ABSTRACT

Title of dissertation: INFERENCE AND CONTROL IN NETWORKS FAR FROM EQUILIBRIUM.

Siddharth Sharma, Doctor of Philosophy, 2023

Dissertation directed by: Professor Doron Levy
Department of Mathematics

This thesis focuses on two problems in biophysics.

1. Inference in networks far from equilibrium.

2. Optimal transitions between network steady-states of unequal dimensions.

The system used for development of the theory and design of computational algorithms is the fully connected and asymmetric version of the widely used Ising model. We begin with the basic concepts of biological networks and their emergence as an analytical paradigm over the last two decades due to advancements in high-throughput experimental methods. Biological systems are open and exchange both energy and matter with their environment. Their dynamics are far from equilibrium and don't have well characterized steady-state distributions. This is in stark contrast to equilibrium dynamics with the Maxwell-Boltzmann distribution describing the histogram of microstates. The development of inference and control algorithms in this work is for nonequilibrium steady-states without detailed balance.

Inferring the Ising model far from equilibrium requires solving the inverse problem in statistical mechanics. As opposed to using a known Hamiltonian to solve for the macro-

scopic averages, we calculate the couplings and fields, i.e., model parameters, given the microstates or stochastic snapshots as inputs. We first demonstrate a time-series calculation for the inverse problem and use Poisson and Pólya-Gamma latent variables to construct a quadratic likelihood function which is then maximized using the expectation-maximization algorithm. In addition to the main calculation, properties of the Pólya-Gamma variables are used to solve logistic regression on a Gaussian mixture. This has applications to problems like clustering and community detection.

Not all available data in biology is time-ordered. In fact for some systems, e.g., gene-regulatory networks, most of the data is not in time-series. The solution to the inverse problem for such systems (data) is qualitatively different as it involves solving for the thermodynamic arrow of time. The present work uses the definition of a sufficient statistic based on equivalence classes to design a likelihood function through the disjoint cycles of the permutation group. The geometric intuition is provided using dihedral group of the same order. We state and prove that our likelihood function is minimally sufficient and present an optimization algorithm with computational results.

The second problem, i.e., optimal network control is solved using optimal transport. We recognize that biological networks have the property to grow and shrink while remaining functional and robust. Recent works that have continued the progress made by earlier seminal results have concentrated on systems which do not undergo transitions that alter their dimensions. For example, a network increasing or decreasing its number of nodes. The connection between thermodynamics and optimal transport is well established through the Wasserstein metric being the minimal dissipation for stochastic dynamics. This result depends on narrow convergence which requires that the system size remains the same.

Recently introduced Gromov-Wasserstein metric defined on a space of metric measure spaces, makes it possible to design optimal paths between probability distributions of different sizes. In context of networks, the GW metric can define geodesics between two

network nonequilibrium steady-states with different number of vertices. The last two chapters discuss the mathematical concepts and results that are required to develop the GW metric on networks and the computational algorithms that follow as a result. We define the probability measures and loss functions as per the physical properties of the Ising model and demonstrate a geodesic calculation between two networks of different sizes.

INFERENCE AND CONTROL IN NETWORKS FAR FROM EQUILIBRIUM.

by

Siddharth Sharma

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Doron Levy, Chair/Advisor
Professor Maria Cameron
Professor Abba Gumel
Professor Wojciech Czaja
Professor William F. Fagan, Dean's Representative

Dedication

To my parents.

Acknowledgments

First and foremost, I would like to thank my advisor Prof. Doron Levy for teaching me how to approach and make progress on a research problem. My journey towards this doctorate would not have been possible without his constant guidance and encouragement. Coming from chemical engineering, the transition to applied mathematics and biophysics had a learning curve. No matter how hopelessly stuck I felt at any point, his experience and feedback always put me back on track. I hope to have an opportunity in the future to support others with the same enthusiasm and positive energy.

I would like to thank the members of my dissertation committee, Prof. Maria Cameron, Prof. Abba Gumel, Prof. Wojciech Czaja and Prof. Bill Fagan for agreeing to serve and guiding me through different stages of the program. Dr. Cameron's course on stochastic methods was my introduction to the subject and I am thankful for her dedication towards her students learning the material, as it made application a natural next step. Prof. Czaja's textbook on Integration theory was used as the texbook for the first ever course I took on measure theory. I have used what I learnt through it to understand previous works and also while developing our method. Prof. Gumel's contributions to the field, especially on the COVID-19 pandemic are an illustration of the significance of quantitative approaches and their applications. My interactions with Prof. Fagan have been illuminating. The applications of quantitative methods to ecology and conservation biology made me realize that I should look for a wider application space for methods that appear specialized and field specific at first glance. I am fortunate to have you all on my committee and hope to emulate your example in the future.

The Maryland Biophysics Program in the Institute for Physical Science and Technology is my home at UMD. An interdisciplinary program with the breadth and openness to new ideas is apt for students from diverse backgrounds to have experiences outside of their traditional areas of learning. Having been pushed out of my comfort zone expanded my skill

set and I feel incredibly grateful for the opportunity. The support of the program directors, Prof. Arpita Upadhyaya and Prof. Jeffrey Klauda has been reassuring throughout. I must mention Souad Nejjar, the biophysics program coordinator for helping all of us from the day of our admission to our graduation.

The Department of Mathematics is where I have worked, learned and taught throughout my PhD. It is the warmth and kindness of people here that makes difficult appear simple. I have been fortunate to work under some incredible course chairs. They have developed me as a teacher and made me a better guide to the students, first as their discussion TA and later as an instructor. I would like to express my gratitude to Prof. Mike Boyle, Prof. David Levermore, Prof. Chris Laskowski, Prof. Amin Gholampour, Matt Griffin, Susan Mazzullo, Karen Mclaren, Casey Cremins and Nathan Manning. I would be remiss not to mention all my students. Your faith in me as your teacher was a huge encouragement throughout. I cannot thank you enough and wish you all the best.

I now want to recognize the staff at the Math department. Whenever I needed help they were always there for me going above and beyond what was required. Jackie, Bill, Martha, Agi, Sara, Liz , Ruth and Christina. You all are cherished and will always be.

My masters thesis advisor Prof. Sudeep Punnathanam at the Department of Chemical Engineering, Indian Institute of Science Bangalore guided me through my first project. My work at IISc exposed me to the scientific way of thinking and what it takes to do serious research. Prof. Punnathanam, Prof. Prabhu Nott, Prof. K Ganapathy Ayappa and Prof. Sanjeev Kumar Gupta shaped my thinking process and prepared me for things ahead. My scientific journey began with the guidance and support of my undergarduate advisors, Prof. Tapan Sarkar and Prof. Uttam Kumar Mandal at the University School of Chemical Technology, Guru Gobind Singh Indraprastha University. Their belief in me helped me take those first steps that have led me to write this thesis.

I have met many friends throughout my time at UMD. Their contribution to this thesis is

immense and instrumental. Whether it be through the RIT meetings on complex networks, machine learning or applied statistics or just discussions about research problems we were working on. It makes one realize that they are not alone and everyone around is dealing with issues relating to research or outside it, making success meaningful when it eventually arrives. I want to thank Pranav, Teju, Shashank and Harshithaa for all the movie nights, *bored* games, birthday cakes and surprises. I don't know what the future holds but with you guys around, I am sure it will be fun.

I was an international student at UMD with no family in the area and really fortunate to have people who looked after me when they had no obligation to do so. I would like to thank Drs. Miriam and Eric Cohen for their love and support and treating me as one of their own. I was just a friend to Sam and your kindness was and still is truly appreciated. I wish all of you the very best. A huge thanks to Sriram and Shanti Kalaga. Ramu uncle and Shanti aunty do not know when to stop while helping someone. A part of me does not want to tell them because some things should never change.

Last but by no means the least, I would like to thank my family for their unconditional love, support and sacrifices that have led to this thesis. Mom, Dad, Arjun, Shuchita, Varenya and Varush. You all have made this thesis possible and it is as much yours as it is mine. I dedicate this thesis to my parents as it is truly a labor of their love.

Table of Contents

Chapter 1: Prologue

Three questions pertinent to this thesis are answered below. The aim is to contextualize methods and results presented in the chapters that follow.

Why should we study biological networks?

The progress in understanding individual components of biological systems such as proteins, genes and neurons, has led to deeper insights about the structure and function of systems that are made from them [3, 55, 71, 112]. The interplay of multiple such components leading to stationary distributions of their collections further emphasized the need for multivariate analysis resulting from observations of multiple degrees of freedom. Network models that included the simultaneous action and evolution of multiple nodes interacting through pairwise couplings were an outcome of such observations [48, 118]. These models and their statistics are an efficient tool to gain further insights [17, 69] into the thus far unknown aspects of living systems and to consequently progress therapeutic modalities, tackling some defining challenges [21, 39, 61, 76, 84, 97] of our time. This is contingent upon the understanding of biological networks and our capacity for targeted control of their components.

Can we control biological networks?

The answer to this question depends on the prevailing capabilities for *in-vivo* editing of biological systems [33]. Due to the pioneering efforts of experimental groups, those capabilities have quantum leaped in the last two decades. The most notable being the CRISPR-Cas9 endonuclease [68, 115] method using palindromic repeats to cleave and rejoin DNA

fragments. Other tools such as zinc-finger nuclease [122], transcription activator-like effector nucleases [78] and other meganucleases [114] have also developed in parallel. In addition, the capacity to interfere at the level of RNA is also present in the form of RNA interference [120]. The promise of these techniques is to induce a paradigm shift in therapeutic regimes for human diseases where instead of interfering in cell structure and function through chemical agents, i.e., drugs, the correction can be carried out at the level of transcription or translation within the central dogma. The minutest lack of accuracy and/or precision could lead to a collapse of such a strategy as robustness is a crucial property of gene-expression. Any errors, especially in a networked structure are catastrophic. In fact, off-target mutations are a sub-topic of gene-editing science [24, 44, 67, 135] while also being a consistent feature of neural circuit manipulations [91]. The issues that arise have nothing to do with the nature of editing itself but can be characterized as relaxations of the network post perturbation, i.e., a new steady-state resulting due to targeted edits leading to additional unexpected off-target mutations. So, as per state of the art at the time of writing this document, biological networks can be edited but not controlled with the accuracy required for prospective therapeutic applications.

How should a biological network be controlled?

A prerequisite for designing a process to go to a desired final state is the knowledge of the initial state. For networks, this is encoded in the coupling matrix, i.e. the matrix of all the pairwise interactions or edges of the network. An observer cannot directly infer the magnitudes of these edges as the only data available is the collective and simultaneous behavior of all the nodes or a subset of nodes that is of interest. On the other hand, edits are made directly to the edges or vertices and the behavioral change is consequential. This implies that an inference of the coupling matrix is the first step towards network control as it is the only way to characterize the interactions, besides the statistical properties of nodes, which are not directly controllable as they are collective in nature and cannot be mapped to specific

nodes or edges. This first step requires inferring the coupling matrix from a set of observations of the networks behavior [103, 133, 134]. A networks steady state determines the type of inference method used for obtaining its coupling matrix. Living systems, through active transport [100][1] exchange matter and energy with their surroundings and therefore do not obey detailed balance (see chapter 2). This has implications for the inference algorithm which cannot use thermal equilibrium and the resulting Maxwell-Boltzmann distribution [38, 89]. So, for biological networks the inference of the coupling matrix needs to be carried out far from equilibrium. In addition, the set of network observations are sometimes not available in time-series [1, 111] and a fundamentally different approach is needed for those type of data where the network snapshots are not arranged as per the thermodynamic arrow of time [108].

With the initial state coupling matrix inferred, the next step is to drive the network to the desired set of couplings and by extension its new steady state distribution. Pioneering efforts [80, 101, 102, 138] have led to numerical algorithms capable of optimally driving between network nonequilibrium steady states. In all of these works, the size of the network remains the same across the transition path. This is in contrast with biology where nodes are added and deleted as a functional feature [131]. This makes biological networks dynamic not only in node states but also node number. This, yet again calls for a different approach where all the dynamic features of the network steady state are accounted for. The resulting methods needs to be informed by the coupling matrices of the initial and final state while allowing for them to have different dimensions. Such a procedure, when implemented would lead to an optimal control of a network nonequilibrium steady state and be a template for biological network control.

[1] Active transport is the movement of molecules or ions across a cell membrane from a region of lower concentration to a region of higher concentration-against the concentration gradient. Active transport requires cellular energy to achieve this movement. There are two types of active transport: primary active transport that uses adenosine triphosphate (ATP), and secondary active transport that uses an electrochemical gradient.

The work presented here uses statistical learning and optimal transport to accomplish the above mentioned tasks.

Chapter 2: Inference and Control in biological networks

2.1 Networks and Biology

Networks are ubiquitous in biology. Although the notion of networks in biology has been that of an analytical tool representing binary interactions among bio-entities, it is experimentally confirmed that such structures exist. A network, also known as a **graph**, consists of two types of components

1. **Nodes or vertices**: These are the basic units that make the network. In biological networks, these can be genes, proteins, neurons, metabolites, epigenetic components or even individual cells when cell signalling network behavior is investigated.

2. **Edges:** These are the connections between the nodes.

If the edges in a network are directed, i.e., pointing in only one direction, the network is called a *directed network* (or a directed graph, sometimes digraph for short). When drawing a directed network, the edges are typically drawn as arrows indicating the direction, as illustrated in Fig.(2.1). If all edges are bidirectional, or undirected, the network is an undirected network (or undirected graph), as illustrated in Fig.(2.2).

Biological networks are generally directed in nature as some edges are almost always directed. For example, the TP53 network shown in Fig.(2.3) has both directed and undirected edges and is therefore, by definition, a directed network. This would be the case for most networks where functional control is relegated to a particular set of nodes as the

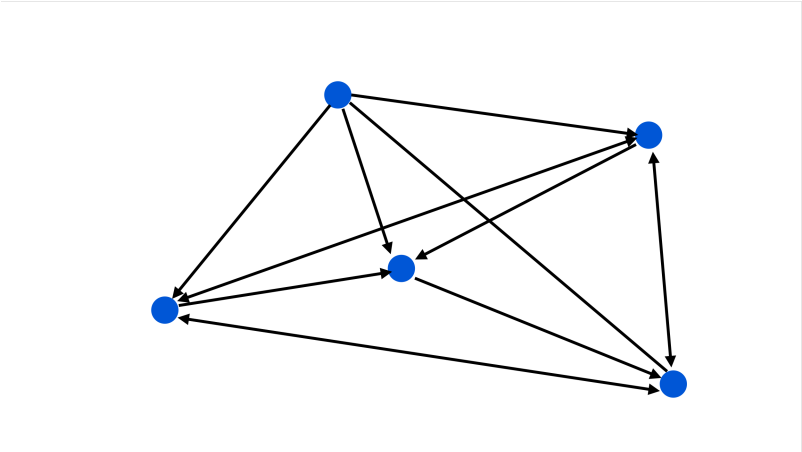


Figure 2.1: A directed network represented by edges with directions. All edges with an arrow along one end are the *directed* edges, whereas the edges with arrows on both ends are undirected.

flow of information would have a bias. This almost universal characteristic of biological networks can be of an advantage for inferring the edge matrix of networks depending upon the quality of data and the belief in clinical/experimental observations. In general, one must assume that the network is undirected and that the entire edge matrix has non-zero non-diagonal entries. The evolution data of the network used to compute the edge matrix would reveal the directed nature of some edges, usually through near zero values for on of the directions. For networks where no prior information about the edges is known, the only option is to assume an undirected edge matrix. For example, directly imaging protein evolution in the brain [95] reveals real time spike activity of neurons (Fig.(2.4)) with no prior information of synaptic connections between them, making it necessary to label all edges as undirected.

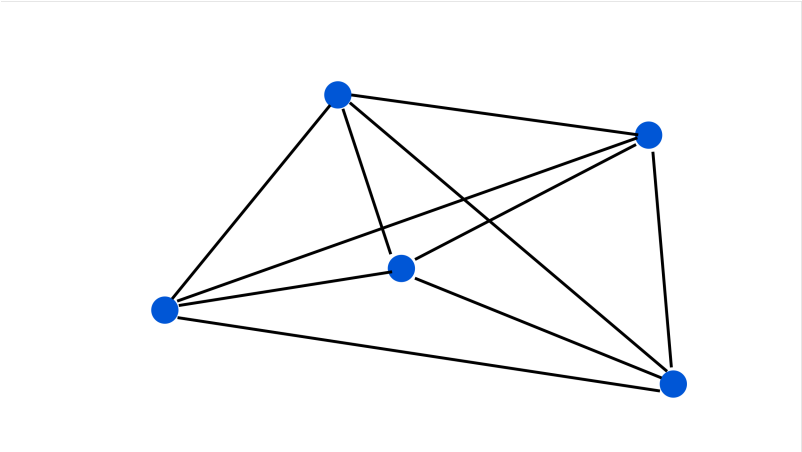


Figure 2.2: An undirected network represented by edges without directions. All edges are assumed to be bi-drectional.

2.1.1 Symmetry, detailed balance and thermal equilibrium

We now discuss the symmetry or the lack thereof in the coupling matrices[1] for biological networks. From the very characterization of a coupling matrix as directed, its asymmetry is a direct consequence. The asymmetry of biological networks, can be understood from the viewpoint of stochastic dynamics of networks with Hamiltonians defined through coupling matrices.

Let T be the $n \times n$ Markov matrix corresponding to a coupling matrix. The elements of T are the $n \times n$ conditional/transition probabilities between two states of the network, i.e, T_{ij} is the probability of going from state i to state j. The probabilities $p_i T_{ij} = p_j T_{ji}$ mean that there is no preference between states i and j for the network. Here, p_i and p_j are the probabilities of states i and j in the stationary distribution (histogram of microstates). For the stationary distribution, the solution to the eigenvalue problem $T\,p = \lambda\,p$ is obtained for

[1] A coupling matrix $\boldsymbol{J}$, contains all the pairwise interactions for a network. It is always an $n \times n$ square matrix where $n \in \mathbb{N}$, is the number of nodes in the network

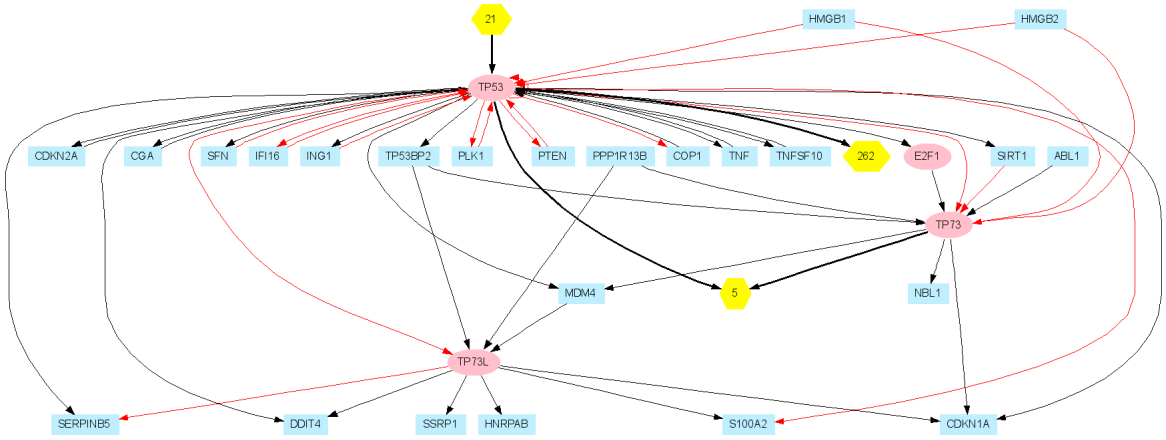


Figure 2.3: Gene Regulatory Network of TF family $p53$ in *homo sapiens*(Image courtesy: Cold Spring Harbor Laboratory, New York). The blue rectangles are genes. The pink ellipses are transcription factors. The yellow hexagons are clustered genes with the number of genes in the cluster as the hexagon label. The red lines indicate that the protein DNA binding is known. The *edges* of this network are not all undirected so by definition it is a directed network. The clustered genes with transcription factors in green are listed as follows.

21 $\longrightarrow$ TP53: PPP1R13L, BTBD14B, APEX1, IRF8, NSEP1, GTSE1, PIN1, ING3, PPM1A, PPP1CB, CARF, SSTR3, THRB, UBE2D2, UBE2D3, YY1, CUL5, PCAF, TNFRSF7, MAGED1, TNFRSF5.

TP53$\longrightarrow$ 262: CEBPA, E2F2, E2F3, E2F4, E2F5, EGR1, ESR1, ETS1, ETS2, FOXD1, FOS, MYC, ATF3, NFIC, NFKB1, BCL2, RELA, BCL3, BRCA1, BRCA2, STAT1, STAT3, TOPORS, ABCC4, TRIM22, PRG3, NDRG1, PIAS3, CDX1, HTATIP, SIVA, PLK2, CHEK1, CHEK2, BIRC8, TP53AP1, CHUK, CLK1, ABCC2, PLK3, KLF6, CSF1, CSPG2, CTNNB1, CTSD, DGKA, DGKB, DGKG, DDB2, GADD45A, AFP, ARID3A, DUSP1, EEF1A1, EEF1A2, EGFR, EPHA2, EIF4EBP1, HIPK1, ERBB2, AKT1, EZH2, FANCC, PTK2B, FBLN1, UNC5B, FBLN2, ALDH1A3, EFEMP1, FDXR, FKBP3, DKK1, KIN, P2RX2, FNBP4, PRG1, GAMT, GAS1, BBC3, SESN1, C20orf10, GCLM, GML, GPX1, RPA4, HD, HIC1, HLA-A, HLA-B, HMGN2, APAF1, APC, APCS, HRAS, BIRC2, BIRC3, HSPA1A, HSPA4, BIRC4, BIRC5, HSPCB, IGF1R, IGF2, IGFBP3, KLK3, TNFRSF6, IL1B, IL2, IL4, IL6, IL8, AQP1, IL11, IL15, IL15RA, JAK2, KAI1, APITD1, RHOA, STMN1, LGALS3, LIG1, LIG2, LIG3, LIG4, TACSTD1, MAP4, MCM2, MGMT, MIF, MKI67, MMP1, MMP2, MMP13, ABCC1, MSH2, JST, COX1, MUC2, BIRC1, NFATC4, NHLH2, NINJ1, NINJ2, NME1, NOS2A, NOS3, NTHL1, P2RX1, P2RX3, P2RX4, P2RX5, RRM2B, PCNA, POLK, LISCH7, PEG3, ABCB1, PIK3CA, PMAIP1, PML, POLD1, ATR, PPP1R7, PPP2R4, SMPD3, PRKAB1, MAPK1, MAPK8, MAPK9, PRODH, PSEN1, PCBP4, BAG1, PTGS2, PTHLH, BIRC6, BAI1, BAI2, BAK1, PTPN13, BAX, CARD12, RAD51, RAD51L1, ACTA2, RB1, CCND1, BCL2L1, RPS6KA1, BDKRB1, S100A4, S100B, CX3CL1, PERP, BLM, WIG1, SIAH1, SLC19A1, SMPD1, SMPD2, SOD2, SSTR2, SYK, TAF9, TBP, TBXAS1, BTG1, TERT, TFRC, TGFA, TGFB1, TGM1, TIMP3, TK1, TOP2A, TP53BP1, TUBA1, TYMS, TYR, TYRP1, WRN, XPC, XRCC5, BTG2, BIRC7, CALM2, PDRG1, CASP1, CASP3, SESN2, MAD1L1, CASP8, AMID, DGKE, DGKD, CDC14B, CDC14A, CAV1, ABCC3, TNFRSF10C, TNFRSF10B, TNFRSF10A, CFLAR, IER3, CCNA2, CCNA1, CCNB1, MBD4, CCNE1, CCNG1, CCNH, P2RXL1, PTTG1, TP53INP1, LITAF, GDF15, TP53I11, EI24, TP53I3, CDC25C, MVP.

TP53 + TP73$\longrightarrow$ 5: HIPK2, MDM2, THBS1, VEGF, CABLES1.

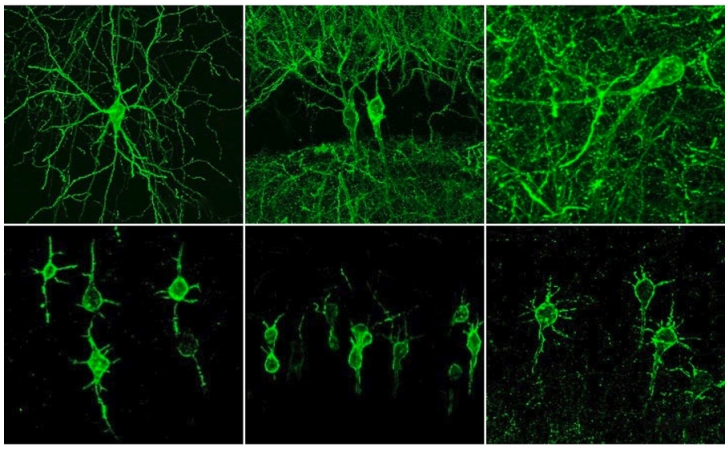


Figure 2.4: Fluorescent imaging of neuron populations reveals their spike patterns [95]. In the top row, the probe employed images all synaptic activity. In the bottom row, the neurons are isolated as the probe specifically accumulates in the neuron which screens the activity from the axons. If a time-series of such a system is recorded with no prior synaptic information for the recorded neurons, then is prudent to assume an undirected edge or coupling network.

$\lambda = 1$[2]. The distribution p that solves $Tp = p$ is the stationary distribution. For a symmetric T which is the result of a symmetric coupling matrix, p is the Maxwell-Boltzmann distribution.

$$p(H_i) = \frac{e^{-\beta H_i}}{Z}, \tag{2.1}$$

where $Z = \sum_{i=1}^{M} e^{-\beta H_i}$ is the partition function for a system with M microstates. This distribution is a characteristic of systems with symmetric transition matrices corresponding to **detailed balance** or a lack of a probability current within the system. This means that $p_i T_{ij} = p_j T_{ji} \ \ \forall i, j \in M$.

The thermodynamic implication here is that a system which obeys detailed balance is in thermal equilibrium. At this point it is good to review the basic tenets of equilibrium. Equilibrium describes an averaged behavior over many realizations of the same system

[2]This condition is actually called global balance. The requirement here is the symmetry of the matrix T. The condition $p_i T_{ij} = p_j T_{ji}$ mentioned is actually known as *local detailed balance* and is a stronger condition.

under similar conditions or multiple systems under the same set of conditions. The primary characteristics are

1. Equilibrium is not static and must not be conflated with kinetic arrest. It refers to dynamics with constant averaging.

2. There is no net flow in any real, conformation or population space of either material, probability or reactions.

The former is a characterization of thermal equilibrium as dynamic rather than static. The latter condition is a manifestation of detailed balance. This is because detailed balance is the balance of flows between any (and every) pair of states that can be defined. Not just "detailed" infinitesimal states, but a balance of flows among tiny states implies a balance of flows among global states which are groups of tiny states. The "flow" as defined here refers to the motion of material or trajectories/probability, depending on the situation at hand. A few illustrative examples are of help here

(a) **A Solution mixture:** The first step is to divide our system into multiple sub-volumes. We now consider two adjacent sub-volumes out of these i.e. sub-volumes i and j (See Fig.(2.5(a))). In just these two sub-volumes, some set of molecules will move from region to the other in a given time interval. For a solution at equilibrium, there will be other systems in which the opposite flow occurs. Averaging over all systems, there is no net flow of any type of molecule between any two sub-volumes in equilibrium. This is a detailed balance.

(b) **Molecule conformations:** In the conformation space of a single molecule, if only two conformal states i and j are available, then there will be an equal number of $i \longrightarrow j$ and $j \longrightarrow i$ transitions over any interval of time. See Fig.(2.5(b)).

(c) **Chemical Reaction(s):** For a single reaction the equilibrium is determined by the ratio of the reaction rates, defined by the well-known equilibrium constant $K_{eq} = \frac{k_f}{k_b}$. In Fig.(2.5(c)), it is $K_{eq} = \frac{k_{ij}}{k_{ji}}$.

As it is evident, equilibrium dynamics vary in the type of quantities that change over a period of observation for different systems. What is a consistent feature is that the time average and ensemble average both approach the same average measurement. This is explained by Daniel. M. Zuckerman as

> In equilibrium, time averaging and ensemble averaging will yield the same result. To see this, consider a solution containing many molecules diffusing around and perhaps exhibiting conformational motions as well. Assume the system has been equilibrating for a time much longer than any of its intrinsic timescales (inverse rates). Because finite-temperature motion in a finite system is inherently stochastic, over a long time each molecule will visit different regions of the container and also different conformations - in the same proportion as every other molecule. If we take a snapshot at any given time of this equilibrium system, the "ensemble" of molecules in the system will exhibit the same distribution of positions and conformations as a long single trajectory of any individual molecule. This has to be true because the snapshot itself results from the numerous stochastic trajectories of the molecules that have evolved over a long time [136].

This coinciding of time and ensemble averages is a result of detailed balance as the locally stochastic sub-volume level time measurements lead to the total volume, i.e., ensemble level measurements. This collective averaging would be the same as if the total system was measured and averaged over time.

With the nature of equilibrium described along with its connection to detailed balance and

symmetric coupling matrices, the next step is identify the nature of biological networks and their steady states. The fact that directed networks are the norm in biophysics means that detailed balance is not a possibility in most situations and should definitely not be an assumption for any type of inference or control algorithm unless there is strong experimental/clinical evidence to do so.

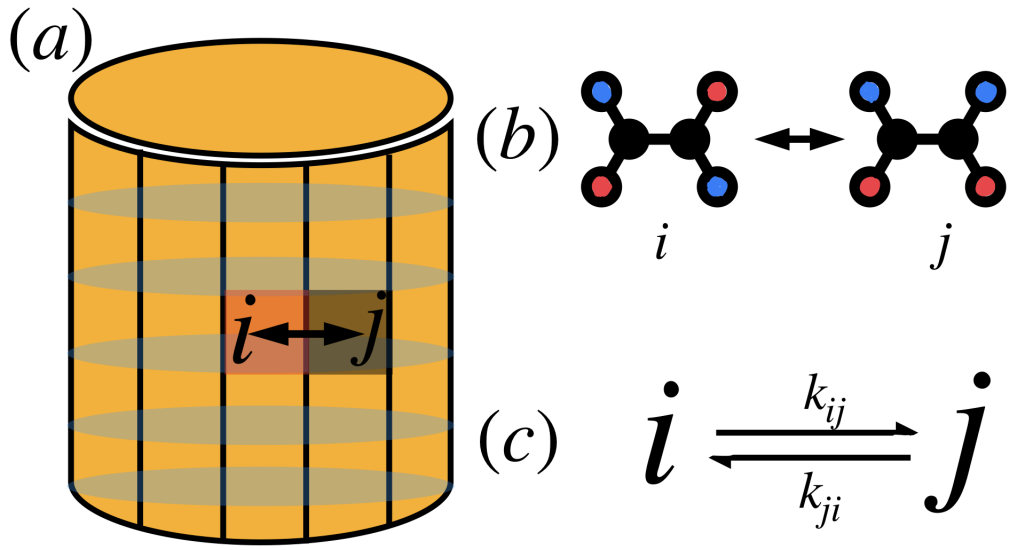


Figure 2.5: A schematic of the three illustrative examples of equilibrium/detailed balance.

2.1.2 Irreversible processes and the Nonequilibrium steady-state

If a system is left isolated, i.e., no matter or energy exchange occurs in the long time limit, then it would eventually equilibrate. This itself excludes biological systems from equilibration as the consistent exchange of matter and energy is a hallmark of most processes at almost every scale. From the ATP energy exchange and storage to the ion-exchange mechanisms of cellular membrane voltage gates, to the numerous enzymatic processes that abound within the cell organalles, all require energy and matter to be transported across the cytoplasm to maintain what is known as cellular homeostasis. In fact, the earliest works on abiogenesis elucidate principal scenarios such as the primordial soup where the

exchange of matter and energy within an initial aqueous mixture of inorganic compounds leads to synthesis of organic compounds and eventually amino acids. Charles Darwin in his letter to Dalton Hooker commented:

> It is often said that all the conditions for the first production of a living being are now present, which could ever have been present. But if (and oh what a big if) we could conceive in some warm little pond with all sort of ammonia and phosphoric salts,—light, heat, electricity present, that a protein compound was chemically formed, ready to undergo still more complex changes, at the present such matter would be instantly devoured, or absorbed, which would not have been the case before living creatures were formed [35].

Subsequent works by the likes of Urey and Miller [86] proved the existence of such mechanisms under laboratory settings, making it evident that exchange of matter and energy was a feature of the first organic processes. This means that the characterization of steady-states of biological systems do not constitute thermal equilibrium at the system scale and consequently detailed balance and network edge-symmetry.

Looking at thermodynamic details, the processes with energy and mass exchange as a feature would generally have some irreversible steps, i.e., events which have a very low probability of occurring in the reverse directions, e.g., chemical reaction equilibrium with skewed reaction rate ratios, thereby pushing the reaction essentially in only one direction. One such example is enzyme kinetics with the Michealis-Menten kinetics as one of the basic models. For an enzyme E and substrate S, the binding leads to the enzyme-substrate complex ES. The next step is the catalysis where the enzyme-substrate complex leads to the product P.

The reaction can be written as

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k}_b}{\overset{\mathrm{k}f}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k}c} \mathrm{P} + \mathrm{E}$$

In the second step, the backward reaction rate is so small as compared to the forward rate that the reaction is considered irreversible. This makes the enzyme catalysis irreversible in the catalytic step. This means that the ensemble average of any subset of phase-space of this process would not give an ensemble (or time) averages where the backward step i.e. $\mathrm{P} + \mathrm{E} \longrightarrow \mathrm{ES}$ is going to happen at a rate comparable to its forward counterpart, making equilibration impossible. This is just a very basic example of how a system has irreversible steps leading to a steady state that is not characterized by detailed balance. Such a steady state is called a **Nonequilibrium Steady-State**.

Directed networks have asymmetric coupling matrices, don't obey detailed balance and operate in a nonequilibrium steady-state (NESS). Relevant examples for Biophysics include Gene regulatory networks, networks of neurons, metabolic networks, etc. Since the steady-state of a system far from equilibrium is generally not characterized, therefore the algorithms for inferring biological networks have to be designed not for thermal equilibrium but for a NESS. Having established the nature of the steady-states in which a biological network is expected to exist and its impact on the numerical methods for inference and control, the next task is to investigate the need and significance for such processes and the challenges they present for precision medicine in systems biology.

2.2 Precision Medicine and the Nonequilibrium Steady-State

Therapeutic procedures are by and large disease specific, but due to different kinds of variations at the population or individual level, this "one-size fits all" approach has its

shortcomings. The FDA defines **precision medicine**[3], sometimes known as "personalized medicine" as an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles. The goal of precision medicine is to target the right treatments to the right patients at the right time. The idea is to move away from the idea of the "average patient" and to quantify the differences between population groups and individuals based on the Next-generation sequencing (NGS) data. The nature of biological networks is of significance for such a goal. Far from equilibrium dynamics need to be inferred before any type of control algorithm can be designed. In addition the lack of availability of time-ordered NGS data poses a different set of challenges. Solutions to these problems that stem from the nature of the steady state distributions of biological networks and the experimental/clinical data available from them, are crucial to the success of precision medicine. In context of the present work, we attribute the prospect of successful therapeutic algorithms to the following three sequential tasks.

1. The experimental procedures for obtaining high-quality NGS data
2. Inferring the coupling matrices of networks of interest.
3. Designing control algorithms for networks far from equilibrium with asymmetric coupling matrices, specific to the experimental techniques at hand.

The accuracy and precision of first task lies in the advancements in experimental methods for data acquisition and processing. The progress has been considerable and as observers we are optimistic that future recording techniques [106] would be even more conducive to statistical analysis and model recognition. One such avenue is the development of $in-vivo$ single-cell methods [29] that do not destroy the cell. This combined with gene editing techniques has a paradigm shifting potential for downstream processing as time-series data

[3] https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine

with targeted edits would be a possibility.

The two remaining tasks are the focus of this work, albeit for very simple model systems. We are hopeful that the extendable methods we propose would be useful for additional complexities arising with the evolution of data sets. We briefly discuss the three steps mentioned as follows.

2.2.1 Big data in network biology: New insights into pathology

The central dogma of molecular biology is the standard first principle in genomics. It was the first insight into the specificity of protein synthesis. In 1970, Francis Crick defined it as the residue by residue transfer of sequential information. This reinstated his earlier description form 1958 about the unidirectional nature of the process where the sequence information is passed from double-stranded DNA to single-stranded RNA and finally to proteins but any kind of retrieval, i.e., information transfer in the reverse direction is impossible. To briefly state the process: Proteins are biopolymers that are central to the functionality of the cell. Processes ranging from metabolism to reproduction to antigen defense are all dependent on proteins. The sequences for every synthesized protein are encoded in specific regions of the DNA called genes. These genes are transcribed by polymerase enzymes to produce mesenger-RNAs which are translated by ribosomes to produce proteins.The entire collection of proteins produced by a genome is called the *gene-expression* of a cell. This is a stationary characteristic of the cell and by extension the organism which is more or less invariant over laboratory observations. This control is mainly attributed to proteins that are extraneous to the DNA called transcription factors which bind to the promoter region of a gene i.e. the site on the DNA where the transcription of gene commences. Finally, there is combinatorial control [15] exerted over the gene-expression through multiple transcription factors controlling the expression of multiple genes.

Advances in microarray techniques [109] has made gene expression mapping a common reference for research over the past two decades. The pivot of almost all high-throughput expression mapping techniques is the process of generating complimentary DNA(cDNA) from mRNA using the enzyme reverse transcriptase. These cDNA libraries are then used as a measure of individual gene expression. The two main classes of methods used for mapping eukaryotic gene expression are

1. **Microarray DNA sequencing:** These are microchips with small DNA sequences planted on them to act as probes [121]. After reverse transcription of the mRNA, the cDNA sequences are fluorescently labelled and are allowed to bind with their complimentary DNA sequences on the microarray. The amount of cDNA on a particular probe gives an estimate of the concentration of its complimentary mRNA and by extension the expression of the particular gene associated with it. An issue that results from the methodology is that mRNA concentrations have to be averaged within in a population of cells.

2. **RNA-seq:** This method [125] is similar to DNA sequencing up to the cDNA synthesis. Afterwards, it uses sequencing adaptors attached to the cDNA fragments to read short sequences. These read sequences are then aligned with the genome and classified into three types:

    - Exonic reads: These are completely contained within the exons of the reference genome.
    - Junction reads: These are the reads lying on the junction between two genes.
    - Poly(A) end-reads: Polyadenylation marks the end of transcription of gene as it adds multiple molecules of adenine to the end of mRNA transcribed. When the reads are aligned as complimentary to these, they are called poly(A) end-reads.

Using these reads, an expression profile of each gene in the library can be mapped [54]. It would be known from the final sequence that which genes are active and by how much. The main advantage of RNA-seq is the direct measurement of mRNA levels as opposed to DNA-microarrays.

The above mentioned techniques have revolutionized the understanding of genotype-phenotype correlations at an unprecedented rate. It has been possible to identify loci all across genomes that are significant in a wide variety of diseases [53]. See Fig.(2.6) for an example.

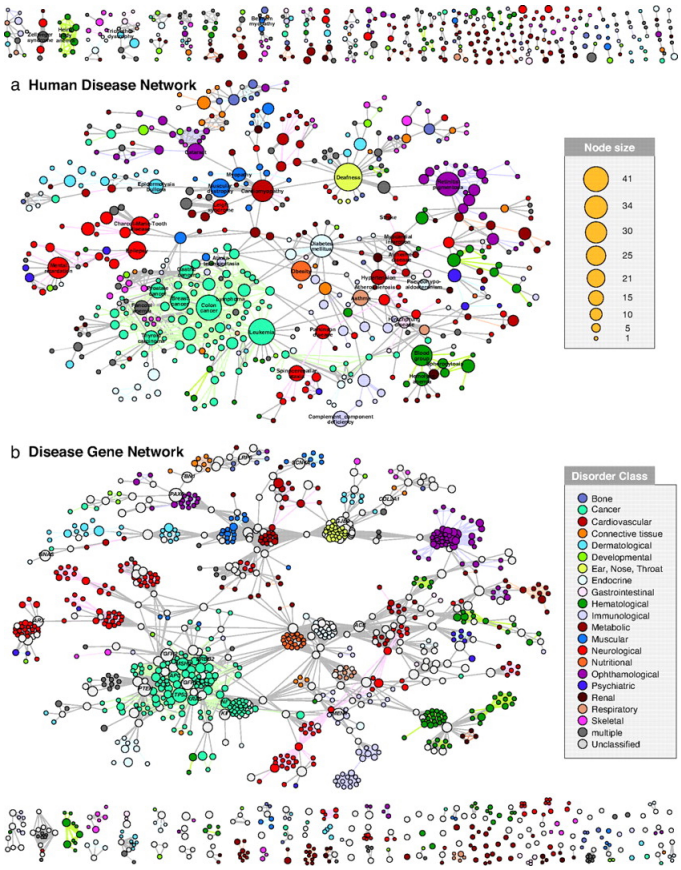


Figure 2.6: The HDN and the DGN. (a) In the HDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs, the name of the 22 disorder classes being shown on the right. A link between disorders in the same disorder class is colored with the corresponding dimmer color and links connecting different disorder classes are gray. The size of each node is proportional to the number of genes participating in the corresponding disorder, and the link thickness is proportional to the number of genes shared by the disorders it connects [52].

This has lead to a surge in studies trying to pin down mechanisms of processes that were previously partially understood. Issues ranging from aging [113] to ailments like alzhiemers [85], diabetes [63] and heart disease [14] have been studied in unprecedented detail. Parallel to these developments, there have been instances where the apparent genotypes were not particularly aligned with their observed phenotypic variants [31, 43, 65, 137]. This was expected from a system with a high number of degrees of freedom. The next question was: How to uncover the complexity that was not explained by the standard

model of DNA(gene)-RNA-Protein chain given by Crick?

