the reformulated SDA terminated at 20 iterations. The estimates in green were obtained with the persistence length assumed constant, which leads to the expected error in the contour length magnitudes. The estimates recomputed locally to improve accuracy are shown in blue. This is done after estimating the corresponding persistence lengths.

The reformulated SDA was applied to simulated data containing 8 unfolding events with additive Gaussian noise ($\mathcal{N}(0, \sigma = 10 \text{ pN})$) in the measured forces. The contour lengths were incremented from 90 nm in steps of 30 nm, and the persistence lengths incremented from 115 pm in steps of 35 pm. These simulation parameters were chosen to represent a typical scenario encountered in our experiments.

In Fig. 3.2.1, the estimates $\hat{L}$ from the SDA are compared with the ground truth. The algorithm was initialized with $p_o = 130 \text{ pm}$, and thus the errors in the estimate for $\hat{L}$ are lesser in the region where P was close to this value. Although, the errors appear to increase in other regions, we note that the location of steps in the estimate match those of the truth. These regions are then used to find $\hat{P}$ and improve the estimates of $\hat{L}$, as shown in Fig. 3.2.2 and Fig. 3.2.1 respectively. Fig. 3.2.3 compares the true and measured forces with those computed from the estimates $\hat{L}$ and $\hat{P}$.

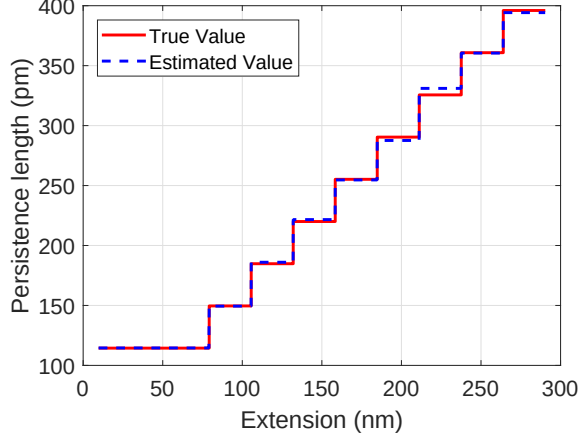


Fig. 3.2.2: Persistence length estimates compared with the actual values when applying the SDA on simulated data.

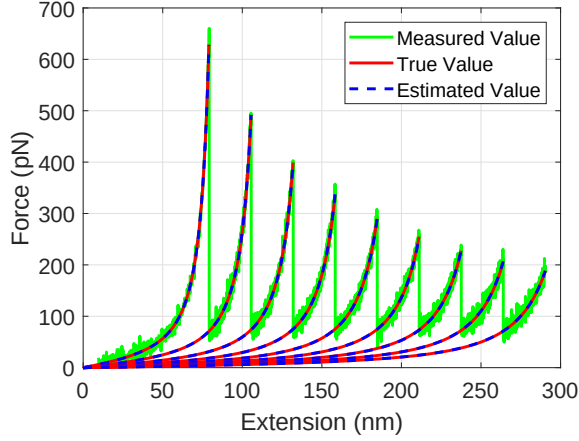


Fig. 3.2.3: Estimates of the force versus extension curve compared with the true and measured curves from the simulated system.

3.2.2 Effectiveness of Data Analysis: Experiments

The reformulated SDA was used to analyze the data from single molecule force spectroscopy on the I27O AFM reference protein from AthenaES, which is a combination of 8 repeats of the domain Ig 27 from human titin. The MFP-3D AFM from Oxford Instruments was used with Bio-cantilevers (BL-RC-150VB) from Asylum Research with a mean spring constant of 6 pN nm^{-1} . Further, the actual cantilever spring constant is measured by analyzing their thermal fluctuations. All the experiments were conducted at a temperature of 298 K, with the protein solution applied to a

fresh gold substrate. The AFM cantilever's tip is then pressed against the substrate with a force in the range of 0.5nN to 2.0nN for a duration of 3s. The tip is retracted away at a constant speed. Whenever different ends of a segment of the protein are adsorbed to the tip and to the substrate, a tensile force is applied on the segment. To ensure that only a single molecule is present between the tip and the substrate, the concentration of the protein solution is kept low (typically below 200 nM l^{-1}).

The estimates of the most probable persistence length (P) and the most probable contour length per domain (ΔL) from analyzing over 3000 experiments with a 6% success rate are compared in Table 3.2.1 with the reported values in [1, 63]. A sample force curve and its estimation is shown in Fig. 3.2.4. The automated analysis of the data using the SDA yielded estimates within acceptable tolerances.

Table 3.2.1: Titin Parameters: Reported vs SDA estimates

	Reported	Estimated with SDA	% Error
ΔL (nm)	28.4 (see [1])	28.04	-1.3
P (pm)	300 ± 70 (see [63])	292	-2.7

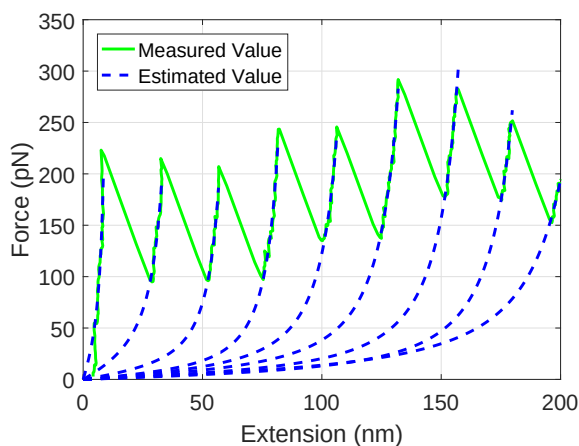


Fig. 3.2.4: Sample force versus extension curve from pulling experiments on titin. The force estimates from the WLC models fit after each unfolding event are overlaid on the measurements.

3.3 Muscular Dystrophy

Muscular Dystrophies (MD) are a set of genetic disorders that result in partial functioning or degradation of skeletal and cardiac muscles [34]. Duchenne Muscular Dystrophy (DMD) is the most common form of MD, occurring in 1 in 3500-6000 males at birth [9, 10]. DMD has devastating consequences for those affected, who experience progressive muscle weakness from early childhood. Many are wheelchair bound by age 12 and suffer fatal cardiac or respiratory failures in their late teens to mid twenties [34]. With medical assistance their life expectancy is about 27 to 28 years [10, 64].

Becker Muscular Dystrophy (BMD) is a less severe and rarer form of MD occurring in about 1 in 18000 male births [35]. Both DMD and BMD have been linked to a defect in the same gene (called the DMD gene) that codes for dystrophin [34], a muscle protein. The DMD gene is located on the X chromosome, hence primarily affects the male population, although 5-10% of the female carriers of the gene are reported to show some symptoms [34, 9]. The Duchenne and Becker Muscular Dystrophies have no cure.

3.4 Dystrophin and Utrophin: Structure and Function

Dystrophin is present in striated muscle cells and interacts with the dystrophin-glycoprotein complex (DGC), a set of proteins found at the membrane of muscle cells [65]. The disease causing defects in the DMD gene lead to the loss of function or absence of dystrophin, and subsequently to muscle cell membrane damage and muscle fiber death [66]. Dystrophin is hypothesized to act as a molecular spring or shock absorber to mechanically stabilize the muscle cell membranes during muscle contractions [67].

Utrophin is a muscle protein homologous to dystrophin and is found in muscles during fetal development and is eventually replaced by dystrophin after birth [68]. Up-regulation of utrophin is being investigated as a potential therapy for DMD [69]. Since dystrophin and utrophin have important mechanical functions in the muscle cell, atomic force microscopy (AFM) based single molecule force spectroscopy provides a crucial methodology for their nano-mechanical characterization. Only three studies reported the mechanical characterization of dystrophin prior to this work and were

limited to small recombinant fragments of dystrophin [70, 71, 72]. There were no studies on the mechanical characterization of larger fragments and full length utrophin prior to the work described in this chapter [73].

Dystrophin and utrophin are long chain like molecules, composed of folded domains linked by non-linear springs. When stretched, the non-linear springs provide a restoring force, and the folded domains unfold, abruptly changing the restoring force. Here, by 'unfold', we refer to the process in which a protein loses its native 3-dimensional structure. Specifically, dystrophin consists of an N-terminal actin binding domain (ABD1), a central rod domain consisting of 24 triple helix spectrin-type repeats (R1 to R24), a second actin binding (ABD2, contained in spectrin type repeats R11 to R17), and a C-terminal (CRCT) domain, with 5 unstructured 'hinge' regions as shown in Fig. 3.4.1A [67]. Utrophin consists of an N-terminal domain, a central rod domain consisting of 22 triple helix spectrin-type repeats (R1 to R22), a cysteine rich domain (CR), and a C-terminus (CT), with 5 unstructured hinge regions as shown in Fig. 3.4.1B [74].

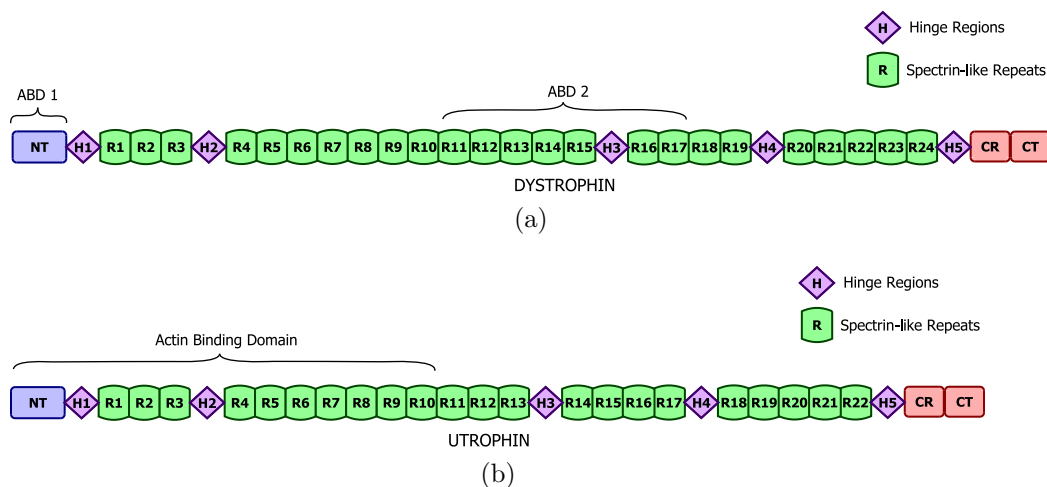


Fig. 3.4.1: Schematic of (a) dystrophin and (b) utrophin.

3.5 Single Molecule Force Spectroscopy of Utrophin

3.5.1 Atomic Force Microscopy Protocol for Utrophin Characterization

The single molecule force spectroscopy experiments were performed utilizing a MFP-3D atomic force microscope (AFM) from Oxford Instruments. The AFM setup contained a flexible cantilever with a sharp tip, a laser-photodiode based sensor which measures the position of the cantilever tip, and a piezo electric nano-positioner which can move the substrate in three spatial directions with respect to the cantilever base [75]. We used a soft BioLever (BL-RC-150VB) from Asylum Research with a typical spring constant of 6 pN/nm and a tip radius of 25 ± 12 nm coated with Cr/Au. A droplet (100 μ l) of purified protein solution (at concentrations ranging from 12.5–200 nM) was deposited on freshly cleaved mica substrate and allowed to settle for 10 minutes before commencing the experiment to ensure adsorption of proteins to the mica surface. The cantilever tip was made to approach the mica surface, pressed against the surface for 3 seconds with a force ranging from 1000 – 2000 pN and then retracted with a specified velocity. For the experiments on the utrophin constructs, a retraction velocity of 1 μ m/s was used. The approach-retraction cycle was repeated multiple times. During such a cycle, if a part of a protein molecule was attached to the cantilever tip with another part adsorbed on the substrate, the intermediate segment was stretched during the retraction phase of the cantilever. The pulling force on the molecule was balanced by the force on the cantilever which is estimated from the product of the deflection of the cantilever tip and the spring constant of the cantilever. The spring constant of the cantilever was determined before the pulling experiments by analyzing the thermal response of cantilever deflection [76]. Forced extension caused the folded domains in the molecule to unfold, characterized by the saw-tooth pattern observed in the cantilever deflection vs separation curve [2]. To eliminate possibilities of pulling on multiple proteins, initial force extension data were monitored for instances where the number of unfolding events exceeded available domains for a single molecule. The sample was diluted until the number of unfolding events remained in the expected range. Additionally, the success rate of pulling was maintained between 5 – 25%. Data from 300 – 500 successful force spec-

troscopy experiments (with at least one identifiable unfolding event) were collected for each protein construct which were used to determine the statistical behavior of the unfolding forces.

3.5.2 Data Analysis and Contour Length Filters

Data from each of the successful protein pulling experiments were collected and analyzed using a heuristic algorithm written in the MATLAB software from MathWorks®. Due to the irregular nature of the full length proteins utrophin and spectrin, which differs from the engineered protein Titin I270, the heuristic algorithm was found to be more effective than the reformulated SDA. Collected data included the separation of the cantilever base from the surface of the substrate from which the extension of the molecule was calculated. Data also included the measured deflection of the cantilever tip as a function of separation. Since the cantilever deflection is proportional to the force applied, the measured deflection can be converted to the force acting on the cantilever by multiplying the spring constant of the cantilever. Any observed force peak in the data was required to be greater than three standard deviations from the approach curve (3σ) to be considered a true unfolding event. The first significant peak is corrupted by the adhesion forces between the tip of the AFM cantilever and the substrate/protein and thus was discarded. Similarly, the last significant peak was not considered for analysis as it most likely represents the detachment of the protein molecule from the tip of the cantilever. The extension of protein molecules in between successive unfolding events was fit with the worm like chain (WLC) models (see 3.1). Parameters of the model are the persistence length (P) and the contour length (L), with statistics obtained for each utrophin construct (see Fig. 3.8.7-3.8.8). As a validation of our experimental setup and data analysis protocols, we obtained the statistics of the persistence length and contour length for the protein titin I270. Values of the most likely unfolding force, contour length, and persistence length were found to be 212 pN, 26.4 nm, and 304 pm respectively, which matched well with the reported values [1] of 225 pN, 28.4 nm, and 390 pm, respectively (see Fig. 3.8.1). To further lessen the impact of measurement noise and to remove chances of analysis and conclusions getting corrupted by pulling of multiple protein strands the analysis was repeated on filtered data. Three filters of increasing strengths, Type A, B, and C, based on the expected contour length were used. A single force-extension curve considered for analysis consisted of one or more unfolding events. The contour length

increment for unfolding event i is defined as $\Delta L_{c_i} := L_{c_i} - L_{c_{i-1}}$, where L_{c_i} is the contour length before unfolding event i occurs, measured in nm. Events with a negative contour length increment were removed from further analysis even when filters were not used. The Type A filter discards an entire force-extension curve if it contained at least one event with a negative contour length increment estimate. The Type B filter, in addition to Type A filtering, removes any events whose estimated contour length was outside expected bounds. The Type C filter, in addition to Type A filtering, also discards the entire force-extension curve if it contained at least one event with estimated contour length falling outside expected bounds. The expected contour length contribution for a domain was taken as the product of the number of amino acid residues in the domain, which is known, with the average contour length contribution of a single residue (0.299 nm [2]). For the first unfolding event, the smallest domain (with the smallest number of amino acids) in the molecule formed the lower bound of the contour length estimate, and the largest domain, the upper bound. Similarly, for the second unfolding event, the lower bound was obtained from the two smallest domains in the molecule, and the upper bound from the two largest domains. The contour length bounds for the subsequent n^{th} event were similarly computed from the contour length contributions of the n smallest and n largest domains in the molecule. We further applied a tolerance factor of 50% to the bounds obtained from expected contour length increments to accommodate the errors in the parameters estimates due to measurement noise. The presence of unstructured regions within the central rod domain of utrophin, termed “hinges”, are not well characterized and it is unknown whether these domains contribute to unfolding events. In our filtering process, we used the expected contour length increments of these hinge regions ranging from 5.4 – 22.8 nm as the lower limit, which includes the traditional hinge regions 1-4 depicted in the schematic (Fig. 3.6.1a) as well as an unstructured region between repeats 13 and 14 [74]. The filtered data lead to the same results as the unfiltered data, further reinforcing our conclusions (see Supplementary Fig. S14, S15, and S16 of [73]).

3.5.3 Kolmogorov-Smirnov Analysis

The sample distributions ($D_i = \{f_1, f_2, \dots, f_{n_i}\}$) of unfolding forces of different unfolding events were compared using the Two Sample Kolmogorov-Smirnov (KS) test. When comparing two sample distributions D_i and D_j , the KS metric (H_{KS}) is a

measure of the distance between them [60]. H_{KS} is given by Equation 2,

$$H_{KS} = \max_f |S_i(f) - S_j(f)|,$$

where, $S_i(f)$ is the empirical distribution function of the sample force distributions (D_i) defined by: $S_i(f) := \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbb{I}_{f_k \leq f}$, where, n_i is the total number of elements in D_i . Here $\mathbb{I}_{f_k \leq f}$ is the indicator function defined by:

$$\mathbb{I}_{f_k \leq f} = \begin{cases} 1 & \text{if } f_k \leq f, \\ 0 & \text{otherwise.} \end{cases}$$

H_{KS} ranges from 0 to 1, with 0 indicating identical distributions, and 1 indicating maximally dissimilar distributions.

3.6 Utrophin Characterization Results

We obtained AFM data for a titin reference protein (Fig. 3.8.1) and spectrin (Fig. 3.8.2) that corroborate previously published values [1, 2]. AFM data for Utr NT-R10, Utr R11-CT, and utrophin, all showed characteristic saw-tooth patterns of individual domain unfolding (Fig. 3.6.1b-d). From representative force versus extension curves and from the force vs unfolding event statistics (Fig. 3.8.3 and Fig. 3.6.2), we observe that the unfolding forces for utrophin proteins was significantly larger than the unfolding forces reported for spectrin and for fragments of dystrophin (for example, 90–130 pN in Fig. 3.8.4 versus < 50 pN [70, 71, 72]). To rule out the possibility of pulling multiple proteins, we adjusted the protein concentrations used for pulling experiments. When using higher concentrations, a large fraction of the cases were observed where the number of unfolding events exceeded the number of available domains in a single molecule, suggesting unfolding of multiple proteins is being observed (Fig. 3.8.5). The protein concentrations used in our experiments were lowered until such observations were absent. The concentrations were further reduced to ensure that the probability of successful pulls were in the 5 – 25% range, a success percentage used in previous studies to greatly reduce the probability of pulling multiple proteins [32]. Further evidence of single molecule pulling is demonstrated by estimating the contour length increase resulting from unfolding events, which was

found to be uniform (Fig. 3.8.6). Thus, multi-protein pulls, if they exist, contribute only minimally to the data we report. To determine statistical consistency, we collected data for 300–500 successful pulling experiments (with at least one identifiable unfolding event) for each protein construct and analyzed distributions for unfolding force, contour length increment, and persistence length (Fig. 3.8.4, 3.8.7, 3.8.8). The overall unfolding force distributions of the utrophin constructs do not show significant differences (Fig. 3.8.4a-c). However, distributions of the unfolding force as a function of the unfolding event count within individual force traces revealed different behaviors for Utr NT-R10 and Utr R11-CT (Fig. 3.6.2). For Utr NT-R10, the statistics of the unfolding forces for different unfolding event counts were similar with overlapping distributions (Fig. 3.6.2a), where the unfolding force distribution for the first unfolding event is similar to all unfolding events between the second and the eighth, demonstrating a brittle behavior. Force distributions for Utr R11-CT (Fig. 3.6.2b) and full-length utrophin (Fig. 3.6.2c) were right-shifted and broadened as the unfolding event increased, demonstrating a stiffening spring behavior. To quantify these differences, we performed the Kolmogorov-Smirnov (KS) test that gives a measure of similarity between two distributions (Fig. 3.6.3) [60]. A KS distance of zero (green) between a pair of distributions indicates identical distributions and a value of one (red) indicates maximally different distributions. The results of this statistical test comparing the distribution of each unfolding event with every other unfolding event within each protein construct and across protein constructs demonstrate that the unfolding behavior of Utr NT-R10 is statistically different from Utr R11-CT and full-length utrophin. Here, the maximum KS distance between the 1st unfolding event and any other unfolding event within Utr NT-R10 (Fig. 3.6.3a) is 0.30 in contrast to a maximum KS distance of 0.87 and 0.91 for Utr R11-CT and full-length utrophin, respectively (Fig. 3.6.3b-c). The KS tests also show that all the unfolding events of Utr NT-R10 have similar distributions of forces for unfolding (Fig. 3.6.3a) whereas Utr R11-CT and utrophin force distributions become more dissimilar with increasing unfolding events (Fig. 3.6.3b-c). KS test cross comparisons between different utrophin constructs demonstrate that the distributions of Utr NT-R10 exhibit increasing dissimilarity with increasing unfolding events compared to Utr R11-CT and full-length utrophin as indicated by heat maps (Fig. 3.6.3d-e). In contrast, the heat map comparing Utr R11-CT and full-length utrophin demonstrates similar distributions with increasing unfolding events. Overall, these data show uniform unfolding

forces upon extension for Utr NT-R10 and a nonlinear increase in unfolding forces upon extension for Utr R11-CT and full-length utrophin. The mechanical differences demonstrated between Utr NT-R10 and Utr R11-CT are in striking contrast to previously published circular dichroism data showing nearly identical thermal melt profiles for Utr NT-R10 and Utr R11-CT20. One possible explanation for the different mechanical properties of Utr NT-R10 and Utr R11-CT is that their 10 – 12 homologous spectrin-like repeats are controlled by long-range intra-protein communication from the unique ABD1 and/or CRCT terminating domains. To test the possibility of long-range effects we characterized the nano-mechanics of three new utrophin constructs (Fig. 3.6.4a) by without ABD1 (Utr R1-10), CRCT (Utr R11-22) or without ABD1 and CRCT (Utr R1-22). Additionally, we characterized a construct containing 10 spectrin-like repeats that span both the N-terminal and C-terminal halves (Utr R6-15). In the absence of terminal ABD1 and/or CRCT domains, Utr R-10, Utr R11-22, and Utr R1-22 constructs maintained the brittle, or stiffening behaviors (Fig. 3.6.4b-e) observed with the original terminal domain-containing constructs (Fig. 3.6.2-3.6.3). Application of the Kolmogorov-Smirnov test to compare the unfolding force distributions with Utr NT-R10, Utr R11-CT, and full-length utrophin confirm that the termini do not significantly contribute to the mechanical behavior of utrophin (Fig. 3.8.9-3.8.10). Interestingly, Utr R6-15 displays brittle mechanical behavior despite containing 5 spectrin-like repeats associated with the stiffening-spring behaving C-terminal half of utrophin. To summarize the AFM results, we plotted the mode of the distribution of unfolding forces as a function of unfolding events obtained from 300 – 500 successful pulling experiments for each of the six utrophin constructs (Fig. 3.6.5a, Fig. 3.8.11). Figure 3.6.5b provides a plot of the KS distance for every unfolding event compared to the distribution of the first unfolding event distribution of Utr NT-R10 and demonstrates the statistical difference between the N-terminal half (red and green plots) and C-terminal half (blue and cyan plots) of utrophin independent of their unique terminating domains (ABD1 and CRCT, respectively). Despite their high sequence and structural homology, our results suggest the sequences of spectrin-like repeats within the N-terminal and C-terminal halves of utrophin encode information that dramatically influences their respective mechanical behaviors. Additionally, Fig. 3.6.5 reveals that full-length utrophin, both with and without its terminating domains (pink and black plots), exhibit mechanical behavior directed by the C-terminal half whereas a construct containing both N-terminal and

C-terminal half spectrin-like repeats exhibits mechanical behavior directed by the N-terminal half (Utr R6-15, black star plot). The stiffening spring behavior and high forces of unfolding in utrophin are similar to the mechanical properties of titin rather than spectrin (Fig. 3.8.1-3.8.2). The biological implication of these findings is that utrophin may be too stiff to be functioning as a spring at the sarcolemma. Rather, the mechanical properties of utrophin may be suited to function as part of an in-series elastic element with titin at the myotendinous junction where it is localized in adult skeletal muscle [77, 78].

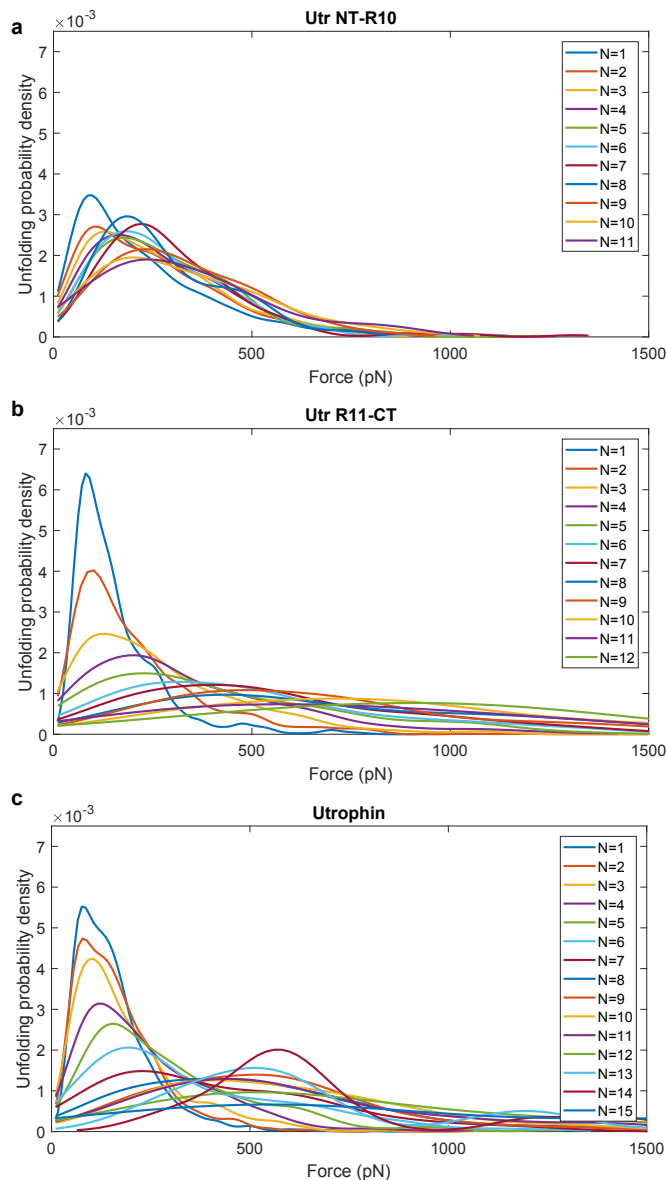


Fig. 3.6.2: Unfolding force distributions for utrophin terminal constructs reveal markedly different mechanical behaviors. a-c, Plots of the probability distribution of unfolding force vs the unfolding force for Utr NT-R10 (a), Utr-R11-CT (b), and full-length utrophin (c) collected from 3 batches of 300-500 successful pulls per batch. Here, ‘N’ represents the unfolding event count. For example, the distribution corresponding to $N = 4$ represents the distribution of the 4th unfolding event.

