# Probabilistic Investigation of Non-equilibrium Neural Network and Decision-making Models

by

**Nematollah Farhadi Mahalli**

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Physics

KOÇ ÜNİVERSİTESİ

September, 2023

**Probabilistic Investigation of Non-equilibrium Neural Network and Decision-making Models**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

**Nematollah Farhadi Mahalli**

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Alkan Kabakčioğlu (Advisor)

_____

Prof. Özgür E. Müstecaplıoğlu

_____

Assist. Prof. Onur Pusuluk

Date: _____

*To all Brave Women of Iran who have inspired the world with their unwavering strength. This work is dedicated to your courage and the hope for a brighter future that you embody.*

# ABSTRACT

**Probabilistic Investigation of Non-equilibrium Neural Network and Decision-making Models**
**Nematollah Farhadi Mahalli**
**Master of Science in Physics**
**September, 2023**

The Prisoner's Dilemma game (PDG) is commonly used as an experimental tool to study the probabilistic aspects of human decision-making. Numerous behavioral tests have been undertaken over an extended period of time to investigate this game, revealing a consistent contravention of the well-established "sure thing principle." This concept holds significant importance within the rational theory of decision-making. The violation in question can be elucidated by quantum probabilistic models, which attribute it to a second-order interference phenomenon. This particular effect is outside the scope of explanation provided by classical probability theory. In this study, we utilize the framework of generalized probabilistic theories and analyze the phenomenon of interference from the perspective of quantum information theory in order to ascertain its origin. Specifically, we propose a modification to an established quantum probabilistic model by employing the density matrix formalism. We investigate varying levels of classical and quantum uncertainty pertaining to a player's anticipation of another player's action in the Prisoner's Dilemma Game (PDG). This allows us to illustrate that the presence of quantum coherence in the player's initial prediction and its conversion to probabilities during the dynamics is what enables the explanation of the violation. Furthermore, we explore the significance of additional quantum information-theoretical measures, such as quantum entanglement, in the process of decision-making. In conclusion, we put forth a novel expansion of the Probability Distribution Generator (PDG) framework, encompassing three distinct choices. This extension aims to facilitate a comprehensive evaluation of the predictive capabilities of quantum probability theory in comparison to a broader probabilistic theory that encompasses it as a specific instance while also demonstrating third-order interference phenomena.

# ÖZETÇE

**Denge Dışı Sinir Ağlarının ve Karar Verme Modellerinin Olasılığa Dayalı Incelenmesi**
**Nematollah Farhadi Mahalli**
**Fizik, Yüksek Lisans**
**eylül 2023**

Mahkumun kilemi oyunu (PDG), insann karar verme srecinin olaslksal ynlerini incelemek iin yaygn olarak deneysel bir ara olarak kullanlr. Bu oyunu aratrmak iin uzun bir sre boyunca ok sayda davran testi yapld ve bu, kkl "kesin olan ey ilkesinin" tutarl bir ekilde ihlal edildiini ortaya kard. Bu kavram rasyonel karar verme teorisi ierisinde byk nem tamaktadr. Sz konusu ihlal, onu ikinci dereceden bir giriim olgusuna balayan kuantum olaslk modelleriyle aklanabilir. Bu zel etki, klasik olaslk teorisinin salad aklamann kapsam dndadr. Bu almada, genelletirilmi olaslk teorilerinin erevesinden yararlanyoruz ve giriim olgusunu, kkenini tespit etmek iin kuantum bilgi teorisi perspektifinden analiz ediyoruz. Spesifik olarak, younluk matrisi formalizmini kullanarak yerleik bir kuantum olaslk modelinde bir deiiklik neriyoruz. Mahkumun kilemi Oyununda (PDG) bir oyuncunun baka bir oyuncunun eylemiyle ilgili tahminiyle ilgili eitli dzeylerde klasik ve kuantum belirsizlii aratryoruz. Bu, oyuncunun ilk tahmininde kuantum tutarllnn varlnn ve bunun dinamikler srasnda olaslklara dntrlmesinin, ihlalin aklanmasn salayan ey olduunu gstermemize olanak tanr. Ayrca, karar verme srecinde kuantum dolakl gibi ilave kuantum bilgisiteorik nlemlerinin nemini aratryoruz. Sonu olarak, Olaslk Datm Oluturucu (PDG) erevesinin farkl seenei kapsayan yeni bir geniletmesini ortaya koyduk. Bu uzant, kuantum olaslk teorisinin ngr yeteneklerinin, onu belirli bir rnek olarak kapsayan ve ayn zamanda nc dereceden giriim olayların da gsteren daha geni bir olaslk teorisine kyasla kapsaml bir deerlendirmesini kolaylatrmay amalamaktadr. Fizikiler SGD eitim srecini modellemede Fokker-Planck denklemlerini (FPE) kullanmaya alyorlar. Grltl bir ortamda duraan bir durum iin FPE zmlerini kullanan bir matematiksel model araclyla mini gruplama srecini anlattk. Bu proje srasnda karlatmz zorluklardan biri, gzlemlenen duraan durumu yeniden oluturmak iin mini-toplu grlty n-boyutlu faz uzayndaki genel bir koordinat dnm biimiyle ilikilendiren basit, fiziksel

bir model oluturmakt. Projemde, ok deikenli normal dalmdan kaynaklanan kaynak grlts olarak uzayda basit bir kaydrma uygulanarak bunun yeniden retilebileceini gsterdim. Sonularmz, grltnn karakterinin, sistemin Boltzmann'a uygun veya Boltzmann'a uygun olmayan bir duraan duruma m geeceini belirlediini gstermektedir. Ek olarak, MNIST veri setini kullandm ve gerek bir veri setindeki ML'nin modelimiz tarafndan tahmin edilen duraan durumlarla nasl sonulanabileceini grmek iin bunlar Julia'da mini gruplar halinde eittim. Aslnda, gzlemlenen duraan durumun yeniden retilmesi, kovaryans matrisinin, olaslk akmnn ve difzyon matrisinin zelliklerinin kontrol edilmesini gerektirir.

# ACKNOWLEDGMENTS

I wish to extend my heartfelt thanks to Prof. Alkan Kabakčioğlu for his exceptional guidance and kindness throughout this journey. His mentorship has enriched both my academic and personal growth. I am also grateful to Dr. Onur Pusuluk for his dedicated contributions to our work. Together, they have not only shaped this research but also exemplified the professionalism I aspire to emulate. Their mentorship goes beyond academics, and I am fortunate to have learned from these remarkable individuals.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| PDG | Prisoner's Dilemma game |
| GPT | General Probability Theory |
| CPT | Classical Probability Theory |
| QPT | quantum Probability Theory |
| SGD | Stocahstic Gradient Descent |
| FPE | Fokker-Planck Equation |
| ML | Machine Learning |

# Chapter 1

# INTRODUCTION TO QUANTUM DECISION-MAKING

Is it possible to derive meaningful interpretations from the statistical results of measuring a physical system when we lack a comprehensive understanding of its inherent characteristics and governing principles? Generalized probabilistic theories (GPTs) offer a mathematical framework to address this question. These theories are referenced in literature such as [Popescu and Rohrlich, 1994], [Hardy, 2001], [Barrett, 2007], [Paterek et al., 2010], and [Mller, 2021]. They provide a structured approach to modelling experimental scenarios by breaking them down into distinct components: preparations, transformations, and measurements, each labelled for clarity. In instances where these theories are applied, preparations are effectively represented using probability vectors. Similarly, both transformations and measurements are characterized within the scope of classical probability theory (CPT) using stochastic matrices. Conversely, when dealing with quantum probability theory (QPT), preparations are symbolized by density matrices. The transformations of these density matrices are orchestrated by completely-positive trace-preserving maps, while the role of measurements is assumed by operations referred to as positive operator-valued measures.

Expanding upon these intricate and abstract mathematical structures, the framework of Generalized Probabilistic Theories (GPTs) facilitates a comprehensive investigation into the underlying physical principles that underpin a specific measurement statistic. In the pioneering work presented in Reference [Popescu and Rohrlich, 1994], Popescu and Rohrlich introduced the concept that a set of axioms rooted in physical motivation could serve as the foundation from which Quantum Probability Theory (QPT) could be derived. Their focus was primarily directed towards aspects

of nonlocality and relativistic causality. However, their investigation concluded that while these axioms contributed to the theoretical landscape of QPT [Popescu and Rohrlich, 1994, Popescu, 2014], they also resulted in an expansion of QPT's potential theoretical scope. To achieve this objective, they undertook an assessment of the violations of the CHSH-type Bell inequality [Bell, 1964, Clauser et al., 1969], where the allowable values are bounded by 2 and $2\sqrt{2}$ under the realms of Classical Probability Theory (CPT) and QPT [Cirel'son, 1980, Tsirel'son, 1987, Khalfin and Tsirelson, 1992]. Nevertheless, it emerged that the upper limit imposed by nonlocality and relativistic causality was 4 [Popescu and Rohrlich, 1994], thereby establishing that while these principles were necessary, they remained insufficient to encapsulate the entirety of QPT. This discovery also indicated that nonlocality and relativistic causality could potentially lead to the emergence of Generalized Probabilistic Theories (GPTs) that extend beyond the confines of quantum theory, showcasing distinct nonclassical attributes akin to those present in QPT. By considering a diverse array of axiomatic choices, the delineation of QPT from the broader collection of non-quantum and beyond-quantum GPTs is made possible [Hardy, 2001, Masanes and Mller, 2011, Rau, 2011, Chiribella et al., 2011]. Moreover, it's noteworthy that a more lenient set of axioms capable of isolating QPT can also harmoniously align with the principles of Classical Probability Theory (CPT) [Masanes and Mller, 2011].

Furthermore, it is important to note that instances, where Classical Probability Theory (CPT) falls short in accurately predicting measurement statistics, extend beyond solely Bell inequality experiments. A noteworthy example pertains to empirical observations indicating that individuals often deviate from a fundamental principle in rational decision theory known as the "sure-thing principle" [Savage, 1954, Pearl, 2016]. This departure from rational behavior is particularly evident in decision-making scenarios involving uncertainty [Tversky and Kahneman, 1974, Tversky and Shafir, 1992, Shafir and Tversky, 1992, Croson, 1999, Li and Taplin, 2002], such as instances encountered in the context of the Prisoner's Dilemma game [Shafir and Tversky, 1992, Croson, 1999, Li and Taplin, 2002]. Quantum Probability Theory (QPT) models addressing decision-making processes provide an explanation for this violation, attributing it to a phenomenon termed quantum interference effect [Buse-

meyer et al., 2006, Pothos and Busemeyer, 2009, Khrennikov and Haven, 2009, Tesa, 2020, Pothos and Busemeyer, 2022]. Interestingly, this quantum interference effect has found utility within decision theory as an explanatory framework for various other phenomena [Wang and Busemeyer, 2013, Aerts et al., 2013, Yearsley and Busemeyer, 2016, Busemeyer and Wang, 2017, Khrennikov et al., 2018, Basieva et al., 2021]. However, the root cause of this interference and its interplay with uncertainty remain open questions. In this context, our objective is to establish an approach that facilitates the exploration of these inquiries, harnessing the capabilities of the Generalized Probabilistic Theories (GPTs) framework.

The structure of the paper is as follows: We will begin by elucidating the appropriate manner of representing a general mental state in order to effectively model the deviation from the sure thing principle using Quantum Probability Theory (QPT) in Section 2. Moving forward to Section 3, we will deliver a comprehensive analytical illustration attributing the violation exclusively to the transitions between probabilities and quantum coherence during the initial prediction. The intricate connections encompassing quantum uncertainty, quantum coherence, and quantum entanglement will be thoroughly explored in Section 4. To supplement these discussions, Section 5 will furnish numerical findings derived from a diverse range of initial mental states. Lastly, Section 6 will delve into the potential of leveraging the Generalized Probabilistic Theories (GPT) framework to amplify existing models grounded in Quantum Probability Theory (QPT), alongside proposing methodologies to assess and affirm their efficacy.

# Chapter 2

# QUANTUM MODEL SYSTEM

Maintaining the universality of our approach, it is feasible to construct a representation of a decision-making process through the lens of the widely recognized two-player game known as the "Prisoner's Dilemma." Within this game's framework, a player's decision-making isn't exclusively rooted in logical deliberation; it also hinges upon their anticipation of the counterpart's course of action.

Consider a scenario where each participant in the game is tasked with selecting one option from a pair of choices, represented as $c$ and $d$. In this context, the mechanics behind a player's decision-making procedure can be effectively portrayed utilizing Classical Probability Theory (CPT) in the following manner. The probability vector

$$\vec{p} = \{p_{dd}, p_{dc}, p_{cd}, p_{cc}\}, \tag{2.1}$$

Conveys the psychological disposition of the player. In this context, $p_{ij}$ symbolizes the joint probability $p(B = i, A = j)$, where $A$ and $B$ denote the player's decision and their anticipatory projection concerning the other player's choice. The connection between marginal probabilities and joint probabilities is established by virtue of Bayes' theorem, which states that

$$p(B = i, A = j) = p(A = j|B = i)\, p(B = i)\,, \tag{2.2}$$

where $p(A = j|B = i)$ is the conditional probability, the probability for the player to choose action $j$ when she/he is sure the other player has chosen action $i$.