Gene-regulatory networks(GRNs) [7] were one of the most prominent answers to this question. A GRN is defined as a collection of genes or parts of genes that interact to control a particular cellular function. This leap of thought from genes as individual switches [60] to parts of a larger system is not uncharacteristic of biology and even less so for systems biology. Different cellular functions are known to be controlled by a collection of molecules acting in conjunction to form signaling pathways providing cohesion and control. Extending that idea to genes has made it possible to explain the observed correlations to a greater extent. It is natural that the investigations that followed gravitated towards the least understood phenotypes [94].

Understanding Cancer with GRNs

Cancer is a collection of diseases which are characterized by the ability of cells to divide uncontrolled with a potential to spread to secondary locations in addition to the initial site of proliferation. This makes it a condition resulting from gene dysregulation [16, 98, 99] since processes like cellular division, motility, cohesion, mortality, etc., are transcribed and tightly controlled. This gives cancer cells a sense of free will where they do everything from evading the immune system to arresting healthy cell functionality so that their own survival and proliferation is paramount [58]. As a result, in cases where detection is late prognosis could be poor. An interrogative opinion about the lack of understanding about cancer could be stated as follows.

> *Different aspects of neoplastic pathology are initiated, progressed and regulated by distinct parts of the genome. Yet it is possible to have a unifying principle consisting of almost universal symptoms appearing as if they were*

co-regulated by a common information processing unit. [58]

The common processing unit here implies a refined layer of information processing that consists of not only individually transcribed genes but also a set of interactions among different genes all over the genome which significantly distinguishes their collective expression from the logical sum. There are several instances in which including regulatory genomics in studies may be of benefit. One primary example is that of cancer.

Cancer proliferation and progression are multistage processes with simultaneous as well as sequential appearance of characteristic changes to processes that are critical to cellular function. It is therefore standard clinical practice that therapies are multi-pronged and therefore parallely disruptive to all the progression factors. A schematic of this fact is presented in Fig.(2.7) where the ten hallmarks of cancer are shown with their respective therapeutic protocols [58].

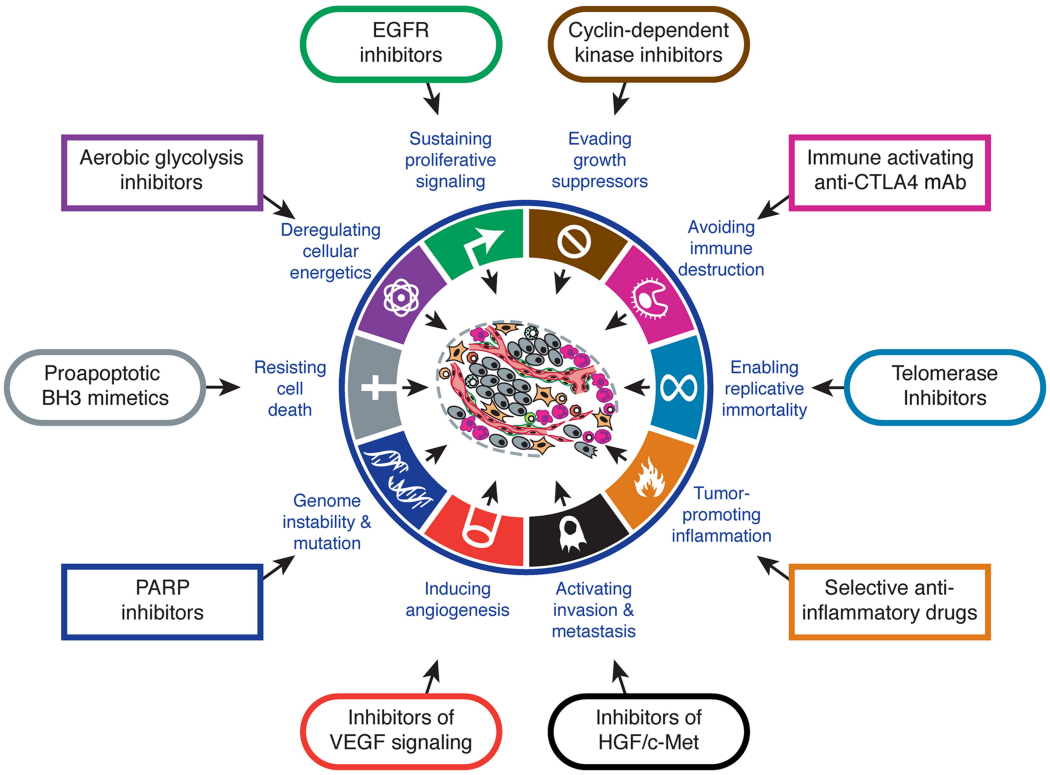


Figure 2.7: Therapeutic targeting of the hallmarks of cancer [58].

It is well known that many cancer patients are over-treated [8, 62] and during therapy, even healthy cells can be disrupted or destroyed. In addition, standard therapeutic approaches are not guaranteed to disrupt malignant cells. Any analytical necessity would enable identifying the main targets for a specific type of malignancy and then to design the minimally disruptive yet effective therapies that do not adversely affect cells that are neither a part of neoplasia nor are assisting in progression.

2.2.2 Identifying and characterizing gene interactions: The inverse Ising problem

Life at its core creates higher order structures and more complex functional organization from smaller individual units [107]. Clearly, multi-cellular organisms are more complex than single cells. At first glance this seems to be a violation of the second law of thermodynamics as life, almost like a struggle against entropy, is creating lower entropy structures using constituent units which collectively have a higher entropy as compared to the structures or organisms they constitute. This would be true if one assumes thermal equilibrium. However, in absence of thermal equilibrium, higher biological complexity is a valid outcome. One may find this as an all pervasive characteristic of life if its origins are looked at more carefully. The first unicellular organisms appeared on earth at three main sites: Tidal pools, geothermal vents in the deep oceans and water below the ice sheets. All three of these places have one thing in common, i.e., they all sit on energy gradients. These energy gradients provided the essential *free energy* to construct more complex structures. This did not violate the second law as the first living cells were open systems exchanging matter and energy with their surroundings. This basic trait of living systems should not be neglected for modeling biological systems, including GRNs, networks of neurons and other relevant classes of biological networks. A far from equilibrium model informed from

the physics of the system is required to correctly model most biological systems. This is contextual to the type of dynamics being observed. For example, if the node states of a network are classified as binary, then the model of network interactions needs to be a binary state network interaction model. If on the other hand the observation is that of communities of nodes of multiple states, then a multi-species model is needed. The present work focuses mainly on the former case with the fully-connected Ising model far from equilibrium.

The theory of complex networks intersects significantly with those a random graphs and biological networks are no exception. As mentioned earlier, the goal here is to identify a simple approximation for biological networks and to derive numerical methods for inference and control. The particular type of random graph used to model binary state interactions is the fully connected Ising model of N nodes where the state of each spin s_i is denoted by

$$\sigma_i(t) = \{+1, -1\} \quad \forall t \in [0, \infty). \tag{2.2}$$

The notation in Eq.(2.2) denotes that the nodes can have binary up or down spins as their states in a continuous-time stochastic process. So, we have

$$\boldsymbol{\sigma}(t) = [\sigma_1(t), \sigma_2(t), \cdots, \sigma_N(t)], \tag{2.3}$$

as an array of time varying spins connected by an asymmetric coupling matrix. This matrix $\boldsymbol{J}$ along with the external field vector $\boldsymbol{\theta}$, represent the Ising Hamiltonian:

$$E(\boldsymbol{\sigma}(t)|\boldsymbol{J}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \theta_i \sigma_i(t) + \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \sigma_i \sigma_j \quad . \tag{2.4}$$

The local field at each vertex is written as

$$H_i(t) = \theta_i(t) + \sum_{j=1}^{N} J_{ij}\sigma_j(t) \tag{2.5}$$

In the present case, self-couplings are allowed, i.e., $J_{ii} \neq 0$. The task at hand is to infer the square matrix $\boldsymbol{J}$ and the field vector $\boldsymbol{\theta}$, given a set of network configurations

$$[\boldsymbol{\sigma}(t_1), \boldsymbol{\sigma}(t_2), \boldsymbol{\sigma}(t_3), \cdots, \boldsymbol{\sigma}(t_M)] \quad \text{where} \quad t_1 < \cdots\cdots < t_M, \tag{2.6}$$

denoting the time-sequenced nature of the network configurations. This is the inverse problem in statistical mechanics and it is the central method of inference used in the present work for the binary state biological networks considered in their original or binarized form. In a forward problem, the Hamiltonian is given and the configurations of the systems, i.e., microstates are generated through an analytical or numerical solution. The one-dimensional Ising model has an analytical solution but a fully connected Ising graph far from equilibrium needs to numerically simulated to observe its behavior. The method used for the same in this work is the asynchronous Glauber dynamics explained is the next chapter. In an inverse problem on the other hand, the observables are the microstates and the unknown parameters are the interactions between the spins. The goal of the numerical method is to learn the interaction matrix from statistical properties of the data available. In case of the Ising model, these are the magentisation

$$m_i = \langle \sigma_i \rangle, \tag{2.7}$$

and the pairwise correlations

$$c_{ij} = \langle \sigma_i \sigma_j \rangle, \tag{2.8}$$

where the angular brackets denote ensemble averages. The subsequent chapters include a discussion of the inverse problem under two classifications of the available data, i.e., time-series and time-disordered data. The former uses is the standard inverse problem while the latter needs a deeper discussion of the properties of a set of configurations under the action of the permutation group leading to a approximation of the thermodynamic arrow of time. The discussion at the point of definition and the solution is self-contained so we defer the introduction to the detailed problem to the subsequent chapters in both cases.

2.2.3 Network control far from equilibrium

The definition of network control varies as per the diversity of systems and the environment. Thus it is of importance that a contextual definition is provided. Consider two network steady-states with different probability distributions, i.e., relative frequency distributions of microstates. **The problem of network control is to optimally drive a network from one steady-state to another**. The two steady-state distributions of the individual Markov chains are called the initial and target distributions respectively. A basic idea to establish a sense of distance between two distributions p and q of a continuous random variable x, is their relative entropy or the Kullback-Leibler divergence

$$KL(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \quad . \tag{2.9}$$

If x is discrete, then

$$KL(p||q) = \sum_i p_i(x) \log\left(\frac{p_i(x)}{q_i(x)}\right) \quad . \tag{2.10}$$

Relative entropy is a quantity encountered frequently in information theory and machine learning. To demonstrate the meaning behind the relative entropy, we calculate it for two

normal distributions $\mathcal{N}(0,2)$ and $\mathcal{N}(2,2)$.

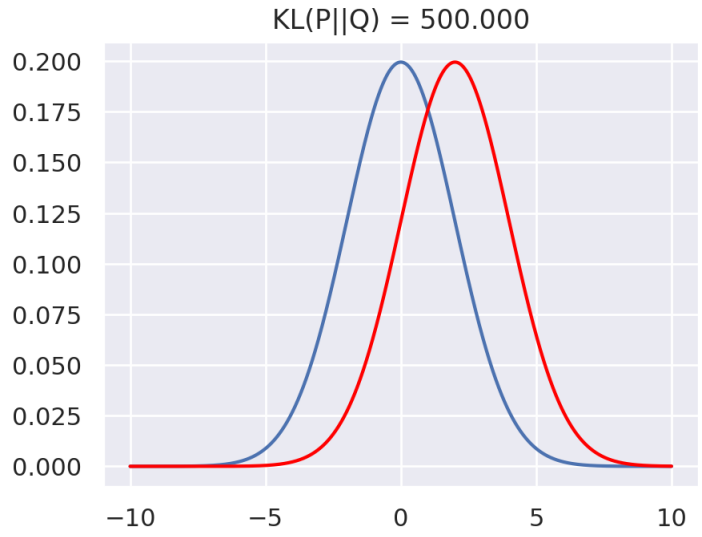


Figure 2.8: The KL-divergence between $\mathcal{N}(0,2)$ and $\mathcal{N}(2,2)$. The relative entropy is a measure of the overlap between the two distributions.

The divergence value is 500 (see Fig.(2.8)) and as can be seen in the formulae above which only contain probability distributions, it is a dimensionless quantity. The divergence is higher when the overlap is lesser as shown in Fig.(2.9). $KL-$divergence is not symmetric and when p and q are switched as shown in Fig.(2.10).

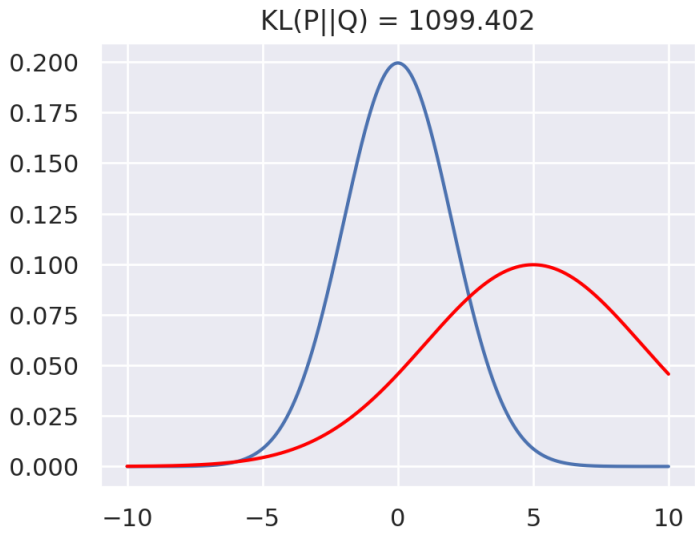


Figure 2.9: The *KL*-divergence between $p = \mathcal{N}(0,2)$ and $q = \mathcal{N}(5,4)$. The relative entropy is a measure of the overlap between the two distributions and is greater than the case in Fig.(2.8).

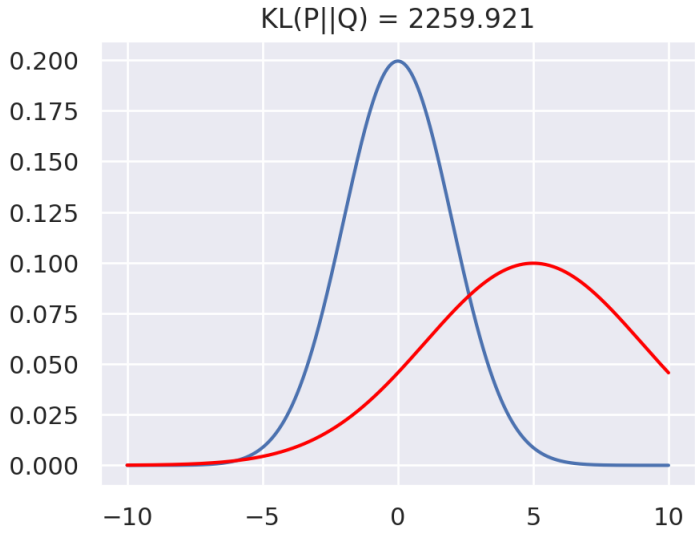
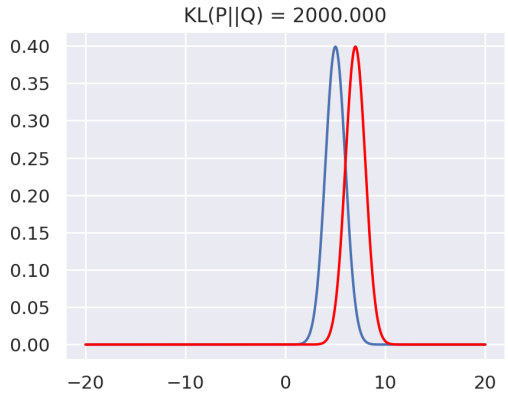
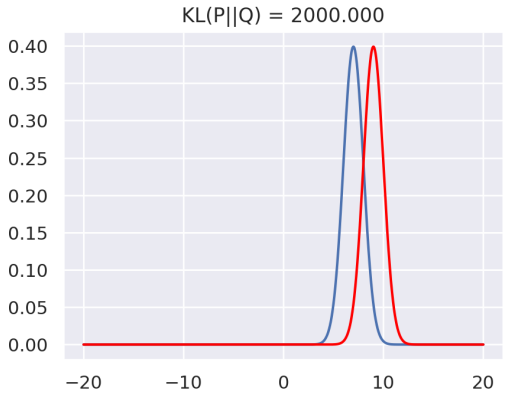


Figure 2.10: The *KL*-divergence between $p = \mathcal{N}(4,5)$ and $q = \mathcal{N}(0,2)$. The relative entropy is a measure of the overlap between the two distributions and is different from the case in Fig.(2.9) as relative entropy is asymmetric.

As this behavior of the relative entropy is observed, it begs the question how sensitive it is to translation along the random variable x. It is shown in Fig.(2.11a) and Fig.(2.11b).



(a) The KL-divergence between $p = \mathcal{N}(5,1)$ and $q = \mathcal{N}(7,1)$



(b) The KL-divergence between $p = \mathcal{N}(7,1)$ and $q = \mathcal{N}(9,1)$

Figure 2.11: An $x-$translation of the random variable leaves the $KL-$divergence unchanged to 2000. This means that the same distribution overlap at any μ_x for p and q results in the same relative entropy.

This property of the $KL-$divergence of measuring just overlap and not translation can also be extended to the fact that the overlap measured for two distributions with almost no overlap is going to approach infinity no matter their separation along the random variable axis. This is not ideal for physical systems as the translation of the mean has physical implications. To take an elementary example of the velocity distribution of a gas, a shift in the mean of the distribution would indicate an increase in the average velocity, i.e., temperature and if two positions of the curve are treated as the same, then the relative entropy of the two distributions fails to capture the true energetic cost of the transition in terms of either work done or heat absorbed. Taking the example of cancer, studies have shown that not only structure and function but the levels of proteins markedly change when compared to healthy tissue [23, 105]. If a protein interaction network of such proteins is considered, then higher average values of node specific protein concentrations is observed. This is an indication of a shift in the value of the first moment of the distribution of concentration and again it is insufficient to just consider the shape of the distribution and not its displacement about the original mean.

Optimal transport: A basic introduction for biological network control

In terms of a network far from equilibrium, if the probability distribution of microstates, i.e., the steady state distribution is changed, the relative entropy is not the ideal method to calculate the *effective cost* of that change. Especially, if the difference between the initial and the target distributions is not just of shapes but also of their respective means. In fact if relative entropy is used in such a case, then the calculated cost only takes into account the overlap and fails to capture the translation of the mean even if the shape remains the same. This raises the question, "How does one **optimally** transport a probability distribution?" The answer(s) lie in *optimal transport*. The details of the theoretical concepts are given in the last two chapters of this thesis but here we try to develop an introductory discussion of the basic ideas and extend the same to biological networks.

One can imagine a steady-state probability distribution of a network as a sequence of *likelihood masses* where each microstate is assigned its own mass, making the histogram akin to a sand pile. A method to shift the pile to another location in a different shape is required and the simplest way to do so is to define a cost function, which multiplied by the individual masses gives the work required to move the mass from one location to another in addition to also moving it among the microstates, i.e., change the shape of the sand pile. The French mathematician Gaspard Monge wrote this specific problem for *probability sand piles* and the most efficient way of moving them. The original Monge problem consists of finding the optimal way for rearranging a Borel probability measure μ_I on $\mathbb{R}^d$ onto a Borel probability measure μ_T on $\mathbb{R}^d$ against the cost function $c(\boldsymbol{z}) = ||\boldsymbol{z}||$. The work involved by a particle of mass $d\mu_T(\boldsymbol{x})$ moving from a point $\boldsymbol{x}$ to a point $r(\boldsymbol{x})$ along a smooth path $t \longrightarrow g(t,\boldsymbol{x})$

between time 0 and 1 is

$$\int_0^1 ||\dot{g}(t,\boldsymbol{x})||dtd\mu_I(\boldsymbol{x}) \ . \tag{2.11}$$

Hence, the total work for the entire pile is

$$L[\boldsymbol{g}] := \int_{\mathbb{R}^d}\int_0^1 ||\dot{g}(t,\boldsymbol{x})||dtd\mu_I(\boldsymbol{x}) \ . \tag{2.12}$$

We have to account for the fact that the pile of likelihoods still has the same mass, i.e.,

$$\mu_I[r^{-1}(A)] = \mu_T(A) \tag{2.13}$$

for all Borel sets $A \in \mathbb{R}^d$. The Monge formulation of the problem is somewhat less useful as compared to the Kantorovich formulation which consists of transport plans of the form $\mu(\boldsymbol{x},\boldsymbol{y})$. A transport plan is a joint probability distribution with marginals μ_I and μ_T. They are cartesian products on $\boldsymbol{X}\times\boldsymbol{X}$ and the projections of a transport plan on each axis are the steady states μ_I and μ_T. The transportation price for a given plan is an integral of the metric as a function of two variables, and the task is to minimize this value over all admissible transport plans: in obvious notation it takes the form

$$K_r(\mu_I,\mu_T) := \inf\left[\int_{\boldsymbol{X}}\int_{\boldsymbol{X}} r(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x},\boldsymbol{y})\right] \tag{2.14}$$

This thesis only deals with finite systems i.e. biological networks and therefore by definition $\mu(\boldsymbol{x},\boldsymbol{y})$ is a probability measure. Here, $r(\boldsymbol{x},\boldsymbol{y})$ is the matrix of the cost of transporting a unit of sand from one point to another, but it is not necessarily a metric. In finite systems, the Kantorovich formulation is better suited and therefore we develop our methods later based on this approach. A property of this formulation is that it requires the distri-

butions μ_I and μ_T to be of the same size, i.e., the number of nodes of the network in the two steady states should be equal. This is not true for biological networks. Microarray data has constantly demonstrated the recruitment of new genetic and epigenetic factors leading to dysregulation and consequently compromised or lost functionality. This aspect of biological networks we refer to as codimensionality of their Markov Chains and its impact on optimal transport is discussed in the next section.

2.3 Network growth in biology

Spatial biological networks, i.e., networks where the vertices are bioentities in real space and endowed with a metric, exist at various scales. From GRNs to ecosystems, they perform various significant functions from signal transduction to nutrient transport. The performed functions are a result of the network architecture and are a direct consequence of the developmental processes. During critical developmental processes, the size of these functional networks changes to include new functions and exclude redundancies as needed. This variation in size is important to understand both development and disease. At any scale, it is difficult to observe network growth experimentally and therefore theoretical models and simulations based on them are to be relied upon. The first model of biological network growth was published in 1967 [30] and there has been constant progress ever since. The main issue is the domain specificity of different models and a consequent lack of a generalized growth model. There are results for development of new edges between preexisting nodes which is apt for modeling branching neurons or leaf vascular networks. Then there is another class of models where specific nodes are targeted to grow outwards with branching and merging. These models are well-suited for neural network development [22, 70, 81], branching of organ structures [59] or even fibrous materials [49]. This crucial lack [9] of a unified growth model means that optimal control cannot be guaranteed if a

particular network growth model is applied to a certain scenario. It is reasonable to look for alternative approaches and we again turn towards optimal transport and the Kantorovich formulation.

We still consider the same example and follow the same notation we used for developing the Kantorovich problem with one difference: The dimensions of μ_I and μ_T are different. This means that instead of

$$K_r(\mu_I, \mu_T) := \inf \left[\int_{\boldsymbol{X}} \int_{\boldsymbol{X}} r(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x}, \boldsymbol{y}) \right] \tag{2.15}$$

We have

$$GW_r(\mu_I, \mu_T) := \frac{1}{2} \inf_{\mu(X \times Y) \in \Gamma} \left[\int_{\boldsymbol{X} \times \boldsymbol{Y}} \int_{\boldsymbol{X} \times \boldsymbol{Y}} [\omega(x_i, x_j) - \omega(y_i, y_j)]^2 d\mu(x_i, y_i) d\mu(x_j, y_j) \right] \tag{2.16}$$

5.

2.4 Synopsis

We introduced complex networks and classified them as directed and undirected. Biological networks were explained to be directed and examples from $p53$ and neuron populations were given to support the categorization. This was followed by a brief introduction to the thermodynamics of biological networks and their far from equilibrium dynamics. Their steady states are distinguishable from thermal equilibrium by the lack of detailed balance and the resulting Maxwell-Boltzmann distribution. Next, the fully connected asymmetric Ising model was introduced as an approximation to certain types of biological networks. The dynamics, albeit simple, is an appropriate platform for testing numerical algorithms while also being extendable to more complex phenomena, e.g., multiple species of vertices

instead of just two. Precision medicine is a significant component of the future treatment protocols for many human diseases. Its requirement and the steps necessary for its realization include inference and control of biological networks. This means the capability to infer and optimally control nonequilibrium steady states. This is the significance of the two problems solved in this thesis. The inverse problem in statistical mechanics is to be solved for the Ising model to accomplish the task of inference far from equilibrium. This needs to be done for time-ordered as well as time-disordered data as certain systems, e.g., gene-regulatory networks, do not provide enough time-series data. For control, the usage of optimal transport as opposed to relative entropy is discussed in context of driving between steady state distributions that not only differ in their shapes but also their location on the horizontal axis. Finally, biological networks grow and shrink as they evolve in structure and function. We emphasize the need to incorporate this feature in any numerical method that aims to optimally drive between two network steady states. All the above tasks, in the sequence they are discussed, will be accomplished in the next four chapters.

Chapter 3: The inverse problem far from equilibrium

3.1 Introduction

In chapter 2 we introduced biological systems as open systems that exchange material and energy with their surrounding environment. Next, the fully-connected Ising model was shown to be an appropriate choice for a pairwise interactions that would be driving the dynamics and functionality of such biological networks. The fully-connected Ising model is a good approximation for several types of networks, e.g., neuron interaction networks, somatic mutation maps between normal and disease states, etc., while being an acceptable approximation of binarized versions of others. Binarization is a field on its own where random variables are mapped to the field of integers, specifically the set $\{-1, +1\}$. The methods mentioned here are specific to binary data but the basic principle can be easily extended to multi-species mixture networks, e.g., GRN's with more than two values of vertices. The model is deceptively complex even with the binary version because of the generally intractable nature of the systems steady state. Nevertheless, the algorithm is general and can be extended to multi-species models.

We start with formulating the inverse problem far from equilibrium. We are following the traditional setup of maximum likelihood estimation [42, 87]. The problem in its original form is tough to solve even with considerable computational resources and therefore we use latent variable [25, 73, 77, 96] approximations into the original likelihood function making it quadratic [40, 51] in the estimated parameters. The latent variables used merit

their own discussion. Finally, the effect of variations in data length, standard deviation of the coupling matrix, and probability rate of node selection are shown so a hyperparameter [46, 130] dependence of the inference method can be mapped.

3.2 Structure of the Problem

We consider a fully connected asymmetric Ising graph of N nodes where the state of each spin s_i is denoted by

$$\sigma_i(t) = \{+1, -1\} \quad \forall t \in [0, \infty) \tag{3.1}$$

The notation in Eq.(3.1) refers to the nodes that can have binary up or down spins as their states in a continuous-time stochastic process. Accordingly,

$$\boldsymbol{\sigma}(t) = [\sigma_1(t), \sigma_2(t), \cdots, \sigma_N(t)], \tag{3.2}$$

is an array of time varying spins connected by an asymmetric coupling matrix. This matrix $\boldsymbol{J}$ along with the external field vector $\boldsymbol{\theta}$, represent the Ising Hamiltonian:

$$E(\boldsymbol{\sigma}(t)|\boldsymbol{J}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \theta_i \sigma_i(t) + \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \sigma_i \sigma_j \ . \tag{3.3}$$

The local field at each vertex is written as

$$H_i(t) = \theta_i(t) + \sum_{j=1}^{N} J_{ij} \sigma_j(t) \ . \tag{3.4}$$

In the present case, self-couplings are allowed, i.e., $J_{ii} \neq 0$. The task at hand is to

infer the square matrix $\boldsymbol{J}$ and the field vector $\boldsymbol{\theta}$, given a set of network configurations $[\boldsymbol{\sigma}(t_1), \boldsymbol{\sigma}(t_2), \boldsymbol{\sigma}(t_3), \cdots, \boldsymbol{\sigma}(t_M)]$ where $t_1 < \cdots\cdots < t_M$, denoting the time-sequenced nature of the network configurations.

3.3 Asynchronous Glauber dynamics

The term asynchronous [134] here refers to flipping only one spin in a certain time interval Δt. This type of dynamics would be observed for biological systems like gene-regulatory networks where the flipping of individual nodes is slow on observed time-scales as compared to the thermal motion in the surrounding environment. Synchronous updates make more sense of artificially constructed systems like social networks, financial markets, etc.

For the system described above, the probability of a spin to be selected for flipping over a time period Δt is given by $\gamma \Delta t$. Here, $\gamma > 0$ is the probability rate of spin-selection which remains constant over the entire observation time. The quantity γ can be thought of as a property of the network itself. This also gives the probability of a spin not selected for flipping to be $1 - \gamma \Delta t$. The probability of spin being flipped post selection is given as

$$P_i^f(t) = \frac{e^{-(\sigma_i(t) H_i(t))}}{2\cosh(H_i(t))} \quad . \tag{3.5}$$

This gives the probability of a spin not being flipped during the interval Δt as

$$1 - \gamma \Delta t + \gamma \Delta t \left(1 - P_i^f(t)\right) \quad . \tag{3.6}$$

Since the flipping of a spin in a particular configuration is incumbent upon that configuration resulting as the outcome of a previous transition, therefore the flip and no-flip probabilities are conditional probabilities of a Markov Chain. Thus, the total probability of a

time-series of network configurations can be written as

$$P\left(\boldsymbol{\sigma}_N\left([0,t]\right)|\boldsymbol{J},\boldsymbol{\theta}\right)=\prod_{(i,t)\in F}\left\{\gamma\Delta t\frac{e^{-(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)}\right\}\times\prod_{(i,t)\in NF}\left\{1-\gamma\Delta t+\gamma\Delta t\frac{e^{(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)}\right\}\ . \tag{3.7}$$

The first term on the right-hand side is the multiplication of all the flipping probabilities of nodes which were selected for flipping ($(i,j)\in F$). The second term consists of multiplication probabilities of nodes which were either not selected for flipping or the ones which were selected but not flipped ($(i,j)\in NF$).

3.4 Likelihood for a time-series of network configurations

As mentioned earlier, the approach discussed here is of likelihood maximization. To that end we have

Theorem 3.1. *For a kinetic Ising model with coupling matrix $\boldsymbol{J}$, external fields $\boldsymbol{\theta}$ and spin selection probability rate γ, the complete data likelihood for a time sequenced set of configurations over a time interval $(0,T]$ is given by*

$$\mathscr{L}\left(\boldsymbol{\sigma}_N\left([0,t]\right)|\boldsymbol{J},\boldsymbol{\theta}\right)=\prod_{(i,t)\in F}\left\{\frac{e^{-(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)}\right\}\times\prod_{i=1}^{N}\exp\left(\gamma\int_0^T\left\{\frac{e^{(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)}-1\right\}dt\right)\ . \tag{3.8}$$

Proof. The time sequence total probability is given by equation 3.7. We now use the definition of the likelihood function as

$$\mathscr{L}\left(\theta;x\right)=\prod_i f\left(x_i;\theta\right)=\prod_i\frac{dP(\theta)}{dm}\ , \tag{3.9}$$

where the last term on the right is the Radon-Nikodym derivative [45, 57, 126] of the

probability function P w.r.t the Lebesgue measure m [12]. At this point, we need to observe the difference between the flipping and non-flipping terms. The flipping terms are the points within the time-series where the spins flipping are selected with probability $\gamma\Delta t$ and the spin selected is flipped through Glauber dynamics. The rest of the time, the spins which get selected are not flipped. It is therefore clear that the former are discrete events while the latter are continuous accumulations of probability terms in a product along the arrow of time. This is when we use the Radon derivative to construct the likelihood function. First, we consider the flipping term. The Lebesgue measure is $\gamma\Delta t$. This means that this term is equal to

$$\mathscr{L}^f = \frac{dP^f}{d(\gamma\Delta t)} = \frac{\frac{dP^f}{d(\Delta t)}}{\frac{d(\gamma\Delta t)}{d(\Delta t)}} = \prod_{i,t\in f} \frac{e^{\sigma_{i(t)}H_{i(t)}}}{2\cosh H_i(t)} \quad . \tag{3.10}$$

As for the second term, the lack of flipping event occurs over each time interval Δt and as long as the flipping does not occur, the probability product keeps accumulating new no-flip probabilities. Consider the Radon derivative

$$\frac{dP^{nf}}{d(\gamma\Delta t)} = \frac{\frac{dP^{nf}}{d(\Delta t)}}{\frac{d(\gamma\Delta t)}{d(\Delta t)}} = \prod_{i,t\in nf} \left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh H_i(t)} - 1 \right) \quad . \tag{3.11}$$

It can be observed that there are no-flip events for all N spins in this product and all of them can be considered as decomposition's of a multivariate Poisson process [72] which has the probability rate represented by the product on the right hand side. This leads to a direct definition of the likelihood function for a single spin as follows

$$\frac{dP}{dt} = \gamma\left(\frac{e^{(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)P \ .$$

$$\implies \frac{dP}{P} = \gamma\left(\frac{e^{(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)dt \ .$$

$$\implies \log P\big|_{P(t=0)}^{P(t=T)} = \gamma\int_0^T\left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)dt \ .$$

$$\log\mathscr{L}_i^{nf} - \log 1 = \gamma\int_0^T\left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)dt \ .$$

$$\mathscr{L}_i^{nf} = \exp\left[\gamma\int_0^T\left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)dt\right] \ .$$

The complete no-flip likelihood for all N spins can then be written as,

$$\mathscr{L}^{nf} = \prod_{i=1}^{N}\exp\left[\gamma\int_0^T\left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right)dt\right] \ . \tag{3.12}$$

This gives the complete-data likelihood as,

$$\mathscr{L}\left(\boldsymbol{\sigma}_N\left([0,t]\right)|\boldsymbol{J},\boldsymbol{\theta}\right) = \mathscr{L}^f \times \mathscr{L}^{nf}$$

$$= \prod_{i,t\in F}\left\{\frac{e^{-(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)}\right\} \times \prod_{i=1}^{N}\exp\left(\gamma\int_0^T\left\{\frac{e^{(\sigma_i(t)H_i(t))}}{2\cosh\left(H_i\left(t\right)\right)} - 1\right\}dt\right) \ . \tag{3.13}$$

This completes the proof. □

The optimization of this likelihood by gradient ascent with respect to the couplings and fields will solve the inverse problem. The issue is the complexity of the solution which has two sources:

1. The no-flip integration being inside an exponential function.

2. The hyperbolic cosines in the denominators of both the flipping and no-flip terms.

We note that this computational issue arises because they have the fields and couplings are the arguments of the exponential and hyperbolic functions. This makes inverting these functions an issue while solving .

3.5 Latent Variables

A latent variable in context of a statistical model is a variable that is not observed or directly measured. Such variables can only be inferred when the model is reformulated in a form which includes them. Such models that aim to explain the values taken by the observed variables through processes described by the latent variables are known as latent variable models. This is precisely what we aim to do in the present case. The original model of the likelihood function gave us the observed variables, i.e., the couplings $\boldsymbol{J}$ and the external fields $\boldsymbol{\theta}$, in a form which makes their inference computationally intractable. So, we describe the same complete-data likelihood through latent variables which remedy the above problem. This leads us to define variables that eliminate the exponential function of the no-flip integral and to define another set of latent variables for elimination of the hyperbolic cosine. We begin with the former.

3.5.1 Poisson random variables for no-flip likelihood function

We decompose the no-flip integral in a manner similar to its construction. Recall that while constructing the integral, we used the Radon derivative to get the probability rate of the Poisson process of a spin not flipping for a time interval. Vice-versa, we order the intervals in which a spin remains unchanged and order them as $n \in \{0, 1, \cdots, n_{\max}\}$. For a spin σ_i that remains unflipped over an interval n denoted by σ_i^n, the value of local field H_i^n remains constant over n. This means that the no-flip Glauber probability integral can be

approximated as

$$\int_0^T \left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} \right) dt \approx \sum_{n=0}^{n_{max}} \left(\frac{e^{\sigma_i^n H_i^n}}{2\cosh\left(H_i^n\right)} \right) \left(t_{n+1} - t_n\right) \quad . \tag{3.14}$$

At this stage, the no-flip Glauber probability integral can be approximated as

$$\int_0^T \left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)} \right) dt \approx \sum_{n=0}^{n_{max}} \left(\frac{e^{\sigma_i^n H_i^n}}{2\cosh\left(H_i^n\right)} \right) \left(t_{n+1} - t_n\right) \quad . \tag{3.15}$$

At this stage we look at the moment generating function of the Poisson distribution [72]

Theorem 3.2. *Let X be a discrete random variable which is Poisson distributed with parameter λ where $\lambda \in \mathbb{R}_+$. Then the moment generating function M_X of X is given by $M_X(t) = e^{\lambda(e^t-1)}$.*

Proof. For a Poisson distribution

$$P(X = n) = \frac{\lambda^n e^{-\lambda}}{n!} \quad . \tag{3.16}$$

By definition

$$\begin{aligned} M_X\left(t\right) = \mathbb{E}\left[e^{tX}\right] = \sum_{n=0}^{\infty} P(X=n)e^{tn} \implies M_X\left(t\right) &= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} e^{tn} \\ = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\left(\lambda e^t\right)^n}{n!} &= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)} \quad . \end{aligned} \tag{3.17}$$

□

By theorem 3.2 for the moment generating function, we get the no-flip integral for a

single spin as

$$\exp\left[\gamma\int_0^T\left(\frac{e^{\sigma_i(t)H_i(t)}}{2\cosh\left(H_i\left(t\right)\right)}-1\right)dt\right]=\prod_{n=1}^{n_{max}}\sum_{\rho_i^n=0}^{\infty}\left(\frac{e^{\sigma_i^nH_i^n}}{2\cosh\left(H_i^n\right)}\right)^{\rho_i^n}e^{-\gamma(t_{n+1}-t_n)}\frac{\gamma\left(t_{n+1}-t_n\right)^{\rho_i^n}}{\rho_i^n!}\quad . \tag{3.18}$$

Here the variables σ_i^n are Poisson distributed for each spin i and each time interval n of σ_i being constant. The complete product of sums is the total probability of spin i being constant over any number of times slices within the n^{th} interval of constant σ_i . The product just extends the same to all the intervals of constant σ_i with the notation, $t_0=0$ and $t_{n_{max}+1}=T$. So the product is that of moment generating functions over the entire set $n\in\{0,1,\cdots,n_{\max}\}$.

3.5.2 Põlya-Gamma distribution

Consider the example of a logistic regression. The probability distribution being binomial, the likelihood function can be written down as

$$\mathscr{L}\left(p\right)\propto\prod_{n=1}^{N}p^{y_n}\left(1-p\right)^{1-y_n}=p^{\Sigma y_n}\left(1-p\right)^{N-\Sigma y_n}\quad . \tag{3.19}$$

The probabilities are parameterized as follows.

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots\cdots+\beta_Dx_D=\boldsymbol{\beta}^T\boldsymbol{x}\ . \tag{3.20}$$

Solving for p we get,

$$p = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}} \quad , \tag{3.21}$$

and

$$1 - p = \frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}} \quad . \tag{3.22}$$

So, the likelihood function can be written as

$$\mathscr{L}(\beta) = \frac{[e^{\boldsymbol{\beta}^T \boldsymbol{x}}]^{\Sigma y_n}}{[1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}]^N} \quad . \tag{3.23}$$

The likelihood function in Eq.(3.23), when used for Bayesian inference, would require a prior distribution multiplied with it and the resulting product to be normalized. For this reason, the regression is intractable. To remedy this Polson et al. introduced an augmentation scheme using Põlya-Gamma random variables [96]. For $\omega \sim PG(b,c)$ with $b > 0$ and $c \in \mathbb{R}$, the distribution has the form

$$PG(\omega|b,c) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{\left(k - \frac{1}{2}\right)^2 + \left(\frac{c}{2\pi}\right)^2} \quad . \tag{3.24}$$

when $g \sim gamma(b,1)$ are gamma distributed random variables. The moments of this density can be all written in closed form despite the complicated form of the function itself. This property comes in handy as the expectation is written as

$$\mathbb{E}[\omega] = \frac{b}{2c} \tanh\left(\frac{c}{2}\right) \quad . \tag{3.25}$$

It is important in the present context to have the closed form expressions for the expectation of the Põlya-Gamma variable. To get to a point where this result can be proven, we first

derive two significant results which are as follows.

Theorem 3.3. *For* $\omega \sim PG(b,c)$

$$\frac{(e^{\psi})^{a}}{(1+e^{\psi})^{b}} = 2^{-b} e^{\kappa\psi} \int_{0}^{\infty} e^{-(\omega\psi^{2}/2)} p(\omega)\, d\omega \ . \tag{3.26}$$

Proof. Let us begin by writing

$$a = \kappa + \frac{b}{2}$$

Next, we use the above transformation in the left-hand side of Eq.(3.26).

$$\begin{aligned}\frac{(e^{\psi})^{a}}{(1+e^{\psi})^{b}} = \frac{e^{\left(\kappa\psi+\frac{\psi b}{2}\right)}}{(1+e^{\psi})^{b}} = \frac{e^{\kappa\psi}e^{\frac{\psi b}{2}}}{(1+e^{\psi})^{b}} = \frac{e^{\kappa\psi}(e^{\frac{\psi}{2}})^{b}}{(1+e^{\psi})^{b}} = \frac{e^{\kappa\psi}}{\left(\frac{1+e^{\psi}}{e^{\frac{\psi}{2}}}\right)^{b}} \\ = \frac{e^{\kappa\psi}}{\left(e^{-\frac{\psi}{2}}+e^{\frac{\psi}{2}}\right)^{b}} = \frac{e^{\kappa\psi}}{\left(2\cosh\left(\frac{\psi}{2}\right)\right)^{b}} = \frac{2^{-b}e^{k\psi}}{\cosh^{b}\left(\frac{\psi}{2}\right)}\end{aligned} \tag{3.27}$$

At this stage, we need another fact about Põlya-gamma distributions. It is that they are subset of a class of distributions which are infinite convolutions of the gamma distribution.

$$PG(\omega|1,0) = \frac{1}{2\pi^{2}} \sum_{k=1}^{\infty} \frac{g_{k}}{\left(k-\frac{1}{2}\right)^{2}} \ . \tag{3.28}$$

When $g_k \sim Gamma(1,0)$ are mutually independent gamma distributed random variables.