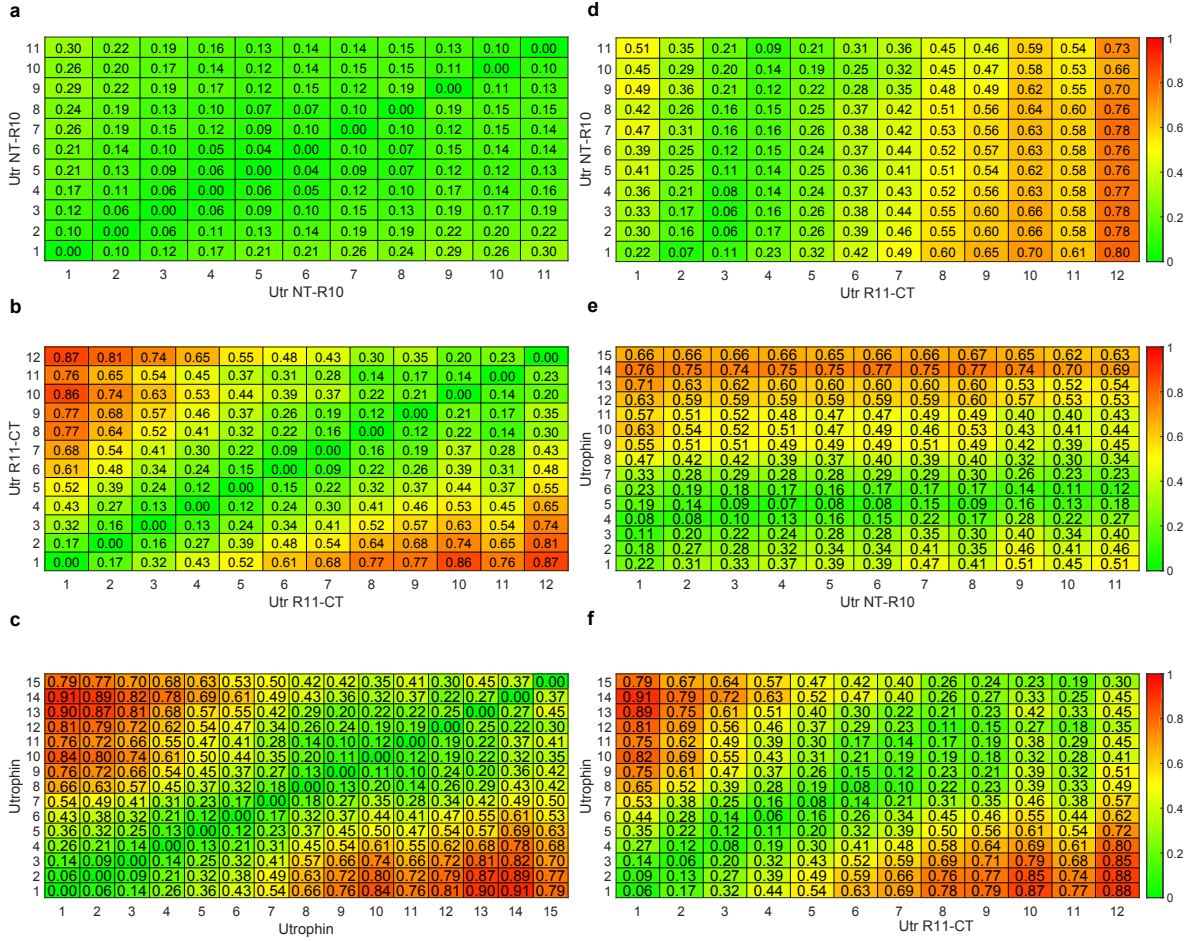


Fig. 3.6.3: Kolmogorov-Smirnov (KS) test metrics. KS test metrics which compare the distribution of unfolding forces for different unfolding event counts. The color-bars (heat maps) represent the KS metric with a 0 value (green) indicating similarity and value of 1 representing maximum dissimilarity (red). Each of the horizontal and vertical axes represents unfolding event counts from the corresponding molecule. The sub-figures (a), (b), and (c) compare the distributions within the same molecule (self-comparison). The sub-figures (d), (e), and (f) with horizontal and vertical axes from different molecules demonstrate cross-comparisons.

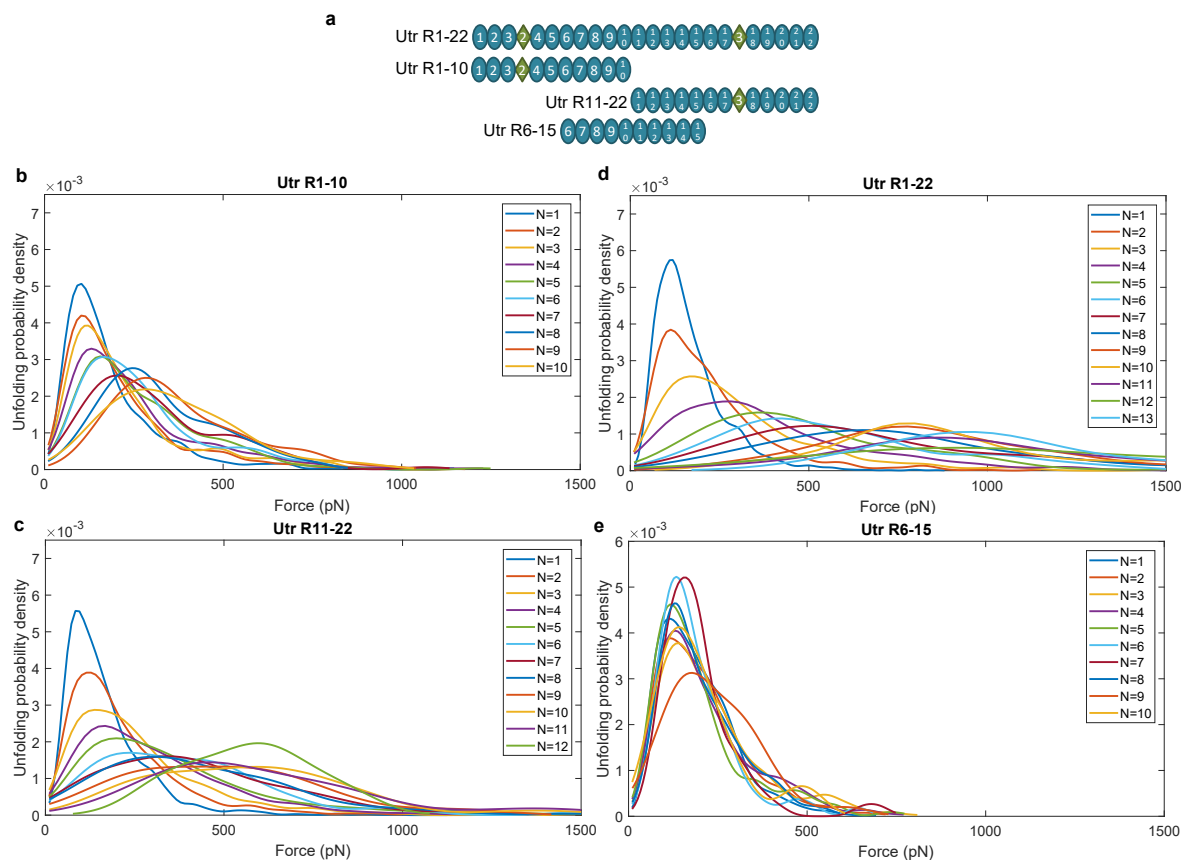


Fig. 3.6.4: Unique mechanical behavior of utrophin halves is maintained upon deletion of terminal domains. a, Schematic of utrophin constructs lacking terminal domains and a 10-repeat construct spanning both N- and C-terminal halves. b-e, Plots of probability distributions vs unfolding force of each unfolding event (N) for Utr R1-10 (b), Utr R11-22 (c), Utr R1-22 (d), and Utr R6-15 (e).

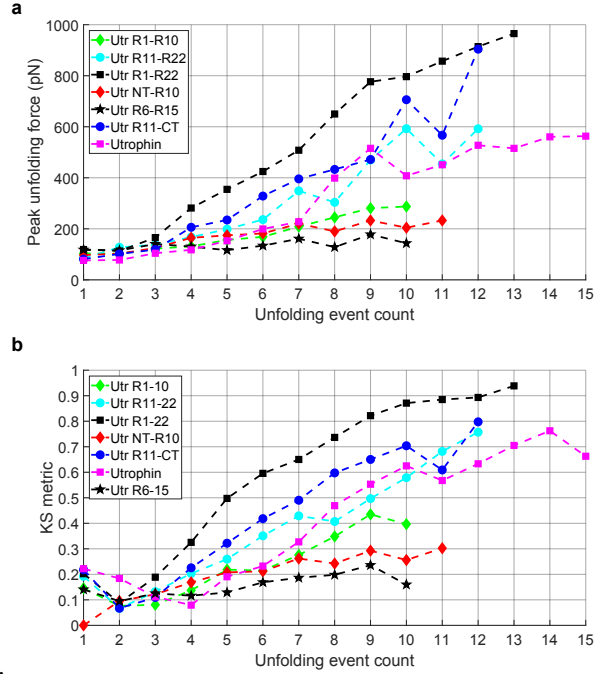


Fig. 3.6.5: Summary of utrophin mechanical behavior. a, Comparison of modes of the unfolding force distributions vs unfolding event count shows brittle vs stiffening behavior of utrophin constructs. Peak (or mode) of the distribution of unfolding forces expressed as a function of the number of domains unfolded (unfolding event count). Data are combined over 3 batches of 300-500 successful pulling experiments per batch for each protein construct and more detailed statistics are provided in Supplemental Figure S11. b, KS metrics for the distributions of unfolding force corresponding to different unfolding event counts from each of the utrophin constructs. The KS metrics are computed by using the force distribution of the first unfolding event of Utr NT-R10 as the reference distribution.

3.7 Conclusion

This chapter extends the use of the step detection algorithm for estimating parameters of a model for polymer chains. Furthermore, it shows the efficacy of the methods developed on force spectroscopy data obtained using atomic force microscopy. We report the first nano-mechanical characterization of full length and partial constructs of the muscle protein utrophin using AFM based force spectroscopy. Our experimental protocol was validated using the I27O titin reference protein and the softer human spectrin protein. Our data suggest that utrophin is stiffer than was expected based on earlier force spectroscopy data available for spectrin type repeats. Further we observe

an unexpected 'brittle' unfolding pattern for the N-terminal half of utrophin, wherein the unfolding forces do not change with increasing unfolding domains. These findings indicate utrophin might have a function similar to titin at the myotendinous junction. Post-translational phosphorylation by the insect cell expression system has now been suggested as a reason for the higher unfolding forces observed [79]. We provide the filters and algorithms that help automate the analysis of the large amounts of data that arise from force spectroscopic studies and ensure that our data are from single molecules of proteins.

3.8 Supporting Information

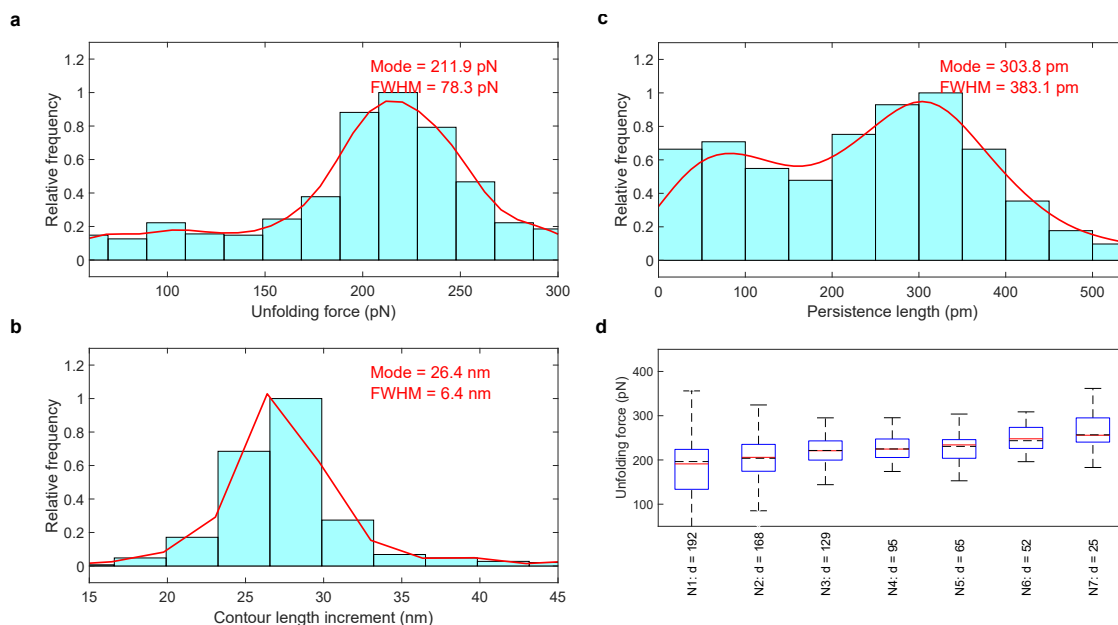


Fig. 3.8.1: AFM statistics of titin I27O (an AFM reference protein with 8 repeats of the Ig 27 domain of human titin). a, Histogram plot of unfolding forces with the mode at 212 pN when pulled at 1 $\mu\text{m/s}$ with a cantilever spring constant of 6.67 pN/nm. This matches closely with the reported value of 225 pN [1]. b, Histogram plot of contour length increments indicates that the most probable value is 26.4 nm, while 28.4 ± 0.3 nm was the corresponding reported value [1]. c, The peak persistence length was measured to be 304 pm compared to the reported value of 390 ± 70 pm [1]. d, Box plots of the unfolding forces classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. The unfolding forces for titin I27O range from 200 pN to 300 pN (a 50% increase) for unfolding counts of 1 to 8, respectively.

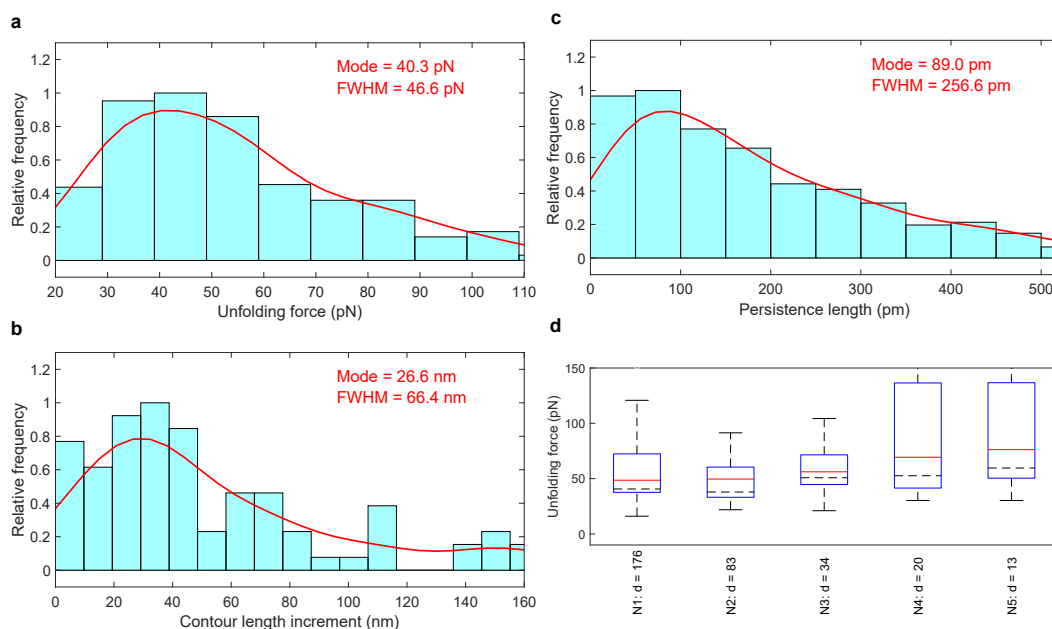


Fig. 3.8.2: AFM statistics of spectrin. a, Unfolding force histogram when pulled at $0.25 \mu\text{m/s}$ with a cantilever of spring constant 6.88 pN/nm with a most likely peak value of 40 pN compared to 28 pN reported in the literature under similar conditions [2]. b, Contour length increment histogram plot with a peak value of 26.6 nm , consistent with a value of 31.7 nm from the earlier report [2]. c, Persistence length histogram plot with a peak measured 89 pm . d, Box plots of the unfolding forces classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. The unfolding forces vary from 26 pN for the first unfolding event to 40 pN for the sixth (a 54% increase), indicating that the molecule exhibits a mildly stiffening spring behavior.

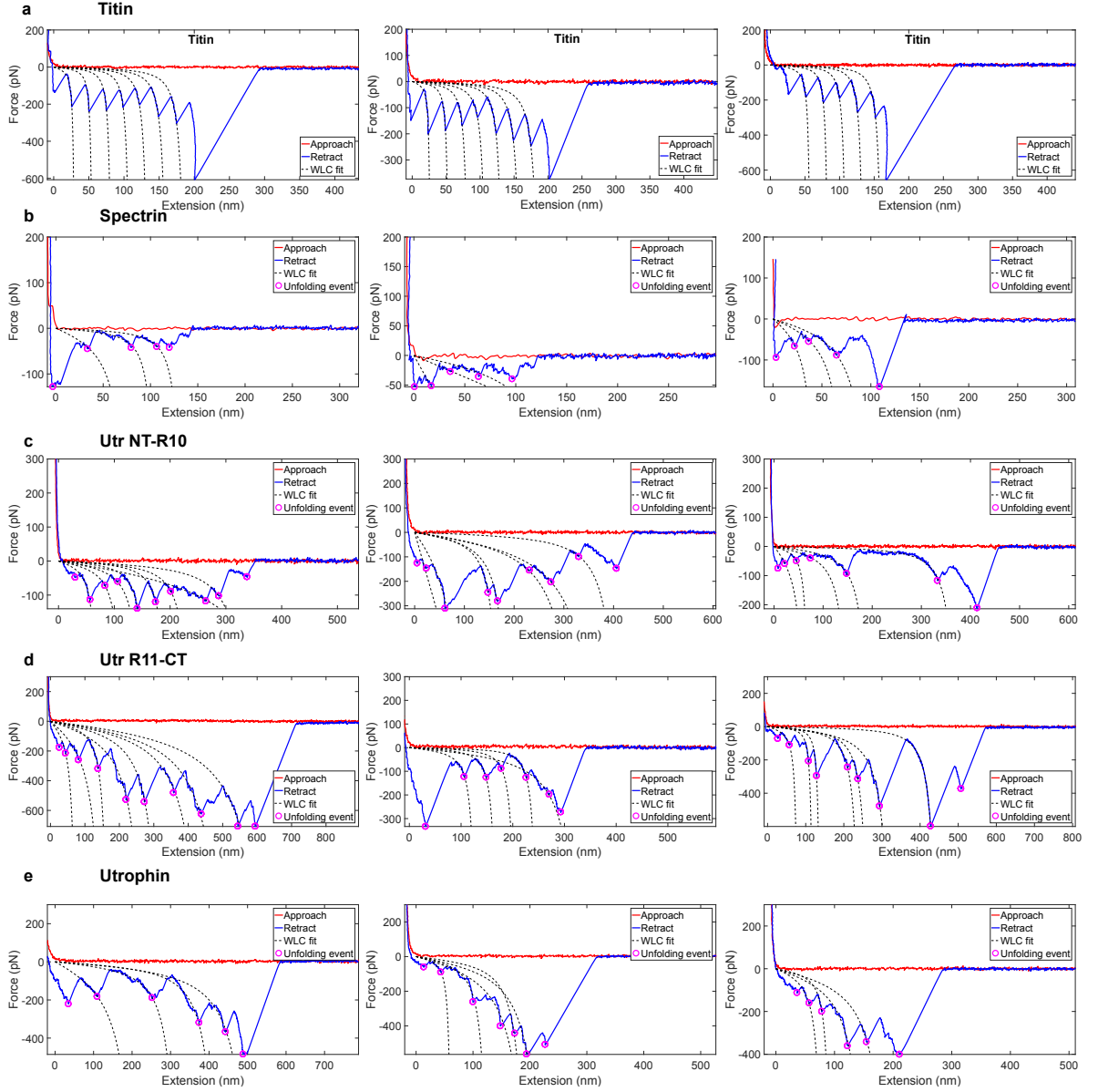


Fig. 3.8.3: Additional force vs extension curves with WLC fits. Candidate force vs extension curves of titin (a), spectrin (b), Utr NT-R10 (c), Utr R11-CT (d), and full-length utrophin (e). Approach - force on the cantilever as it approaches the substrate; retract - force on the cantilever as the molecule is extended; WLC fit - fit of the worm like chain model, which relates the force exerted on the molecule to its extension; unfolding event - force minima corresponding to domain unfolding.

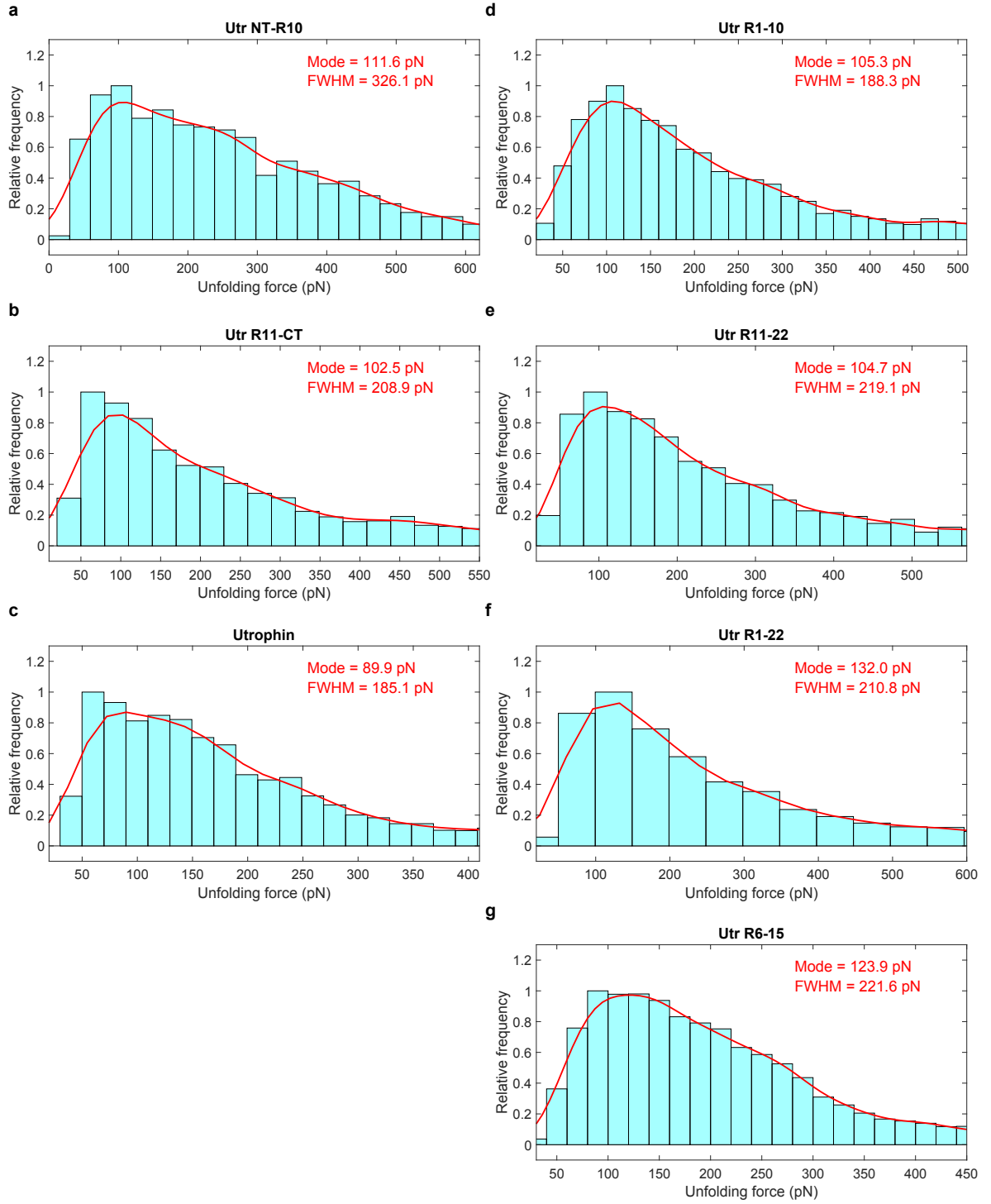


Fig. 3.8.4: Unfolding force histograms for the seven utrophin constructs. The unfolding forces are reported for a pulling speed of $1 \mu\text{m/s}$ with cantilever spring constants between $5.04 - 9.06 \text{ pN/nm}$.