Hence, the state of action (prediction) for the player can be derived from the mental state represented by equation (2.1), following the exclusion of marginal pre-

diction (action) probabilities through the subsequent calculations:

$$\vec{p}_A = \{\sum_i p_{id}, \sum_i p_{ic}\}, \tag{2.3}$$

$$\vec{p}_B = \{\sum_j p_{dj}, \sum_j p_{cj}\}, \tag{2.4}$$

In this context, the first component of $\vec{p}_A$ ($\vec{p}_B$) corresponds to the marginal probability $p(A = d)$ ($p(B = d)$).

The formalism of density matrices allows us to view Classical Probability Theory (CPT) as a limiting case of Quantum Probability Theory (QPT) and recover all predictions made by probability vectors. For instance, the mental state given in Eq. (2.1) can be represented as a $4 \times 4$ diagonal density matrix:

$$\rho = \text{diag}(\vec{p}), \tag{2.5}$$

The partial traces of this density matrix over $B$ and $A$ lead to the following action and prediction states:

$$\rho_A \equiv \text{tr}_B[\rho] = \text{diag}(\vec{p}_A), \tag{2.6}$$

$$\rho_B \equiv \text{tr}_A[\rho] = \text{diag}(\vec{p}_B). \tag{2.7}$$

In contrast, Quantum Probability Theory (QPT) offers predictions that encompass more general states, as demonstrated by the following expression:

$$\rho_A = \begin{pmatrix} p_A & \lambda_A \\ \lambda_A^* & 1 - p_A \end{pmatrix}, \tag{2.8}$$

This configuration is also considered a valid density matrix for the action state. The diagonal elements correspond to the probabilities $p_A = p(A = d)$ and $1-p_A = p(A = c)$. A state is labeled as pure when $p_A = |\psi_d|^2$, $1 - p_A = |\psi_c|^2$, and $\lambda_A = \psi_d \psi_c^*$. In such cases, the state can be expressed as the self-outer product of a state vector $|\psi\rangle_A = \psi_d |d\rangle_A + \psi_c |c\rangle_A$, which translates to $\rho_A = |\psi\rangle\langle\psi|_A$, where $\langle\psi|$ represents the Hermitian conjugate of $|\psi\rangle$, and the basis vectors $|d\rangle$ and $|c\rangle$ are defined as:

$$|d\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |c\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.9}$$

In instances where $\rho_A$ cannot be expressed as a self-outer product, it is referred to as a mixed state, implying that the action state constitutes a probabilistic combination of two or more state vectors. An illustrative instance of this concept is evident in the classical action state (2.6), which serves as a representative example of mixed states. Specifically, it can be equivalently written as $p_A|d\rangle\langle d|_A + (1-p_A)|c\rangle\langle c|_A$. Nonetheless, it is important to note that $|\lambda_A|^2 \leq \sqrt{p_A\,(1-p_A)}$ in a general scenario.

In a comparable vein, the depiction of a generic prediction state is as follows:

$$\rho_B = \begin{pmatrix} p_B & \lambda_B \\ \lambda_B^* & 1-p_B \end{pmatrix}, \tag{2.10}$$

Within this context, $p_B = p(B=d)$, $1-p_B = p(B=c)$, and it holds that $|\lambda_B|^2 \leq \sqrt{p_B\,(1-p_B)}$.

The occurrence of non-zero off-diagonal elements within Eqs. (2.8) and (2.10) is recognized as *quantum coherence*. Although these elements do not possess a probabilistic nature, their interchange with the diagonal elements is not proscribed in a general Quantum Probability Theory (QPT) transformation. As a result, these elements have the potential, in principle, to contribute to the ultimate probabilities within a decision-making process.

In cases where the comprehensive cognitive state of the player cannot be formulated in the structure exemplified by

$$\rho = \sum_k q_k\,\rho_B^{(k)} \otimes \rho_A^{(k)}, \tag{2.11}$$

it is characterized as *inseparable* or *quantum entangled*. To elaborate, the interrelationship between the player's action and prediction encompasses non-classical correlations commonly referred to as *quantum entanglement*. This aspect of quantum information theory signifies a specific manifestation of quantum coherence, which transcends the confines of localized subsystems to be jointly shared among them. A more comprehensive exploration of the nuances surrounding quantum coherence and entanglement is slated for Section 4.

# Chapter 3

# ANALYTICAL RESULTS

The probability vector presented in equation (2.1) proves insufficient in characterizing the infringement of the sure-thing principle (STP). In situations where the player is certain about the choice $i$ made by the other player, the likelihood of the player selecting action $d$ becomes $p_{A|B=i} \equiv p(A = d|B = i)$. Conversely, when uncertainty pertains to the other player's action, this probability transitions into an average, as depicted below:

$$p_{A|B=u} = p_B\, p_{A|B=d} + (1 - p_B)\, p_{A|B=c}, \tag{3.1}$$

Here, the subscript $u$ signifies uncertainty. This equation embodies the STP: If $p_{A|B=d} = p_{A|B=c} = 1$, then, when faced with uncertainty, the player should unequivocally opt for action $d$. However, based on empirical observations [Shafir and Tversky, 1992, Croson, 1999, Li and Taplin, 2002], equation (3.1) fails to hold, as $p_{A|B=u}$ generally falls below the average of $p_{A|B=d}$ and $p_{A|B=c}$. Consequently, individuals are capable of violating the STP, at least within experimental contexts, thereby highlighting the inadequacy of Bayesian Classical Probability Theory (CPT) in elucidating the intricacies of decision-making.

We will now proceed to illustrate how the violation of the sure-thing principle (STP) can be observed within the realm of Quantum Probability Theory (QPT). We shall embark on demonstrating this phenomenon. Let the cognitive state of the player at time $t$ be denoted as $\rho^{(\alpha)}(t) = \varepsilon_t[\rho^{(\alpha)}(0)]$, wherein $\varepsilon_t$ signifies a completely-positive trace-preserving map, and $\alpha = \{d, c, u\}$. In this context, the situation $\alpha = u$ indicates a scenario where an exact prediction is lacking right from the outset of the decision-making process. Conversely, for instances where $\alpha \neq u$, the player possesses comprehensive information regarding the action chosen by the other player. This implies that the initial predictive state can be succinctly represented

as $\mathrm{tr}_A[\rho^{(\alpha\neq u)}] = |\alpha\rangle\langle\alpha|_B$. Furthermore, we shall assume a consistent initial action state across all $u$ scenarios.

The subsequent relationship among the three distinct initial states can be readily demonstrated as shown below:

$$\rho^{(u)}(0) = p_B\,\rho^{(d)}(0) + (1-p_B)\,\rho^{(c)}(0) + \chi(0)\,, \tag{3.2}$$

This equivalence holds under the condition that the uncertainty does not necessitate any initial classical or quantum correlations between the action and prediction. In essence, $\rho^{(u)}(0)$ aligns with $\rho_B(0)\otimes\rho_A(0)$. The final term in the aforementioned equation takes the form of

$$\chi(0) = \begin{pmatrix} 0 & 0 & p_A\,\lambda_B & \lambda_A\,\lambda_B \\ 0 & 0 & \lambda_A^*\,\lambda_B & (1-p_A)\,\lambda_B \\ p_A\,\lambda_B^* & \lambda_A\,\lambda_B^* & 0 & 0 \\ \lambda_A^*\,\lambda_B^* & (1-p_A)\,\lambda_B^* & 0 & 0 \end{pmatrix}, \tag{3.3}$$

Furthermore, the linearity property of $\varepsilon_t$ maintains the structure of the decomposition of $\rho^{(u)}$ across all time intervals, which can be expressed as

$$\rho^{(u)}(t) = p_B\,\rho^{(d)}(t) + (1-p_B)\,\rho^{(c)}(t) + \chi(t)\,. \tag{3.4}$$

The trace of $\chi(t)$ should always amount to zero, which stems from the stipulations that $\mathrm{tr}[\rho^{(u)}(t)] = \mathrm{tr}[\rho^{(d)}(t)] = \mathrm{tr}[\rho^{(c)}(t)] = 1$. If all the diagonal elements within $\chi(t)$ also hold a value of zero, then $\chi(t)$ contributes exclusively to quantum coherence and has no impact on the probabilities. This serves to ensure that the probabilities $p_{A|B=u}$, $p_{A|B=d}$, and $p_{A|B=c}$ strictly adhere to the sure-thing principle (STP). Conversely, when non-zero diagonal elements are present in $\chi(t)$, there's a potential for the violation of this principle. Utilizing Equation (3.4), we can establish a well-defined connection linking the probabilities, leading to the explicit relationship depicted below:

$$p_{A|B=u} = p_B\,p_{A|B=d} + (1-p_B)\,p_{A|B=c} + \delta\,, \tag{3.5}$$

where

$$p_{A|B=\alpha} \equiv p_{A|B=\alpha}(t) = \sum_{i=d,c} \langle i_B\otimes d_A|\rho^{(\alpha)}(t)|i_B\otimes d_A\rangle\,, \tag{3.6}$$

and

$$\delta \equiv \delta(t) = \sum_{i=d,c} \langle i_B \otimes d_A | \chi(t) | i_B \otimes d_A \rangle \,. \tag{3.7}$$

Equation (3.5) stands apart from (3.1) due to the inclusion of $\delta(t)$. The presence of this term introduces a discrepancy between the two equations. Whenever $\delta(t)$ assumes a non-zero value, it signifies a violation of the sure-thing principle (STP). Consequently, the condition that not only ensures but also encapsulates the fulfillment of STP throughout all dynamics is the absence of quantum coherence within the initial prediction state, denoted as $\delta(t) = 0$ for all $t$ if $\lambda_B = 0$. However, when $\lambda_B$ is non-zero, even though $\delta$ initiates with a value of zero, it might eventually progress to a non-zero value as the decision-making process unfolds. Such instances of revival in $\delta$ can plausibly be attributed to potential conversions between quantum coherence and probabilities during the trajectory of closed or open system dynamics of the cognitive state, characterized by the completely-positive trace-preserving map $\varepsilon_t$.

# Chapter 4

# INFORMATION-THEORETICAL CONCEPTS

As elaborated in Section 3, we establish that the violation of the sure-thing principle (STP) through predictions influenced by uncertainty necessitates the involvement of quantum coherence. However, the connection between uncertainty and quantum coherence requires examination. Consequently, we will delve into this aspect from the perspective of quantum information theory.

## *4.1 Quantum Uncertainty*

The extent of disorder, ambiguity, or unpredictability inherent in a system prepared in the state $\rho$ can be assessed through the notion of *von Neumann entropy* [Neumann, 1927]. This entropy is defined as follows:

$$S[\rho] = -\mathrm{tr}[\rho \log_2 \rho]. \tag{4.1}$$

The value of $S[\rho]$ becomes zero for pure states, implying the absence of uncertainty or randomness when $\rho$ can be represented as the outer product of a single state vector. Conversely, $S[\rho]$ achieves its maximum value for mixed states, where the eigenvalues constitute a uniform probability distribution. This indicates the highest level of randomness within a quantum state.

Upon inserting the eigendecomposition $\rho = \sum_j \epsilon_j |\epsilon_j\rangle\langle\epsilon_j|$ into (4.1), the outcome is $S[\rho] = -\sum_j \epsilon_j \log_2 \epsilon_j$, where $\{\epsilon_j, |\epsilon_j\rangle\}$ constitute the eigenspectrum of $\rho$. Consequently, $S[\rho]$ can be interpreted as the quantum equivalent of Shannon's information entropy [Shannon, 1948a, Shannon, 1948b]. This signifies that it is also related to the average quantum information content present within $\rho$ [Schumacher, 1995].

## *4.2 Quantum Coherence*

Quantum coherence manifests as an embodiment of quantum nonlocality and is quantified by the degree of quantum superposition inherent in a state relative to a fixed orthonormal basis. In the ensuing discussion, we will explore two appropriate quantification measures for quantum coherence, both originally introduced in Ref [Baumgratz et al., 2014]. The first measure is termed the $l_1$ *norm of coherence*, formulated as follows:

$$C_{l_1}[\rho] = \sum_{i \neq j} |\rho_{ij}|, \tag{4.2}$$

The second measure is known as the *relative entropy of coherence*:

$$C_{RE}[\rho] = \min_{\varsigma \in \mathrm{IC}} S[\rho||\varsigma] = S(\rho_d) - S(\rho), \tag{4.3}$$

Here, $S[\rho||\varsigma]$ represents the *relative von Neumann entropy* of $\rho$ with respect to $\varsigma$, IC represents the ensemble of all incoherent states, and $\rho_d$ signifies the diagonal component of the density matrix.