Now, we take the Laplace transform of $\omega \sim PG(b,0)$ as

$$\mathbb{E}\left[e^{-\omega t}\right] = \prod_{i=1}^{t} \left(1 + \frac{t}{2\pi^2\left(k-\frac{1}{2}\right)^2}\right)^{-b} \tag{3.29}$$

From the Weierstrass Factorization theorem, we can write the product on the right hand side of Eq.(3.29) as

$$\frac{1}{\cosh^b\left(\sqrt{\frac{t}{2}}\right)} \tag{3.30}$$

Next, we replace t with $\frac{\psi^2}{2}$ giving us,

$$\mathbb{E}\left[e^{-(\omega\psi^2)/2}\right] = \cosh^{-b}\left(\frac{\psi}{2}\right) \tag{3.31}$$

This gives

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = \frac{2^{-b}e^{k\psi}}{\cosh^b(\psi/2)} = 2^{-b}e^{\kappa\psi}\mathbb{E}[e^{-(\omega\psi^2)/2}] \tag{3.32}$$

The expectation in Eq.(3.32) is with respect to $\omega \sim PG(b,0)$. Applying the definition of expectation to we get:

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}e^{\kappa\psi}\mathbb{E}\left[e^{-(\omega\psi^2)/2}\right] = 2^{-b}e^{\kappa\psi}\int_0^{\infty} e^{-(\omega\psi^2)/2}p(\omega)d\omega \ , \tag{3.33}$$

where $\kappa = a - \frac{b}{2}$ and $p(\omega) = PG(\omega|b,0)$. This proves the result. □

The second result deals with the exponential tilting of $PG(b,0)$. It is as follows.

Theorem 3.4. *$PG(\omega|\psi) \sim PG(b,\psi)$.*

Proof. The expectation of $e^{-(\omega\psi^2)/2}$ is

$$\mathbb{E}[e^{-(\omega\psi^2)/2}] = \int_0^\infty e^{-(\omega\psi^2)/2} p(\omega)\, d\omega \ . \tag{3.34}$$

Next we normalize a Põlya-gamma distribution for a tilting parameter c as,

$$PG(\omega|b,c) = \frac{e^{-\frac{c^2}{2}\omega} PG(\omega|b,0)}{\mathbb{E}[e^{-(c^2\omega)/2}]} \tag{3.35}$$

For $c = \psi$ we get

$$PG(\omega|b,\psi) = \frac{e^{(-\psi^2\omega)/2} \times PG(\omega|b,0)}{\mathbb{E}[e^{-(\psi^2\omega)/2}]} = PG(b,\psi) \ . \tag{3.36}$$

The denominator acts as a normalization constant for the tilted probability density i.e. a partition function. This completes the proof. □

At this point one would wonder how the results thus far would help us solve the problem of logistic regression or for that matter, the inverse problem. The utility of the Põlya-Gamma augmentation for logistic regression is that it helps us in constructing a Gibbs sampler for the likelihood. Lets take a quick detour towards the concept of Gibbs sampling. Note that for our purpose, it is not essential to construct a Gibbs sampler but just have closed form expressions for expectation of Põlya-Gamma variables. We look at Gibbs sampling to better understand the solution to logistic regression using latent variables. The context of the original solution is actually what makes the variables themselves applicable to the inverse problem. *PG* augmentation helps to stratify Gaussian mixtures. We want to use the variables for the hyperbolic cosine denominators which by design contain all the different types of transitions as products. The cosine hyperbolic is just the closed form of their averages and we want to transition that to expectations of exponential functions.

3.5.3 Gibbs Sampling

Gibbs Sampling is a Markov-chain Monte Carlo method used to obtain a sequence of observations (system configurations) from a multivariate probability distribution when direct sampling from the distribution is not feasible. In the present case, we use a Gibbs sampler as a method to solve logistic regression which is an inference problem. Recall from Eq.(3.23) that the likelihood function for all observations of the Bernoulli random variable is

$$\mathscr{L}(\beta) = \frac{[e^{\boldsymbol{\beta}^T \boldsymbol{x}}]\Sigma^{y_n}}{[1+e^{\boldsymbol{\beta}^T \boldsymbol{x}}]^N} \quad . \tag{3.37}$$

This gives the contribution of the n^{th} observation to be

$$\mathscr{L}_n(\beta) = \frac{[e^{\boldsymbol{\beta}^T \boldsymbol{x}_n}]^{y_n}}{[1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_n}]} \quad . \tag{3.38}$$

from Eq.(3.33) we can write the above likelihood as

$$\begin{aligned} \mathscr{L}_n(\beta) = \frac{[e^{\boldsymbol{\beta}^T \boldsymbol{x}_n}]^{y_n}}{[1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_n}]} &= -2e^{(\kappa_n \boldsymbol{\beta}^T \boldsymbol{x}_n)} \mathbb{E}_{p(\omega_n|1,0)} \left[e^{-\frac{\omega_n (\boldsymbol{\beta}^T \boldsymbol{x}_n)^2}{2}} \right] \\ &= -2e^{(\kappa_n \boldsymbol{\beta}^T \boldsymbol{x}_n)} \int_0^\infty e^{-\frac{\omega_n (\boldsymbol{\beta}^T \boldsymbol{x}_n)^2}{2}} p(\omega_n|1,0) d\omega_n \quad , \end{aligned} \tag{3.39}$$

where $\kappa_n = y_n - \frac{1}{2}$. The basic tenet of a Gibbs sampler is to condition on one variable while changing the others. In this present case, we condition on ω_n, i.e., we keep it constant. This

makes the n^{th} observation's contribution to the likelihood Gaussian in $\boldsymbol{\beta}$. This means that

$$p(\boldsymbol{\beta}|\Omega,y) = p(\beta)\prod_{n=1}^{N}\mathscr{L}_n(\boldsymbol{\beta}|\omega_n) \ , \tag{3.40}$$

where $\Omega = diag(\omega_1,\omega_2,\cdots,\omega_N)$ is a diagonal matrix with its elements as the N latent, i.e, Polya-Gamma variables. It is also to be noted that if we condition on the latent variables, the expectation is constant. This means that

$$p(\boldsymbol{\beta}|\Omega,y) = p(\beta)\prod_{n=1}^{N}\mathscr{L}_n(\boldsymbol{\beta}|\omega_n) \propto p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{(\kappa_n\boldsymbol{\beta^T x_n})}\times e^{-\frac{\omega_n(\boldsymbol{\beta}^T\boldsymbol{x}_n)^2}{2}}$$

This product can be transformed by completion of squares.

$$p(\boldsymbol{\beta}|\Omega,y) \propto p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{(\kappa_n\boldsymbol{\beta^T x_n})}\times e^{-\frac{\omega_n(\boldsymbol{\beta}^T\boldsymbol{x}_n)^2}{2}} = p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{(\kappa_n\boldsymbol{\beta^T x_n})-\frac{\omega_n(\boldsymbol{\beta}^T\boldsymbol{x}_n)^2}{2}}$$

$$= p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta^T x_n})^2-2\frac{\kappa_n\boldsymbol{\beta}^T x_n}{\omega_n}\right)} = p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta^T x_n})^2-2\frac{\kappa_n\boldsymbol{\beta}^T x_n}{\omega_n}+\left(\frac{\kappa_n}{\omega_n}\right)^2-\left(\frac{\kappa_n}{\omega_n}\right)^2\right)} \ .$$

At this point we recall that the conditioning is with respect to $\boldsymbol{\omega}$ and therefore the least term i.e. $-\left(\frac{\kappa_n}{\omega_n}\right)^2$ is just a constant giving the following result

$$p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta^T x_n})^2-2\frac{\kappa_n\boldsymbol{\beta}^T x_n}{\omega_n}+\left(\frac{\kappa_n}{\omega_n}\right)^2-\left(\frac{\kappa_n}{\omega_n}\right)^2\right)} \tag{3.41}$$

$$\propto p(\boldsymbol{\beta})\prod_{n=1}^{N}e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta^T x_n})-\left(\frac{\kappa_n}{\omega_n}\right)\right)^2} \ .$$

At this point we can write

$$p(\boldsymbol{\beta}|\Omega,y) = p(\beta)\prod_{n=1}^{N}\mathscr{L}_n(\boldsymbol{\beta}|\omega_n) \propto p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x_n})-\left(\frac{\kappa_n}{\omega_n}\right)\right)^2} \quad . \tag{3.42}$$

Next, $\frac{\kappa_n}{\omega_n}$ is a quotient of variables on the same index set and can be combined to one single variable z_n i.e. $\boldsymbol{z} = \left(\frac{\kappa_1}{\omega_1}, \frac{\kappa_2}{\omega_2}, \cdots\cdots, \frac{\kappa_n}{\omega_n}\right)$ giving the quadratic form on the R.H.S below

$$\begin{aligned} p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{\omega_n}{2}\left((\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x_n})-\left(\frac{\kappa_n}{\omega_n}\right)\right)^2} &= p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{1}{2}\left((\boldsymbol{z}-\boldsymbol{X\beta})^T\boldsymbol{\Omega}(\boldsymbol{z}-\boldsymbol{X\beta})\right)} \\ = p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{1}{2}\left([\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{X^{-1}z})]^T\boldsymbol{\Omega}[\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{X^{-1}z})]\right)} &= p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{1}{2}\left((\boldsymbol{\beta}-\boldsymbol{X^{-1}z})^T\boldsymbol{X}^T\boldsymbol{\Omega X}(\boldsymbol{\beta}-\boldsymbol{X^{-1}z})\right)} \end{aligned} \tag{3.43}$$

Finally, we can write the proportionality relation

$$p(\boldsymbol{\beta}|\Omega,y) \propto p(\boldsymbol{\beta})\prod_{n=1}^{N} e^{-\frac{1}{2}\left((\boldsymbol{\beta}-\boldsymbol{X^{-1}z})^T\boldsymbol{X}^T\boldsymbol{\Omega X}(\boldsymbol{\beta}-\boldsymbol{X^{-1}z})\right)} \quad . \tag{3.44}$$

This means that if the prior $p(\boldsymbol{\beta})$ is Gaussian, then the posterior $p(\boldsymbol{\beta}|\Omega,y)$ can be written as a Gaussian in $\boldsymbol{\beta}$ too. This enables us to sample $\boldsymbol{\Omega}$ given $\boldsymbol{\beta}$ and then $\boldsymbol{\beta}$ given $\boldsymbol{\Omega}$. This is exactly what a Gibbs sampler is. The main idea can be written as a pair of time-discretized probability relations:

$$\boldsymbol{\Omega}(t+1) \sim p(\boldsymbol{\Omega}|\boldsymbol{\beta}(t)) \quad , \tag{3.45}$$

$$\boldsymbol{\beta}(t+1) \sim p(\boldsymbol{\beta}|\boldsymbol{\Omega}(t+1)) \quad . \tag{3.46}$$

Our next task is to identify where these densities would come from. The first one, i.e., $p(\boldsymbol{\Omega}|\boldsymbol{\beta}(t))$ is the exponential tilting relation we proved in Theorem 3.4. More specifically

$$p(\omega_n|\boldsymbol{\beta}) \sim PG(1, \boldsymbol{\beta}^T x_n) \ . \tag{3.47}$$

Recall that $\boldsymbol{\Omega} = diag(\omega_1, \omega_2, \cdots, \omega_N)$ and therefore the sampling would be along the diagonal of the latent matrix for each observations' contribution. The second relation requires a derivation related to the summation of quadratic forms. To that end

Lemma 3.1.

$$(\boldsymbol{X} - \boldsymbol{\alpha})^T \boldsymbol{\Psi}^{-1} (\boldsymbol{X} - \boldsymbol{\alpha}) + (\boldsymbol{X} - \boldsymbol{\zeta})^T \boldsymbol{\Phi}^{-1} (\boldsymbol{X} - \boldsymbol{\zeta}) \tag{3.48}$$

can be written as the sum of a single quadratic form and constant term that is independent of $\boldsymbol{X}$.

Proof. First we expand both terms as

$$(\boldsymbol{X} - \boldsymbol{\alpha})^T \boldsymbol{\Psi}^{-1} (\boldsymbol{X} - \boldsymbol{\alpha}) = \boldsymbol{X}^T \boldsymbol{\Psi}^{-1} \boldsymbol{X} - 2\boldsymbol{\alpha}^T \boldsymbol{\Psi}^{-1} \boldsymbol{X} + \boldsymbol{\alpha}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\alpha} \ , \tag{3.49}$$

$$(\boldsymbol{X} - \boldsymbol{\zeta})^T \boldsymbol{\Phi}^{-1} (\boldsymbol{X} - \boldsymbol{\zeta}) = \boldsymbol{X}^T \boldsymbol{\Phi}^{-1} \boldsymbol{X} - 2\boldsymbol{\zeta}^T \boldsymbol{\Phi}^{-1} \boldsymbol{X} + \boldsymbol{\zeta}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\zeta} \ . \tag{3.50}$$

Next we rearrange terms as

$$\begin{aligned} &(\boldsymbol{X} - \boldsymbol{\alpha})^T \boldsymbol{\Psi}^{-1} (\boldsymbol{X} - \boldsymbol{\alpha}) + (\boldsymbol{X} - \boldsymbol{\zeta})^T \boldsymbol{\Phi}^{-1} (\boldsymbol{X} - \boldsymbol{\zeta}) \\ &= \boldsymbol{X}^T (\boldsymbol{\Psi}^{-1} + \boldsymbol{\Phi}^{-1}) \boldsymbol{X} - 2(\boldsymbol{\alpha}^T \boldsymbol{\Psi}^{-1} + \boldsymbol{\zeta}^T \boldsymbol{\Phi}^{-1}) \boldsymbol{X} + (\boldsymbol{\alpha}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\alpha} + \boldsymbol{\zeta}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\zeta}) \end{aligned} \tag{3.51}$$

This is quadratic in $\boldsymbol{X}$. We can write a more condensed form.

$$(\boldsymbol{X}-\boldsymbol{\alpha})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\alpha})+(\boldsymbol{X}-\boldsymbol{\zeta})^T\boldsymbol{\Phi}^{-1}(\boldsymbol{X}-\boldsymbol{\zeta})=\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}-2\boldsymbol{B}\boldsymbol{X}+\boldsymbol{C}\ . \tag{3.52}$$

where

$$\boldsymbol{\Psi}^{-1}+\boldsymbol{\Phi}^{-1}=\boldsymbol{A}\ , \tag{3.53}$$

$$\boldsymbol{\alpha}^T\boldsymbol{\Psi}^{-1}+\boldsymbol{\zeta}^T\boldsymbol{\Phi}^{-1}=\boldsymbol{B}\ , \tag{3.54}$$

$$\boldsymbol{\alpha}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}+\boldsymbol{\zeta}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\zeta}=\boldsymbol{C}\ . \tag{3.55}$$

Moreover,

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}-2\boldsymbol{B}\boldsymbol{X}+\boldsymbol{C}=(\boldsymbol{X}-\boldsymbol{A}^{-1}\boldsymbol{B})\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{A}^{-1}\boldsymbol{B})-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B}+\boldsymbol{C}\ . \tag{3.56}$$

□

Inside the exponential function, the L.H.S of Eq.(3.56) leads to a kernel that is proportional to the Gaussian kernel with mean $\boldsymbol{A}^{-1}\boldsymbol{B}$ and covariance $\boldsymbol{A}^{-1}$. The terms $\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B}$ and $\boldsymbol{C}$ do not contain the variable $\boldsymbol{X}$ so they give rise to multiplicative constants and hence the proportionality. To be specific to the notation we used in Eq.(3.44), we consider two Gaussians quadratic in $\boldsymbol{\beta}$. One with mean $\boldsymbol{d}$ and covariance D, and the other with mean $\boldsymbol{X}^{-1}\boldsymbol{z}$ and covariance $\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X}$. With respect to Eq.(3.56), the quantities for these two Gaussians are as follows

$$\boldsymbol{A}=(\boldsymbol{D}^{-1}+\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X})^{-1}\ , \tag{3.57}$$

$$\boldsymbol{B}=\boldsymbol{D}^{-1}\boldsymbol{d}+(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X})\boldsymbol{X}^{-1}\boldsymbol{z}=\boldsymbol{D}^{-1}\boldsymbol{d}+\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{z}=\boldsymbol{D}^{-1}\boldsymbol{d}+\boldsymbol{X}^T\boldsymbol{\kappa}\ . \tag{3.58}$$

The second equation is transformed first due to cancellation and then due to the fact that $\boldsymbol{z} = \left(\frac{\kappa_1}{\omega_1}, \frac{\kappa_2}{\omega_2}, \cdots\cdots, \frac{\kappa_n}{\omega_n}\right)$ and $\boldsymbol{\Omega} = diag(\omega_1, \omega_2, \cdots, \omega_N)$. Now we get back to the second equation in the Gibbs sampling couple Eq.(3.45) and write down the second conditional probability as

$$p(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{B_\omega}, \boldsymbol{A_\omega}) \tag{3.59}$$

where

$$\boldsymbol{A_\omega} = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \boldsymbol{\Omega} \boldsymbol{X})^{-1} \tag{3.60}$$

$$\boldsymbol{B_\omega} = \boldsymbol{A_\omega}(\boldsymbol{X}^T \boldsymbol{\kappa} + \boldsymbol{D}^{-1} \boldsymbol{d}) \tag{3.61}$$

If $\boldsymbol{\Omega}$ can be sampled, then the above two equations give a conditionally Gaussian likelihood centered at $\boldsymbol{X}^{-1}\boldsymbol{z}$ and with covariance $\boldsymbol{X}^T \boldsymbol{\Omega} \boldsymbol{X}$.

3.5.4 Gibbs Sampling for a logistic regression using Põlya-Gamma latent variables

As an illustration of the Põlya-Gamma augmentation, we solve a logistic regression problem through a Gibbs sampling algorithm based on latent variable augmentation of the posterior and by extension the likelihood function. The task is to classify points from two different normal distributions. The algorithm is as follows

1. Sample 1000 points each from two normal distributions. The ones used for this example are $\mathcal{N}(0,1)$ and $\mathcal{N}(5,1.5)$. These are

2. Distribute both Gaussians according to a Binomial with parameter p. The value used here is $p = 0.5$. This follows from the concept of sampling distributions based on the

central limit theorem.

3. Initialize the Gibbs sampler with the first step of calculating the $\mathbf{\Omega}$ diagonal matrix. This is done through a draw from the Põlya-Gamma distribution $p(\omega_n|\boldsymbol{\beta}) \sim PG(1,\boldsymbol{\beta}^T x_n)$.

4. Next, use the Gaussian distribution $p(\boldsymbol{\beta}|\mathbf{\Omega},\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{B_\omega},\boldsymbol{A_\omega})$ according to Eq.(3.60) and use these to obtain $\mathbf{\Omega}$ to apply step 3 again.

5. Repeat 3 and 4 in an iterative manner to achieve the classification of points from both the Gaussians.

We demonstrate the application of above steps to construct a Gibbs sampler for classifying data within a binary Gaussian mixture. The two distributions sampled for 1000 points i.e. the ground truth are shown in Fig.(3.1). The Põlya-Gamma augmented gibbs sampler classifies points as per logistic regression for 1000 iterations. The resulting labelled data is shown in Fig.(3.2). The incorrectly classified points lie mainly along the intersection of both the distributions.

3.5.5 Moments of the Põlya-Gamma distribution.

We now use the results obtained previously to prove three key facts about the Põlya-Gamma distribution. The first one is the formula for expected value and the second is the resulting corollary that Põlya-Gamma distribution is an infinite convolution of Gamma random variables.

Lemma 3.2. *The expectation value of* $PG(\omega|b.c)$ *is given by* $\mathbb{E}[\omega] = \frac{b}{c}\tanh\frac{c}{2}$.

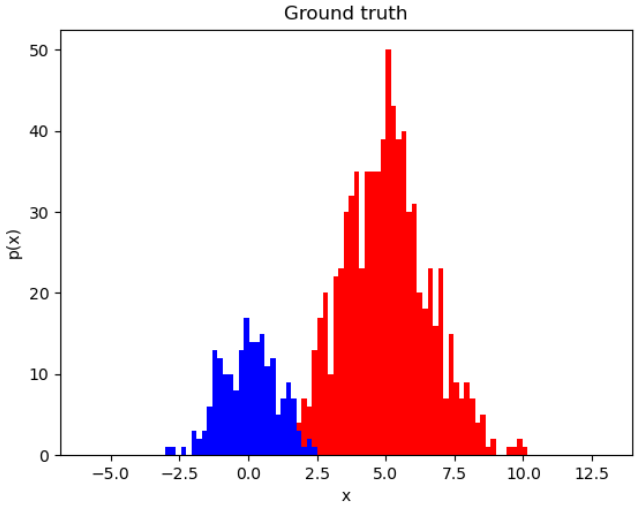


Figure 3.1: The two normal distribution histograms shown in blue $\{\mathcal{N}(0,1)\}$ and red $\{\mathcal{N}(5,1.5)\}$.

Proof. Rewriting Eq.(3.29) and Eq.(3.30) we get

$$\mathbb{E}\left[e^{-\omega t}\right] = \prod_{i=1}^{t}\left(1+\frac{t}{2\pi^2\left(k-\frac{1}{2}\right)^2}\right)^{-b} = \frac{1}{\cosh^b\left(\sqrt{\frac{t}{2}}\right)} \ .$$

This gives

$$\mathbb{E}\left[e^{-\omega t}\right] = \frac{1}{\cosh^b\left(\sqrt{\frac{t}{2}}\right)} \ . \tag{3.62}$$

Next we recall that

$$PG(\omega|b,c) = \frac{e^{-\frac{c^2}{2}\omega} PG(\omega|b,0)}{\mathbb{E}[e^{-(c^2\omega)/2}]} \ . \tag{3.63}$$

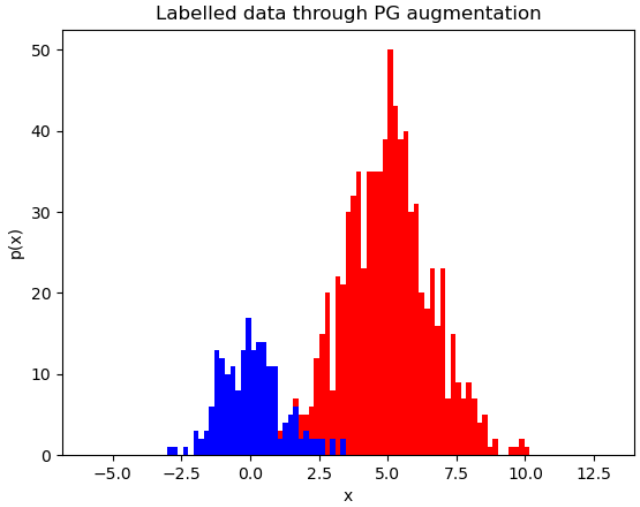


Figure 3.2: The classification of the all points according to their original sampling distributions using a Gibbs sampler constructed using Põlya-Gamma latent variables. The augmented sampler was iterated for 1000 cycles i.e. steps 3 and 4 of the algorithm. The incorrectly classified points are mainly in the region of intersection of both the distributions i.e. around $\mu \pm 3\sigma$ and further.

From Eq.(3.62) and Eq.(3.64), we get that

$$PG(\omega|b,c) = \frac{e^{-\frac{c^2}{2}\omega} PG(\omega|b,0)}{\cosh^{-b}\left(\frac{c}{2}\right)} = e^{-\frac{c^2}{2}\omega} PG(\omega|b,0) \times \cosh^{b}\left(\frac{c}{2}\right) \quad . \tag{3.64}$$

Taking Laplace transform again

$$\mathbb{E}\left[e^{\omega t}\right] = \cosh^{-b}\left(\sqrt{\frac{\frac{c^2}{2}+t}{2}}\right) \times \cosh^{b}\left(\frac{c}{2}\right) \quad . \tag{3.65}$$

This is a moment generating function and differentiating with respect to t would give us

the moments of ω with respect to $PG(\omega|b,c)$ as

$$\frac{d\mathbb{E}[e^{\omega t}]}{dt} = \frac{b}{2c}\sinh^{-b}\left(\sqrt{\frac{\frac{c^2}{2}-t}{2}}\right)\times\cosh^{b}\left(\frac{c}{2}\right) \quad . \tag{3.66}$$

Taking $t = 0$

$$\mathbb{E}[\omega] = \frac{b}{2c}\tanh\left(\frac{c}{2}\right) \quad . \tag{3.67}$$

□

Next we prove that Põlya-Gamma distributions are infinite convolutions of Gamma distributions.

Corollary 3.1. *The Põlya-Gamma distributions are infinite convolutions of Gamma random variables.*

Proof. Rewriting Eq.(3.65) we have

$$\mathbb{E}\left[e^{\omega t}\right] = \frac{\cosh^{b}\left(\frac{c}{2}\right)}{\cosh^{b}\left(\sqrt{\frac{\frac{c^2}{2}+t}{2}}\right)} \quad . \tag{3.68}$$

Applying the Weierstrass factorization theorem

$$\mathbb{E}\left[e^{\omega t}\right] = \frac{\cosh^b\left(\frac{c}{2}\right)}{\cosh^b\left(\sqrt{\frac{\frac{c^2}{2}+t}{2}}\right)} = \prod_{k=1}^{\infty}\left(\frac{1+\frac{\frac{c^2}{2}}{2(k-\frac{1}{2})^2\pi^2}}{1+\frac{\frac{c^2}{2}+t}{2(k-\frac{1}{2})^2\pi^2}}\right) = \prod_{k=1}^{\infty}\left(1+d_k^{-1}t\right)^{-b} \quad . \quad (3.69)$$

where $d_k = 2\left(k-\frac{1}{2}\right)^2\pi^2 + \frac{c^2}{2}$. Each term in this product is the Laplace transform of a Gamma distribution. This gives $PG(\omega|b,c)$ as an infinite convolution

$$\omega \sim \sum_{k=1}^{\infty}\frac{\Gamma(b,1)}{d_k} = \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{\Gamma(b,1)}{2\left(k-\frac{1}{2}\right)^2\pi^2+\frac{c^2}{2}} \quad . \quad (3.70)$$

□

Next, we derive the variance of the Põlya-Gamma distribution. For that we already have the Moment Generating function from Eq.(3.69). We state the result as follows.

Corollary 3.2. $\sigma^2\left[PG(\omega|b,c)\right] = \frac{b}{4c^3}\frac{(\sinh c - c)}{\cosh^2\frac{c}{2}}$.

Proof. We rewrite Eq.(3.69)

$$\mathbb{E}\left[e^{\omega t}\right] = \frac{\cosh^b\left(\frac{c}{2}\right)}{\cosh^b\left(\sqrt{\frac{\frac{c^2}{2}+t}{2}}\right)} = \prod_{k=1}^{\infty}\left(\frac{1+\frac{\frac{c^2}{2}}{2(k-\frac{1}{2})^2\pi^2}}{1+\frac{\frac{c^2}{2}+t}{2(k-\frac{1}{2})^2\pi^2}}\right) = \prod_{k=1}^{\infty}\left(1+d_k^{-1}t\right)^{-b} \quad . \quad (3.71)$$

taking logarithm yields the cumulant generating function.

$$\Phi(t) = \sum_{k=1}^{\infty} -b \log\left(1 + \frac{t}{d_k}\right) \quad . \tag{3.72}$$

The Taylor series of $\Phi(t)$ around $t = 0$ can be written as

$$\Phi(t) = -\mu_1' t + \frac{1}{2!}\left(\mu_2' - (\mu_1')^2\right) t^2 + \cdots \tag{3.73}$$

where μ_k' represents the raw moment of order k as is standard notation on cumulant generating functions. In Eq.(3.73), the coefficient of t is the mean and that of t^2 is the variance. Now we expand the Taylor series as

$$\Phi(t) = \sum_{k=1}^{\infty} -b\left(\frac{t}{d_k} + \frac{t^2}{2d_k} + \cdots\right) = -b\sum_{k=1}^{\infty}\frac{t}{d_k} - b\sum_{k=1}^{\infty}\frac{t^2}{2d_k^2} + \cdots \tag{3.74}$$

Applying the Weierstrass Factorization Theorem yields

$$\mu_1' = \frac{b}{2c}\tanh\frac{c}{2} \quad . \tag{3.75}$$

$$\left(\mu_2' - (\mu_1')^2\right) = \frac{b}{4c^3}\frac{(\sinh c - c)}{\cosh^2\frac{c}{2}} \quad . \tag{3.76}$$

□

The contour plots of both the moments are shown in Fig.(3.3) and Fig.(3.4). It is to be noted that the functions are not defined at $c = 0$. Limiting values i.e. $c \longrightarrow 0$ are possible though and give the estimates for the mean and variance values.

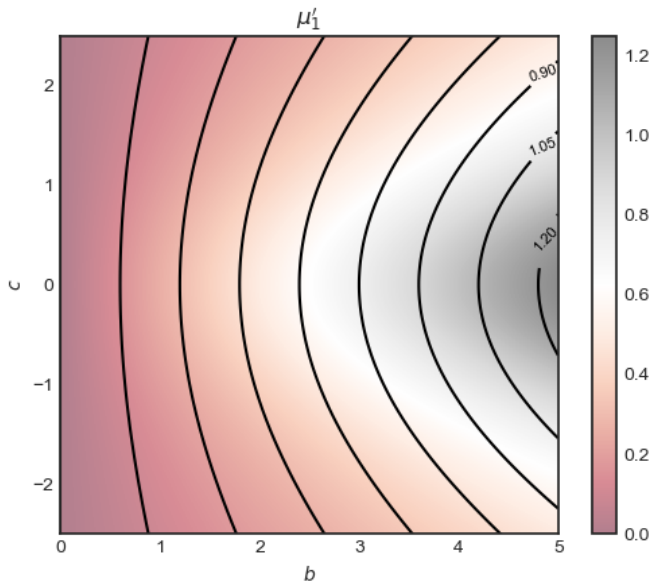


Figure 3.3: Contour plot of $\mathbb{E}[PG(\omega|b,c)]$ with b and c as the variables. It can be seen how the exponential tilting changes the value of the mean.

3.5.6 The latent variable form of the Likelihood function

To write down the augmented likelihood function in terms of the latent variables ρ and ω from the Poisson and Põlya-Gamma distributions, we need to divide the augmentation into two parts, i.e., flip and no-flip times. The flip times only need the hyperbolic cosine augmentation through the Põlya-Gamma variables while the no-flip times require both. This is where the efficiency gain of the computation lies as the no-flip part is the one with the bulk of complexity given that it has the exponential for the integration as well as the hyperbolic cosine which is a standard feature of the local field. By Eq.(3.30) we have

$$\int_0^\infty e^{-2\omega x^2} PG(\omega|b,0)d\omega = \cosh^{-b}(x) \quad . \tag{3.77}$$

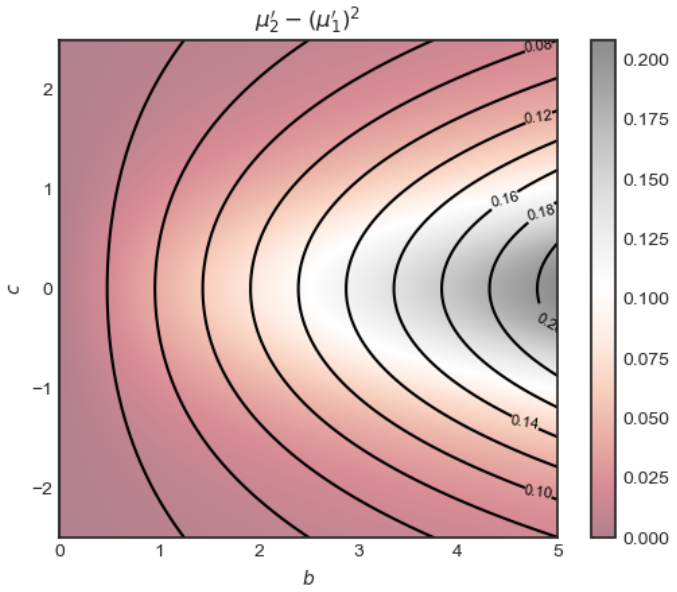


Figure 3.4: Contour plot of $\sigma^2[PG(\omega|b,c)]$ with the shape factors as the contour variables.

This is the expectation value that we would assign to the hyperbolic cosines in the flip and the no-flip parts of the likelihood function. In general, we write

$$P(\boldsymbol{\sigma}_N([0,t])|\boldsymbol{J},\boldsymbol{\theta}) = \sum_{\rho}\int \mathscr{L}(\{s,\rho,\omega\}([0,t])|\boldsymbol{J},\boldsymbol{\theta})d\omega \ , \tag{3.78}$$

and the likelihood function $\mathscr{L}(\{s,\rho,\omega\}([0,t])|\boldsymbol{J},\boldsymbol{\theta})$ is the augmented form of Eq.(3.8) written as

$$\mathscr{L}(\{s,\rho,\omega\}([0,t])|\boldsymbol{J},\boldsymbol{\theta}) = \prod_{(i,t)\in F} e^{\left[-\sigma_i H_i(t)-\omega_i(H_i(t))^2\right]PG(\omega_i|1,0)} \tag{3.79}$$
$$\times\prod_{i,n} e^{\left[\rho_i^n(\sigma_i^n H_i^n-\log 2)-2(H_i^n)^2\omega_i^n\right]}\times e^{-\gamma(t_{n+1}-t_n)}\frac{\gamma(t_{n+1}-t_n)^{\rho_i^n}}{\rho_i^n!}\times PG(\omega_i^n|\rho_i^n,0) \ .$$

The exponents of the exponential functions in this product are at most quadratic and consequently this is a much more tractable form of the complete data likelihood. Due to the presence of latent variables it is favorable to apply an optimization program that can ac-

commodate the latent variable augmented form of the likelihood. One such procedure is the Expectation-Maximization Algorithm and we describe it in the next section, marking the transition from augmentation to inference in solving the inverse problem.

3.6 The Expectation-Maximization Algorithm

The expectation-maximization algorithm is a numerical method for obtaining the maximum likelihood estimate (MLE) of model parameters in presence of latent variables. In its original form, the algorithm is meant to obtain MLE for missing variables. It works just as well if we treat the variables we are not supposed to observe, i.e., latent variables, as missing variables. The algorithm also works for other problems that can be posed as missing data problems such as mixture estimation. This section explores the concepts behind the workings of the EM-algorithm and its applications.

3.6.1 Convex functions

Let f be a real-valued function defined an an interval $I[a,b]$. f is said to be convex on I if $\forall x_1, x_2 \in I$ and a real number $\lambda \in [0, I]$.

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad , \tag{3.80}$$

Note that the equation of the secant between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is $y - f(x_1) = \frac{f(x_2)-f(x_1)}{x_2-x_1}(x-x_1)$. This means the definition of convexity is that the value of function evaluated in the interval I should not exceed the y value of the secant line between the end-points of I. In addition, the function $-f$ is strictly concave if f is strictly convex.

Theorem 3.5. *If $f(x)$ is twice differentiable on $[a,b]$ and $f''(x)$ on $[a,b]$, then $f(x)$ is convex on $[a,b]$.*

Proof. For

$$x \leq y \in [a,b], \;\; \lambda \in [0,1] \;\; ,$$

let

$$z = \lambda y + (1-\lambda) x \;\; .$$

By definition, $f(x)$ is convex iff

$$f(\lambda y + (1-\lambda) x \leq \lambda f(y) + (1-\lambda) f(x)$$

Note that

$$f(z) = \lambda f(z) + (1-\lambda) f(z) \leq \lambda f(y) + (1-\lambda) f(x) \tag{3.81}$$

Taking λ and $1-\lambda$ terms to the same sides of the inequality the definition of convexity can be written as

$$\lambda [f(z) - f(y)] \leq (1-\lambda) [f(x) - f(z)] \tag{3.82}$$

The Lagrange's mean value theorem states that $\exists s \in [x,z]$ such that $\frac{f(z)-f(x)}{z-x} = f'(s)$ Similarly, $\exists t \in [z,y]$, such that $\frac{f(y)-f(z)}{y-z} = f'(t)$. Since,

$$f''(x) \geq 0, \;\; \forall x \in [a,b] \quad \text{and} \quad s \leq t \;\; ,$$

therefore $f'(s) \leq f'(t)$ Next, we write

$$z = \lambda y + (1-\lambda) x \tag{3.83}$$

which gives

$$\begin{aligned} z - \lambda z + \lambda z = \lambda y + (1-\lambda) x \ , \\ (1-\lambda)(z-x) = \lambda (y-z) \end{aligned} \tag{3.84}$$

$$\begin{aligned} (1-\lambda) \frac{f(z) - f(x)}{z-x} &= f'(s)(1-\lambda) \\ (1-\lambda)(f(z) - f(x)) &= f'(s)(1-\lambda)(z-x) \\ (1-\lambda)(f(z) - f(x)) \leq f(t)(1-\lambda)(z-x) &= f'(t)\lambda(y-z) \\ (1-\lambda)(f(z) - f(x)) &\leq \lambda (f(y) - f(z)) \\ (1-\lambda)[f(x) - f(z)] &\geq \lambda [f(z) - f(y)] \end{aligned} \tag{3.85}$$

which is the condition for convexity. This completes the proof. □

Corollary 3.3. $-\log x$ *is strictly convex on* $(0, \infty)$

Proof. If $f(x) = \log(x)$ then $f''(x) = \frac{1}{x^2} > 0, \forall x \in (0, \infty)$. From Theorem 3.5 $f(x)$ is strictly convex. By extension $\log x$ is concave on $(0, \infty)$. □

Theorem 3.6 (Jensen's Inequality)**.** *Let $f(x)$ be a convex function defined on I. If $x_1, x_2, x_3 .... x_n \in I$ with $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_n > 0$ and $\sum_{i=1}^{n} \lambda_i = 1$ then*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i \right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

Proof. $n = 1$ is a triviality. $n = 2$ is the definition of convexity as it is proven above. For a general n, we need to use induction. To that end, assume the theorem is true for some n. Then

$$\begin{aligned}
f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) &= f\left(\lambda_{n+1}x_{n+1} + \sum_{i=1}^{n}\lambda_i x_i\right) \\
&= f\left(\lambda_{n+1}x_{n+1} + (1-\lambda_{n+1})\frac{1}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right) \\
&\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\frac{1}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right) \\
&= \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\sum_{i=1}^{n}\frac{\lambda_i x_i}{(1-\lambda_{n+1})}\right) \\
&\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1})\sum_{i=1}^{n}\frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) \\
&= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^{n}\lambda_i f(x_i) = \sum_{i=1}^{n+1}\lambda_i f(x_i) \quad . \qquad (3.86)
\end{aligned}$$

This proves the inequality. □

Now, we can use the fact that $\log x$ is a concave function and write the following relation which would be handy later in our main derivation. We use the fact that $\log x$ is a concave function and write

$$\log\left(\sum_{i=1}^{n}\lambda_i x_i\right) \geq \sum_{i=1}^{n}\lambda_i \log(x_i) \qquad (3.87)$$

giving us a lower-bound on the logarithm of a summation. We are now in a position to prove the key results for the EM-algorithm.