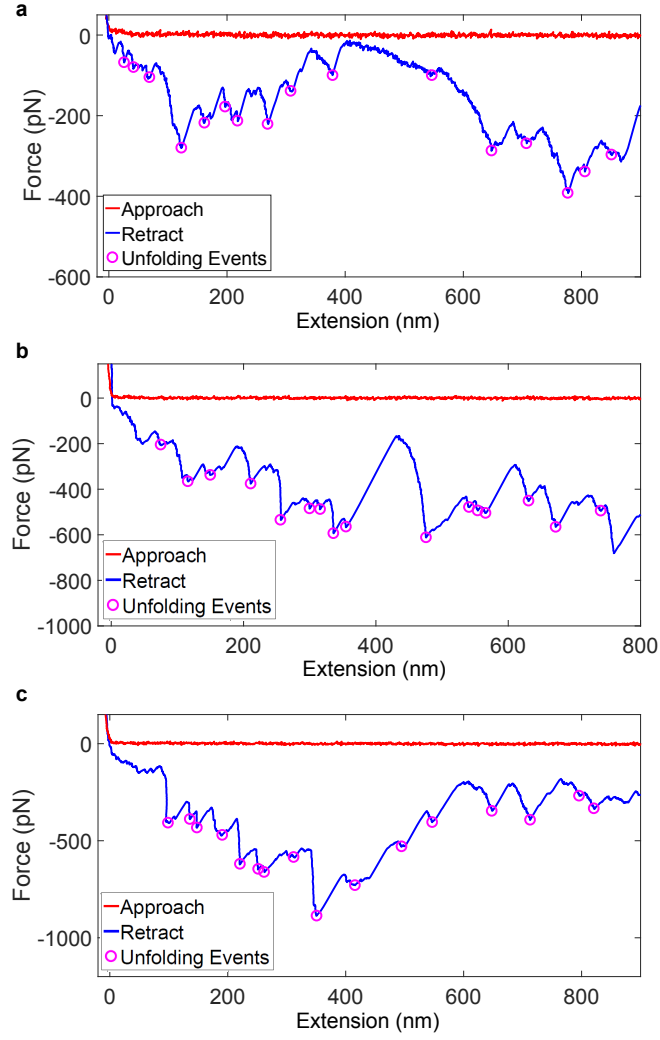


Fig. 3.8.5: Force extension characteristics of multiple protein pulls. a-c, Example force traces of Utr NT-R10 at high concentrations (500 nM) with unfolding events exceeding the number of available domains for a single molecule (11 domains for Utr NT-R10).

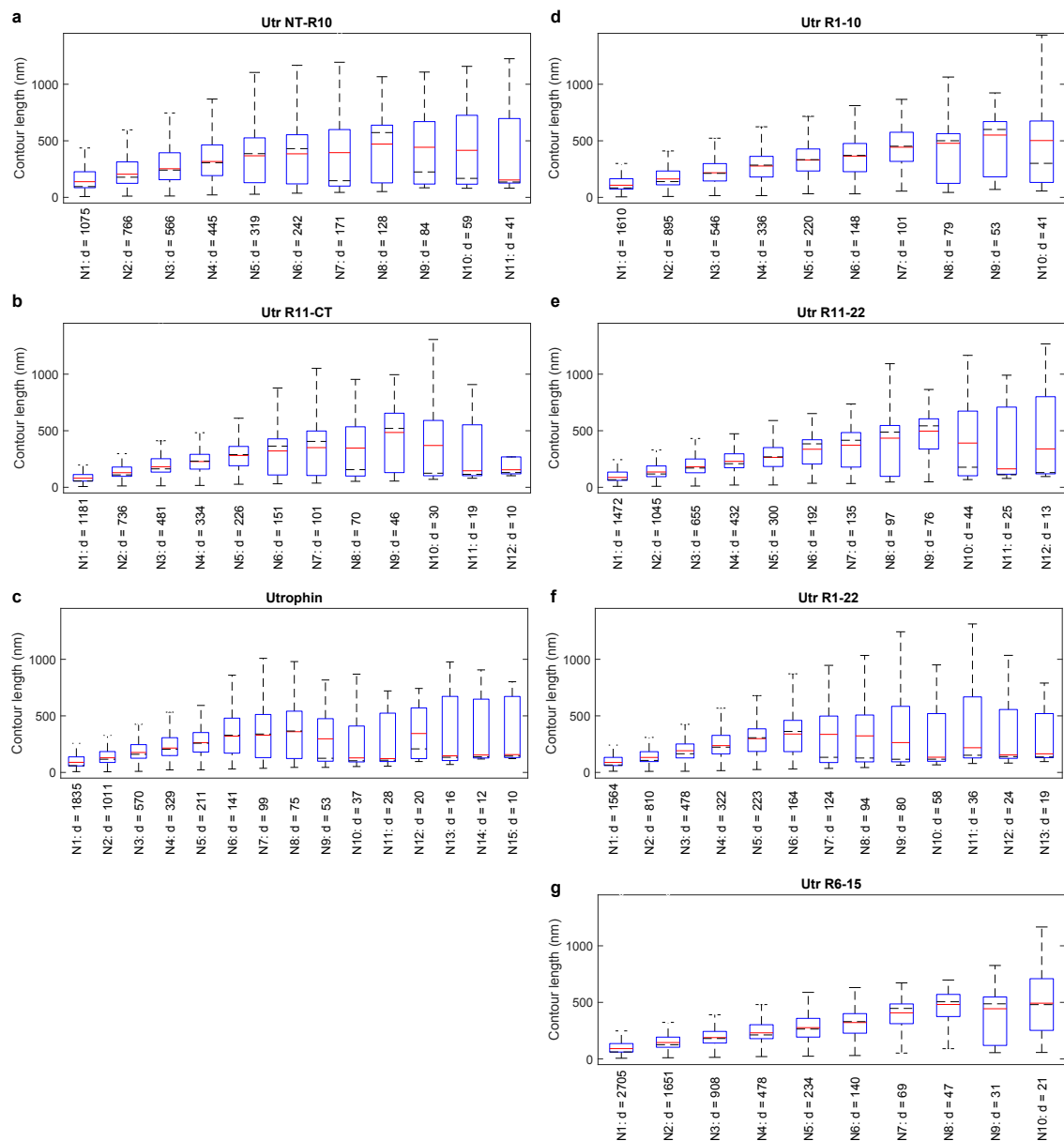


Fig. 3.8.6: Box plots of contour lengths as a function of unfolding count. Box plots of the contour length classified based on the unfolding count, with the red line indicating the median value and the dotted black line indicating the mode. The edges of the box represent the 25% and 75% percentiles, with the whisker plots marking the minimum and maximum recorded values excluding the outliers. These figures show contour length increasing linearly with unfolding events, where unfolding events are represented as ' N_i ', $i = 1, 2, 3, \dots$ and ' d ' denotes the number of pulling experiments that resulted in ' N_i ' events. A nearly linear increase in the contour length is expected as the unfolding events occur.

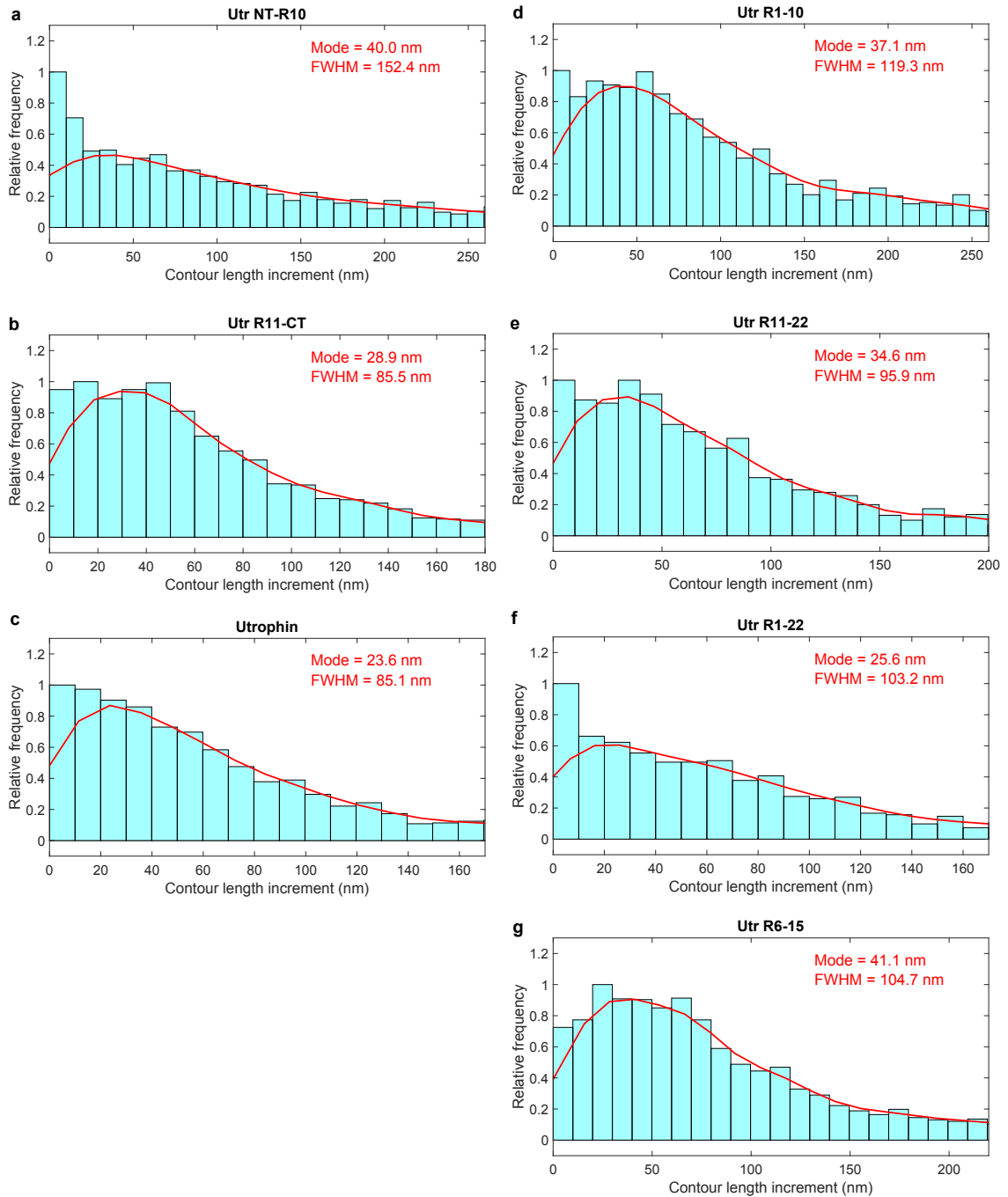


Fig. 3.8.7: The most probable contour length increments range between 23.6 to 41.1 nm. In previous reports, spectrin repeats of 106 amino acids had an average contour length increment of 31.7 nm [2] which gives a predicted contour length increment of 32.6 nm for utrophin repeats of 109 average amino acids.

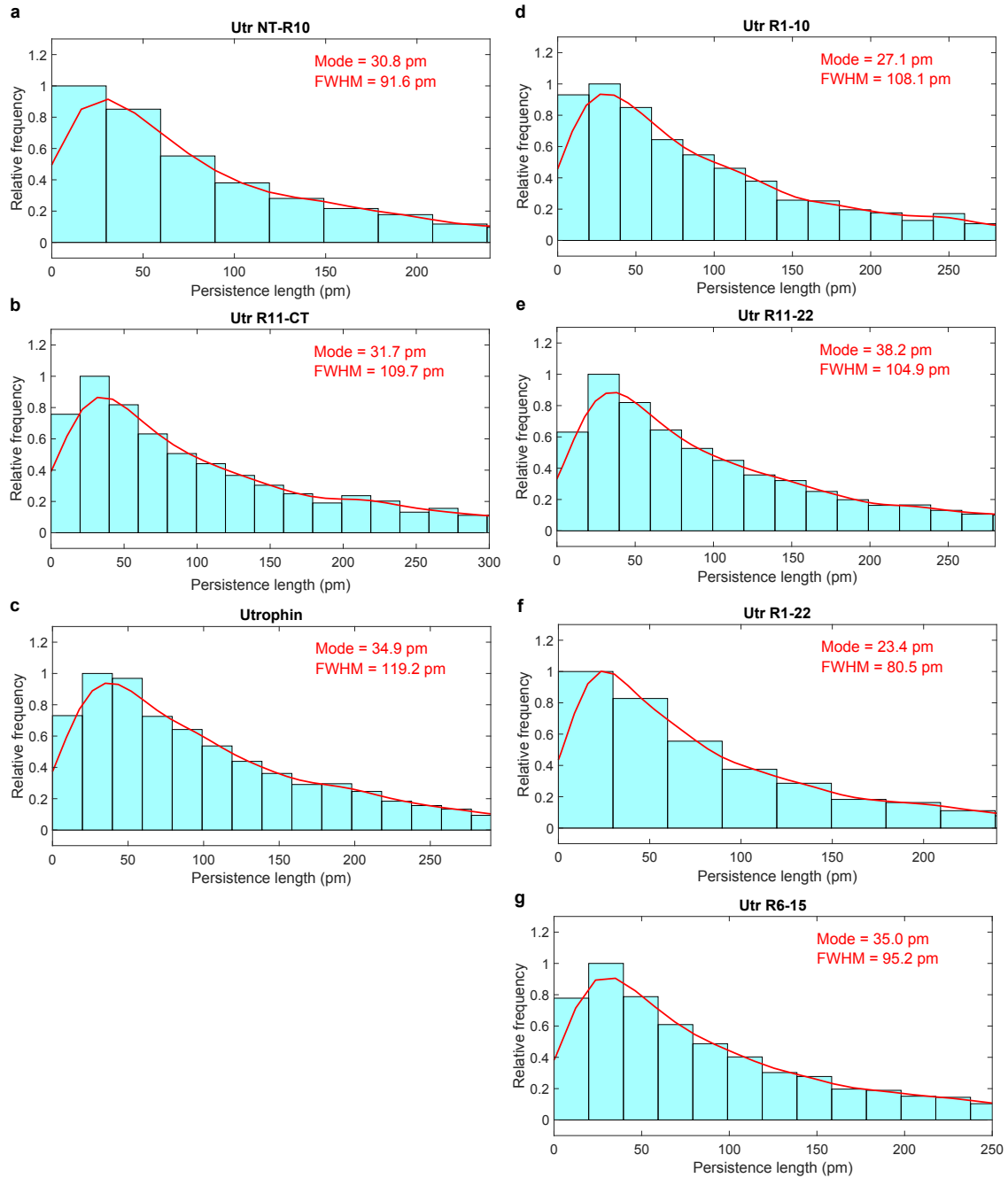


Fig. 3.8.8: Persistence length histograms for the seven utrophin constructs. The most probable persistence lengths were found to be between 23.4 to 38.2 pm for the utrophin constructs.

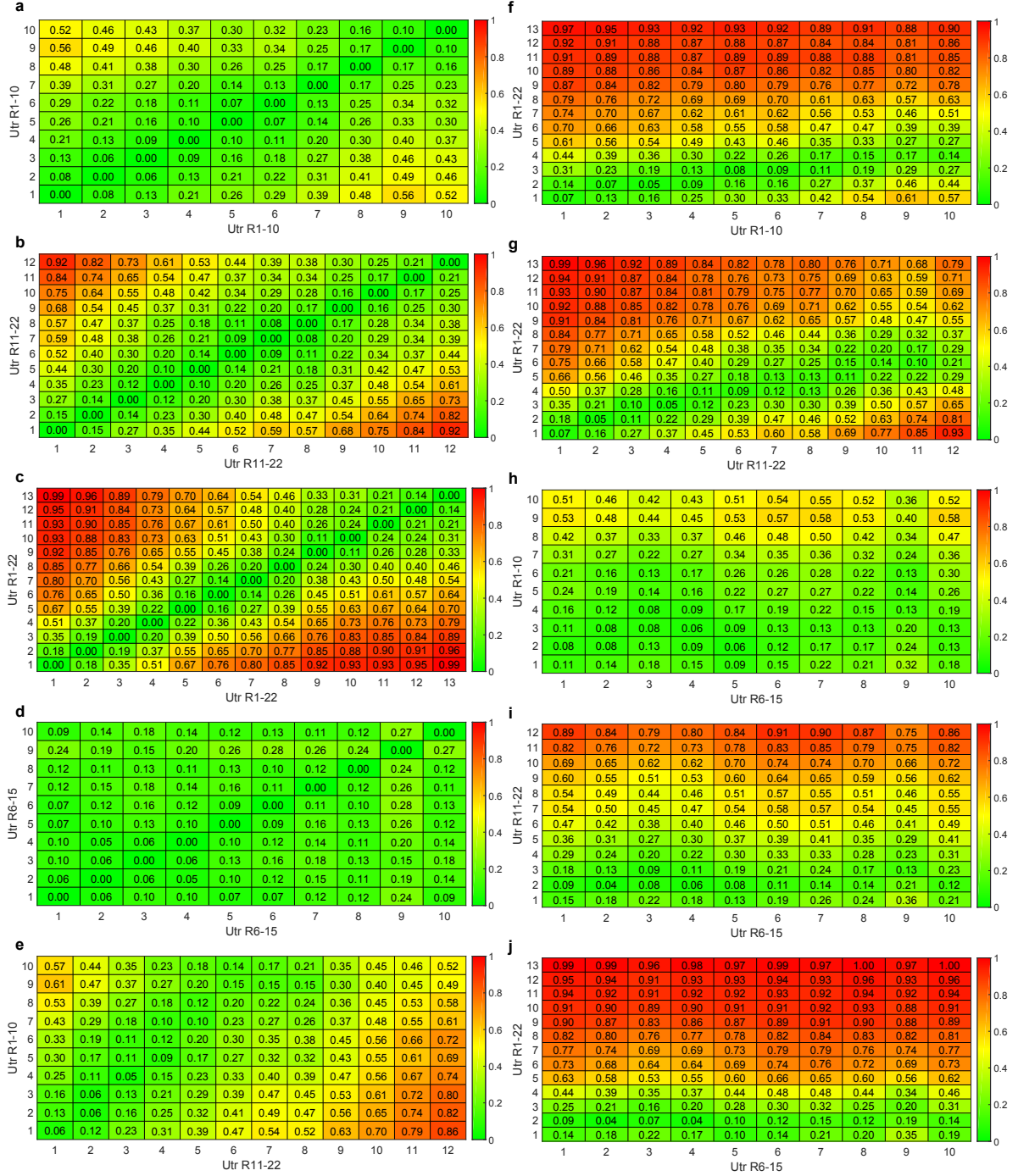


Fig. 3.8.9: Heat maps of Kolmogorov-Smirnov (KS) statistics. KS test metrics which compare the distribution of unfolding forces for different unfolding event counts. The color-bars (heat maps) represent the KS metric with a 0 value (green) indicating similarity and value of 1 representing maximum dissimilarity (red).

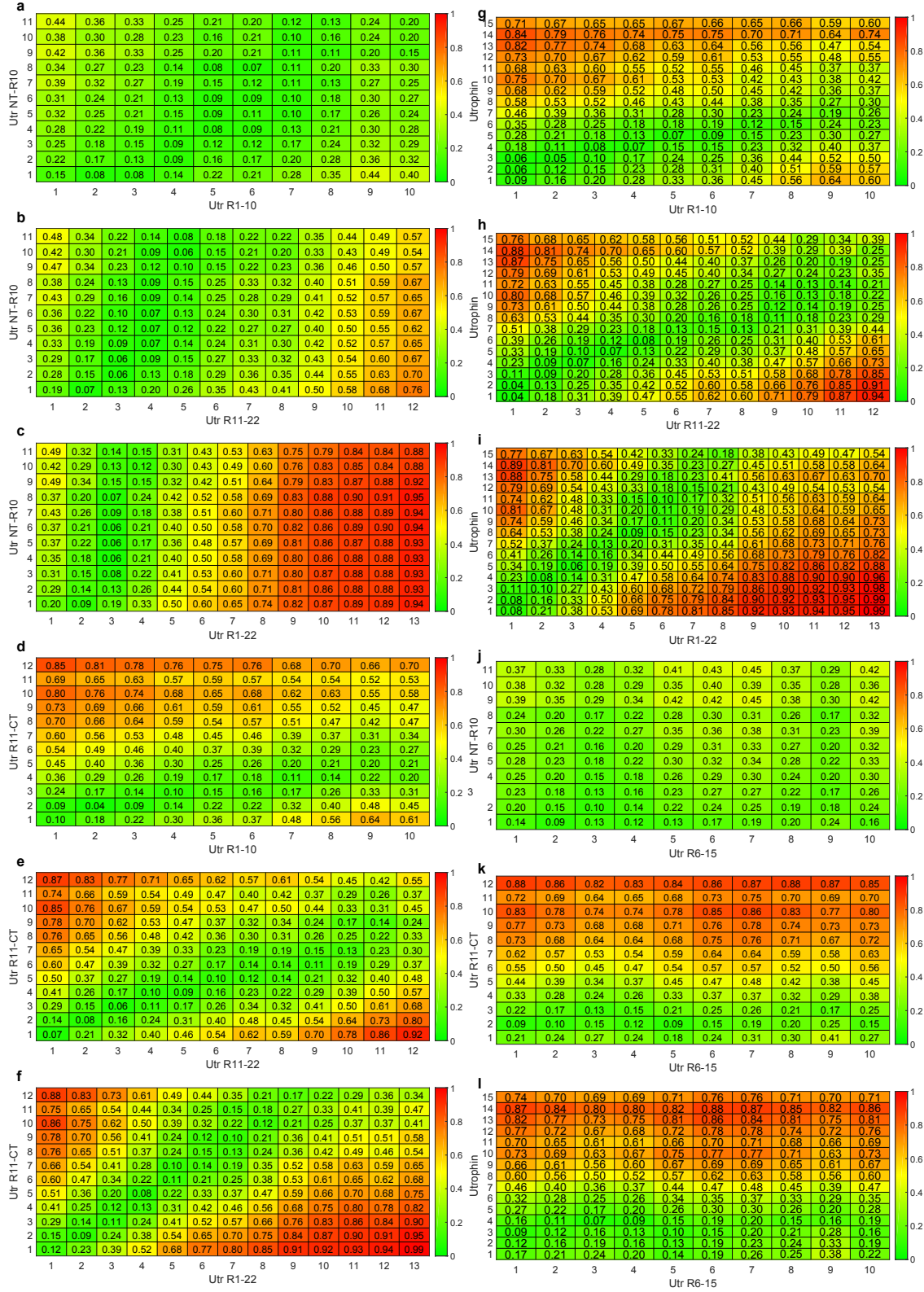


Fig. 3.8.10: Heat maps of KS test metrics which compare the distribution of unfolding forces for different unfolding event counts.

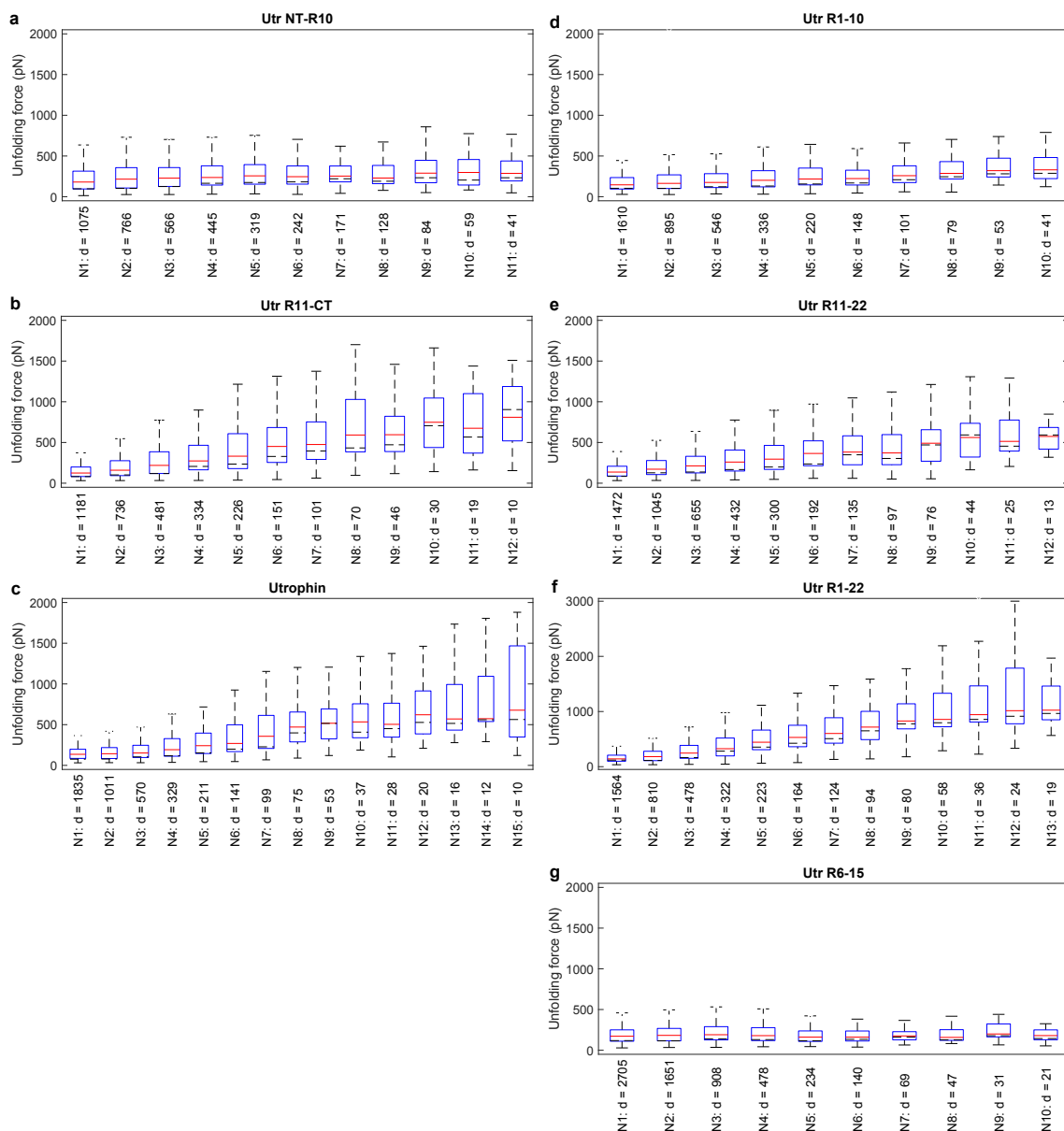


Fig. 3.8.11: Box plots of unfolding forces as a function of unfolding count. Utr NT-R10, Utr R1-R10, and Utr R6-15 (a, d, and g) can be seen to exhibit brittle behavior, wherein the unfolding force for the 10th unfolding event is comparable to that of the 1st unfolding event. Utr R11-CT and UTR R11-R22 show a stiffening spring behavior, wherein the unfolding forces have a 3 to 5-fold increase for the 10th unfolding event compared to the 1st (b and e). Full-length utrophin and Utr R1-22 show a strongly stiffening spring behavior with 4 to 6-fold increases for the 10th unfolding events compared to their 1st (c and f). The relative frequencies of each unfolding event are represented in Fig. 3.8.13.