The measure $C_{l_1}$ takes into account distinct off-diagonal terms in $\rho$ individually, treating each off-diagonal element as indicative of nonlocality between two basis states. On the other hand, $C_{RE}$ does not distinguish between these off-diagonal terms and emphasizes the overall nonclassical character, offering a holistic view. Moreover, $C_{RE}$ quantifies how much the removal of all coherence would increase uncertainty within the system.

## *4.3 Quantum Correlations*

Quantum correlations emerge as a result of quantum coherence that is distributed across subsystems without being confined to specific locations. In this context, we will focus solely on a specific form of quantum correlation known as quantum entanglement.

The *entropy of entanglement* [Bennett et al., 1996a], defined as

$$E_E[|\psi\rangle\langle\psi|] = (S \circ \mathrm{tr}_{A(B)})[|\psi\rangle\langle\psi|], \tag{4.4}$$

serves as the unique measure for quantifying entanglement in pure bipartite states [Popescu and Rohrlich, 1997]. As a result, the uncertainty within subsystems increases as the entanglement in the pure joint state becomes more pronounced.

The concept of entanglement entropy can be extended to account for mixed states by utilizing a convex roof approach. This expansion is referred to as the *entanglement of formation* [Bennett et al., 1996b],

$$E_F[\rho] = \min \left( \sum_j r_j \, E_E[|\psi_j\rangle\langle\psi_j|] \right), \tag{4.5}$$

where the minimum is taken over all possible decompositions of $\rho$ into pure states that satisfy $\rho = \sum_j r_j \, |\psi_j\rangle\langle\psi_j|$.

# Chapter 5

# NUMERICAL RESULTS

Even though their associations with quantum coherence are explored in Section 4, it remains complex to establish definitively whether quantum uncertainty and entanglement contribute to the violation of the STP. This section delves into the examination of uncertainty and entanglement dynamics within the player's mental state. For this purpose, we will initiate the dynamics using four distinct initial states that showcase varying distributions of uncertainty and coherence.

## *5.1 Initial Mental States*

We consider initial mental states that lack correlations between their components,

$$\rho^{(\alpha)}(0) = \rho_B^{(\alpha)}(0) \otimes \rho_A^{(\alpha)}(0), \tag{5.1}$$

where $\alpha = \{u, d, c\}$ as detailed in Section 3. Each subsystem's description relies on only two parameters, $\lambda_{A/B}$ and $p_{A/B}$, as explained in Eqs. (2.8-2.10) respectively. In cases where precise prediction characterizes the beginning of the decision-making process, $p_B$ assumes values of either zero or one, resulting in $\mathrm{tr}_A[\rho^{(\alpha \neq u)}] = |\alpha\rangle\langle\alpha|_B$. Moreover, to ensure the initial action state's immunity to uncertainty, $\rho_A^{(u)}(0) = \rho_A^{(d)}(0) = \rho_A^{(c)}(0)$. Our attention is directed towards four representative cognitive states, outlined as follows.

**1) Classical-Classical States:** The initial set of cognitive states that we are examining lack initial quantum coherence, which is localized in either the prediction or action subsystem. However, there is maximum uncertainty present within both subsystems. The resulting cognitive states, denoted as Case 1 in Table 5.1, correspond to the initial states previously utilized in a CPT Model by Pothos and Busemeyer [Pothos and Busemeyer, 2009]. Furthermore, we investigate different

levels of uncertainty by reducing the uniformity of the marginal probability distributions (as shown in Case 1* in Table 5.1).

**2) Classical-Quantum States:** In contrast to the previous situation, we introduce initial quantum coherence into the action subsystem. This introduction, represented as Case 2 in Table 5.1, leads to the establishment of maximum local coherence within this subsystem, effectively eliminating all uncertainty.

**3) Quantum-Classical States:** In this situation, we introduce quantum coherence into the initial prediction state. Through maximizing the infusion of coherence, we can render the state $\rho_B^{(u)}$ pure, effectively eliminating uncertainty in the prediction. Consequently, this approach does not correspond to a lack of precise prediction at the commencement of the decision-making process. This renders Case 3 indicative of an inappropriate mental state. However, when the introduced quantum coherence is only half of the maximum attainable quantity, a substantial degree of uncertainty (81%) persists in the prediction (see Case 3* in Table 5.1).

**4) Quantum-Quantum States:** The final category of mental states we are investigating is characterized by the presence of quantum coherence initially localized both in the prediction and action subsystems. In the scenario of maximal quantum coherence, no uncertainty is present in either of these subsystems. This specific mental state is labelled as Case 4 in Table 5.1, which corresponds to the initial states previously employed in a QPT model developed by Pothos and Busemeyer [Pothos and Busemeyer, 2009]. However, unlike Case 3, this mental state is not considered appropriate. As a result, we explore states that display a combination of coherence and uncertainty in the prediction, accomplished by reducing the magnitude of $\lambda_B$, indicated as Case 4* in Table 5.1.

### *5.2 Quantum Dynamics*

Throughout this discussion and in the subsequent sections, our central emphasis is directed towards the distinct dynamics exhibited by mental states as outlined in the work by Busemeyer and Pothos [Pothos and Busemeyer, 2009]. The aim of our inquiry is to delve into the potential impact that quantum uncertainty and entanglement might exert on the violation of the sure-thing principle (STP).

More precisely, we characterize the time evolution of the mental state using a unitary dynamics described by the equation

$$\rho(t) = \boldsymbol{U}(t)\rho(0)\boldsymbol{U}^{\dagger}(t)\,, \tag{5.2}$$

where this temporal progression is driven by a Hamiltonian operator $\boldsymbol{H}$, as outlined in the subsequent expression:

$$\boldsymbol{U}(t) \equiv \exp(-i\boldsymbol{H}t/\hbar)\,. \tag{5.3}$$

In the foregoing context, the value of $\hbar$ is taken to be 1, and the Hamiltonian operator is divided into two components as indicated by the equation:

$$\boldsymbol{H} = \boldsymbol{H}_A + \boldsymbol{H}_B\,, \tag{5.4}$$

where each of these terms represents distinct interactions involving the prediction and action of the player. The initial interaction grants predictive influence over the action, and it is formulated as follows:

$$\boldsymbol{H}_A = \sum_{i=d,c} |i\rangle\langle i|_B \otimes \boldsymbol{H}_{Ai}\,, \tag{5.5}$$

with

$$\boldsymbol{H}_{Ai} = \frac{1}{\sqrt{1+\mu_i^2}} \begin{pmatrix} \mu_i & 1 \\ 1 & -\mu_i \end{pmatrix}\,, \tag{5.6}$$

where $\mu_i$ represents a constant associated with the pay-offs linked to actions $d$ and $c$ when prediction $i$ is considered. In simpler terms, $\boldsymbol{H}_A$ is incorporated into the Hamiltonian with the specific intention of blending the joint probabilities $p_{id} = \langle i_B \otimes d_A|\rho(t)|i_B \otimes d_A\rangle$ and $p_{ic} = \langle i_B \otimes c_A|\rho(t)|i_B \otimes c_A\rangle$ as functions of both $\mu_i$

and $t$. Furthermore, it affects the quantum coherence carried by the mental state. Nevertheless, it lacks the capability to generate quantum entanglement between the initially separate prediction and action states.

We suppose that both constants, $\mu_d$ and $\mu_c$, share the same value of $\mu$ as demonstrated in [Pothos and Busemeyer, 2009]. This implies that the pay-offs are not influenced by the prediction and remain detached from it.

The second term in the Hamiltonian (5.4) can be expressed as follows:

$$\boldsymbol{H}_B = \sum_{j=d,c} \boldsymbol{H}_{Bj} \otimes |j\rangle\langle j|_A \,, \tag{5.7}$$

where

$$\boldsymbol{H}_{Bj} = -\frac{\gamma}{\sqrt{2}} \begin{pmatrix} \nu_j & 1 \\ 1 & -\nu_j \end{pmatrix} , \tag{5.8}$$

Given $\nu_d = -\nu_c = 1$, the term $\boldsymbol{H}_B$ combines the joint probabilities $p_{dj} = \langle d_B \otimes j_A|\rho(t)|d_B \otimes j_A\rangle$ and $p_{cj} = \langle c_B \otimes j_A|\rho(t)|c_B \otimes j_A\rangle$ as a function of $\gamma$ and $t$. This operation fundamentally alters the influence and interaction between the player's prediction and action. Unlike $\boldsymbol{H}_A$, the impact of $\boldsymbol{H}_B$ goes beyond managing the quantum coherence present in the mental state; it also establishes quantum entanglement between initially separate prediction and action states.

Taking into account the combined effects of $\boldsymbol{H}_A$ and $\boldsymbol{H}_B$, the overall Hamiltonian dictating the unitary evolution outlined in Eq. (5.2) is established. However, once the player has definitively made a decision to select a specific action, the final phase of the decision-making process is modelled by an abrupt measurement operation. In this context, we envision the execution of a projective measurement on the action state. This measurement acts to erase all existing quantum coherence and correlations within the action state, resulting in its collapse to the uncorrelated state $|j\rangle_B$ with a probability of $p_{dj} + p_{cj}$.

### *5.3 Uncertainty, Coherence, and Correlations*

In this section, we delve into an examination of the unitary evolutions of the initial mental states as outlined in Table 5.1, while taking into account the Hamiltonian

outlined in Eq. (5.4). Our earlier exploration in Section 3 established that in order to observe a violation of STP, the requirement for a non-zero value of $\delta$ is crucial. This condition points to the existence of quantum coherence in the initial prediction, and this analytical finding remains unaltered by the dynamics of the mental state. To further validate this discovery, we carried out numerical simulations with $\gamma = 1.74$ and $\mu_d = \mu_c = 0.59$. The results are summarized in Tables 5.2 and 5.3, and visually depicted in Figure 10.1. In alignment with our analytical analysis, we observe that the violation of STP only occurs when the initial prediction showcases quantum coherence (Cases 3, 3$^*$, 4, and 4$^*$).

The discussion in Section 4 raises another question concerning the violation of STP. Can the dynamics of a closed system generate entanglement between prediction and action, potentially influencing the violation? However, a thorough examination of the results summarized in Table 5.3 demonstrates that quantum entanglement does not significantly contribute to this violation. In Case 1, where STP remains unviolated, entanglement does not arise. In contrast, Cases 1$^*$ and 2 indicate the possibility of entanglement between prediction and action. A similar pattern emerges in the scenarios where STP is violated. For instance, in Case 3$^*$, the mental state retains its separability, while entanglement is observed in the remaining cases. Overall, these outcomes suggest that although the initial local quantum coherence in prediction can evolve into quantum entanglement between prediction and action over time, it does not directly play a substantial role in the violation.

Upon comparison between the outcomes presented in Table 5.1 and Table 5.3, an additional important aspect becomes evident concerning the behavior of the mental state's evolution governed by Eq. (5.4). This evolution possesses the capability to convert the initial quantum coherence present in both prediction and action into quantum entanglement. Moreover, a particularly fascinating observation is that even when the initial mental state lacks quantum coherence, the action and prediction can still develop entanglement. Case 1$^*$ serves as a notable instance of this phenomenon, where the initial action and prediction states are characterized by classical, albeit nonuniform, probability distributions. Intriguingly, if the dynamics of the mental state adhere to quantum principles, both quantum coherence and entanglement can

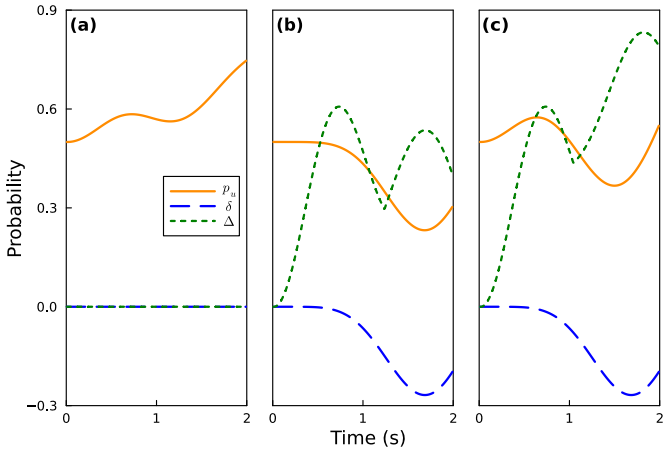


Figure 5.1: The time-dependent behavior of the probabilistic quantities $p_u \equiv p_{A|B=u}$, $\delta$, and $\Delta \equiv \sum_i |\chi_{ii}|$ is illustrated using solid orange, dashed blue, and dotted green curves for the following cases: (a) Case 2, (b) Case 3*, and (c) Case 4*.

.

still arise.

In conclusion, it can be stated that the violation of STP is not influenced by prediction uncertainty or action-prediction entanglement. The initial quantum coherence present in the prediction state, as discussed in Section 3, undergoes a conversion into probabilities during the evolution of the mental state. This process clarifies the observed violation of STP in the original QPT model proposed in Ref. [Pothos and Busemeyer, 2009]. Moreover, our analytical discussion in Section 3 gains empirical support through the numerical calculations depicted in Figure 10.1. The diagonal elements of the matrix $\chi$ in Eq. (3.3), as shown, exhibit non-zero values over time in cases where STP is violated. An especially significant finding emerges when examining the violation of STP: it initiates at the point where a change becomes evident in the behavior of $\Delta$, the sum of these probabilistic quantities. Although this aspect lies outside the scope of this paper, we intend to explore this intriguing result in future studies.