3.6.2 Derivation of the EM-algorithm

Let a parameterized family with the parameter set $\boldsymbol{\theta}$ produce a vector $\boldsymbol{X}$. The task we have at hand is to estimate $\boldsymbol{\theta}$ such that $P(\boldsymbol{X}|\theta)$ is maximum, i.e., the MLE for $\boldsymbol{\theta}$. We define the log-likelihood function

$$L(\boldsymbol{\theta}) = \log\left(P(\boldsymbol{X}|\boldsymbol{\theta})\right) \quad . \tag{3.88}$$

The probability is defined as conditioned on $\boldsymbol{\theta}$ but the likelihood function is considered to be a function of theta given $\boldsymbol{X}$. This means that the estimation of parameters is dependent of the available data. In addition, our aim is to maximize L in an iterative manner, i.e., if we estimate $L(\boldsymbol{\theta}_n)$, then the update from the algorithm $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_n)$. This is equivalent to maximizing the difference

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_n) = \log P(\boldsymbol{X}|\boldsymbol{\theta}) - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \quad . \tag{3.89}$$

Thus far in this standard MLE procedure, we have not considered the possibility of variables we do not directly observe but the EM procedure includes these latent variables quite naturally. We begin this second leg of latent variable inclusion by considering variable vector $\boldsymbol{Z}$ and single realization of the same vector by $\boldsymbol{z}$. The actual likelihood estimate is now written as a total probability with latent variables conditioned on the observed variables as

$$P(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta})\, P(\boldsymbol{z}|\boldsymbol{\theta}) \quad . \tag{3.90}$$

We rewrite the difference relation

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_n) = \log \sum_{\boldsymbol{z}} P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta})\, P(\boldsymbol{z}|\boldsymbol{\theta}) - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \quad . \tag{3.91}$$

The presence of the logarithm of a suggests the use of Jensen's inequality, namely the relation in Eq.(3.87). The next task is to identify the λ_is, i.e., the multiplicative constants that are always positive and sum to 1. To accomplish this we see that $P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta})$ are probability measures. This means that they are positive and sum to 1 for a particular $\boldsymbol{z}$. That is

$$\begin{aligned} \sum_{\boldsymbol{z}} P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) = 1 \ , \\ P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) \geq 0 \ . \end{aligned} \tag{3.92}$$

So we have

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_n) &= \log \sum_{\boldsymbol{z}} P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) P(\boldsymbol{z}|\boldsymbol{\theta}) - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \\ &= \log \sum_{\boldsymbol{z}} P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) P(\boldsymbol{z}|\boldsymbol{\theta}) \frac{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}_n)}{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}_n)} - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \\ &= \log \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) P(\boldsymbol{z}|\boldsymbol{\theta})}{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}_n)} P(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}_n) - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \\ &\geq \sum_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) P(\boldsymbol{z}|\boldsymbol{\theta})}{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}_n)} - \log P(\boldsymbol{X}|\boldsymbol{\theta}_n) \\ &= \sum_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}) P(\boldsymbol{z}|\boldsymbol{\theta})}{P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\theta}_n) P(\boldsymbol{X}|\boldsymbol{\theta}_n)} \stackrel{\Delta}{=} \Delta(\boldsymbol{\theta}|\boldsymbol{\theta_n}) \ . \end{aligned} \tag{3.93}$$

This gives

$$L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}_n) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta_n}) \ . \tag{3.94}$$

We now define a function which is equivalent to the right-hand side of eq. 3.94

$$l(\boldsymbol{\theta}|\boldsymbol{\theta_n}) = L(\boldsymbol{\theta}_n) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta_n}) \ , \tag{3.95}$$

giving

$$L(\boldsymbol{\theta}) \geq l(\boldsymbol{\theta}|\boldsymbol{\theta_n}) \ . \tag{3.96}$$

In addition, we observe the following

$$l(\boldsymbol{\theta_n}|\boldsymbol{\theta_n}) = L(\boldsymbol{\theta}_n) + \Delta(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) \ . \tag{3.97}$$

We now substitute for $\Delta(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n)$ using Eq.(3.93) and get the expression

$$l(\boldsymbol{\theta_n}|\boldsymbol{\theta_n}) = L(\boldsymbol{\theta}_n) + \sum_z P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X}|z,\boldsymbol{\theta}_n)\, P(z|\boldsymbol{\theta}_n)}{P(\boldsymbol{X}|z,\boldsymbol{\theta}_n)\, P(\boldsymbol{X}|\boldsymbol{\theta}_n)} \ . \tag{3.98}$$

Using the chain rule for conditional probability we get

$$= L(\boldsymbol{\theta}_n) + \sum_z P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X},z|\boldsymbol{\theta}_n)}{P(\boldsymbol{X},z|\boldsymbol{\theta}_n)} = L(\boldsymbol{\theta}_n) + \sum_z P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log 1 = L(\boldsymbol{\theta}_n) \ . \tag{3.99}$$

This result gives us the insight we need for our iteration process, which is that at $\boldsymbol{\theta} = \boldsymbol{\theta}_n$, we have , $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n) = L(\boldsymbol{\theta}_n)$. In addition, the upper bound on $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ is $L(\boldsymbol{\theta})$. This makes the goal of the algorithm to increment $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ as any increment in $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ is effectively an increment in the value of $L(\boldsymbol{\theta})$. This allows to consider maximizing the likelihood as an iterative process. We now derive the recursion relation for this process.

$$\theta_{n+1} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{\theta}_n) = \arg\max_{\boldsymbol{\theta}} \left[L(\boldsymbol{\theta}_n) + \sum_z P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X}|z,\boldsymbol{\theta})\, P(z|\boldsymbol{\theta})}{P(\boldsymbol{X}|z,\boldsymbol{\theta}_n)\, P(\boldsymbol{X}|\boldsymbol{\theta}_n)} \right] \tag{3.100}$$

The terms that do not contain $\boldsymbol{\theta}$ can be dropped as they are inconsequential in the maxi-

mization.

$$\theta_{n+1} = \arg\max_{\boldsymbol{\theta}} \left[\sum_{z} P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log \frac{P(\boldsymbol{X},z,\boldsymbol{\theta})\,P(z,\boldsymbol{\theta})}{P(z,\boldsymbol{\theta})P(\boldsymbol{\theta})} \right] \quad .$$

$$= \arg\max_{\boldsymbol{\theta}} \left[\sum_{z} P(z|\boldsymbol{X},\boldsymbol{\theta}_n) \log P(\boldsymbol{X},z|\boldsymbol{\theta}) \right] = \arg\max_{\boldsymbol{\theta}} \left[\mathbb{E}_{z|\boldsymbol{X},\boldsymbol{\theta}_n} \log P(\boldsymbol{X},z|\boldsymbol{\theta}) \right] \qquad (3.101)$$

Eq.(3.101) is the basis for the numerical procedure we are about to undertake. The expectation of the log-likelihood of the available data as a function of latent variables conditioned on the actual (overt) parameters with respect to the latent variable distribution conditioned on the data and actual variables. At this point we define the two repeating steps of the EM algorithm:

1. **E- step** : Calculate the expected value $\mathbb{E}_{z|\boldsymbol{X},\boldsymbol{\theta}_n} \log P(\boldsymbol{X},z|\boldsymbol{\theta})$.

2. **M-step** : Maximize the conditional expectation calculated in the E- step with respect to the actual parameters $\boldsymbol{\theta}$.

At this point we must establish the advantage of the EM algorithm versus direct maximization of the log-likelihood. The intended use we have for this procedure is to express the time-series likelihood in terms of Poisson and P$\tilde{o}$lya-Gamma variables and to then follow the above two steps repeatedly. We now go on to derive the corresponding numerical algorithm.

3.6.3 EM algorithm for the inverse Ising problem

We apply the concepts discussed in the previous section to construct a EM scheme for a fully connected Ising model. The idea is to evaluate the expectation of the likelihood with respect to the latent variables conditioned on the coupling matrix. This cost-function is then maximized within a tolerance value. We look at the expectation calculation parallel

to our derivation of the EM-algorithm as follows.

E-step

By definition, the expectation of the log-likelihood

$$l(\boldsymbol{J},\boldsymbol{\theta}|\boldsymbol{J}_m,\boldsymbol{\theta}_m) = \sum_{\rho}\int \mathscr{L}(\{\boldsymbol{s},\boldsymbol{\rho},\boldsymbol{\omega}\}([0,t])|\boldsymbol{J},\boldsymbol{\theta}) p(\boldsymbol{\rho},\boldsymbol{\omega}|\{\boldsymbol{s}\}([0,t]),\boldsymbol{J}_m,\boldsymbol{\theta}_m) d\omega \quad , \quad (3.102)$$

where the conditional distribution

$$p(\boldsymbol{\rho},\boldsymbol{\omega}|\{\boldsymbol{s}\}([0,t]),\{\boldsymbol{J}_m,\boldsymbol{\theta}_m\}) = p(\boldsymbol{\omega}|\{\boldsymbol{s},\boldsymbol{\rho}\}([0,t]),\{\boldsymbol{J}_m,\boldsymbol{\theta}_m\}) \times P(\boldsymbol{\rho}|\{\boldsymbol{s}\}([0,t]),\{\boldsymbol{J}_m,\boldsymbol{\theta}_m\}) \quad . \quad (3.103)$$

The first term on the right-hand side of eq. 3.103 can be written as

$$p(\boldsymbol{\omega}|\{\boldsymbol{s},\boldsymbol{\rho}\}([0,t]),\{\boldsymbol{J}_m,\boldsymbol{\theta}_m\}) = \prod_{i,t\in F} PG(\omega_i(t)|1,2H_i(t)) \times \prod_{i,n} PG(\omega_i^n|\rho_i^n,2H_i^n) \quad . \quad (3.104)$$

Here the first term is for the flipping times, while the second is result of expressing the hyperbolic cosines in non-flipping terms with the ρ_i^n being the Poisson variable for the i^{th} node being constant in its state for the n^{th} interval. Note that the second term of this product is the tilted Põlya-Gamma distribution we defined in eq. 3.35 with the variable now as ω_i^n. Therefore,

$$PG(\omega_i^n|b,c) = \frac{e^{-\frac{c^2}{2}\omega_i^n} PG(\omega_i^n|b,0)}{\mathbb{E}\left[e^{-(c^2\omega_i^n)/2}\right]} = \frac{e^{-\frac{c^2}{2}\omega_i^n} PG(\omega_i^n|b,0)}{\cosh^{-b}\left(\frac{c}{2}\right)} \quad . \quad (3.105)$$

This achieves the latent variable transformation for the hyperbolic cosine for the entire time-series making use of both the tilted and regular form of the Põlya-Gamma distribution.

The second term of Eq.(3.103) pertains to the Poisson variables, i.e., ρ_i^n conditioned on the data, coupling matrix and fields. This was written exactly as it is in the latent likelihood Eq.(3.79) giving

$$P(\boldsymbol{\rho}|\{\boldsymbol{s}\}([0,t]),\{\boldsymbol{J}_m,\boldsymbol{\theta}_m\}) = \tag{3.106}$$
$$\prod_{i,n} e^{\left[\rho_i^n(\sigma_i^n H_i^n - \log 2) - 2(H_i^n)^2\omega_i^n\right]} \times e^{-\gamma(t_{n+1}-t_n)}\frac{\gamma(t_{n+1}-t_n)^{\rho_i^n}}{\rho_i^n!} \times PG(\omega_i^n|\rho_i^n,0) \ .$$

The Põlya-Gamma variables appear again as a hyperbolic cosine is still there in the no-flip Glauber terms. In addition to the above, the following expectation values are needed:

1. Flipping time Põlya-Gamma expectation

$$\langle\omega_i(t)\rangle = \frac{1}{4H_i(t)}\tanh(H_i(t)) \ . \tag{3.107}$$

2. Non-flipping time Põlya-Gamma expectation

$$\langle\omega_i^n\rangle = \frac{\rho_i^n}{4H_i^n}\tanh(H_i^n) \tag{3.108}$$

3. Non-flipping time Poisson expectation

$$\langle\rho_i^n\rangle = \gamma(t_{n+1}-t_n)\frac{e^{\sigma_i^n H_i^n}}{2\cosh(H_i^n)} \tag{3.109}$$

These averages have been calculated by applying lemma 3.2 and the expectation formula for a Poisson variable. With the expectation integral, i.e., the cost-function set, we go the maximization procedure.

M-step

The central idea is to have the scheme

$$\boldsymbol{J}_{m+1}, \boldsymbol{\theta}_{m+1} = \underset{\{\boldsymbol{J}_m, \boldsymbol{\theta}_m\}}{\text{argmax}} \quad l(\boldsymbol{J}, \boldsymbol{\theta} | \boldsymbol{J}_m, \boldsymbol{\theta}_m) \quad . \tag{3.110}$$

At this point, we recall that the log-likelihood in Eq.(3.79) is quadratic in $\boldsymbol{J}, \boldsymbol{\theta}$. The maximization of this quadratic form would lead to a system of N linear equations to be solved simultaneously. This gives

$$\boldsymbol{A}_i \boldsymbol{J}_{i\cdot} = \boldsymbol{b}_{i\cdot} \quad , \tag{3.111}$$

where $\boldsymbol{J}_{i\cdot} = [\theta_i, J_{i1}, J_{i2}, \cdots J_{iN}]^T$. The coefficients are written as

$$A_{ijk} = 4 \left(\sum_{t \in F_i} \langle \omega_i(t) \rangle \sigma_k(t) \sigma_j(t) + \sum_n \langle \omega_i^n \rangle \sigma_k^n \sigma_j^n \right) \quad , \tag{3.112}$$

and

$$b_{ij} = - \left(\sum_{t \in F_i} \sigma_i(t) \sigma_j(t) + \sum_n \langle \rho_i^n \rangle \sigma_i^n \sigma_j^n \right) \quad , \tag{3.113}$$

where F_i is the set of all times where spin σ_i got flipped. This framework is the result of differentiation followed by separation of terms with and without the couplings and fields. As one can see, it naturally lends itself to a structure very similar to a multivariate Gaussian and also uses the Covariance matrix. We now solve the inverse Ising problem in the nonequilibrium steady state. The algorithm used for the calculations is shown in Algorithm 1. The collection of averages over the $E-$step as well as the maximization of the augmented likelihood are in a while loop conditioned on a tolerance value, which we set as

10^{-4} in this example.

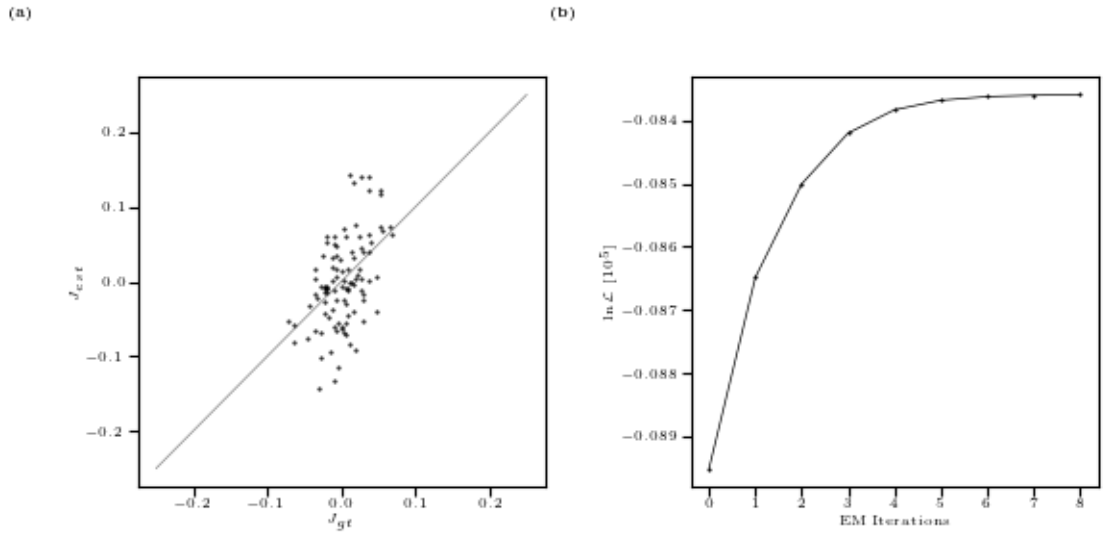


Figure 3.5: (a) The estimated coupling matrix J_{est} vs the ground truth for the bench-marking simulation J_{gt} for $N = 10, T = 1000$, $\gamma = 1$ and $\alpha = 0.1$. The scatter plot is juxtaposed against the line $y = x$. (b) The complete-data likelihood $(\times 10^{-5})$ evolution along the iterations of the EM algorithm. The likelihood saturates to a maximum value until step tolerance is achieved.

3.7 Calculations and Results

Our first calculation is for a network of 10 nodes with $\gamma = 1$ and α i.e. the *spread factor* of the Gaussian distribution that generates the coupling matrix being 0.1. We are looking for bench-marking calculations where we compare a ground truth we obtain from a simulation with our maximum-likelihood estimation. We also demonstrate the evolution of our augmented likelihood with each iteration. The size of the dataset is determined by the time for which our simulation runs to generate the microstates. We now categorize our discussion based on the variation of the parameters T, γ and α.

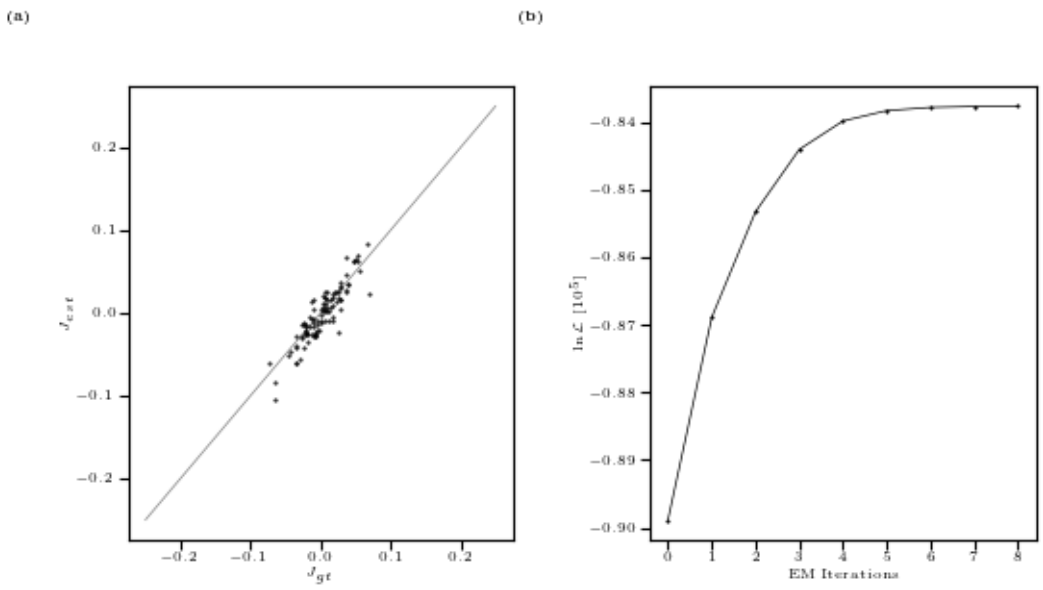


Figure 3.6: (a) The estimated coupling matrix J_{est} vs the ground truth for the bench-marking simulation J_{gt} for $N = 10, T = 10000$, $\gamma = 1$ and $\alpha = 0.1$. The scatter plot is juxtaposed against the line $y = x$. (b) The complete-data likelihood $(\times 10^{-5})$ evolution along the iterations of the EM algorithm. The likelihood saturates to a maximum value until step tolerance is achieved.

3.7.1 Dependence of reconstruction error on simulation parameters

The reconstruction error defined as

$$RE = \frac{||J_{est} - J_{gt}||_2}{||J_{gt}||_2} \quad , \tag{3.114}$$

is the ratio of the 2-norm of the difference between the estimated (J_{est}) and ground truth (J_{gt}) couplings to the 2-norm of the ground truth couplings. This ratio decreases as the estimated values of couplings get closer to the actual ground truth values which have been used for the monte carlo simulations to generate the data used as input. We now discuss each result.

Estimated vs. ground truth values of the coupling matrix

We show the calculations for $N = 10, \gamma = 1, \alpha = 0.1$ and very the values of T, i.e. the length of simulation time. This gives us different values for the length of our benchmarking data set. The first result for $T = 1000$ is shown in Fig.(3.5). The graph on the right is that of the likelihood function as the iterations for the numerical method are increased. The better the estimate, the lower the reconstruction error (RE) and the points are closer to the line $x = y$ for the graph of J_{est} vs J_{gt}. This can be seen through a comparison among Fig.(3.6), Fig.(3.7) and Fig.(3.8).

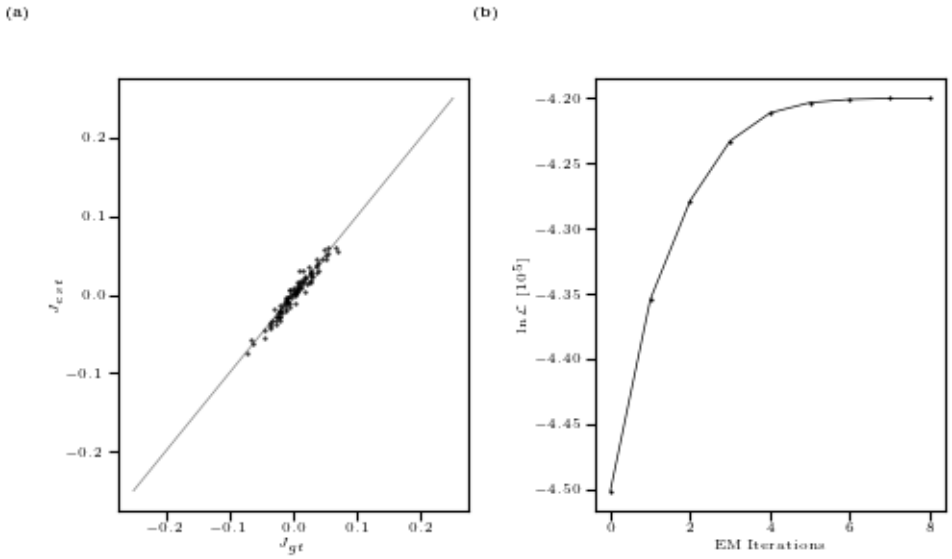


Figure 3.7: (a) The estimated coupling matrix J_{est} vs the ground truth for the bench-marking simulation J_{gt} for $N = 10, T = 50000$, $\gamma = 1$ and $\alpha = 0.1$. The scatter plot is juxtaposed against the line $y = x$. (b) The complete-data likelihood $(\times 10^{-5})$ evolution along the iterations of the EM algorithm. The likelihood saturates to a maximum value until step tolerance is achieved.

Variation of the reconstruction error with the length of input data

The value of *RE* decreases with an increase in the the number of data points, i.e, the number of microstate snapshots available. Fig.(3.9) shows how the numerical approximations get better as the length of the available data increases. In fact, for a network of size N, the asymmetric coupling matrix with N^2 independent elements has to be determined by a histogram of at least 2^N independent stochastic snapshots. Therefore for a 10 node network, a minimum of $2^{10} = 1024$ snapshots are required to account for all possible network states. The algorithm shows marked improvement as the number of available data points are much larger than that. It is to be expected as the histogram representing the steady state is more accurate as the number of data points increases.

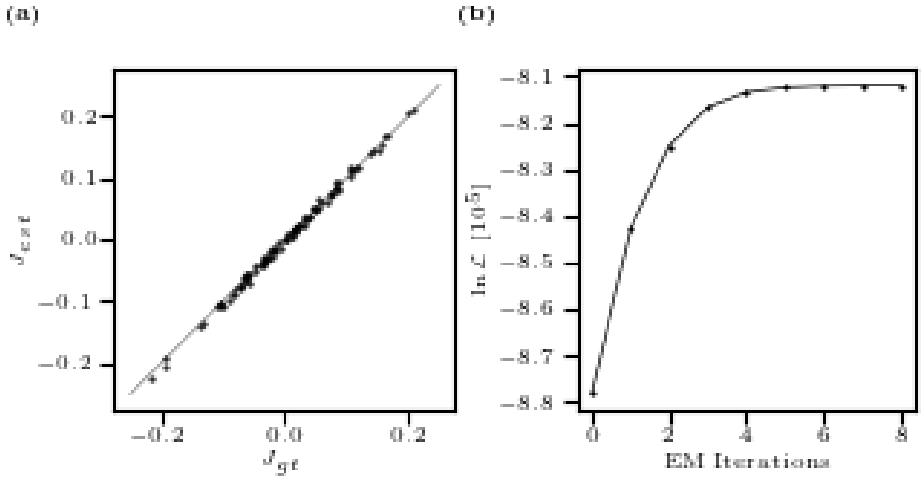


Figure 3.8: (a) The estimated coupling matrix J_{est} vs the ground truth for the bench-marking simulation J_{gt} for $N = 10, T = 100000$, $\gamma = 1$ and $\alpha = 0.1$. The scatter plot is juxtaposed against the line $y = x$. (b) The complete-data likelihood ($\times 10^{-5}$) evolution along the iterations of the EM algorithm. The likelihood saturates to a maximum value until step tolerance is achieved.

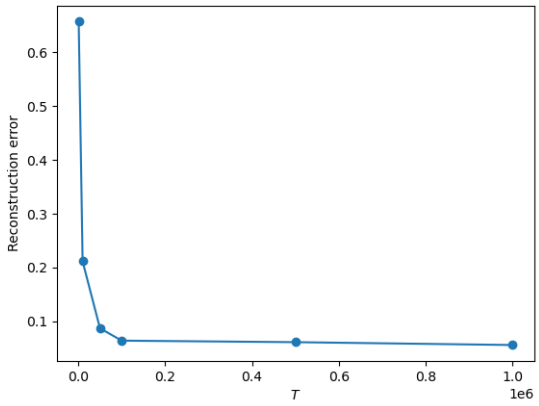


Figure 3.9: The reconstruction error as a function of the size of the data-set, expressed through the simulation time for the ground truth Glauber dynamics. For the reference set, $\alpha = 0.3$ and $\gamma = 1$

Variation of Reconstruction error with the standard deviation of the coupling matrix.

In our calculations, the input set is generated by running a monte carlo simulation with a ground truth coupling matrix J_{gt}, which is generated through a standard normal. We vary the standard deviation α of this normal distribution and calculate its effect on the reconstruction error for the case of $T = 100000$. As we see in Fig.(3.11), there is a marked decrease in the value of $R.E$ for higher values of α.

Variation in the reconstruction error with the probability rate γ

The probability rate is the same for each node in the network. Within this framework, it can be different for each node. The higher the probability rate, the more the probability of selection of a spin and hence it frequency of flipping over the period of observation. For a higher γ, there would be more flip terms as compared to non-flip terms. The resulting effect is very similar to having a larger data set. The graph in Fig.(3.10) shows the dependence of $R.E$ on γ.

Algorithm 1 Algorithm for MLE for a fully connected Ising model

Require: $N, T \in \mathbb{Z}_+$ and $\gamma, \alpha \in \mathbb{R}_+$

Ensure: $N \times \gamma \times T =$ Number of updates $>> N$

$J_{gt} \leftarrow [\mathcal{N}(0,1)]_{N\times N}$ ▷ This is the coupling matrix

$\theta_{gt} \leftarrow [\mathcal{N}(0,1)]_N^T$ ▷ These are the field(threshold) values at individual nodes.

Data generated $\leftarrow$ Glauber MCMC

Separate flipping and no-flip parts.

$\rho_i^n \leftarrow$ Poisson distributed no-flip time intervals for spin i

$\omega_i^n \leftarrow \cosh(\sigma_i^n H_i^n)$ PG variables for no-flip data

$\omega_i(t) \leftarrow \cosh(\sigma_i(t) H_i(t))$ PG variables for flip data

llk$\leftarrow -\infty$

$J_est = [0]_{N+1\times N}$ ▷ Initialize the fields an couplings as one matrix

while not converged **do**

oldllk=llk

$\rho_\tau, \omega\tau, H_\tau \leftarrow$E-step(flip data,$J_{est}, \gamma$)

$J_{est} \leftarrow$M-step(ρ_τ, ω_τ, flip intervals, flip nodes, Correlation matrix(flip data))

llk $\leftarrow$ log likelihood(flip data, J_{est}, flip intervals, flip node, γ)

converged=(llk-oldllk)/$(N \times T) \leq 10^{-4}$

end while

output$\leftarrow J_{est}$

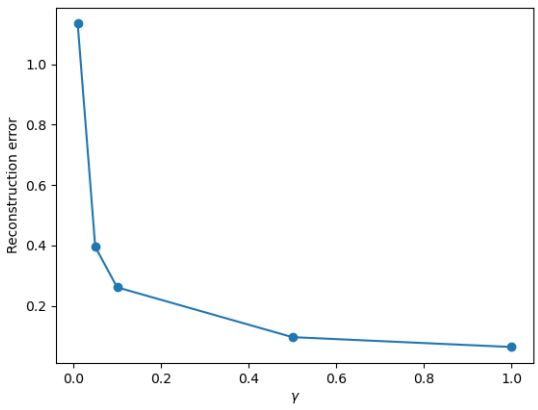


Figure 3.10: The reconstruction error as a function of the probability rate γ of the network Glauber dynamics. For the reference set, $\alpha = 0.3$ and $T = 100000$

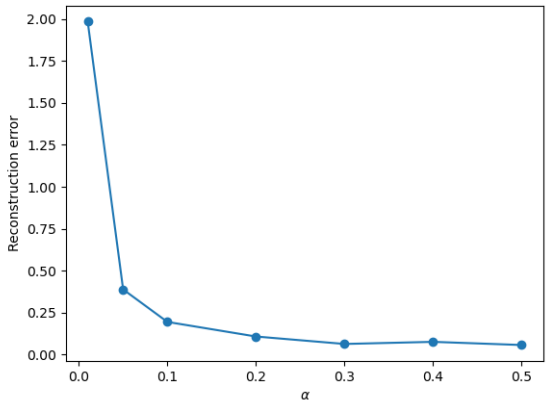


Figure 3.11: The reconstruction error as a function of the standard deviation of the Gaussian distribution i.e. α of the network coupling matrix. For the reference set, $\gamma = 1$ and $T = 100000$

3.8 Synopsis

We set out to demonstrate a latent variable based calculations for solving the inverse Ising problem for a time-series data. The latent variable approach was taken to reduce the computational complexity of the likelihood. The integral of exponential functions in the complete-data likelihood in Eq.(3.8) was approximated by using its structural similarity to the moment generating function of a Poisson distribution in Eq.(3.18) whereas P$\tilde{o}$lya-gamma random variables in Eq.(3.26) were used to transform the hyperbolic cosine terms in he denominator of the Glauber transition probabilities. The P$\tilde{o}$lya-gamma variables were introduced to reduce the computational complexity of the likelihood function for logistic regression. We demonstrate this solution method through a Gibbs sampling procedure for a binary Gaussian mixture. Although the main purpose of this solution is to explain the properties and applications of the distribution, it can be applied to complex networks for clustering and community detection problems. The time-series calculations are used to show the dependence of the reconstruction error *RE* on the parameters of the Ising model used to generate the data. It should be noted that in a real-world machine learning prob-

lem, these would be hyperparamters which would be tuned to minimize the difference in ensemble averages of order parameters between the data generated by J_{est} and the input.

Chapter 4: The inverse problem in absence of time-ordered data

4.1 Introduction

In the introduction to the inverse Ising problem, we described a certain class of the experimental and clinical data sets that do not lend themselves to a time-series format. The two main examples that were discussed were those of gene-regulatory network data recorded through single cell micro-arrays and neuronal spike trains that come from different sources and do not have a common origin point in time. Such data is also important for mapping the mechanisms and biophysical origins of several categories of human pathology such as cancer, diabetes, autoimmune disorders, Alzheimer's disease, etc. In addition, there is a lack of thorough understanding of some basic mechanisms of cellular dynamics and behavior under different environmental conditions. For example, the behavior of the brain on larger scales is still bringing new details with ever deeper insights possible due to improved experimental techniques. On the side of gene-regulation, there is still lack of clarity about epigenetic behavior of several non-coding constituents in special network states like colorectal cancer, triple negative breast carcinoma, etc. All this makes the problem of time-disorder a significant one in context of network inference. To develop the theory, we go back to the history of definition of a sufficient statistic. It is the insight given by this definition which leads us to formulate a likelihood which represents the empirical distribution through conditional probabilities as in the case for time-series data. This task is computationally intensive so we would still use latent variable augmentation for our

calculations.

4.2 Profile likelihood

The idea of profile likelihood is central to the development of the numerical method for the time-disordered inverse Ising problem. We first introduce the concept informally. Consider a set of n samples with the distribution $\hat{p} = (\hat{p}_1, \hat{p}_2, \cdots \hat{p}_n)$. Maximize the probability of observing the empirical distribution $\hat{p}$ up to a relabeling. $\kappa(\hat{p}) \stackrel{\Delta}{=} (\hat{p}_{\kappa(1)}, \hat{p}_{\kappa(2)}, \cdots \hat{p}_{\kappa(n)})$ where κ is some permutation. This means defining a likelihood of observing the same empirical distribution when the constituent elements have been rearranged in a form, which is different from the original one. This likelihood is the profile likelihood and its maximization, i.e., PML and its distribution p^* is written as

$$p^* = \underset{p}{\operatorname{argmax}} \sum_{\kappa} e^{-nKL(\kappa(\hat{p})||p)} \quad . \tag{4.1}$$

This is the maximization of the negative exponential to the Kullback-Liebler divergence so that the estimated distribution p is closest to each permutation of the elements and the sum of the exponential leads to the profile maximum likelihood distribution. The profile likelihood is a permutation likelihood and therefore if the brute force version of a problem is solved, then the complexity is $O(n!)$ which makes PML computations computationally prohibitive. As a result, approximations schemes are introduced based on a case by case basis. In the present case, we are interested in an irreducible and aperiodic Markov chain, i.e., Glauber dynamics and therefore, our approximation to the profile likelihood would be based on that. Before that, we need to formalize some theoretical aspects of PML in context of it being a sufficient statistic for the empirical distribution. As a first step, lets informally define a sufficient statistic.

Definition 4.1 (Sufficient statistic-I)**.** [47] A statistic is said to be sufficient with respect to a statistical model and its associated parameters if no other statistic that can be calculated from the same sample provides any additional information about the value of the model parameters.

To further develop this idea we need to be familiar with two concepts, the first one being the invariance of a statistic under labeling. Thereafter, we look at a recent notation concept in statistical physics known as a permutation glass [128]. The former is required to formulate the definition of the likelihood function we would develop and use while the latter will associate the likelihood function to a physical system through the statistical concept of quenched disorder. In the following sections, familiarity with the basic concepts of groups of finite order is assumed. In absence of the same, any of the references [5, 10, 18, 41] and [20] are a great source.

4.2.1 Permutation group

Definition 4.2 (Permutation group)**.** A permutation group is a finite group G whose elements are permutations of a given set and whose group operation is composition of permutations in G.

The order of a group being the number of elements in it, the order of a permutation group would always divide $n!$. This follows from Lagrange's theorem which we state as follows.

Theorem 4.1 (Lagrange)**.** *If H is a subgroup of G, then $|G| = [G:H].|H|$. Here $|G|$ denotes the order of G, while $[G:H]$ is the index of H in G, i.e., the number of left-cosets of H in G.*

Proof. Let $x, y \in G$ and call them equivalent if $\exists h \in H$ such that $x = yh$. This is because the left cosets of H in G are the equivalence classes of a certain equivalence relation on G. This

leads to the facts that the left cosets form a partition of G. Each left coset aH has the same cardinality as H because $x \mapsto ax$ defines a bijection $H \to aH$ (the inverse is $y \mapsto a^{-1}y$). The number of left cosets is $[G:H]$ and therefore $|G| = [G:H].|H|$. □

Definition 4.3 (Symmetric group)**.** The group of all permutations of a set M is called the symmetric group of M. It is denoted as $Sym(M)$

Being a subgroup of a symmetric group, all that is necessary for a set of permutations to satisfy the group axioms and be a permutation group is that it contain the identity permutation, the inverse permutation of each permutation it contains, and be closed under composition of its permutations. The intuition of the order of a permutation group dividing $n!$ comes from the fact that the number of left cosets of P in S_n i.e. $[P:S_n]$ would be a factor of the order of S_n, i.e., $n!$. It is clear from the above definitions that a permutation group is a subgroup of the symmetric group. If $M = \{1,2,3,4,\cdots n\}$ then $Sym(M)$ is called the symmetric group of n letters and is denoted by S_n.

Cauchy and cyclic notations

For $M = \{1,2,3,4,\cdots n\}$, the permutations are bijections of M and they can be represented by Cauchy's two-line notation. If σ is a permutation of M, then

$$\begin{pmatrix} 1 & \cdots & n \\ \sigma(1) & \cdots & \sigma(n) \end{pmatrix} . \tag{4.2}$$

(4.3)

For instance, a particular permutation of $\{1,2,3,4,5\}$ can be written as

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix} \quad . \tag{4.4}$$

(4.5)

This means $\sigma(1) = 2, \sigma(2) = 5, \sigma(3) = 4, \sigma(4) = 3$ and $\sigma(5) = 1$. Another way to write a permutation is in cyclic form. Consider the permutation

$$\begin{pmatrix} 3 & 2 & 5 & 1 & 4 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix} \quad . \tag{4.6}$$

This means $\sigma(1) = 2, \sigma(2) = 5, \sigma(3) = 4, \sigma(4) = 3$ and $\sigma(5) = 1$. The cyclic notation for this is $(1,2,5)(3,4)$. As it is evident, both forms are equivalent and represent the same fundamental transitions, i.e., permutations of the elements of a set. Moving forward, one can also denote a permutation or a set of permutation as a mapping. For instance a set $\boldsymbol{X}$ when acted upon by a permutation maps as a bijection say, $T(x) = t$. Having some sense of the permutation group, we now develop concept of a profile likelihood and see if it sufficient as a statistic.

4.3 Sufficient statistic: Kolmogorov and Hãjek definitions

Consider the statistical model

$$\mathscr{E} = \{\mathscr{X}, P_\theta, \theta \in \boldsymbol{\Theta}\} \quad . \tag{4.7}$$

The set $\mathscr{X}$ denotes the observations as $X \sim P_\theta$ for $\theta \in \mathbf{\Theta}$ and $X in \mathscr{X}$. Let $F(\theta) : \mathbf{\Theta} \mapsto \mathscr{Y}$ be a measurable function which induces a metric $d(F, \hat{F}) : \mathscr{Y} \times \mathscr{Y} \mapsto \mathbb{R}_+$. This metric d is a loss function. Lets consider a statistic T which we are assuming to be sufficient. If that is the case, then a PML approach would aim to maximize the probability of $T(x) = t$ occurring in the data rather than the probability of x itself as we did thus far in the MLE approach.

Definition 4.4 (PML estimator)**.** The profile maximum likelihood estimator of θ is defined as

$$\hat{\theta}_T(t) \stackrel{\Delta}{=} \underset{\theta}{\operatorname{argmax}}\ P_\theta(T(X) = t) \quad . \tag{4.8}$$

Next we tie this definition of the estimator into the PML by defining a performance estimate for the likelihood function through the measurable function F.

Theorem 4.2. *Consider the statistical model*

$$\mathscr{E} = \{\mathscr{X}, P_\theta, \theta \in \mathbf{\Theta}\} \quad . \tag{4.9}$$

Let $T = T(X)$ be a statistic such that $T : \mathscr{X} \mapsto \mathscr{T}$ and $|\mathscr{T}| < \infty$. Let $F(\theta) : \mathbf{\Theta} \mapsto \mathscr{Y}$ be a measurable function. Suppose there exists an estimator $\hat{F} : \mathscr{T} \mapsto \mathscr{Y}$ such that

$$\sup_{\theta \in \mathbf{\Theta}} P_\theta(d(F(\theta), \hat{F}(T)) > \varepsilon) < \delta \quad . \tag{4.10}$$

Then

$$\sup_{\theta \in \mathbf{\Theta}} P_\theta(d(F(\theta), F(\hat{\theta}_T)) > 2\varepsilon) < \delta |\mathscr{T}| \quad . \tag{4.11}$$

Theorem 4.2 puts into light an issue with the sufficiency of a parameter. In general, one would say that with the estimator of $\hat{F}(T)$ well defined and the cardinality $|\mathscr{T}|$ be small

enough, T as a statistic is sufficient. But what happens if there are nuisance parameters, i.e., any set of parameters which is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest. This is a very important question in statistical decision theory and was addressed by Kolmogorov as follows.

Definition 4.5 (Sufficiency:Kolmogorov)**.** [75] A statistic $T = T(X)$ is called sufficient for $F(\theta)$, if the posterior distribution of $F(\theta)$ given $X = x$, depends only on $T = t$ and on the prior distribution of θ.

There was an issue with this definition when $F(\theta)$ is not constant. If that is the case and T is sufficient for $F(\theta)$, then T is sufficient for θ as well. This makes the definition void and therefore it should be revised/corrected. This was done by H$\tilde{a}$jek as follows.

Definition 4.6 (Sufficiency:H$\tilde{a}$jek)**.** [56] A statistic $T = T(X)$ is called sufficient for $F(\theta)$ if there exists some other functional $R(\theta)$ such that $F(\theta) = F_1(R(\theta))$, and the following is satisfied:

1. The distribution of T will depend on $R(\theta)$ only, i.e. $P_\theta dT = P_{R(\theta)} dT$.

2. There exists a distribution $Q_R \in P_R$ such that T is sufficient for the family $\{Q_R\}$, where P_R is the convex hull of the distributions $\{P_\theta : R(\theta) = R\}$.

This does put forward the question about the determination of sufficiency of T with respect to $F(\theta)$. To state the standard method to define a sufficient statistic, we must state one concept regarding equivalence classes for a group.

Definition 4.7 (orbit)**.** [41] Let G be a group acting on a set X. The orbit of an element $x \in X$ is defined as $Orb(x) := \{y \in X : \exists g \in G : y = g * x\}$m where $*$ denotes the binary operation of the group, i.e., the group action. Thus the orbit of an element is all its possible destinations under the group action.

It can be seen that if there is a relation that the group action induces between x and its image y under the group action, then orbit, $Orb(x)$ is the equivalence class of x under the relation $x\mathscr{R}y$ induced by the group action. Coming back to the definition of a sufficient statistic, we let $G = g$ be a group of surjective transformations of the X space on itself. We define an event as G-invariant, if $gA = A$ for all $g \in G$. The set of G-invariant events is a sub$-\sigma-$algebra $\mathscr{B}$, and a measurable function f is $\mathscr{B}$-measurable if and only if $f(gx) = f(x)$ for all $g \in G$.

Definition 4.8. Consider two sets of transformations $g \in G$ acting upon X. Define a family of probability distributions with parameters τ given as P_τ such that

$$P_{\tau,g}(X \in A) = P_\tau(gX \in A) \ . \tag{4.12}$$

Letting $\theta = (\tau, g)$, the statistic corresponding to the sub-σ-algebra of G-invariant events is sufficient for τ in the sense of Def.(4.6). In other words, it is sufficient for τ to look at the set of equivalence classes X/G, while $Gx \in X/G$ denotes the equivalence class (orbit) of x, i.e.,

$$Gx \stackrel{\Delta}{=} \{gx : g \in G\} \ . \tag{4.13}$$

We now look at applying the above definition to a the case of a fully connected Ising model while the data of configurations is not available as a time series. For that we need to find a G where the following two tasks will be performed successfully.

1. A G-invariant likelihood functions is constructed. This means that this likelihood contains all the equivalence classes for each $x \in A$, i.e., the data under the group action $gx : g \in G$.

2. MLE for the likelihood function would give the correct coupling matrix and inciden-

tally infer the thermodynamic arrow of time.

4.4 Disjoint cycles

Recall from the cycle notation that a cycle of length s denoted $(a_1, a_2, \cdots a_s)$ where $a_1, a_2, \cdots, a_s \in 1, 2, ..., n$ are distinct elements is a permutation σ such that $\sigma(a_1) = a_2$, $\sigma(a_2) = a_3, \cdots \sigma(a_{s-1}) = a_s$, and $\sigma(a_s) = 1$ and where all other elements in the permutation are mapped to themselves. Now suppose that we have two cycles of elements from $1, 2, ..., n$. One way to compare these cycles is to see whether they "move" any elements in common. If two cycles do not move any elements in common, then we give the pair of cycles a special name which we define below.

Definition 4.9 (Disjoint cycles)**.** If $(a_1, a_2, \cdots a_s)$ and $(b_1, b_2, \cdots b_t)$ are cycles of $1, 2, ..., n$ then these cycles are said to be disjoint if $a_i \neq b_j$ for all $i \in 1, 2, ..., s$ and for all $j \in 1, 2, ..., t$.

We now prove a very important theorem regarding disjoint cycles which would help us with invariance property of our likelihood later.

Theorem 4.3. *[41][Disjoint cycles commute] If the pair of cycles $\alpha = (a_1, a_2, \cdots, a_m)$ and $\beta = (b_1, b_2, \cdots, b_n)$ have no entries in common, then $\alpha\beta = \beta\alpha$.*

Proof. Suppose α and β are permutations of $S = \{a_1, a_2, \cdots, a_m, c_1, c_2, \cdots c_l, b_1, b_2, \cdots, b_n\}$, where the c's are left fixed by the permutation cycles. If $x = a_i$ for some i and β fixes all the a elements, then

$$\alpha\beta(a_i) = \alpha(\beta(a_i)) = \alpha(a_i) = a_{i+1} \ .$$
$$\beta\alpha(a_i) = \beta(\alpha(a_i)) = \beta(a_{i+1}) = a_{i+1} \ .$$

So $\alpha\beta = \beta\alpha$ on the a elements. Similarly, if $y = b_i$ for some i and α fixes all the b elements,

then

$$\alpha\beta(b_i) = \alpha(\beta(b_i)) = \alpha(b_{i+1}) = b_{i+1} \ .$$
$$\beta\alpha(b_i) = \beta(\alpha(b_i)) = \beta(a_i) = a_{i+1} \ .$$

Finally both α and β fix the c elements so we have

$$\alpha\beta(c_i) = \alpha(\beta(c_i)) = \alpha(c_i) = c_i \ .$$
$$\beta\alpha(c_i) = \beta(\alpha(c_i)) = \beta(c_i) = c_i \ .$$

Thus $\alpha\beta(x) = \beta\alpha(x), \ \ \forall x \in S$ □

After proving the above result about disjoint cycles and combining it with the version of sufficient statistic, we would be able to construct the likelihood function for time-disordered data using the concept of a permutation glass [128].