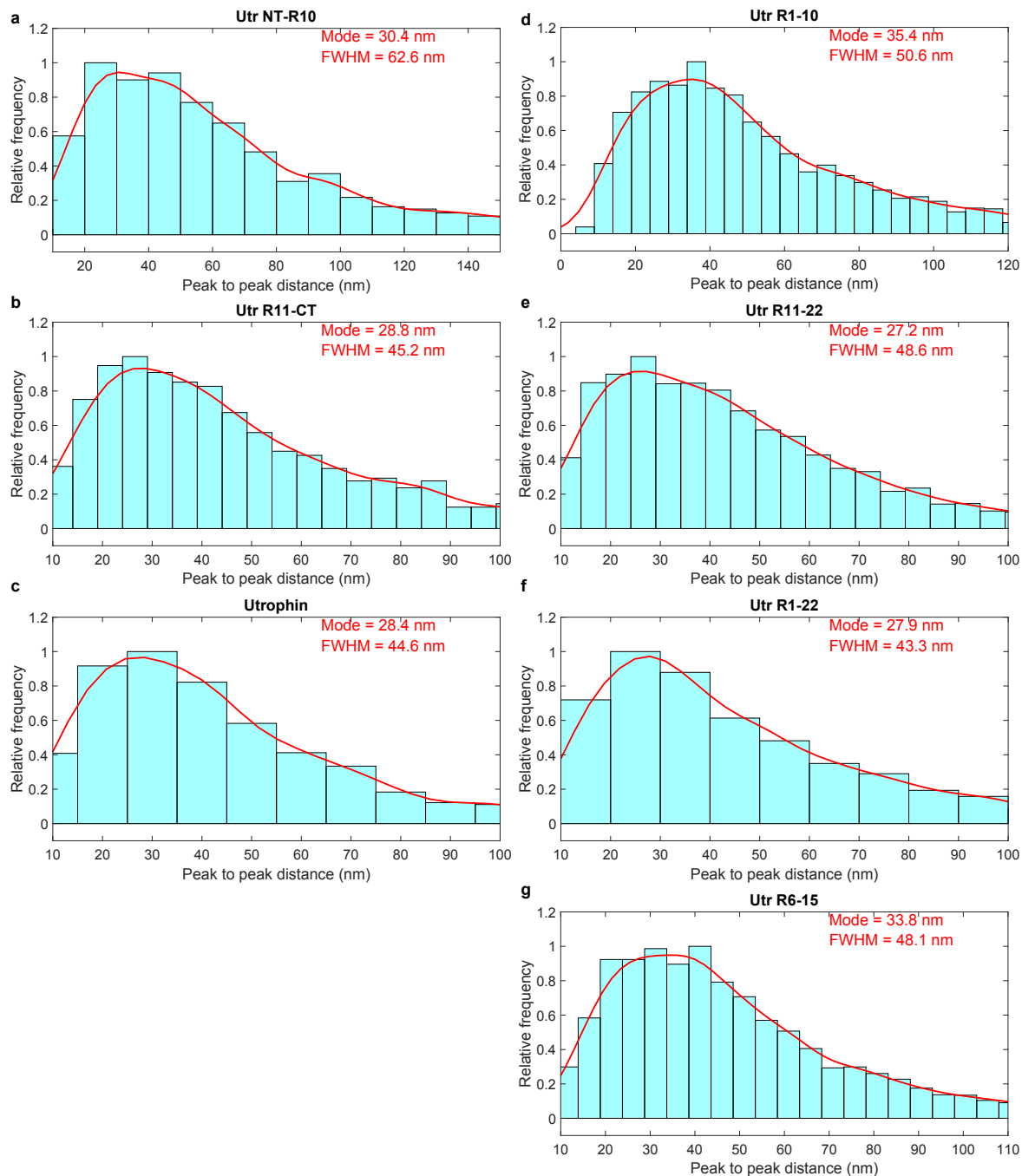


Fig. 3.8.12: Peak to peak distance. Distribution of the elongation of the utrophin constructs in between successive unfolding events. Each construct exhibits an asymmetric distribution with single peak (mode).

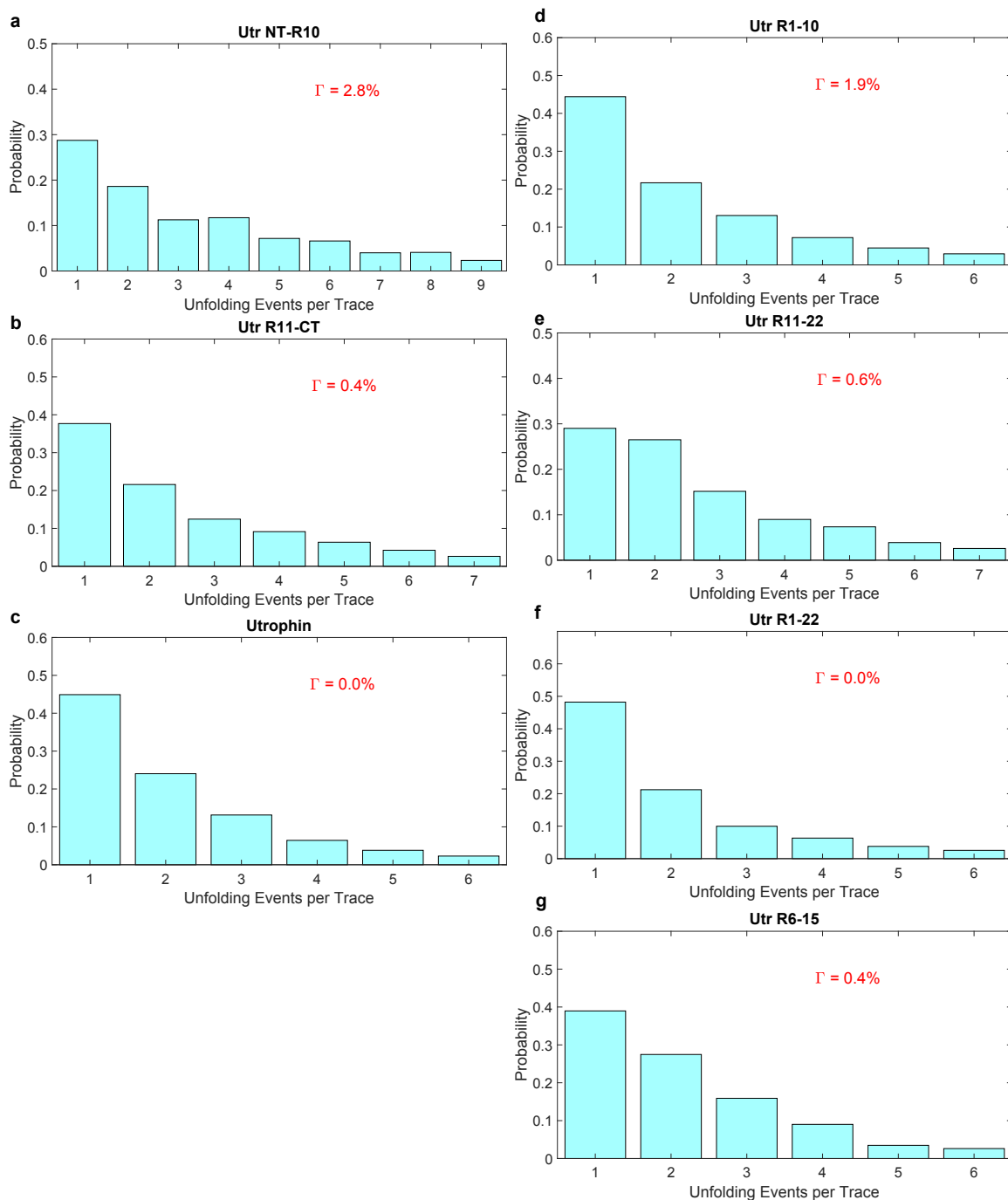


Fig. 3.8.13: Unfolding events per trace. The probabilities of observing a particular number of events (vertical axes) consistently reduce with the increasing unfolding event counts (horizontal axes). The value Γ inset in the figures represents the probability of observing force vs extension curves where the total number of unfolding events exceeds the expected number of folded domains for the given protein construct. The highest value of Γ observed is 2.8% (a).

Chapter 4

Quantifying Errors in the Jarzynski Estimator

In chapter 3, we presented the single molecule nano-mechanical characterization of the muscle protein utrophin. We followed the state of the art techniques to conduct the study. The key property reported in our study and other similar studies, the unfolding force, is not truly intrinsic to the protein. The unfolding force depends on a multitude of experimental conditions such as the temperature, loading rate, spring constant of the probe etc. On the other hand, free energy differences between two states of protein, is independent of experimental conditions. Estimating the free energy is made more complex when the experiments on the system are done in a non-equilibrium manner, as is the case in many single molecule force spectroscopic studies. In this chapter, we provide methods to quantify the errors in the estimates of free energy differences obtained from finite samples of non-equilibrium experimental data.

4.1 Introduction

Non-equilibrium processes are commonplace in a wide range of systems in physics, chemistry, and biology. They frequently provide a convenient and often the only means to study the thermodynamic and equilibrium properties of systems at the small scale [80]. For example, the equilibrium free energy differences between states is used for characterizing many important biomolecular systems [81]. Molecular motors, such as kinesin and dynein, are known to operate in the non-equilibrium regime

to convert chemical free energy in the form of ATP molecules, to useful work in carrying intracellular cargo [82, 83, 84]. Such cellular transport phenomena play a vital role in understanding the diseases such as cancers, deafness, and several neurodegenerative disorders [37]. Similarly, the nano-mechanics of shock absorbing muscle proteins, such as dystrophin and utrophin are studied via single molecule force spectroscopy. Here, characterizing the free energy landscape is crucial to unraveling the mechanisms of incurable diseases such as Duchenne Muscular Dystrophy caused by the malfunctioning or absent muscle protein dystrophin [73, 85]. Single molecule force spectroscopy while providing a fast and convenient experimental method for studying such molecules, excites them far from equilibrium [86]. Here, extracting the equilibrium free energy information from non-equilibrium experiments is thus of critical interest.

The Jarzynski equality links the equilibrium free energy differences between two states of a system to the non-equilibrium work required to move the system between the states [87]. The equality is given by:

$$e^{-\beta\Delta F} = \langle e^{-\beta W} \rangle, \quad (4.1)$$

where ΔF is the free energy difference between two states of interest, and $\beta := \frac{1}{k_B T}$ with Boltzmann constant k_B and the temperature T . The angular brackets $\langle \cdot \rangle$ denote an ensemble average obtained in the limit of infinite number of samples. The Jarzynski estimator employing a finite number of work samples, W_i , $i = 1, 2, \dots, N \in \mathbb{N}$, is given by:

$$\Delta \hat{F}_N := -\frac{1}{\beta} \ln \left[\frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} \right], \quad (4.2)$$

which yields an approximate estimate of the free energy difference ΔF with an error, $\varepsilon_N := \Delta F - \Delta \hat{F}_N$, between the true free energy difference and the estimate, when the number of samples is N .

One of the earliest results in quantifying the quality of the Jarzynski estimator is presented in [88], wherein the bias and variance errors in the Jarzynski estimates are studied under the assumption that the work samples are taken from a Gaussian distribution; such an assumption holds primarily under near equilibrium conditions. In [88], a numerical simulation study was used to develop empirical formulae for the bias and variance errors that depended on the knowledge of the dissipative work. The authors of [89] and [90] establish that the number of samples required for a desired

accuracy increases exponentially with an increase in system size and in proportion to the average of the dissipated work. The investigation in [91] refines the analysis and provides an approximate lower bound on the percentage error in the estimate. However, an upper bound is not provided.

The analysis in [92] proposes a maximum likelihood based technique to improve the convergence rate of the estimates when the distribution of the work is known and Gaussian. In [90] and [93], the use of experiments that move the system from an initial state A to a final state B and from an initial state B to a final state A are shown to improve efficiency of the free energy calculations. Further, [93] suggests strategies, such as the use of diagnostic plots, the Bennett Acceptance Ratio [94], and the use of intermediate states to increase the robustness of the computations. In cases where free energies are computed via simulation studies, [95] gives a re-sampling technique in which some trajectories are eliminated and replaced with clones, whose work values have a greater impact on the final result.

The state-of-the-art work referenced above leave considerable gaps in quantifying the error of the Jarzynski estimator. Here, the shortcomings of the existing work include: 1) assumptions of near-equilibrium conditions [88], or 2) the information on experimental variables, such as the need of the true free energy difference [91] or the knowledge of the true mean of the dissipated work [88, 90], or 3) the prerequisite of special experimental conditions that lead to Gaussian work distributions [92]. Plenty of systems undergo non-equilibrium state transitions that do not fall under these restrictive criteria [80]. For example, [96] reports asymmetric non-Gaussian work distributions for atomically flat surfaces undergoing friction when away from equilibrium. Recently, [97] observed for the first time, gamma distributed work when studying adsorption free energies of the A3 peptide on gold surface using non-equilibrium single molecule pulling experiments.

To overcome the limitations in the existing literature and to meet the challenges involved, we provide a comprehensive method to compute the confidence error bounds on the estimates from a sample limited Jarzynski estimator. Our method: 1) is applicable on far from equilibrium experiments, 2) can operate on a wide range of unknown work distributions, 3) has tunable confidence levels, 4) depends only on the experimentally measured quantities, i.e. the measured work done.

This chapter is organized as follows: 1) A rigorous analytical framework for quantifying the errors produced in computing free energy differences as a result of having

a finite amount of data is presented in Section 4.2. 2) Two methods for determining confidence bounds with provable guarantees and a technique to combine the methods to leverage their strengths are described in Section 4.3. 3) The methods are empirically verified and experimentally validated in Section 4.4 and Section 4.5. 4) The interpretation of the results and future directions of this work are given in Section 4.6.

4.2 The Jarzynski Estimator

The Jarzynski equality developed in [87] translates an ensemble of non-equilibrium measurements of work done to move a system between two states, to the equilibrium free energy differences between the states. In this section, we briefly review the Jarzynski equality and explicitly state the assumptions under which the equality applies. We then show that the ensemble of work involved are independent and identically distributed (i.i.d). This provides us with the framework to develop the confidence bounds.

Consider two Hamiltonian systems $\mathcal{A}$ and $\mathcal{B}$. Here $\mathcal{A}$ represents a system of interest and $\mathcal{B}$ represents the environment. The primary focus will be on the system $\mathcal{A}$. Let $X_{\mathcal{A}}(t) := (q_{\mathcal{A}}(t), p_{\mathcal{A}}(t))$ be the states of the system $\mathcal{A}$, where $q_{\mathcal{A}}(t) \in \mathbb{R}^K$, $K \in \mathbb{N}$, are the generalized coordinates and $p_{\mathcal{A}}(t) \in \mathbb{R}^K$ are the generalized momentums of $\mathcal{A}$, and t represents time. Let $\lambda(t) \in \mathbb{R}$ be an externally controlled variable that can affect the states $X_{\mathcal{A}}(t)$. We denote a trajectory using $\{\cdot\}$, for example, $\{X_{\mathcal{A}}\}_0^\tau := \{X_{\mathcal{A}}(t) | 0 \leq t \leq \tau\}$, represents the trajectory of all the states $X_{\mathcal{A}}(t)$ from $t = 0$ to $t = \tau$. We represent the Hamiltonian by $H_{\mathcal{A}}(q_{\mathcal{A}}(t), p_{\mathcal{A}}(t), \lambda(t)) : \mathbb{R}^{2K+1} \rightarrow \mathbb{R}$. We make the following assumption about the two systems $\mathcal{A}$ and $\mathcal{B}$.

Assumption 4.1. *a) The energy in system $\mathcal{A}$ is finite.*

b) Energy absorptions or donations from $\mathcal{A}$ to $\mathcal{B}$ do not change the temperature $T_{\mathcal{B}}$ of $\mathcal{B}$.

c) The dynamics of $\mathcal{A}$ and $\mathcal{B}$ are Markovian, that is

$$p(X_{\mathcal{A}}(t+dt) | \{X_{\mathcal{A}}\}_0^t) = p(X_{\mathcal{A}}(t+dt) | X_{\mathcal{A}}(t)) \text{ for all } t.$$

Now consider an experiment, wherein $\mathcal{A}$ and $\mathcal{B}$ start at thermal equilibrium at $t = 0$, and are at a temperature $T_{\mathcal{A}} = T_{\mathcal{B}} = T$. Subsequently, the system $\mathcal{A}$ is perturbed away from the equilibrium by a protocol in which $\lambda(t)$ is varied in a fixed

trajectory $\{\lambda\}_0^\tau$ from $\lambda(0) = u$ to $\lambda(\tau) = v$. During this perturbation when $0 \leq t \leq \tau$, the temperature of $\mathcal{A}$ is free to be different from that of $\mathcal{B}$, i.e. $T_{\mathcal{A}} \neq T_{\mathcal{B}} = T$. After this perturbation, the system $\mathcal{A}$ is equilibrated with $\mathcal{B}$, i.e. at $t \gg \tau$, $T_{\mathcal{A}} = T_{\mathcal{B}} = T$.

At $t = 0$, when $\mathcal{A}$ is in thermal equilibrium with $\mathcal{B}$, the probability of finding the system $\mathcal{A}$ in a state $X_{\mathcal{A}}(0) = x_{\mathcal{A}}^0$ is given by $p(X_{\mathcal{A}}(0) = x_{\mathcal{A}}^0) = \frac{e^{-\beta H_{\mathcal{A}}(x_{\mathcal{A}}^0, \lambda(0))}}{Z}$; where $Z := \int e^{-\beta H_{\mathcal{A}}(X_{\mathcal{A}}, \lambda(0))} dX_{\mathcal{A}}$ is the partition function and $\beta = \frac{1}{k_B T}$. The change in $H_{\mathcal{A}}(X_{\mathcal{A}}(t), \lambda(t))$ due to a change $d\lambda$ is defined as the work done on the system and is given by:

$$dW := d\lambda \frac{\partial H_{\mathcal{A}}(X_{\mathcal{A}}, \lambda)}{\partial \lambda}.$$

The work done on the system over the entire trajectory $\{\lambda\}_0^\tau$, is given by

$$W = \int dW = \int_0^\tau \frac{d\lambda}{dt} \frac{\partial H_{\mathcal{A}}(X_{\mathcal{A}}, \lambda)}{\partial \lambda} dt.$$

We now consider N independent and identical realizations of this experiment, where:

1. The systems $\mathcal{A}_i$ and $\mathcal{B}_i$ are identical to $\mathcal{A}$ and $\mathcal{B}$ respectively for every $i = 1, 2, \dots, N$.
2. The system $\mathcal{A}_i$ is in thermal equilibrium with $\mathcal{B}_i$ at temperature T at $t = 0$, for every $i = 1, 2, \dots, N$.
3. The parameter λ_i is varied from $\lambda_i(0) = u$ to $\lambda_i(\tau) = v$ along the same trajectory $\{\lambda\}_0^\tau$. That is

$$\{\lambda_i\}_0^\tau = \{\lambda\}_0^\tau, \text{ for all } i = 1, 2, \dots, N.$$

Then, for each realization $i = 1, 2, \dots, N$;

$$W_i = \int_0^\tau \frac{d\lambda_i}{dt} \frac{\partial H_{\mathcal{A}_i}(X_{\mathcal{A}_i}, \lambda_i)}{\partial \lambda_i} dt, \quad (4.1)$$

where $X_{\mathcal{A}_i}$ represents the states of the system $\mathcal{A}_i$. Further, the initial canonical distributions for each realization are the same. That is

$$\begin{aligned} p_i(X_{\mathcal{A}_i}(0) = x_{\mathcal{A}}^0) &= p(X_{\mathcal{A}}(0) = x_{\mathcal{A}}^0) \\ &= \frac{e^{-\beta H_{\mathcal{A}}(x_{\mathcal{A}}^0, \lambda(0))}}{Z}, \end{aligned}$$

for every $i = 1, 2, \dots, N$. In addition, since the systems $\mathcal{A}_i$ and $\mathcal{B}_i$ are identical to $\mathcal{A}$ and $\mathcal{B}$ for every $i = 1, 2, \dots, N$, the transition probability functions for each realization are the same. That is,

$$p_i(X_{\mathcal{A}_i}(t+dt)|X_{\mathcal{A}_i}(t)) = p(X_{\mathcal{A}}(t+dt)|X_{\mathcal{A}}(t)),$$

for each $i = 1, 2, \dots, N$.

Thus, in each realization i , the trajectory $\{X_{\mathcal{A}_i}\}_0^\tau$ will be sampled from the same joint distribution. Since the trajectory $\{X_{\mathcal{A}_i}\}_0^\tau$ in a realization i determines the work done W_i in that realization, $p_i(W_i = w) = \rho(w)$, for every $i = 1, 2, \dots, N$. This implies that the random variables $W_1, W_2, \dots, W_N$ are independent and identically distributed. Under the conditions described, the Jarzynski equality states that [87], for each $i = 1, 2, \dots, N$,

$$\begin{aligned} e^{-\beta\Delta F} &= \langle e^{-\beta W_i} \rangle \\ &= \int_{-\infty}^{\infty} p_i(w) e^{-\beta w} dw. \end{aligned}$$

Invoking the Strong Law of Large Numbers [98],

$$P\left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} = e^{-\beta\Delta F}\right) = 1.$$

The convergence in probability of summation $\frac{1}{N} \sum_{i=1}^N e^{-\beta W_i}$ to $e^{-\beta\Delta F}$ provides the basis for the practical Jarzynski estimator $\Delta\hat{F}_N := -\frac{1}{\beta} \ln \left[\frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} \right]$ that approximates the free energy change. We define the error ε_N in the free energy estimate as $\varepsilon_N := \Delta F - \Delta\hat{F}_N$. With the framework developed here, we now describe the confidence bounds on ε_N .

4.3 Confidence Bounds

Based on Assumption 4.1, the work is finite and thus there exist $W_{min}, W_{max} \in \mathbb{R}$ such that $W_{min} \leq W_i \leq W_{max}$ for all $i = 1, 2, \dots, N$, where $|W_{min}|, |W_{max}| < \infty$. In the previous section, we established that the measured values of work W_i , $i = 1, 2, \dots, N$ are independent and identically distributed random variables. We leverage the properties of random variables to develop confidence bounds on the error ε_N as

follows.

4.3.1 Via Mean and Range

The Jarzynski estimator as given in (4.2), primarily computes the approximate mean of $e^{-\beta W_i}$. To find a bound on the error ε_N , we obtain a bound for the error between an approximate mean and the true mean of a random variable. The Empirical Bernstein Inequality as given in [99], permits us to develop such a bound. We state the theorem next.

Theorem 4.1. Empirical Bernstein Inequality: *Let $X_1, X_2, \dots, X_N$ be independent and identically distributed real valued random variables, with mean $E[X_i] = \mu$ for all $i = 1, 2, \dots, N$, each $X_i \in [0, R]$, $R > 0$. Then the following inequality holds with probability at least $1 - \delta$:*

$$|\bar{X}_N - \mu| \leq \bar{s}_N \sqrt{\frac{2 \ln(3/\delta)}{N}} + R \frac{3 \ln(3/\delta)}{N}.$$

Here,

$$\bar{X}_N := \frac{1}{N} \sum_{i=1}^N X_i,$$

denotes the empirical mean and

$$\bar{s}_N := \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X}_N)^2}$$

denotes the empirical standard deviation of $X_1, X_2, \dots, X_N$.

As can be seen, the Empirical Bernstein Inequality depends on the empirical mean, the empirical variance, and the range R of the random variable. Estimating R for a system under study will be discussed in Section 4.3.3. We now show how to find confidence bounds on ΔF using the Bernstein Inequality.

Lemma 4.1. *Let W be the non-equilibrium work done on a system and let it follow some unknown distribution $\rho(W)$. Let $W_i \in [W_{\min}, W_{\max}]$, $i = 1, 2, \dots, N$ be the i.i.d and real valued samples drawn from $\rho(W)$. Define $Y_i := e^{-\beta W_i}$ for every $i = 1, 2, \dots, N$, $Y := e^{-\beta W}$, and $\Delta \hat{F}_N := -\frac{1}{\beta} \ln \left[\frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} \right]$. Then, for any $\delta \in (0, 1)$, the following inequality holds with probability at least $(1 - \delta)$:*

$$-\frac{1}{\beta} \ln \left(e^{-\beta \Delta \hat{F}_N} - \Psi_N \right) \geq \Delta F \geq -\frac{1}{\beta} \ln \left(e^{-\beta \Delta \hat{F}_N} + \Psi_N \right),$$

where

$$\begin{aligned} \Psi_N &:= \bar{\sigma}_N \sqrt{\frac{2 \ln(3/\delta)}{N}} + R \frac{3 \ln(3/\delta)}{N}, \\ R &= e^{-\beta W_{\min}} - e^{-\beta W_{\max}}, \end{aligned}$$

$$\bar{\sigma}_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y}_N)^2}, \quad \bar{Y}_N = \frac{1}{N} \sum_{i=1}^N e^{-\beta W_i}.$$

Here the upper bound on ΔF exists only when $\Psi_N < e^{-\beta \hat{F}_N}$.