Table 5.1: The unique characteristics of the different initial states are determined by the parameters $p_{B/A}$ and $\lambda_{B/A}$. The levels of quantum information content and uncertainty present in the prediction and action are measured using quantities such as von Neumann entropy ($S$) and the $l_1$ norm of coherence ($C_{l_1}$) across seven distinct cases. Importantly, instances where $S(B)$ becomes zero for $\alpha = u$ should not be considered as viable mental states.

| Case No. | $\alpha$ | $p_B$ | $\lambda_B$ | $C_{l_1}(B)$ | $S(B)$ | $p_A$ | $\lambda_A$ | $C_{l_1}(A)$ | $S(A)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $u$ | 1/2 | 0 | 0 | 1 | 1/2 | 0 | 0 | 1 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| 1* | $u$ | 1/3 | 0 | 0 | 0.92 | 3/5 | 0 | 0 | 0.97 |
| | $d$ | 1 | 0 | 0 | 0 | 3/5 | 0 | 0 | 0.97 |
| | $c$ | 0 | 0 | 0 | 0 | 3/5 | 0 | 0 | 0.97 |
| 2 | $u$ | 1/2 | 0 | 0 | 1 | 1/2 | 1/2 | 1 | 0 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |
| 3 | $u$ | 1/2 | 1/2 | 1 | 0 | 1/2 | 0 | 0 | 1 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| 3* | $u$ | 1/2 | 1/4 | 0.5 | 0.81 | 1/2 | 0 | 0 | 1 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 0 | 0 | 1 |
| 4 | $u$ | 1/2 | 1/2 | 1 | 0 | 1/2 | 1/2 | 1 | 0 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |
| 4* | $u$ | 1/2 | $i/4$ | 0.5 | 0.81 | 1/2 | 1/2 | 1 | 0 |
| | $d$ | 1 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |
| | $c$ | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 1 | 0 |

Table 5.2: We calculate the mean uncertainty within the considered mental states over the time span $t = 0$ to $2\pi$. The extent of quantum uncertainty is evaluated using the von Neumann entropy ($S$). The "STP" column indicates whether each case adheres to or violates the STP within the time interval $(0, 2\pi)$. The last column provides the total correlations shared between the action and prediction states, which is defined as $I(A:B) = S(A) + S(B) - S(AB)$.

| Case | STP | $\alpha$ | $S(B)$ | $S(A)$ | $S(AB)$ | $I(A:B)$ |
|---|---|---|---|---|---|---|
| 1 | S | $u$ | 1 | 1 | 2 | 0 |
| | | $d$ | 0.65 | 1 | 1 | 0.65 |
| | | $c$ | 0.65 | 1 | 1 | 0.65 |
| 1* | S | $u$ | 0.96 | 0.99 | 1.89 | 0.06 |
| | | $d$ | 0.66 | 0.99 | 0.97 | 0.68 |
| | | $c$ | 0.66 | 0.98 | 0.97 | 0.66 |
| 2 | S | $u$ | 0.88 | 0.70 | 1 | 0.59 |
| | | $d$ | 0.41 | 0.40 | 0 | 0.81 |
| | | $c$ | 0.50 | 0.50 | 0 | 1.00 |
| 3 | NS | $u$ | 0.81 | 0.80 | 1 | 0.61 |
| | | $d$ | 0.67 | 1 | 1 | 0.67 |
| | | $c$ | 0.67 | 1 | 1 | 0.67 |
| 3* | NS | $u$ | 0.96 | 0.95 | 1.81 | 0.10 |
| | | $d$ | 0.67 | 1 | 1 | 0.67 |
| | | $c$ | 0.67 | 1 | 1 | 0.67 |
| 4 | NS | $u$ | 0.53 | 0.53 | 0 | 1.05 |
| | | $d$ | 0.41 | 0.41 | 0 | 0.81 |
| | | $c$ | 0.50 | 0.50 | 0 | 1.00 |
| 4* | NS | $u$ | 0.76 | 0.61 | 0.81 | 0.56 |
| | | $d$ | 0.41 | 0.41 | 0 | 0.81 |
| | | $c$ | 0.50 | 0.50 | 0 | 1.00 |

Table 5.3: We assess the mean quantum properties of the considered mental states over the time range $t = 0$ to $2\pi$. The level of quantum coherence is gauged using two metrics: the $l_1$ norm of coherence ($C_{l_1}$) and the relative entropy of coherence ($C_{RE}$). Concurrently, the quantification of quantum entanglement is carried out using the entanglement of formation ($E_F$). The "STP" column indicates whether each case adheres to or violates the STP within the interval $(0, 2\pi)$.

| Case | STP | $\alpha$ | $C_{l_1}(B)$ | $C_{l_1}(A)$ | $C_{l_1}(AB)$ | $C_{RE}(AB)$ | $E_F(AB)$ |
|---|---|---|---|---|---|---|---|
| 1 | S | $u$ | 0 | 0 | 0 | 0 | 0 |
| | | $d$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| | | $c$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| 1* | S | $u$ | 0.15 | 0.10 | 0.52 | 0.09 | 0 |
| | | $d$ | 0.39 | 0.09 | 1.30 | 0.84 | 0.04 |
| | | $c$ | 0.40 | 0.12 | 1.32 | 0.85 | 0.03 |
| 2 | S | $u$ | 0.31 | 0.43 | 1.40 | 0.86 | 0.08 |
| | | $d$ | 0.66 | 0.64 | 2.39 | 1.60 | 0.41 |
| | | $c$ | 0.58 | 0.59 | 2.34 | 1.57 | 0.50 |
| 3 | NS | $u$ | 0.42 | 0.39 | 1.36 | 0.85 | 0.11 |
| | | $d$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| | | $c$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| 3* | NS | $u$ | 0.21 | 0.19 | 0.68 | 0.15 | 0 |
| | | $d$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| | | $c$ | 0.38 | 0 | 1.17 | 0.83 | 0 |
| 4 | NS | $u$ | 0.55 | 0.64 | 2.59 | 1.72 | 0.53 |
| | | $d$ | 0.66 | 0.64 | 2.39 | 1.59 | 0.41 |
| | | $c$ | 0.58 | 0.59 | 2.34 | 1.57 | 0.50 |
| 4* | | $u$ | 0.48 | 0.54 | 1.72 | 1.00 | 0.14 |
| | | $d$ | 0.66 | 0.64 | 2.39 | 1.59 | 0.41 |
| | | $c$ | 0.58 | 0.59 | 2.34 | 1.57 | 0.50 |

# Chapter 6

# OUTLOOK

In this study, we present a comprehensive viewpoint on the classical and quantum decision-making models proposed by Busemeyer and Pothos in their work [Pothos and Busemeyer, 2009]. Our methodology utilizes the unified framework of Generalized Probabilistic Theories (GPTs), allowing us to represent both classical and quantum mental states using density matrices. To capture the evolution of these mental states over time, we adopt the Hamiltonian employed in the Quantum Probabilistic Theory (QPT) model. Upon analyzing the data provided in Table 5.1 and Table 5.3, we discover that this Hamiltonian possesses the capacity to transform an initial classical mental state into a quantum one. To address this phenomenon, our intention is to develop an open-system approach that maintains the classical nature of all initial states. This can be achieved by formulating a master equation that incorporates the Hamiltonian presented in Eq. (5.4).

Busemeyer and Pothos offered an explanation for the violation of the STP within the QPT model by attributing it to the second-order interference effect [Pothos and Busemeyer, 2009]. However, their conclusion was based on an inappropriate representation of the absence of exact prediction using a particular mental state. In our investigation, we delved into the dynamics of diverse mental systems. Through both theoretical analysis and numerical simulations, we demonstrated that the root cause of the interference resulting in STP violation is the presence of quantum coherence in the initial prediction. Furthermore, we unraveled the underlying mechanism: the quantum coherence in the mental state of the initial prediction transforms into probabilities during the passage of time. Importantly, we established that the emergence of quantum uncertainty in the prediction over time or the establishment of quantum entanglement between prediction and action states do not contribute to this mechanism.

The initial mental state employed in the QPT model proposed by Busemeyer and Pothos [Pothos and Busemeyer, 2009] was identified as Case 4 in Table 5.1. Apart from its lack of prediction uncertainty, as evidenced by the data in Table 5.3, this state allows for the emergence of quantum entanglement between the action and prediction subsystems as time elapses. Nevertheless, the question of whether it is possible to partition a mental system during the decision-making process in a way that attributes nonlocality remains unresolved. To avoid delving into speculative discourse, our investigation suggests replacing the mental state in the QPT model with a state, such as the one described in Case 3*, which consistently maintains its separability.

The adoption of the GPT framework in this context provides us with a plethora of insights that extend beyond the analysis of existing probabilistic models for decision-making processes. As previously mentioned in the Introduction, phenomena such as second-order interference are not confined solely to Quantum Probability Theory (QPT). Various Generalized Probabilistic Theories (GPTs), whether they are non-quantum or beyond-quantum, could manifest similar phenomena. In fact, as originally proposed by Sorkin in Ref. [Sorkin, 1994], it is plausible that certain GPTs might encompass interference phenomena of third-order or even higher. Given this consideration, it prompts the question of whether these types of GPTs could potentially outperform QPT in effectively modeling human decision-making mechanisms in games such as the Prisoner's Dilemma (PDG).

The concept of higher-order interference can be illuminated by drawing a comparison with the well-known double-slit experiment, as discussed in Ref. [Mller, 2021]. Let's consider a scenario with two slits, where we are interested in the probability of detecting a particle on a screen when both slits are open, denoted as $P_{12}$. Similarly, there are probabilities $P_1$ and $P_2$ associated with detecting a particle when only slit 1 or slit 2 is open, respectively. According to Classical Probability Theory (CPT), the expectation would be $P_{12} = P_1 + P_2$. However, Quantum Probability Theory (QPT) predicts an inequality $P_{12} \neq P_1 + P_2$. Now, let's expand this experiment by introducing a third slit. Interestingly, both QPT and CPT lead to the same equation: $P_{123} = P_{12} + P_{13} + P_{23} - P_1 - P_2 - P_3$. Nonetheless, when we consider a

Generalized Probabilistic Theory (GPT) that accounts for third-order interference, we discover that $P_{123} \neq P_{12} + P_{13} + P_{23} - P_1 - P_2 - P_3$. A similar analogy can be drawn between the choices made by players in the Prisoner's Dilemma (PDG) and the slits in the double-slit experiment. In this context, we contemplate the outcome if we were to expand the PDG to incorporate three choices for each player. Would we observe an equality that aligns with both CPT and QPT? Alternatively, could we witness a violation that cannot be accounted for by either CPT or QPT?

If experiments conducted using a three-choice Prisoner's Dilemma (PDG) reveal violations that cannot be accounted for within the Quantum Probability Theory (QPT) model, it becomes essential to transition from the density matrix formalism to the density cube formalism [Dakić et al., 2014]. In this context, our goal is to expand the QPT model, initially proposed by Busemeyer and Pothos and analyzed here using quantum information theory methods, to encompass the three-choice PDG. Through this generalization, we can gain valuable insights into whether alternative frameworks beyond QPT are required to accurately model decision-making processes.

# Chapter 7

# INTRODUCTION TO MACHINE LEARNING AND FOKKER-PLANCK EQUATIONS

Currently, neural network-based machine learning has had a significant impact on various domains, such as image recognition, finance, manufacturing, physics, engineering, and biology [Carleo and et al., 2019, Goodfellow et al., 2016, Mehta and et al., 2019, Seifert, 2012, Bryngelson et al., 1995, Veitshans et al., 1997, Dill et al., 2017, Mustonen and Lssig, 2010, Hofrichter et al., 2017]. These applications are based on a model that trains a neural network, which involves determining the model's parameters that minimize the difference between the known truth, derived from a set of training input data, and the model's predicted output. This difference is the loss function, which relies on the network's parameters, which can be extensive, numbering in the billions or more for highly intricate problems. Consequently, the training process becomes a quest to locate the lowest point within a complex, multi-dimensional landscape outlined by the loss function [Li et al., 2018].

The stochastic Gradient Descent (SGD) method optimizes neural network parameters iteratively by using random subsets of training data, aiding convergence and enhancing efficiency in complex learning tasks [Robbins and Monro, 1951].

The stochastic gradient descent (SGD) approach approximates the loss function by selecting a randomly selected subset, referred to as a minibatch, from the complete training dataset during each training iteration. The utilization of this minibatch is to compute the gradient in relation to the parameters of the neural network. The parameters are subsequently modified in a downward direction based on an expected gradient [Hofrichter et al., 2017].