4.5 Permutation-glass

Williams and Shakhnovich [127, 128] define the permutation glass as a model analogous to quenched disorder models such as spin glasses. The disorder here is vicinal, i.e., the zero energy ground state puts all states in their lowest energy or highest probability neighborhood. The permutations therefore get assigned disorder potentials based on the extent of vicinal disagreement compared to the ground state. Fig.(4.1) and Fig.(4.2) are the schematics in the original work used to illustrate the concept. The original work deals with the distributions for such systems at thermal equilibrium and as expected it ends up with the Boltzmann distribution at the steady state. The present work focuses on nonequilibrium steady states. This means that we will not have any standard characterization of

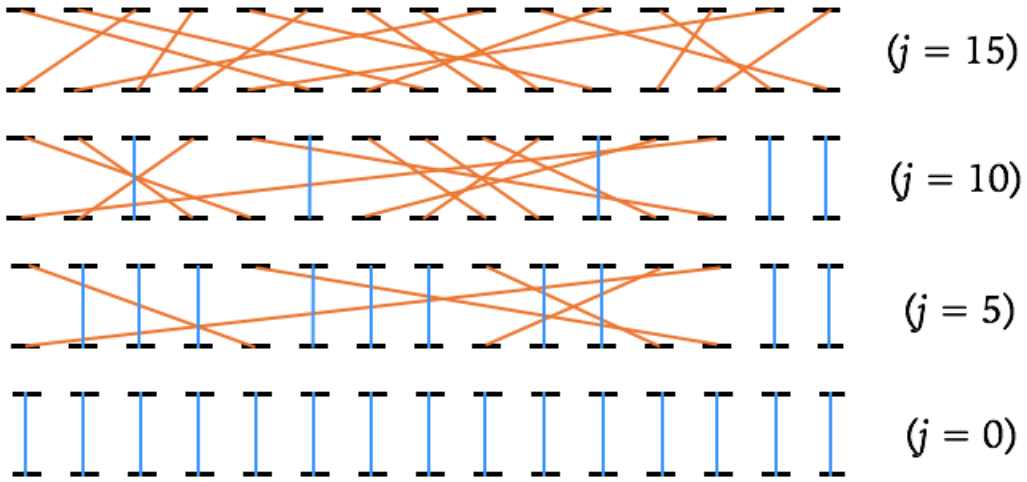


Figure 4.1: The permutation graph[128] depiction of four microstates in a permutation system with $N = 15$. In each graph, j is equivalent to the number of diagonal lines in the permutation graph. The number of "correct" connections are shown as vertical lines.

the the probability distribution like a CDF or MGF. All we have is a histogram of the microstates constituting the empirical distribution of macrostate. The basic statistical physics of the permutations of unique microstates postulates a disorder energy based on the ordering of states. Consider the set of N unique microstates $\{\omega_1, \omega_2, \cdots, \omega_N\}$ where we assign the zero disorder energy to the ordering $(\omega_1, \omega_2, \cdots, \omega_N)$. If a state ω_i is not in its **zero ordering position** θ_i, then the arrangement, i.e., vicinal order has the energy λ_i. This gives us a disorder Hamiltonian

$$\mathscr{H}(\{\theta_i\}) = \sum_{i=1}^{N} \lambda_i \mathbb{I}_{\theta_i \neq \omega_i} \quad , \tag{4.14}$$

where $\mathbb{I}$ is the indicator function. Now that the zero time-disorder energy is defined as the permutation where the $\mathbb{I}_{\theta_i \neq \omega_i} = 0, \forall i \in \{1, 2, \cdots N\}$, the scenarios where it has a finite value can be identified as the disorder permutations. In a way, these permutations derive their energy cost from a quenched disorder decided upon by the thermodynamic arrow of time. If they have vicinal distributions of microstates that are disordered as per the

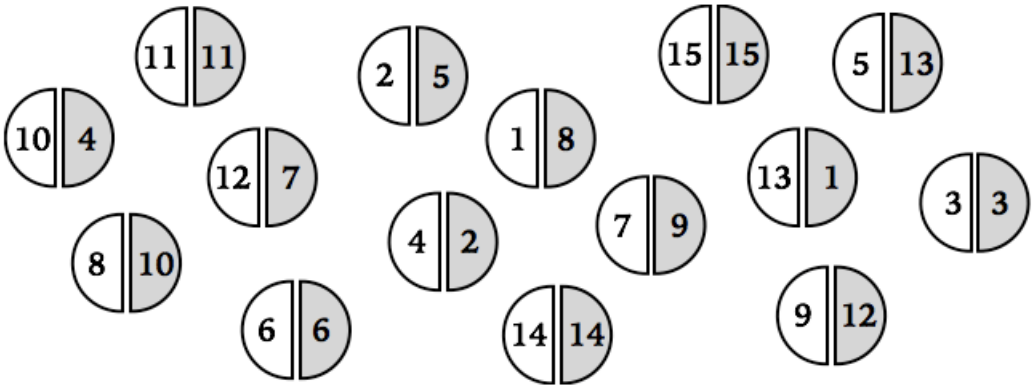


Figure 4.2: "Matching problem" depiction[128] of a $j = 10$ microstate for a $2N = 30$ permutation system. The spatial location of each pair is not important in determining the energy of the state. For this state, the matching pairs are $3, 6, 11, 14$, and 15.

arrow of time, then the quenched disorder of the dynamical system, defined for networks by the coupling matrix $(\boldsymbol{J}, \boldsymbol{\theta})$, is different from the arrangement with zero disorder energy. Thus, each steady-state distribution of microstates has a unique zero-disorder permutation corresponding to the thermodynamic arrow of time. It is along this arrow of time that the unique $(\boldsymbol{J}, \boldsymbol{\theta})$ would produce the zero-disorder permutation, if the forward problem was solved, e.g., an MC simulation or a set of experimental observations. Our goal is to perform a procedure over a minimal set of transitions that form the basis of the permutation glass and solve for the $(\boldsymbol{J}, \boldsymbol{\theta})$. Our calculation of the arrow of time is indirect in the sense that we derive the model parameters corresponding to the steady state distribution and the solution of the system, analytical or numerical, would generate the permutation corresponding to zero time-disorder energy, i.e., the arrow of time.

The first task from this point on is to have a geometric intuition for the minimal set over which the computation is to be performed.

4.5.1 The Dihedral group: D_n

The question asked here is the following.

What are the generators of a permutation glass?

The answer lies in the minimal set of disjoint cycles that form a basis for the entire permutation set. We define the group action as a permutation, i.e., one element moved to another. This action is done as incrementally additive, i.e., for a set of microstates $\{1,2,3\cdots M\}$, the cycles $(1,2,3\cdots)$ and $(1,3,5,\cdots)$ represent two disjoint sets of permutations as the elements are moved to distinct targets in both the sets.

Definition 4.10 (Rigid motion)**.** A rigid motion is a distance-preserving transformation, such as a rotation, a reflection, and a translation, and is also called an isometry.

Definition 4.11. For $n \geq 3$, the dihedral group D_n is defined as the rigid motions taking a regular n-gon back to itself, with the operation being composition.

The symmetries of a polygon becomes significant for equivalence of its configurations upon reflections. The lines of symmetry of the polygons $n = 3,4,5$ are shown in Fig.(4.3). These reflections lie in D_3, D_4 and D_5 for the triangle, square and pentagon respectively. In addition to reflections, a rotation by a multiple of $\frac{2\pi}{n}$ radians around the center carries the polygon back to itself, so D_n contains some rotations. These type of transitions are of interest for construction of likelihoods with respect to transition probabilities. Some properties of the dihedral group need to discussed before the construction of a permutation glass based likelihood function. The first step in that direction would be examine the number of elements in D_n, i.e., the total number of reflections and rotations. Afterwards, the relationship between rotations and reflections can be examined. The creation of a minimal set of transitions on a permutation glass comes out naturally from the concept of conjugacy classes.

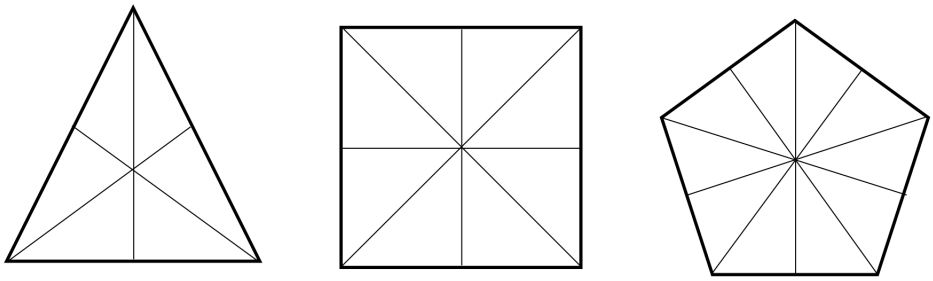


Figure 4.3: The symmetries of a triangle, square and polygon with respect to reflections. When the polygons are reflected with respect to any of their symmetry lines, the original configuration remains unchanged.

The order of D_n

Points in the plane at a specified distance from a given point form a circle, e.g., $x^2 + y^2 + ax + by + c = 0 \quad \forall a, b.c \in \mathbb{R}$, so points with specified distances from two given points are the intersection of two circles, which is two points (non-tangent circles) or one point (tangent circles). For instance, the red points in the figure below have the same distances to each of the two black points.

Lemma 4.1. *Each point on a regular polygon is determined by its distance from the adjacent vertices.*

Proof. In Fig.(4.4), let the red dots be adjacent vertices of a regular polygon. Therefore, the line segment connecting them is an edge of the polygon. Also, the polygon is entirely on one side of the line through the red dots. Consequently, the two black dots can't both be on the polygon, which means each point on the polygon is distinguished from all other

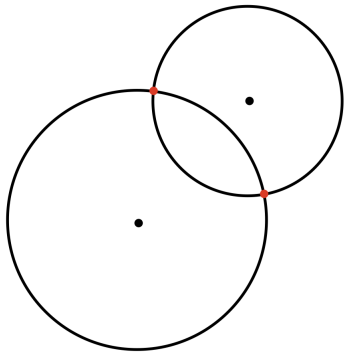


Figure 4.4: A non-tangent intersection of two circles.

points on the polygon by its distances from two adjacent vertices. □

Theorem 4.4. *The order of* $D_n = 2n$

Proof. First, we provide a visual proof. We separate the intuitive argument by odd and even vertex polygons. It can be seen that for both the cases, i.e., odd and even vertex regular $n-$gon, there are two symmetries

1. n symmetries by rotation of all vertices.

2. Further n symmetries by rotation on any of the axes of symmetry followed by reflection.

Both of these are demonstrated in Fig.(4.5) and Fig.(4.6). Firstly, we examine the number of fixed vertices with each type of rigid motion.

1. A non-identity rigid motion fixes no vertex.

2. An identity rigid motion fixes all vertices.

3. A reflection fixes three vertices.

For odd n, there is a reflection across the line connecting each vertex to the midpoint of the opposite side. This is a total of n reflections. They are different because each one fixes a different vertex. For even n, there is a reflection across the line connecting each pair of opposite vertices ($\frac{n}{2}$ reflections) and across the line connecting mid-points of opposite sides (another $\frac{n}{2}$ reflections). The number of these reflections is $\frac{n}{2}+\frac{n}{2}=n$. They are different because they have different types of fixed points on the polygon: different pairs of opposite vertices or different pairs of midpoints of opposite sides. This completes the

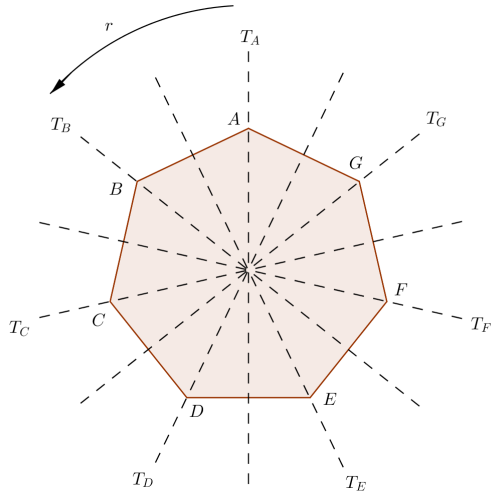


Figure 4.5: The symmetries of a heptagon. The corresponding diheral group would be D_7. In general for odd vertices we have $D_n = 2n+1$

visual proof. Another argument can be made on the basis of lemma 4.1. Consider A and B as two adjacent vertices of a regular n-gon. Let $g \in D_n$ be a rigid motion. The group action of $g \in D_n$ would have to preserve vertex adjacency and therefore $g(A)$ and $g(B)$ are adjacent vertices. See Fig.(4.7).

Having proved that $g(A)$ and $g(B)$ are adjacent vertices, one can see that for each choice for $g(A)$, there are two choices for $g(B)$, i.e., the vertex on either side of $g(A)$. Since there are a total of n vertices, therefore $|D_n| \leq 2n$. Combined with the result earlier from the visual

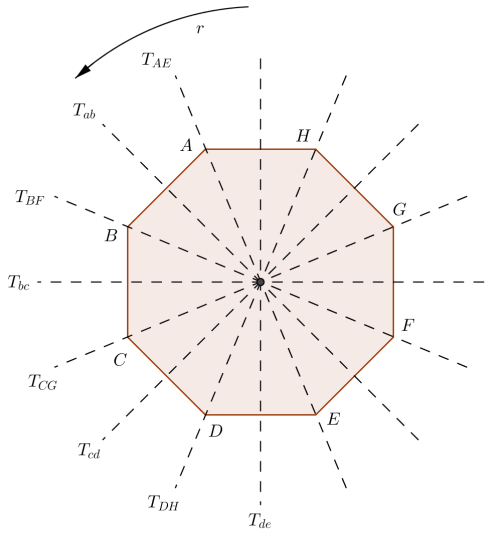


Figure 4.6: The symmetries of a octagon. The corresponding diheral group would be D_8. In general for even vertices we have $D_n = 2n$

proof, we have

$$|D_n| = 2n \ . \tag{4.15}$$

This completes the proof. □

Relationship between rotations and reflections

We denote r as a counterclockwise rotations of $\frac{2\pi}{n}$ radians, with the consideration that r is dependent on n. A schematic of symmetries of reflection of reflection for $n = 3,4,5$ is shown in Fig.(4.8). We now state and prove two results.

Theorem 4.5. *The rotations in D_n are $1, r, r^2 \cdots, r^{n-1}$.*

Proof. The only issue here is whether all the rotations are different from each other and since the number of rotations in D_n is n, therefore the rotations are different. □

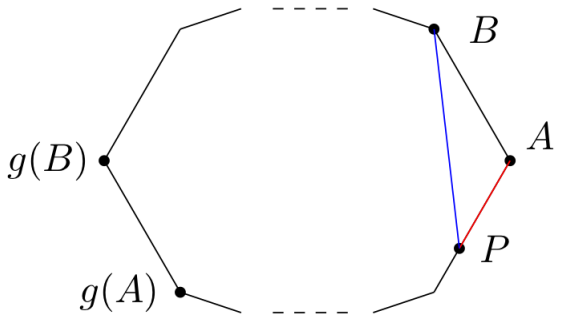


Figure 4.7: For each point P on the polygon, the location of $g(P)$ is determined by $g(A)$ and $g(B)$, because the distances of $g(P)$ from the adjacent vertices $g(A)$ and $g(B)$ equal the distances of P from A and B, and therefore $g(P)$ is determined on the polygon by lemma 4.1. To count $|D_n|$ it thus suffices to find the number of possibilities for $g(A)$ and $g(B)$.

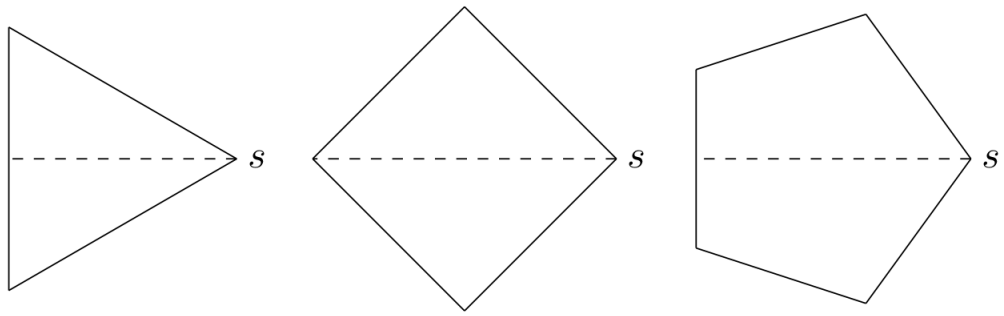


Figure 4.8: Axes of reflection symmetry for regular polygons $n = 3,4,5$. It can be seen that for each case the number of axes is n.

Next, we denote a reflection along a line passing through two opposite vertices as s and observe as earlier that s has order 2. This means $s^2 = 1$ and consequently $s^{-1} = s$.

Theorem 4.6. *The reflection in D_n are $s, rs, sr^2, \cdots, sr^{n-1}$.*

Proof. By theorem 4.5, all the reflections are different as each of them is just a reflection s multiplied by a rotation which are all different. To answer the question if any of the $r^k s$ is a rotation we observe that for $r^k s = r^l$ we have $r^{l-k} = s$ but s is not a rotation. Since out of the $2n$ elements n are rotations and none of the other n of the $r^k s$ elements are rotations, therefore they are all reflections. Also, it should be noted that there is no *rotation-reflection*

term. If an element is of the form $r^k s$ or sr^k, then by definition it is a reflection. Fig.(4.9) shows the geometric interpretation of such terms for $n = 3, 4, 5$. □

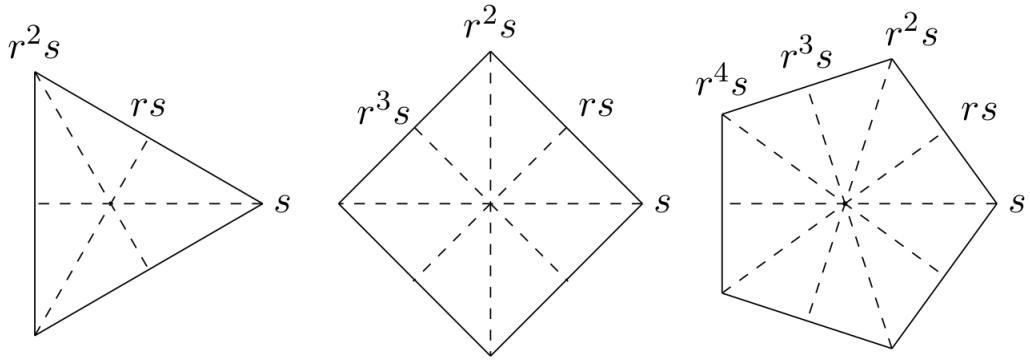


Figure 4.9: Lines of reflection for a regular n-gons for $n = 3, 4, 5$ with clockwise rotations around the polygon starting from a vertex fixed by s. If s is the reflection across the line through the rightmost vertex then rs is the next line of reflection counterclockwise. This repeats incrementally for n terms.

Theorem 4.7. $srs^{-1} = r^{-1}$

Proof. It is known that rs is reflection. See Fig.(4.9). This means

$$(rs)^2 = 1$$

$$rsrs = 1$$

$$srs = r^{-1}$$

but we know that s being a reflection has order 2, i.e., $s^2 = 1$ which gives $s = s^{-1}$. So finally, we have

$$srs^{-1} = r^{-1} \tag{4.16}$$

completing the proof. □

Conjugacy classes

Let x, y and z are subsets of a set A which consist of only blue, green and black balls. Now consider the relation $\sim$ defined as "has the same color as". Then $\sim$ is a an equivalence relation and the set of equivalence classes of this relation is given by the quotient space $A/\sim$ as the set all ball colors. In context of the topic at hand, we ask the following question.

What are the equivalence classes of the group actions of D_n?

The similarity or equivalence of geometric structures are called **conjugacy classes**. These are stated through the following theorem.

Theorem 4.8. *The conjugacy classes of D_n are the following.*

1. *If n is odd.*
   (a) *The identity element:* $\{1\}$.
   (b) $\frac{n-1}{2}$ *conjugacy classes of size*$= 2$*:* $\{r^{\pm 1}\}, \{r^{\pm 2}\}, \cdots, \{r^{\pm(n-1)/2}\}$
   (c) *All n reflections* $r^i s, \quad \forall i \in \{0, 1, 2, \cdots, (n-1)\}$.
2. *If n is even.*
   (a) *Two conjugacy classes of size 1:* $\{1\}, \{r^{(n/2)}\}$
   (b) $\frac{n}{2} - 1$ *conjugacy clases of size 2:* $\{r^{\pm 1}\}, \{r^{\pm 2}\}, \cdots, \{r^{\pm(n/2-1)}\}$.
   (c) *Two classes of reflections:* $\{r^{2is}\}$ *and* $\{r^{2i+1}s\}, \quad \forall 0 \leq i \leq \frac{n}{2} - 1$

Proof. The theorem says the following.

1. Except for the identity and $\{r^{n/2}\}$ for even n, every rotation is conjugate to its inverse.

2. Reflections are conjugate for odd n but split into two conjugacy classes for even n. These are the lines of reflection passing through opposing vertices($r^{even}s$) and the lines that pass through vertices and mid-points of the corresponding opposite sides($r^{odd}s$). See Fig.4.9

This means we have to prove these two facts about the conjugacy classes of D_n Every element of D_n is r^i or $r^i s$ for some integer i. Therefore to find the conjugacy class of an element g we will compute $r^i g r^{-i}$ and $(r^i s)g(r^i s)^{-1}$. Consider the equation

$$r^i r^j r^{-i} = r^j$$
$$(r^{-i}s)r^j(r^i s)^{-1} = r^{-j}$$

This shows that the conjugate of r^j is r^{-j}. The next task is to prove that this is the only conjugate. For that consider, $r^i s r^{-1} = r^{2i}s$ and its direct extension $(r^i s)s(r^{-1}s) = r^{2i}s$. As i varies, $r^{2i}s$ runs through the reflections in which r occurs with an exponent divisible by 2. If n is odd then every integer modulo n is a multiple of 2. This means that when n is odd

$$r^{2i}s = r^k s, \quad \forall i,k \in \mathbb{Z}$$

that is, every reflection in D_n is conjugate to s. The case is different for an even n as only half of the reflections are conjugate to s, the other half are conjugate to rs

$$r^i(rs)r^{-i} = r^{2i+1}s$$
$$(r^i s)(rs)(r^i s)^{-1} = r^{2i-1}s$$

This completes the proof □

Now, that we have an algebraic intuition of the concept of equivalence classes, we

proceed to establish the type of class based likelihood function to solve for the arrow of time from a permutation glass as per theorem 4.8.

4.6 Deriving the likelihood function for a permutation glass

Consider a set of microstates $\{1,2\cdots,M\}$ for a fully connected Ising graph with N nodes. The first task is to put the microstates as vertices of a regular M-gon. This makes the entire structure an $M \times N$ configuration representation. The rotations r^i represent the moving vertices from one position to the other. As we have shown in theorem 4.5, the reflections and rotations are related and therefore it suffices to consider only the latter for the purposes of the present discussion. The goal is to construct an invariant sufficient statistic with respect to the conjugacy classes of the M-gon. Earlier discussion about profile maximum likelihood, which is the same as a permutation glass likelihood, introduced the idea of a G-invariant sufficient statistics. The invariance with respect to conjugacy classes of the M-gon for a likelihood function would make it sufficient[1]

The procedure we follow is informed by the physics of the system. The transition matrix consists of the conditional probabilities derived from glauber dynamics and the unknown variables ,i.e., the couplings($\boldsymbol{J}$) and fields($\boldsymbol{\theta}$) are therefore contained in it. The numerical method should be able to derive $(\boldsymbol{J},\boldsymbol{\theta})$ from just the $M \times N$ time-disordered data matrix. That means the columns of the matrix which correspond to the network states are not ordered in time. We now construct this likelihood.

[1] In fact, such a function would be minimally sufficient because the likelihood is based on transition probabilities of an $M \times M$ Markov Chain(see Chapter-2). At a minimum for all such possible transitions, M^2 conditional probabilities exist. If a partition of such a chain is coarsened, then the sufficiency is lost. It is impossible to get the correct histogram of microstates, if all of their relative transition probabilities are not available.

4.6.1 Level sets of the Profile likelihood probability distribution

We define the level set of a probability distribution p as

$$X = \{x : p_x = u\} \tag{4.17}$$

The motivation for the method lies in what X contains for a dataset. The distribution would have the tendency to put a set of points x with same empirical count on the same X. In fact it does that within $O(\sqrt{M})$ for the complete set of M elements. This means that the microstates with the same empirical counts will be(approximately) in the same level set. If we can find a way to measure the weight of all microstates on the level sets of the distribution, then it would be possible to construct a likelihood based on those weights. The conjugacy classes of D_M are a way to do just that. The empirical distribution is our steady-state distribution and we would have the probability currents of the nonequilibrium steady-state lead to the relative counts of the microstates which would be different from the case where detailed balance would have those counts described by the Maxwell-Boltzmann distribution(see chapter 2).

The likelihood we aim to construct has to be based on transition probabilities as that is the only possible measure we have for time-disordered data. To connect that to the counts of micorstates we observe that

Lemma 4.2. *For a fully connected Ising graph with a coupling matrix $\boldsymbol{J}$ and fields $\boldsymbol{\theta}$, the permutation glass is completely generated by the conjugacy classes of D_M, where M is the number of available snapshots of the system.*

Proof. Let $\boldsymbol{P}$ be the Markov matrix(Glauber chain) of the Ising model. We place the snapshots i.e. available microstates of the chain on the vertices of a regular n-gon. The vicinal distribution is generated for each element $i \in M$ by the conditional probability $P_{ki} \forall k \in \boldsymbol{M}$.

The permutations are generated by the orbits of $k \in M$. The element P_{ki} will have the value $P_{k_p i} \forall k_p \in \boldsymbol{P}$ repeated as per the frequency of k in M. More precisely, the number of times an element is repeated in the permutation glass therefore is in the same proportion to its frequency in M. It can be seen that this representation is of the size M^2 as opposed to $M!$. The size of each orbit is M due to the orbit-stabilizer theorem [41]. The orbit of each element is of size M as the conjugacy classes move each vertex to every position on the polygon and there are M vertices. Recall that the identity is of size one and not two in this case as we do not differentiate between reflected configurations due to the fact that the vicinal distribution is invariant with respect to reflections. We call this $O(M^2)$ representation as the **orbit-reduced form** of the permutation glass. □

Another observation is the equivalence of this construction with the disjoint cycles of the permutation group.

Corollary 4.1. *The orbit-reduced form of a permutation-glass is the set of disjoint cycles of a permutation group of order M with respect to the labels of each element.*

Proof. Consider the element $x_k \in S_M$, $\forall k \in M$. As x_k is a configuration of the fully connected Ising graph, the orbit of x_k is the set of configurations that can transition to x_k by group action. Assuming that all transitions are of finite probability or all elements of the transition matrix are nonzero, we now construct the following set of cycles.
All cycles will have x_1 as the initial and by definition the final configuration. The rest of the elements of each cycle $c_j, \forall j \in M$ would be $\{x_k \in c_j : k = \frac{\mathbb{Z}}{j\mathbb{Z}}\}$. For instance, the cycle $c_r = (x_1, x_{1+r}, x_{1+2r}, \cdots)$. The dihedral group D_M is a subgroup of the symmetric group S_M and the cycles c_j are the minimal set of disjoint cycles with all the orbits of the Markov Chain □

An advantage of the above construction is that it holds even when some transition probabilities are zero. This is carried forward to any likelihood constructions based on the above

as the profile likelihood is based only on transition probabilities. We wish to have optimal inference and for that we would need the likelihood to be minimally sufficient. We discuss the basic concepts briefly.

4.6.2 Minimal Sufficiency

In definition 4.6, we introduced the concept of sufficiency and how a statistic on X can be defined as sufficient when it is invariant with respect to a group action Gx defined on a quotient space X/G of the equivalence classes. Sufficiency of a statistic ensures that it is providing the maximum possible information from the data available and no other statistic will be able to provide any additional information.
The concept is silent on the optimal nature of a statistic and for that we define minimal sufficiency as follows.

Definition 4.12 (Minimal sufficiency)**.** A sufficient statistic T is minimal if for every sufficient statistic T' and for every $x, y \in X, T(x) = T(y)$ whenever $T'(x) = T'(y)$. In other words, T is a function of T', i.e., $\exists f$ such that $T(x) = f(T'(x))\ \forall x \in X$.

The following theorem is a test for sufficiency

Theorem 4.9. *Let $\{p(x;\theta), \theta \in \Omega\}$ be a family of densities with respect to some measure μ. Note that μ is a Lebesgue measure for continuous distribution and a counting measure for discrete distribution. Suppose that there exists a statistic T such that for every $x, y \in X$:*

$$p(x;\theta) = C_{x,y}\, p(y;\theta) \Leftrightarrow T(x) = T(y) \quad \forall C_{x,y} \in \mathbb{R} \ . \tag{4.18}$$

Then T is minimally sufficient.

Proof. We begin by noticing that the theorem is not assuming that T is even sufficient and therefore that needs to be proven. We are going to use the *Neyman-Fischer factorization*

theorem to prove that and then proceed to the minimal sufficiency.

Let us define the range of T as $T(X) = \{t : t = T(x)\}$ for some $x \in X$. For each $t \in T(X)$, consider the preimage $A_t = \{x : T(x) = t\}$ and select an arbitrary representative x_t from each A_t. Then, for any $y \in X$ we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$. By the definition of A_t this implies that $T(y) = T(x_{T(y)})$. The theorem therefore implies

$$p(y;\theta) = C_{y,x_{T(y)}} p(x_{T(y)};\theta) = h(y) g_\theta(T(y)) \ , \tag{4.19}$$

giving us the sufficiency of T by the Neyman-Fischer factorization condition. Next we see that for any sufficient statistic T', by the factorization condition we have

$$p(x;\theta) = \tilde{h}(x)\tilde{g}_\theta(T'(x)) \ . \tag{4.20}$$

Suppose $T'(x) = T'(y)$ for any $x, y \in X$. Then, multiplying and dividing by $\tilde{h}(y)$ gives

$$p(x;\theta) = \tilde{h}(y)\tilde{g}_\theta(T'(y))\frac{\tilde{h}(x)}{\tilde{h}(y)} = p(y,\theta)C_{x,y} \ . \tag{4.21}$$

By the assumption of the theorem, this means $T(x) = T(y)$ and therefore

$$T'(x) = T'(y) \implies T(x) = T(y) \quad \forall x, y \in X \ , \tag{4.22}$$

for any sufficient statistic T'. This means T is minimally sufficient. □

Having the theoretical tools we have thus far, we should construct a likelihood function which can be proven to be minimally sufficient. To do so we must ensure the sufficiency using a set of equivalence classes. Next, we must use theorem 4.9 to make sure the computation would be optimal. It is of significance that we work with an optimal statistic as the task at hand is computationally intensive i.e. $O(M!)$ for the original permutation glass

and $O(M^2)$ for the orbit-reduced version. Even though the improvement is significant, M^2 is still computationally intensive and we have to make sure that nothing better is available before we proceed.

4.6.3 Orbit-reduced Ising likelihood function

The results thus far would facilitate the construction a minimal sufficient statistic i.e the orbit-reduced likelihood function. We use lemma 4.2 to establish the sufficiency with respect to the H$\tilde{a}$jek definition, i.e., definitions 4.6 and 4.8, to actually base the construction on a quotient space rather than the entire permutation set. This gives us sufficiency as a property of the construction. For minimal sufficiency, the relationship between an orbit-reduced and permutation likelihood would need to be that of proportionality as in theorem 4.9. With these two tasks accomplished, we would then make an information theoretic connection to approach the likelihood function as a relationship between the steady-state distribution and an arbitrary one. The procedure is as follows.

Theorem 4.10. *Let M be the index set of stochastic snapshots of a fully-connected asymmetric Ising model with its cardinality $|M|$. The parameters of the model are the coupling matrix $\boldsymbol{J}$ and fields $\boldsymbol{\theta}$. Let the Glauber transition Matrix be $\boldsymbol{P}$ with snapshot labels $i, j \in \{1, 2, \cdots, |M|\}$ indexing the transition probabilities p_{ij}. Then the log-likelihood function:*

$$\mathscr{L}(M; \boldsymbol{J}, \boldsymbol{\theta}) = \frac{1}{|M|} \log \left(\frac{1}{|M|} \prod_{j=1}^{|M|} p_{c_j} \right) \ , \tag{4.23}$$

where p_{c_j} is the set of sequential transition probabilities defined in the sense of corollary 4.1, is a minimal sufficient statistic.

Proof. 1. $\mathscr{L}$ is sufficient:

By the definition of p_{ij} as per corollary 4.1 is a statistic defined by equivalence classes of the permutation group as the sequence of microstates $x_i \forall i \in \{1,2,\cdots,|M|\}$ is $\prod_{r=1}^{|M|} c_r$ where $c_r = (x_1, x_{1+r}, x_{1+2r}, \cdots)$. These are all the disjoint cycles of a the permutation group of order $|M|$. From a geometric standpoint, these are the $|M|^2$ elements of the rotational conjugacy classes. Therefore, by definition 4.8 $\mathscr{L}$ is sufficient.

2. $\mathscr{L}$ is minimally sufficient:

   Let $\mathscr{P}$ be the probability distribution associated with $\mathscr{L}$ and define $\mathscr{P}'$ the distribution for another sufficient likelihood function $\mathscr{L}'$. By definition,

$$\mathscr{P} = \prod_{i \in NF} p_i(c_{NF} \in c_j) \times \prod_{i \in F} p_i(c_F \in c_j) \quad \forall c_{NF} \bigcup c_F = \bigcup_j c_j \ , \tag{4.24}$$

   where we distribute the probabilities within the conjugacy classes into two categories. The probabilities that correspond to a spin flip(F) and the ones that correspond to no flipping(NF) of a spin, leaving two labels with the same snapshot. Also, $\mathscr{L}'$ is sufficient. Therefore it has to contain all the transition probabilities corresponding to $\bigcup_j c_j$. So $\mathscr{L}'$ can contain more terms than $\mathscr{L}$ but not less. So we have

$$\mathscr{P}' = \prod_{i \in NF} p'_i(c_{NF} \in c'_j) \times \prod_{i \in F} p'_i(c_F \in c'_j) \quad \forall c_{NF} \bigcup c_F = \bigcup_j c'_j \ , \tag{4.25}$$

   where

$$\bigcup_j c_j \subset \bigcup_j c'_j \ . \tag{4.26}$$

From Eq.(4.24), Eq.(4.25) and Eq.(4.26) we get

$$\mathscr{P}'(\boldsymbol{J},\boldsymbol{\theta}) = \alpha(\boldsymbol{J},\boldsymbol{\theta})\mathscr{P}(\boldsymbol{J},\boldsymbol{\theta}) \ , \tag{4.27}$$

where $\alpha(\boldsymbol{J},\boldsymbol{\theta})$ contains all the terms that are in $\mathscr{P}'$ but not in $\mathscr{P}$. If we can prove that α is not a function of $\boldsymbol{J},\boldsymbol{\theta}$ for a particular steady-state distribution, then we have $\mathscr{L}'$ as a minimal sufficient statistic. But this is true as both $\mathscr{P}$ and $\mathscr{P}'$ are generated by the same steady state distribution. This is because the same nonequilibrium steady-state would generate the same proportion of frequencies for two sufficiently large time intervals and therefore lead to the same relative frequencies of microstates. Therefore,

$$\mathscr{P}'(\boldsymbol{J},\boldsymbol{\theta}) = \alpha(c,c')\mathscr{P}(\boldsymbol{J},\boldsymbol{\theta}) \ , \tag{4.28}$$

where $\alpha(c,c')$ indicates the dependence of α on the relative sizes of the two distributions. Note that $0 < \alpha(c,c') < 1$ as the proportionality constant itself is product of Glauber transition probabilities. From theorem 4.9 we have $\mathscr{L} = \mathscr{L}'$.

□

Now that we have derived our likelihood function and proved that it is minimally sufficient, we apply it to compute the coupling matrix for a set of time-disordered snapshots of a neural network. We use the fact that $\mathscr{L}'$ is defined on a quotient space to use parallel computing and accelerate our convergence.

4.7 Inference using the orbit-reduced likelihood function

The computation on N^2 glauber terms with an exponential in the numerator and hyperbolic cosine in the denominator is intensive. We propose an efficient algorithm for the

same using the fact that the conjugacy classes contain all orbits. This makes it possible to compute element wise in parallel. The algorithm is as follows

Algorithm 2 Algorithm for MLE for a fully connected Ising model with time-disordered data

Require: $N, T \in \mathbb{Z}_+$ and $\gamma, \alpha \in \mathbb{R}_+$ as hyperparameters
Ensure: $N \times \gamma \times T =$ Number of updates $= j >> N$
$J_{gt} \leftarrow [\mathscr{N}(0,1)]_{N\times N}$ ▷ This is the coupling matrix
$\theta_{gt} \leftarrow [\mathscr{N}(0,1)]_N^T$ ▷ These are the field(threshold) values at individual nodes.
Data generated $\leftarrow$ Glauber MCMC
for $i \in \{1, 2\cdots, j\}$ calculate $p(j \longrightarrow i)$ for each i in parallel
Calculate $\mathscr{L} = \frac{1}{j}\sum_{i=1}^{j} \frac{1}{j} \log(p(j \longrightarrow i))$
$J_est = [0]_{N+1\times N}$ ▷ Initialize the fields an couplings as one matrix
while not converged **do**
oldllk=$\mathscr{L}$
oldllk$\xrightarrow{\boldsymbol{J},\boldsymbol{\theta}}$ $\mathscr{L}$(BFGS optimization step)
converged=$(\mathscr{L} - oldllk) \leq 10^{-4}$
end while
output $\leftarrow J_{est}$

The addition of different orbit probabilities in parallel makes the calculation of the likelihood function efficient. The optimization with respect to the couplings $\boldsymbol{J}$ and fields $\boldsymbol{\theta}$ is done using the Broyden–Fletcher–Goldfarb–Shanno algorithm available as a part of the optimize routine in the package SciPy. We now present the results.

4.7.1 Numerical results

The first calculation is that of a neural network of size $N = 10$. The number of available snapshots is 10^7 and the the reconstruction error $\alpha = 0.65738$. The plot of J_{gt} i.e. the ground truth coupling matrix vs. the estimated values J_{est} is shown in Fig.(4.10).

Next we investigate the dependence of the reconstruction error α on the number of available snapshots ,i.e., the size of available data. This is significant as it gives an estimate of the length of data required for a desired accuracy level. It is natural that the amount of

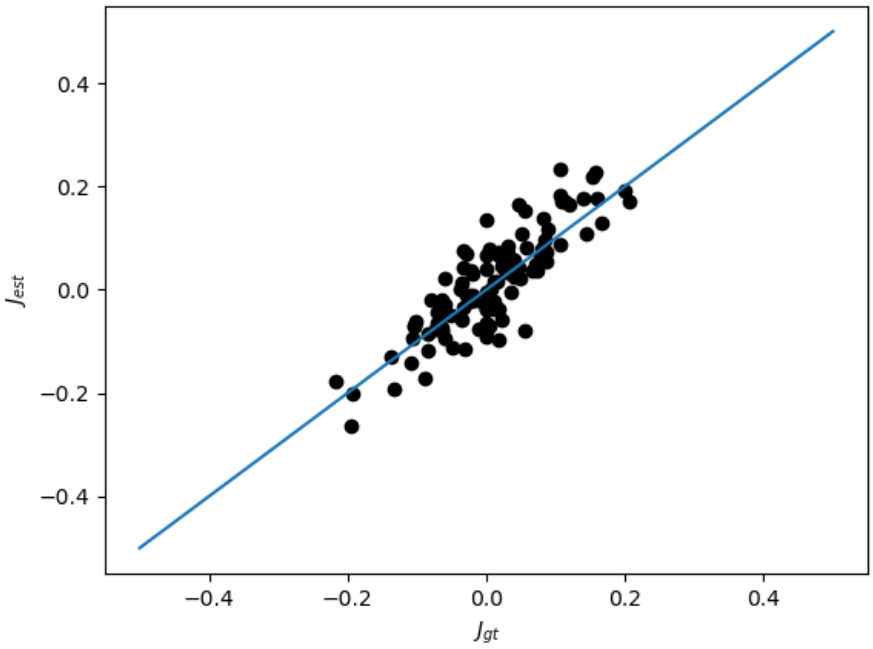


Figure 4.10: A plot of the estimated coupling matrix, J_{est} (y-axis) vs. the ground truth matrix J_{gt}. The $BFGS$ algorithm gave an estimated value of 0.65738 for 10 million snapshots.

data required for this algorithm is large as compared to time-series computations as each permutation contained in a permutation glass is a combinatorial bijection of a time-series data of the same size. This means the we are computing over a minimum number of such bijections through a minimal sufficient statistic.