Proof. We note that Y_i are i.i.d and real valued. Let $Z_i := Y_i - e^{-\beta W_{\max}}$, for all $i = 1, 2, \dots, N$, and $Z := Y - e^{-\beta W_{\max}}$. We observe that $Z \in (0, R)$, where $R = e^{-\beta W_{\min}} - e^{-\beta W_{\max}}$. We use the Bernstein inequality from Theorem 4.1 to find the bounds on $|\bar{Z}_N - \mathbb{E}[Z]|$, where $\bar{Z}_N = \frac{1}{N} \sum_{i=1}^N Z_i$. From the Jarzynski equality, it follows that,

$$\begin{aligned} \bar{Z}_N - \mathbb{E}[Z] &= \bar{Y}_N - \mathbb{E}[Y] \\ &= \frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} - e^{-\beta \Delta F}. \end{aligned}$$

We observe that with probability $(1 - \delta)$, the following inequality holds:

$$\left| \frac{1}{N} \sum_{i=1}^N e^{-\beta W_i} - e^{-\beta \Delta F} \right| \leq \bar{\sigma}_N \sqrt{\frac{2 \ln(3/\delta)}{N}} + R \frac{3 \ln(3/\delta)}{N}.$$

Rearranging and applying the logarithm, gives

$$-\frac{1}{\beta} \ln \left(e^{-\beta \Delta \hat{F}_N} - \Psi_N \right) \geq \Delta F \geq -\frac{1}{\beta} \ln \left(e^{-\beta \Delta \hat{F}_N} + \Psi_N \right),$$

where

$$\Psi_N := \bar{\sigma}_N \sqrt{\frac{2 \ln(3/\delta)}{N}} + R \frac{3 \ln(3/\delta)}{N}.$$

Since the term $e^{-\beta \Delta \hat{F}_N} - \Psi_N$ must be positive for the logarithm to be valid, we see that the upper bound is valid only when $\Psi_N < e^{-\beta \hat{F}_N}$. This completes the proof. $\square$

The bounds derived in Lemma 4.1 are valid for any distribution $\rho(W)$ and are practically computable entirely from the measured samples W_i . We observe that $\lim_N \Psi_N \rightarrow 0$ at a rate of at least $\frac{1}{\sqrt{N}}$. This implies that the upper and lower bounds given by Lemma 4.1 converge to the true free energy difference ΔF in the limit. In addition, when $R \gg \bar{\sigma}_N$, the convergence rate of Ψ_N could be as fast as $\frac{1}{N}$. For small N , the upper bound derived in Lemma 4.1 doesn't exist when $\Psi_N \geq e^{-\beta \Delta \hat{F}_N}$. Thus for situations when $\Delta \hat{F}_N$ is large, the number N needs to be high for obtaining an upper bound. This might not be feasible in certain situations. We derive an alternate bound which overcomes this issue next.

4.3.2 Via the Cumulative Distribution Function

Our approach here is to use the uncertainty quantification on the empirical cumulative distribution function (CDF) of the work measurements. The Dvoretzky-Kiefer-Wolfowitz inequality [100] quantifies the uncertainty in the empirical CDF of a random variable. This inequality was subsequently made tighter by Massart in [101]; we use this version of the inequality for our result and henceforth refer to this inequality as the Dvoretzky-Kiefer-Wolfowitz-Massart (DKWM) Inequality. The DKWM Inequality is described below.

Theorem 4.2. DKWM Inequality: *Let $X_1, X_2, \dots, X_N$ be i.i.d random variables with a continuous cumulative distribution function $\Phi(x)$, with $x \in \mathbb{R}$. Let $\hat{\Phi}_N(x)$ be the empirical cumulative distribution function defined by*

$$\hat{\Phi}_N(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{(X_i \leq x)}, \quad x \in \mathbb{R}$$

with

$$\mathbb{I}_{(X_i \leq x)} := \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases},$$

representing the indicator function. Then the following inequality holds without any restrictions on α

$$P\left(\sqrt{N} \sup_x |\hat{\Phi}_N(x) - \Phi(x)| > \alpha\right) \leq 2e^{-2\alpha^2}.$$

Theorem 4.2 thus gives the interval in which the true cumulative distribution $\Phi(x)$

lies. Indeed,

$$P\left(\sqrt{N} \sup_x \left| \hat{\Phi}_N(x) - \Phi(x) \right| > \alpha\right) \leq 2e^{-2\alpha^2},$$

implies

$$P\left(\sqrt{N} \left| \hat{\Phi}_N(x) - \Phi(x) \right| > \alpha\right) \leq 2e^{-2\alpha^2},$$

holds for all x . Then setting $\eta = 2e^{-2\alpha^2}$, we see that with probability at least $1 - \eta$,

$$\sqrt{N} \left| \hat{\Phi}_N(x) - \Phi(x) \right| \leq \alpha,$$

or

$$\hat{\Phi}_N(x) - \xi \leq \Phi(x) \leq \hat{\Phi}_N(x) + \xi,$$

with $\xi = \alpha/\sqrt{N} = \sqrt{\frac{1}{2N} \ln(2/\eta)}$.

Lemma 4.2. *Let $W_i \in [W_{\min}, W_{\max}]$ for all $i = 1, 2, \dots, N$, be the work values sampled from an unknown distribution $\rho(W)$. Let $\Phi(W)$ be the cumulative distribution function of W , and $\hat{\Phi}_N(W)$ the empirical cumulative distribution function, constructed using W_i . Given an arbitrary $a, b \in [W_{\min}, W_{\max}]$, the following inequalities hold with probability $1 - \eta$:*

$$-\frac{1}{\beta} \ln(U) \geq \Delta F \geq -\frac{1}{\beta} \ln(L),$$

where

$$\begin{aligned} U &:= e^{-\beta a} \left[\hat{\Phi}_N(a) - \hat{\Phi}_N(W_{\min}) - 2\xi \right] \\ &\quad + e^{-\beta W_{\max}} \left[\hat{\Phi}_N(W_{\max}) - \hat{\Phi}_N(a) - 2\xi \right], \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} L &:= e^{-\beta W_{\min}} \left[\hat{\Phi}_N(b) - \hat{\Phi}_N(W_{\min}) + 2\xi \right] \\ &\quad + e^{-\beta b} \left[\hat{\Phi}_N(W_{\max}) - \hat{\Phi}_N(b) + 2\xi \right], \end{aligned} \tag{4.2}$$

if $L, U > 0$, and $\xi := \sqrt{\frac{\ln(2/\eta)}{2N}}$.

Proof. The Jarzynski equality gives us the following relationship

$$e^{-\beta \Delta F} = \int_{W_{\min}}^{W_{\max}} \rho(W) e^{-\beta W} dW.$$

With a choice of an intermediate variable a such that $W_{min} \leq a \leq W_{max}$, we get:

$$\begin{aligned}
e^{-\beta \Delta F} &= \int_{W_{min}}^a \rho(W) e^{-\beta W} dW + \int_a^{W_{max}} \rho(W) e^{-\beta W} dW \\
&\geq e^{-\beta a} \int_{W_{min}}^a \rho(W) dW + e^{-\beta W_{max}} \int_a^{W_{max}} \rho(W) dW \\
&= e^{-\beta a} [\Phi(a) - \Phi(W_{min})] \\
&\quad + e^{-\beta W_{max}} [\Phi(W_{max}) - \Phi(a)].
\end{aligned} \tag{4.3}$$

Invoking Theorem 4.2, we know that with probability $1 - \eta$,

$$\hat{\Phi}_N(W) - \xi \leq \Phi(W) \leq \hat{\Phi}_N(W) + \xi,$$

where $\xi = \sqrt{\frac{\ln(2/\eta)}{2N}}$. Thus, we have

$$\begin{aligned}
e^{-\beta \Delta F} &\geq e^{-\beta a} [\hat{\Phi}_N(a) - \xi - (\hat{\Phi}_N(W_{min}) + \xi)] \\
&\quad + e^{-\beta W_{max}} [\hat{\Phi}_N(W_{max}) - \xi - (\hat{\Phi}_N(a) + \xi)] \\
&= e^{-\beta a} [\hat{\Phi}_N(a) - \hat{\Phi}_N(W_{min}) - 2\xi] \\
&\quad + e^{-\beta W_{max}} [\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(a) - 2\xi].
\end{aligned}$$

Applying the logarithm and rearranging the terms, we get:

$$\Delta F \leq -\frac{1}{\beta} \ln(U), \tag{4.4}$$

with

$$\begin{aligned}
U &:= e^{-\beta a} [\hat{\Phi}_N(a) - \hat{\Phi}_N(W_{min}) - 2\xi] \\
&\quad + e^{-\beta W_{max}} [\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(a) - 2\xi],
\end{aligned}$$

if $U > 0$. Note that the upper bound on ΔF doesn't exist for $U \leq 0$. Starting with the definition $e^{-\beta \Delta F} = \int_{W_{min}}^{W_{max}} \rho(W) e^{-\beta W} dW$ and following identical steps, we find

the lower bound of ΔF as follows.

$$\begin{aligned}
e^{-\beta \Delta F} &= \int_{W_{min}}^b \rho(W) e^{-\beta W} dW + \int_b^{W_{max}} \rho(W) e^{-\beta W} dW \\
&\leq e^{-\beta W_{min}} \int_{W_{min}}^b \rho(W) dW + e^{-\beta b} \int_b^{W_{max}} \rho(W) dW \\
&= e^{-\beta W_{min}} [\Phi(b) - \Phi(W_{min})] + e^{-\beta b} [\Phi(W_{max}) - \Phi(b)].
\end{aligned}$$

Using Theorem 4.2, we get:

$$\begin{aligned}
e^{-\beta \Delta F} &\leq e^{-\beta W_{min}} [\hat{\Phi}_N(b) + \xi - (\hat{\Phi}_N(W_{min}) - \xi)] \\
&\quad + e^{-\beta b} [\hat{\Phi}_N(W_{max}) + \xi - (\hat{\Phi}_N(b) - \xi)] \\
&= e^{-\beta W_{min}} [\hat{\Phi}_N(b) - \hat{\Phi}_N(W_{min}) + 2\xi] \\
&\quad + e^{-\beta b} [\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(b) + 2\xi].
\end{aligned}$$

This relationship gives us the lower bound

$$\Delta F \geq -\frac{1}{\beta} \ln(L),$$

where

$$\begin{aligned}
L &:= e^{-\beta W_{min}} [\hat{\Phi}_N(b) - \hat{\Phi}_N(W_{min}) + 2\xi] \\
&\quad + e^{-\beta b} [\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(b) + 2\xi].
\end{aligned}$$

Since the cumulative distribution function $\hat{\Phi}_N(W)$ is monotonic and non-decreasing, we observe that $L > 0$ for finite N . Thus the lower bound is well defined for all $N > 0$. This completes the proof. $\square$

Remark: In implementing the bounds given in Lemma 4.2, we select $a^* = \arg \min_a [-\frac{1}{\beta} \ln(U)]$ and $b^* = \arg \max_b [-\frac{1}{\beta} \ln(L)]$ to obtain the sharpest values for the bounds.

It would appear that the upper bound given in 4.1 suffers from the same problem of non-computability for small N as seen in the previous result. However, in practice U is computable for a relatively small N . From 4.1, we observe that $U \geq e^{-\beta W_{max}} [\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(W_{min}) - 4\xi]$. By definition of the empirical CDF,

$\hat{\Phi}_N(W_{max}) - \hat{\Phi}_N(W_{min}) = 1$. Therefore, the upper bound is computable when $\xi < \frac{1}{4}$, or in other words when $N > 8 \ln(2/\eta)$. For example, if η is chosen to be 0.05 (for a 95% confidence bound), $N > 29.51$, implying that the upper bound is computable after 30 samples. This is independent of the choice of a and b , as well as the nature of the CDF.

To get the benefits of convergence to the ground truth of Lemma 4.1 and the improved computability offered by Lemma 4.2, we combine the bounds as follows.

Theorem 4.3. *Let $W_i \in [W_{min}, W_{max}]$, $i = 1, 2, \dots, N$ be the i.i.d and real valued non-equilibrium work measurements sampled from some unknown distribution $\rho(W)$. Then, for any $\delta, \eta \in (0, 1)$, the following holds with probability at least $(1 - \delta)(1 - \eta)$:*

$$\mathcal{U} \geq \Delta F \geq \mathcal{L}, \quad (4.5)$$

with the upper bound $\mathcal{U}$ and lower bound $\mathcal{L}$ defined as follows:

(i) The upper bound

$$\mathcal{U} := \min[\mathcal{U}_a, \mathcal{U}_b];$$

$$\mathcal{U}_a := \begin{cases} -\frac{1}{\beta} \ln(e^{-\beta \Delta \hat{F}_N} - \Psi_N) & e^{-\beta \Delta \hat{F}_N} > \Psi_N \\ \infty & e^{-\beta \Delta \hat{F}_N} \leq \Psi_N \end{cases},$$

and

$$\mathcal{U}_b := \begin{cases} -\frac{1}{\beta} \ln(U) & U > 0 \\ \infty & U \leq 0 \end{cases}.$$

(ii) The lower bound

$$\mathcal{L} := \max \left[-\frac{1}{\beta} \ln(e^{-\beta \Delta \hat{F}_N} + \Psi_N), -\frac{1}{\beta} \ln(L) \right].$$

Here $\bar{Y}_N$ and Ψ_N are as defined in Lemma 4.1. Further U, L are as defined in Lemma 4.2.

Proof. The bound in Lemma 4.1 holds with probability at least $(1 - \delta)$ and the bound in Lemma 4.2 holds with probability at least $(1 - \eta)$. We observe that the bounds in 4.5 are the sharper of the two bounds given in Lemmas 4.1 and 4.2. Thus, for 4.5 to hold, we assume that both of the original bounds must be satisfied. This proves the assertion. $\square$

Note: For ease of use, given a user specified confidence percentage $z \in (0, 100)$ in the bounds on ΔF , we set $\delta = \eta = \zeta$, where $\zeta := 1 - \left(\frac{z}{100}\right)^{1/2}$.

We have shown via Lemmas 4.1, 4.2, and Theorem 4.3, confidence bounds $\mathcal{U} \geq \Delta F \geq \mathcal{L}$ on ΔF with provable guarantees. These bounds work with non-restrictive assumptions and require only the work measurements W_i , $i = 1, 2, \dots, N$ to be known. Given the bounds on ΔF , we can obtain the bounds on the error $\varepsilon_N = \Delta F - \Delta \hat{F}_N$ as $\mathcal{U} - \Delta \hat{F}_N \geq \varepsilon_N \geq \mathcal{L} - \Delta \hat{F}_N$ with the same confidence. In the next section, we demonstrate an empirical verification and experimental validation of these bounds.

4.3.3 Estimating the Range of Work

The bounds we have proposed require the limits $[W_{min}, W_{max}]$ of the work measurements W_i . We provide the following methods to estimate the upper and lower limits of work W_i .

Range of Measurements

Given the measurements of work W_i , $i = 1, 2, \dots, N$, the authors in [102] provide the estimates of the lower and upper limit as:

$$\widehat{W}_{min} := \min_i W_i,$$

$$\widehat{W}_{max} := \max_i W_i.$$

We found these estimates to be inaccurate for small sample sizes. The following two methods work well in a wider range of settings.

Instrument Limits

In many experiments at the single molecule scale, the work is done using instruments whose physical limits are well known. The physical limits can be used to estimate the maximum and minimum work that can be performed by the instrument. The knowledge of the instruments and experiments gives us 1) the maximum duration of the experiment τ_m , 2) the maximum absolute force applied, $\mathcal{F}_m := \max \left| \frac{\partial H_{A_i}(X_{A_i}, \lambda_i)}{\partial \lambda_i} \right|$, and 3) the maximum absolute rate with which the external variable can be changed,

$\nu_m := \max \left| \frac{d\lambda_i}{dt} \right|$. Then from 4.1, for all $i = 1, 2, \dots, N$, we have,

$$\begin{aligned} |W_i| &= \left| \int_0^\tau \frac{d\lambda_i}{dt} \frac{\partial H_{\mathcal{A}_i}(X_{\mathcal{A}_i}, \lambda_i)}{\partial \lambda_i} dt \right| \\ &\leq \int_0^{\tau_m} \nu_m \mathcal{F}_m dt \\ &= \nu_m \mathcal{F}_m \tau_m. \end{aligned}$$

Therefore, we can estimate the limits of the work done as

$$\widehat{W}_{min} := -\nu_m \mathcal{F}_m \tau_m,$$

$$\widehat{W}_{max} := \nu_m \mathcal{F}_m \tau_m.$$

Empirical Chebyshev Inequality

A probabilistic estimate for the limits and range of work can be obtained using the Empirical Chebyshev Inequality. The following theorem is originally due to Saw et. al. [103], and later modified by Kaban [104], who gives a version of the Empirical Chebyshev Inequality that uses only the empirical estimates. The theorem is valid for *weakly exchangeable* random variables. The definition as used by Saw et. al. [103] is given in Definition 4.1.

Definition 4.1. Weak exchangeability: A set of random variables $X_1, X_2, \dots, X_N$ are said to be weakly exchangeable if the joint probability $P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_N \leq x_N\}$ is invariant over all permutations of $x_1, x_2, \dots, x_N$ for all values of $x_1, x_2, \dots, x_N$.

Theorem 4.4. Empirical Chebyshev Inequality: Let $X_1, X_2, \dots, X_N$, and X with $N \geq 2$ be weakly exchangeable samples (i.i.d implies weakly exchangeable but the reverse is not true) from an unknown distribution, such that $P\{X_1 = X_2 = \dots = X_N = X\} = 0$. Then, for any $\epsilon > 0$,

$$P\{|X - \bar{X}_N| \geq \epsilon |\bar{X}_N|\} \leq \frac{N+1}{N} \frac{1}{\epsilon^2} \left(\frac{\bar{s}_N}{\bar{X}_N} \right)^2 + \frac{1}{N}.$$

Here $\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$, is the empirical mean, $\bar{s}_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X}_N)^2}$ is the empirical standard deviation.

We note that the measurements of work $W_i, i = 1, 2, \dots, N$ are i.i.d. Setting the right hand side of the inequality in Theorem 4.4 to $\gamma \in [0, 1]$, and solving for ϵ_γ yields $\epsilon_\gamma = \sqrt{\frac{N+1}{N\gamma-1} \left(\frac{\bar{s}_N}{\bar{W}_N}\right)^2}$. Here $\bar{W}_N = \frac{1}{N} \sum_{i=1}^N W_i$ and $\bar{s}_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (W_i - \bar{W}_N)^2}$. This implies that the inequality $\widehat{W}_{min} < W < \widehat{W}_{max}$ holds with probability at least $1 - \gamma$, where:

$$\widehat{W}_{min} := \bar{W}_N - \epsilon_\gamma |\bar{W}_N|,$$

$$\widehat{W}_{max} := \bar{W}_N + \epsilon_\gamma |\bar{W}_N|.$$

In practice, we recommend choosing probability threshold γ to be an order smaller the probability threshold ζ chosen for the Jarzynski bound, i.e. $\gamma < \zeta/10$.

4.4 Methods

We verified and validated the proposed confidence bounds empirically using 1) synthesized data, 2) simulations, and 3) experiments. In each of these cases, non-equilibrium work samples $W_i, i = 1, 2, \dots, N$ are collected and the free energy difference $\Delta\hat{F}_N$ computed using the Jarzynski estimator (see 4.2). Then the bounds, which are computed with only the work samples, are compared with the ground truth of the free energy difference ΔF . These steps are repeated multiple times under identical conditions to obtain the statistical behavior of the confidence bounds while keeping the sample number N fixed. To obtain the characterization for a range of sample numbers, the process is repeated for different values of N . We now describe the methods by which the three techniques generate the work samples.

Synthesized Data

The free energy difference ΔF between two states and the work W required to move the system between these states are related by:

$$W = \Delta F + w_d, \tag{4.1}$$

where w_d denotes the dissipated work [105]. If the states are fixed, then ΔF is a constant. However, the dissipated work is stochastic and follows some unknown distribution $w_d \sim \rho_{w_d}(w_d)$. For the verification, we used the truncated normal, a bi-modal distribution (the mixture of two truncated normals), and the beta distribution as possible candidates for $\rho_{w_d}(w_d)$. We generate work samples W_i by fixing ΔF and adding to it randomly generated samples of dissipated work w_d from a fixed candidate distribution for $\rho_{w_d}(w_d)$. To ensure the Jarzynski equality in 4.1 applies to these samples, the distribution ρ_{w_d} is selected such that $\langle e^{-\beta w_d} \rangle = 1$. There are multiple advantages to synthesizing data in this manner. The ground truth free energy ΔF is known. The distribution of work $\rho(W)$ is very flexible. It is possible to quantify and control the distance from equilibrium by adjusting the mean of the dissipated work $\bar{w}_d$.

Simulations

Here we generate non-equilibrium work samples W_i using Monte Carlo simulations. We chose to use two different systems whose dynamics are modeled by Langevin equations. They are: 1) the Spring Mass system described in [106] and 2) the Particle in a Time-Multiplexed Potential discussed in [107]. The detailed description of these systems is as follows.

The Spring Mass System

Fig. 4.4.1 shows a schematic of the Spring Mass system. It consists of a mass m connected to wall and a controllable handle (represented by a blue circle) via two identical springs, with identical spring constants k and rest lengths l_o . The position of the mass with respect to the wall is given by $q_A(t)$ and the position of the controllable handle is $\lambda(t)$. The momentum of the mass is given by $p_A(t) = m\dot{q}_A(t)$. The wall, the two springs, and the mass comprise the system of interest $\mathcal{A}$ and are placed in an environment $\mathcal{B}$. The temperature of $\mathcal{A}$ and $\mathcal{B}$ are T_A and T_B respectively. In a non-equilibrium experiment, $\mathcal{A}$ and $\mathcal{B}$ are equilibrated to temperature T and the handle is moved according to a schedule $\{\lambda\}_0^\tau$. The equations that model the states (q_A, p_A) are given by:

$$\begin{aligned}\dot{q}_{\mathcal{A}}(t) &= p_{\mathcal{A}}(t)/m, \\ \dot{p}_{\mathcal{A}}(t) &= -2k(q_{\mathcal{A}}(t) - l_o) + k\lambda(t).\end{aligned}\tag{4.2}$$

When the handle position is changed in the form of a ramp, i.e. $\lambda(t) = (L_1 - L_0) \frac{t}{\tau} + L_0$; $0 \leq t \leq \tau$, where L_0, L_1 are the initial and final positions of $\lambda(t)$, the free energy change between the start and end points is given by [106] as,

$$\Delta F = \frac{k}{4} (L_1^2 - L_0^2).$$

In simulations, the work done on the system is computed using the relationship [105]:

$$W_i = \int_0^\tau k [\lambda(t) - 2(q_{\mathcal{A}}(t) - l_o)] \dot{q}_{\mathcal{A}}(t) dt.$$

Here W_i is the work done on the i^{th} instance with the initialization $(q_{\mathcal{A}}^i(0), p_{\mathcal{A}}^i(0))$ sampled from the canonical distribution of $\mathcal{A}$.

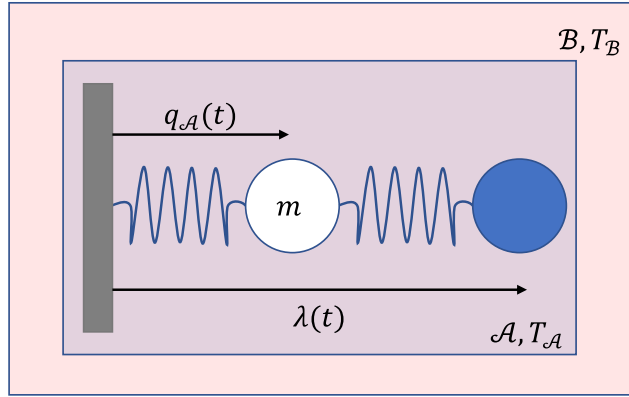


Fig. 4.4.1: Schematic of the Spring Mass system

Particle in a Time-Multiplexed Potential

Radiation pressure from lasers can be used to trap transparent micro-spheres in optical tweezers. The potential created by such a system applies a force that holds the micro-sphere at the focus of the laser. The force can be approximated as $f(x) = k(x - x_o)$, where k is the spring constant of the optical trap, x the position of the micro-sphere, and x_o the focus of the laser [108]. Fig. 4.4.2A shows the

schematic of such a particle in an optical trap and the resulting parabolic potential $\mathcal{U}(x)$ is shown in Fig. 4.4.2B.

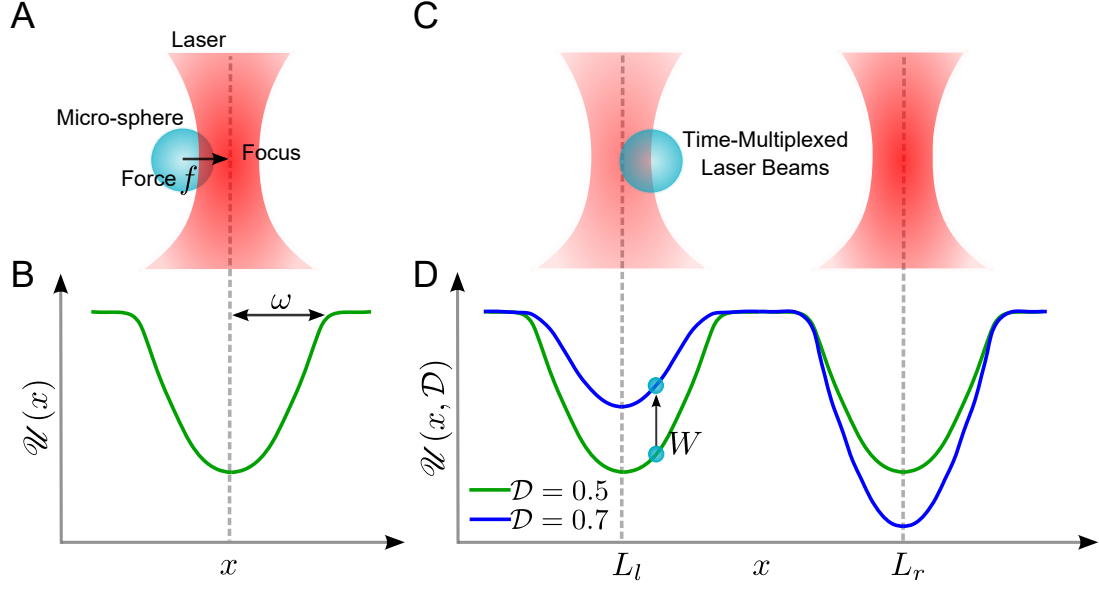


Fig. 4.4.2: Time-Multiplexed Potentials created by optical tweezers. (A) A micro-sphere trapped by a laser beam experiencing a restoring force f towards the focus. (B) The parabolic potential created by the laser trap of width ω . (C) The resultant of a laser beam rapidly multiplexed between two different foci - L_l and L_r and (D) the resultant time averaged potential. Changing the duty ratio $\mathcal{D}$ from 0.5 to 0.7 results in a reshaping of the potential and W work is done on the micro-sphere.