The Fokker-Planck equation is a fundamental concept in the field of stochastic processes and statistical physics. It offers a robust mathematical framework for char-

acterizing the temporal changes of probability distributions in systems influenced by random fluctuations. The partial differential equation discussed in this context has its roots in the research conducted by Adriaan Fokker and Max Planck. It has been widely utilized in several scientific fields, including molecular dynamics, finance, and neuroscience [Welling and Teh, 2011, Hansen et al., 1993].

The fundamental essence of the Fokker-Planck equation is in its capacity to describe the temporal evolution of a probability density function, taking into account the combined effects of deterministic drift forces and stochastic influences. The aforementioned phenomenon embodies an equilibrium between the aforementioned constituents, encompassing the influences of diffusion and drift in systems characterized by stochastic fluctuations. The solutions of the equation provide valuable insights into the enduring characteristics of these probabilistic systems, offering a deeper understanding of concepts such as equilibrium states, stability, and the formation of intricate patterns [Radons et al., 1990].

The Fokker-Planck equation finds extensive utility in several domains, including the representation of particle motion within a fluid medium, elucidation of stock price fluctuations within financial markets, and comprehension of neuronal firing patterns in the field of neuroscience. The answers provided not only enhance our comprehension of the fundamental stochastic processes but also facilitate the ability to make predictions regarding the behavior of the system over a period of time. In general, the Fokker-Planck equation serves as a fundamental tool for elucidating the complex relationship between stochasticity and determinism in diverse natural and manmade systems [Risken, 1996].

The Fokker-Planck equation is a widely used mathematical tool that is frequently utilized to get a deeper understanding of both Stochastic Gradient Descent (SGD) and diverse natural phenomena. The value of this concept resides in its capacity to provide a structured framework for understanding the similarities between these ostensibly disparate events. Within the domain of machine learning, the utilization of equations serves to enhance the understanding of the functioning of stochastic gradient descent (SGD) algorithms. This facilitates the examination of convergence qualities as well as the evaluation of the balance between exploration and exploita-

tion. In addition to its computational domain, the equation has been widely utilized in the field of biology to effectively model protein dynamics and evolutionary mechanisms. The wide range of applications highlights the importance of this phenomenon in enriching our comprehension of various intricate systems [Mustonen and Lssig, 2010, Best and Hummer, 2011, Kimura and et al., 1954, Radons et al., 1990]. Researchers argue that the stationary state of stochastic gradient descent (SGD) is of a nonequilibrium character, utilizing the framework of Fokker-Planck equations [Hansen et al., 1993, Chaudhari and Soatto, 2018].

We use FPE to model SGD with minibatching. The point is that minibatching loss function is not exactly equal to the tru loss function of the data. In othere words, there is a noise in the minibatch loss function. In spite of this difference between the loss functions, minibatching takes us to the almost the same stationary state as the true one. In our model, we add a noise which is rooted in the parameters and results in a noise in loss functions. This noise which is generated from the same distribution at each time step of SGD, enables us to engineer the stationary state features. This model is capable of simply modelling training process, and also, may give a way to make the trainign process faster.

In the following, we talk about neural network and SGD, and then explain the equations of SGD by inibatching. In the next chapter, we begin with random variables, and we derive FPE from scratch. In the next chapter, we see the solutions and stationary state distribution of FPE. Then, we bring a numerical model from which we train our data. Finally, we construct our model and show how the source noise in the parameter can arrange the final features of the stationary state.

# Chapter 8

# MACHINE LEARNING

Machine learning utilizes neural networks to simulate the cognitive learning mechanism of the human brain. Neural networks are comprised of interconnected nodes that are organized into layers. The input data is initially presented to the neural network through the input layer, and then it undergoes a series of transformations as it passes through succeeding layers, where these transformations are facilitated by weighted connections that emulate the functionality of synapses. During the training process, the neural network modifies the weights by comparing its output with the expected results. This iterative process involves decreasing the discrepancy between the output and the expected results using optimization methods.

Neural networks demonstrate exceptional proficiency in the identification of patterns and relationships within datasets, owing to its inherent ability to unveil complex features through the utilization of hidden layers. The layers of the network gradually abstract intricate information, facilitating the network's ability to generalize from training examples to novel data. The learning process persists until the network attains a desirable level of accuracy on the training data, so equipping it with the ability to generate well-informed predictions on previously unknown data. The adaptable nature of this methodology enables neural networks to achieve exceptional performance in a wide range of tasks, including but not limited to image and speech recognition, natural language processing, and predictive analytics. As a result, it significantly contributes to the progress and development of diverse academic disciplines.

### *8.1 Neural Network*

Within this context, we encounter a neural network (NN) comprising a multitude of neurons, specifically N in number. These neurons are characterized by an array of parameters $\boldsymbol{\theta} = \theta_\alpha$, where $\alpha$ spans a range from 1 to N. The training of this neural network hinges upon the utilization of a set of M input-output pairs, denoted as $Z \equiv \boldsymbol{z}_i \equiv (\boldsymbol{x}_i, \boldsymbol{y}_i)$, with the index i varying from 1 to M. It is worth noting that the parameters of these neurons, encapsulated within $\boldsymbol{\theta}$, alongside the input vectors $\boldsymbol{x}_i$ and output vectors $\boldsymbol{y}_i$, are all expressed as real-valued vectors. The training process unfolds across numerous iterative steps. During these steps, neurons are provided with input, $\boldsymbol{x_i}$, with the expectation that they yield predefined outputs, $\boldsymbol{y}_i$. The actual output of each neuron is constructed through the application of a function intricately linked to both the input and the neuron parameters at each time step. This function is denoted as $f(\boldsymbol{\theta}; \boldsymbol{x}_i, \boldsymbol{y}_i)$.

However, it is important to acknowledge that the actual outcome may not be a perfect match with the anticipated result. To quantify this discrepancy, a vital metric known as the loss function, symbolized as $\boldsymbol{\mathcal{L}}(\boldsymbol{\theta})$, is introduced. The loss function is designed to capture the extent of disparity between the actual and expected outputs, and its formulation is presented as follows:

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}) + R(\boldsymbol{\theta}) \tag{8.1}$$

Herein, it is noteworthy that $R(\boldsymbol{\theta})$ constitutes a constraint term, independent of the data, whereas $L(\boldsymbol{\theta})$ is elucidated as follows:

$$L(\boldsymbol{\theta}) \equiv l_i(\boldsymbol{\theta}; \boldsymbol{x}_i, \boldsymbol{y}i) \qquad = \frac{1}{M} \sum i = 1^M [f(\boldsymbol{\theta}; \boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{y}_i]^2 \tag{8.2}$$

In essence, the overarching objective of the training process revolves around the minimization of the loss function, striving to align the predicted outcomes with the actual results. The forthcoming elucidation shall delve into the mechanisms by which the neural network undertakes this critical task.

### *8.2 Stochastic Gradient Descent*

The training process unfolds through the dynamic adjustment of the neural network's parameters, denoted as $\boldsymbol{\theta}$. This crucial task is accomplished through the application of the gradient descent algorithm, which plays a pivotal role in updating these parameters. The gradient descent process can be mathematically represented as follows:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \eta \boldsymbol{\nabla} L(\boldsymbol{\theta}^{t}) \tag{8.3}$$

In this equation, it's essential to note that the value of $\boldsymbol{\theta}$ at time step $t+1$ is intricately linked to its counterpart at time step $t$, with the modification guided by the gradient of the loss function at the same time step. This gradient serves as a compass, directing the network towards optimal parameter values. However, the magnitude of this adjustment is governed by a crucial factor known as the learning rate, symbolized as $\eta$. This learning rate dictates the step size of parameter updates, influencing the convergence speed and stability of the training process, and it's a key hyperparameter that demands careful tuning for effective neural network training [Adhikari et al., 2023].

### *8.3 Mini Batch*

As detailed in the preceding section, the imperative objective of minimizing the loss function necessitates the iterative application of the gradient descent relation, Eq.(8.3), at each time step, accompanied by the computation of the loss function over the entire array of neurons at each of these temporal intervals. However, this undertaking can be computationally demanding, especially when dealing with a substantial number of neurons, leading to significant time and resource consumption. To alleviate this computational burden, a methodological approach emerges: computing the loss function over a subset of neurons, termed a mini-batch, denoted as $L_{B_T}(\boldsymbol{\theta}) = (1/m) \sum_{i \in B_t} l_i(\boldsymbol{\theta}; \boldsymbol{x}_i, \boldsymbol{y}_i)$, with $m$ representing the mini-batch size. This modification considerably reduces the computational cost associated with each time step, facilitating more efficient training procedures. However, the introduction of

mini-batches introduces a fundamental alteration in the gradient relation, Eq.(8.3). The gradient of the loss function for a mini-batch can be expressed as follows:

$$\begin{aligned}\boldsymbol{\nabla} L(\boldsymbol{\theta}^t) &= \frac{1}{M}\sum_{i=1}^{M} l_i(\boldsymbol{\theta}^t; \boldsymbol{x}_i^t, \boldsymbol{y}_i^t) \\ &= \boldsymbol{\nabla}\left[\frac{1}{m}\sum_{i\in B_t} l_i(\boldsymbol{\theta}^t; \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{1}{M-m}\sum_{i\notin B_t} l_i(\boldsymbol{\theta}^t; \boldsymbol{x}_i, \boldsymbol{y}_i)\right]\end{aligned} \tag{8.4}$$

Nevertheless, during the execution of the training program, the true loss function remains uncomputed. Consequently, in the absence of precise knowledge regarding the true loss function at each time step, a stochastic term is introduced into the gradient of the mini-batch loss function to account for this discrepancy. This adjustment results in a modification of the gradient descent relation, as presented below:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \boldsymbol{\nabla} L_{B_t}(\boldsymbol{\theta}^t) + \boldsymbol{\xi}^t \tag{8.5}$$

The parameter introduced here, denoted as $\boldsymbol{\xi}^t$, essentially quantifies the difference between the true and mini-batch loss functions. This component is commonly referred to as a noise term, and its formulation is defined as follows:

$$\boldsymbol{\xi}^t \equiv \eta \boldsymbol{\nabla}\left[L(\boldsymbol{\theta}^t) - L_{B_t}(\boldsymbol{\theta}^t)\right] \tag{8.6}$$

This noise term plays a pivotal role in mitigating the uncertainty introduced by approximating the true loss function with the mini-batch loss function during the training process, enhancing the robustness and adaptability of the neural network training algorithm [Adhikari et al., 2023].

# Chapter 9

# THE FOKKER-PLANCK EQUATION

## *9.1 Random Variables*

From an applications perspective, a "variable" refers to any entity that consistently possesses a measurable or samplable "value". The determination of a variable's value is typically contingent upon the specific context in which it is observed or measured. Typically, it is customary to have a clear definition of the sampling environment in order to ensure that the value acquired from each given sample is uniquely determined. In a broad sense, when variable X possesses the characteristic that its value is uniquely determined by its sampling context, it is referred to as a "sure variable". If the sampling context of a variable does not provide a unique determination of its value, it may be appropriate to refer to that variable as "random."

The concept of a random variable encompasses more than just the uncertainty of values obtained by contextual sampling. $X$ is considered a random variable if and only if there exists a function $P$ of a real variable $x$ such that $P(x)\Delta x$ is approximately equal to the probability of observing the value of $X$ within the interval $[x, x + \Delta x)$, up to first order in $\Delta x$. In the small limits of $\Delta x$, the probability will be $P(x)dx$. Thus, $P(x)$ which is the density of probability of finding variable $X$ at the value $x$, is called as "density function" of $X$. (ref 1991, Gillepse)

It is assumed that the value of X at an initial time t0 is constant, denoted as X(t0) = x0. However, the prediction of X at any time t ¿ t0 is subject to probabilistic uncertainty. In other words, X(t) for a given t ¿ t0 is considered a random variable. For every sequence of distinct time points $t_0, t_1, t_2, ..., t_n$ where $t_0 < t_1 < t_2 < ... < t_n$, it is possible to establish a joint density function for $n$ random variables $X(t_0), X(t_1), ..., X(t_n)$. This joint density function may be created for each $i = 1, 2, ..., n$.

The expression $P(x_n, t_n; x_{n-1}, t_{n-1}, ...; x_1, t_1|x_0, t_0)dx_n dx_{n-1}...dx_1$ represents the joint probability of the random variable $X(t_i)$ falling within the interval $[x_i, x_i+dx_i)$, given the initial condition $X(t_0) = x_0$. Therefore, it can be concluded that $X(t)$ is a stochastic process.

Hence, it is possible to regard a stochastic process, denoted as $X$, at two distinct time points, $X(t_1)$ and $X(t_2)$, as two distinct random variables. The term *continuous Markov process* is used to describe a stochastic process $X(t)$ that satisfies the following conditions:

i) The infinitesimal change in the variable $X$ between time $t$ and $t + dt$ is solely determined by the values of $t$, $dt$, and $X(t)$. Hence, the "conditional increment" in variable $X$ can be delineated as follows:

$\Xi(dt; x, t) = X(t + dt) - X(t)$

In which $x$ is $X(t)$ and $\Xi$ is a function of $t$, $dt$, and $x$. Therefore, the $\Xi$ exhibits the property of being "memoryless".

ii) $X$ is continuous and $\Xi(dt; x, t)$ goes to 0 when $dt \to 0$ for all $x$ and $t$.