4.8 Discussion

We began with the Kolmogorov and Hãjek definitions of sufficiency and to explain the development of the latter, discussed the concept of equivalence classes. The permutation group was used as an example and the result on the commutativity of disjoint cycles was stated and proved. For a geometric perspective, the dihedral group and its conjugacy classes were discussed. The quotient space X/G denoting a partition of X was used the construct

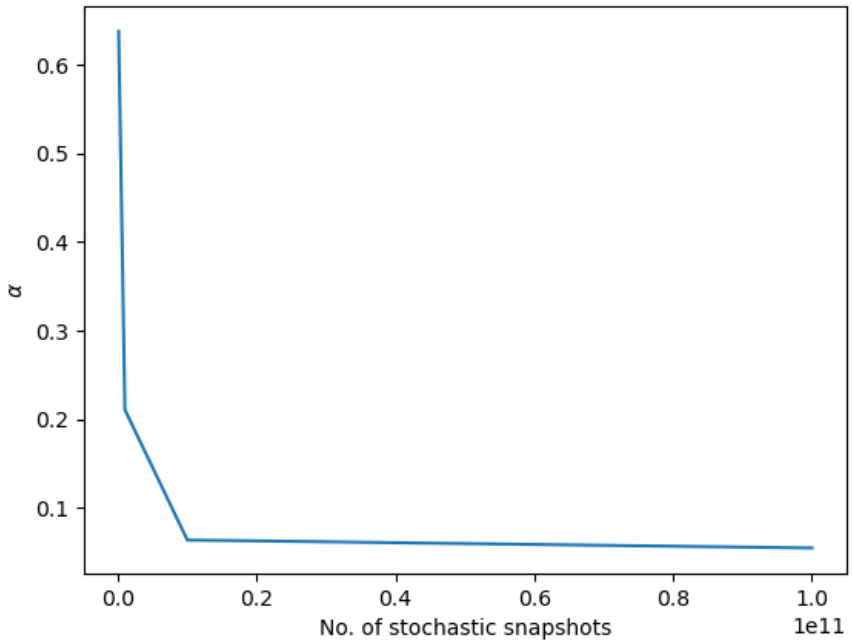


Figure 4.11: A plot of the reconstruction error α vs. the number of stochastic snapshots. The amount of data required for the orbit-reduced likelihood is larger as compared to a MLE for time-series computation.

the proof of lemma 4.2. Having developed the tools required, we proved that a likelihood constructed on the quotient space is minimal sufficient(Theorem 4.10). Finally, a parallel algorithm was used to calculate the likelihood function and the results were presented.

Chapter 5: The Gromov–Wasserstein distance between networks

5.1 Introduction

The idea of distance between two entities is central to our understanding of the difference between them. In a physical setting like classical mechanics, the displacement of an object is a measure of difference between the initial and final coordinates of its center of mass. In an abstract setting, for instance two strings or character sequences of equal length, the Hamming distance is the number of points where those two entities differ. For example, the Hamming distance between *Varenya* and *Vaarush* is 5. In statistics, the entities between which distances are sought are probability distributions. The most common of these distance measures is the relative entropy or Kullback–Leibler divergence [32] from a probability distribution Q to another P, which is

$$D(P||Q) = \sum_{x \in \mathscr{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \quad ,$$

for $x \in \mathscr{X}$ where $\mathscr{X}$ is a probability space. For networks, the above notions of distance do not suffice as the definition of what could constitute the distance between two networks is not apparent given that one cannot endow a network with a natural distance like displacement or an abstract one like the KL-divergence. Moreover, biological network dynamics includes growth and shrinkage as a functional feature during significant processes like cancer, morphogenesis and ageing. This puts an additional requirement on the metric to be

defined between two networks of different sizes, leading to steady state distributions of different dimensions. Optimal driving between nonequilibrium steady state has seen extensive developments over the last two decades but they have been mainly concentrated on probability distributions with the same support [80, 101, 138]. We follow [116] to write an essential theoretical framework for biological network dynamics. This would add to the existing network growth models [92, 129] while being grounded in the order parameters of the fully connected asymmetric Ising model [83, 90]. The connection between thermodynamics and optimal transport is well described in literature [6, 11, 36, 66, 88] and computational works also exist for the same [74]. Our approach contrasts these in terms of a property of measurable functions known as narrow convergence [4] which is an outcome of the topology induced by the Wasserstein metric. Since biological networks grow and shrink, therefore narrow convergence is not available for any metric which is to be defined. This is the reason a new kind of metric, i.e., the Gromov-Wasserstein metric [82] is chosen for such networks.

5.2 Basic concepts

In this section, the construction of a metric between two networks is described. Beginning with some basic definitions, the concepts of Hausdorff distance, isometric embedding and Wasserstein distance are explained with examples. A combination of these leads to the Gromov-Hausdorff distance. Finally, metric measure spaces are introduced to accommodate the notion of coupling measures along with a metric. This leads to the definition of the Gromov-Wasserstein distance.

5.2.1 Metric Spaces and isometric embedding

Definition 5.1 (Metric Space)**.** A metric space (X, d_X) consists of a set of points X and a distance function $d_X : X \times X \longrightarrow \mathbb{R}_+$ which satisfies the following properties

1. For every $x, y \in X$, $d_X(x,y) \geq 0$.
2. For every $x \in X$, $d_X(x,x) = 0$.
3. For every $x, y \in X$ such that $x \neq y$, $d_X(x,y) > 0$.
4. For every $x, y \in X$, $d_X(x,y) = d_X(y,x)$.
5. The triangle inequality: For every $x, y, z \in X$, $d_X(x,y) + d_X(y,z) \geq d_X(x,z)$

The above definition sometimes is relaxed with respect to condition 3, i.e., two distinct points in X are allowed to have $d_X = 0$. In such cases, the set X is said be endowed with a *pseudometric* or a *semimetric*.

Definition 5.2 (Embedding)**.** Given two metric spaces (X, d_X) and (Y, d_Y), a map $\phi : X \longrightarrow Y$ is called an embedding.

Definition 5.3 (Isometric embedding)**.** An embedding is called distance-preserving or *isometric* if for all $x, x' \in X$, $d_X(x,x') = d_Y(\phi(x), \phi(x'))$

Let $\mathscr{B}(X)$ be the Borel $\sigma-$algebra of X. Given two metric spaces (X, d_X) and (Y, d_y) with measures μ_X and μ_Y respectively. Now consider the product space $X \times Y$ with the product measure $\mu_X \otimes \mu_Y$. This product measure is defined on $(X \times Y, \mathscr{B}(X) \times \mathscr{B}(Y))$. In particular for subsets $A, B \in (X \times Y)$ that generate the $\sigma-$algebra, we have a product measure defined as $\mu_X \otimes \mu_Y(A \times B) = \mu_X(A)\mu_Y(B), \ \ \forall A \in \mathscr{B}(X), \ \ B \in \mathscr{B}(Y)$. Any network as defined in this work has three underlying components.

1. The space of nodes, denoted by X.

2. A set of interactions between each pair of nodes. This is represented through the adjacency matrix of the network and we represent it as ω_X.

3. A set of probability measures assigned to each node, denoted by the set μ.

This gives us a descriptor (X, ω, μ) for a network. To further characterize a network as defined above, we need to establish properties of this triplet.

5.2.2 Polish spaces

Definition 5.4 (Cauchy Space)**.** A sequence $x_1, x_2, x_3 \cdots \in (X, d)$ is Cauchy if for $\forall \varepsilon \in \mathbb{R}^+$ $\exists N \in \mathbb{Z}^+$ such that $\forall m, n \in \mathbb{Z}^+$ and $m, n > N$, the distance $d(x_m, x_n) < \varepsilon$.
What is implied by the above definition is that the terms of the sequence get close to each other in such a way that they seem to approach a limit. When such a limit exists inside X for every Cauchy sequence belonging to X, it makes X a *Cauchy space.*

Definition 5.5 (Separable metric space)**.** A topological space is defined as set of points which have sense of closeness defined among them but not a concrete notion of distance to measure it. One way to measure closeness of two set of points is through the notion of neighborhoods. A topological space which contains a countable dense subset is called separable. This means that there exists a sequence $\{x\}_{n=1}^{\infty}$ consisting of all elements of the space such that every nonempty open subset of the space contains at least one element of $\{x\}_{n=1}^{\infty}$.
This means that all finite spaces or separable. The space of nodes X defined earlier will be a countable dense subset of itself and hence separable.

Definition 5.6 (Completely metrizable space)**.** A space X is completely metrizable if there exists at least one d such that (X, d) is Cauchy. The topology T induced by d then defines

the topological space (X,T). i.e., the separation of points is now defined by the metric d and the space is Cauchy.

Definition 5.7 (Homeomorphism)**.** Consider two topological spaces X and Y. A mapping $f : X \longrightarrow Y$ is a Homeomorphism if

1. f is bijective.
2. f is continuous.
3. f^{-1} is continuous.

If such a mapping f exists, then X and Y are homeomorphic to each other. It is straightforward that this is an equivalence relation.

Definition 5.8 (Polish spaces and Measure networks)**.** A Polish space X_p is a completely metrizable space that is separable. This means that X_p is homeomorphic to a complete metric space (Cauchy) that has a countable dense subset.

In our definition of network space (X,ω,μ), the space X is a Polish space. The metric ω induces a topology on X. At this point, it is visible that any choice of ω that maintains the definition of X being a Polish space is acceptable. This means that we can move beyond adjacency or interaction matrices and tackle any kind of measured class of ω. This is useful in situations where a direct measurement of coupling matrices is not available and we have to depend on other measures of nodal distances. Most of biological data belongs to this category. It is understood that the nodal measure μ would From hereon we refer to X,ω,μ as a *measure network*.

5.2.3 The notion of *Closeness* between two measure networks

One of the prerequisites of the above definition of measure networks, the goal of which is to define a distance between two such triplets is to define which networks are considered

as the same. This leads to two cases.

1. A measure network is invariant under relabeling.

2. If there is a node that splits into multiple nodes such that the incoming and outgoing nodal distances remain the same, then the resulting network is the same as the original one.

The above conditions are formalised through the notion of isomorphisms. This is done through pushforward and pullback measures.

5.2.4 Pushforward or image measure and the change of variables formula

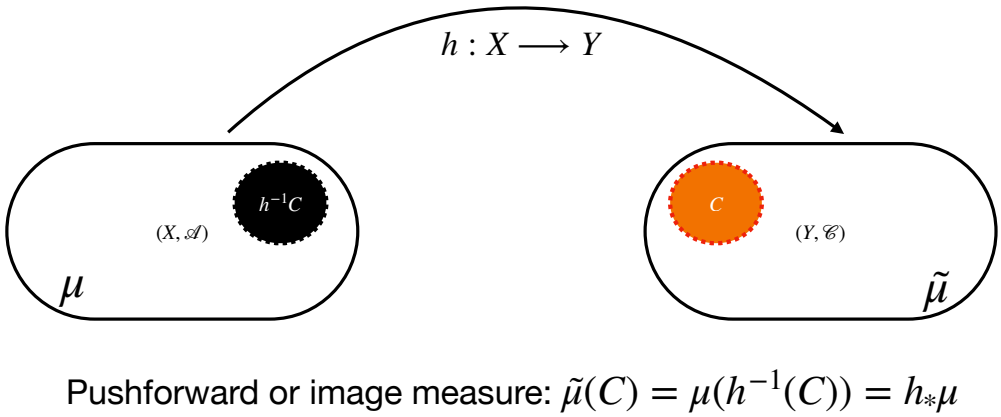
Definition 5.9 (Pushforward measure)**.** Given the measure space (X, μ_X) and Y and a Borel-measurable map $h : X \longrightarrow Y$. The pushforward of μ_X is defined as $h_* \mu_X(\mathscr{A}) = \mu_X(h^{-1}(\mathscr{A}))$ where $\mathscr{A}$ is the Borel $\sigma-$algebra on (X, μ_X). The construction essentially is that of a Borel measure on Y given the sigma algebra $\mathscr{A}$ and measure μ_X defined on X. This proceeds through the preimage of a subset $C \in \mathscr{C}$ generated by the inverse map h^{-1}. A schematic of the process is shown in Fig.(5.1).

Definition 5.10 (Pullback metric)**.** Suppose that you have two spaces X and Y,a metric d on Y, and a function $h : X \longrightarrow Y$. The pullback metric is the following metric on X

$$h^* d(x^{(1)}, x^{(2)}) = d(h(x^{(1)}), h(x^{(2)})), \;\; x^{(1)}, x^{(2)} \in X \;\; . \tag{5.1}$$

Thus, we define a metric on X by mapping points over to Y and taking the distance there.

It can be seen that the process of constructing a pullback metric somewhat mirrors the construction of the pushforward measure. In the latter, we construct a measure on Y using the preimage of a subset of Y on X through h^{-1}. We go from a subset on X to a subset on



[t]

Figure 5.1: A schematic representation of the pushforward measure between two measure spaces. The two polish spaces X and Y have sigma algebras $\mathscr{A}$ and $\mathscr{C}$ respectively. The bijective map $h : X \longrightarrow Y$ has the inverse mapping from Y to X on a subset C of Y denoted by $h^{-1}(C)$. The pushforward measure is the method to construct a measure on Y through the measure already defined on X. This is done through the preimage of the subset $C \in Y$ i.e. $h^{-1}(C)$ and measuring it with μ. The notation $h_*\mu$ is preferred as it establishes the significance of h in the definition.

Y and hence *push forward* a measure.

In the case of a pullback, we are given the images on Y through $h : X \longrightarrow Y$ and we *pull back* the metric on X through the inverse function, with the final result being a metric. Notice, that we did not begin with any knowledge of the metric on X.

5.2.5 Network Isomorphisms

With pushforwards and pullbacks defined, we are in a position to create a notion of *closeness* between networks. To that end, let $\mathscr{N}$ be the space of metric measure spaces i.e. measure networks.

Definition 5.11 (Strong Isomorphism)**.** Given (X, ω_X, μ_X), $(Y, \omega_Y, \mu_Y) \in \mathscr{N}$ and a Borel-measurable bijection $\varphi : X \longrightarrow Y$, with φ^{-1} denoting the inverse map. Let $\omega_X(x,x')$ denote the distance between the nodes x and x' with $\{x,x'\} \in X$. The two network spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) are strongly isomorphic if

1. $\omega_X(x,x') = \omega_Y(\varphi(x), \varphi(x'))\ \ \forall x,x' \in X$.

2. $\varphi_* \mu_X = \mu_Y$.

The strong isomorphism is denoted as $X \cong Y$.

From the viewpoint of practical applications, strong isomorphisms would be rare as the definition is strong and borders on equality. For a measure of similarity between network spaces, a weaker set of conditions could suffice. The parallel here is from the zeroth law of thermodynamics which states that if two thermodynamic systems are each in thermal equilibrium with a third system, then they are in thermal equilibrium with each other. Accordingly, thermal equilibrium between systems is a transitive relation. In case of metric measure spaces, we aim for sustaining a "transitive" definition of isomorphism given a third

measure space (Z, μ_Z). The definition cannot be exact like the zeroth law as isomorphism needs a pushforward and pullback.

Definition 5.12 (Weak Isomorphism)**.** The two network spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) are *weakly* isomorphic if $\exists (Z, \mu_Z)$ along with $f : Z \longrightarrow X$ and $g : Z \longrightarrow Y$ such that.

1. $f_* \mu_Z = \mu_X, \;\; g_* \mu_Z = \mu_Y \quad .$
2. $||f^* \omega_X - g^* \omega_Y||_\infty = 0 \;\; .$

Here, $f^* \omega_X$ is the pullback weight of from the map $(z, z') \mapsto \omega_X(f(z), f(z'))$ and $g^* \omega_Y$ is from $(z, z') \mapsto \omega_Y(g(z), g(z'))$ with the notations running parallel to the Definition 5.10.

The pullbacks mentioned have to be measurable. For this lets consider the following.

Theorem 5.1 (Measurable Pullbacks)**.** *The pullbacks defined in definition 5.12 are measurable.*

Proof. A measure μ_z on the metric measure space z is a linear functional from a vector space of functions $\Phi(Z)$ on Z. This gives

$$\mu_Z : \Phi(Z) \longrightarrow \mathbb{R}; \quad \varphi \mapsto \langle \mu_Z, \varphi \rangle := \int_Z \varphi(z) \mu_Z(dz) \;\; .$$

We can view the space of measures $M(Z)$ on Z as dual to $\Phi(Z)$. The dual operation on this space of measures is the pushforward. Thus if $f : Z \longrightarrow X$ is a map then we get two linear maps

$$f^* : \Phi(X) \longrightarrow \Phi(Z), \;\; f_* : M(Z) \longrightarrow M(X)$$

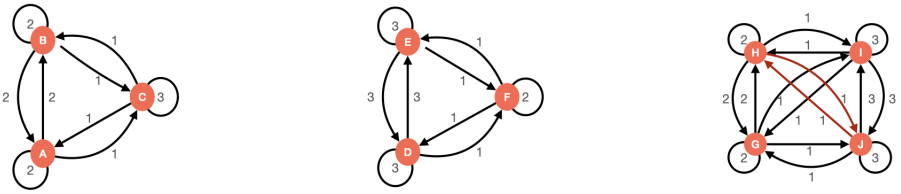
related as follows,

$$\langle \mu_Z, f^* \varphi \rangle = \langle f_* \mu_Z, \varphi \rangle, \;\; \forall \varphi \in \Phi(X), \mu_Z \in M(Z)$$

This program is analogous for the network space Y. □

The pullback f^* acts on spaces of functions via the equality $f^*\varphi = \varphi \circ f$. It is the dual of the pushforward f_* in the sense of duality of locally convex topological spaces.

We can illustrate this through the example networks in Fig.(5.2). Three networks X, Y and Z with different weight matrices and probability vectors are weakly isomorphic to each other, generating pushforwards and measurable pullbacks. In this particular example, the network Z maps surjectively onto X and Y. For X, let the map $f : Z \longrightarrow X$ induce the pushforward μ_X and analogously $g : Z \longrightarrow Y$ would induce μ_Y for Y. The surjection maps both I and J to C. The pullbacks $f_*\omega_X$ and $g_*\omega_Y$ both generate ω_Z as dual maps to their respective pushforwards ω_X and ω_Y.



(X) (Y) (Z)

$$\omega_X(A,B,C) = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \quad \mu_X = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} \qquad \omega_Y(D,E,F) = \begin{pmatrix} 3 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \quad \mu_Y = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} \qquad \omega_Z(G,H,I,J) = \begin{pmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 3 \end{pmatrix}, \quad \mu_Z = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

Figure 5.2: The network Z maps surjectively on X and Y. Considering X and Z, the surjection maps I and J both to C as the incoming and outgoing edges as well as self-edges are the same for all three while G and H are the same as A and B. The pullbacks i.e. ωs can be mapped to Z according to Theorem 5.1.

5.2.6 Isomosrphisms in context of real network data

Consider the network space (X, ω_X, μ_X). Two nodes $\{x, x'\} \in X$, are said to be ε similar if for $\varepsilon > 0$,

$$\|\omega_X(x,.) - \omega_X(x',.)\|_\infty < \varepsilon, \quad \|\omega_X(.,x) - \omega_X(.,x')\|_\infty < \varepsilon$$

The rationale of the ε inequalities lies in the internal and external *perceptions* of x and x'. If $\omega_X(x,x') = \omega_X(x',x) = \omega_X(x,x) = \omega_X(x',x')$, then the nodes x and x' have the same internal perceptions. To have the same external perceptions, all the incoming and outgoing edges need to be the same for both x and x'. The ε inequalities accommodate a reasonable relaxation to the same in context of real-world data which is subject to instrumentation (calibration) as well as measurement errors.

Given a $k \times k$ weight matrix $M_k \in \mathbb{R}^{k \times k}$ and $k \times 1$ vector μ such that $\sum_{i=1}^{k} \mu_M = 1$, a network with k nodes denoted by (M_k, μ_M) is obtained. Another network $(N_j, \nu_N) \cong^s (M_k, \mu_M)$ if and only if $j = k$ and $\exists\ P$, a permutation matrix such that $N_j = PM_KP^T$, and $P\mu_M = \nu_N$.

5.3 Measure couplings

In this section we formalize the notion of measure couplings on metric measure spaces. More specifically, we define what are known as *transport plans* between network spaces in context of optimal transport. We follow the construction in [116] and extend through the previous sections.

Definition 5.13 (Measure coupling)**.** Given two network spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y), any probability measure μ on the product space $X \times Y$ (equipped with the product topology

and product $\sigma-$field) satisfying

$$(\pi_X)_*\mu = \mu_X, \quad (\pi_Y)_*\mu = \mu_Y \ , \tag{5.2}$$

is called a coupling of measures μ_X and μ_Y. The network measure vectors μ_X and μ_Y will be the marginals of the coupling (transport plan) μ. Here π_X and π_Y are projections from $X \times Y$ to X and Y respectively.

The pushforwards in Eq.(5.2) are defined on the Borel $\sigma-$algebras of both the networks. This makes it possible to write them as

$$\mu(A_X \times Y) = \mu_X(A_X) \ \ \forall A_X \in \mathscr{B}(X) \ . \tag{5.3}$$

$$\mu(A_Y \times X) = \mu_Y(A_Y) \ \ \forall A_y \in \mathscr{B}(Y) \ . \tag{5.4}$$

The collections of all couplings between (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) will be denoted as $\mathscr{C}(\mu_X, \mu_Y)$. The set $\mathscr{C}(\mu_X, \mu_Y)$ is nonempty as because it always contains the **product coupling** $\mu = \mu_X \otimes \mu_Y$, which is unique as

$$\mu(A_X \times A_Y) = \mu_X(A_X).\mu_Y(A_Y), \ \ \forall A_X \in \mathscr{B}(X), \ \ A_Y \in \mathscr{B}(Y) \ . \tag{5.5}$$

If one of the probabilities is a Dirac measure, then $\mu = \mu_X \otimes \mu_Y$ is the only coupling, i.e., $\mathscr{C}(\delta_{x_o}, \mu_Y) = \{\delta_{x_o} \otimes \mu_Y\}$

Lemma 5.1. *Given (μ_X and μ_Y), the set of all couplings $\mathscr{C}(\mu_X, \mu_Y)$ is a nonempty compact subset of $\mathscr{P}(X \times Y)$, the set of all probability measures on $X \times Y$ equipped with the weak topology.*

Proof. As the projection maps π_X and π_Y in Eq.(5.13) are continuous and $\mathscr{C}(\mu_X, \mu_Y)$ is a closed subset of $\mathscr{P}(X \times Y)$, the result is an application of Prokhorov's theorem (Theorem

5.3).

For every map $f : X \longrightarrow Y$ with $f_*\mu_X = \mu_Y$ a coupling of μ_X and μ_Y can be written as

$$\mu = (Id, f)_*\mu_X \quad .$$

If $\mu_X = \mu_Y$ as $X = Y$, then for the choice $f = Id$, we get the diagonal coupling

$$d\mu(x,y) = d\delta_x(y)d\mu_X(x) \quad .$$

In general, for (Z, ω_Z, μ_Z) and measurable maps $g : Z \longrightarrow X$ and $h : Z \longrightarrow Y$ with $g_*\mu = \mu_X$ and $h_*\mu = \mu_Y$. A coupling of μ_X and μ_Y is given by

$$\mu(x,y) = (g,u)_*\mu \quad .$$

□

One can now look at the concept of coupling from the perspective of correspondence between two measure spaces. The support for the couplings between two metric-measure space will be a subset of $X \times Y$ for the network spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y).

Couplings as Markov Kernels

Couplings as defined above refer to a joint probability distribution of μ_X and μ_Y so as to admit a disintegration through a Markov kernel. This is achieved as follows.

The metric measure spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) have their respective Borel probability measures μ_X and μ_Y. The coupling $d\mu(x,y)$ is admitted as a map from $X \times Y$ to $\mathscr{S}$.

$$d\mu(x,y) = d\mu_x(y)d\mu_X(x) \quad .$$

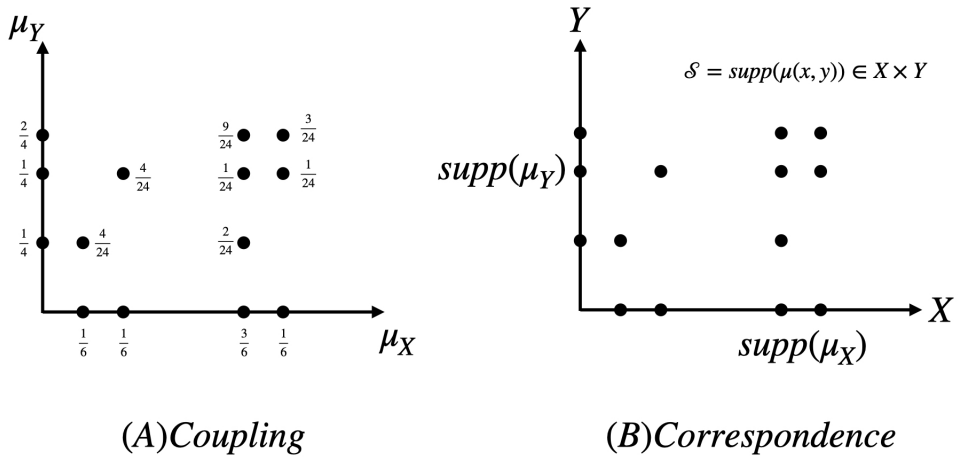


Figure 5.3: The coupling has the correspondence between (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) as its underlying structure. The couplings are an optimal quantitative representation of a correspondence set $\mathscr{S} \in X \times Y$ which is the support set of the optimal couplings between the two metric measure spaces.

This is how couplings are defined as measures on $X \times Y$. This *disintegration* of probability measures is the basis of *soft-coupling* as it replaces $\varepsilon-$isometries, i.e., $f : X \longrightarrow Y$, between the two metric-measure spaces (X, ω_X, μ_X) and (Y, ω_Y, μ_Y). The kernel facilitates the mapping of $\mu_X \in X$ to $\mu_x(y) \in Y$ instead of the map $f : X \longrightarrow Y$ mapping points directly from X to Y. This new disintegration of a measure on X into one on Y depending on the coordinates mapped provides *probability maps* irrespective of the codimensionality of X and Y in $\mathscr{N}$, i.e., the space of all networks. This is the measure map which we refer to as a soft-coupling. Further development of this concept is required as two issues need to be addressed.

1. A series of couplings need to form a connected map so that multiple spaces could be soft-coupled.

2. An optimal coupling set has to exist and to define such a set, a metric optimal condi-

tion needs to be defined between any two metric measure spaces.

We begin these tasks by stating the gluing lemma for couplings.

Theorem 5.2 (Gluing lemma)**.** *Let $X_1, X_2, \cdots X_k$ be polish spaces and $\mu_1, \mu_2, \cdots \mu_k$ be respective probability measures on the $\sigma-$fields. Then for every choice of couplings $\mu \in \mathscr{C}(\mu_{i-1}, \mu_i),\ \forall i \in \{1,2,3,\cdots k\},\ \exists!\, \mu \in \mathscr{P}(X_1, X_2, \cdots X_k)$ such that*

$$(\pi_{i-1}, \pi_i)_* \mu = \mu_i, \qquad \forall i \in \{1,2,3,\cdots k\} \ .$$

μ is defined as the gluing of the couplings $\mu_1, \mu_2, \cdots \mu_k$ and is denoted as

$$\mu = \mu_1 \boxtimes \mu_2 \cdots \boxtimes \mu_k \ .$$

In particular μ has marginals $\mu_1, \mu_2, \cdots \mu_k$, i.e., $(\pi_i)_* \mu = \mu_i,\ \forall i \in \{1,2,3,\cdots k\}$.

Proof. The result for $k = 2$ [124] is presented first. The basic idea for $k > 2$ is to form a sequence of probability measures using the Markov kernel. Central to this scheme is the idea of **disintegration of measure**. If π is a probability measure on $X \times Y$ with marginal μ_X on X, then there exists a measurable application $x \mapsto \pi_x, \forall x \in X$ from X to $\mathscr{P}(Y)$, uniquely determined $d\mu(X)$ almost everywhere, such that

$$\pi = \int_X (\delta_x \otimes \pi_x) d\mu(x) \ .$$

This means that a probability mapping from X to Y through $u \in (C)(X \times Y)$

$$\int_{X \times Y} u(x,y) d\pi(x,y) = \int_X \left[\int_Y u(x,y) d\pi_x(y) \right] d\mu(x) \ ,$$

or for any measurable set $A \subset X \times Y$

$$\pi[A] = \int_X \pi_x[A_x] d\mu(x) \ ,$$

where

$$A_x = \{y \in Y : (x,y) \in A\} \ .$$

Next we go to three Polish spaces to take the first gluing step. Consider a third space Z and define probability maps/transport plans $\pi(x,y) \in \mathscr{C}(X \times Y)$ and $\pi(y,z) \in \mathscr{C}(Y \times Z)$. The common marginal here is μ_Y so it will be used to disintegrate both probability maps. This occurs through measurable applications $\pi_{xy:y} : Y \mapsto \mathscr{P}(Y)$ and $\pi_{yz:z} : Y \mapsto \mathscr{P}(Z)$ such that

$$\pi_{xy} = \int_Y \pi_{xy:y} \otimes \delta_y d\mu_Y(y) \ ,$$

and

$$\pi_{yz} = \int_Y \delta_y \otimes \pi_{yz:z} d\mu_Y(y) \ .$$

Finally, the three space glued coupling is formed as

$$\pi(X \times Y \times Z) = \int_Y \pi_{xy:y} \otimes \delta_y \otimes \pi_{yz:z} d\mu_Y(y) \ .$$

□

5.3.1 Distortion of a coupling

We again work with (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) as two metric-measure networks. Let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ and consider a probability space $(X \times Y)^2$ equipped with the product mea-

sure $\mu \otimes \mu$. The $p-$distortion function is defined as

$$dis_p(\mu) = \left(\int_{X\times Y} \int_{X\times Y} |\omega_X(x,x') - \omega_Y(y,y')|^p d\mu(x,y) d\mu(x',y') \right)^{\frac{1}{p}} ,$$

in the sense of l^p-spaces. In the present work we would be concerned mainly with the case $p = 2$. The distortion is a metric defined on the space of spaces and its optimization would lead to the optimal coupling set. We now state that theory. First, we need tightness of probability measures within the network spaces. For that we need the following result

Theorem 5.3 (Prokhorov's Theorem)**.** *Let X be a Polish space. Then $P \subseteq \mathscr{P}(X)$ is tight, if and only if its closure is compact in $\mathscr{P}(X)$.*

Proof. We first prove the backward argument. ($\Longleftarrow$):

Claim: If $U_1, U_2, \cdots$ are open sets in X that cover X and if $\varepsilon > 0$ then there exists a $k \geq 1$ such that

$$\mu \left(\bigcup_{i=1}^{k} U_i \right) > 1 - \varepsilon, \ \ \forall \mu \in P \ .$$

To prove by contradiction, suppose that for every $k > 1$ there is a $\mu_k \in P$ with $\mu_k \left(\bigcup_{i=1}^{k} U_i \right) \leq 1 - \varepsilon$. As P^c is compact, there is a $\mu \in P^c$ and a subsequence $\mu_{k_j} \implies \mu$. For any $n \geq 1$, $\bigcup_{i=1}^{k} U_i$ is open. Hence

$$\mu \left(\bigcup_{i=1}^{n} U_i \right) \leq \liminf_{j\to\infty} \mu_{k_j} \left(\bigcup_{i=1}^{n} U_i \right) \leq \liminf_{j\to\infty} \mu_{k_j} \left(\bigcup_{i=1}^{k_j} U_i \right) \leq 1 - \varepsilon \ .$$

But $\bigcup_{i=1}^{\infty} U_i = X$, so $\mu(\bigcup_{i=1}^{n} U_i) \to \mu(X) = 1$ as $n \to \infty$. This a contradiction, proving the claim. As the next step we start with $\varepsilon > 0$ and take $D = \{a_1, a_2, \cdots\}$ to be dense in X. For every $m \geq 1$, the open balls $B\left(a_i, \frac{1}{m}\right), i = \{1, 2, \cdots\}$ cover X. So by the claim we just proved,

there is a k_m such that

$$\mu\left(\bigcup_{i=1}^{k_m} B\left(a_i, \frac{1}{m}\right)\right) < 1 - \varepsilon 2^{-m}, \ \ \forall \mu \in P \ .$$

now we define

$$K := \bigcap_{m=1}^{\infty} \bigcup_{i=1}^{k_m} \bar{B}\left(a_i, \frac{1}{m}\right) \ .$$

To make K totally bounded we observe that it is closed and then see that for each $\delta > 0$ we can take $m > \frac{1}{\delta}$ and $K \subset \bigcup_{i=1}^{k_m} \bar{B}(a_i, \delta)$. This makes K compact as X is complete. Now, for every $\mu \in P$

$$\begin{aligned}\mu(X \setminus K) = \mu\left(\bigcap_{m=1}^{\infty}\left[\bigcup_{i=1}^{k_m} \bar{B}\left(a_i, \frac{1}{m}\right)\right]^c\right) &\leq \sum_{m=1}^{\infty} \mu\left(\left[\bigcup_{i=1}^{k_m} \bar{B}\left(a_i, \frac{1}{m}\right)\right]^c\right) \\ &= \sum_{m=1}^{\infty}\left(1 - \mu\left(\bigcup_{i=1}^{k_m} \bar{B}\left(a_i, \frac{1}{m}\right)\right)\right) < \sum_{m=1}^{\infty} \varepsilon 2^{-m} = \varepsilon\end{aligned}$$

Hence P is tight.

□

The forward proof is much more delicate and has prerequisites. Namely,

1. Compactification of metric spaces.

2. Riesz representation theorem.

First, we look at compactification. To that end, lets rephrase the forward proof slightly. X is a compact metric space and therefore every set of Borel probability measures on it is tight and therefore $\mathscr{P}(X)$ is tight. Thus the forward result implies that $\mathscr{P}(X)$ is compact whenever X is compact. This would be our proof. As we stated earlier, there are some prerequisites for the proof and we now gather them.

Proposition 1. *If (X,d) is a metric space, then $(\mathscr{P}, d_P)$ is a compact metric space.*

Proof. Let us denote the set of continuous and bounded functions on X as

$$C_b(X) := \{f : X \longrightarrow \mathbb{R} : \text{f is continuous and bounded}\}$$

This implies that each $f \in C_b(X)$ is integrable with respect to any finite Borel measure on X. Also, since X is compact, therefore

$$C_b(X) = C(X) = \{f : X \longrightarrow \mathbb{R} : \text{f is continuous}\} \ .$$

This means that $C(X)$ is a Banach space under the supremum norm defined by

$$||f||_\infty = \sup_{x \in X} |f(x)| \ .$$

We denote by $C(X)'$ the dual Banach space of $C(X)$ and consider

$$\Phi := \{\phi \in C(X) : ||\phi|| \leq 1, \phi(\mathbb{1}) = 1, \phi(f) \geq 0, \forall f \in C(X), \text{ with } f \geq 0\} \ ,$$

with $\mu \in \mathscr{P}(X)$ define $\phi_\mu(f) := \int f d\mu$, with $f \in C(X)$. Now, $T : \mu \to \phi_\mu$ is a bijection from $\mathscr{P}(X)$ onto Φ. Moreover, T is a sequential homeomorphism relative to the weak* topology on Φ. This follows straightforwardly from the Riesz representation theorem. At this point we invoke Alaoglu's theorem which reads as follows

Theorem 5.4 (Banach–Alaoglu)**.** *For a topological vector space X with continous dual space X', the polar set*

$$U^0 = \{x' \in X' : \sup_{x \in U} |x'(x)| \leq 1\}$$

of any neighborhood U of origin in X is compact in weak* topology $\sigma(X', X)$ on X'. More-

over, U^0 is the polar of U with respect to the canonical system $\langle X, X^{\#}\rangle$ and it is also a compact subset of $(X^{\#}, \sigma(X^{\#}, X))$.

A specialized version of this would be, If X is a normed space then the closed unit ball in the continuous dual space X' (endowed with its usual operator norm) is compact with respect to the weak* topology. We omit the proof but use the specialized version for our map Φ. Banach–Alaoglu theorem implies that B' defined as

$$B' = \{\phi \in C(X)' : ||\phi|| \leq 1\} \ ,$$

is weak* compact and therefore so is Φ since it is weak* closed in B'. This proves the proposition by leading to the fact that $\mathscr{P}(X)$ is compact since Φ is sequentially compact. It is to be noted that the converse is also true. □

We now take the next step in the compactification of a probability space by proving the existence of a homeomorphic map as follows.

Lemma 5.2. *If (X, d) is a separable metric space then there exists a compact metric space (Y, δ) and a map $T : X \longrightarrow Y$ such that T is a homeomorphism from X onto $T(X)$.*

Proof. To begin the argument, we need to define Y and T as follows. Let $Y := [0,1]^{\mathbb{N}} = \{(\xi_i)_{i=1}^{\infty} : \xi_i \in [0,1] \forall i\}$ and

$$\delta(\xi, \eta) := \sum_{i=1}^{\infty} 2^{-i} |\xi_i - \eta_i|, \quad \xi, \eta \in Y \ .$$

This makes δ a metric on Y with its topology being the topology of coordinate-wise convergence and (Y, δ) is compact. Let $D = \{a_1, a_2, \cdots\}$ be dense in X and define

$$\alpha_i(x) := \min\{d(x, a_i), 1\}, \ \ x \in X, i = 1, 2 \cdots$$

Then for each k, $\alpha_K : X \to [0,1]$ is continuous. For $x \in X$ define

$$T(x) := (\alpha_i(x))_{i=1}^{\infty} \in Y \ .$$

Claim: for any $C \subset X$ closed and $x \notin C$ there exists $\varepsilon > 0$ and i such that

$$\alpha_i(x) \leq \frac{\varepsilon}{3}, \quad \alpha_i(y) \geq \frac{2\varepsilon}{3}, \ \ \forall y \in C \ .$$

To prove the claim, take $\varepsilon := \min d(x,C), 1 \in (0,1]$. Take i such that $d(a_i,x) < \frac{\varepsilon}{3}$, then $\alpha_i(x) \leq \frac{\varepsilon}{3}$ and for $y \in C$ we have

$$\begin{aligned} \alpha_i(y) = \min\{d(y,a_i),1\} \geq \min\{(d(y,x) - d(x,a_i)),1\} \geq \min\{(d(x,C) - \frac{\varepsilon}{3}),1\} \\ \geq \min\{\frac{2\varepsilon}{3},1\} = \frac{2\varepsilon}{3} \end{aligned}$$

This proves that T is injective as if $x \neq y$, then $\exists$ an i such that $\alpha_i(x) \neq \alpha_i(y)$. The way we defined $T : X \longrightarrow T(X)$, it is surjective, so it is a bijection. For proving the lemma, we need to show sequential convergence, i.e. ,

$$x_n \longrightarrow x \iff T(x_n) \longrightarrow T(x)$$

To that end, if $x_n \longrightarrow x$ then $\alpha_i(x_n) \longrightarrow \alpha_i(x), \forall i$ and therefore $\delta(T(x_n), T(x)) \to 0$, as $n \to \infty$. □

We now complete the proof of Prokhorov's theorem.

Theorem 5.5 (Forward direction for Prokhorov's Theorem)**.** *If (X,d) is a separable metric space and $\Gamma \in \mathscr{P}(X)$ is tight, then $\bar{\Gamma}$ is compact.*

Proof. We first observe that both Γ and $\bar{\Gamma}$ are tight in X. Now, let $\varepsilon > 0$ and K be a compact

subset of X such that $\mu(K) \geq 1-\varepsilon, \forall \mu \in \Gamma$. Then for every $\mu \in \bar{\Gamma}$, there is a sequence $(\mu_n)_n$ in Γ that converges to μ giving

$$\mu(K) \geq \limsup_{n\to\infty} \mu_n(K) \geq 1-\varepsilon$$

Next, let $(\mu_n)_n$ be a sequence in $\bar{\Gamma}$ and we aim to show that $(\mu_n)_n$ has a convergent subsequence. Repeating our earlier construction for homeomorphic maps, let (Y,δ) be a compact metric space and a map $T : X \longrightarrow Y$ such that T is a homeomorphism from X onto $T(X)$. For, $\mathscr{B} \in \mathscr{B}(Y), T^{-1}(\mathscr{B})$ is Borel in X. Define, $\nu_n(B) := \mu_n(T^{-1}B) \ \ , B \in \mathscr{B}, \forall n \in \mathbb{N}$. Then $\nu \in \mathscr{P}(Y) \ \ \forall n$. As Y and $\mathscr{P}(X)$ are compact metric spaces, there is s ν such that there is a compact subsequence $\nu_{n_k} \to \nu$ in $\mathscr{P}(Y)$. We need to construct a map to bring ν back to X. To be exact, a translation of ν on a measure on X. For this, set $Y_0 = T(X)$.

Claim: There exists a set $E \in \mathscr{B}(Y)$ such that $E \subset Y_0$ and $\nu(E) = 1$. This means that ν is concentrated on Y_0.

Proof. Let $\nu_0(A) := \nu(A \cap E), \ \ A \in \mathscr{B}(Y_0)$. Why this works is because $A \in \mathscr{B}(Y_0) \implies A \cap E \in \mathscr{B}(E) \implies A \cap E \in \mathscr{B}(Y)$ as E is a Borel subset of Y. Also, ν_0 is a finite Borel measure on Y_0 and $\nu(E) = \nu_0(E) = 1$. This provides us with the measure translation

$$\mu(A) = \nu_0(T(A)) = \nu_0((T^{-1})^{-1}(A)), \ \ A \in \mathscr{B}(X)$$

This gives $\mu \in \mathscr{P}(X)$ and now we have to show that $\mu_{n_k} \to \mu$ in $\mathscr{P}(X)$. Let C be closed in X. Then $T(C)$ is closed in $T(X) = Y_0$ and need not be closed in Y. This means $\exists Z \in Y$ closed with $Z \cap Y_0 = T(C)$. This means $C = \{x \in X : T(x) = T(C)\} = \{x \in X : T(x) \in Z\} =$

$T^{-1}(Z)$, because there are no points in $T(C)$ outside Y_0, and $Z \cap E = T(C) \cap E$. This gives

$$\begin{aligned}\limsup_{k\to\infty}\ \mu_{n_k}(C) = \limsup_{k\to\infty}\ \nu_{n_k}(Z) &\leq \nu(Z) = \nu(Z\cup E) + \nu(Z\cup E^c) \\ &= \nu(T(C)\cap E) + 0 = \nu_0(T(C)) = \mu(C)\ ,\end{aligned}$$

and consequently $\mu_{n_k} \to \mu$. Finally, to prove the claim we use tightness of $\bar{\Gamma}$. For each $m \geq 1$ take K_m compact in X such that $\mu(K_m) \geq 1 - \dfrac{1}{m}\ \ \forall \mu \in \Gamma$. Then $T(K_m)$ is a compact subset of Y hence closed in Y, so

$$\nu(T(K_m)) \geq \limsup_{k\to\infty}\ \nu_{n_k}(T(K_m)) \geq \limsup_{k\to\infty}\ \mu_{n_k}(K_m) \geq 1 - \frac{1}{m}\ .$$

Now we take $\bigcup_{m=1}^{\infty} K_m$. This gives $E \in \mathscr{B}(Y)$ and $\nu(E) \geq \nu(K_m) \forall m$. So $\nu(E) = 1$, proving the claim. □

This completes the proof of Prokhorov's theorem, hence paving the way for the compactness of couplings in the space of network spaces. In the next chapter, we use this fact to construct maps between to isomorphism classes and characterize the geodesics. □

Lemma 5.3 (Compactness of couplings)**.** *Following the notation in previous sections, $\mathscr{C}(\mu_X \times \mu_Y)$ is compact in $\mathscr{P}(X \times Y)$.*

Proof. This proof requires the following lemma.