By multiplexing the focus of the laser rapidly between two locations, say L_r and L_l , the time averaged potential can be shown to form a double well as shown in Fig. 4.4.2C and Fig. 4.4.2D. The depth of the two potential wells can be controlled by the amount of time the laser spends in the two locations. Let t_r and t_l be the amount of time the laser spends in the right and left wells respectively, before switching to the other well. We define the duty ratio as $\mathcal{D} := \frac{t_r}{t_r + t_l}$ and a Boolean variable $r(t)$ to indicate if the laser focus is at location L_r . The value of $r(t)$ is 1 if the laser is at location L_r at time t and 0 if the laser is at L_l at t . The potential $\tilde{\mathcal{U}}(x, r)$ is described by [107]:

$$\tilde{\mathcal{U}}(x, r) = \begin{cases} \frac{1}{2}k(x - L_r)^2 + \mathcal{U}_b, & \text{if } |x - L_r| \leq \omega, r = 1, \\ \frac{1}{2}k(x - L_l)^2 + \mathcal{U}_b, & \text{if } |x - L_l| \leq \omega, r = 0, \\ \frac{1}{2}k\omega^2 + \mathcal{U}_b, & \text{otherwise,} \end{cases} \quad (4.3)$$

where ω is the width of the potential well, $\mathcal{U}_b$ is a constant. A particle in such a laser trap, can be modeled as a Brownian particle in a time multiplexed double well potential. The dynamics of the particle are then described by

$$c \frac{dx}{dt} = \nu(t) - \frac{\partial \widetilde{\mathcal{U}}(x(t), r(t))}{\partial x}. \quad (4.4)$$

Here c is the coefficient of viscosity and $\nu(t)$ is a noise term that implicitly models the particle's collisions with the particles in its environment. Specifically, $\nu(t)$ is Gaussian random variable such that $\langle \nu(t) \rangle = 0$ and $\langle \nu(t), \nu(t') \rangle = \frac{2c}{\beta} \delta(t - t')$ for all $t, t' \in \mathbb{R}$, with $\delta(\cdot)$ being the Dirac-delta function.

By trapping such a particle with some initial duty ratio $\mathcal{D}_i$ and switching the duty ratio to a final value $\mathcal{D}_f$, we can perform work on it. If the time to switch from $\mathcal{D}_i$ to $\mathcal{D}_f$ is much smaller than the time required for the particle to equilibrate with its environment, we can perform non-equilibrium work. Sekimoto [109] provides a way to compute the work done on the particle described by

$$W = \mathcal{U}(x(t_j), \mathcal{D}(t_j^+)) - \mathcal{U}(x(t_j), \mathcal{D}(t_j^-)). \quad (4.5)$$

Here t_j is the time at which the duty ratio is changed, with $t_j^- = t_j - dt$, and $t_j^+ = t_j + dt$ the times just before and after the change. In 4.5, the time averaged potential $\mathcal{U}(x, \mathcal{D})$ derived from 4.3 is used and is given by

$$\mathcal{U}(x, \mathcal{D}) = \begin{cases} \frac{\mathcal{D}}{2} k (x - L_r)^2 + \frac{1-\mathcal{D}}{2} k \omega^2 + \mathcal{U}_b, & \text{if } |x - L_r| \leq \omega, \\ \frac{1-\mathcal{D}}{2} k (x - L_l)^2 + \frac{\mathcal{D}}{2} k \omega^2 + \mathcal{U}_b, & \text{if } |x - L_l| \leq \omega, \\ \frac{1}{2} k \omega^2 + \mathcal{U}_b, & \text{otherwise.} \end{cases} \quad (4.6)$$

Monte Carlo simulations of the system can be done using 4.4 and the samples of work generated using 4.5. The free energy difference between the potentials created by the duty ratios $\mathcal{D}_i$ and $\mathcal{D}_f$ can be computed analytically (see 4.4) and can be used to verify the confidence bounds on the estimate.

Experiments

We use optical tweezers to perform non-equilibrium experiments. The optical tweezers consist of a 1064 nm trapping laser and a 850 nm detection laser. An acousto-optic

deflector (AOD) is used to control the position of the focus of the trapping laser in the x-y plane. The detection laser, in conjunction with a quadrature photodiode (QPD) is used to measure the location of a micro-sphere caught in the optical trap. For the experiment protocol, we used $\mathcal{D}_i = 0.5$ and $\mathcal{D}_f = 0.75$, with the time to switch from one duty ratio to the other kept below $1\mu s$. The multiplexing period $t_r + t_l$ of the laser was set to $20\mu s$. The two foci of the laser were at $L_l = -500\text{ nm}$ and $L_r = 500\text{ nm}$. The width ω of the potential wells were determined to be 80 nm and the spring constant at full power of the laser was estimated to be 13.3 pN/nm . A dilute solution of polystyrene beads with an average diameter of $1\mu\text{m}$ in water was used at room temperature ($T = 297\text{ K}$). The average time constant of the beads was experimentally found to be 1 ms using a step test.

4.5 Results

We now present the validation data obtained from the empirical study conducted on the confidence bounds.

Synthesized Data

Fig. 4.5.1 shows the performance of the three confidence bounds described in Section 4.3. To verify the bounds function correctly under varied conditions, three candidate distributions $\rho(W)$ for the non-equilibrium work were chosen: the truncated normal distribution, the bimodal distribution, and the beta distribution. While keeping the free energy difference ΔF fixed to $10k_B T$, the mean of the dissipated work $\langle w_d \rangle$ was varied from $2k_B T$ to $20k_B T$. Thus, the ratio $\frac{\langle w_d \rangle}{\Delta F}$, which represents the distance from equilibrium was varied from 0.2 to 2.0. This was done for each of the candidate distributions. Further, the number of samples of work N , was changed from 20 to 10^4 .

For each case, where the distribution of work, the ratio $\frac{\langle w_d \rangle}{\Delta F}$, and N are fixed, $M = 250$ batches of work samples were generated. Upper and lower confidence bounds are computed for each batch of work samples. Thus, the empirical observations for each case are derived from $M = 250$ independent calculations of the bounds. We chose a confidence level of 95%.

In Fig. 4.5.1a, we show using box plots, batches of Combined Bounds computed for the cases where N is varied, for the truncated normal distribution of work and

$\frac{\langle w_d \rangle}{\Delta F} = 0.5$. Boxes are drawn such that the center line represents the median value of a batch. The edges of the boxes represent the inter-quartile range, which is 50% of the population from both sides of the median. The whiskers show the minimums and maximums after excluding outliers. The outliers are data points which lie outside of three standard deviations from the median, assuming a Gaussian distribution for the data. The graph on the right in Fig. 4.5.1a, shows the histograms of the $M = 250$ upper bounds, free energy estimates $\Delta \hat{F}_N$, and the lower bounds computed for the case where $N = 864$. In this example, all of the upper and lower bounds contain the true ΔF , which is shown by the dotted line passing through $10 k_B T$. Since we are using confidence bounds, in some instances, the bounds may fail to contain the true ΔF . We track such 'fail' cases and have included them in our data. This empirical study shows that the probability of seeing such bound failures is less than 5%, as is guaranteed and expected from confidence bounds determined at 95% (see SI Fig. 4.7.1). The box plots in this figure start only from $N = 20$, because the bounds are not computable for lesser number of samples as alluded to earlier.

In Fig. 4.5.1b, the behavior of the three bounds are summarized for all the cases tested. The graphs on the first column show the distributions of non-equilibrium work used. The second column shows the performance of the Bounds using Mean, the third column the Bounds using CDF, and the fourth column the Combined Bounds. The graphs in the second, third, and fourth columns are determined as follows. The upper and lower boundaries of the shaded region of a particular color represent the modes of the $M = 250$ values of the upper and lower bounds computed for a particular case. We observe that the Bounds using Mean and the Combined Bounds converge to the true ΔF as the number of samples of work increases. Further, all the bounds are sharper and converge faster when the experiment is closer to the equilibrium, that is, when $\langle w_d \rangle$ is smaller. Out of the three bounds, we note that the Combined Bounds have the benefits of the other two - the convergence property of the Bounds using Mean, and the better computability ($N \sim 50$) of the Bounds using CDF. We remark that all of the bounds, for each of the cases were computed strictly from the work values W_i , without any knowledge of true ΔF or the nature of the work distributions. The range of the work values was assumed to be known obtained using instrument limits. The performance of the bounds on the synthetic data using range estimation using measurements and using the empirical Chebyshev inequality are shown in SI Fig. 4.7.7 to 4.7.10.

In Fig. 4.5.2, we have provided a performance comparison of our bounds with the method given by Gore-Ritort-Bustamante in [88]. We refer to this approach as the 'GRB' method in this document. The GRB method is applicable only when the distribution of work is normal. It can be seen that the bounds produced by the GRB method, while sharper for distributions of work close to equilibrium ($\langle w_d \rangle / \Delta F = 0.2, 0.5$), are inaccurate close to 50% of the time (see SI Fig. 4.7.2). Our bounds give a predictable and tunable confidence level. For distributions of work far from equilibrium ($\langle w_d \rangle / \Delta F = 1.0, 2.0$), the GRB bounds are more conservative.

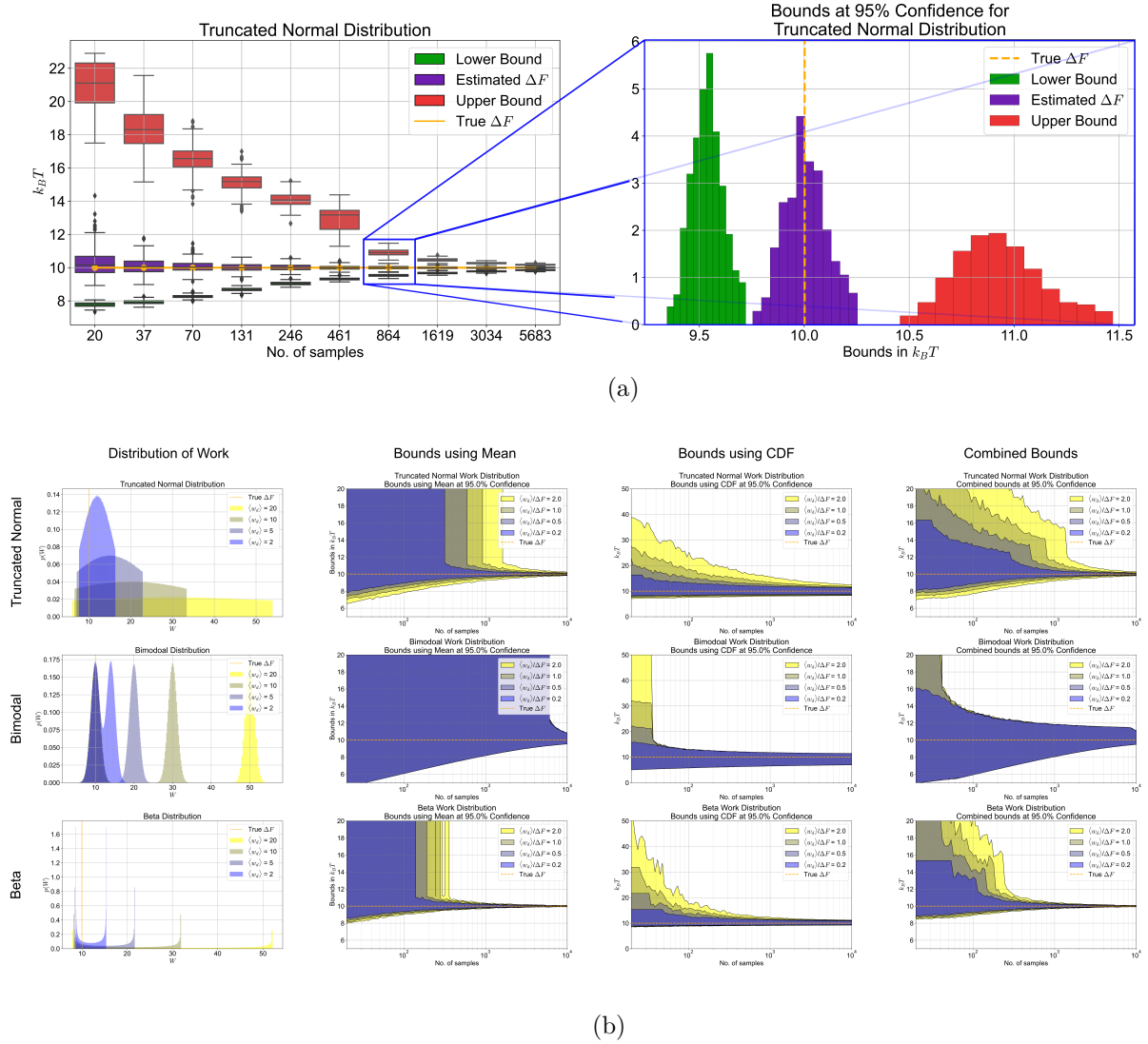


Fig. 4.5.1: A summary of the performance and characteristics of the bounds on the synthetic data. (a) A detailed view of the performance results. The box plot shows (left) showcases the performance of the Combined Bounds at 95% confidence level on work distributions that are truncated normals, with $\langle w_d \rangle / \Delta F = 0.5$ as the number of samples are increased. Each box plot represents one case of the synthetic data validation, wherein each case comprises of a collection of 250 computations of the bounds. An example case ($N = 864$) is shown in more detail on the right. The summarized view of the bounds from the synthetic data study, with varying distributions, varying $\langle w_d \rangle$, and increasing N , is shown in (b). The performance curves have been obtained using 250 computations for each case and interpolating the modes of the data obtained in each case.

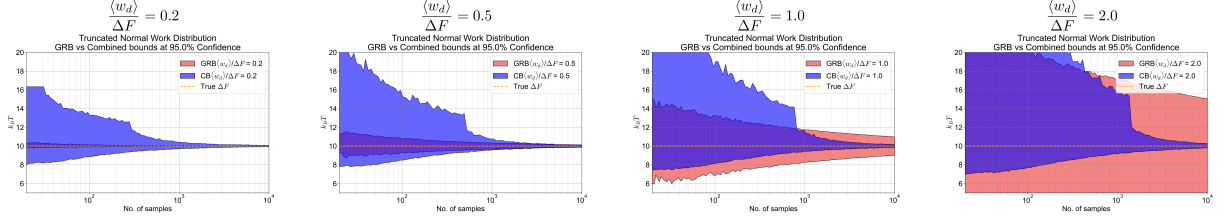


Fig. 4.5.2: Comparison of our combined bounds with the GRB method on synthetic data with truncated normal distributions of work. The performance curves have been obtained using 250 computations for each case and interpolating the modes of the data obtained in each case.

The Spring Mass System

The confidence bounds on the work values obtained from the simulated Spring Mass system are shown in Fig. 4.5.3. The work values are normally distributed. The mean of the dissipated work $\langle w_d \rangle$ is varied by changing the speed with which $\lambda(t)$ in Fig. 4.4.1 is changed. The faster speeds lead to higher dissipated work and thus are farther from the equilibrium. Similar to the analysis on the synthesized data, the bounds corresponding to each case were computed $M = 250$ times on freshly generated simulated data sets with new initializations. The modes of these bounds are plotted as the upper and lower boundaries for the regions displayed in Fig. 4.5.3. The graphs show that the bounds contain and converge to the true ΔF , with the Combined Bounds becoming computable around 30 samples of work and converge to within 10% of ΔF around 300 – 3000 samples, depending on the dissipated work. The confidence bounds were found to contain the true value 100% of the time for all the cases in this example (see SI Fig. 4.7.3).

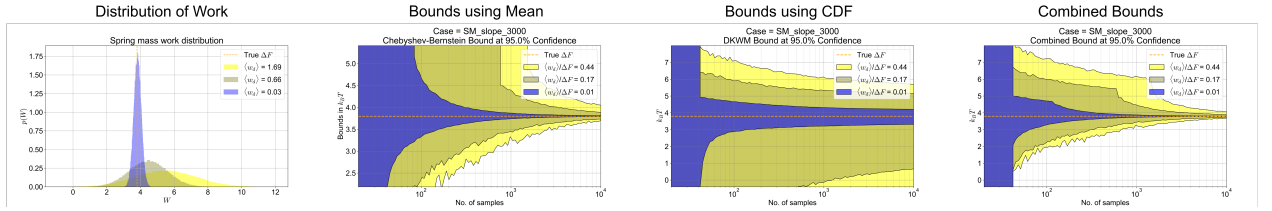


Fig. 4.5.3: Distributions of work and the corresponding bounds obtained from the simulated Spring Mass system.

Particle in a Time-Multiplexed Potential

Fig. 4.5.4A, shows the work values obtained in moving a $1\mu m$ particle in an optical potential at 50% duty ratio, to one at 30% duty ratio using optical tweezers. The work is seen to be beta distributed with two distinct modes. Fig. 4.5.4B, shows the modes of the 95% confidence bounds computed on $M = 250$ batches of work values (with different sample numbers N). The Combined Bounds are the sharpest and contain the true ΔF in all of the cases tested here (see SI Fig. 4.7.6). The true ΔF is computed by using a large batch of 10^7 work values and is ensured to match the expected analytical and simulated results (see SI Fig. 4.7.5). The ratio $\frac{\langle w_d \rangle}{\Delta F}$ was computed to be 1.03, indicating the experiment was done in the non-equilibrium setting.

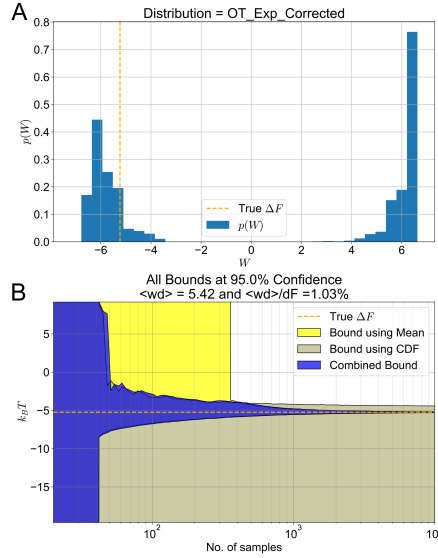


Fig. 4.5.4: (A) Distribution of the work done on a microsphere by changing the shape of the double well optical potential, from an initial duty ratio of 50% to a final duty ratio of 30%. (B) Confidence bounds and the true ΔF applied to the experimental work measurements.

4.6 Discussion

Free energy differences are an important quantity in a wide range of applications including the study of systems at the nano-scale. The experiments conducted to measure this quantity are often easier done under the non-equilibrium setting. While

the Jarzynski equality provides a way to link the equilibrium free energy differences to the non-equilibrium work measurements, it is asymptotic. Is it possible to quantify the errors in the estimates when only a finite number of work samples are available? In this chapter, we provided two methods to obtain confidence intervals on the finite sample estimates and consequently the errors. We fused the the two methods to combine the advantages of each technique - computability at small sample sizes and convergence to the ground truth. The confidence bounds developed in this chapter are the first of their kind and have several advantages over previously proposed solutions. They require only the measured work samples. We demonstrated that they can successfully be used on a wide range of work distributions without having any prior knowledge on the nature of distributions or the true free energy difference. This can be seen in Fig. 4.5.1, wherein the combination of our techniques produces 95% confidence bounds that are within 20% of the true ΔF at around 500 – 1500 samples of work. For a fixed number of samples of work, our bounds become less sharp with increasing distance from the equilibrium. However, the width of the bounds grow slowly with the distance from the equilibrium. For example, at $N = 1000$ samples and $\langle w_d \rangle / \Delta F = 0.2$ when the true $\Delta F = 10 k_B T$, we observe the width of the combined bounds are $0.69 k_B T$ for the truncated normal distribution, $3.75 k_B T$ for the bimodal distribution, and $0.40 k_B T$ for the beta distribution of work. The corresponding widths for $N = 1000$ samples, $\langle w_d \rangle / \Delta F = 2.0$, and true $\Delta F = 10 k_B T$ are: $6.67 k_B T$ for truncated normal distribution, $3.77 k_B T$ for the bimodal distribution, and $0.71 k_B T$ for the beta distribution of work. Therefore, for a 10 fold increase in the distance from equilibrium, the width of bounds increase by a factor of 9.66 for the truncated normal, 1.005 for the bimodal distribution, and 1.76 for the beta distribution. The comparison with the GRB method showcases this robustness to the distance from equilibrium even better. In Fig. 4.5.2, we observe that for $N = 1000$, $\langle w_d \rangle / \Delta F = 0.2$, and true $\Delta F = 10 k_B T$, the width of the GRB bound is $0.054 k_B T$. At $N = 1000$ and $\langle w_d \rangle / \Delta F = 2.0$ the width of the GRB bound is $14.12 k_B T$, which is an increase by a factor of 261 where our bounds increased by a factor of 9.66. Note also that the GRB bounds do not have a tunable confidence level, the confidence is set to a default level of 50% and the bounds are guaranteed only when the distribution of work is normal.

The distributions we selected for the synthetically generated work samples reflect the nature of distributions of work observed in real systems. In Fig. 4.5.3, we provide the distributions of work done on a Spring Mass system whose schematic is

shown in Fig. 4.4.1. The Spring Mass system is a proxy for force spectroscopy of single molecules of proteins using atomic force microscopes and produces normally distributed work. Our bounds work under this setting with a growth in the width of the bounds by a factor of ~ 30 when the distance to equilibrium increases by a factor of 44, corroborating our observations on synthetic data. Using optical tweezers to do work on a particle in a time-multiplexed well (see Fig. 4.4.2), we show a real example where the non-equilibrium work done is beta distributed. Such a multi-modal distribution cannot be approximated by a normal distribution and hence methods developed for normal distributions are not applicable. In contrast, our method can be successfully applied, as seen in Fig. 4.5.4. Our method produces 95% confidence bounds whose width converges to less than 20% of the true ΔF at around $N = 400$ samples.

We believe the techniques provided in this chapter are an important step towards improving reproducibility of research and increasing confidence on measurements made at the nano-scale. Future directions of this research are to improve the sharpness of the bounds. This might be possible by further optimizing the intermediate steps or by using the bounds to detect dominant events which contribute more to errors in the estimates. The developments in this chapter are available as an user friendly toolbox in Python with tunable confidence levels.