### *9.2 Kramers-Moyal Equations*

To derive the Kramers-Moyal equation, we need to define a function as below.

$$f(x) \equiv P(x + \xi, t + \Delta t|x, t)P(x, t|x_0, t_0) \tag{9.1}$$

In which $x$ and $\xi$ are Markov processes. Using Tayor expansion, we can write:

$$f(x - \xi) = f(x) + \sum_{n=1}^{\infty} \frac{(-\xi)^n}{n!} \frac{\partial^n}{\partial x^n} f(x) \tag{9.2}$$

It can be shown that:

$$\begin{aligned} &P(x, t + \Delta t|x_0, t_0) \\ &= \int_{-\infty}^{\infty} d\xi P(x + \xi, t + \Delta t|x, t)P(x, t|x_0, t_0) \\ &+ \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n} \left[ \int_{-\infty}^{\infty} d\xi P(x + \xi, t + \Delta t|x, t)P(x, t|x_0, t_0) \right] \end{aligned} \tag{9.3}$$

Thus, the derivative of $P$ is derived simply.

$$\frac{P(x,t+\Delta t|x_0,t_0)-P(x,t|x_0,t_0)}{\Delta t}$$

$$=\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial x^n}\left[\left(\frac{1}{\Delta t}\int_{-\infty}^{\infty}d\xi\xi^n P(x+\xi,t+\Delta t|x,t)\right)P(x,t|x_0,t_0)\right] \tag{9.4}$$

At the limit $\Delta t \to 0$,

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty}d\xi\xi^n P(x+\xi,t+\Delta t|x,t)$$

$$=\frac{1}{dt}\int_{-\infty}^{\infty}d\xi\xi^n P(x+\xi,t+dt|x,t)$$

$$\equiv B_n(x,t) \tag{9.5}$$

Therefore, it results in:

$$\frac{\partial}{\partial t}P(x,t|x,t)=\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial x^n}\left[B_n(x,t)P(x,t|x,t)\right]. \tag{9.6}$$

This is known as the (forward) Kramers-Moyal equation [Gillespie, 1996].

### *9.3 Langevin Equation*

As it is said before, each $X(t_i)$ is a stochastic process, therefore, we can say that each $\Xi_i(dt/n;X(t_{i-1}),t_{i-1})$ is a Markov process. Hence, we can write the $\Xi(dt;X(t),t)$ as a summation of all of these processes.

$$\Xi(dt;X(t),t)=\sum_{i=1}^{n}\Xi_i(dt/n;X(t_{i-1}),t_{i-1}) \tag{9.7}$$

or

$$\Xi(dt;x,t)=\sum_{i=1}^{n}\Xi_i(dt/n;x,t) \tag{9.8}$$

Consequently, we can find the mean and the variance below.

$$mean\{\Xi(dt;x,t)\}=\sum_{i=1}^{n}mean\{\Xi_i(dt/n;x,t)\}]$$

$$var\{\Xi(dt;x,t)\}=\sum_{i=1}^{n}var\{\Xi_i(dt/n;x,t)\}]$$

Since all $\Xi_i(dt/n; x, t)$s have the same density function, all means and variances are the same.

$$mean\{\Xi(dt; x, t)\} = n.mean\{\Xi_i(dt/n; x, t)\}]$$

$$var\{\Xi(dt; x, t)\} = n.var\{\Xi_i(dt/n; x, t)\}]$$

Due to the above equations, it is simple to show that the mean and the variance can be written by using independent functions as follows.

$$mean\{\Xi(dt; x, t)\} = A(x, t)dt$$

$$var\{\Xi(dt; x, t)\} = D(x, t)dt$$

It is observed that the random variable $\Xi(dt/n; x, t)$ may be represented as the summation of $n$ statistically independent and identically distributed random variables. Each of these random variables has a clearly defined mean, denoted as $A(x, t)(dt/n)$, and a clearly defined variance, denoted as $D(x, t)(dt/n)$. The central limit theorem states that as the number of summands, denoted as $n$, increases to a sufficiently big value, the sum of these summands may be approximated to a normal random variable with increasing accuracy. However, it is worth noting that the variable $n$ can be chosen to have arbitrarily high values. As a result, it follows that the summation $\Xi(dt; x, t)$ must precisely correspond to a normal distribution.

$$\Xi(dt; x, t) = N(A(x, t)dt, D(x, t)dt) \tag{9.9}$$

Besides, it is simple to show that for a normal distribution, $N = N(0, 1)$, and two variables, $a$ and $d$, independent from N, we have:

$$Nd^{1/2} + a = N(d^{1/2}.0 + a, (d^{1/2})^2.1) = N(a, d) \tag{9.10}$$

As a result,

$$\Xi(dt; x, t) = N[D(x, t)dt]^{1/2} + A(x, t)dt \tag{9.11}$$

Which is the other form of:

$$X(t + dt) = X(t) + A(X(t), t)dt + D^{1/2}(X(t), t)N(t)(dt)^{1/2} \tag{9.12}$$

This is the standard form of the Langevin equation [**?**].

### *9.4 Fokker-Planck Equation*

In the preceding section, it was demonstrated that in the context of a continuous Markov process denoted as $X(t)$, characterized by the functions $A(x,t)$ and $D(x,t)$, the propagator $\Xi(dt;x,t)$ follows a normal distribution with a mean of $A(x,t)dt$ and a variance of $D(x,t)dt$. As is commonly understood, when considering a random variable $Y$ that follows a normal distribution $P(x)$ with a mean of $m$ and a variance of $a^2$, the $n^{th}$ moment can be expressed in the following manner.

$$\langle Y^n \rangle = n! \sum_{k=0 \ (k \ even)}^{n} \frac{m^{n-k}(a^2)^{k/2}}{(n-k)!(k/2)!2^{k/2}} \tag{9.13}$$

By utilizing this relationship, it is feasible to deduce the $n^{th}$ moment of $\Xi(dt;x,t)$.

$$\langle \Xi^n(dt;x,t) \rangle = n! \sum_{k=0 \ (k \ even)}^{n} \frac{[A(x,t)]^{n-k}[D(x,t)]^{k/2}dt^{n-k/2})}{(n-k)!(k/2)!2^{k/2}} \tag{9.14}$$

When, $n = 1, 2$,

$$\langle \Xi(dt;x,t) \rangle = A(x,t)dt$$

$$\langle \Xi^2(dt;x,t) \rangle = D(x,t)dt$$

According to the Eq. (9.5),

$$\langle \Xi(dt;x,t) \rangle = A(x,t)dt = B_1(x,t)dt$$

$$\langle \Xi^2(dt;x,t) \rangle = D(x,t)dt = B_2(x,t)dt$$

Finally, the Eq. (9.6) leads us to

$$\frac{\partial}{\partial t}P(x,t|x_0,t_0) = -\frac{\partial}{\partial x}[A(x,t)P(x,t|x_0,t_0)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[D(x,t)P(x,t|x_0,t_0)] \tag{9.15}$$

This is the (forward) Fokker-Planck equation (FPE). The other form of FPE can be written by defining a probability current function, $J(x,t;x_0,t_0)$.

$$J(x,t;x_0,t_0) = A(x,t)P(x,t|x_0,t_0) - \frac{1}{2}\frac{\partial}{\partial x}[D(x,t)P(x,t|x_0,t_0)] \tag{9.16}$$

Which changes the FPE to the form below.

$$\frac{\partial}{\partial t}P(x,t|x_0,t_0) = -\frac{\partial}{\partial x}J(x,t;x_0,t_0) \tag{9.17}$$

The given expression exhibits the structure of a continuous equation governing the probability density function $P(x,t|x_0,t_0)$, where the function $J(x,t;x_0,t_0)$ assumes the function of a probability current. When the equation is integrated across the x-interval $[a,b]$, the result is obtained.

$$\frac{\partial}{\partial t}\left[\int_a^b dx P(x,t|x_0,t_0)\right] = J(a,t;x_0,t_0) - J(b,t;x_0,t_0) \tag{9.18}$$

Based on the given assertion, the time derivative of the probability function $Prob\{X(t) \in [a,b)\}$ can be expressed as the discrepancy between the influx rate of the "occupation probability" into the interval $[a,b)$ at $x = a$ and the outflow rate of the same probability from the interval $[a,b)$ at $x = b$. The phenomenon elucidated by FPE is formally referred to as the diffusion process in mathematical terms. The vector $A(x,t)$ is commonly referred to as the drift vector, whereas the matrix $B(x,t)$ is commonly known as the diffusion matrix. By definition, the diffusion matrix, as represented by Eq. (9.5), is both positive semidefinite and symmetric.

The majority of continuous Markov processes that are of practical significance exhibit temporal homogeneity, thus, $A(x,t) = A(x)$ and $D(x,t) = D(x)$, and FPE is written as below.

$$\frac{\partial}{\partial t}P(x,t|x_0,t_0) = -A(x)\frac{\partial}{\partial x}P(x,t|x_0,t_0) + \frac{1}{2}D(x)\frac{\partial^2}{\partial x^2}P(x,t|x_0,t_0) \tag{9.19}$$

### *9.4.1 A Precise Description*

The concept of FPE entails a phenomenon wherein the stochastic process $X(t)$ exhibits a continuous sample path. In order to provide further clarification, we will proceed with the computation of the probability $P$ under the condition that $P(z,t|y,t) = \delta(z-y)$. Therefore,

$$\frac{\partial}{\partial t}P(z,t|y,t') = -A(z)\frac{\partial}{\partial z}P(z,t|y,t') + \frac{1}{2}D(z)\frac{\partial^2}{\partial z^2}P(z,t|y,t') \tag{9.20}$$

The equation's solution, when subjected to the initial condition, yields the following result.

$$P(z, t + \Delta t|y, t) = (2\pi)^{-N/2} \left(\det[D(y)]\right)^{1/2}$$
$$\times \exp\left(-\frac{1}{2}\frac{[z - y - A(y)\Delta t]^T [D(y)]^{-1} [z - y - A(y)\Delta t]}{\Delta t}\right) \tag{9.21}$$

The probability distribution being referred to is the Gaussian distribution, characterized by a variance matrix denoted as $D(y)$ and a mean value of $y + A(y)\Delta t$. The aforementioned relationship illustrates a system that is in motion with a drift, characterized by a velocity denoted as $A(y)$. Additionally, this system experiences a Gaussian fluctuation, which is superimposed upon the drift. The fluctuation is described by a covariance matrix denoted as $D(y)\Delta t$. Therefore, it is possible to compose

$$y(t + \Delta t) = y(t) + A(y(t))\Delta t + \beta(t)(\Delta t)^{1/2} \tag{9.22}$$

In which, $\langle \beta(t) \rangle = 0$, and $\langle \beta(t)\beta(t)^T \rangle = D(y)$.( [Gillespie, 1991, Srkk and Solin, 2019])

# Chapter 10

# EARTHQUAKE MODEL

Following our discussion on Stochastic Gradient Descent (SGD) in Neural Networks (NN), as well as the utilization of the Fokker-Planck equation (FPE) as a differential equation to elucidate stochastic processes, we shall establish a comparison between the Langevin equation, denoted as Eq. (9.12), and the evolution equation of SGD, represented as Eq. (8.5). The similarity between the stochastic gradient descent (SGD) evolution equation and the Langevin equation becomes apparent when we treat the time increment $dt$ as a unit value. Therefore, the change in the loss function $\Delta L(\theta)$ exhibits behavior similar to that of a drift term, exerting a force on the system that opposes the direction of the gradient of $L(\theta)$. The diffusion matrix, denoted as $D(\theta)$, is formed through the construction of the matrix $\langle \xi \xi^T \rangle$. Therefore, we can use the results of FPE for SGD analysis.