Lemma 5.4. *[123] Let X and Y be two Polish spaces. Let $P_X \subseteq \mathscr{P}(X)$ and $P_Y \subseteq \mathscr{P}(Y)$ be tight in their respective spaces. Then the set $\mathscr{C}(P_X, P_X) \subseteq \mathscr{P}(X \times Y)$ of couplings with marginals P_X and P_Y is tight in $\mathscr{P}(X \times Y)$.*

We omit the proof. Lemma 5.4 states that the couplings are tight and we have to prove that they are compact. We have to apply Prokhorov's theorem (Theorem 5.3). Consider the singletons $\{\mu_X\}$ and $\{\mu_Y\}$. They are closed and compact in $\mathscr{P}(X)$ and $\mathscr{P}(Y)$. By

theorem 5.3, they are tight. Now consider $\mathscr{C}(P_X, P_X) \subseteq \mathscr{P}(X \times Y)$. This set is obtained by the intersecting the preimages of continuous projections onto the marginals μ_X and μ_Y and therefore is closed. By lemma 5.4 it is tight. By applying theorem 5.3 again, it is compact. □

Compactness of couplings is a crucial component of the proof of existence of optimal couplings. Without compactness, the infimum of the metric does not necessarily exist which is a fact an optimal metric relies upon. In thermodynamics, the Wasserstein metric induces the minimum entropy topology due to compactness of the measure couplings and the same is required here in the absence of narrow convergence. The second necessary component is of course the continuity of the distortion functional. We cannot guarantee the existence of the metric at all points (network spaces) in space of (network)spaces if the integral over the pseudomanifold is not continuous.

5.3.2 Continuity of the distortion functional

Lemma 5.5. *[27][Continuity of the distortion functional] Let* $1 \leq p < \infty$*, with* (X, ω_X, μ_X)*,* $(Y, \omega_Y, \mu_Y) \in [\mathscr{N}]$*. The distortion functional* dis_p *is continuous on* $\mathscr{C}(\mu_X, \mu_Y)$*. For* $p = \infty$*,* dis_∞ *is lower semicontinuous.*

Proof. Uniform limit of continuous functions is continuous. The strategy is to make a sequence of continuous functions converge to dis_p, making it continuous. We are only considering Polish spaces with finite measures and therefore as it is true for L^p spaces, bounded continuous functions are dense. We pick two such functions. $\omega_X^n = L^p\left(\mu_X^{\otimes 2}\right)$ and $\omega_Y^n = L^p\left(\mu_Y^{\otimes 2}\right)$ such that $\|\omega_X - \omega_X^n\|_{L^p(\mu_X \otimes \mu_X)} \leq \frac{1}{n}$, and $\|\omega_Y - \omega_Y^n\|_{L^p(\mu_Y \otimes \mu_Y)} \leq \frac{1}{n}$. For each $n \in \mathbb{N}$ we define the functional $dis_p^n : \mathscr{C}(\mu_X, \mu_Y) \to \mathbb{R}_+$ as $dis_p^n(\nu) = \|\omega_X^n - w_Y^n\|_{L^p(\nu \otimes \nu)}$. We also note that this function is bounded on the probability space. Next, the narrow topology on $\mathscr{P}(X \times Y)$ is induced by a distance and therefore it is enough to show sequential

continuity to prove that dis_p^n is continuous. Let $\nu \in \mathscr{C}(\mu_X, \mu_Y)$ and let $(\nu_m)_{m\in\mathbb{N}}$ be a sequence on $\mathscr{C}(\mu_X, \mu_Y)$ converging narrowly to ν. Then $\nu_m \otimes \nu_m$ converges narrowly to $\nu \otimes \nu$ [13, 110]. This gives

$$\lim_{m\to\infty} dis_p^n(\nu_m) = \lim_{m\to\infty} \left(\int_{X\times Y} \int_{X\times Y} |\omega_X^n - \omega_Y^n|^p \, d\nu_m \otimes d\nu_m \right)^{\frac{1}{p}}, \tag{5.6}$$

and with the bounded and continuous integrand gives

$$\lim_{m\to\infty} dis_p^n(\nu_m) = \left(\int_{X\times Y} \int_{X\times Y} |\omega_X^n - \omega_Y^n|^p \, d\nu \otimes d\nu \right)^{\frac{1}{p}} = dis_p^n(\nu), \tag{5.7}$$

from the definition of narrow convergence. This gives sequential continuity and therefore by the definition of narrow convergence, continuity of $dis_p^n(\nu)$. Now, we have to show that $dis_p^n(\nu)$ is uniformly convergent. For that we start with $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ and

$$\left\| dis_p(\mu) - dis_p^n(\mu) \right\| = \left| \|\omega_X - \omega_Y\|_{L^p(\nu\otimes\nu)} - \|\omega_X^n - \omega_Y^n\|_{L^p(\nu\otimes\nu)} \right|. \tag{5.8}$$

From Minkowski inequality [12], we get

$$\leq \left| \|\omega_X - \omega_X^n\|_{L^p(\mu_X\otimes\mu_X)} - \|\omega_Y - \omega_Y^n\|_{L^p(\mu_Y\otimes\mu_Y)} \right| \leq \frac{2}{n}. \tag{5.9}$$

This shows that the distortion functional is the uniform limit of continuous functions and is therefore uniformly continuous. The lower semicontinuity of dis_p^n follows from Jensen's inequality as we are working in probability spaces. □

Now that we have proven the continuity of the distortion functional over the network space along with the compactness of measure couplings, the next step is use the gluing lemma to prove the existence of optimal couplings.

5.3.3 Existence of optimal couplings

First, we need a definition

Definition 5.14 (optimal couplings)**.** Let (X, ω_X, μ_X), $(Y, \omega_Y, \mu_Y) \in [\mathscr{N}]$ and $p \in [1, \infty]$. A coupling $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ is optimal if $dis_p(\mu) = \inf_{\nu \in C(\mu_X, \mu_Y)} dis_p(\nu)$.

We now have the following theorem.

Theorem 5.6 (Existence of optimal $\mu \in \mathscr{C}(\mu_X, \mu_Y)$)**.** *Let $1 \leq p < \infty$, with two measure networks (X, ω_X, μ_X), $(Y, \omega_Y, \mu_Y) \in [\mathscr{N}]$. Then there exists a minimizer of $dis_p(.)$ in $\mathscr{C}(X \times Y)$ it is the optimal coupling.*

Proof. The work for this proof has already been done. We invoke lemmas 5.5 and 5.3 and observe that $dis_p(\mu)$ is lower semicontinous and compact. This implies that it has an infimum on $\mathscr{C}(\mu_X, \mu_Y)$. □

We can now define the optimal metric from the infimum we have just proven to exist.

5.4 Gromov-Wasserstein distance

Definition 5.15 (Gromov-Wasserstein distance)**.** [82, 116] The Gromov-Wasserstein distance between two networks (X, ω_X, μ_X), $(Y, \omega_Y, \mu_Y) \in [\mathscr{N}]$ is given as

$$d_{\mathscr{N}_p}(X, Y) = \frac{1}{2} = \inf_{\mu \in C(\mu_X, \mu_Y)} dis_p(\mu) \tag{5.10}$$

We observe that this distance is bounded [27]. In addition, it is a *pseudometric* [117]. Now that we have our main tool ready, it is time to test it through some simple example computations.

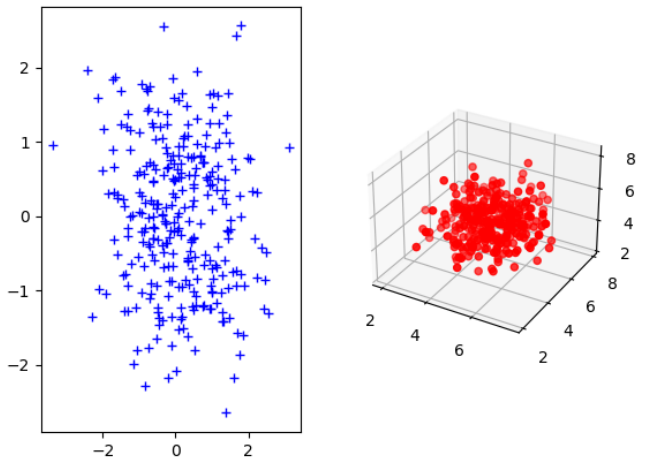


Figure 5.4: Plots of two normal distribution samples. The first distribution is two-dimensional and the points are generated on a plane with $\mathcal{N}(0,1)$, while the second is three-dimensional and the points are generated in a volume with $\mathcal{N}(4,1)$.

5.4.1 Computing Gromov-Wasserstein distance between two distributions

We illustrate the GW metric calculation through two gaussian distributions of different sizes. This is the example for the optimal transport library for the Python language. Once we have this example calculation correct, then we can extend the same to a network calculation. The first task is plot the two distribution samples between which the GW metric is to be calculated. A two and a three dimensional set of points sampled from two normal distributions is shown in Fig.(5.4). The second plot shown in Fig.(5.5) shows the heat maps of the loss functions of both distributions. The first distribution would have these as line segments in a plane while the second would have them in a volume. What is shown in the maps is the magnitude of the lengths of those line segments. We now plot the GW metric. The GW metric is calculated through the Sinkhorn algorithm [34] and is cumulative for all the points in the final calculation. The plot shows the individual contribution for each point. The Borel probability measures are taken from the uniform distribution. The color

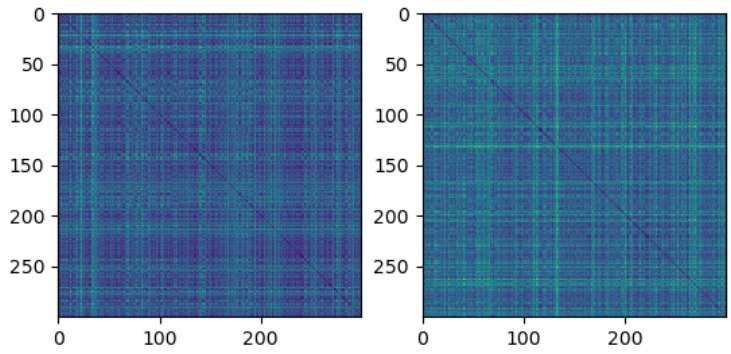


Figure 5.5: Heat maps for the ωs for two normal distribution samples. The magnitude of the lengths of pairwise distances is encoded in color.

coded points represent magnitudes for each of the 300 sampled points. Another version of the GW metric is through a pointwise calculation. It is shown in Fig.(5.7). As the final step, we compute the Wasserstein metric in the next section for the case of narrow convergence. The support for both distributions has the same dimensions for this case.

5.4.2 Computing Wasserstein distance

In this section we compute the Wasserstein distance between two distributions. We recall that the Wasserstein metric is a distance between two distributions with the same dimensions and support. This would mean that we need narrow convergence induced by the Wasserstein metric. This setup is referred to as *weak topology*.

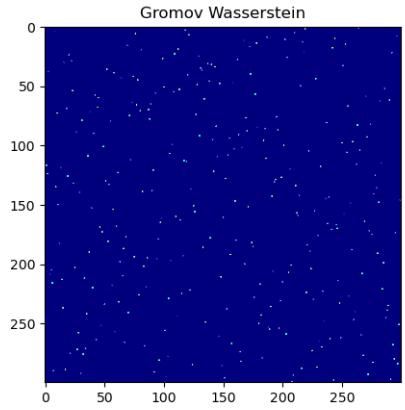


Figure 5.6: The Gromov-Wasserstein metric calculated for each of the 300 sampled points. The colors represent the magnitude of GW pairs. The probability measures are uniform.

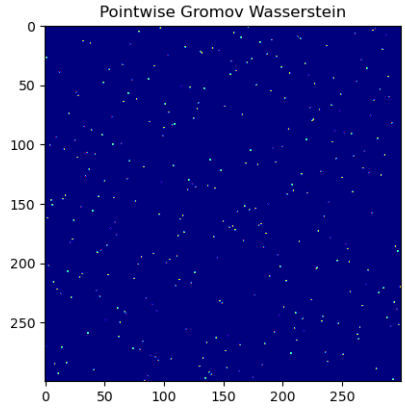


Figure 5.7: The pointwise Gromov-Wasserstein metric calculated for each of the 300 sampled points. The colors represent the magnitude of GW pairs. The probability measures are uniform.

1-D optimal transport with the Wasserstein metric.

We start with two sampling distributions for unfair binomial trails and consider two Gaussian distributions namely $\mathcal{N}(50,10)$ and $\mathcal{N}(100,20)$. The distributions are shown in Fig.(5.8)

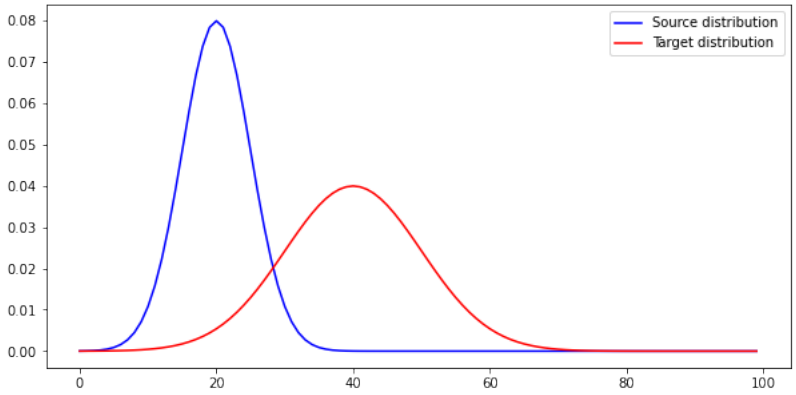


Figure 5.8: The two gaussians between which the Wasserstein metric is to be calculated. The source distribution is blue is $\mathcal{N}(20,5)$ while the target distribution is$\mathcal{N}(40,10)$. The conditions for narrow convergence are satisfied in the metric space containing these distributions.

The next step is to define the metric for the wasserstein quadrature. We do this through a loss matrix which is the normalized matrix made out of distance pairs between each point of the two distributions. In our case we go for the euclidean distance i.e. $M_{ij} = \sqrt{(x_1 - x_2)^2}$. This matrix is always square due to the condition that the distributions are of the same size. The two distributions are shown as two axes along the coordinate axes in Fig.(5.9) with the distance matrix in inset.

Next we calculate the Wasserstein metric by solving the Monge-Kantranovich problem. The problem can be stated as the calculation of an optimal map between two distributions a and b as an optimization problem stated as follows.

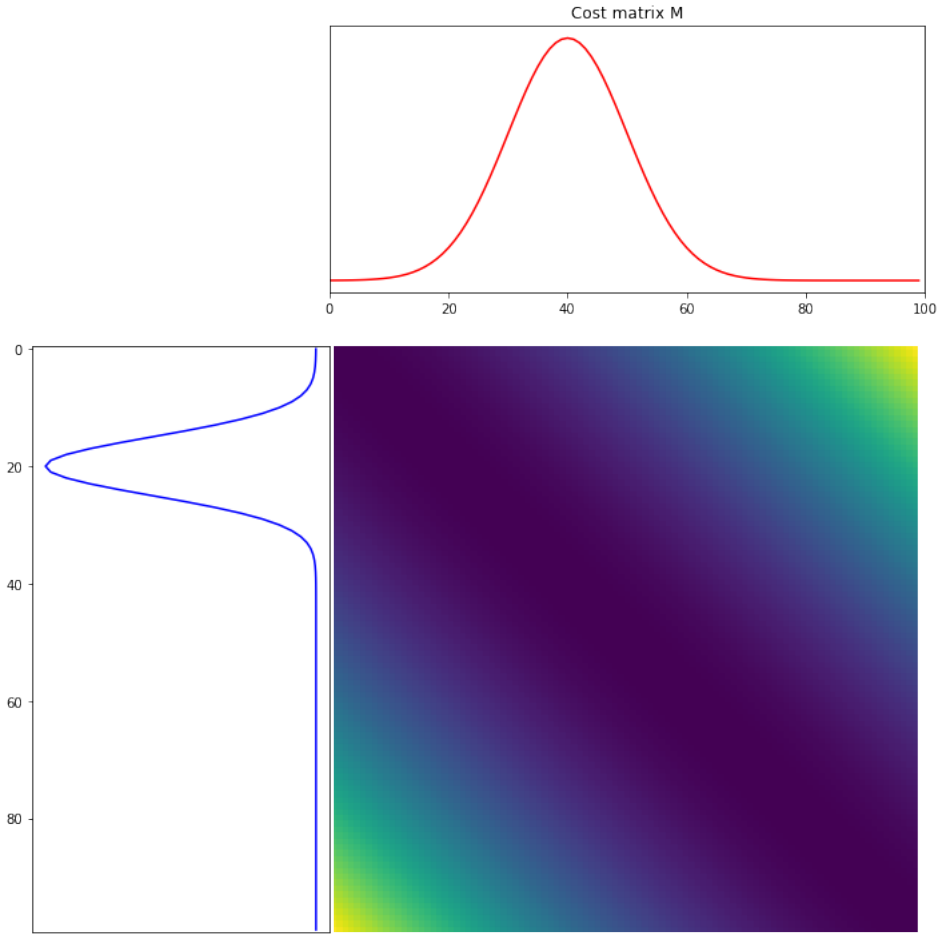


Figure 5.9: The distance matrix **M** shown along with the two normal distributions. The magnitude of the metric i.e. elements of the matrix is represented by the darkness of the color. The area to the left of any point $X = a$ in the distributions is cumulative probability $P(X \leq a)$

$$\gamma = \underset{\gamma}{\operatorname{argmin}} \langle \gamma, M \rangle_F$$

$$\gamma \mathbf{1} = a$$

$$\gamma^T \mathbf{1} = \boldsymbol{b}$$

$$\gamma \geq 0$$

where M is the cost matrix. It is to be noted that the transportation plan is a positive map with its marginals as the source and target distributions.

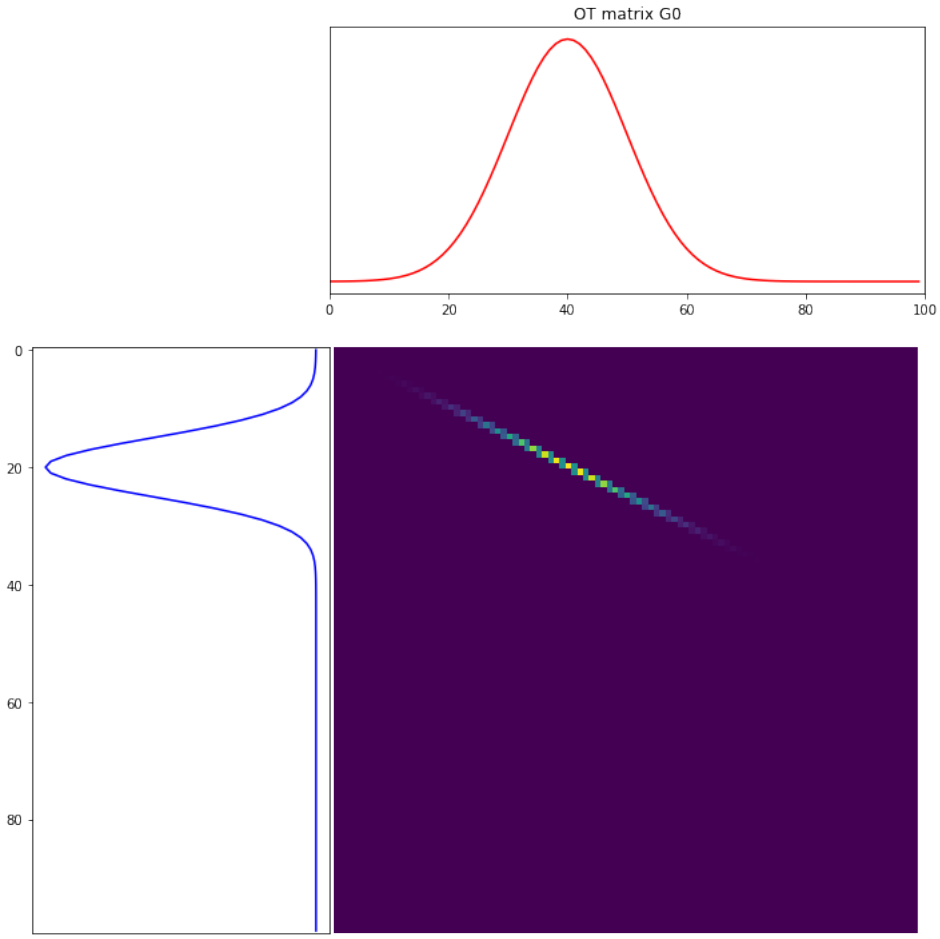


Figure 5.10: The EMD matrix **M** shown along with the two normal distributions. The magnitude of the metric, i.e., elements of the matrix is represented by the darkness of the color. The area to the left of any point $X = a$ in the distributions is cumulative probability $P(X \leq a)$

$2D$ optimal transport between empirical distributions

In real life data, it is often the case that we only have an empirical distribution available to us, the characterization of which is generally not feasible. We look at two such distributions in Fig.(5.11). Our task is to create an optimal coupling set from the blue set to the red

one.

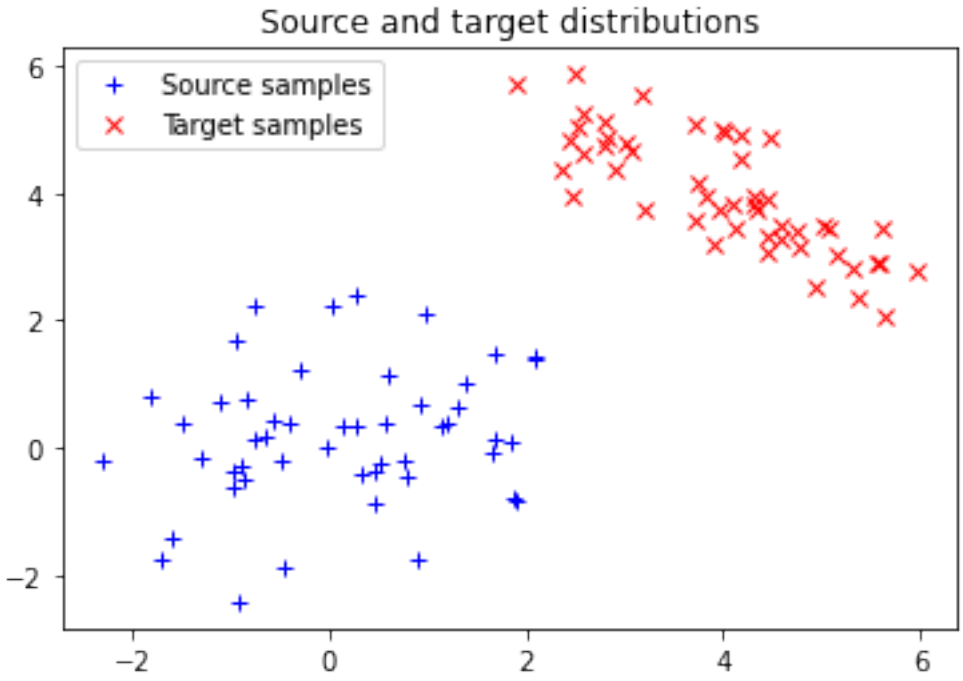


Figure 5.11: The two empirical distributions coming from a data set. We do not use the fact that these are Gaussian in our calculations. The source and target distributions are shown in blue and red.

Since this a $2D$ dataset, we again use the euclidean distances between points as the elements of the cost matrix, shown in Fig.(5.12).

Next, we compute the optimal coupling matrix. The coupling is a map between a pair of points with each from the source and the target distribution. The yellow dots in Fig.(5.13) show the couplings.

Finally, we plot the couplings, i.e., the transport plan along with the source and target coordinates in Fig.(5.14).

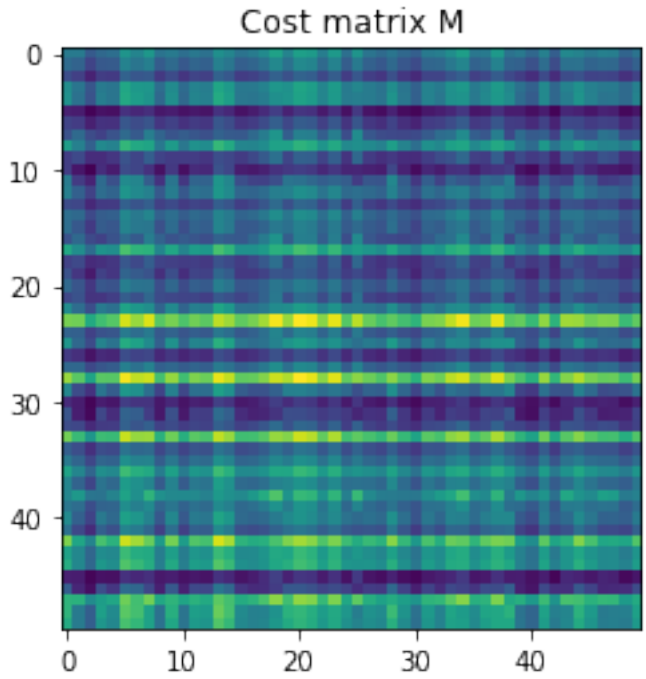


Figure 5.12: The euclidean distance cost matrix for the empirical data set. This is a square matrix as the support of both distributions and their size are the same.

5.5 Synopsis

This chapter had the goal of developing the basic mathematical tools for calculation of Gromov-Wasserstein distances between fully connected asymmetric Ising models. We wrote down the theory for general probability spaces of which the Ising networks are a type. As a first step, we introduced basic concepts of network of metric spaces. This was followed by defining metric measure spaces of the form (X, ω_X, μ_X) which contain a Borel probability measure associated with each node in addition to the functions ω_X. Basic idead about Polish spaces followed with pushforward measures and pullback functions defined to develop the idea of weakly isomorphic network spaces. This was followed by the theory of measure couplings and their compactness was proved through a complete explanation and proof of Prokhorov's theorem. This led the theorem about the existence

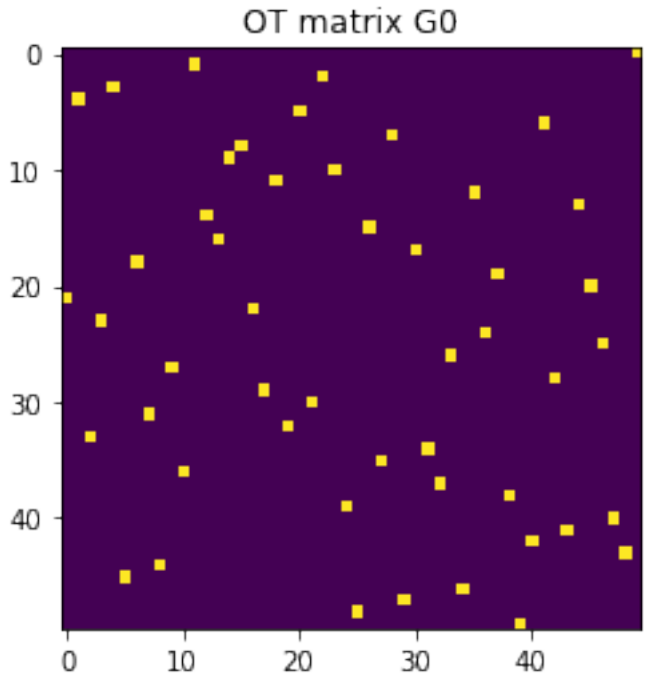


Figure 5.13: The coupling matrix for the empirical distributions. The yellow dots are couplings, i.e., joint distribution maps between the elements.

of optimal couplings by Strum [116], ultimately proving the existence of a minimizer and hence the definition of the GW metric. The final step was to present a few illustrative calculations on the calculation of the GW and the Wasserstein metrics. The tools developed in the this chapter give us the capability to undertake calculations for specific measures and integrands in the next chapter. The concepts discussed will be specifically developed for Ising networks.

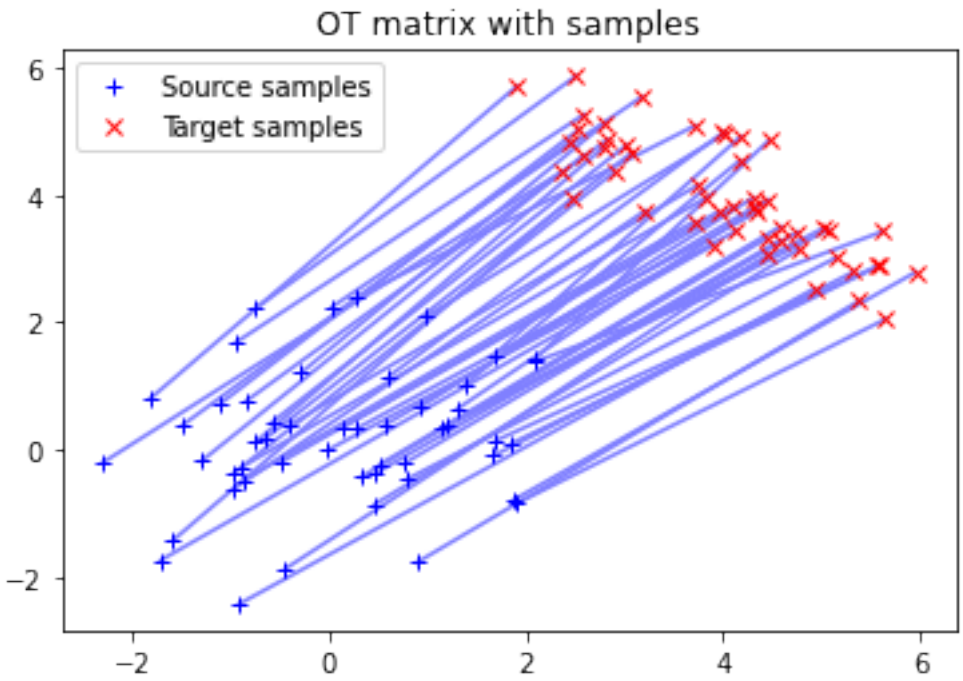


Figure 5.14: The euclidean distance cost matrix for the empirical data set. This is a square matrix as the support of both distributions and their size are the same.

Chapter 6: Optimal driving between network steady states

6.1 Introduction

In the previous chapter, we stated the framework for calculation of Gromov-Wasserstein distance between two networks in a space of measure networks. The goal of this chapter would be to obtain the same for a specific type of system, namely a fully connected asymmetric neural network. This is uncharted territory in terms of statistical mechanics of these highly complex nonlinear systems and no standard characterization of the steady-state distribution exists in general. As a result, a numerical method needs to be constructed to efficiently drive the system between two probability distributions that exist in different dimensions. This constitutes the following steps.

1. Define the network as a finite dimensional Polish space X.

2. Define a Borel probability measure μ_X^i on each node i.

3. Establish the notion of metric between the two steady-state distributions. As we would see later, we don't need the strict definition of a metric d_X. We would still be able to construct transport plans as if the space(manifold) of spaces(measure networks) is a metric space (X, d_X).

4. Construct a geodesic between any two networks $(X, d_X, \mu_X$ and (Y, d_Y, μ_Y) on the space of all metric measure spaces(neural-networks).

The first step is the same as the theoretical framework stated and explained in section 5.2.2 and we refer the reader to it. As for the next two tasks i.e. the node indexed Borel probability measure μ_X and the metric notion d_X, we refer to the concepts in section 5.3. The change here as compared to the general theory is to focus on a specific system i.e. a fully connected asymmetric neural network. Hence, the main goals here would be to specify what quantities would be playing the roles of μ_X and d_X for networks far from equilibrium. We would extend some concepts based in the theory of spin-glasses to accomplish the above. It deems mentioning that the present works considers the definitions of the probability measures and metrics as central to the application of optimal transport theory to nonequlibrium steady states and considers it incorrect to use any bijective mapping and supported probability distribution as d_X and μ_X respectively. The subsequent sections complete with tasks from step 2 onward.

6.2 Borel-Probability measures

The space of vertices is a topological space. The measure construction needs to be such that the measure is defined on all open sets within it. Thus the to be constructed probability measure needs to be defined on the smallest $\sigma-$algebra of X. More precisely,

$$\text{For } (X,d,\mu),\ \mu : B(x) \to (0,1] \text{ such that } \mu(\phi) = 0 \text{ and } B_1, B_2, B_3 \cdots \in B \text{ are disjoint}$$
$$\mu\left(\bigcup_{i=1}^{\infty} Bi\right) = \sum_{i=1}^{\infty} \mu(Bi) \tag{6.1}$$

The concepts developed in section 5.3 would suffice for the mathematical definitions stated above. We now look at what possible quantities would be assigned as μ_X for the measure networks in the space of all networks $[\mathscr{N}]$.

6.2.1 Edwards-Anderson order-parameter for spin-glasses

Is there a measure of node fidelity? That is, if we observe a certain node across microstates pertaining a nonequibrium steady-state, could we measure its tendency to flip across the interval of observation. Such a measure would identify two-phases of a node, i.e., completely stationary and constantly flipping, as the two extremes and everything in between would be assigned an intermediate value characterizing the node by the number of flips which are a characteristic of the steady-state distribution. For example, a close to frozen $T \to 0$ neural network, i.e., kinetically arrested network would barely flip a node while a network in a high-temperature state would be the exact opposite and constantly flip the node with the frequency of flipping as a function of the macrostate parameters of the steady-state. Therefore, fidelity can be an order-parameter. An obvious choice for fidelity is the mean of node states across the arrow of time, i.e., local magnetization

$$m_i = \langle x_i \rangle \ , \tag{6.2}$$

where the angular brackets indicate ensemble average. This measure would assign values close to $+1$ or -1 to vertices that do not undergo changes in their state that often across the interval of measurement as compared to the closer to zero values to the ones that undergo frequent flipping. The role of local magnetization as a measure of node fidelity is quite straightforward but it admits negative values and for that reason is not directly admissible as a Borel probability measure. To that we end, we look at the **Vertex Edwards-Anderson order parameter.**

$$q_i = \langle m_i \rangle^2 \ , \tag{6.3}$$

and define

$$\mu_i = P(q_i) \quad , \tag{6.4}$$

through a histogram. We note that for multiple steady-states, we end up with a simplex

$$\sum_N := \{P(q_i) \in \mathbb{R}^+; \sum_{i=1}^{N} P(q_i) = 1\} \ , \tag{6.5}$$

when all the probability distributions are of size $N, \forall i \in \{1,2,\ldots N\}$. Our goal is to drive between probability distribution of different dimensions, such as $a \in \sum_N, b \in \sum_M$. We couple a and b through a coupling map defined as $\mathscr{C}_{a,b} := \{K \in (\mathbb{R}_+)^{N\times M}; K[\mathbb{1}]_M = a, K^T[\mathbb{1}]_N = b\}$ where, $a \in \sum_N, b \in \sum_M$, which makes it a joint probability distribution of the vertex Edwards-Anderson order parameters for the respective network steady states. We now need a statistical measure of these couplings which would be a relative quantity for the initial and final steady states.

6.2.2 Entropy of a coupling map

The entropy g a $N \times M$ matrix K can be written as

$$H(K) := -\sum_{i=1}^{N}\sum_{j=1}^{M} K_{i,j}\left(\log\left(K_{ij}\right) - 1\right) \ . \tag{6.6}$$

Given that our coupling maps are non-square matrices in general, this is the definition we follow for the coupling maps defined above. This framework generalizes our definition of couplings in a discrete setting for probability distributions of unequal size.

6.3 The cost function: A notion of the metric d_X

The theory of optimal transport does not make any distinction between two metrics d_X and d_Y as long as they are correctly defined. In the present setup, the only requirement we impose on this so called cost function is that it be square-integrable. The reason being that our definition is general for Lebesgue space, l^p but we operate in $p = 2$ for gauge invariance. In addition, we also need to account for statistical properties of real-world neural networks (Ising graphs) and systems approximated by them. Out of those, the most important is the tendency of these systems to undergo continuous phase transitions. This makes it necessary to integrate a square function which is also an order parameter. Finally, We see to it that our integrand is conjugate to the local fidelity as it is defined above. To begin with we define

$$\xi_{ij}^{X} = J_{ij}\langle x_i x_j\rangle - J_{ij}\langle x_i\rangle\langle x_j\rangle \quad . \tag{6.7}$$

This is the weighted two point correlation function. Its asymmetric, is not zero if $i = j$ unless the diagonal of the coupling matrix is zero and it does not obey triangle inequality. This means that Eq.(6.7) is not a metric. In addition, it does not have gauge invariance. As it stands, ξ_{ij} is a measure of edge fidelity. The reason being the ensemble average of energy is going to be very different for two nodes connected by edge J_{ij} if the nodes themselves flip often and simultaneously. This synchronized flipping-makes the first term large and the second term smaller as the mean mode magnetization decreases for both nodes if flipping is frequent.

Simultaneous flipping of two nodes is a measure of the reliability of the edge connecting them. It is to be noted that this is different from coupling magnitude as the coupling matrix is a cumulatively constitutive entity with respect to the network. This means that it acts

a whole entity giving rise to the observed behavior. There is no way to isolate individual edge effects but only observe the cumulative effect of the entire coupling matrix.
Edge reliability on the other hand means the extent to which an edge correlates the behavior of the nodes it connects with respect of its own magnitude. That is, if the edge is high in magnitude then do the nodes demonstrate a strong correlation in their behavior. The actual cause of the correlations in node behavior is indeed intractable in the network setting but all that this measure encodes is a match between the correlation of the node behavior with respect to the edge magnitude. For high-frequency flipping, this prompts four cases.

1. **Large edge-magnitude; high-frequency synchronous flipping**: In this case edge fidelity is large and edge is said to be reliable. The second term vanishes but the first term approaches the edge magnitude.

2. **Large edge-magnitude; high-frequency asynchronous flipping**: Both the first and second term vanish as both the ensemble averages go to zero.

3. **Small edge-magnitude; high-frequency synchronous flipping**:The first term is larger and approaches the edge magnitude while the second term vanishes. The edge is unreliable as there is synchronous flipping even with a small coupling magnitude.

4. **Small edge-magnitude; high-frequency asynchronous flipping**: Both terms vanish as the ensemble averages go to zero. The edge is reliable as the flipping is asynchronous which is consistent with the small edge magnitude/coupling constant.

There is a range of intermediate behaviors possible as far a *edge-reliability* is concerned.

6.4 Optimal driving between nonequilibrium steady-states of unequal dimensions

The Borel probability measure for node-fidelity and the edge-reliability function are conjugate order-parameters. When node-fidelity changes, the edge-reliability varies in conjunction as it is dependent on the frequency of flipping. Our system is in a non equilibrium steady-state so linear-response theory is not applicable. but we can still attempt to write an integral

$$H_{x,y} = \int_{X\times Y}\int_{X\times Y} \left(\xi^X_{ij} - \xi^Y_{kl}\right)^2 d\mu\left(x_i, y_k\right) d\mu\left(x_j, y_l\right) \quad . \tag{6.8}$$

This is the cost of going from state X to state Y. For minimal dissipation, an infimum of this integral is required. It can be seen that such a minimization gives the Gromv-Wasserstein metric. We now state the tensor notation and then arrive at the metric.

6.4.1 Tensor notation

We define the entropy generation in driving from a network X to a network Y in the discrete setting as

$$H_{X,Y} = \langle \left(\xi^X - \xi^Y\right)^2, \mu_{x,y} \rangle \quad . \tag{6.9}$$

The edge-reliabilities are square asymmetric materials and the square of their differences is a 4-tensor. We define

$$\Gamma_{ijkl} := \left(\xi^X_{ij} - \xi^Y_{kl}\right)^2 \quad . \tag{6.10}$$

With this notation, the minimum entropy generation path is defined as

$$d_{X,Y} = \min_{\mu_{x,y} \in C_{x,y}} \sum_{ijkl} \Gamma_{ijkl} \mu^{ik}_{x,y} \mu^{jl}_{x,y} \quad . \tag{6.11}$$

We can write d_{XY} through product notation as

$$d_{X,Y} = \min_{\mu_{x,y} \in C_{X,Y}} \langle \Gamma(\xi_{i,j}, \xi_{k,l}) \otimes \mu_{x,y}, \mu_{x,y} \rangle \quad \forall i,j,k,l \in X,Y \quad , \tag{6.12}$$

where $\Gamma\left(\xi_{i,j}, \xi_{k,l}\right) = \left(\xi^X_{i,j} - \xi^Y_{k,l}\right)^2$ for our gauge-invariant case of L^2. This leads to efficient computation due to the following result by Solomon et al. [93].