4.7 Supporting Information

4.7.1 Time Averaged Potential of a Time-Multiplexed Optical Trap

The instantaneous potential $\tilde{\mathcal{U}}(x, r)$ generated by an optical trap that is switched between two locations, centered at L_r and L_l , with width ω is described by [107] and given in 4.3. This is restated here for convenience:

$$\tilde{\mathcal{U}}(x, r) = \begin{cases} \frac{1}{2}k(x - L_r)^2 + \mathcal{U}_b, & \text{if } |x - L_r| \leq \omega, r = 1, \\ \frac{1}{2}k(x - L_l)^2 + \mathcal{U}_b, & \text{if } |x - L_l| \leq \omega, r = 0, \\ \frac{1}{2}k\omega^2 + \mathcal{U}_b, & \text{otherwise.} \end{cases}$$

Here $\mathcal{U}_b$ is a constant, $r(t)$ represents the location of the laser focus at time t , and k is the spring constant of the laser trap. The location of the laser focus is at L_r when $r(t) = 1$ and at L_l when $r(t) = 0$. Let t_r and t_l be the amount of time the focus of the laser spends in the locations L_r and L_l respectively. Define $T := t_r + t_l$, the switching time period. The duty ratio $\mathcal{D}$ defined as $\mathcal{D} := \frac{t_r}{t_r + t_l}$. To find the time averaged potential $\mathcal{U}(x, \mathcal{D})$ we consider the following three cases:

- Average potential of the right well, i.e. $|x - L_r| \leq \omega$.

$$\begin{aligned}\mathcal{U}(x, \mathcal{D}) &= \frac{1}{T} \int_0^T \widetilde{\mathcal{U}}(x, r) dt \\ &= \frac{1}{T} \int_0^{\mathcal{D}T} \left[\frac{1}{2}k(x - L_r)^2 + \mathcal{U}_b \right] dt + \frac{1}{T} \int_{\mathcal{D}T}^T \left[\frac{1}{2}k\omega^2 + \mathcal{U}_b \right] dt \\ &= \frac{\mathcal{D}}{2}k(x - L_r)^2 + \frac{(1 - \mathcal{D})}{2}k\omega^2 + \mathcal{U}_b.\end{aligned}$$

- Average potential of the left well, i.e. $|x - L_l| \leq \omega$.

$$\begin{aligned}\mathcal{U}(x, \mathcal{D}) &= \frac{1}{T} \int_0^T \widetilde{\mathcal{U}}(x, r) dt \\ &= \frac{1}{T} \int_0^{\mathcal{D}T} \left[\frac{1}{2}k\omega^2 + \mathcal{U}_b \right] dt + \frac{1}{T} \int_{\mathcal{D}T}^T \left[\frac{1}{2}k(x - L_l)^2 + \mathcal{U}_b \right] dt \\ &= \frac{(1 - \mathcal{D})}{2}k(x - L_l)^2 + \frac{\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b.\end{aligned}$$

- Average potential of the unlit region, i.e. $|x - L_r| > \omega$ and $|x - L_l| > \omega$.

$$\begin{aligned}\mathcal{U}(x, \mathcal{D}) &= \frac{1}{T} \int_0^T \widetilde{\mathcal{U}}(x, r) dt \\ &= \frac{1}{2}k\omega^2 + \mathcal{U}_b.\end{aligned}$$

Therefore,

$$\mathcal{U}(x, \mathcal{D}) = \begin{cases} \frac{\mathcal{D}}{2}k(x - L_r)^2 + \frac{1 - \mathcal{D}}{2}k\omega^2 + \mathcal{U}_b, & \text{if } |x - L_r| \leq \omega, \\ \frac{1 - \mathcal{D}}{2}k(x - L_l)^2 + \frac{\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b, & \text{if } |x - L_l| \leq \omega, \\ \frac{1}{2}k\omega^2 + \mathcal{U}_b, & \text{otherwise.} \end{cases}$$

4.7.2 Free Energy Difference for a Particle in a Time-Multiplexed Optical Trap

The free energy of a particle in a bounded optical trap can be computed using the relationship given by:

$$F = -\frac{1}{\beta} \ln [Z(\mathcal{D})],$$

where, $Z(\mathcal{D})$ is the partition function defined as:

$$Z(\mathcal{D}) := \int_{-\infty}^{\infty} e^{-\beta \mathcal{U}^*(x, \mathcal{D})} dx.$$

Here $\mathcal{U}^*(x, \mathcal{D})$ is the time averaged potential of a bounded optical trap and is given by:

$$\mathcal{U}^*(x, \mathcal{D}) = \begin{cases} \frac{\mathcal{D}}{2}k(x - L_r)^2 + \frac{1-\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b, & \text{if } |x - L_r| \leq \omega, \\ \frac{1-\mathcal{D}}{2}k(x - L_l)^2 + \frac{\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b, & \text{if } |x - L_l| \leq \omega, \\ \frac{1}{2}k\omega^2 + \mathcal{U}_b, & \text{if } |x| \leq M, |x - L_r| > \omega, |x - L_l| > \omega, \\ \infty & \text{otherwise.} \end{cases}$$

The infinite potential at $-M$ and M represents the bounding walls at those locations, implying the probability of finding the particle outside of the bounds is zero. In the experimental setting, $M > \max\{|L_r|, |L_l|\} + \omega$. Therefore,

$$Z(\mathcal{D}) = (2M - 4\omega)e^{-\beta[\frac{1}{2}k\omega^2 + \mathcal{U}_b]} \quad (4.1)$$

$$+ \frac{C_l}{2} \left(\frac{\pi}{\alpha_l}\right)^{\frac{1}{2}} [\text{erf}(\omega\sqrt{\alpha_l}) - \text{erf}(-\omega\sqrt{\alpha_l})] \quad (4.2)$$

$$+ \frac{C_r}{2} \left(\frac{\pi}{\alpha_r}\right)^{\frac{1}{2}} [\text{erf}(\omega\sqrt{\alpha_r}) - \text{erf}(-\omega\sqrt{\alpha_r})], \quad (4.3)$$

with $C_l := e^{-\beta(\frac{\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b)}$, $C_r := e^{-\beta(\frac{1-\mathcal{D}}{2}k\omega^2 + \mathcal{U}_b)}$, $\alpha_l := \beta\frac{(1-\mathcal{D})}{2}k$, and $\alpha_r := \beta\frac{\mathcal{D}}{2}k$.

The free energy difference due to switching the duty ratio from an initial value of $\mathcal{D}_i$ to a final value of $\mathcal{D}_f$ is thus given by:

$$\Delta F := F(\mathcal{D}_f) - F(\mathcal{D}_i) = -\frac{1}{\beta} \ln \left[\frac{Z(\mathcal{D}_f)}{Z(\mathcal{D}_i)} \right]. \quad (4.4)$$

4.7.3 Performance Figures

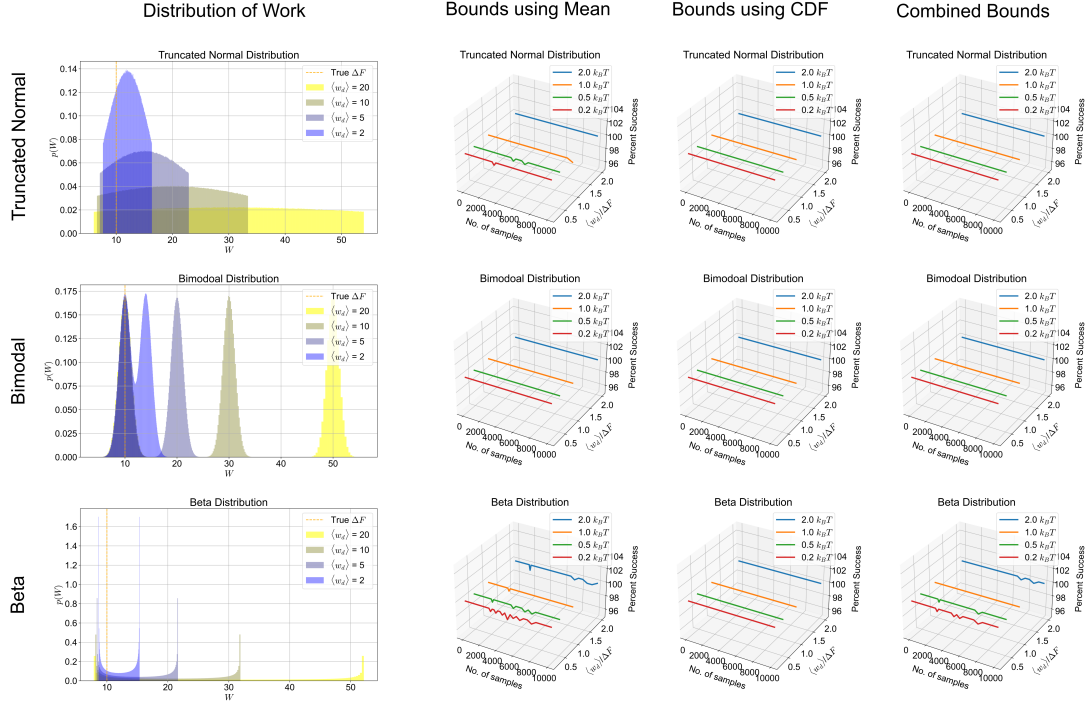


Fig. 4.7.1: Success rates for the bounds on the synthesized data with a known range.

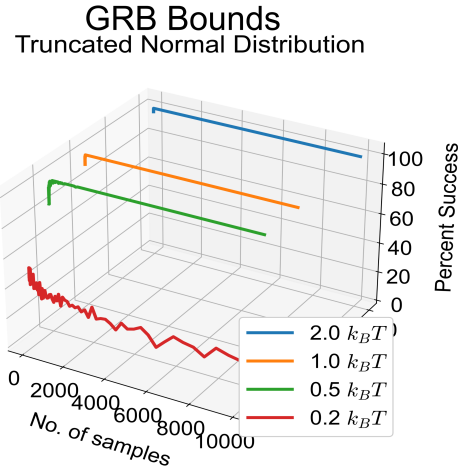


Fig. 4.7.2: Success rates for the GRB bounds on the synthesized data with a known range.

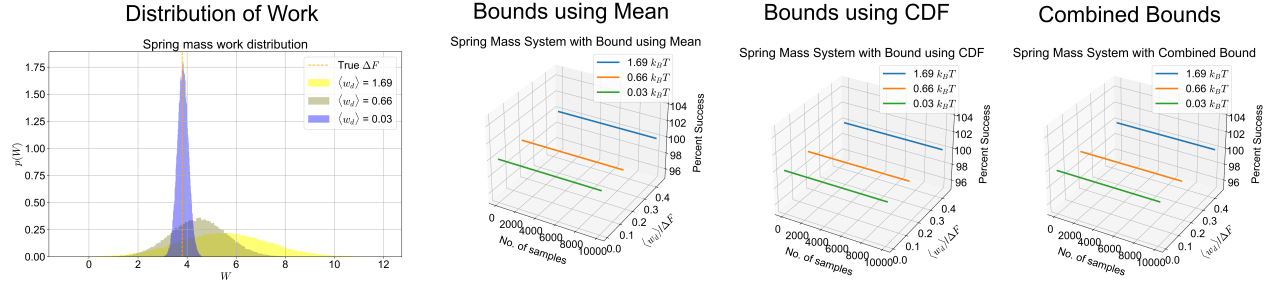


Fig. 4.7.3: Success rates for the bounds on the Spring Mass data.

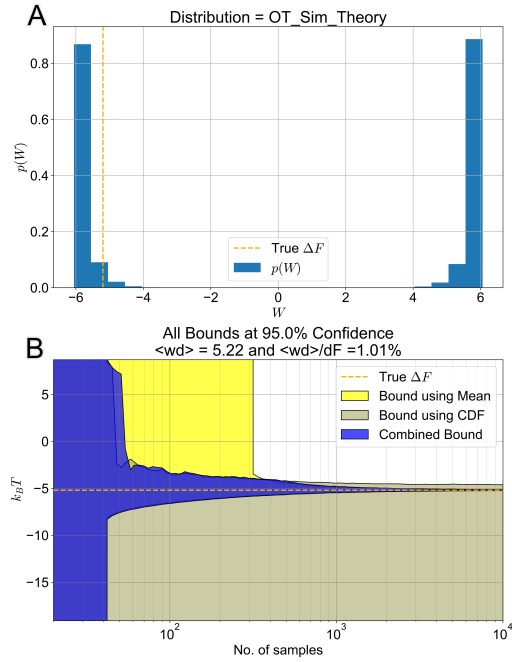


Fig. 4.7.4: Particle in a Time-Multiplexed Potential: Simulated Results

OT_Sim_Theory with $\langle wd \rangle = 5.22$ and $\langle wd \rangle/dF = 1.01\%$

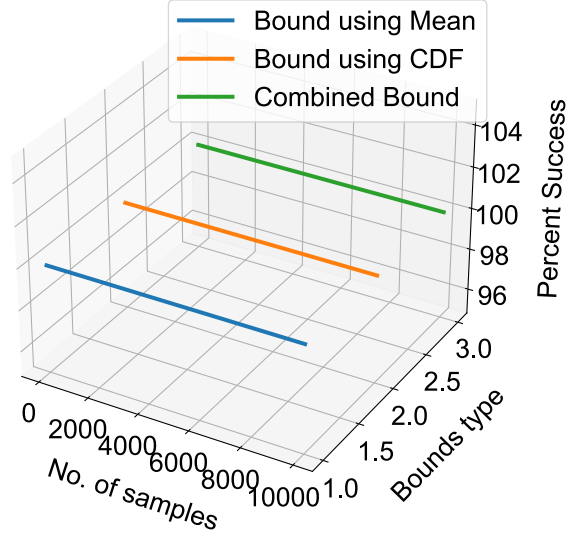


Fig. 4.7.5: Success rates for the bounds on the simulated data from the optical tweezers system.

OT_Exp_Corrected with $\langle wd \rangle = 5.42$ and $\langle wd \rangle/dF = 1.03\%$

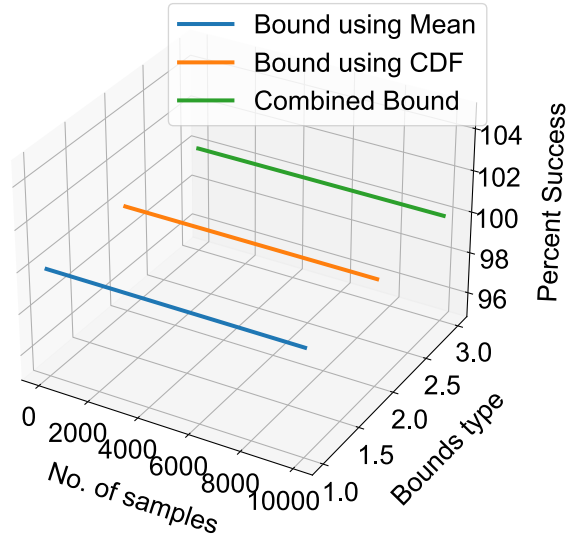


Fig. 4.7.6: Success rates for the bounds on the experimental data from the optical tweezers.

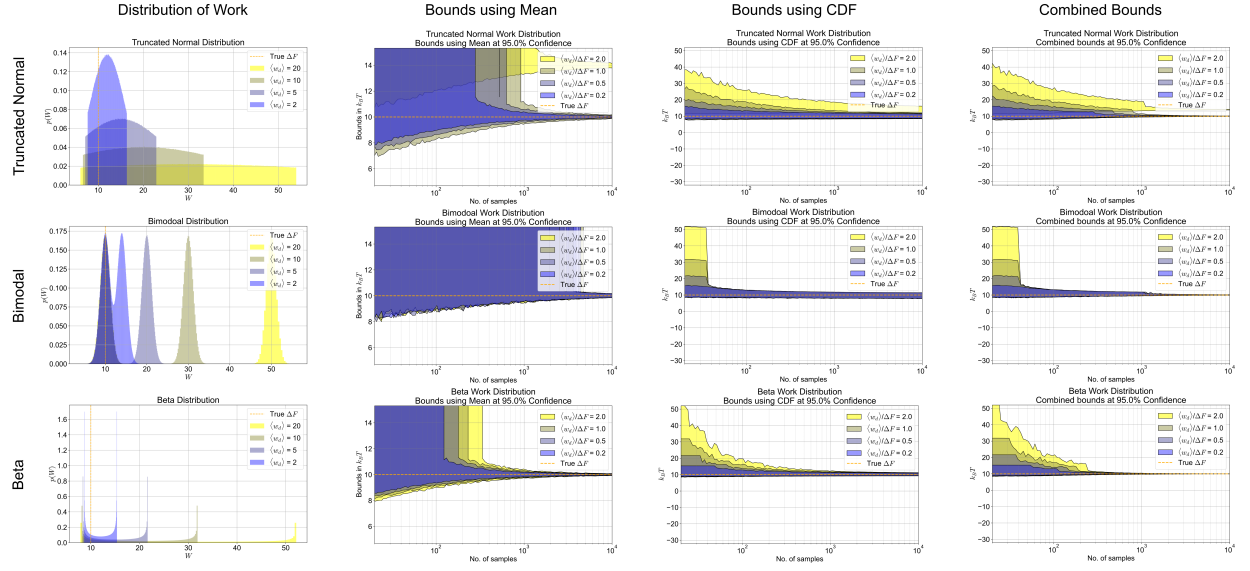


Fig. 4.7.7: Performance and characteristics of the bounds on the synthesized data with range estimated using the range of measurements method.

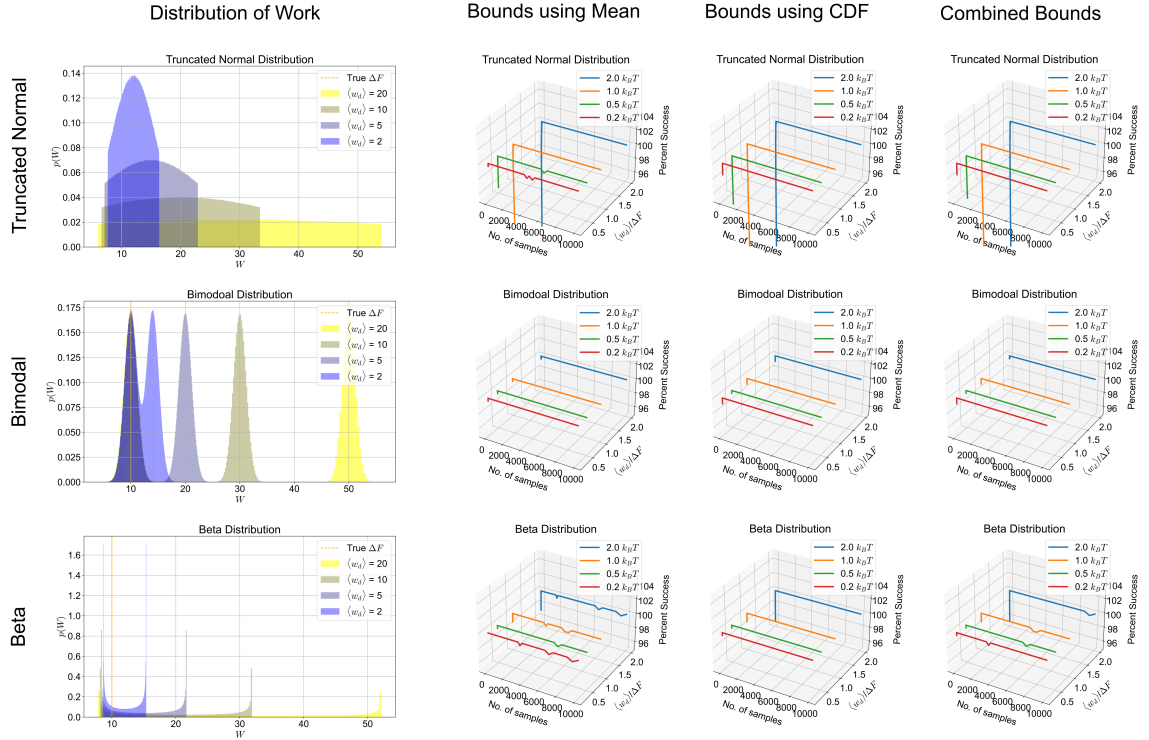


Fig. 4.7.8: Success rates for the bounds on the synthesized data with range estimated using the min-max method.

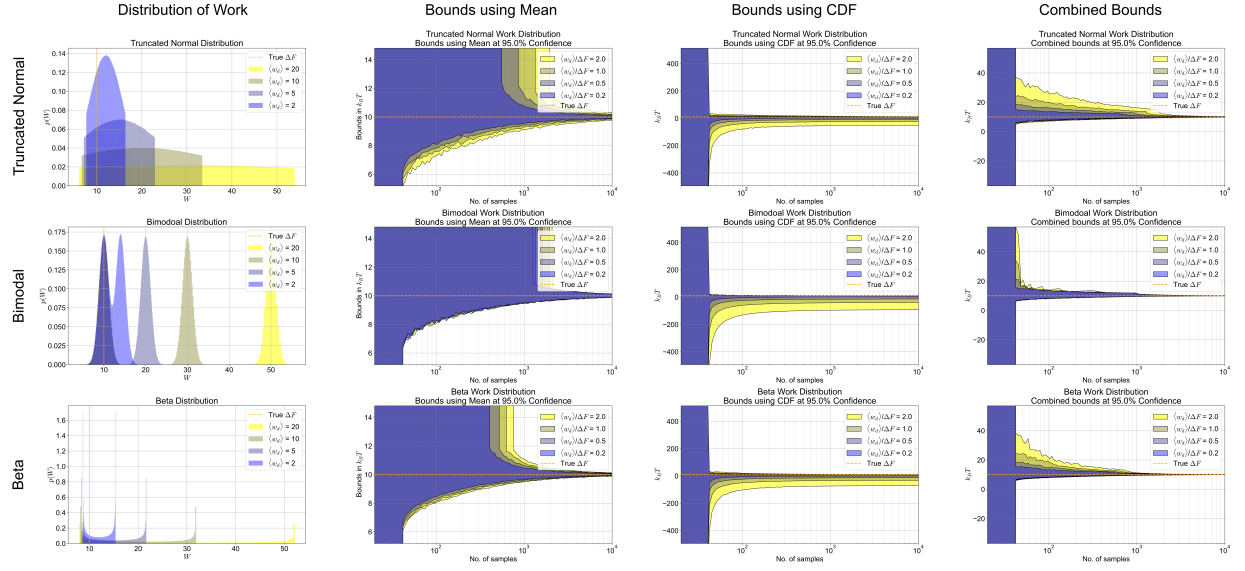


Fig. 4.7.9: Performance and characteristics of the bounds on the synthesized data with range estimated using the empirical Chebyshev inequality.

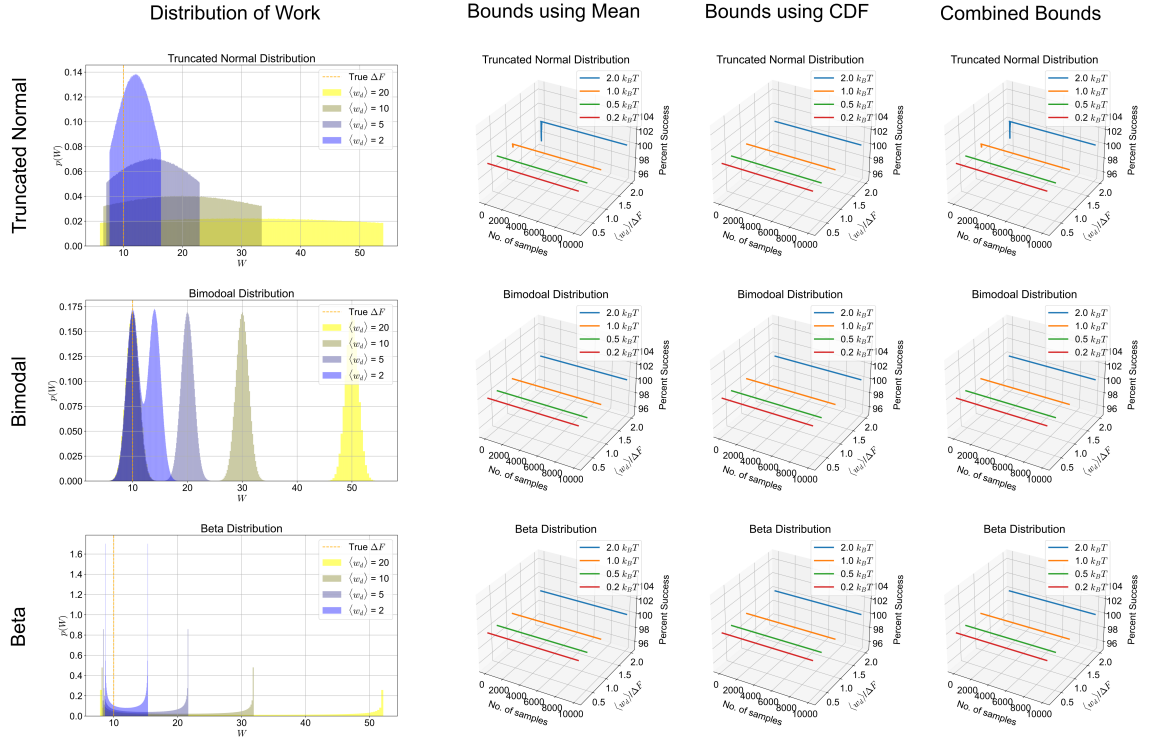


Fig. 4.7.10: Success rates for the bounds on the synthesized data with range estimated using the empirical Chebyshev inequality.

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

In the first part of this dissertation, we demonstrated an iterative change detection algorithm. The algorithm works with minimal prior information to meet the challenges of change detection in single molecule studies. We provided a theoretical framework for understanding how the algorithm 'learns' from data and from its own previous estimate. We provided theoretical guarantees on the algorithm's ability to improve the posterior probability with each iteration. We corroborated our analysis with an empirical performance study on a wide variety signals. Our algorithm is versatile and is adaptable for a range of experimental conditions such as low signal to noise ratio, sensor distortion, and non-linear parameter identification.

In the second part, we showcased how to modify our algorithm to help automate data analysis required in single molecule force spectroscopic studies of proteins. We then presented the techniques involved in the nano-mechanical characterization of the muscle protein utrophin. Our nano-mechanical characterization of the large multi-domain protein, utrophin, is the first study of its kind. Besides the tools for automated data analysis, we provide the experimental protocols and diagnostic tools needed to ensure only single molecules of the protein are being observed. We validated our protocols as well as our tools on the engineered protein titin and the well studied soft protein spectrin. Our study of utrophin yielded unexpected results - that utrophin is stiffer than expected and that it exhibits heterogeneous behavior along its length. The results, which led to the hypothesis that utrophin is more similar to titin rather than spectrin could not have been observed from bulk studies, wherein, all fragments of

utrophin have near identical characteristics. Our tools and techniques have enabled the comparative study of utrophin derived using different expression systems and helped suggest post-translation phosphorylation as a potential mechanism of tuning the stiffness of utrophin [79].

Free energy, an important intrinsic property of proteins as well as many small scale systems. This is different from the parameters such as unfolding forces, which are dependent on experimental conditions. In the final part of the dissertation, we develop confidence bounds for the free energy calculated using the Jarzynski equality from a finite number samples of non-equilibrium work. We use optical tweezers as a test bench to produce non-equilibrium work. We show that the work measured can be well characterized analytically, yet produced a beta distributed non-equilibrium work that cannot be approximated by a Gaussian distribution. Our confidence bounds do not require prior information on the free energy or assumptions about the nature of the distributions. We show the convergence properties of our bounds using both synthetic data, as well as real data.

5.2 Future Scope

There are multiple avenues to extend the work presented in this thesis. One direction is directed force spectroscopy, either using AFMs or optical tweezers, to improve the interpretability of the data. In directed force spectroscopy, a protein under study is modified to have directional chemical tags on either end, while the instrument probes contain the complementary chemical tags that will link with the tags. This technique, along with a fingerprint molecule in sequence with the protein will help ensure only one molecule is being studied and will provide specificity on the exact domain of the protein being studied. This is in contrast to our undirected force spectroscopic study, which obfuscates information on the domains being unfolded.

The use of optically excitable fingerprints (fluorophore) in conjunction with Total Internal Reflection Microscopy (TIRF) is another method to improve specificity on the data being collected. The change detection algorithm can be expanded for image analysis or for characterizing the fluorophores used in the TIRF based single molecule force spectroscopy. The confidence bounds developed for the finite data Jarzynski estimator can be expanded for reconstructing free energy landscapes from non-equilibrium experiments, such as for the method provided in [86]. The use of

the confidence bounds for the more general Crooks equality [110] in energy landscape reconstruction will expand the scope of this work.