## *10.1 Stationary States*

The provided system dynamics eventually converges to a stable state characterized by the weight distribution $P^s(\theta)$ following an initial period of transient behavior during which the process of "learning" takes place. For a sufficiently small value of $\eta$, the stationary state can be characterized by variations of magnitude $O(\eta)$ around a certain value $\theta_0$. This value represents a minimum of the effective loss function, which we define as the reference point. The Taylor expansion of the probability current around the point $\theta_0 \approx 0$ results in

$$J(\boldsymbol{\theta}) \approx -\eta P H_0 . \boldsymbol{\theta} - D_0 . \boldsymbol{\nabla} P \tag{10.1}$$

Next, we examine the temporal progression of the primary and secondary cumulants

of the probability distribution function $P(\boldsymbol{\theta}; t)$.

$$\bar{\boldsymbol{\theta}} \equiv \int d\boldsymbol{\theta}\boldsymbol{\theta} P(\boldsymbol{\theta}) \tag{10.2}$$

$$\Sigma_{\alpha\beta} \equiv \int d\boldsymbol{\theta}\theta_\alpha\theta_\beta P(\boldsymbol{\theta}) \tag{10.3}$$

These equations lead us to the following results:

$$\frac{d\bar{\boldsymbol{\theta}}}{dt} = -\eta H_0.\bar{\boldsymbol{\theta}} \tag{10.4}$$

$$\frac{d\Sigma}{dt} = -\eta(H_0\Sigma + \Sigma H_0) + 2D_0 \tag{10.5}$$

Assuming that the function $P(\boldsymbol{\theta})$ exhibits rapid decay. At the stationary state, it is observed that the value of the vector $\bar{\boldsymbol{\theta}} = 0$, and

$$\Sigma H_0 + H_0\Sigma = 2\eta^{-1}D_0 \tag{10.6}$$

The concept under consideration assumes the role of a broad and encompassing fluctuation. The dissipation theorem establishes a relationship between the size of stationary fluctuations, denoted by $\Sigma$, and the diffusion matrix $D_0$. (reference 28 version 1) The provided equation represents a comprehensive fluctuation-dissipation relation [28], which can be mathematically determined by solving for the variable $\Sigma$ in respect to the variables $H_0$ and $D_0$. The multivariate Gaussian distribution is the stationary distribution that satisfies the Fokker-Planck equation.

$$P^s(\boldsymbol{\theta}) = \frac{\exp[\boldsymbol{\theta}\Sigma^{-1}\boldsymbol{\theta}/2]}{\sqrt{(2\pi)^N \det(\Sigma)}} \tag{10.7}$$

### *10.2 Training Model*

During the process of training a particular dataset, a significant consequence arises. The observed outcome can be attributed to the iterative learning approach employed on the dataset, emphasizing the significant influence of the training procedure on the original data. The first example is a straightforward regression problem designed to minimize the dimension of the parameter vector $\boldsymbol{\theta}$. The objective is to reconstruct a Gaussian distribution curve using a limited number of samples that are subject to noise. The training dataset consists of 200 pairs of

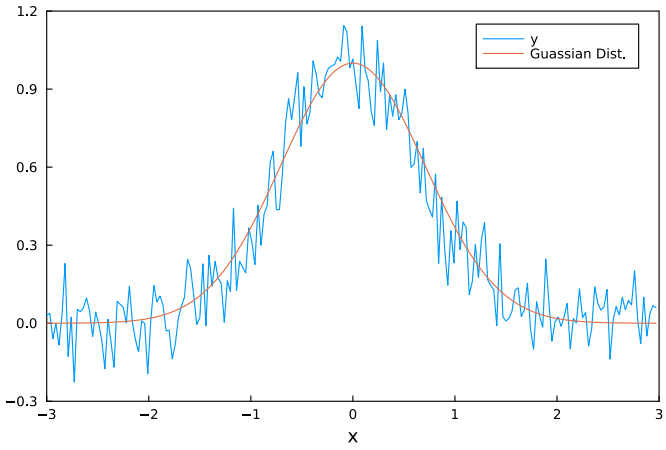


Figure 10.1: The training data, $x_{train}$ and $y_{train}$. The $y_{train}$s are generated from a noisy Gaussian distribution.

input-output values, denoted as $(x_i; y_i)$. The input values, denoted as $x_i$, are calculated using the formula $x_i = x_{min} + i \cdot dx$, where $x_{min}$ is equal to -3 and $dx$ is equal to 0.03. The corresponding output values, denoted as $y_i$, are calculated using the formula $y_i = \exp(-x_i^2) + r_i$, where $r_i = 0.1 \cdot N(0, 1)$. The neural network consists of a pair of units, each responsible for executing the computation $h_k(x_i) = \tanh(w_{k,1}x + w_{k,2})$ on the input variable $x$. The computed output, denoted as $\tilde{y}_i = w_4 + \sum_{k=1,2} w_{3,k}h_k(x_i)$. We initiate the process with a randomly selected set of weights denoted as $\boldsymbol{\theta}^T = [w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}, w_{3,1}, w_{3,2}w_4]$. To update these weights, we utilize the stochastic gradient descent (SGD) technique with replacement, employing a minibatch size of $m = 10$. The sample loss function has been established as

$$l(\boldsymbol{\theta}; x_i, y_i) = (\tilde{y}_i(\boldsymbol{\theta}; x_i) - y_i)^2 \tag{10.8}$$

and a regularization $R(\boldsymbol{\theta}) = 10^{-3}||\boldsymbol{\theta}||^2$.
The training data is derived from a Gaussian distribution with noise, as depicted in Fig. (10.2).

The training process by using these training data leads us to a stationary state.

### *10.3 Model*

In this particular section, we construct a theoretical framework that facilitates the modelling of stationary states in the context of stochastic gradient descent (SGD). In order to do this, we formulate the Earthquake model by introducing a noise component into the stochastic gradient descent (SGD) evolution equation, as opposed to the noise arising from minibatching in Equation (8.5). The noise-generating distribution has the ability to influence the characteristics of the stationary state.

### *10.4 The Most General Transformation*

Infinitesimal transformations, including translation, rotation, and scaling, can be expressed as a sum that encompasses all characteristics of any transformation. The original coordinate system is denoted by $\theta$s, while the transformed coordinate system is represented as $u$s.

$$\boldsymbol{u} = \boldsymbol{\theta} + \boldsymbol{\zeta}(\boldsymbol{\theta}, \{\psi_i\}) \tag{10.9}$$

In which, $\boldsymbol{\zeta}$ is a function of $\theta_i$s and a bunch of random variables, $\psi_i$s. The actual loss function at the point $\boldsymbol{\theta}$ is $L(\boldsymbol{\theta})$.

$$L(\boldsymbol{\theta}) = \sum_j \frac{h_j}{2}\theta_j^2 \tag{10.10}$$

$h_i$s are constants and properties of the loss function which build the Hessian matrix, shown by $\boldsymbol{H}$, and $\boldsymbol{H}_{ij} = h_i\delta_{ij}$, as one key feature of the loss landscape. After transformation, the loss function in the transformed coordinate system would be written as $L(\boldsymbol{u})$ which has a new functional form in terms of $\boldsymbol{\theta}$ like $L'(\boldsymbol{\theta})$.

$$L'(\boldsymbol{\theta}) = \sum_j \frac{h_j}{2}[\theta_j + \zeta_j(\boldsymbol{\theta}, \{\psi_i\})]^2 \tag{10.11}$$

Based on the provided definition, computing the noise term is a straightforward process.

$$\boldsymbol{\xi}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \eta\left(\boldsymbol{\nabla}L'(\boldsymbol{\theta}) - \boldsymbol{\nabla}L(\boldsymbol{\theta})\right) \tag{10.12}$$

In other words,

$$\xi_j(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \eta h_j\left([\theta_j + \zeta_j(\boldsymbol{\theta}, \{\psi_i\})]\left[1 + \frac{\partial\zeta_j(\boldsymbol{\theta}, \{\psi_i\})}{\partial\theta_j}\right] - \theta_j\right) \tag{10.13}$$

### *10.4.1 The simplest case*

As efforts are always towards more simplified models, the first simplification to this model can be the assumption that $\boldsymbol{\zeta}$ is not a function of $\boldsymbol{\theta}$. Accordingly, the noise term (Eq. 10.13) will be more simplified.

$$\xi_j(\boldsymbol{\zeta}) = \eta h_j \zeta_j(\{\psi_i\}) \tag{10.14}$$

The next simplification is the assumption that $\boldsymbol{\zeta}$ is a stochastic parameter and does not depend on any other random parameter besides $\boldsymbol{\psi}$. Consequently, we can derive the general form for the noise term.

$$\boldsymbol{\xi}(\boldsymbol{\zeta}) = \eta \boldsymbol{H} \boldsymbol{\zeta} \tag{10.15}$$

## *10.5 Diffusion Matrix*

According to the definition, it is possible to derive the diffusion matrix for the introduced noise term.

$$\boldsymbol{D}_{ij} \equiv \frac{1}{2}\langle \xi_i(\boldsymbol{\zeta})\xi_j(\boldsymbol{\zeta})\rangle \tag{10.16}$$

$$= \frac{\eta^2}{2} h_i \langle \zeta_i \zeta_j \rangle h_j \tag{10.17}$$

As evident, the calculation involves the two-point correlation function of $\zeta_i$, denoted as $\langle \zeta_i \zeta_j \rangle$. This correlation function essentially represents the covariance matrix of the distribution of $\zeta_i$. Hence, the covariance matrix, denoted as $\boldsymbol{\Omega}$, is defined as follows.

$$\boldsymbol{\Omega}_{ij} \equiv \langle \zeta_i \zeta_j \rangle \tag{10.18}$$

Hence, the diffusion matrix is represented as a function of $\boldsymbol{\Omega}$ and $\boldsymbol{H}$.

$$\boldsymbol{D} = \frac{\eta^2}{2} \boldsymbol{H} \boldsymbol{\Omega} \boldsymbol{H} \tag{10.19}$$

## *10.6 Stationary State Covariance Matrix*

Our focus lies in understanding the distribution of $\theta_i$ values in the stationary state. To do this, we need to calculate the covariance matrix of the stationary state, denoted as $\boldsymbol{\Sigma}$. This distribution function can be determined by utilizing the diffusion

matrix, where $\boldsymbol{\Sigma}ij = 2\eta^{-1}(h_i + h_j)^{-1}\boldsymbol{D}ij$. Therefore, deriving the general form of the stationary state covariance matrix is a straightforward process.

$$\boldsymbol{\Sigma}_{ij} = \eta \frac{h_i h_j}{h_i + h_j} \boldsymbol{\Omega}_{ij} \tag{10.20}$$

### *10.7 Probability Current*

The probability current, represented by $\boldsymbol{J}$, is defined as $\boldsymbol{J} \equiv \boldsymbol{\Sigma H} - \boldsymbol{H\Sigma}$. In this model, finding the matrix for the probability current can be done quite easily.

$$\boldsymbol{J}_{ij} = -\eta \frac{h_i h_j}{h_i + h_j} (h_i - h_j) \boldsymbol{\Omega}_{ij} \tag{10.21}$$

# Chapter 11

# CONCLUSION

## *11.1 Numerical Results*

Several crucial criteria must be met by our theoretical model. The initial requirement is that our Earthquake model must be capable of replicating the numerical outcomes observed during the training phase. To put it differently, the Earthquake model should yield an identical distribution for the stationary state as that achieved through numerical training. The numerical training process reveals that the stationary state should be characterized by a specific seven-dimensional covariance matrix and mean point. The Earthquake model has the capacity to reproduce this identical distribution for the stationary state. It's worth noting that this is achieved by incorporating a noise source, which means that the distribution of the noise source can predict the stationary state distribution. In this particular case, our focus is narrowed down to just two dimensions. The numerical results from the training process indicate that the covariance matrix of the stationary state is as follows:

$$\Sigma_{ss} = \begin{bmatrix} 16.4 & 6.57 \\ 6.56 & 17.6 \end{bmatrix} \times 10^{-11}$$

The mean value for these dimensions is:

$$\mu_{ss} = \text{-} \begin{bmatrix} 9.09 \times 10^{-1} \\ 8.83 \times 10^{-3} \end{bmatrix}$$

The Hessian matrix at the minimum of the loss function is:

$$H_{ss} = \begin{bmatrix} 1.42 \times 10^{4} & 3.69 \times 10^{3} \\ 3.69 \times 10^{3} & 2.00 \times 10^{4} \end{bmatrix}$$

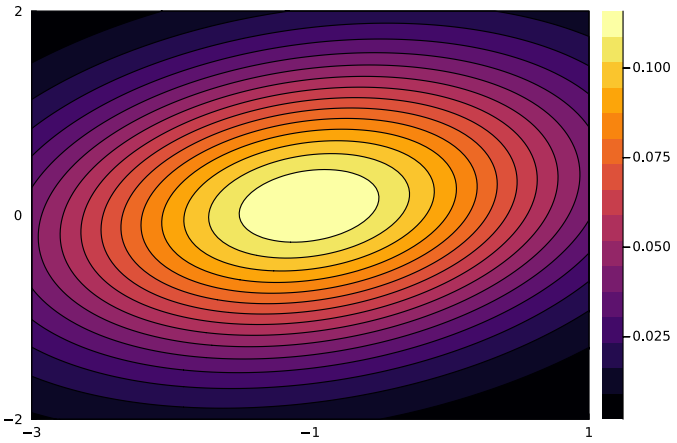


Figure 11.1: This stationary state in the eigen space can be visualized in Figure (11.1). The stationary state distribution is characterized by a covariance matrix, $\Sigma^E_{ss} = [11.4, 3.55, 3.55, 22.6] \times 10^{-11}$, and a mean value at the origin of the coordinate system, $[-908, -8.84] \times 10^{-3}$. The source pertains to the 6th and 7th dimensions.