Theorem 6.1. *If the loss function i.e.* Γ *can be written as*

$$\Gamma(a,b) = f_1(a) + f_2(b) - h_1(a)\, h_2(b) \quad , \tag{6.13}$$

for functions (f_1, f_2, h_1, h_2)*, then for any* $\mu \in \mathscr{C}_{X,Y}$

$$\Gamma\left(\xi^X_{ij}, \xi^Y_{k,l}\right) \otimes \mu_{x,y} = f_1(\xi^X_{ij}) \mu_X 1^T_N + 1_M \mu^T_Y f_2\left(\xi^Y_{kl}\right) - h_1(\xi^X) \mu_{x,y} h_2(\xi^Y)^T \quad , \tag{6.14}$$

giving a loss function independent of $\mu_(X,Y)$ *; the joint distribution which is the transport map or the network-flow map and dependent only on the marginals. (These are the node-fidelity vectors in our case))*

Proof. Lets compute the tensor product term by term. First we have

$$\sum_j f_1\left(\xi^X_{i,j}\right)\sum_l \mu^{j,l}_{x,y} = \left(f_1(\xi^X)\left(\mu_X\right)\right)_i = A_{ij} \ .$$

Next

$$\sum_l f_2\left(\xi^Y_{k,l}\right)\sum_j \mu^{j,l}_{x,y} = \left(f_2(\xi^Y)\left(\mu_Y\right)\right)_j = B_{ij} \ ,$$

and the cross term

$$\sum_j h_1\left(\xi^X_{ij}\right)\sum_l h_2\left(\xi^Y_{k,l}\right)\mu_{jl} = \left(h_1\left(\xi^X\right)\left(h_1(\xi^Y)\,\mu^T_{x,y}\right)^T\right)_{ij} = C_{ij} \ .$$

This proves the result. □

Remark: If the original form of the problem is considered , then we have the computation in $O\left(M^2N^2\right)$ but if we use Eq.(6.14), then we have $O\left(M^2N+N^2M\right)$. We use this fact to formulate a matrix form of the tensor notation.

6.4.2 Matrix formulation of the tensor product

We use marginalization to reduce the indices of the tensor, resulting in a more tractable computational form.

Lemma 6.1. *[28, 93]*

$$d_{X,Y} = \min_{\mu_{x,y}\in C_{x,y}} \sum_{ijkl} \Gamma_{ijkl}\mu^{ik}_{x,y}\mu^{jl}_{x,y} = \left(\langle \mu_X, \xi_X.^{\wedge 2}\mu_X\rangle + \langle \mu_Y, \xi_Y.^{\wedge 2}\mu_Y\rangle - 2Tr(C^T\xi_X^T C\xi_Y)\right) \ , \tag{6.15}$$

where C is the set of measure couplings.

Proof. First we write the marginals as

$$P(q_X) := p, \text{ and}$$
$$P(q_Y) := q \ .$$

The GW metric is

$$d_N(X,Y) = \frac{1}{2} \min_{C \in \mathscr{C}(p,y)} \left(\sum_{ijkl} \left| \xi_{ik} - \xi_{jl} \right|^2 C_{kl} C_{ij} \right)^{\frac{1}{2}} \ . \tag{6.16}$$

After expanding the square we have three terms and we write them separately as

$$\sum_{ijkl} \xi_{ik}^2 C_{kl} C_{ij} = \sum_{ik} \xi_{ik}^2 p_k p_i = \sum_i p_i \sum_k \xi_{ik}^2 p_k = \sum_i p_i \left(\xi_{X\cdot}^{\wedge 2} p \right)_i = \langle p, \xi_{X\cdot}^{\wedge 2} p \rangle \ ,$$

and

$$\sum_{ijkl} \xi_{jl}^2 C_{kl} C_{ij} = \sum_{jl} \xi_{jl}^2 q_l q_j = \sum_l q_l \sum_j \xi_{jl}^2 q_j = \sum_l q_l \left(\xi_{Y\cdot}^{\wedge 2} q \right)_l = \langle q, \xi_{Y\cdot}^{\wedge 2} q \rangle \ .$$

Finally the $2XY$ term is written as

$$-2 \sum_{ijkl} \xi_{ik} \xi_{jl} C_{kl} C_{ij} = -2 \sum_{il} \left(\xi_X C \right)_{ik} \left(C \xi_Y \right)_{il} = -2 \langle \xi_X C, C \xi_Y \rangle = -2 Tr \left(C^T \xi_X^T C \xi_Y \right) \ .$$

The last equality is due to the Frobenious norm. □

6.4.3 Entropic regularization

We define the entropy of the transport map as

$$H\left(\mu_{X,Y}\right) := -\sum_{i=1}^{N}\sum_{j=1}^{M}\mu_{i,j}\left(\log\left(\mu_{ij}\right)-1\right) \quad . \tag{6.17}$$

Where the map is from this steady state X to Y. The elements $\mu_{X,Y}$ denote the joint probability distribution with marginals μ_X and μ_Y. The node-fidelity histograms are mapped optimally through the infimum of the inner-product.

$$H_{X,Y} = \langle\left(\xi^X - \xi^Y\right)^2, \mu_{x,y}\rangle \quad , \tag{6.18}$$

giving the GW metric as

$$d_{X,Y} = \underset{\mu_{x,y}\in C_{X,Y}}{\operatorname{argmin}}\left(H_{X,Y}\right) \quad . \tag{6.19}$$

This nonconvex optimization problem is entropically regularized as

$$d_{X,Y} = \underset{\mu_{x,y}\in C_{X,Y}}{\operatorname{argmin}}\left(H_{X,Y}\right) - \kappa H(\mu_{X,Y}) \quad . \tag{6.20}$$

Here κ is a user defined constant of regularization. The solution to this problem reads

$$d_{X,Y} = AKB \quad \text{where} \quad K := e^{\frac{\left(\xi^X-\xi^Y\right)^2}{\varepsilon}} \in \mathbb{R}_+^{M\times N} \tag{6.21}$$

The rectangular matrix K is called the Gibbs kernel of the cost matrix. The matrices A and B are the solutions of the Sinkhorn iterations [34].

$$A = \frac{\mu_x}{KB} \tag{6.22}$$

$$B = \frac{\mu_y}{K^T A} \tag{6.23}$$

The division is component-wise. Having a algorithm for the calculation of the GW metric, we now derive optimal paths between two network steady-states. Before that, an issue of steady-state equivalence needs to be addressed. As it was shown in chapter 4, the equivalence classes for Ising models are descibed by their weak isomorphisms. We formalize the construction as follows.

6.4.4 Equivalent steady-states with different dimensions

The next step is to identify which networks, characterized by their steady-states, are equivalent to each other. The main aim here is to develop a method based on such an equivalence , which makes is possible for a network in a particular steady state to change its number of vertices. For this task we call upon the concept of weakly isomorphic network spaces defined in Chapter 4. In particular for two network spaces (X, ξ_X, μ_X) and (Y, ξ_Y, μ_Y). we define a common expansion network (Z, μ_Z) with the probability map

$$\pi_X : Z \to X \quad , \tag{6.24}$$

$$\pi_Y : Z \to Y \quad , \tag{6.25}$$

with the corresponding push forward measures

$$(\pi_X)_{\#}\mu_Z = \mu_X \quad , \tag{6.26}$$

$$(\pi_Y)_{\#}\mu_Z = \mu_Y \quad , \tag{6.27}$$

and pullback functions

$$(\pi_X)^* \xi_X\left(x_i, x_j\right) = \xi_X\left(\pi_X\left(x_i\right), \pi_X\left(x_j\right)\right) \quad , \tag{6.28}$$

$$(\pi_Y)^* \xi_Y(y_k, y_l) = \xi_Y\left(\pi_X\left(y_k\right), \pi_Y\left(y_l\right)\right) \quad , \tag{6.29}$$

satisfying the identity

$$\left\|(\pi_X)^* \xi_X - (\pi_Y)^* \xi_Y\right\|_\infty = 0 \quad , \tag{6.30}$$

giving $d_{XY} = 0$ and making the two network steady-states X and Y weakly isomorphic ,i.e., $X \cong^w Y$. Weak isomorphisms provide us the ability to the coupling probabilities between two networks to measure preserving maps between networks. This is because the new set of nodes redistribute the Edwards- Anderson parameters into a common space (push forward measures) while simultaneously defining a corresponding square integrable functions (pullback functions). Note that this makes the space of spaces a pseudo metric space as the condition $d_{X,Y} = 0$ requires the networks X and Y to be weakly isomorphic. The method we use to access the dimension change for an Ising model is that of blowup and alignment proposed by Chowdhry and Needham [28]. The following definitions are required.

Definition 6.1 (Blowup)**.** Let (X, ξ_X, μ_X) and (Y, ξ_Y, μ_Y) be two Ising networks. Let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ and define a vector $\boldsymbol{u} := (u_x)_{x \in X}$ with $u_x := |\{y \in Y : \mu(x,y) > 0\}|$. Similarly

$\boldsymbol{v} := (v_y)_{y \in Y}$ with $v_y := |\{x \in X : \mu(x,y) > 0\}|$. A union of the indices

$$X[\boldsymbol{u}] := \bigcup_{x \in X} \{(x,i) : 1 \leq i \leq u_x\} \quad ,$$

and

$$Y[\boldsymbol{v}] := \bigcup_{y \in Y} \{(x,j) : 1 \leq j \leq v_y\} \quad .$$

Next, we define the blowup edge-reliability functions. First we do it for (X, ξ_X, μ_X). Fix $x \in X$ and let $y' := \{y_1, y_2, \ldots, y_{u_x}\} \in Y$ such that $\mu(x,y') > 0$ with $\mu_{X[u]}(x,i) := \mu(x,y_i)$. This takes care of the probability measures. For the square integrable functions we define the index sets with $l, m \in X$ as $1 \leq i \leq u_l$ and $1 \leq j \leq u_m$ and define

$$\xi_{X[\boldsymbol{u}]}((l,i),(m,j)) = \xi_X(l,m) \quad , \tag{6.31}$$

giving $\left(X[\boldsymbol{u}], \xi_{X[\boldsymbol{u}]}, \mu_{X[\boldsymbol{u}]}\right)$ as the blown up Ising network for (X, ξ_X, μ_X).
Similarly, $(Y[\boldsymbol{v}], \xi_{Y[\boldsymbol{v}]}, \mu_{Y[\boldsymbol{v}]})$ can be defined for (Y, ξ_Y, μ_Y).

By the very definition, the blowups are weakly isomorphic to their original networks. The next step after expanding a network is to redistribute its probability measures along the support of the target vector distribution. This is done as follows.

Definition 6.2 (Alignment)**.** A permutation matrix is a square binary matrix for which exactly one entry of each row and each column is one and the rest are zeros. Let X and Y be networks with m and n nodes. We binarize the $m \times n$ matrix μ of optimal couplings between them . The binarized matrix, denoted as $\mathbf{1}_{\boldsymbol{\mu}>\mathbf{0}}$ has the value 1 wherever $\mu_{XY} > 0$ and zeros elsewhere . Let $\hat{X}$ and $\hat{Y}$ be the blown up versions of X and Y. Let $\mathbf{1}_{\hat{\mu}>0}$ be the new binarization matrix and in this case we intentionally define it as a permutation matrix.

Then, we assign new values to $\hat{Y}$ and $\hat{\mu}$ as

$$\hat{Y} \leftarrow \mathbf{1}_{\hat{\mu}>0} \hat{Y} \mathbf{1}_{\hat{\mu}>0}^{T} \text{ and } \hat{\mu} \leftarrow \mathbf{1}_{\hat{\mu}>0} \hat{\mu} \ , \tag{6.32}$$

giving $\hat{\mu} = diag\left(\mu_{\hat{X}}\right)$. Finally we get

$$d_N\left(\hat{X},\hat{Y}\right) = \sum_{i,j=1}^{n} \left|\xi_{\hat{X}}\left(x_i,x_j\right) - \xi_{\hat{Y}}\left(yi,yj\right)\right|^2 \mu_{\hat{X}}\left(x_i\right)\mu_{\hat{X}}\left(xj\right) \ . \tag{6.33}$$

We are now in a position to derive the functions on equivalence classes in metric measure spaces which would be geodesics.

6.4.5 Geodesics between two networks far from equilibrium

Geodesics between two network spaces $(X,\xi_X,\mu_X$ and (Y,ξ_Y,μ_X) are defined through their isomorphism classes $[X]$ and $[Y]$ in the space of network spaces $[N]$. A geodesic from $[X]$ to $[Y]$ in $[N]$ is a continuous map γ such that $\gamma(0) = [X]$ from a closed interval $[0,T]$ into $[N]$. The following relation holds

$$d_N\left(\gamma(0),\gamma(t)\right) = \frac{|t-S|}{T} \cdot d_N\left([X],[Y]\right) \quad \forall s,t \in [0,T] \ . \tag{6.34}$$

Geodesics by the above definition are not unique but can always be constructed. We now look at constructing γ for $[X],[Y] \in [N]$. Let $\mu_{X,Y}$ be the optimal coupling matrix. For each $t \in [0,T]$ we define $\gamma(t) := [X \times Y, \Xi_t, \mu_{X\times Y}]$, which is a function on the space of spaces . We have define the counterpart of the edge-reliability function for the geodesics as

$$\Xi_t\left(\left(x_i,y_k\right),\left(x_j,y_l\right)\right) := (1-t)\,\xi_X\left(x_i,x_j\right) + t\xi_Y\left(y_k,y_l\right) \ . \tag{6.35}$$

The above form is chosen because $\gamma(1) = Y$. Note that $\gamma(t)$ is defined on $X \times Y$ while Ξ is always defined on $X \times Y \times X \times Y$. With the theoretical tools available to us for computing geodesics, we now do so for two fully connected Ising graphs.

6.5 Numerical results

We follow [28] to give a $m+n$ size function for the geodesic instead of the expected $m \times n$ for two networks X and Y of sizes m and n respectively. The advantage of their method is that it does not render the computation intractable for moderately large networks. The computation here uses ξ_X and ξ_Y for calculating Ξ at different values of t. Also, only the marginal distributions μ_X and μ_Y are needed. With only the order-parameters needed as the initial inputs for both X and Y. What would be available to us as observables are the values of μ as the joint distribution while the geodesics are constructed as a network flow process. We summarize the results in three heat maps. First we plot the initial and final coupling matrices in Fig.(6.1). Note that the two matrices are asymmetric.

Next, we plot the edge-reliability matrices for both initial (X) and final (Y) steady states. The matrices ξ_X and xi_Y are ensemble averages by definition. As shown in Fig.(6.2), the matrices are asymmetric due to them being Hadamard products of the covariance and coupling matrices with the latter being asymmetric unlike the former which is symmetric by definition. Finally, we plot the geodesic $\gamma(t)$ with 10 iteration steps. The function $\gamma(t)$ is defined on the space of network spaces with X and Y being the initial and final points. We begin with a network weakly isomorphic to X at $t = 0$ and proceed to Y at $t = T$. The size of the geodesic function is $Z+Y$ where Z is isomorphic to X but is the same size as Y. Fig.(6.3) shows the mapping γ at different points during the path of least dissipation. This is an optimal control algorithm far from equilibrium. We note that edge-reliability is a type of coupling weighted variance function. So this algorithm would avoid critical points as

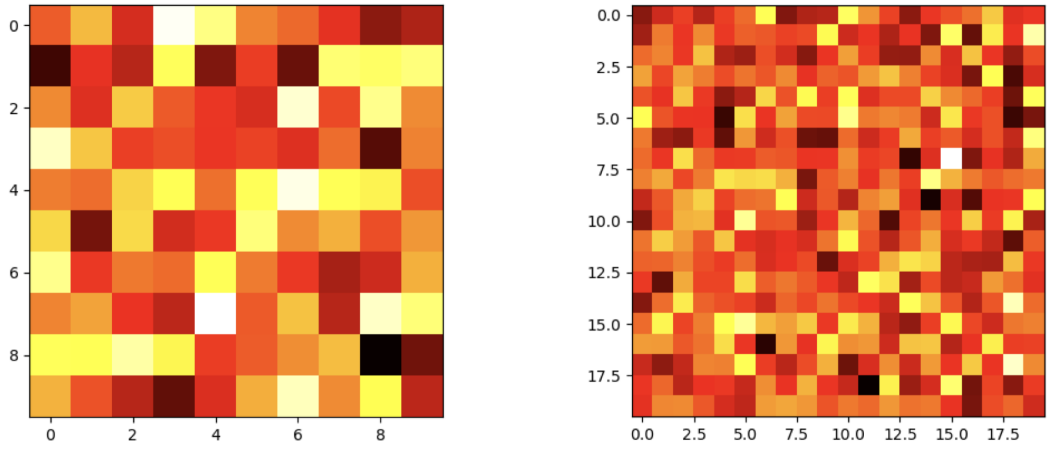


Figure 6.1: Heat-maps of initial (J_X) and final (J_Y) coupling matrices on the left and right respectively. Both the square matrices are asymmetric. J_X is a 10×10 matrix whereas J_Y is 20×20.

the Gromov-Wasserstein metric is an infimum of an integral by definition.

6.6 Synopsis

After discussing the mathematical machinery required for driving between network nonequilibrium steady states of different dimensions in chapter 5, this chapter focused on the development of Gromov-Wasserstein metrics specific to Ising models far from equilibrium. The development of algorithms for calculating the GW metrics have mainly concentrated on machine learning and shape classification [28, 93]. There has been a absence of specific physics-informed square integrable functions and probability measures. For the fully connected asymmetric Ising model, we proposed the edge-reliability function as the covariance of node states along each edge along with the histogram of the Edwards-Anderson order parameter as the Borel probability measure. The GW metric we propose is the infimum of the integral of the reliability with respect to the joint distribution of the EA

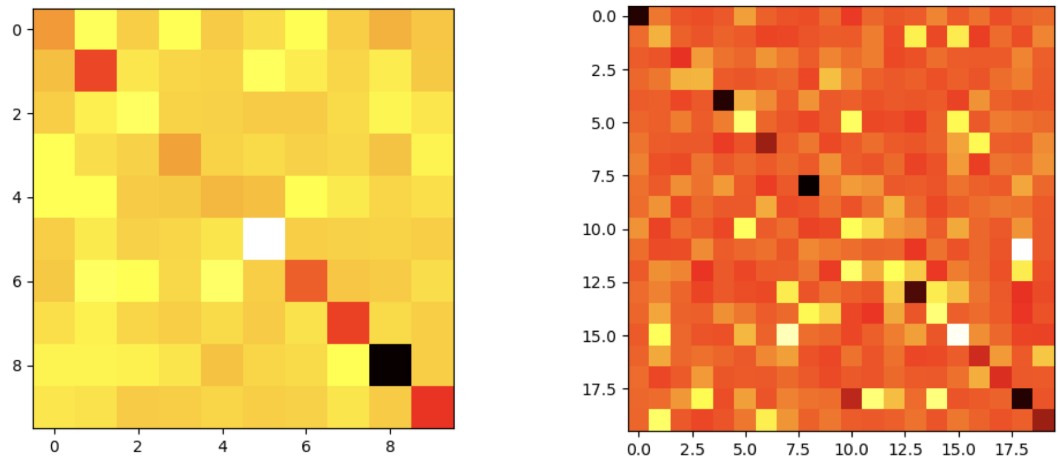


Figure 6.2: Heat-maps of initial (ξ_X) and final (ξ_Y) edge-reliability matrices on the left and right respectively.

parameter on the space of networks of the form (X, ξ_X, μ_X). The calculation is performed on the matrix form of the tensor product as shown in Eq.(6.4.3).

The aim is to drive optimally between two network steady states of unequal dimensions i.e. two networks of different sizes far from equilibrium. An essential task therefore is to define equivalence classes across dimensions throughout the pseudo-manifiold[1] of networks. We follow [28] to define blow-ups and alignments to construct Ising isomorphism classes through a redistribution of ξ and μ for two networks through a permutation matrix (see definitions 6.1 and 6.2). With this framework, we define the geodesics on the space of Ising networks. The calculations are presented in Fig.(6.1), Fig.(6.2) and Fig.(6.3).

[1]The space of network spaces is quotiented by the equivalence classes of weakly isomorphic networks $(X/\cong^w)$ and is therefore an **Riemannian orbifold** [19, 28].

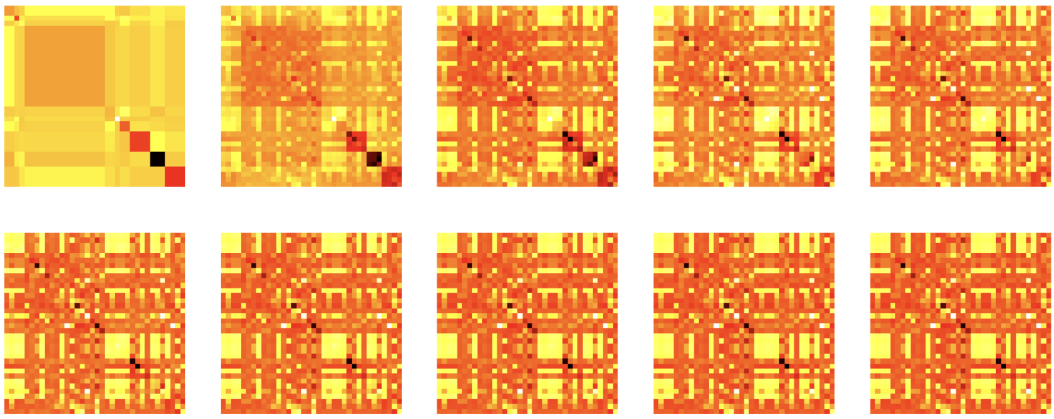


Figure 6.3: A sequence of ten heat maps of the geodesic function $\gamma(t)$. The geodesic is of size $\cong^w X + Y = Y + Y = 40$, where by $\cong^w X$ we indicate the blown up and realigned version of X with the size of Y, i.e., 20. The map is from $X \times Y \to \cong^w X + Y$. The number of iterations can be increased for a finer trajectory.

Chapter 7: Epilogue: A discussion about the results

The work presented in this thesis applies statistical learning and optimal transport to biological networks. The results can be classified into two main categories.

1. Inference of coupling matrices of fully connected asymmetric Ising models.

2. Optimal driving between nonequilibrium steady states of unequal dimensions.

In chapter-2, we discuss fundamental concepts about biological networks. Beginning with the binary classification of directed and undirected networks, we state how biological networks should be considered as directed. This leads to the idea of asymmetric coupling matrices due to lack of detailed balance and presence of active transport in biological systems. The idea of precision medicine is then explored in conjunction with accurate and precise *in-vivo* edits. We propose that solution to inverse problems and optimal network control are of significance to the success these strategies.

The solution to the inverse problem for the kinetic Ising model has been achieved with the availability of time-series data [64, 132, 134]. The likelihood maximization is computationally intense and the use of latent variables was recently introduced [26, 40, 51, 96]. We reproduced the results for time-series data using latent variables in chapter-3 and even more significnatly, illustrated the inner workings of the latent model by illustrating a P$\tilde{o}$lya-Gamma based Gibbs-sampling procedure. This has applications to phase transitions as it separates distributions in a mixture.

In chapter-4 we proposed a likelihood function based on orbits of the Markov chain gov-

erning the asymmetric Ising model. We proved that this likelihood is a minimally sufficient statistic and showed the results for the solution of the inverse problem. To our best knowledge, the solution to the inverse problem far from equilibrium and without time-series data has only been successfully achived by two other studies [37, 38]. Our method uses group theory and can accommodate parallel computing for more efficient computation.

In chapter-5, we follow the ideas from Strum [116] and Chowdhury [27] to write the mathematical machinery needed for optimal driving between nonequilibrium steady states for Ising models and unequal dimensions. We start with the basic concepts of metric measure spaces and look at weak isomorphisms as a tool to change the number of vertices in a networks while maintaining a notion of equivalence between them. We write down the necessary conditions for the Gromov-Wasserstein distance to be used for optimal driving.

In chapter-6, we extend these ideas to asymmetric Ising networks by defining *edge-reliability* and *node-fidelity* as order parameters for the network. The integral for entropy generation when written using these directly results in a Gromov-Wasserstein distance. We then derive the geodesics for these networks using the ideas in [28] while maintaining a connection to the thermodynamics of the network through the order parameters.

This work presents solutions to network inference and control from statistical learning and optimal transport respectively. There have been other approaches [2, 50, 79, 104, 119] to both these problems over the past two decades and the ones presented here are focused on biological networks. The true dynamics of a biological network will have network growth simultaneously occurring with variations in the coupling matrix. The next steps in our opinion should be aimed at designing comprehensive methods that combine the statistics and dynamics of the network and consequently unifying the the optimal transport and statistical learning through a single line of approach rather than a sequential workflow as the present work does. We hope that the methods and results in this document serve as an adequate initial point for the same. We thank the body of work that preceded our study and made

it possible for us to design the numerical methods and obtain the results. The accuracy and precision of future works on biological network control will be the determinants of the eventual success of precision medicine. We wish our colleagues who will undertake this task our very best.

Bibliography

[1] Atte Aalto, Lauri Viitasaari, Pauliina Ilmonen, Laurent Mombaerts, and Jorge Gonçalves. Gene regulatory network inference from sparsely sampled noisy data. *Nature communications*, 11(1):1–9, 2020.

[2] Chaguki T Abdallah and Herbert G Tanner. Complex networked control systems: Introduction to the special section. *IEEE Control Systems Magazine*, 27(4):30–32, 2007.

[3] Eric Alm and Adam P Arkin. Biological networks. *Current opinion in structural biology*, 13(2):193–202, 2003.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[5] M. A. Armstrong. *Groups and symmetry*, volume Undergraduate texts in mathematics. Springer, 2010.

[6] Erik Aurell, Carlos Mejía-Monasterio, and Paolo Muratore-Ginanneschi. Optimal protocols and optimal transport in stochastic thermodynamics. *Physical review letters*, 106(25):250601, 2011.

[7] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[8] Christopher E Barbieri, Sylvan C Baca, Michael S Lawrence, Francesca Demichelis, Mirjam Blattner, Jean-Philippe Theurillat, Thomas A White, Petar Stojanov, Eliezer Van Allen, Nicolas Stransky, et al. Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. *Nature genetics*, 44(6):685–689, 2012.

[9] Marc Barthelemy. *Morphogenesis of spatial networks*. Springer, 2018.

[10] Alan F. Beardon. *Algebra and geometry*. Cambridge University Press, 2005.

[11] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[12] John J Benedetto and Wojciech Czaja. *Integration and modern analysis*, volume 15. Springer, 2009.

[13] Patrick Billingsley. Convergence of probability measures john wiley & sons. *INC, New York*, 2(2.4), 1999.

[14] Benoit G Bruneau. The developmental genetics of congenital heart disease. *Nature*, 451(7181):943–948, 2008.

[15] Nicolas E Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9):5136–5141, 2003.

[16] Michael F Buckley, KJ Sweeney, JA Hamilton, RL Sini, DL Manning, RI Nicholson, CK Watts, EA Musgrove, RL Sutherland, et al. Expression and amplification of cyclin genes in human breast cancer. *Oncogene*, 8(8):2127–2133, 1993.

[17] Ali Cakmak and Gultekin Ozsoyoglu. Mining biological networks for unknown pathways. *Bioinformatics*, 23(20):2775–2783, 2007.

[18] Peter J. Cameron. *Introduction to algebra*. Oxford University Press, 2008.

[19] Francisco C Caramello Jr. Introduction to orbifolds. *arXiv preprint arXiv:1909.08699*, 2019.

[20] Nathan C. Carter. *Visual group theory*, volume Classroom resource materials. Mathematical Association of America, 2009.

[21] Chung-Ta Chang, Tin-Yun Ho, Ho Lin, Ji-An Liang, Hui-Chi Huang, Chia-Cheng Li, Hsin-Yi Lo, Shih-Lu Wu, Yi-Fang Huang, and Chien-Yun Hsiang. 5-fluorouracil induced intestinal mucositis via nuclear factor-κb activation by transcriptomic analysis and in vivo bioluminescence imaging. *PloS one*, 7(3):e31808, 2012.

[22] Alain Chédotal and Linda J Richards. Wiring the brain: the biology of neuronal guidance. *Cold Spring Harbor perspectives in biology*, 2(6):a001917, 2010.

[23] Johanna Chiche, M Christiane Brahimi-Horn, and Jacques Pouysségur. Tumour hypoxia induces a metabolic shift causing acidosis: a common feature in cancer. *Journal of cellular and molecular medicine*, 14(4):771–794, 2010.

[24] Seung Woo Cho, Sojung Kim, Yongsub Kim, Jiyeon Kweon, Heon Seok Kim, Sangsu Bae, and Jin-Soo Kim. Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickases. *Genome research*, 24(1):132–141, 2014.

[25] Hee Min Choi and James P Hobert. The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

[26] Davin Choo, Yuval Dagan, Constantinos Daskalakis, and Anthimos Vardis Kandiros. Learning and testing latent-tree ising models efficiently. *arXiv preprint arXiv:2211.13291*, 2022.

[27] Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

[28] Samir Chowdhury and Tom Needham. Gromov-wasserstein averaging in a riemannian framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 842–843, 2020.

[29] Elena Cid and M Liset. Methods for single-cell recording and labeling in vivo. *Journal of Neuroscience Methods*, 325:108354, 2019.

[30] Dan Cohen. Computer simulation of biological pattern generation processes. *Nature*, 216(5112):246–248, 1967.

[31] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[32] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.

[33] David Benjamin Turitz Cox, Randall Jeffrey Platt, and Feng Zhang. Therapeutic genome editing: prospects and challenges. *Nature medicine*, 21(2):121–131, 2015.

[34] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[35] Charles Darwin and Francis Darwin. *Autobiography and selected letters*, volume 479. Courier Corporation, 1958.

[36] Andreas Dechant and Yohei Sakurai. Thermodynamic interpretation of wasserstein distance. *arXiv preprint arXiv:1912.08405*, 2019.

[37] Simon L Dettmer and Johannes Berg. Inferring the parameters of a markov process from snapshots of the steady state. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(2):023403, 2018.

[38] Simon L Dettmer, H Chau Nguyen, and Johannes Berg. Network inference in the nonequilibrium steady state. *Physical Review E*, 94(5):052116, 2016.

[39] Nadezhda T Doncheva, Yassen Assenov, Francisco S Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nature protocols*, 7(4):670–685, 2012.

[40] Christian Donner and Manfred Opper. Inverse ising problem in continuous time: A latent variable approach. *Physical Review E*, 96(6):062104, 2017.

[41] David S. Dummit and Richard M. Foote. *Abstract algebra*. Wiley, 3rd ed edition, 2004.

[42] Scott R Eliason. *Maximum likelihood estimation: Logic and practice*. Number 96. Sage, 1993.

[43] Akl C Fahed, Bruce D Gelb, JG Seidman, and Christine E Seidman. Genetics of congenital heart disease: the glass half empty. *Circulation research*, 112(4):707–720, 2013.

[44] Yuriy Fedorov, Emily M Anderson, Amanda Birmingham, Angela Reynolds, Jon Karpilow, Kathryn Robinson, Devin Leake, William S Marshall, and Anastasia Khvorova. Off-target effects by sirna can induce toxic phenotype. *Rna*, 12(7):1188–1196, 2006.

[45] Paul David Feigin. Maximum likelihood estimation for continuous-time stochastic processes. *Advances in Applied Probability*, 8(4):712–736, 1976.

[46] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[47] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[48] Kapil G Gadkar, Rudiyanto Gunawan, and Francis J Doyle. Iterative approach to model identification of biological networks. *BMC bioinformatics*, 6(1):1–20, 2005.

[49] Gerd Gaiselmann, Ralf Thiedmann, Ingo Manke, Werner Lehnert, and Volker Schmidt. Stochastic 3d modeling of fiber-based materials. *Computational Materials Science*, 59:75–86, 2012.

[50] Jianxi Gao, Yang-Yu Liu, Raissa M D'souza, and Albert-László Barabási. Target control of complex networks. *Nature communications*, 5(1):1–8, 2014.

[51] Surbhi Goel. Learning ising and potts models with latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 3557–3566. PMLR, 2020.

[52] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[53] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[54] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[55] João Guerra and Inês Lynce. Reasoning over biological networks using maximum satisfiability. In *International conference on principles and practice of constraint programming*, pages 941–956. Springer, 2012.

[56] Jaroslav Hájek. On basic concepts of statistics. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 139–162. Univ. California Press, Berkeley, Calif., 1967.

[57] Paul R Halmos and Leonard J Savage. Application of the radon-nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20(2):225–241, 1949.

[58] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[59] Edouard Hannezo, Colinda LGJ Scheele, Mohammad Moad, Nicholas Drogo, Rakesh Heer, Rosemary V Sampogna, Jacco Van Rheenen, and Benjamin D Simons. A unifying theory of branching morphogenesis. *Cell*, 171(1):242–255, 2017.

[60] Jeff Hasty, David McMillen, Farren Isaacs, and James J Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, 2(4):268–279, 2001.

[61] Alireza Heidari. Integrative approach to biological networks for emerging roles of proteomics, genomics and transcriptomics in the discovery and validation of human colorectal cancer biomarkers from dna/rna sequencing data under synchrotron radiation. *Transcriptomics*, 5(2):e117, 2017.

[62] EAM Heijnsdijk, Arno der Kinderen, EM Wever, Gerrit Draisma, MJ Roobol, and HJ De Koning. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. *British journal of cancer*, 101(11):1833–1838, 2009.

[63] Matthias Heinig, Enrico Petretto, Chris Wallace, Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R Langley, Anja Bauerfeind, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467(7314):460–464, 2010.

[64] John A Hertz, Yasser Roudi, Andreas Thorning, Joanna Tyrcha, Erik Aurell, and Hong-Li Zeng. Inferring network connectivity using kinetic ising models. *BMC neuroscience*, 11(1):1–2, 2010.

[65] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature genetics*, 41(2):178–186, 2009.

[66] Sosuke Ito. Geometric thermodynamics for the fokker-planck equation: Stochastic thermodynamic links between information geometry and optimal transport. *arXiv preprint arXiv:2209.00527*, 2022.

[67] Aimee L Jackson and Peter S Linsley. Recognizing and avoiding sirna off-target effects for target identification and therapeutic application. *Nature reviews Drug discovery*, 9(1):57–67, 2010.

[68] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012.

[69] Björn H Junker and Falk Schreiber. *Analysis of biological networks*. John Wiley & Sons, 2011.

[70] Marcus Kaiser. Mechanisms of connectome development. *Trends in Cognitive Sciences*, 21(9):703–717, 2017.

[71] Alon Keinan, Ben Sandbank, Claus C Hilgetag, Isaac Meilijson, and Eytan Ruppin. Fair attribution of functional contribution in artificial and biological networks. *Neural computation*, 16(9):1887–1915, 2004.

[72] John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

[73] Arto Klami. Polya-gamma augmentations for factor models. In *Asian Conference on Machine Learning*, pages 112–128. PMLR, 2015.

[74] Patrice Koehl, Marc Delarue, and Henri Orland. Statistical physics approach to the optimal transport problem. *Physical review letters*, 123(4):040603, 2019.

[75] A. Kolmogoroff. Sur l'estimation statistique des paramètres de la loi de Gauss. *Bull. Acad. Sci. URSS. Sér. Math. [Izvestia Akad. Nauk SSSR]*, 6:3–32, 1942.

[76] Inna Kuperstein, Luca Grieco, David PA Cohen, Denis Thieffry, Andrei Zinovyev, and Emmanuel Barillot. The shortest path is not the one you know: application of biological network resources in precision oncology research. *Mutagenesis*, 30(2):191–204, 2015.

[77] Scott Linderman, Matthew J Johnson, and Ryan P Adams. Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, 28, 2015.

[78] Andrew M Scharenberg, Philippe Duchateau, and Julianne Smith. Genome engineering with tal-effector nucleases and alternative modular nuclease technologies. *Current gene therapy*, 13(4):291–303, 2013.

[79] Chuang Ma, Han-Shuang Chen, Ying-Cheng Lai, and Hai-Feng Zhang. Statistical inference approach to structural reconstruction of complex networks from binary time series. *Physical Review E*, 97(2):022301, 2018.

[80] Dibyendu Mandal and Christopher Jarzynski. Analysis of slow transitions between nonequilibrium steady states. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(6):063204, 2016.

[81] Susan Maskery and Troy Shinbrot. Deterministic and stochastic elements of axonal guidance. *Annu. Rev. Biomed. Eng.*, 7:187–221, 2005.

[82] Facundo Mémoli. Spectral gromov-wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 256–263. IEEE, 2009.

[83] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[84] Tijana Milenković, Vesna Memišević, Anthony Bonato, and Nataša Pržulj. Dominating biological networks. *PloS one*, 6(8):e23016, 2011.

[85] Jeremy A Miller, Michael C Oldham, and Daniel H Geschwind. A systems level analysis of transcriptional changes in alzheimer's disease and normal aging. *Journal of neuroscience*, 28(6):1410–1420, 2008.

[86] Stanley L Miller and Harold C Urey. Organic compound synthesis on the primitive earth: Several questions about the origin of life have been answered, but much remains to be studied. *Science*, 130(3370):245–251, 1959.

[87] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[88] Muka Nakazato and Sosuke Ito. Geometrical aspects of entropy production in stochastic thermodynamics based on wasserstein distance. *Physical Review Research*, 3(4):043093, 2021.

[89] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[90] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press, 2001.

[91] Timothy M Otchy, Steffen BE Wolff, Juliana Y Rhee, Cengiz Pehlevan, Risa Kawai, Alexandre Kempf, Sharon MH Gobes, and Bence P Ölveczky. Acute off-target effects of neural circuit manipulations. *Nature*, 528(7582):358–363, 2015.

[92] Torsten Johann Paul and Philip Kollmannsberger. Biological network growth in complex environments: A computational framework. *PLOS Computational Biology*, 16(11):e1008003, 2020.

[93] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[94] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[95] Kiryl D Piatkevich, Seth Bensussen, Hua-an Tseng, Sanaya N Shroff, Violeta Gisselle Lopez-Huerta, Demian Park, Erica E Jung, Or A Shemesh, Christoph Straub, Howard J Gritton, et al. Population imaging of neural activity in awake behaving mice. *Nature*, 574(7778):413–417, 2019.

[96] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[97] Sumanta Ray and Sanghamitra Bandyopadhyay. A nmf based approach for integrating multiple data sources to predict hiv-1–human ppis. *BMC bioinformatics*, 17(1):1–13, 2016.

[98] John C Reed. Dysregulation of apoptosis in cancer. *Journal of clinical oncology*, 17(9):2941–2941, 1999.

[99] Daniel R Rhodes, Terrence R Barrette, Mark A Rubin, Debashis Ghosh, and Arul M Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research*, 62(15):4427–4433, 2002.

[100] Th. Rosenberg and Monica Johanson. On accumulation and active transport in biological systems. i. thermodynamic considerations. *Acta Chemica Scandinavica*, 2:14–33, 1948.

[101] Grant M Rotskoff and Gavin E Crooks. Optimal control in nonequilibrium systems: Dynamic riemannian geometry of the ising model. *Physical Review E*, 92(6):060102, 2015.

[102] Grant M Rotskoff, Gavin E Crooks, and Eric Vanden-Eijnden. Geometric approach to optimal nonequilibrium control: Minimizing dissipation in nanomagnetic spin systems. *Physical Review E*, 95(1):012148, 2017.

[103] Yasser Roudi and John Hertz. Mean field theory for nonequilibrium network reconstruction. *Physical review letters*, 106(4):048702, 2011.

[104] Justin Ruths and Derek Ruths. Control profiles of complex networks. *Science*, 343(6177):1373–1376, 2014.

[105] Ruth Sager. Expression genetics in cancer: shifting the focus from dna to rna. *Proceedings of the National Academy of Sciences*, 94(3):952–955, 1997.

[106] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[107] Erwin Schrödinger et al. *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press, 1992.

[108] Alireza Seif, Mohammad Hafezi, and Christopher Jarzynski. Machine learning the thermodynamic arrow of time. *Nature Physics*, 17(1):105–113, 2021.

[109] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[110] Albert N Shiryaev. *Probability-1*, volume 95. Springer, 2016.

[111] Chao Sima, Jianping Hua, and Sungwon Jung. Inference of gene regulatory networks using time-series data: a survey. *Current genomics*, 10(6):416–429, 2009.

[112] Florian Sohler, Daniel Hanisch, and Ralf Zimmer. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, 2004.

[113] Lucinda K Southworth, Art B Owen, and Stuart K Kim. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS genetics*, 5(12):e1000776, 2009.

[114] Barry L Stoddard. Homing endonucleases: from microbial genetic invaders to reagents for targeted dna modification. *Structure*, 19(1):7–15, 2011.

[115] Paulina Strzyz. Crispr–cas9 wins nobel. *Nature Reviews Molecular Cell Biology*, 21(12):714–714, 2020.

[116] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. 2020.

[117] Karl-Theodor Sturm et al. On the geometry of metric measure spaces. *Acta mathematica*, 196(1):65–131, 2006.

[118] Yuriy Sverchkov and Mark Craven. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS computational biology*, 13(6):e1005466, 2017.

[119] Gábor Szederkényi, Julio R Banga, and Antonio A Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC systems biology*, 5(1):1–15, 2011.

[120] Katrin Tiemann and John J Rossi. Rnai-based therapeutics–current status, challenges and prospects. *EMBO molecular medicine*, 1(3):142–151, 2009.

[121] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Saldaña. Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine*, 13(9):527–541, 2007.

[122] Fyodor D Urnov, Edward J Rebar, Michael C Holmes, H Steve Zhang, and Philip D Gregory. Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, 11(9):636–646, 2010.

[123] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[124] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[125] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[126] Darren J Wilkinson. *Stochastic modelling for systems biology*. Chapman and Hall/CRC, 2018.

[127] Mobolaji Williams. Statistical physics of the symmetric group. *Physical Review E*, 95(4), apr 2017.

[128] Mobolaji Williams. Permutation glass. *Physical Review E*, 97(1), jan 2018.

[129] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2012.

[130] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[131] Chang hun You, Lawrence B. Holder, and Diane J. Cook. Learning patterns in the dynamics of biological networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 977–986, New York, NY, USA, 2009. Association for Computing Machinery.

[132] Hong-Li Zeng, Mikko Alava, Erik Aurell, John Hertz, and Yasser Roudi. Maximum likelihood reconstruction for ising models with asynchronous updates. *Physical review letters*, 110(21):210601, 2013.

[133] Hong-Li Zeng and Erik Aurell. Inverse ising techniques to infer underlying mechanisms from data. *Chinese Physics B*, 29(8):080201, 2020.

[134] Hong-Li Zeng, Erik Aurell, Mikko Alava, and Hamed Mahmoudi. Network inference using asynchronously updated kinetic ising model. *Physical Review E*, 83(4):041135, 2011.

[135] Xiao-Hui Zhang, Louis Y Tee, Xiao-Gang Wang, Qun-Shan Huang, and Shi-Hua Yang. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, 4:e264, 2015.

[136] Daniel M Zuckerman. *Statistical physics of biomolecules: an introduction*. CRC press, 2010.

[137] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[138] Patrick R Zulkowski, David A Sivak, and Michael R DeWeese. Optimal control of transitions between nonequilibrium steady states. *PloS one*, 8(12):e82754, 2013.