Bibliography

- [1] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, “Mechanical and chemical unfolding of a single protein: a comparison,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 7, pp. 3694–3699, 1999, bibtex*{publisher:’National Acad Sciences’,__markedentry:’[siva:6]’}.
- [2] M. Rief, J. Pascual, M. Saraste, and H. E. Gaub, “Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles,” *Journal of molecular biology*, vol. 286, no. 2, pp. 553–561, 1999, bibtex*{publisher:’Elsevier’,__markedentry:’[siva:6]’}.
- [3] E. M. Puchner and H. E. Gaub, “Force and function: probing proteins with AFM-based force spectroscopy,” *Current opinion in structural biology*, vol. 19, no. 5, pp. 605–614, 2009, bibtex*{publisher:’Elsevier’,__markedentry:’[siva:6]’}.
- [4] F. M. Fazal and S. M. Block, “Optical tweezers study life under tension,” *Nature photonics*, vol. 5, no. 6, pp. 318–321, 2011, bibtex*{publisher:’Nature Research’,__markedentry:’[siva:6]’}.
- [5] M. Ludwig, M. Rief, L. Schmidt, H. Li, F. Oesterhelt, M. Gautel, and H. Gaub, “AFM, a tool for single-molecule experiments,” *Applied Physics A: Materials Science & Processing*, vol. 68, no. 2, pp. 173–176, 1999, bibtex*{publisher:’Springer’,__markedentry:’[siva:6]’}.
- [6] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, “Reversible unfolding of individual titin immunoglobulin domains by AFM,” *science*, vol. 276, no. 5315, pp. 1109–1112, 1997, bibtex*{publisher:’American Association for the Advancement of Science’,__markedentry:’[siva:6]’}.

- [7] C. Bustamante, Y. R. Chemla, N. R. Forde, and D. Izhaky, “Mechanical processes in biochemistry,” *Annual Review of Biochemistry*, vol. 73, pp. 705–748, 2004.
- [8] J. B. Udgaonkar, “Introducing the Mechanical Forces in Biochemistry Special Issue,” *Biochemistry*, vol. 58, no. 47, pp. 4655–4656, Nov. 2019. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.biochem.9b00999>
- [9] K. Bushby, R. Finkel, D. J. Birnkrant, L. E. Case, P. R. Clemens, L. Cripe, A. Kaul, K. Kinnett, C. McDonald, S. Pandya, J. Poysky, F. Shapiro, J. Tomezsko, and C. Constantin, “Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management,” *The Lancet Neurology*, vol. 9, no. 1, pp. 77–93, Jan. 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1474442209702716>
- [10] S. Rall and T. Grimm, “Survival in Duchenne muscular dystrophy,” *Acta Myologica*, vol. 31, no. 2, pp. 117–120, Oct. 2012. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3476855/>
- [11] K. Hares, J. S. Miners, A. J. Cook, C. Rice, N. Scolding, S. Love, and A. Wilkins, “Overexpression of Kinesin Superfamily Motor Proteins in Alzheimers Disease,” *Journal of Alzheimer’s Disease*, vol. 60, no. 4, pp. 1511–1524, Nov. 2017.
- [12] N. Hirokawa and Y. Tanaka, “Kinesin superfamily proteins (KIFs): Various functions and their relevance for important phenomena in life and diseases,” *Experimental Cell Research*, vol. 334, no. 1, pp. 16–25, May 2015.
- [13] A. J. Lucanus and G. W. Yip, “Kinesin superfamily: roles in breast cancer, patient prognosis and therapeutics,” *Oncogene*, vol. 37, no. 7, pp. 833–838, Feb. 2018.
- [14] S. Bottaro and K. Lindorff-Larsen, “Biophysical experiments and biomolecular simulations: A perfect match?” *Science*, vol. 361, no. 6400, pp. 355–360, Jul. 2018, publisher: American Association for the Advancement of Science. [Online]. Available: <https://www.science.org/doi/full/10.1126/science.aat4010>

- [15] E. Callaway, “What’s next for AlphaFold and the AI protein-folding revolution,” *Nature*, vol. 604, no. 7905, pp. 234–238, Apr. 2022, bandiera_abtest: a Cg_type: News Feature Number: 7905 Publisher: Nature Publishing Group Subject_term: Structural biology, Computational biology and bioinformatics, Biological techniques, Drug discovery. [Online]. Available: <https://www.nature.com/articles/d41586-022-00997-5>
- [16] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, and P. D. Adams, “AlphaFold predictions: great hypotheses but no match for experiment,” *Biochemistry*, preprint, Nov. 2022. [Online]. Available: <http://biorxiv.org/lookup/doi/10.1101/2022.11.21.517405>
- [17] P. V. Cornish and T. Ha, “A Survey of Single-Molecule Techniques in Chemical Biology,” *ACS Chemical Biology*, vol. 2, no. 1, pp. 53–61, Jan. 2007. [Online]. Available: <https://pubs.acs.org/doi/10.1021/cb600342a>
- [18] T. Ha, “Single-molecule methods leap ahead,” *Nature methods*, vol. 11, no. 10, pp. 1015–1018, 2014.
- [19] S. M. Altmann, “Pathways and intermediates in forced unfolding of spectrin repeats,” *Structure*, vol. 10, 2002. [Online]. Available: [https://doi.org/10.1016/S0969-2126\(02\)00808-0](https://doi.org/10.1016/S0969-2126(02)00808-0)
- [20] S. Rajaganapathy, J. Melbourne, and M. V. Salapaka, “Change detection using an iterative algorithm with guarantees,” *Automatica*, vol. 136, p. 110075, Feb. 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0005109821006038>
- [21] S. Rajaganapathy, J. Melbourne, T. Aggarwal, R. Shrivastava, and M. V. Salapaka, “Learning and Estimation of Single Molecule Behavior,” in *2018 Annual American Control Conference (ACC)*, Jun. 2018, pp. 5125–5130, iSSN: 2378-5861.
- [22] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, “Circular binary segmentation for the analysis of array-based DNA copy number data,” *Biostatistics*, vol. 5, no. 4, pp. 557–572, Oct. 2004.

- [23] A. M. Snijders, J. Fridlyand, D. A. Mans, R. Segraves, A. N. Jain, D. Pinkel, and D. G. Albertson, "Shaping of tumor and drug-resistant genomes by instability and selection," *Oncogene*, vol. 22, no. 28, pp. 4370–4379, Jul. 2003.
- [24] T. A. Graubert, P. Cahan, D. Edwin, R. R. Selzer, T. A. Richmond, P. S. Eis, W. D. Shannon, X. Li, H. L. McLeod, J. M. Cheverud, and T. J. Ley, "A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome," *PLoS Genetics*, vol. 3, no. 1, p. 9, 2007.
- [25] M. J. McConnell, M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall, and F. H. Gage, "Mosaic Copy Number Variation in Human Neurons," *Science*, vol. 342, no. 6158, pp. 632–637, Nov. 2013.
- [26] E. S. Page, "On problems in which a change in a parameter occurs at an unknown point," *Biometrika*, vol. 44, no. 1/2, pp. 248–252, 1957.
- [27] D. Picard, "Testing and estimating change-points in time series," *Advances in applied probability*, vol. 17, no. 4, pp. 841–867, 1985.
- [28] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
- [29] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, Mar. 2005.
- [30] J.-F. Ducr  Robitaille, L. A. Vincent, and G. Boulet, "Comparison of techniques for detection of discontinuities in temperature series," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 23, no. 9, pp. 1087–1101, 2003.
- [31] K. Svoboda, C. F. Schmidt, B. J. Schnapp, and S. M. Block, "Direct observation of kinesin stepping by optical trapping interferometry," *Nature*, vol. 365, no. 6448, pp. 721–727, 1993.

- [32] A. F. Oberhauser, P. K. Hansma, M. Carrion-Vazquez, and J. M. Fernandez, “Stepwise unfolding of titin under force-clamp atomic force microscopy,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 468–472, 2001.
- [33] B. C. Carter, M. Vershinin, and S. P. Gross, “A comparison of step-detection methods: how well can you do?” *Biophysical journal*, vol. 94, no. 1, pp. 306–319, 2008.
- [34] A. E. Emery, “The muscular dystrophies,” *The Lancet*, vol. 359, no. 9307, pp. 687–695, Feb. 2002.
- [35] K. M. D. Bushby, M. Thambyayah, and D. Gardner-Medwin, “Prevalence and incidence of Becker muscular dystrophy,” *The Lancet*, vol. 337, no. 8748, pp. 1022–1024, Apr. 1991.
- [36] N. Hirokawa, S. Niwa, and Y. Tanaka, “Molecular Motors in Neurons: Transport Mechanisms and Roles in Brain Function, Development, and Disease,” *Neuron*, vol. 68, no. 4, pp. 610–638, Nov. 2010.
- [37] H. L. Sweeney and E. L. Holzbaur, “Motor Proteins,” *Cold Spring Harbor Perspectives in Biology*, vol. 10, no. 5, p. a021931, May 2018.
- [38] M. A. Little and N. S. Jones, “Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 467, no. 2135, pp. 3088–3114, Nov. 2011.
- [39] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.
- [40] W. Hua, E. C. Young, M. L. Fleming, and J. Gelles, “Coupling of kinesin steps to ATP hydrolysis,” *Nature*, vol. 388, no. 6640, p. 390, 1997.
- [41] N. J. Carter and R. Cross, “Mechanics of the kinesin step,” *Nature*, vol. 435, no. 7040, p. 308, 2005.
- [42] B. M. Sadler and A. Swami, “Analysis of multiscale products for step detection and estimation,” *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 1043–1051, 1999.

- [43] J. W. Kerssemakers, E. L. Munteanu, L. Laan, T. L. Noetzel, M. E. Janson, and M. Dogterom, “Assembly dynamics of microtubules at molecular resolution,” *Nature*, vol. 442, no. 7103, p. 709, 2006.
- [44] T. Aggarwal, D. Materassi, R. Davison, T. Hays, and M. Salapaka, “Detection of steps in single molecule data,” *Cellular and molecular bioengineering*, vol. 5, no. 1, pp. 14–31, 2012.
- [45] F. E. Mullner, S. Syed, P. R. Selvin, and F. J. Sigworth, “Improved hidden Markov models for molecular motors, part 1: basic theory,” *Biophysical journal*, vol. 99, no. 11, pp. 3684–3695, 2010.
- [46] S. Syed, F. E. Mullner, P. R. Selvin, and F. J. Sigworth, “Improved hidden Markov models for molecular motors, part 2: extensions and application to experimental data,” *Biophysical journal*, vol. 99, no. 11, pp. 3696–3703, 2010.
- [47] C. Truong, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods,” *Signal Processing*, vol. 167, p. 107299, Feb. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0165168419303494>
- [48] E. L. Dumont, C. Do, and H. Hess, “Molecular wear of microtubules propelled by surface-adhered kinesins,” *Nature nanotechnology*, vol. 10, no. 2, pp. 166–169, 2015.
- [49] S. M. Pang, S. Le, and J. Yan, “Mechanical responses of the mechanosensitive unstructured domains in cardiac titin,” *Biology of the Cell*, vol. 110, no. 3, pp. 65–76, Mar. 2018.
- [50] H. Cai, *Biomimetic Nanoarchitectures for the Study of T Cell Activation with Single-Molecule Control*. Columbia University, 2016.
- [51] L. C. Bryan, D. R. Weilandt, A. L. Bachmann, S. Kilic, C. C. Lechner, P. D. Odermatt, G. E. Fantner, S. Georgeon, O. Hantschel, V. Hatzimanikatis, and others, “Single-molecule kinetic analysis of HP1-chromatin binding reveals a dynamic network of histone modification and DNA interactions,” *Nucleic Acids Research*, 2017.
- [52] R. Killick, P. Fearnhead, and I. A. Eckley, “Optimal Detection of Change-points With a Linear Computational Cost,” *Journal of the American Statistical*

- Association*, vol. 107, no. 500, pp. 1590–1598, Dec. 2012. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/01621459.2012.737745>
- [53] S. Arlot, A. Celisse, and Z. Harchaoui, “A Kernel Multiple Change-point Algorithm via Model Selection,” p. 56, 2019.
 - [54] P. Fryzlewicz, “Wild binary segmentation for multiple change-point detection,” *The Annals of Statistics*, vol. 42, no. 6, Dec. 2014.
 - [55] —, “Unbalanced Haar Technique for Nonparametric Function Estimation,” *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1318–1327, Dec. 2007. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1198/016214507000000860>
 - [56] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.
 - [57] J. A. Bilmes and others, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.
 - [58] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004.
 - [59] I. Goodfellow, Y. Bengio, and Courville, *Deep Learning*. MIT Press, 2016.
 - [60] J. R. Winkler, “Numerical recipes in C: The art of scientific computing, second edition,” *Endeavour*, vol. 17, no. 4, p. 201, Jan. 1993. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/016093279390069F>
 - [61] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.
 - [62] R. Law, P. Carl, S. Harper, P. Dalhaimer, D. W. Speicher, and D. E. Discher, “Cooperativity in forced unfolding of tandem spectrin repeats,” *Biophysical journal*, vol. 84, no. 1, pp. 533–544, 2003, bibtex*{publisher:’Elsevier’,__markedentry:’[siva.6]’}.
 - [63] A. E. Systems, “I27O AFM Reference Protein,” AthenaES, 1450 South Rolling Road Baltimore, MD 21227 USA, Tech. Rep. Catalog Number: 0304, 2011, bibtex*{__markedentry:’[siva.6]’}.

- [64] L. Passamano, A. Taglia, A. Palladino, E. Viggiano, P. DâAmbrosio, M. Scutifero, M. R. Cecio, V. Torre, F. D. Luca, E. Picillo, O. Paciello, G. Piluso, G. Nigro, and L. Politano, “Improvement of survival in Duchenne Muscular Dystrophy: retrospective analysis of 835 patients,” p. 5.
- [65] J. M. Ervasti, “Costameres: the Achillesâ heel of Herculean muscle,” *J. Biol. Chem.*, vol. 278, 2003. [Online]. Available: <https://doi.org/10.1074/jbc.R200021200>
- [66] I. N. Rybakova, J. R. Patel, and J. M. Ervasti, “The Dystrophin Complex Forms a Mechanically Strong Link between the Sarcolemma and Costameric Actin,” *J. Cell Biol.*, vol. 150, 2000. [Online]. Available: <https://doi.org/10.1083/jcb.150.5.1209>
- [67] J. M. Ervasti, “Dystrophin, its interactions with other proteins, and implications for muscular dystrophy,” *Biochim. Biophys. Acta*, vol. 1772, 2007. [Online]. Available: <https://doi.org/10.1016/j.bbadis.2006.05.010>
- [68] C. Rigoletto, “Utrophin expression during human fetal development,” *Int. J. Dev. Neurosci.*, vol. 13, 1995. [Online]. Available: [https://doi.org/10.1016/0736-5748\(95\)00039-J](https://doi.org/10.1016/0736-5748(95)00039-J)
- [69] S. Guiraud, H. Chen, D. T. Burns, and K. E. Davies, “Advances in genetic therapeutic strategies for Duchenne muscular dystrophy,” *Exp. Physiol.*, vol. 100, 2015. [Online]. Available: <https://doi.org/10.1113/EP085308>
- [70] N. Bhasin, R. Law, G. Liao, D. Safer, J. Ellmer, B. M. Discher, H. L. Sweeney, and D. E. Discher, “Molecular Extensibility of Mini-dystrophins and a Dystrophin Rod Construct,” *Journal of Molecular Biology*, vol. 352, no. 4, pp. 795–806, Sep. 2005. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0022283605008776>
- [71] C. C. Krieger, N. Bhasin, M. Tewari, A. E. X. Brown, D. Safer, H. L. Sweeney, and D. E. Discher, “Exon-skipped dystrophins for treatment of Duchenne muscular dystrophy: Mass spectrometry mapping of most exons and cooperative domain designs based on single molecule mechanics,” *Cytoskeleton*, vol. 67, no. 12, pp. 796–807, Dec. 2010. [Online]. Available: <http://doi.wiley.com/10.1002/cm.20489>

- [72] S. Le, M. Yu, L. Hovan, Z. Zhao, J. Ervasti, and J. Yan, “Dystrophin As a Molecular Shock Absorber,” *ACS Nano*, vol. 12, no. 12, pp. 12 140–12 148, Dec. 2018. [Online]. Available: <http://pubs.acs.org/doi/10.1021/acsnano.8b05721>
- [73] S. Rajaganapathy, J. L. McCourt, S. Ghosal, A. Lindsay, P. M. McCourt, D. A. Lowe, J. M. Ervasti, and M. V. Salapaka, “Distinct mechanical properties in homologous spectrin-like repeats of utrophin,” *Scientific Reports*, vol. 9, no. 1, pp. 1–11, Mar. 2019. [Online]. Available: <https://www.nature.com/articles/s41598-019-41569-4>
- [74] S. J. Winder, T. J. Gibson, and J. Kendrick-Jones, “Dystrophin and utrophin: the missing links!” *FEBS Lett.*, vol. 369, 1995. [Online]. Available: [https://doi.org/10.1016/0014-5793\(95\)00398-S](https://doi.org/10.1016/0014-5793(95)00398-S)
- [75] Q. Binnig and Gerber., “Atomic force microscope,” *Phys. Rev. Lett.*, vol. 56, 1986. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.56.930>
- [76] J. E. Sader, J. W. M. Chon, and P. Mulvaney, “Calibration of rectangular atomic force microscope cantilevers,” *Rev. Sci. Instrum.*, vol. 70, 1999. [Online]. Available: <https://doi.org/10.1063/1.1150021>
- [77] K. Ohlendieck, “Dystrophin-related protein is localized to neuromuscular junctions of adult skeletal muscle,” *Neuron*, vol. 7, 1991. [Online]. Available: [https://doi.org/10.1016/0896-6273\(91\)90301-F](https://doi.org/10.1016/0896-6273(91)90301-F)
- [78] J. V. Welser, “Myotendinous junction defects and reduced force transmission in mice that lack alpha7 integrin and utrophin,” *Am. J. Pathol.*, vol. 175, 2009. [Online]. Available: <https://doi.org/10.2353/ajpath.2009.090052>
- [79] M. P. Ramirez, S. Rajaganapathy, A. R. Hagerty, C. Hua, G. C. Baxter, J. Vavra, W. R. Gordon, J. M. Muretta, M. V. Salapaka, and J. M. Ervasti, “Phosphorylation alters the mechanical stiffness of a model fragment of the dystrophin homologue utrophin,” *Journal of Biological Chemistry*, vol. 299, no. 2, Feb. 2023, publisher: Elsevier. [Online]. Available: [https://www.jbc.org/article/S0021-9258\(22\)01290-X/abstract](https://www.jbc.org/article/S0021-9258(22)01290-X/abstract)
- [80] C. Bustamante, J. Liphardt, and F. Ritort, “The Nonequilibrium Thermodynamics of Small Systems,” p. 7.

- [81] N. Kolimi, A. Pabbathi, N. Saikia, F. Ding, H. Sanabria, and J. Alper, “Out-of-Equilibrium Biophysical Chemistry: The Case for Multidimensional, Integrated Single-Molecule Approaches,” *The Journal of Physical Chemistry B*, vol. 125, no. 37, pp. 10 404–10 418, Sep. 2021. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.jpccb.1c02424>
- [82] R. Ait-Haddou and W. Herzog, “Brownian Ratchet Models of Molecular Motors,” *Cell Biochemistry and Biophysics*, vol. 38, no. 2, pp. 191–214, 2003. [Online]. Available: <http://link.springer.com/10.1385/CBB:38:2:191>
- [83] A. I. Brown and D. A. Sivak, “Theory of Nonequilibrium Free Energy Transduction by Molecular Machines,” *Chemical Reviews*, vol. 120, no. 1, pp. 434–459, Jan. 2020. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00254>
- [84] C.-B. Li and S. Toyabe, “Efficiencies of molecular motors: a comprehensible overview,” *Biophysical Reviews*, vol. 12, no. 2, pp. 419–423, Apr. 2020. [Online]. Available: <http://link.springer.com/10.1007/s12551-020-00672-x>
- [85] K. Bushby, R. Finkel, D. J. Birnkrant, L. E. Case, P. R. Clemens, L. Cripe, A. Kaul, K. Kinnett, C. McDonald, S. Pandya, J. Poysky, F. Shapiro, J. Tomezsko, and C. Constantin, “Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management,” *The Lancet Neurology*, vol. 9, no. 1, pp. 77–93, Jan. 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1474442209702716>
- [86] G. Hummer and A. Szabo, “Free Energy Surfaces from Single-Molecule Force Spectroscopy,” p. 10.
- [87] C. Jarzynski, “Nonequilibrium Equality for Free Energy Differences,” *Physical Review Letters*, vol. 78, no. 14, pp. 2690–2693, Apr. 1997. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.78.2690>
- [88] J. Gore, F. Ritort, and C. Bustamante, “Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 22, pp.

12 564–12 569, Oct. 2003. [Online]. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.1635159100>

- [89] R. C. Lua and A. Y. Grosberg, “Practical Applicability of the Jarzynski Relation in Statistical Mechanics: A Pedagogical Example,” *The Journal of Physical Chemistry B*, vol. 109, no. 14, pp. 6805–6811, Apr. 2005. [Online]. Available: <https://pubs.acs.org/doi/10.1021/jp0455428>
- [90] C. Jarzynski, “Rare events and the convergence of exponentially averaged work values,” *Physical Review E*, vol. 73, no. 4, p. 046105, Apr. 2006. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.73.046105>
- [91] N. Y. Halpern and C. Jarzynski, “Number of trials required to estimate a free-energy difference, using fluctuation relations,” *Physical Review E*, vol. 93, no. 5, p. 052144, May 2016, arXiv: 1601.02637. [Online]. Available: <http://arxiv.org/abs/1601.02637>
- [92] M. Arrar, F. M. Boubeta, M. E. Szretter, M. Sued, L. Boechi, and D. Rodriguez, “On the accurate estimation of free energies using the jarzynski equality,” *Journal of Computational Chemistry*, vol. 40, no. 4, pp. 688–696, Feb. 2019. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1002/jcc.25754>
- [93] A. Pohorille, C. Jarzynski, and C. Chipot, “Good Practices in Free-Energy Calculations,” *The Journal of Physical Chemistry B*, vol. 114, no. 32, pp. 10 235–10 253, Aug. 2010. [Online]. Available: <https://pubs.acs.org/doi/10.1021/jp102971x>
- [94] C. H. Bennett, “Efficient estimation of free energy differences from Monte Carlo data,” *Journal of Computational Physics*, vol. 22, no. 2, pp. 245–268, Oct. 1976. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0021999176900784>
- [95] N. M. Roussey and A. Dickson, “Enhanced Jarzynski free energy calculations using weighted ensemble,” *The Journal of Chemical Physics*, vol. 153, no. 13, p. 134116, Oct. 2020. [Online]. Available: <http://aip.scitation.org/doi/10.1063/5.0020600>

- [96] J. R. Gomez-Solano, C. July, J. Mehl, and C. Bechinger, “Non-equilibrium work distribution for interacting colloidal particles under friction,” *New Journal of Physics*, vol. 17, no. 4, p. 045026, Apr. 2015. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1367-2630/17/4/045026>
- [97] Z. Kuang, K. M. Singh, D. J. Oliver, P. B. Dennis, C. C. Perry, and R. R. Naik, “Gamma estimator of Jarzynski equality for recovering binding energies from noisy dynamic data sets,” *Nature Communications*, vol. 11, no. 1, p. 5517, Dec. 2020. [Online]. Available: <http://www.nature.com/articles/s41467-020-19233-7>
- [98] D. P. Bertsekas and J. N. Tsitsiklis, “Introduction to Probability,” p. 284.
- [99] J.-Y. Audibert, R. Munos, and C. Szepesvari, “Tuning Bandit Algorithms in Stochastic Environments,” in *Algorithmic Learning Theory*, M. Hutter, R. A. Servedio, and E. Takimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4754, pp. 150–165. [Online]. Available: http://link.springer.com/10.1007/978-3-540-75225-7_15
- [100] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, “Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator,” *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 642–669, 1956, publisher: Institute of Mathematical Statistics. [Online]. Available: <https://www.jstor.org/stable/2237374>
- [101] P. Massart, “The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality,” *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990, publisher: Institute of Mathematical Statistics. [Online]. Available: <https://www.jstor.org/stable/2244426>
- [102] V. Mnih, C. Szepesvári, and J.-Y. Audibert, “Empirical Bernstein stopping,” in *Proceedings of the 25th international conference on Machine learning - ICML '08*. Helsinki, Finland: ACM Press, 2008, pp. 672–679. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1390156.1390241>
- [103] J. G. Saw, M. C. K. Yang, and T. C. Mo, “Chebyshev Inequality with Estimated Mean and Variance,” p. 4, 1984.

- [104] A. Kabłaj, “Non-parametric detection of meaningless distances in high dimensional data,” *Statistics and Computing*, vol. 22, no. 2, pp. 375–385, Mar. 2012. [Online]. Available: <http://link.springer.com/10.1007/s11222-011-9229-0>
- [105] C. Jarzynski, “Equalities and Inequalities: Irreversibility and the Second Law of Thermodynamics at the Nanoscale,” *Annual Review of Condensed Matter Physics*, vol. 2, no. 1, pp. 329–351, Mar. 2011. [Online]. Available: <http://www.annualreviews.org/doi/10.1146/annurev-conmatphys-062910-140506>
- [106] H. Håjar and J. M. O. de Zǎrate, “Jarzynski’s equality illustrated by simple examples,” *European Journal of Physics*, vol. 31, no. 5, pp. 1097–1106, Sep. 2010. [Online]. Available: <https://iopscience.iop.org/article/10.1088/0143-0807/31/5/012>
- [107] S. Talukdar, S. Bhaban, and M. V. Salapaka, “Memory erasure using time-multiplexed potentials,” *Physical Review E*, vol. 95, no. 6, p. 062121, Jun. 2017. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevE.95.062121>
- [108] J. Baillieul and T. Samad, Eds., *Encyclopedia of Systems and Control*. Cham: Springer International Publishing, 2021. [Online]. Available: <https://link.springer.com/10.1007/978-3-030-44184-5>
- [109] K. Sekimoto, “Energetics of thermal ratchet models,” *arXiv:cond-mat/9611005*, Nov. 1996, arXiv: cond-mat/9611005. [Online]. Available: <http://arxiv.org/abs/cond-mat/9611005>
- [110] G. E. Crooks, “Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences,” *Physical Review E*, vol. 60, no. 3, pp. 2721–2726, Sep. 1999. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.60.2721>