At this point, with this information in hand, we can configure the noise source to replicate the same numerical results. By transforming the Hessian matrix into its eigen space, we have:

$$H^E_{ss} = \begin{bmatrix} 1.24 & 0 \\ 0 & 2.18 \end{bmatrix} \times 10^4$$

Now, we must transition everything into the eigen space of the Hessian matrix, resulting in the covariance matrix for the stationary state:

$$\Sigma^E_{ss} = \begin{bmatrix} 11.4 & 3.55 \\ 3.55 & 22.6 \end{bmatrix} \times 10^{-11}$$

The stationary state in the eigen space can be visualized as in Figure (11.1). This leads to a source noise covariance matrix as follows:

$$\Omega = \begin{bmatrix} 18.4 & 4.49 \\ 4.49 & 20.6 \end{bmatrix} \times 10^{-15}$$

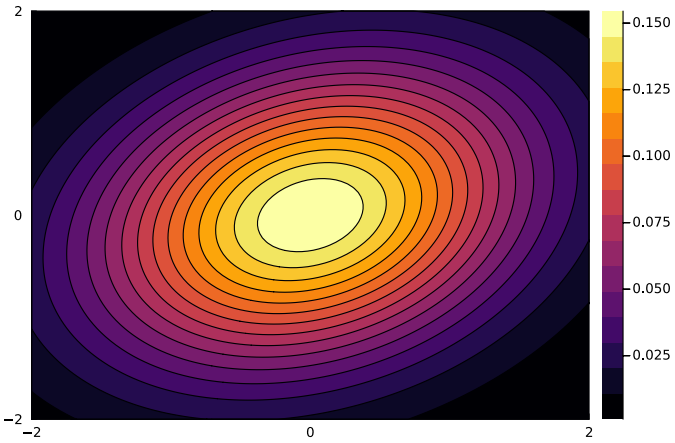


Figure 11.2: The mean of the distribution remains at the origin. Consequently, we can visualize the source noise as shown in Figure (11.2). The source noise distribution is characterized by a covariance matrix, $\Omega = [18.4, 4.49, 4.49, 20.6] \times 10^{-15}$, with a mean value at the origin of the coordinate system. The source pertains to the 6th and 7th dimensions.

The mean of the distribution is the origin. Thus, we can plot the source noise as in the Figure (11.2). By incorporating such a noise distribution into our model, we can accurately replicate the stationary state distribution. Furthermore, it becomes feasible to generate the diffusion matrix and probability current matrix. The quantities mentioned for the numerical case are as follows:

$$J = \begin{bmatrix} 0 & -3.33 \\ 3.33 & 0 \end{bmatrix} \times 10^{-14}$$

The true probability current derived from the numerical code is as follows:

$$J_{ss} = \begin{bmatrix} 0 & -3.33 \\ 3.33 & 0 \end{bmatrix} \times 10^{-14}$$

# Chapter 12

# FUTURE WORK

There are several aspects that can be addressed as a consequence of the earthquake model. The primary objective of this study is to develop a straightforward model for the analysis of the machine learning training process. The results of this model have the potential to reduce the overall cost of the training process or facilitate the engineering of the stationary state. In the model, we established a comparison between the training process conducted without the use of minibatching and the training process conducted with minibatching. To clarify, the act of introducing noise can be likened to the process of selecting a minibatch. Hence, by integrating the processes of noise addition and minibatching, we may examine the circumstances in which noise can induce stationary states characterized by Boltzmann equilibrium.

# BIBLIOGRAPHY

[Adhikari et al., 2023] Adhikari, S., Kabakçioglu, A., Strang, A., Yuret, D., and Hinczewski, M. (2023). Machine learning in and out of equilibrium. *arXive*.

[Aerts et al., 2013] Aerts, D., Gabora, L., and Sozzo, S. (2013). Concepts and their dynamics: A quantum-theoretic modeling of human thought. *Top. Cogn. Sci.*, 5:737–772.

[Barrett, 2007] Barrett, J. (2007). Information processing in generalized probabilistic theories. *Phys. Rev. A*, 75:032304.

[Basieva et al., 2021] Basieva, I., Khrennikov, A., and Ozawa, M. (2021). Quantum-like modeling in biology with open quantum systems and instruments. *Biosystems*, 201:104328.

[Baumgratz et al., 2014] Baumgratz, T., Cramer, M., and Plenio, M. B. (2014). Quantifying coherence. *Phys. Rev. Lett.*, 113:140401.

[Bell, 1964] Bell, J. S. (1964). On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1:195–200.

[Bennett et al., 1996a] Bennett, C. H., Bernstein, H. J., Popescu, S., and Schumacher, B. (1996a). Concentrating partial entanglement by local operations. *Phys. Rev. A*, 53:2046–2052.

[Bennett et al., 1996b] Bennett, C. H., DiVincenzo, D. P., Smolin, J. A., and Wootters, W. K. (1996b). Mixed-state entanglement and quantum error correction. *Phys. Rev. A*, 54:3824–3851.

[Best and Hummer, 2011] Best, R. B. and Hummer, G. (2011). Diffusion models of protein folding. *Physical Chemistry Chemical Physics*, 13:16902–16911.

[Bryngelson et al., 1995] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Genetics*, 21:167–195.

[Busemeyer and Wang, 2017] Busemeyer, J. and Wang, Z. (2017). Is there a problem with quantum models of psychological measurements? *PLoS One*, 8:12.

[Busemeyer et al., 2006] Busemeyer, J. R., Wang, Z., and Townsend, J. T. (2006). Quantum dynamics of human decision-making. *J. Math. Psychol.*, 50(3):220–241. Special Issue: Jean-Claude Falmagne: Part II.

[Carleo and et al., 2019] Carleo, G. and et al. (2019). Machine learning and the physical sciences. *arXiv preprint*.

[Chaudhari and Soatto, 2018] Chaudhari, P. and Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE.

[Chiribella et al., 2011] Chiribella, G., D'Ariano, G. M., and Perinotti, P. (2011). Informational derivation of quantum theory. *Phys. Rev. A*, 84:012311.

[Cirel'son, 1980] Cirel'son, B. S. (1980). Quantum generalizations of bell's inequality. *Lett. Math. Phys.*, 4:93–100.

[Clauser et al., 1969] Clauser, J. F., Horne, M. A., Shimony, A., and Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884.

[Croson, 1999] Croson, R. T. (1999). The disjunction effect and reason-based choice in games. *Organ. Behav. Hum. Decis. Process.*, 80(2):118–133.

[Dakić et al., 2014] Dakić, B., Paterek, T., and Č. Brukner (2014). Density cubes and higher-order interference theories. *New J. Phys.*, 16:023028.

[Dill et al., 2017] Dill, K., Jernigan, R. L., and Bahar, I. (2017). *Protein Actions: Principles and Modeling*. Garland Science.

[Gillespie, 1991] Gillespie, D. T. (1991). Markov processes: An introduction for physical scientists.

[Gillespie, 1996] Gillespie, D. T. (1996). The mathematics of Brownian motion and Johnson noise. *American Journal of Physics*, 64(3):225–240.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[Hansen et al., 1993] Hansen, L. K., Pathria, R., and Salamon, P. (1993). Stochastic dynamics of supervised learning. *Journal of Physics A: Mathematical and General*, 26:63.

[Hardy, 2001] Hardy, L. (2001). "Quantum theory from five reasonable axioms".

[Hofrichter et al., 2017] Hofrichter, J., Jost, J., and Tran, T. D. (2017). *Information Geometry and Population Genetics*, volume 62. Springer.

[Khalfin and Tsirelson, 1992] Khalfin, L. A. and Tsirelson, B. S. (1992). Quantum/classical correspondence in the light of bell's inequalities. *Found. Phys.*, 22:879–948.

[Khrennikov et al., 2018] Khrennikov, A., Basieva, J. I., Pothos, E. M., and Yamoto, I. (2018). Quantum probability in decision making from quantum information representation of neuronal states. *Sci. Rep.*, 8:16225.

[Khrennikov and Haven, 2009] Khrennikov, A. Y. and Haven, E. (2009). Quantum mechanics and violations of the sure-thing principle: The use of probability interference and other concepts. *J. Math. Psychol.*, 53(5):378–388. Special Issue: Quantum Cognition.

[Kimura and et al., 1954] Kimura, M. and et al. (1954). *Stochastic Processes and Distribution of Gene Frequencies under Natural Selection*. Citeseer.

[Li et al., 2018] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. pages 6389–6399.

[Li and Taplin, 2002] Li, S. and Taplin, J. (2002). Examining whether there is a disjunction effect in prisoners dilemma games. *Chinese J. Psychol.*, 44:2546.

[Masanes and Mller, 2011] Masanes, L. and Mller, M. P. (2011). A derivation of quantum theory from physical requirements. *New J. Phys.*, 13(6):063001.

[Mehta and et al., 2019] Mehta, P. and et al. (2019). A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*.

[Mustonen and Lssig, 2010] Mustonen, V. and Lssig, M. (2010). Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences*, 107:4248–4253.

[Mller, 2021] Mller, M. P. (2021). Probabilistic theories and reconstructions of quantum theory. *SciPost Phys. Lect. Notes*, page 28.

[Neumann, 1927] Neumann, J. v. (1927). Thermodynamik quantenmechanischer gesamtheiten. *Nachrichten von der Gesellschaft der Wissenschaften zu Gttingen, Mathematisch-Physikalische Klasse*, 1927:273–291.

[Paterek et al., 2010] Paterek, T., Dakić, B., and Č Brukner (2010). Theories of systems with limited information content. *New J. Phys.*, 12(5):053037.

[Pearl, 2016] Pearl, J. (2016). The sure-thing principle. *J. Causal Inference*, 4(1):81–86.

[Popescu, 2014] Popescu, S. (2014). Nonlocality beyond quantum mechanics. *Nature Phys.*, 10:264–270.

[Popescu and Rohrlich, 1994] Popescu, S. and Rohrlich, D. (1994). Quantum nonlocality as an axiom. *Found. Phys.*, 24:379.

[Popescu and Rohrlich, 1997] Popescu, S. and Rohrlich, D. (1997). Thermodynamics and the measure of entanglement. *Phys. Rev. A*, 56:R3319–R3321.

[Pothos and Busemeyer, 2009] Pothos, E. M. and Busemeyer, J. R. (2009). A quantum probability explanation for violation of 'rational' decision theory. *Proc. R. Soc. B*, 276:2171.

[Pothos and Busemeyer, 2022] Pothos, E. M. and Busemeyer, J. R. (2022). Quantum cognition. *Annu. Rev. Psychol.*, 73:749–778.

[Radons et al., 1990] Radons, G., Schuster, H. G., and Werner, D. (1990). Fokker-planck description of learning in backpropagation networks. In *International Neural Network Conference*, pages 993–996. Springer.

[Rau, 2011] Rau, J. (2011). Measurement-based quantum foundations. *Found Phys.*, 41:380–388.

[Risken, 1996] Risken, H. (1996). *The Fokker-Planck Equation*. Springer.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

[Savage, 1954] Savage, J. (1954). The foundations of statistics. john wiley & sons, inc., 1954, 294 pp. *Naval Research Logistics Quarterly*, 1(3):236–236.

[Schumacher, 1995] Schumacher, B. (1995). Quantum coding. *Phys. Rev. A*, 51:2738–2747.

[Seifert, 2012] Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems, and molecular machines. *Reports on Progress in Physics*, 75:126001.

[Shafir and Tversky, 1992] Shafir, E. and Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cogn. Psychol.*, 24(4):449–474.

[Shannon, 1948a] Shannon, C. E. (1948a). A mathematical theory of communication. *Bell Labs Tech. J.*, 27(3):379–423.

[Shannon, 1948b] Shannon, C. E. (1948b). A mathematical theory of communication. *Bell Labs Tech. J.*, 27(4):623–656.

[Sorkin, 1994] Sorkin, R. D. (1994). Quantum mechanics as quantum measure theory. *Mod. Phys. Lett. A*, 9:3119–3127.

[Srkk and Solin, 2019] Srkk, S. and Solin, A. (2019). *Applied Stochastic Differential Equations*.

[Tesa, 2020] Tesa, J. (2020). A quantum model of strategic decision-making explains the disjunction effect in the prisoners dilemma game. *Decision*, 7:43–54.

[Tsirel'son, 1987] Tsirel'son, B. S. (1987). Quantum analogues of the bell inequalities. the case of two spatially separated domains. *J. Math. Sci.*, 36:557–570.

[Tversky and Kahneman, 1974] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

[Tversky and Shafir, 1992] Tversky, A. and Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychol. Sci.*, 3(5):305–310.

[Veitshans et al., 1997] Veitshans, T., Klimov, D., and Thirumalai, D. (1997). Protein folding kinetics: Timescales, pathways, and energy landscapes in terms of sequence-dependent properties. *Protein Science*, 2:1–22.

[Wang and Busemeyer, 2013] Wang, Z. and Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Top. Cogn. Sci.*, 5:689–710.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. pages 681–688.

[Yearsley and Busemeyer, 2016] Yearsley, J. M. and Busemeyer, J. R. (2016). Quantum cognition and decision theories: A tutorial. *J. Math. Psychol.*, 74:99–116.